ON THE LIMITS OF SPARSE LINEAR MODEL WITH SHUFFLED DIFFERENTIAL PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse linear regression is a fundamental problem in high-dimensional statistics, and its private counterparts leave a large gap between the central and the local model of differential privacy. Shuffled differential privacy (SDP) sits between the two, offering accuracy closer to the central model without the trust requirement of a central curator. This paper studies sparse linear regression under both non-interactive and interactive SDP protocols. In the non-interactive setting, any non-interactive (ϵ, δ)-SDP protocol is shown to incur an ℓ_2-norm estimation error of $\Omega(\frac{\sqrt{dk}}{n\epsilon})$ with dimension d, sparsity k, and n users, revealing an inherent dimensionality barrier. In addition, a non-interactive algorithm is proposed that achieves an ℓ_2-norm upper bound of $\tilde{\mathcal{O}}(\sqrt{\frac{k}{n}} + \frac{d\sqrt{k}}{n\epsilon})$. To overcome this barrier, we turn to the interactive user-level setting, where each user holds m samples. A two-round interactive algorithm is developed that first recovers the support through a heavy hitter step and then estimates the parameter, achieving an error of $\tilde{\mathcal{O}}(\sqrt{\frac{k}{nm}} + \frac{k}{n\sqrt{m}\epsilon})$ with no polynomial dependence on d. Furthermore, by extending the score attack framework to user-level data, a new lower bound $\Omega(\sqrt{\frac{k\log(d/k)}{nm}} + \frac{k\log(d/k)}{n\sqrt{m}\epsilon})$ is established for any user-level (ϵ, δ)-DP, showing that the proposed algorithm is near-optimal with only a gap of the factor $\log(d/k)$.

1 INTRODUCTION

Protecting data privacy has become a paramount concern in modern information and database systems, which often store highly sensitive personal records, and (sparse) linear models, a cornerstone of both statistics and machine learning, are routinely fitted to such data. Differential privacy (DP) has emerged as the gold standard for privacy-preserving data analysis (Dwork et al., 2006; Dwork & Roth, 2014), providing mathematically verifiable protection that remains resilient to attackers with arbitrary auxiliary information.

The central model of DP assumes the existence of a trusted curator who collects raw data from all individuals and then applies privacy-preserving mechanisms before releasing the results. However, in many distributed settings, users may not trust a central entity with their unencrypted sensitive information. To address this, local differential privacy (LDP) was introduced, shifting the privacy-preserving step to the user's side; each individual perturbs their own data locally before sending it to a server (Duchi et al., 2013; 2014; Kasiviswanathan et al., 2011). Despite this stronger privacy guarantee, a critical issue with LDP is its limited utility, as the significant amount of noise required to protect each individual entry often leads to much lower statistical accuracy compared to the central model. Shuffled differential privacy (SDP) has recently emerged as a promising middle ground that simultaneously addresses these concerns (Cheu et al., 2019; Erlingsson et al., 2019; Balle et al., 2019). By introducing an intermediate shuffler that de-identifies and permutes the local messages before they reach the server, SDP can provide accuracy levels approaching those of the central model while maintaining a distributed architecture that does not require a trusted curator.

Private linear regression has been studied both as stochastic optimization, which minimizes the empirical or population risk, and as statistical estimation, which recovers the parameter $\theta^* \in \mathbb{R}^d$ of the linear model $y = \langle x, \theta^* \rangle + \zeta$ under the ℓ_2-norm error (Kifer et al., 2012; Talwar et al., 2015; Cai et al., 2021). We take the estimation view and focus on the high-dimensional sparse model, where

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

θ^* is k-sparse with $k \ll d$ and d may far exceed the sample size n (Raskutti et al., 2011; Yang et al., 2014; Jain et al., 2014); here the main difficulty under privacy is to avoid an error that scales polynomially with d.

In recent years, there have been many studies of the DP sparse linear model under different privacy notations. Specifically, in the central DP model, the l_2-norm estimation error for k-sparse parameters scales logarithmically with the dimension d, achieving a nearly optimal rate of $\mathcal{O}(\sqrt{\frac{k\log d}{n}} + \frac{k\log d}{n\epsilon})$ (Cai et al., 2021). In sharp contrast, the local DP model imposes a much heavier privacy cost, where the estimation error inevitably suffers from a polynomial dependency on d; as summarized in Table 1, the improved analysis of sparse local-private regression shows that non-interactive and interactive LDP algorithms still obey strong dimension-dependent lower bounds (Zhu et al., 2024). This creates a massive gap that makes LDP algorithms essentially ill-suited for modern high-dimensional scenarios where $d \gg n$. However, on the other side, for SDP, recent studies show that for many problems, such as histogram estimation Balcer & Cheu (2019), their theoretical behaviors are more similar to the central DP, while for other problems, such as parity function learning Cheu & Ullman (2021), the theoretical behaviors are more similar to local DP. Given the large gap on the limitations of sparse linear model between the two models, we have a natural question: does sparse linear regression in the shuffled model bridge the gap toward the efficiency of central DP, or does it still reflect the dimensionality barriers of local DP?

Table 1: Comparison summary of our work with related studies on (ϵ, δ)-DP linear regression under sub-Gaussian data assumptions. Here n, m, k, and d denote the number of users, samples per user, sparsity, and dimension. Note that we can regard $m = 1$ when the privacy unit is *sample*. All bounds are measured by the ℓ_2-norm estimation error. The $\tilde{\mathcal{O}}$ notation suppresses polylogarithmic factors. The full table can be seen in the Appendix A.

Model	**Type**	**Unit**	**Setting**	**Upper Bound**	**Lower Bound**	**Source**
Central	-	sample	general	$\tilde{\mathcal{O}}(\sqrt{\frac{d}{n}} + \frac{d}{n\epsilon})$	$\Omega(\sqrt{\frac{d}{n}} + \frac{d}{n\epsilon})$	Cai et al. (2021)
		sample	k-sparse	$\tilde{\mathcal{O}}(\sqrt{\frac{k}{n}} + \frac{k}{n\epsilon})$	$\Omega(\sqrt{\frac{k\log d}{n}} + \frac{k\log d}{n\epsilon})$	Cai et al. (2021)
Local	Non-interactive	sample	k-sparse	$\tilde{\mathcal{O}}(\frac{d\sqrt{k}}{\sqrt{n}\epsilon})$	$\Omega(\sqrt{\frac{dk\log d}{n\epsilon^2}})$	Zhu et al. (2024)
	Interactive	sample	k-sparse	$\tilde{\mathcal{O}}(\frac{k\sqrt{d}}{\sqrt{n}\epsilon})$	$\Omega(\sqrt{\frac{dk}{n\epsilon^2}})$	Zhu et al. (2024)
		user-level	k-sparse	$\tilde{\mathcal{O}}(\sqrt{\frac{k}{nm}} + \frac{k}{\sqrt{nm}\epsilon})$	$\Omega(\frac{k}{\sqrt{nm}\epsilon})$	Ma et al. (2024)
Shuffle	Non-interactive	sample	k-sparse	$\tilde{\mathcal{O}}(\sqrt{\frac{k}{n}} + \frac{d\sqrt{k}}{n\epsilon})$	$\Omega(\frac{\sqrt{dk}}{n\epsilon})$	**Our Work**
	Interactive	user-level	k-sparse	$\tilde{\mathcal{O}}(\sqrt{\frac{k}{nm}} + \frac{k}{n\sqrt{m}\epsilon})$	$\Omega(\sqrt{\frac{k\log(d/k)}{nm}} + \frac{k\log(d/k)}{n\sqrt{m}\epsilon})$	

In this work, we provide a rigorous view of sparse linear regression under SDP. Our contributions are three-fold.

- We first study the non-interactive protocol. By converting robust (ϵ, δ)-SDP to pan-privacy, we establish a lower bound of $\Omega(\frac{\sqrt{dk}}{n\epsilon})$ on the estimation error measured by the ℓ_2-norm difference, revealing an inherent dimensionality barrier, which is more similar to LDP. We also propose an (ϵ, δ)-SDP algorithm, which achieves an upper bound of $\tilde{\mathcal{O}}(\sqrt{\frac{k}{n}} + \frac{d\sqrt{k}}{n\epsilon})$, characterizing the limits of non-interactive SDP.
- We then study how we can overcome the theoretical barrier with relaxed settings. Specifically, we consider the interactive and user-level SDP protocol, where each user has m samples. We develop a two-round interactive algorithm, U2R (User-level Two-Round Regression), that achieves an upper bound of $\tilde{\mathcal{O}}(\sqrt{\frac{k}{nm}} + \frac{k}{n\sqrt{m}\epsilon})$. Moreover, we develop a novel user-level score attack method and show a lower bound of $\Omega\Big(\sqrt{\frac{k\log(d/k)}{nm}} + \frac{k\log(d/k)}{n\sqrt{m}\epsilon}\Big)$ for

user-level (ϵ, δ)-DP. The two rates differ only by a factor of $\log(d/k)$, indicating that our upper bound is near-optimal and that, with interaction and user-level data, the problem is more closely related to central DP.

- We evaluate U2R on synthetic and real data. Across the privacy budgets ϵ, dimension d, and user counts n, it has the lowest coefficient error and MSE among the private methods.

2 Related Work

Private high-dimensional regression has been studied across three main privacy models. We provide a comprehensive review in Appendix A.

In the central model, sparse estimation typically incurs only a logarithmic dependence on the dimension d (Talwar et al., 2015; Cai et al., 2021), and the score attack of Cai et al. (2023) gives the lower-bound technique that we extend to the user-level regime. The local model is more restrictive, with both non-interactive and interactive estimators exhibiting polynomial dependence on d (Wang & Xu, 2020; Zhu et al., 2024); user-level multi-sample regimes alleviate these costs (Ma et al., 2024). Positioned between the two, the shuffled model leverages privacy amplification (Cheu et al., 2019; Erlingsson et al., 2019; Balle et al., 2019), while multi-message protocols compute a sum with central-model accuracy (Balle et al., 2020), which is the summation step both our algorithms use. On the negative side, Cheu & Ullman (2021) relate shuffled DP to pan-privacy, the route our non-interactive lower bound follows. Shuffled protocols have been studied for primitives such as heavy hitters (Bun et al., 2019), but their application to sparse linear regression remains largely open, which is the gap we fill.

3 Preliminaries

This section introduces the notations, the models, and the problem setting used throughout the paper.

Notations. For an integer d, let $[d] = \{1, \dots, d\}$ and let $\mathbb{R}^d$ be the d-dimensional Euclidean space. For a vector $v \in \mathbb{R}^d$, $\|v\|_p$ is the ℓ_p-norm, and $\|v\|_0$ is the number of non-zero entries. The support of θ is $\mathrm{supp}(\theta) = \{j \in [d] : \theta_j \neq 0\}$, and for a set $S \subseteq [d]$ we write $\theta_S \in \mathbb{R}^{|S|}$ for the restriction of θ to the coordinates in S. Given a matrix $X \in \mathbb{R}^{n \times d}$, let $x_i^\top$ be its i-th row and x_{ij} (or $[X]_{ij}$) be its (i, j)-th entry, which is also the j-th element of x_i. We use $\|X\|_2$ for the spectral norm and $\|X\|_{\infty,\infty} = \max_{i,j} |x_{ij}|$ for the max norm, and $\lambda_{\min}(X)$ for the minimal eigenvalue of X. For an event A, $\mathbb{I}[A]$ denotes the indicator, i.e., $\mathbb{I}[A] = 1$ if A occurs and $\mathbb{I}[A] = 0$ otherwise. A zero-mean random variable X is sub-Gaussian if its sub-Gaussian norm $\|X\|_{\psi_2} = \inf\{c > 0 : \mathbb{E}[\exp(X^2/c^2)] \leq 2\}$ is bounded; a random vector x is sub-Gaussian if $\langle x, u\rangle$ is sub-Gaussian for every unit vector u. We use $O(\cdot)$ and $\Omega(\cdot)$ in the standard sense, and $\widetilde{O}(\cdot)$ hides logarithmic factors.

3.1 Differential Privacy

Definition 1 (Differential Privacy (Dwork et al., 2006)). *A randomized algorithm $\mathcal{M}$ with domain $\mathcal{D}^n$ is (ϵ, δ)-differentially private (DP) if for all measurable sets $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{M})$ and for any two neighboring datasets $D, D' \in \mathcal{D}^n$ that differ in exactly one record, it holds that*

$$\mathbb{P}[\mathcal{M}(D) \in \mathcal{S}] \leq e^\epsilon \cdot \mathbb{P}[\mathcal{M}(D') \in \mathcal{S}] + \delta. \tag{1}$$

Definition 1 is stated with respect to a neighboring relation on $\mathcal{D}^n$. In the sample-level regime a record is a single pair (x_i, y_i), so D and D' differ in one sample. In the user-level regime a record is the whole local dataset U_i of user i, so D and D' differ in all m samples held by one user. All our user-level theorems use the latter relation.

The central model assumes the curator sees the raw data. The shuffled model removes this assumption, and instead places a trusted shuffler between the users and the data collector.

Definition 2 (Local Randomizer, Shuffler, and Analyzer). *A protocol in the shuffled model is a triple $\mathcal{P} = (\mathcal{R}, \mathcal{S}, \mathcal{A})$, where the local randomizer $\mathcal{R} : \mathcal{D} \to \mathcal{Y}^t$ maps one record to t messages, the shuffler $\mathcal{S} : \mathcal{Y}^{nt} \to \mathcal{Y}^{nt}$ applies a uniformly random permutation to the messages it receives, and*

the analyzer $\mathcal{A} : \mathcal{Y}^{nt} \to \mathcal{O}$ maps the shuffled messages to the output of the protocol. On a dataset $D = (z_1, \ldots, z_n)$, the view of the data collector is

$$\mathcal{V}_{\mathcal{P}}(D) = \mathcal{S}\big(\mathcal{R}(z_1), \ldots, \mathcal{R}(z_n)\big), \tag{2}$$

and the output of the protocol is $\mathcal{M}_{\mathcal{P}}(D) = \mathcal{A}(\mathcal{V}_{\mathcal{P}}(D))$.

In the user-level regime, the record of user i is $U_i \in \mathcal{D}^m$, so the local randomizer is $\mathcal{R} : \mathcal{D}^m \to \mathcal{Y}^t$ and the view is $\mathcal{V}_{\mathcal{P}}(D) = \mathcal{S}\big(\mathcal{R}(U_1), \ldots, \mathcal{R}(U_n)\big)$.

Definition 3 (Shuffled Differential Privacy (Cheu et al., 2019; Balle et al., 2019))**.** *A protocol $\mathcal{P} = (\mathcal{R}, \mathcal{S}, \mathcal{A})$ satisfies (ϵ, δ)-shuffled DP (SDP) if the view $\mathcal{V}_{\mathcal{P}}$ satisfies Definition 1, i.e., for any two neighboring datasets D, D' and any measurable event $\mathcal{E} \subseteq \mathcal{Y}^{nt}$,*

$$\mathbb{P}[\mathcal{V}_{\mathcal{P}}(D) \in \mathcal{E}] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{V}_{\mathcal{P}}(D') \in \mathcal{E}] + \delta. \tag{3}$$

Definition 3 implies that $\mathcal{M}_{\mathcal{P}}$ is (ϵ, δ)-DP in the sense of Definition 1.

Definition 4 (Non-interactive and Interactive protocol)**.** *A protocol is non-interactive if every user applies $\mathcal{R}$ to their own record once and sends the resulting messages in a single round, so that $\mathcal{R}$ does not depend on any other user's messages. A protocol is interactive if it proceeds in $T > 1$ rounds, where the randomizer used by a user in round t may depend on the aggregated output released by the analyzer in rounds $1, \ldots, t-1$.*

3.2 Problem Setup

We consider the setting of sparse linear regression. Let $\mathcal{D} = \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ be the data universe and let there be n users in the population. We distinguish two data regimes. In the sample-level regime, each user $i \in [n]$ holds one pair $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. In the user-level regime, each user i holds m i.i.d. samples $U_i = \{(x_{i,j}, y_{i,j})\}_{j=1}^{m}$, which we also write in matrix form as $(X_i, y_i) \in \mathbb{R}^{m \times d} \times \mathbb{R}^m$; the total number of samples is then $N = nm$.

Assumption 1 (Sparse Linear Model)**.** *The samples are i.i.d. realizations of the sparse linear regression model $y = \langle \theta^*, x \rangle + \zeta$, where the distribution of x has mean zero, the noise ζ satisfies $\mathbb{E}[\zeta \mid x] = 0$ and is sub-Gaussian with variance proxy σ^2, and the underlying parameter lies in $\Theta_k = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq k\}$ with $k \leq d$ and $\|\theta^*\|_2 = \mathcal{O}(1)$. In the user-level regime, the samples are i.i.d. both within and across users.*

Assumption 2 (Covariance Structure)**.** *Let $\Sigma = \mathbb{E}[xx^\top]$ be the population covariance matrix. There is a constant $\kappa_x > 0$ such that $\|\Sigma^{-1/2} x\|_{\psi_2} \leq \kappa_x$.*

4 Understanding the Limitations in Non-interactive SDP

In this section, we characterize sparse linear regression under non-interactive shuffled DP. We first establish a minimax lower bound on the ℓ_2-norm estimation error, and then give an algorithm whose bound is decent.

Throughout this section the parameter space is $\widetilde{\Theta}_k = \Theta_k \cap \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1\}$. The ℓ_1 normalization strengthens the ℓ_2 bound of Assumption 1 and is what keeps the sensitivity of a single record under control once each user privatizes their own statistics; it is not needed in Section 5. We also use the following conditions, which are used in this section only.

Assumption 3 (Curvature and Boundedness)**.** *There is a constant $\kappa_\infty > 0$ such that $\|\Sigma w\|_\infty \geq \kappa_\infty \|w\|_\infty$ for any $w \neq 0$, and there are constants $c_x, c_y = O(1)$ such that $\|x\|_\infty \leq c_x$ and $|y| \leq c_y$ almost surely.*

The first condition is the ℓ_∞ curvature used by the soft-thresholding step, and the second lets each record be clipped at a known scale before it is privatized.

4.1 Lower Bound

We begin with the fundamental limits of the non-interactive setting. The standard route to a minimax lower bound under local DP does not apply here. In the local model each message depends on a

single record, so the KL divergence tensorizes, $D_{\mathrm{KL}}(P^n \| U^n) = n \cdot D_{\mathrm{KL}}(P \| U)$, and private Fano's inequality or a direct mutual information bound can be used. In the shuffled model the messages of the n users are permuted before the analyzer sees them, and the mutual information between the input and the view $\mathcal{V}_{\mathcal{P}}$ can be unbounded, so these arguments do not carry over. To solve the problem, we introduce pan-privacy, which constrains the internal state of an algorithm rather than the messages of individual users.

Definition 5 (Pan-Privacy (Cheu & Ullman, 2021)). *An algorithm M processes a stream $Z = (z_1, \ldots, z_n)$ one record at a time: it updates an internal state from z_i and the previous state, and computes an output from the final state. Write $M_I(Z_{\le t})$ for the internal state after the first t records and $M_O(M_I(Z))$ for the output. Then M is (ϵ, δ)-pan-private if for every pair of streams Z, Z' differing in one record, every $t \in [n]$, and every measurable set T of state-output pairs,*

$$\mathbb{P}\big[\big(M_I(Z_{\le t}), M_O(M_I(Z))\big) \in T\big] \le e^{\epsilon} \cdot \mathbb{P}\big[\big(M_I(Z'_{\le t}), M_O(M_I(Z'))\big) \in T\big] + \delta. \tag{4}$$

Pan-privacy protects the record against an adversary that observes the output together with the internal state at one point in the stream. Cheu & Ullman (2021) showed that any algorithm satisfying robust (ϵ, δ)-shuffled DP can be simulated by a (ϵ, δ)-pan-private algorithm on a constant fraction of the users; here, robustness means that the protocol remains (ϵ, δ)-shuffled DP as long as a constant fraction of the users follow it, a condition met by all protocols considered in this work. Therefore, to obtain the utility lower bound of SDP algorithms, it is enough to bound the error of pan-private algorithms.

The lower bound is proved on the following family of hard instances, which specifies a distribution over the sparse parameter θ^* and over the data (x, y).

Definition 6 (Hard Distribution). *Let ν be the target ℓ_2 error and let $\gamma = \nu/\sqrt{k}$. Define $\theta_Z = \gamma Z$ with $Z \in \{+1, 0, -1\}^d$, where each coordinate Z_j is drawn independently according to*

$$\mathbb{P}\{Z_j = +1\} = \frac{k}{4d}, \quad \mathbb{P}\{Z_j = -1\} = \frac{k}{4d}, \quad \mathbb{P}\{Z_j = 0\} = 1 - \frac{k}{2d}. \tag{5}$$

Given θ_Z, a sample $(x, y) \in \{-1, 1\}^d \times \{-1, 1\}$ is generated by drawing x uniformly from the hypercube, so that $\mathbb{P}(x) = 2^{-d}$, and setting $y = \langle x, \theta_Z \rangle + \zeta$ with

$$\mathbb{P}\{\zeta = 1 - \langle x, \theta_Z \rangle\} = \frac{1 + \langle x, \theta_Z \rangle}{2}, \quad \mathbb{P}\{\zeta = -1 - \langle x, \theta_Z \rangle\} = \frac{1 - \langle x, \theta_Z \rangle}{2}. \tag{6}$$

The expected sparsity of θ_Z is $k/2$, and Markov's inequality gives $\|\theta_Z\|_0 \le k$ with probability at least $1/2$; under $\nu \le 1/\sqrt{k}$ we also have $\|\theta_Z\|_1 \le 1$. The construction gives $y \in \{-1, 1\}$ and $\mathbb{E}[\zeta \mid x] = 0$, and since $|\langle x, \theta_Z \rangle| \le 1$, the noise satisfies $|\zeta| \le 2$ almost surely, so the instances lie in the model class of Assumptions 1–3 with parameter in $\widetilde{\Theta}_k$.

In Appendix D.1 we bound the sum over the d coordinates of the total variation distance between the two conditional distributions of the output given $Z_i = +1$ and $Z_i = -1$, which gives the following result.

Theorem 1. *Let $\mathcal{M}^{\mathrm{SDP}}_{\epsilon,\delta}$ be the family of all non-interactive protocols that satisfy robust (ϵ, δ)-shuffled DP with $0 < \epsilon \le \mathcal{O}(1)$ and $\delta = o(1/n)$. Let $Z = \{(x_i, y_i)\}_{i=1}^n$ consist of n i.i.d. samples from the sparse linear model of Assumption 1 with parameter θ^*. Then we have*

$$\inf_{\mathcal{A} \in \mathcal{M}^{\mathrm{SDP}}_{\epsilon,\delta}} \sup_{\theta^* \in \widetilde{\Theta}_k} \mathbb{E}\big[\|\mathcal{A}(Z) - \theta^*\|_2\big] \ge \Omega\Big(\frac{\sqrt{dk}}{n\epsilon}\Big), \tag{7}$$

where the expectation is over the data and the randomness of the protocol.

Central DP admits the rate $\Omega\big(\sqrt{k \log d/n} + k \log d/(n\epsilon)\big)$, while non-interactive local DP obeys $\Omega\big(\sqrt{dk \log d/(n\epsilon^2)}\big)$ (Cai et al., 2021; Zhu et al., 2024). Theorem 1 places non-interactive SDP between the two. Written as $\sqrt{dk}/(n\epsilon)$ against $\sqrt{dk \log d}/(\sqrt{n}\epsilon)$, the shuffled bound is smaller than the local one by a factor $\sqrt{n \log d}$, which is the gain from shuffling. The factor $\sqrt{d}$, however, is unchanged. This is the difference from the central model, where the privacy noise is added once to an aggregate statistic and the dimension enters only through the $\log d$ cost of locating the support; in the local and shuffled models each user randomizes their own record, and shuffling rescales the resulting per-coordinate noise without removing it. Hence the error of any non-interactive SDP protocol carries a polynomial dependence on d, whatever the local randomizer or analyzer, which is the primary bottleneck when $d \gg n$. We next give an algorithm that achieves a nontrivial rate.

4.2 Upper Bound

The sparse estimation is actually an ℓ_0-constrained least squares problem, and this is the fundamental challenge when proposing an efficient non-interactive SDP algorithm. To control local noise scales, we therefore relax the ℓ_0 constraint to an ℓ_1-regularized problem, whose solution can be written in closed form under Assumption 3. While this relaxation will link the error to the dimension d, the shuffler narrows this gap toward central DP. Then we can obtain the solution from the aggregated statistics by solving the normal equation and applying soft-thresholding.

Our non-interactive SDP algorithm is given as Algorithm 2 in Appendix B. We assume the length $D = d(d+1)/2 + d = \Theta(d^2)$.

Theorem 2. *Let $\theta^* \in \widetilde{\Theta}_k$ and Assumptions 1–3 hold, with $0 < \epsilon \leq \mathcal{O}(1)$ and the parameters of Algorithm 2 set as in Appendix B. If*

$$n \geq \Omega\left(\max\left\{\frac{d^2 \log n\sqrt{\log(d/\delta)\log d}}{\epsilon\kappa_\infty},\ \frac{d^2\log d}{\kappa_\infty^2},\ \frac{D\log(D/\delta)}{\epsilon^2}\right\}\right), \tag{8}$$

then Algorithm 2 satisfies (ϵ, δ)-shuffled DP, and with probability at least $1 - \mathcal{O}(d^{-2})$ its output satisfies

$$\|\hat{\theta} - \theta^*\|_2 \leq \widetilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}} + \frac{d\sqrt{k}}{n\epsilon}\right). \tag{9}$$

Privacy follows in two steps. The protocol that receives the rounded statistics directly and perturbs their sum with the binomial noise $\mathrm{Bin}(M, 1/2) - M/2$ is $(\epsilon, \delta/2)$-DP. By the security of the summation protocol, the aggregator's view of the real protocol is within total variation distance $e^{-\epsilon}\delta/8$ of this one, and Lemma F.2 of Balle et al. (2020) turns this distance into the stated (ϵ, δ)-shuffled DP guarantee. For the utility, the rounding error is of the same order as the binomial noise; we bound both in the ℓ_∞-norm and then apply the soft-thresholding bound. The resulting privacy term $d\sqrt{k}/(n\epsilon)$ leaves a gap of $\sqrt{d}$ to the lower bound $\Omega(\sqrt{dk}/(n\epsilon))$ of Theorem 1. The gap comes from the clipping radius: the second-moment statistic has $\|x_i x_i^\top\|_F = \|x_i\|_2^2 = \Theta(d)$, and to privatize it, this scale enters the noise directly. We leave closing this gap to future work; in either case, the polynomial dependence on d is not removed, which is the main barrier of the non-interactive setting.

5 Overcoming the Theoretical Barrier

Section 4 shows that a polynomial dependence on d is unavoidable in the non-interactive setting. In this section we show that the barrier can be overcome when each user holds multiple samples and two rounds of interaction are allowed. Specifically, we give a two-round user-level SDP algorithm whose estimation error is independent of the dimension d once the local sample size m is large enough. In this section, we consider the following additional assumption.

Assumption 4 (Restricted Eigenvalue)**.** *There is a constant $\kappa > 0$ such that $v^\top \Sigma v \geq \kappa\|v\|_2^2$ for every $S \subseteq [d]$ with $|S| \leq k$, and every v with $\|v_{S^c}\|_1 \leq 3\|v_S\|_1$, and there is a constant C such that $\max_{\ell\in[d]} \Sigma_{\ell\ell} \leq 1$ and $\lambda_{\max}(\Sigma_{S,S}) \leq C$ for every $|S| \leq k$.*

This is the restricted eigenvalue condition of Bickel et al. (2009), the standard condition under which the Lasso attains its fast rate, and the cone constant 3 is the one produced by the basic inequality when the regularization parameter dominates the score (van de Geer & Bühlmann, 2009; Negahban et al., 2012). It holds with high probability for sub-Gaussian designs whose population covariance has a bounded minimal eigenvalue (Raskutti et al., 2010), and conditions of this form are standard in the DP sparse regression literature (Cai et al., 2021; Ma et al., 2024; Tian et al., 2025). Taking v supported on S gives $\lambda_{\min}(\Sigma_S) \geq \kappa$.

5.1 Upper Bound

Instead of relaxing the ℓ_0 constraint as in Section 4.2, we reduce the problem to a low-dimensional regression by first estimating the support of θ^*. In the user-level regime, each user turns their m local samples into one local estimate of an active coordinate, and a heavy hitter step aggregates these estimates into the global support set (Bun et al., 2019; Ma et al., 2024).

Algorithm 1 User-level Two-Round Shuffled DP Sparse Regression (U2R)

Input: Local datasets $\{(X_i, y_i)\}_{i=1}^{n}$, sparsity k, privacy parameters (ϵ, δ).
Stage 1: Support Recovery (users $i \in \{1, \dots, n_1\}$)
for each user $i \in \{1, \dots, n_1\}$ **do**
  Solve the Lasso, get a local candidate set $\hat{S}_i$, and draw one index $v_i \sim \mathrm{Unif}(\hat{S}_i)$.
  Send the randomized response of v_i on each coordinate $\ell \in [d]$ through d parallel shufflers.
end for
Aggregator: decode the shuffled messages to get the counts $\{\hat{W}_\ell\}_{\ell \in [d]}$, recover support $\hat{S}$ via heavy hitter identification, and release $\hat{S}$, where $s = |\hat{S}|$.
Stage 2: Constrained Estimation (the remaining $n_2 = n - n_1$ users)
for each user $i \in \{n_1 + 1, \dots, n\}$ **do**
  Compute the local least squares estimate $\tilde{\theta}_i \in \mathbb{R}^s$ on $\hat{S}$, clip it, and send it to the shuffler.
end for
Aggregator: decode the messages of the first half users into a coarse estimate and broadcast it; the second half messages clip around this estimate, and yield the refined estimate $\hat{\theta}_{\hat{S}} \in \mathbb{R}^s$.
Output: the k-sparse estimate $\hat{\theta} \in \mathbb{R}^d$ that equals $\hat{\theta}_{\hat{S}}$ on $\hat{S}$ and zero elsewhere.

We summarize the pipeline in Algorithm 1. In this algorithm, **Stage 1: Support Recovery**, **Stage 2: Constrained Estimation**, and their internal hyperparameters (e.g., clipping thresholds) can be found in Algorithm 3 and Algorithm 4 in the Appendix B. Algorithm 1 is two-round because we count rounds by the messages that users send.

The heavy hitter step of Stage 1 (Algorithm 3) is what keeps the dimension out of the final rate, so we describe it in more detail. The heavy hitter algorithms used for the LDP model (Bun et al., 2019; Ma et al., 2024) estimate the frequency of every coordinate through a prefix tree: the users are split into $\lceil \log d \rceil$ groups, with the users of one group reporting a hashed bit of one level of the tree, and the server scans the levels in order. Two costs come with this. Every report has to be private on its own, so the randomization a user applies is calibrated to their own message, and each level of the tree is estimated from a $1/\log d$ fraction of the users. In the shuffled model, neither cost is needed: the shuffler removes the association between a message and its sender, so the aggregator sees nothing beyond the counts, and each user can answer at a randomization level calibrated to the view $\mathcal{V}_\mathcal{P}$ rather than to a single message. We therefore replace the prefix tree by one round of randomized response over $[d]$. User i reduces their m local samples to a single index v_i, sends the indicator of v_i with every coordinate flipped independently with probability γ, and the aggregator debiases the shuffled counts and keeps the coordinates that pass a threshold. To ensure $\hat{S}$ is the true support, we give Proposition 1.

Proposition 1. *Let Assumptions 1, 2 and 4 hold, let $S^* = \mathrm{supp}(\theta^*)$ with $|S^*| = k$, and let the parameters of Algorithm 3 be set as in Appendix B. Suppose there is $\rho > 0$ with $\min_{\ell \in S^*} |\theta^*_\ell| \geq \rho$, and that*

(i) ***Local sample size.***

$$m \geq \Omega\left(\max\left\{\frac{\sigma^2 k \log d}{\kappa^2 \rho^2},\ \frac{k \log(ed/k)}{\kappa^2},\ k + \log n\right\}\right). \tag{10}$$

(ii) ***User number of Stage 1.***

$$n_1 \geq \Omega\left(\max\left\{\frac{k\sqrt{\log(dn)\log(1/\delta)}}{\epsilon},\ k \log(dn)\right\}\right). \tag{11}$$

Then with probability at least $1 - \mathcal{O}(n^{-2})$,

$$\hat{S} = S^*, \tag{12}$$

so that $s = |\hat{S}| = k$.

Proposition 1 gives more than the containment $S^* \subseteq \hat{S}$ that Stage 2 needs: although the local set $\hat{S}_i$ of a user may differ from S^*, the heavy hitter step returns S^* exactly. In equation 10, the first two

terms make the local selection of one user in Stage 1 correct with probability at least $1 - 1/(32k)$, independently of n, and the third term is what Stage 2 needs for the local least squares estimate on $\hat{S}$. The first term is equivalently the beta-min condition $\rho \geq \Omega\big((\sigma/\kappa)\sqrt{k \log d/m}\big)$, which is standard for support recovery (Wainwright, 2009; Ma et al., 2024) and allows ρ to decrease as m grows.

Theorem 3. *Let $\theta^* \in \Theta_k$, let the conditions of Proposition 1 hold, and let the parameters of Algorithm 1 be set as in Appendix B. If*

$$n\epsilon \geq \Omega\left(\frac{\sqrt{m \log(k/\delta)}}{\sigma}\right), \tag{13}$$

then Algorithm 1 satisfies user-level (ϵ, δ)-shuffled DP, and with probability at least $1 - \mathcal{O}(n^{-2})$ its output satisfies

$$\|\hat{\theta} - \theta^*\|_2 \leq \widetilde{\mathcal{O}}\left(\sqrt{\frac{k}{nm}} + \frac{k}{n\sqrt{m}\epsilon}\right). \tag{14}$$

The dimension d is absent from both terms. By Proposition 1 the second stage runs on the k coordinates of $\hat{S}$, so the vector a user sends has k entries rather than d, and the dimension is eliminated. What remains of d is in the conditions on m and n_1 in Proposition 1. In the synthetic dimension sweep, U2R keeps a low mean relative coefficient error as d increases at fixed k and m, as shown in Table 6. User-level data alone does not suffice: Appendix F.1 shows that under a non-interactive protocol the lower bound for user-level SDP still carries a polynomial dependence on d, so interaction is also essential.

5.2 Lower Bound

We have already shown a promising result in Theorem 3. However, a natural question is whether our upper bound can be further improved. Now we show that our upper bound is already near-optimal.

Theorem 4. *Let $Z = \{U_i\}_{i=1}^n$ be a dataset in which each user $i \in [n]$ holds m i.i.d. samples $U_i = \{(x_{i,j}, y_{i,j})\}_{j=1}^m$ from the exponential family*

$$f_\theta(y \mid x) = h(y, \sigma) \exp\left(\frac{y x^\top \theta - \psi(x^\top \theta)}{c(\sigma)}\right). \tag{15}$$

Let Assumptions 2 and 4 hold and let $\|\psi''\|_\infty \leq c_2$. Let $\mathcal{M}_{\epsilon,\delta}$ be the family of all estimators satisfying user-level (ϵ, δ)-DP with $0 < \epsilon < 1$ and $\delta = o\big(\log(d/k)/(n\sqrt{m \log(nm)})\big)$. If $k \leq \sqrt{d}$, $k \geq \Omega(\log^2(d/k))$ and $n\sqrt{m}\,\epsilon \geq \Omega\big(k^{3/4} \log^{3/2}(d/k)\big)$, then

$$\inf_{\mathcal{A} \in \mathcal{M}_{\epsilon,\delta}} \sup_{\theta^* \in \Theta_k} \mathbb{E}\big[\|\mathcal{A}(Z) - \theta^*\|_2\big] \geq \Omega\left(\sqrt{\frac{k \log(d/k)}{nm}} + \frac{k \log(d/k)}{n\sqrt{m}\epsilon}\right), \tag{16}$$

where the expectation is over the data and the randomness of the estimator.

Theorem 4 defines the fundamental efficiency limits for any differential private estimator in the user-level setting. It is stated for the exponential family, which contains the sparse linear model of Assumption 1 as the special case $\psi(u) = u^2/2$ and $c(\sigma) = \sigma^2$, so the bound applies to the model class we estimate. By Definition 3, the output of any (ϵ, δ)-shuffled DP protocol is (ϵ, δ)-DP in the sense of Definition 1, so Theorem 4 also lower bounds the error of any user-level SDP protocol.

Comparing it with Theorem 3, the two rates agree in n, m, k and ϵ and only differ by the factor $\log(d/k)$: our algorithm is near-optimal, and sparse linear regression in the shuffled model is closer to central DP once interaction and enough user-level data are available.

User-level Score Attack. We now outline the proof. Minimax lower bounds under privacy constraints are difficult to characterize: Fano's inequality and mutual information bounds give no general recipe and are often loose, particularly in the user-level regime. To handle user-level data, we extend the score attack framework of Cai et al. (2023) and introduce the user-level score attack, which is our main technical tool here.

Definition 7 (User-level Score Attack)**.** *For a dataset Z in which each user $i \in [n]$ holds m i.i.d. samples $U_i = \{z_{i,1}, \dots, z_{i,m}\}$, the user-level score attack for an estimator $\mathcal{A}(Z)$ is*

$$A_i := \left\langle \mathcal{A}(Z) - \theta,\ \sum_{j=1}^{m} S_\theta(z_{i,j}) \right\rangle, \tag{17}$$

where $S_\theta(z) = \nabla_\theta \log f_\theta(z)$ is the score statistic.

For the exponential family of Theorem 4, with mean function $\mu(x^\top\theta) = \psi'(x^\top\theta)$, the score statistic is $S_\theta(y, x) = (y - \mu(x^\top\theta))x/c(\sigma)$, so the user-level score attack takes the form

$$A_i = \frac{1}{c(\sigma)} \left\langle \mathcal{A}(Z) - \theta,\ \sum_{j=1}^{m} \big(y_{i,j} - \mu(x_{i,j}^\top\theta)\big)x_{i,j} \right\rangle. \tag{18}$$

Proposition 2. *Let $Z = \{U_1, \dots, U_n\}$ be as in Theorem 4, let $\mathcal{A}(Z)$ be an estimator of θ, and let $Z'_{(i)}$ be the adjacent dataset in which the i-th user's data U_i is replaced by an i.i.d. copy $\widetilde{U}_i$. Write $A_i = \langle \mathcal{A}(Z) - \theta, S_\theta(U_i)\rangle$ and $A'_i = \langle \mathcal{A}(Z'_{(i)}) - \theta, S_\theta(U_i)\rangle$. Then*

(i) ***Soundness:*** *for each $i \in [n]$, $\mathbb{E}[A'_i] = 0$ and*

$$\mathbb{E}|A'_i| \le \sqrt{\mathbb{E}\|\mathcal{A}(Z) - \theta\|_2^2} \cdot \sqrt{\frac{m}{c(\sigma)}\lambda_{\max}\big(\mathbb{E}[\psi''(x^\top\theta)xx^\top]\big)}; \tag{19}$$

(ii) ***Completeness:*** *under standard regularity conditions,*

$$\sum_{i=1}^{n} \mathbb{E}[A_i] = \sum_{j=1}^{d} \frac{\partial}{\partial\theta_j}\mathbb{E}[\mathcal{A}(Z)_j]. \tag{20}$$

With Proposition 2, we bound the expected sum $\sum_{i=1}^{n} \mathbb{E}[A_i]$ from above by Lemma 3 and from below by Lemma 4. Comparing the two gives Theorem 4.

Numerical results. On synthetic data, U2R attains mean relative coefficient error 0.00520 at the default setting, comparable to non-private Lasso. On Kelmarsh, it has the lowest mean test MSE among the private methods at both $\epsilon = 1$ and 4 for each $m \in \{10, 50, 100\}$; Appendix C reports the detailed setup and complete results.

6 Conclusion

We studied sparse linear regression under shuffled DP. In the non-interactive setting, we showed a lower bound on the ℓ_2-norm estimation error of any (ϵ, δ)-SDP protocol and gave an algorithm with a matching form, both bounds carrying a polynomial dependence on d. In the interactive user-level setting, we gave a two-round algorithm that first applies a heavy hitter step and then estimates the parameter, whose error does not depend polynomially on d. By extending the score attack to user-level data, we proved a lower bound for any user-level (ϵ, δ)-DP estimator, which shows that this algorithm is near-optimal. Experiments on synthetic and real-world datasets support these results.

AI use statement

In this work, we used generative AI tools as coding assistants to implement the experimental algorithms. We have not used generative AI tools for drafting sections of the paper, generating synthetic datasets, or proving mathematical claims. Additionally, we used generative AI tools to revising wording and for organizing the bibliography. We have reviewed all AI-assisted work: the authors independently examined and verified all experimental scripts and mathematical derivations. We take full responsibility for the final content of this work, including text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our experiments use the publicly available Kelmarsh wind-farm dataset, released by Cubico Sustainable Investments Ltd under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. We use the dataset in accordance with its license terms and provide appropriate attribution. The experiments analyze wind-turbine operational measurements and involve neither human participants nor personally identifiable information.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that our results can be reproduced from the paper and its appendix. For the theoretical results, the problem setting and assumptions are stated explicitly in Section 3.2 (Assumptions 1 and 2). The additional conditions used in each regime are given in Assumption 3 (Section 4) and Assumption 4 (Section 5). Complete proofs of all claims are provided in the appendix. Theorems 1 and 2 are proved in Appendix D. Proposition 1, Theorem 3, Proposition 2 and Theorem 4 are proved in Appendix E. The supplementary results on non-interactive user-level SDP and dense regression, together with their proofs, are in Appendix F. For the algorithms, Appendix B gives full pseudocode for the non-interactive protocol (Algorithm 2) and for both stages of U2R (Algorithms 3 and 4). Tables 3–5 list every parameter required by the theorems, including clipping radii, rounding precisions, binomial noise parameters, moduli, numbers of shares, and thresholds. For the experiments, Appendix C.1 describes the synthetic data-generating process, the default values and sweep ranges, the number of repetitions, and the evaluation metrics. It also gives the hyperparameters of U2R and of all baselines, and how they were selected. The real-data experiments use the publicly available Kelmarsh wind farm dataset (CC BY 4.0). The same appendix details the full preprocessing pipeline, including feature construction, the date-based public/training/test split, imputation, standardization, and the assignment of rows to simulated users. We do not release source code with this submission. We believe these descriptions are sufficient to reimplement the algorithms and reproduce the reported results.

REFERENCES

Jayadev Acharya, Clément L. Canonne, Himanshu Tyagi, and Ziteng Sun. The role of interactivity in structured estimation. In *Proceedings of the 35th Conference on Learning Theory*, pp. 1328–1355. PMLR, 2022.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 393–403. PMLR, 2021.

Victor Balcer and Albert Cheu. Separating local & shuffled differential privacy via histograms. *arXiv preprint arXiv:1911.06879*, 2019.

Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *Advances in Cryptology – CRYPTO 2019*, pp. 638–667. Springer, 2019.

Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. Private summation in the multi-message shuffle model. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 659–676, 2020.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. *ACM Transactions on Algorithms*, 15(4):1–40, 2019.

T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

T. Tony Cai, Yichen Wang, and Linjun Zhang. Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2307.01801*, 2023.

Emmanuel Candès and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

Albert Cheu and Jonathan Ullman. The limits of pan privacy and shuffle privacy for learning and estimation. In *Proceedings of the 53rd Annual ACM Symposium on Theory of Computing (STOC)*, 2021.

Albert Cheu, Adam Smith, Jonathan Ullman, Maxim Zeber, and Sergey Zhilyaev. Distributed differential privacy via shuffling. In *Advances in Cryptology – EUROCRYPT 2019*, volume 11476 of *Lecture Notes in Computer Science*, pp. 375–403. Springer, 2019.

Amit Daniely and Vitaly Feldman. Locally private learning without interaction requires separation. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

John Duchi, Michael Jordan, and Martin Wainwright. Local privacy and statistical minimax rates. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 429–438, Berkeley, CA, USA, 2013. IEEE. doi: 10.1109/FOCS.2013.53. URL `https://arxiv.org/abs/1302.3203`.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Privacy Aware Learning. *Journal of the ACM*, 61(6):1–57, 2014.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Now Publishers Inc., 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2468–2479. SIAM, 2019.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. 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.

Badih Ghazi, Pasin Manurangsi, Rasmus Pagh, and Ameya Velingker. Private aggregation from fewer anonymous messages. In *Advances in Cryptology – EUROCRYPT 2020*, pp. 798–827. Springer, 2020.

Antonious M. Girgis and Suhas Diggavi. Multi-message shuffled privacy in federated learning. *IEEE Journal on Selected Areas in Information Theory*, 5:12–27, 2024.

Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. 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*, pp. 227–236, 2022.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014.

Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Proceedings of the 33rd Conference on Learning Theory*, pp. 2204–2235. PMLR, 2020.

Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 10633–10660. PMLR, 2022.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private Convex Empirical Risk Minimization and High-dimensional Regression. In *Conference on Learning Theory*, pp. 25.1–25.40. JMLR Workshop and Conference Proceedings, 2012.

Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Sai Suggala. Near optimal private and robust linear regression. *arXiv preprint arXiv:2301.13273*, 2023.

Yuheng Ma, Ke Jia, and Hanfang Yang. Better locally private sparse estimation given multiple samples per user. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR, 2024.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q balls. *IEEE Transactions on Information Theory*, 57(10): 6976–6994, 2011.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.

Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is Interaction Necessary for Distributed Private Learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 58–77. IEEE, 2017.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly Optimal Private LASSO. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Xizhi Tian, Meng Ding, Touming Tao, Zihang Xiang, and Di Wang. Differentially private sparse linear regression with heavy-tailed responses. In *Machine Learning and Knowledge Discovery in Databases: Research Track (ECML PKDD)*, pp. 363–379. Springer, 2025.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression via adaptive clipping. *arXiv preprint arXiv:2207.04686*, 2022.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, 2009. doi: 10.1109/TIT.2009.2016018.

Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. *IEEE Transactions on Information Theory*, 67(2):1182–1200, 2020.

Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10081–10091. PMLR, 2020.

Di Wang, Huangyu Zhang, Marco Gaboardi, and Jinhui Xu. Estimating smooth glm in non-interactive local differential privacy model with public unlabeled data. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 1207–1213. PMLR, 2021.

Eunho Yang, Aurelie Lozano, and Pradeep Ravikumar. Elementary estimators for high-dimensional linear regression. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 388–396. PMLR, 2014.

Kai Zheng, Wenlong Mou, and Liwei Wang. Collect at Once, Use Effectively: Making Non-interactive Locally Private Learning Possible. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 4130–4139. PMLR, 2017.

Liyang Zhu, Meng Ding, Vaneet Aggarwal, Jinhui Xu, and Di Wang. Improved analysis of sparse linear regression in local differential privacy model. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

A Detailed Comparison and Related Work

Table 2 provides a comprehensive comparison of our theoretical results with prior literature on (ϵ, δ)-DP linear regression. This extends the summary provided in the main text by including bounds for specific edge cases (such as 1-sparse models) and detailing previous interactive local methods.

Table 2: Detailed comparison of related studies on (ϵ, δ)-DP linear regression. Data assumptions and parameter definitions are identical to Table 1.

Type	Unit	Setting	Upper Bound	Lower Bound	Source
Central DP					
-	sample	general	$\tilde{\mathcal{O}}(\sqrt{\frac{d}{n}} + \frac{d}{n\epsilon})$	$\Omega(\sqrt{\frac{d}{n}} + \frac{d}{n\epsilon})$	Cai et al. (2021)
	sample	k-sparse	$\tilde{\mathcal{O}}(\sqrt{\frac{k}{n}} + \frac{k}{n\epsilon})$	$\Omega(\sqrt{\frac{k\log d}{n}} + \frac{k\log d}{n\epsilon})$	Cai et al. (2021)
Local DP					
Non-interactive	sample	1-sparse	-	$\Omega(\sqrt{\frac{d\log d}{n\epsilon^2}})$	Wang & Xu (2020)
	sample	k-sparse	$\tilde{\mathcal{O}}(\frac{d\sqrt{k}}{\sqrt{n}\epsilon})$	$\Omega(\sqrt{\frac{dk\log d}{n\epsilon^2}})$	Zhu et al. (2024)
Interactive	sample	general	$\tilde{\mathcal{O}}(\frac{d^{\frac{3}{2}}}{\sqrt{n}\epsilon})$	-	Smith et al. (2017)
	sample	1-sparse	-	$\Omega(\sqrt{\frac{d}{n\epsilon^2}})$	Wang & Xu (2020)
	sample	k-sparse	$\tilde{\mathcal{O}}(\frac{k\sqrt{d}}{\sqrt{n}\epsilon})$	$\Omega(\sqrt{\frac{dk}{n\epsilon^2}})$	Zhu et al. (2024)
	user-level	k-sparse	$\tilde{\mathcal{O}}(\sqrt{\frac{k}{nm}} + \frac{k}{\sqrt{nm}\epsilon})$	$\Omega(\frac{k}{\sqrt{nm}\epsilon})$	Ma et al. (2024)
Shuffle DP					
-	user-level	general	-	$\Omega(\sqrt{\frac{d}{nm}} + \frac{d}{n\sqrt{m}\epsilon})$	**Our Work**
Non-interactive	user-level	k-sparse	-	$\Omega(\sqrt{\frac{k\log(d/k)}{nm}} + \frac{\sqrt{dk}}{n\sqrt{m}\epsilon})$	
	sample	k-sparse	$\tilde{\mathcal{O}}(\sqrt{\frac{k}{n}} + \frac{d\sqrt{k}}{n\epsilon})$	$\Omega(\frac{\sqrt{dk}}{n\epsilon})$	
Interactive	user-level	k-sparse	$\tilde{\mathcal{O}}(\sqrt{\frac{k}{nm}} + \frac{k}{n\sqrt{m}\epsilon})$	$\Omega(\sqrt{\frac{k\log(d/k)}{nm}} + \frac{k\log(d/k)}{n\sqrt{m}\epsilon})$	

Private high-dimensional regression has been studied extensively in the central model of differential privacy. Early work by Chaudhuri et al. (2011) and Kifer et al. (2012) showed that private empirical risk minimization remains feasible even in high dimension, and Talwar et al. (2015) designed a nearly optimal private LASSO algorithm under an ℓ_1 constraint. The central-DP optimization literature has since developed a much sharper understanding of stochastic convex optimization and private linear regression, including optimal-rate results in different geometries and heavy-tailed regimes (Bassily et al., 2019; Feldman et al., 2020; Asi et al., 2021; Wang et al., 2020; Kamath et al., 2022;

Hu et al., 2022; Liu et al., 2023; Varshney et al., 2022). On the statistical side, Cai et al. (2021) and Cai et al. (2023) established minimax rates and lower-bound techniques for private parameter estimation, showing that sample-level central DP only incurs logarithmic dependence on the ambient dimension in sparse problems. The score attack of Cai et al. (2023) is stated for datasets in which each record is one sample; the user-level score attack of Section 5.2 extends it to the regime where a record is the whole local dataset of a user, which is what Theorem 4 needs. These results provide the benchmark that one would ideally like to approach in decentralized settings.

The local model is considerably more restrictive. Foundational results on local privacy and statistical minimax rates are due to Duchi et al. (2013) and Duchi et al. (2014), while later work studied the role of interaction in distributed private learning and local estimation (Smith et al., 2017; Zheng et al., 2017; Daniely & Feldman, 2019; Hu et al., 2022; Acharya et al., 2022; Wang et al., 2021). For sparse linear regression in particular, Wang & Xu (2020) initiated the study of the local model, proving lower bounds in the 1-sparse case and proposing an interactive private iterative hard-thresholding procedure under strong distributional assumptions. The later improved analysis of Zhu et al. (2024) substantially sharpened this picture in two directions: it extended the lower bounds to general k-sparse regression, and it provided both the first explicit non-interactive LDP estimator for sparse regression and a corrected utility analysis for the interactive hard-thresholding line. Their results show that, even after these improvements, sample-level LDP sparse regression still exhibits a polynomial dependence on the ambient dimension d, with non-interactive and interactive upper bounds scaling on the order of $\widetilde{O}(d\sqrt{k}/(\sqrt{n}\epsilon))$ and $\widetilde{O}(k\sqrt{d}/(\sqrt{n}\epsilon))$, respectively. These phenomena are closely related to the non-private high-dimensional estimation literature, where canonical sparsity-aware procedures include the Dantzig selector (Candès & Tao, 2007), elementary estimators (Yang et al., 2014), and iterative hard thresholding (Jain et al., 2014); minimax rates over sparse parameter classes are summarized by Raskutti et al. (2011).

Several papers have therefore looked for stronger primitives or better structural assumptions. On the robust statistics side, clipped or Huberized estimators are now standard tools for controlling heavy-tailed noise and stabilizing private regression (Sun et al., 2020; Kamath et al., 2020); these ideas also appear explicitly in the improved non-interactive LDP construction of Zhu et al. (2024), where clipped sufficient statistics are privatized locally before sparse recovery is performed at the server. Both of our protocols clip in the same way, and Algorithm 2 privatizes the same clipped sufficient statistics, the difference being that the noise is added to the aggregate rather than to each record. On the decentralized side, Ma et al. (2024) moved to the user-level multi-sample regime and proposed the two-stage 2-SLR framework, where each user first locally screens candidate coordinates before a second-stage low-dimensional estimator is applied. Their result is especially relevant to our work because it shows that once multiple samples are grouped per user, the severe dimension dependence of standard sample-level LDP can be substantially alleviated. Algorithm 1 follows the same two-stage structure, and the local screening step of Algorithm 3 is a Lasso together with a threshold, which asks only for the restricted eigenvalue condition of Assumption 4 rather than for uncorrelated features.

Shuffled differential privacy offers a middle ground between central DP and LDP. The basic model and its privacy amplification effect were formalized by Cheu et al. (2019) and Erlingsson et al. (2019), and the amplification view was later refined through the privacy-blanket perspective (Balle et al., 2019). A separate line studies multi-message protocols that compute a sum in the shuffled model with the accuracy of the central model, both for real summation and for aggregation in federated learning (Balle et al., 2020; Ghazi et al., 2020; Girgis & Diggavi, 2024). The summation step of Algorithm 2 and of Stage 2 of Algorithm 1 is of this second type: the aggregator sees a noisy sum rather than n separately randomized messages, so the noise can be calibrated to the ℓ_2 sensitivity of the query. On the negative side, separations and limitations of shuffle privacy have been established for statistical tasks such as histogram estimation, together with the relation between shuffled DP and pan privacy (Balcer & Cheu, 2019; Cheu & Ullman, 2021), which is the route our non-interactive lower bound follows. Heavy-hitter primitives under local noise provide a practical mechanism for support identification in decentralized settings (Bun et al., 2019); Stage 1 of Algorithm 1 replaces the prefix tree of these protocols by one round of randomized response over $[d]$. However, despite this rapid progress, sparse linear regression in the shuffled model has remained largely open: the existing literature either analyzes general shuffled estimation primitives or studies local/user-level baselines rather than giving a sharp treatment of sparse regression itself. Our work fills this gap by

characterizing both the non-interactive sample-level shuffled regime and the interactive user-level shuffled regime for sparse linear models.

B Detailed Algorithms and Parameter Settings

This section gives the pseudocode of the algorithms of Sections 4 and 5, together with the parameter settings used by Theorem 2, Proposition 1 and Theorem 3. Throughout, $D = d(d+1)/2 + d = \Theta(d^2)$, $v(\cdot)$ vectorizes the upper triangle of a symmetric matrix including the diagonal, $\Pi_{B(0,R)}$ is the Euclidean projection onto the ball of radius R, and $c_0, c_1, c_A, c'_A, c_\lambda, c_p, B_\theta, C_\beta, C_\gamma$ are absolute constants.

Both protocols privatize a sum of integer vectors in the same way: each user rounds its statistic to the integer grid, adds a share of a centered binomial noise, and uploads additive shares of the result through parallel shufflers, one per coordinate. The noise is calibrated to the ℓ_2 sensitivity of the integer query rather than to its ℓ_1 sensitivity, which is what keeps the privacy term at d in Theorem 2 and removes d altogether in Theorem 3.

Algorithm 2 runs on all n users at the budget (ϵ, δ). In Algorithm 1 the three groups of users, namely the n_1 users of Stage 1, the n_3 users of Stage 2a and the n_4 users of Stage 2b, are disjoint, so a neighboring pair of datasets changes the view of one group only and each group runs at the full budget (ϵ, δ) by parallel composition. The parameter settings below are stated accordingly. The summation step requires at least 19 participants per group.

B.1 Non-interactive Algorithm

Algorithm 2 Sample-Level Non-Interactive Shuffled DP Sparse Linear Regression

Input: Data $\{(x_i, y_i)\}_{i=1}^n$, sparsity k, privacy parameters (ϵ, δ), parameters of Table 3.
Local procedure (user $i \in [n]$)
Compute $H_i = x_i x_i^\top$, $g_i = x_i y_i$, and set $F_i = (v(H_i); g_i) \in \mathbb{R}^D$.
Clip: $\bar{F}_i = F_i \cdot \min\{1, C_F/\|F_i\|_2\}$.
For each $j \in [D]$, round: $\bar{U}_{i,j} = \lfloor p\bar{F}_{i,j} \rfloor + \mathrm{Ber}(p\bar{F}_{i,j} - \lfloor p\bar{F}_{i,j} \rfloor)$.
For each $j \in [D]$, draw $\zeta_{i,j} \sim \mathrm{Bin}(N_0, 1/2) - N_0/2$ and set $W_{i,j} = (\bar{U}_{i,j} + \zeta_{i,j}) \bmod q$.
For each $j \in [D]$, split $W_{i,j}$ into r additive shares in $\mathbb{Z}_q$ and send them, together with one unshuffled share, through the j-th shuffler.
Shuffler: permute the messages of each of the D streams independently.
Aggregator
For each $j \in [D]$, sum the received shares modulo q to get S_j, and set $S_j \leftarrow S_j - q$ if $S_j > q/2$.
Rescale $\hat{F}_j = S_j/(np)$, fill $\hat{H} \in \mathbb{R}^{d\times d}$ from the first $d(d+1)/2$ entries of $\hat{F}$ and $\hat{g} \in \mathbb{R}^d$ from the last d.
Solve $\tilde{\theta} = \hat{H}^{-1}\hat{g}$, soft-threshold $\hat{\theta} = \mathcal{S}_{\lambda_n}(\tilde{\theta})$, and project $\hat{\theta}$ onto $\{\theta : \|\theta\|_1 \le 1\}$.
Output: the k-sparse estimate $\hat{\theta}$.

The parameter N_0 is even, so the noise of a user is integer valued, and $N_0 \ge 2$, so every user holds at least one Bernoulli trial per coordinate and the shares of the n users add up to a centered $\mathrm{Bin}(nN_0, 1/2)$ noise on each coordinate. The choice of p keeps the rounding error of the same order as the binomial noise, so that both are absorbed into λ_n. The modulus q is large enough that the sum of the uploaded values does not wrap around. Under Assumption 3 we have $\|F_i\|_2 \le c_x^2 d + c_x c_y \sqrt{d}$, so the clipping step does not change F_i once $c_0 \ge c_x^2 + c_x c_y$; the step is kept so that the algorithm also applies to data that are not bounded, where the factor $\log n$ in C_F covers the tail of $\|F_i\|_2$, which then exceeds C_F with probability at most n^{-3}.

B.2 Stage 1: Support Recovery

The window for λ_{loc} is non-empty under the first term of equation 10: the lower end is what dominates the score $\|m^{-1}X_i^\top \zeta_i\|_\infty$ of one user with probability at least $1 - 1/(64k)$, and the upper end is what turns the ℓ_2 rate of the Lasso into the coordinatewise control needed by the threshold $\rho/2$.

Table 3: Parameters of Algorithm 2, used in Theorem 2.

Symbol	Setting	Role
C_F	$c_0\, d \log n$	ℓ_2 clipping radius
p	$\left\lceil \frac{\sqrt{n}\,\epsilon}{d \log n \sqrt{\log(4D/\delta)}} \right\rceil$	rounding precision
$\Delta_2^{(p)}$	$2pC_F + \sqrt{D}$	ℓ_2 sensitivity of the integer query
M	$\left\lceil \frac{32(\Delta_2^{(p)})^2 \log(4D/\delta)}{\epsilon^2} \right\rceil$	total binomial parameter
N_0	$2\lceil M/(2n) \rceil$	binomial parameter per user and coordinate
q	$4n(pC_F + N_0)$	modulus of the group $\mathbb{Z}_q$
r	$\left\lceil \frac{2\log_2(8De^{\epsilon}/\delta) + \log_2 q}{\log_2 n - \log_2 e} \right\rceil + 1$	number of additive shares per coordinate
λ_n	$\frac{c_1}{\kappa_\infty}\left(\sigma\sqrt{\frac{\log d}{n}} + \frac{C_F\sqrt{\log(4D/\delta)\log D}}{n\epsilon}\right)$	soft-thresholding parameter

Table 4: Parameters of Algorithm 3, used in Proposition 1.

Symbol	Setting	Role
λ_{loc}	$c_\lambda\, \sigma\sqrt{\frac{\log d}{m}}$, with $\lambda_{\mathrm{loc}} \le \frac{\kappa\rho}{16\sqrt{k}}$	regularization parameter of the local Lasso
$\rho/2$	—	threshold of the local selection step
γ	$\frac{C_\gamma \log(1/\delta)}{n_1\epsilon^2}$	flipping probability of the randomized response
τ_{HH}	$\frac{n_1}{4k}$	threshold on the debiased counts

The value of γ is the smallest one at which the counts are private. It satisfies $\gamma \le 1/4$ as long as $\delta \ge (dn)^{-Ck^2\epsilon^2}$, so that the debiasing factor $1/(1-2\gamma)$ is at most 2, and under equation 11 the fluctuation of the debiased counts stays below $n_1/(16k)$, so that τ_{HH} separates the two groups of coordinates.

B.3 Stage 2: Constrained Estimation

The $n_2 = n - n_1$ remaining users are split into two groups of sizes $n_3 = \lfloor n_2/2 \rfloor$ and $n_4 = n_2 - n_3$. Both groups upload through the same summation step as Algorithm 2, with s parallel shufflers instead of D. The quantity $\hat{A}$ broadcast with τ is a function of the public parameters and of $\hat{S}$, which has already been released, so broadcasting τ costs no privacy budget.

The two groups of Stage 2 use the same summation step with different sizes and clipping radii. To state their parameters once, we index them by $g \in \{a, b\}$ and set

$$(n_a, R_a) = (n_3, R_0), \qquad (n_b, R_b) = (n_4, \tau),$$

so that $p_g, \Delta_g, M_g, N_0^{(g)}, q_g, r_g$ in Table 5 are the parameters of Stage 2a for $g = a$ and of Stage 2b for $g = b$.

The constant B_θ is any known upper bound on $\|\theta^*\|_2$, which Assumption 1 provides; taking $B_\theta = 1$ is enough on the parameter space we consider, so that $R_0 = \Theta(1)$. The second term in the setting of p_g makes $\Delta_g \le 3p_gR_g$ and $\Delta_g \ge \sqrt{s}\max\{1, \epsilon\}$ hold for every value of s, at the cost of a larger $N_0^{(g)}$, that is of more messages per user; it does not change the noise seen by the aggregator. Indeed the precision p_g cancels against the sensitivity Δ_g, and that noise is $\Theta\big(R_0\sqrt{s\log n\log(4s/\delta)}/(n_3\epsilon)\big)$ for the coarse estimate and $\Theta\big(\tau\sqrt{s\log n\log(4s/\delta)}/(n_4\epsilon)\big)$ for the refined one. Since $\tau = \Theta(\tau_\beta)$ under the condition equation 13, the second scale is what produces the privacy term of Theorem 3, and it involves s rather than d.

Algorithm 3 Heavy Hitter Support Recovery (Stage 1)

Input: Local datasets $\{(X_i, y_i)\}_{i=1}^{n_1}$, sparsity k, signal strength ρ, privacy parameters (ϵ, δ), parameters of Table 4.
Local procedure (user $i \in [n_1]$)
Solve the Lasso $\hat{\theta}^{(i)} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2m}\|y_i - X_i\theta\|_2^2 + \lambda_{\text{loc}}\|\theta\|_1$.
Set $\hat{S}_i = \{\ell \in [d] : |\hat{\theta}_\ell^{(i)}| \geq \rho/2\}$, and $\hat{S}_i = [d]$ if this set is empty.
Draw $v_i \sim \text{Unif}(\hat{S}_i)$.
For each $\ell \in [d]$, send $r_{i\ell} = \mathbb{I}[v_i = \ell] \oplus \text{Ber}(\gamma)$ through the ℓ-th shuffler.
Shuffler: permute the messages of each of the d streams independently.
Aggregator
For each $\ell \in [d]$, form the count $C_\ell = \sum_{i=1}^{n_1} r_{i\ell}$ and debias it: $\hat{W}_\ell = (C_\ell - n_1\gamma)/(1-2\gamma)$.
Set $\hat{S} = \{\ell \in [d] : \hat{W}_\ell \geq \tau_{HH}\}$ and $s = |\hat{S}|$.
Output: the support estimate $\hat{S}$.

Algorithm 4 Clipped Sparse Estimation (Stage 2)

Input: Support $\hat{S}$ with $s = |\hat{S}|$, local datasets of the n_2 users, privacy parameters (ϵ, δ), parameters of Table 5.
Local procedure (every user of Stage 2)
Compute the local least squares estimate $\tilde{\theta}_i = (X_{i,\hat{S}}^\top X_{i,\hat{S}})^{-1} X_{i,\hat{S}}^\top y_i \in \mathbb{R}^s$.
Stage 2a (the first n_3 users)
Clip $\bar{\theta}_i = \Pi_{B(0,R_0)}(\tilde{\theta}_i)$.
For each $j \in [s]$, round $p_a\bar{\theta}_{i,j}$ to the integer grid, add $\text{Bin}(N_0^{(a)}, 1/2) - N_0^{(a)}/2$, reduce modulo q_a, and send r_a additive shares together with one unshuffled share through the j-th shuffler.
Aggregator: decode the shuffled messages, rescale by $n_3 p_a$ to get the coarse estimate $c \in \mathbb{R}^s$, and broadcast $\tau = \tau_\beta + \hat{A}$.
Stage 2b (the remaining n_4 users)
Clip around the coarse estimate: $\Delta_i = \Pi_{B(0,\tau)}(\tilde{\theta}_i - c)$.
For each $j \in [s]$, round $p_b\Delta_{i,j}$ to the integer grid, add $\text{Bin}(N_0^{(b)}, 1/2) - N_0^{(b)}/2$, reduce modulo q_b, and send r_b additive shares together with one unshuffled share through the j-th shuffler.
Aggregator: decode the shuffled messages, rescale by $n_4 p_b$, and add back c to get $\hat{\theta}_{\hat{S}} \in \mathbb{R}^s$.
Output: the estimate $\hat{\theta} \in \mathbb{R}^d$ that equals $\hat{\theta}_{\hat{S}}$ on $\hat{S}$ and zero elsewhere.

C Experimental Setup and Results

We evaluate U2R on synthetic sparse linear regression and nacelle temperature prediction using the Kelmarsh wind farm dataset.

C.1 Experimental Setup

We compare U2R in Algorithm 1 with 2-SLR (Ma et al., 2024), LDP-IHT (Zhu et al., 2024), Sample-SDP in Algorithm 2, and non-private Lasso fitted to all nm observations. U2R and 2-SLR protect users; LDP-IHT and Sample-SDP protect individual observations. U2R and Sample-SDP use shuffled (ϵ, δ)-DP with $\delta = 10^{-5}$; 2-SLR and LDP-IHT use pure ϵ-LDP. For the same protected unit, pure ϵ-DP implies $(\epsilon, 10^{-5})$-DP, so we compare utility at the same numerical budget while retaining the differences in privacy units and trust models.

For synthetic data, each user holds m independent observations

$$x_{ij} \sim \mathcal{N}(0, I_d), \qquad y_{ij} = x_{ij}^\top \theta^* + \zeta_{ij}, \qquad \zeta_{ij} \sim \mathcal{N}(0, \sigma^2),$$

with covariates independent of residuals. We choose k coordinates uniformly, set their coefficients to $1/k$, and set the others to zero. The public defaults are $n = 15000$, $m = 50$, $d = 256$, $k = 8$, $\sigma = 0.25$, and $\epsilon = 4$; each sweep varies one quantity. The realized support is used only for evaluation.

Table 5: Parameters of Algorithm 4, used in Theorem 3. The index $g \in \{a, b\}$ refers to Stage 2a and Stage 2b, with $(n_a, R_a) = (n_3, R_0)$ and $(n_b, R_b) = (n_4, \tau)$.

Symbol	Setting	Role
τ_β	$C_\beta\,\sigma\sqrt{\dfrac{s\log n}{m}}$	uniform bound on $\|\tilde{\theta}_i - \theta^*_{\hat{S}}\|_2$
R_0	$B_\theta + \tau_\beta$	clipping radius of Stage 2a
$\hat{A}$	$c_A\,\sigma\sqrt{\dfrac{s\log n}{n_3 m}} + c'_A\,\dfrac{R_0\sqrt{s\log n\log(4s/\delta)}}{n_3\epsilon}$	error bound of the coarse estimate c
τ	$\tau_\beta + \hat{A}$	clipping radius of Stage 2b
p_g	$\max\left\{\dfrac{c_p\sqrt{n_g\log n}\,\epsilon}{R_g\sqrt{\log(1/\delta)}},\ \dfrac{(\sqrt{s}+2)\max\{1,\epsilon\}}{R_g}\right\}$	rounding precision
Δ_g	$2p_gR_g + \max\{1,\epsilon\}\sqrt{s} + 2$	ℓ_2 sensitivity of the integer query; $\Delta_g \le 3p_gR_g$
M_g	$\left\lceil\dfrac{32\Delta_g^2\log(4s/\delta)}{\epsilon^2}\right\rceil$	total binomial parameter
$N_0^{(g)}$	$2\lceil M_g/(2n_g)\rceil$	binomial parameter per user and coordinate
q_g	$4n_g(p_gR_g + N_0^{(g)})$	modulus of the group $\mathbb{Z}_{q_g}$
r_g	$\left\lceil\dfrac{2\log_2(8se^\epsilon/\delta) + \log_2 q_g}{\log_2 n_g - \log_2 e}\right\rceil + 1$	number of additive shares per coordinate

Each setting has ten repetitions with shared training data and 4000 independent test observations across methods. Matching settings and privacy budgets reuse the data and randomization, with one Lasso fit per repetition. We simulate aggregate mechanism outputs and measure relative coefficient error $\|\hat{\theta} - \theta^*\|_2/\|\theta^*\|_2$, test MSE, support recall $|\hat{S} \cap S^*|/k$, and exact recovery.

U2R uses disjoint user groups of sizes $\lfloor n/2\rfloor$ for selection and two equal remaining groups for coarse and refined estimation. Its local Lasso penalty is $0.01\sigma\sqrt{\log(dn)/m}$, with coefficient threshold $\rho/2$ and count threshold $n_1/(4k)$, where $\rho = 1/k$. The response probability is the larger of $32\log(1/\delta)/(n_1\epsilon^2)$ and the value calibrated from shuffled binary response or the bound for two bits under pure local DP. Coordinates are shuffled independently. Stage 2 follows Algorithm 4 with local OLS, clipping radius and rounding coefficients set to one, and privacy calibration for finite samples, including the summation protocol.

For 2-SLR, selection uses the 64 largest positive marginal correlations, selection size 2, $B = 1$, and 4 bins. LDP-IHT uses $T = 2$ and gradient radius 4; separate development seeds select $\eta = 1.5$, $\tau_x = 1.5$, $\tau_y = 0.75$, $k' = 10$ for $\epsilon < 4$, and $\eta = 2$, $\tau_x = 1$, $\tau_y = 0.5$, $k' = 8$ otherwise. Sample-SDP uses $c_0 = 0.1$ and $c_1 = 2$. Lasso uses five fold cross validation over 300 penalties, no intercept, tolerance 10^{-4}, and at most 3000 iterations. These rules are fixed before the sweeps.

Kelmarsh contains measurements every ten minutes from six turbines over 2017 to 2022. We predict nacelle temperature from 90 predictors: 59 physical mean signals, four cyclic time features, 22 missing value indicators, and five turbine indicators. Whole UTC dates are split into public, training, and test sets in nominal proportions 20%, 70%, and 10%, leaving $1{,}295{,}987$ training and $185{,}416$ test rows after removing missing responses. Separate public dates provide $80{,}000$ rows for preprocessing and $40{,}000$ for validation. Public medians fill missing values; public means and standard deviations scale features and responses, with indicators centered.

Each real data repetition samples nm training rows without replacement and assigns them to $n = 10000$ simulated users. We test $m \in \{10, 50, 100\}$ and $\epsilon \in \{1, 4\}$, sharing rows and groups across methods and budgets. The test set is fixed.[1] U2R uses $\rho = 0.05$, a unit parameter norm bound, and $\sigma = 0.25$. Public validation over all six settings and three separate seeds selects $k = 12$ from $\{8, 10, 12\}$. After count thresholding, U2R keeps at most $\min\{k, m - 2\}$ coordinates ranked by debiased private counts and calibrates Stage 2 to the selected dimension.

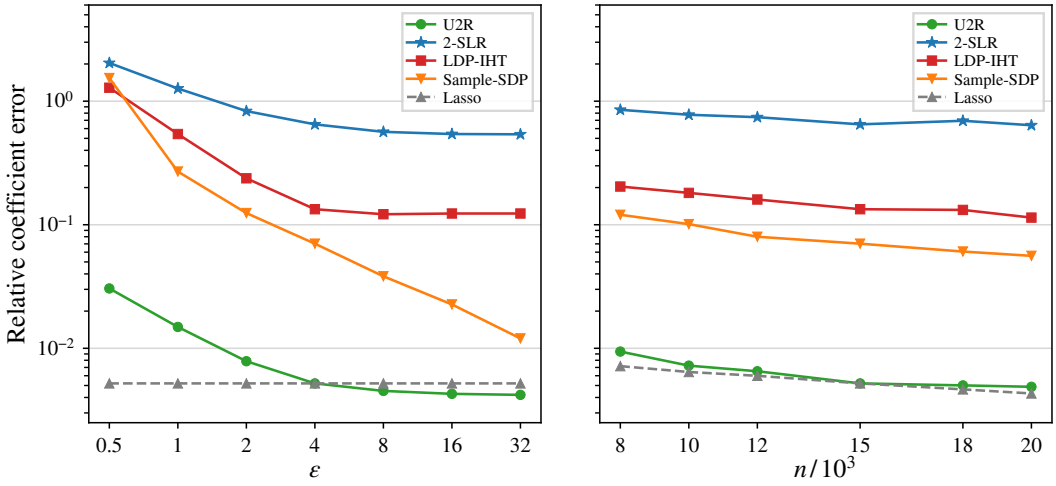
[1] The fixed test dates were also consulted during development.

For real data, 2-SLR uses 16 bins. LDP-IHT uses $T = 5$, $\eta = 0.1$, $\tau_x = 1$, $\tau_y = 0.5$, and $k' = 8$. Sample-SDP uses $c_0 = 0.03$, $c_1 = 1$, and diagonal shift $\max\{0, -\lambda_{\min}(\widetilde{G})\} + 4\sqrt{d}\, s_G$, where s_G is the noise standard deviation per coordinate. Lasso cross validation is grouped by UTC date. Other settings remain unchanged. We report ten repetitions of MSE in the original units and its ratio to the paired Lasso fit, reused across budgets.

C.2 Main Results

Privacy budget and number of users. Figure 1 shows relative coefficient error as ϵ varies in Exp1 and n varies in Exp3. In Exp1, U2R has the lowest mean error among the private methods at every tested budget, decreasing from 0.03050 at $\epsilon = 0.5$ to 0.00420 at $\epsilon = 32$. At $\epsilon = 4$, its error is 0.00520, comparable to non-private Lasso. Most of the reduction occurs between $\epsilon = 0.5$ and 4, with smaller gains at larger budgets.

In Exp3, increasing n from 8000 to 20000 at $m = 50$ reduces U2R's mean error from 0.00941 to 0.00488. Each private method has lower mean error at the larger user count. U2R recovers the exact support in every repetition, so its remaining coefficient error lies on the active coordinates. Section C.3 examines support recovery as n and m vary.



(a) Privacy budget. (b) Number of users.

Figure 1: Relative coefficient error versus ϵ in Exp1, left, and n in Exp3, right, with $m = 50$, $d = 256$, and $k = 8$. Exp1 fixes $n = 15000$; Exp3 fixes $\epsilon = 4$. Lines show means over ten repetitions. Lasso is shown in gray. Both vertical axes use the same logarithmic scale.

Ambient dimension. In Exp2, U2R's mean error ranges from 0.00520 to 0.00630 as d increases from 64 to 512, with $k = 8$ and $m = 50$ fixed, as shown in Table 6. It remains close to non-private Lasso across these dimensions. Sample-SDP's error increases from 0.01766 to 0.14111 over the same range.

Table 6: Exp2: relative coefficient error versus dimension at $n = 15000$, $m = 50$, and $\epsilon = 4$. Entries are means over ten repetitions.

d	U2R	2-SLR	LDP-IHT	Sample-SDP	Lasso
64	0.00567	0.49079	0.13262	0.01766	0.00450
128	0.00630	0.61461	0.12595	0.03549	0.00453
256	0.00520	0.64914	0.13354	0.07027	0.00520
512	0.00561	0.77533	0.18692	0.14111	0.00522

Real data prediction. Table 7 reports test MSE relative to non-private Lasso in Exp5. U2R has the lowest mean ratio among the private methods in all six settings. As m increases from 10 to 100, its ratio decreases from 1.48844 to 1.12972 at $\epsilon = 1$ and from 1.43520 to 1.13226 at $\epsilon = 4$. More observations per user bring U2R closer to Lasso. Increasing ϵ reduces its mean prediction error at $m = 10$ and 50; at $m = 100$, the ratios are similar under both budgets.

Table 7: Exp5: Kelmarsh test MSE divided by the paired non-private Lasso MSE at $n = 10000$ and $d = 90$. U2R and Sample-SDP use $\delta = 10^{-5}$. Entries are mean $\pm$ sample standard deviation over ten repetitions.

Budget	m	U2R	2-SLR	LDP-IHT	Sample-SDP
$\epsilon = 1$	10	1.48844 $\pm$0.05467	11.21592 $\pm$10.06853	3.15212 $\pm$0.72151	2.91723 $\pm$0.41512
	50	1.17335 $\pm$0.03528	3.08118 $\pm$1.14883	2.36752 $\pm$0.39188	1.91000 $\pm$0.31047
	100	1.12972 $\pm$0.02559	3.23649 $\pm$1.62979	2.25214 $\pm$0.45721	1.50741 $\pm$0.12895
$\epsilon = 4$	10	1.43520 $\pm$0.01047	3.07791 $\pm$1.29223	2.44676 $\pm$0.59026	2.05175 $\pm$0.36818
	50	1.15870 $\pm$0.02745	2.08907 $\pm$0.16864	2.36333 $\pm$0.41893	1.48151 $\pm$0.27440
	100	1.13226 $\pm$0.00299	2.19689 $\pm$0.26004	2.32449 $\pm$0.33028	1.37230 $\pm$0.12374

C.3 Additional Results

Table 8: Exp4: Stage 1 support recall and exact recovery. Recall entries are mean $\pm$ sample standard deviation in percentage points; exact recovery is the number of repetitions with $\hat{S} = S^*$ out of ten.

Value	U2R recall in %	2-SLR recall in %	U2R exact	2-SLR exact
Exp4a: vary m; $n = 15000$, $d = 256$, $\epsilon = 4$				
25	3.75 ± 8.44	26.25 ± 13.76	0/10	0/10
50	100.00 ± 0.00	63.75 ± 12.43	10/10	0/10
75	100.00 ± 0.00	72.50 ± 16.46	10/10	1/10
100	100.00 ± 0.00	76.25 ± 12.43	10/10	0/10
Exp4b: vary n; $m = 50$, $d = 256$, $\epsilon = 4$				
8000	100.00 ± 0.00	35.00 ± 12.91	10/10	0/10
10000	100.00 ± 0.00	46.25 ± 10.29	10/10	0/10
12000	100.00 ± 0.00	50.00 ± 10.21	10/10	0/10
15000	100.00 ± 0.00	63.75 ± 12.43	10/10	0/10
18000	100.00 ± 0.00	53.75 ± 15.65	10/10	0/10
20000	100.00 ± 0.00	61.25 ± 12.43	10/10	0/10

Support recovery. Table 8 compares the Stage 1 selectors in Exp4. At $n = 15000$, U2R's mean recall increases from 3.75% at $m = 25$ to 100% at $m = 50$ and remains 100% at $m = 75$ and 100. It recovers the exact support in all ten repetitions at each of these three larger values of m. The 2-SLR recall increases from 26.25% to 76.25% over the same sweep. With $m = 50$, U2R recovers

the exact support in every repetition at all tested user counts from 8000 to 20000. The 2-SLR recall ranges from 35.00% to 63.75%.

Prediction error. At the default setting, U2R and Lasso have mean test MSEs of 0.0629119 and 0.0629186, respectively, close to the residual variance $\sigma^2 = 0.0625$. LDP-IHT and Sample-SDP yield 0.0653332 and 0.0635799. At $\epsilon = 32$, U2R's MSE is 0.0629116; its reduction in coefficient error has little effect on prediction MSE over this budget range.

Prediction in original units. Table 9 gives test MSE in squared degrees Celsius for Exp5. At $\epsilon = 1$, increasing m from 10 to 100 reduces U2R's mean MSE from 3.56667 to 2.70641 and mean RMSE from 1.88827 to $1.64502\,^\circ\mathrm{C}$; mean R^2 increases from 0.87853 to 0.90782. At $m = 100$ and $\epsilon = 4$, the mean MSEs of U2R, Lasso, and Sample-SDP are 2.71249, 2.39565, and 3.28753, respectively.

Table 9: Exp5: Kelmarsh test MSE in squared degrees Celsius. U2R and Sample-SDP use $\delta = 10^{-5}$. Entries are mean $\pm$ sample standard deviation over ten repetitions; Lasso is reused across budgets.

Budget	m	U2R	2-SLR	LDP-IHT	Sample-SDP	Lasso
$\epsilon = 1$	10	3.56667 $\pm$0.13126	26.89088 $\pm$24.15804	7.55337 $\pm$1.72935	6.98929 $\pm$0.98621	2.39625 $\pm$0.00376
	50	2.81101 $\pm$0.08419	7.38104 $\pm$2.75047	5.67181 $\pm$0.93834	4.57575 $\pm$0.74360	2.39572 $\pm$0.00123
	100	2.70641 $\pm$0.06082	7.75405 $\pm$3.90559	5.39543 $\pm$1.09591	3.61122 $\pm$0.30899	2.39565 $\pm$0.00063
$\epsilon = 4$	10	3.43908 $\pm$0.02529	7.37791 $\pm$3.10457	5.86216 $\pm$1.40887	4.91615 $\pm$0.87979	2.39625 $\pm$0.00376
	50	2.77593 $\pm$0.06601	5.00484 $\pm$0.40439	5.66193 $\pm$1.00425	3.54916 $\pm$0.65645	2.39572 $\pm$0.00123
	100	2.71249 $\pm$0.00665	5.26299 $\pm$0.62298	5.56877 $\pm$0.79226	3.28753 $\pm$0.29627	2.39565 $\pm$0.00063

D Omitted Proofs in Section 4

This section provides proofs for the theories in Section 4 of the main text.

D.1 Proof of Theorem 1

We first reduce a robust shuffled DP protocol to a pan-private algorithm, and then prove the lower bound for pan-private algorithms by a hybrid argument over the stream and Assouad's method over the coordinates.

Let $\mathcal{X} = \{\pm 1\}^{d+1}$ be the data universe of Definition 6, let U be the uniform distribution on $\mathcal{X}$, and let $\chi_i(x, y) = yx_i$ for $i \in [d]$. These functions satisfy $\mathbb{E}_U[\chi_i] = 0$ and $\mathbb{E}_U[\chi_i\chi_l] = \mathbb{I}[i = l]$. For $\theta \in \mathbb{R}^d$ let $Q_\theta(x, y) = U(x, y)\big(1 + y\langle x, \theta\rangle\big)$, which is the distribution of one sample in Definition 6 with parameter θ. For $z \in \{-1, 0, 1\}^d$ we write $P_z = Q_{\gamma z}$ and $Z_0(z) = \{i \in [d] : z_i = 0\}$. For $i \in Z_0(z)$ and $b \in \{\pm 1\}$, let $z^{i,b} = z + be_i$. Then

$$P_{z^{i,b}} - P_z = b\gamma\, \chi_i\, U. \tag{21}$$

For a distribution P on $\mathcal{X}$, $M(P^n)$ denotes the distribution of the output of M on n i.i.d. records from P.

Following Cheu & Ullman (2021), a protocol $\mathcal{P}$ is $(\epsilon, \delta, 1/3)$-robust shuffled DP if, for every $\gamma' \geq 1/3$, the shuffled messages of any $\gamma' n$ users that follow the protocol satisfy (ϵ, δ)-DP; this is the robustness in Theorem 1. Let $\mathcal{P}$ be such a non-interactive protocol and let $n' = n/3$. By Cheu & Ullman (2021, Theorem 2.6), there is an (ϵ, δ)-pan-private algorithm $M^{\mathcal{P}}$ that processes a stream of n' records and satisfies, for every distribution P on $\mathcal{X}$,

$$d_{TV}\big(M^{\mathcal{P}}(P^{n'}),\, \mathcal{M}_{\mathcal{P}}(P^n_{(2/9)})\big) \leq e^{-c_1 n}, \qquad P_{(b)} = bP + (1 - b)U,$$

where $c_1 > 0$ is an absolute constant. Since Q_θ is linear in θ, we have $\frac{2}{9}Q_\theta + \frac{7}{9}U = Q_{2\theta/9}$, and $2\theta/9 \in \widetilde{\Theta}_k$ whenever $\theta \in \widetilde{\Theta}_k$. Let Π_1 be the projection onto the unit ℓ_1-ball and define $M = \frac{9}{2}\Pi_1(M^{\mathcal{P}})$. As a post-processing of the output, M is (ϵ, δ)-pan-private. The projection does not increase the distance to any point of the unit ℓ_1-ball, and $\|M - \theta\|_2 \le \frac{11}{2}$ for $\theta \in \widetilde{\Theta}_k$, so

$$\mathbb{E}\big[\|M(Q_\theta^{n'}) - \theta\|_2\big] \le \frac{9}{2}\,\mathbb{E}\big[\|\mathcal{M}_{\mathcal{P}}(Q^n_{2\theta/9}) - \tfrac{2}{9}\theta\|_2\big] + \frac{11}{2}e^{-c_1 n} \qquad \text{for all } \theta \in \widetilde{\Theta}_k. \tag{22}$$

It therefore suffices to show that every (ϵ, δ)-pan-private algorithm M on n' records satisfies

$$\sup_{\theta^* \in \widetilde{\Theta}_k} \mathbb{E}\big[\|M(Q_{\theta^*}^{n'}) - \theta^*\|_2\big] \ge c_2\,\frac{\sqrt{dk}}{n'\epsilon} \tag{23}$$

for a constant $c_2 > 0$. Indeed, combining equation 22 and equation 23 gives $\sup_{\theta^*} \mathbb{E}\|\mathcal{M}_{\mathcal{P}}(Z) - \theta^*\|_2 \ge \frac{2}{3}c_2\frac{\sqrt{dk}}{n\epsilon} - \frac{11}{9}e^{-c_1 n}$. Since $\epsilon \le \mathcal{O}(1)$, we have $\frac{\sqrt{dk}}{n\epsilon} \ge \Omega(\frac{1}{n})$, so for n larger than an absolute constant the last term is at most $\frac{1}{3}c_2\frac{\sqrt{dk}}{n\epsilon}$, which gives the theorem. In the rest of the proof we fix an (ϵ, δ)-pan-private algorithm M with internal states $S_1, \dots, S_{n'}$, and prove equation 23. We take n larger than an absolute constant, so that $\delta = o(1/n)$ gives $n'\delta \le 1/64$.

For $j \in [n']$, a prefix $x_{<j} = (x_1, \dots, x_{j-1})$ and a record $x \in \mathcal{X}$, let $q_j(\cdot \mid x_{<j}, x)$ be the distribution of S_j when the first j records are $(x_{<j}, x)$. Applying Definition 5 at time $t = j$ to two streams that differ in the j-th record shows that, for every fixed $x_{<j}$, the map $x \mapsto q_j(\cdot \mid x_{<j}, x)$ is (ϵ, δ)-DP. By Cheu & Ullman (2021, Lemma 3.7), there is a $(2\epsilon, 0)$-DP map $x \mapsto q'_j(\cdot \mid x_{<j}, x)$ with $d_{TV}\big(q_j(\cdot \mid x_{<j}, x), q'_j(\cdot \mid x_{<j}, x)\big) \le \delta$ for all x. Below we drop $x_{<j}$ from the notation when it is fixed, and write sums over states s; for general state spaces the sums are replaced by integrals of densities with respect to a common dominating measure. Let $\bar{q}'_j(s) = \mathbb{E}_{X\sim U}[q'_j(s \mid X)]$ and

$$h_{j,s}(x) = \frac{q'_j(s \mid x)}{\bar{q}'_j(s)} - 1, \qquad \hat{h}_{j,s}(i) = \mathbb{E}_U\big[h_{j,s}\chi_i\big].$$

Since q'_j is $(2\epsilon, 0)$-DP, $e^{-2\epsilon}\bar{q}'_j(s) \le q'_j(s \mid x) \le e^{2\epsilon}\bar{q}'_j(s)$, so $|h_{j,s}(x)| \le e^{2\epsilon} - 1$. For a distribution P on $\mathcal{X}$, write $q_j \circ P$ for the distribution of S_j when the j-th record is drawn from P. If $P' - P = b\gamma\chi_i U$, then using $\mathbb{E}_U[\chi_i] = 0$,

$$\begin{aligned} d_{TV}(q'_j \circ P', q'_j \circ P) &= \frac{1}{2}\sum_s \Big|\sum_x q'_j(s \mid x)\big(P'(x) - P(x)\big)\Big| \\ &= \frac{\gamma}{2}\sum_s \big|\mathbb{E}_U[q'_j(s \mid X)\chi_i(X)]\big| = \frac{\gamma}{2}\,\mathbb{E}_{s\sim\bar{q}'_j}\big|\hat{h}_{j,s}(i)\big|. \end{aligned} \tag{24}$$

Fix $z \in \{-1, 0, 1\}^d$, $i \in Z_0(z)$ and $b \in \{\pm 1\}$. For $j = 0, \dots, n'$, let H_j be the distribution of the output of M when $X_1, \dots, X_j \sim P_z$ and $X_{j+1}, \dots, X_{n'} \sim P_{z^{i,b}}$, all independent. Then $H_0 = M(P_{z^{i,b}}^{n'})$ and $H_{n'} = M(P_z^{n'})$. In H_{j-1} and H_j the prefix $X_{<j}$ has the same distribution P_z^{j-1}, the suffix $X_{>j}$ has the same distribution $P_{z^{i,b}}^{n'-j}$ and is independent of $(X_{<j}, X_j)$, and only X_j differs. Since M processes the stream one record at a time, its output is a post-processing of $(S_j, X_{>j})$. Therefore

$$d_{TV}(H_{j-1}, H_j) \le \mathbb{E}_{X_{<j}\sim P_z^{j-1}}\, d_{TV}\big(q_j \circ P_{z^{i,b}},\, q_j \circ P_z\big) \le \mathbb{E}_{X_{<j}\sim P_z^{j-1}}\Big[\frac{\gamma}{2}\,\mathbb{E}_{s\sim\bar{q}'_j}\big|\hat{h}_{j,s}(i)\big|\Big] + 2\delta,$$

where the second step replaces q_j by q'_j at a cost of 2δ and then applies equation 21 and equation 24. By the triangle inequality,

$$d_{TV}\big(M(P_{z^{i,b}}^{n'}), M(P_z^{n'})\big) \le \frac{\gamma}{2}\sum_{j=1}^{n'} \mathbb{E}_{X_{<j}\sim P_z^{j-1}}\mathbb{E}_{s\sim\bar{q}'_j}\big|\hat{h}_{j,s}(i)\big| + 2n'\delta. \tag{25}$$

The distribution of $X_{<j}$ and the maps q'_j in equation 25 depend on z but not on i or b.

Since $\chi_1, \dots, \chi_d$ are orthonormal in $L_2(U)$, Bessel's inequality gives $\sum_{i\in[d]} \hat{h}_{j,s}(i)^2 \le \mathbb{E}_U[h_{j,s}^2] \le (e^{2\epsilon} - 1)^2$, and by the Cauchy–Schwarz inequality $\sum_{i\in Z_0(z)} |\hat{h}_{j,s}(i)| \le \sqrt{d}\,(e^{2\epsilon} - 1)$. Summing

equation 25 over $i \in Z_0(z)$ and $b \in \{\pm 1\}$, and exchanging the sum over i with the expectations, which do not depend on i, we obtain

$$T(z) := \sum_{i \in Z_0(z)} T_i(z) \le \gamma n' \sqrt{d}\,(e^{2\epsilon}-1) + 4dn'\delta, \qquad T_i(z) := \sum_{b \in \{\pm 1\}} d_{TV}\big(M(P^{n'}_{z^{i,b}}), M(P^{n'}_{z})\big). \tag{26}$$

Note that equation 26 bounds the sum over coordinates only, and this is all that is used below.

Draw Z as in Definition 6, let $\theta_Z = \gamma Z$, draw the stream $X \sim P^{n'}_Z$ given Z, and let $\hat\theta = M(X)$. For $i \in [d]$ let $G_i = \{\|Z_{-i}\|_0 \le k-1\}$ and $\psi_i = \operatorname{sign}(\hat\theta_i)$, with $\operatorname{sign}(0) = +1$. Condition on Z_{-i} and on $Z_i \neq 0$, and let $z = (Z_{-i}, 0)$, so that $z_i = 0$. Then Z_i is uniform on $\{\pm 1\}$ and $X \sim P^{n'}_{z^{i,Z_i}}$, so by Le Cam's method and the triangle inequality,

$$\mathbb{P}\big(\psi_i \neq Z_i \mid Z_{-i}, Z_i \neq 0\big) \ge \frac{1}{2}\Big(1 - d_{TV}\big(M(P^{n'}_{z^{i,+1}}), M(P^{n'}_{z^{i,-1}})\big)\Big) \ge \frac{1}{2}\big(1 - T_i(z)\big).$$

Let $N = \sum_{i \in [d]} \mathbb{I}[Z_i \neq 0,\, G_i,\, \psi_i \neq Z_i]$. If $N \ge 1$, some i has $Z_i \neq 0$ and $\|Z_{-i}\|_0 \le k-1$, so $\|Z\|_0 \le k$ and $N \le \|Z\|_0 \le k$. Each counted coordinate has $|\hat\theta_i - \gamma Z_i| \ge \gamma$. Hence

$$\|\hat\theta - \theta_Z\|_2\, \mathbb{I}[\|Z\|_0 \le k] \ge \gamma\sqrt{N} \ge \frac{\gamma N}{\sqrt{k}}. \tag{27}$$

Since Z_i is independent of Z_{-i} and $\mathbb{P}(Z_i \neq 0) = \frac{k}{2d}$,

$$\mathbb{E}[N] \ge \frac{k}{4d} \sum_{i=1}^{d} \Big(\mathbb{P}(G_i) - \mathbb{E}\big[T_i\big((Z_{-i}, 0)\big)\big]\Big).$$

We bound the two sums in turn. First, $\|Z_{-i}\|_0 \sim \mathrm{Bin}(d-1, \frac{k}{2d})$ has mean less than $\frac{k}{2}$, so Markov's inequality gives $\mathbb{P}(G_i^c) \le \frac{1}{2}$ and $\sum_i \mathbb{P}(G_i) \ge \frac{d}{2}$. Second, $(Z_{-i}, 0) = Z$ on the event $\{Z_i = 0\}$, which is independent of Z_{-i} and has probability $1 - \frac{k}{2d} \ge \frac{1}{2}$, so by equation 26,

$$\sum_{i=1}^{d} \mathbb{E}\big[T_i\big((Z_{-i}, 0)\big)\big] = \sum_{i=1}^{d} \frac{\mathbb{E}\big[T_i(Z)\,\mathbb{I}[Z_i = 0]\big]}{\mathbb{P}(Z_i = 0)} \le 2\,\mathbb{E}\big[T(Z)\big] \le 2\gamma n' \sqrt{d}\,(e^{2\epsilon} - 1) + 8dn'\delta.$$

Combining the two bounds,

$$\mathbb{E}[N] \ge \frac{k}{8} - \frac{k\gamma n'(e^{2\epsilon} - 1)}{2\sqrt{d}} - 2kn'\delta.$$

We now choose $\gamma = \frac{\sqrt{d}}{8n'(e^{2\epsilon}-1)}$, so that the second term equals $\frac{k}{16}$, and use $n'\delta \le \frac{1}{64}$, so that the third term is at most $\frac{k}{32}$. This gives $\mathbb{E}[N] \ge \frac{k}{32}$, and by equation 27,

$$\sup_{\theta^* \in \widetilde{\Theta}_k} \mathbb{E}\big[\|M(Q^{n'}_{\theta^*}) - \theta^*\|_2\big] \ge \mathbb{E}\big[\|\hat\theta - \theta_Z\|_2\, \mathbb{I}[\|Z\|_0 \le k]\big] \ge \frac{\gamma\sqrt{k}}{32} = \frac{\sqrt{dk}}{256\, n'(e^{2\epsilon} - 1)}.$$

Finally, for $0 < \epsilon \le c_0$ with an absolute constant c_0, we have $e^{2\epsilon} - 1 \le 2e^{2c_0}\epsilon$, which gives equation 23 with $c_2 = \frac{1}{512 e^{2c_0}}$ and completes the proof.

D.2 Proof of Theorem 2

We first give a lemma on the binomial mechanism. It is used in the privacy part of this proof, and again in the proofs of Proposition 1 and Theorem 3.

Lemma 1. *Let $s \ge 1$, $\epsilon > 0$ and $\delta \in (0,1)$ with $\epsilon \le \log(2s/\delta)$, and let $\Delta \ge \sqrt{s}\,\epsilon$. Let $B = (B_1, \ldots, B_s)$ have independent entries $B_j \sim \mathrm{Bin}(M, 1/2)$, where*

$$M \ge \frac{32\Delta^2 \log(2s/\delta)}{\epsilon^2}. \tag{28}$$

Then for any $v, v' \in \mathbb{Z}^s$ with $\|v - v'\|_2 \le \Delta$ and any $E \subseteq \mathbb{Z}^s$,

$$\mathbb{P}[v + B \in E] \le e^{\epsilon}\, \mathbb{P}[v' + B \in E] + \delta.$$

Only an upper bound on $\|v-v'\|_2$ is used, so the condition $\Delta \geq \sqrt{s}\,\epsilon$ can always be met by taking Δ larger. Subtracting a constant from $v+B$, for example $M/2$ on each coordinate, is a post-processing and keeps the guarantee.

Proof of Lemma 1. Let $a = v' - v$ and $A = M/2$. For $t \in \{0, \dots, M\}^s$ define

$$\ell(t) = \sum_{j=1}^{s} \log \frac{\mathbb{P}[B_j = t_j]}{\mathbb{P}[B_j = t_j - a_j]},$$

with $\ell(t) = +\infty$ if $t_j - a_j \notin \{0, \dots, M\}$ for some j. For every output o, $\mathbb{P}[v + B = o] = e^{\ell(o-v)}\,\mathbb{P}[v' + B = o]$, so

$$\mathbb{P}[v + B \in E] \leq e^{\epsilon}\,\mathbb{P}[v' + B \in E] + \mathbb{P}[\ell(B) > \epsilon],$$

and it suffices to show $\mathbb{P}[\ell(B) > \epsilon] \leq \delta$.

Write $w_j = B_j - A$, let $W = \sqrt{M \log(4s/\delta)/2}$, and let $G = \{\max_j |w_j| \leq W\}$. Each w_j is a sum of M independent terms in $[-1/2, 1/2]$, so Hoeffding's inequality gives $\mathbb{P}[|w_j| > W] \leq 2e^{-2W^2/M} = \delta/(2s)$, and $\mathbb{P}[G^c] \leq \delta/2$. We check that $W + \Delta \leq A/2$. From $\Delta \geq \sqrt{s}\,\epsilon$ and equation 28, $M \geq 32 s \log(2s/\delta) \geq 32 \log(2s/\delta)$. Since $\log(4s/\delta) \leq 2\log(2s/\delta)$, we get $W \leq \sqrt{M \log(2s/\delta)} \leq M/\sqrt{32}$. From equation 28 and $\epsilon \leq \log(2s/\delta)$,

$$\Delta \leq \frac{\epsilon\sqrt{M}}{\sqrt{32 \log(2s/\delta)}} \leq \sqrt{\frac{M \log(2s/\delta)}{32}} \leq \frac{M}{32}.$$

Hence $W + \Delta \leq M/4 = A/2$.

Fix a coordinate j with $a_j > 0$ and drop the index j. On G, $B = A + w$ and

$$\frac{\mathbb{P}[B = A + w]}{\mathbb{P}[B = A + w - a]} = \frac{\binom{M}{A+w}}{\binom{M}{A+w-a}} = \prod_{r=1}^{a} \frac{A - w + r}{A + w - a + r}.$$

Let $x_r = (w-r)/A$ and $y_r = (w-a+r)/A$. Since $|w|+a \leq A/2$, we have $x_r < 1$ and $|y_r| \leq 1/2$. Using $\log(1-x) \leq -x$ and $-\log(1+y) \leq -y + y^2$ for $|y| \leq 1/2$, and $|w - a + r| \leq |w| + a$ for $1 \leq r \leq a$,

$$\log \frac{A - w + r}{A + w - a + r} = \log(1 - x_r) - \log(1 + y_r) \leq \frac{a - 2w}{A} + \frac{(|w| + a)^2}{A^2}.$$

Summing over r gives the bound $(a^2 - 2aw)/A + a(|w|+a)^2/A^2$ for this coordinate. If $a_j < 0$, the same bound applied to $M - B_j$, which has the same distribution, gives $(a_j^2 - 2a_j w_j)/A + |a_j|(|w_j| + |a_j|)^2/A^2$; if $a_j = 0$ the term is zero. Summing over j and using $A = M/2$, on G we obtain

$$\ell(B) \leq \frac{2\|a\|_2^2}{M} - \frac{4\langle a, w\rangle}{M} + \frac{4}{M^2} \sum_{j=1}^{s} |a_j| \big(|w_j| + |a_j|\big)^2.$$

We bound the three terms. For the first, by equation 28 and $\epsilon \leq \log(2s/\delta)$,

$$\frac{2\|a\|_2^2}{M} \leq \frac{2\Delta^2}{M} \leq \frac{\epsilon^2}{16 \log(2s/\delta)} \leq \frac{\epsilon}{16}.$$

For the third, $(|w_j| + |a_j|)^2 \leq 2W^2 + 2\Delta^2$ on G, $\sum_j |a_j| \leq \sqrt{s}\,\Delta$ and $W^2 \leq M \log(2s/\delta)$, so with equation 28, $\Delta \geq \sqrt{s}\,\epsilon$ and $\epsilon \leq \log(2s/\delta)$,

$$\begin{aligned} \frac{4}{M^2} \sum_{j=1}^{s} |a_j| \big(|w_j| + |a_j|\big)^2 &\leq \frac{8\sqrt{s}\,\Delta \log(2s/\delta)}{M} + \frac{8\sqrt{s}\,\Delta^3}{M^2} \\ &\leq \frac{\sqrt{s}\,\epsilon^2}{4\Delta} + \frac{\sqrt{s}\,\epsilon^4}{128\,\Delta \log^2(2s/\delta)} \leq \frac{\epsilon}{4} + \frac{\epsilon}{128}. \end{aligned}$$

For the second, $\langle a, w\rangle$ is a sum of sM independent terms, where the terms of coordinate j lie in an interval of length $|a_j|$. By Hoeffding's inequality and equation 28,

$$\mathbb{P}\Big[-\frac{4\langle a,w\rangle}{M} > \frac{\epsilon}{2}\Big] \le \exp\Big(-\frac{2(\epsilon M/8)^2}{M\|a\|_2^2}\Big) \le \exp\Big(-\frac{\epsilon^2 M}{32\Delta^2}\Big) \le \frac{\delta}{2s} \le \frac{\delta}{2}.$$

On G and outside this event, $\ell(B) \le \epsilon/16 + \epsilon/2 + \epsilon/4 + \epsilon/128 < \epsilon$. Hence $\mathbb{P}[\ell(B) > \epsilon] \le \delta/2 + \delta/2 = \delta$. □

Proof of Theorem 2. The parameters are set as in Table 3. We use $n \ge 20$, which the summation step requires, and $\epsilon \le \log(4D/\delta)$, which holds for d larger than an absolute constant since $\epsilon \le \mathcal{O}(1)$. By the choice of p and C_F, $pC_F \ge c_0\sqrt{n}\,\epsilon/\sqrt{\log(4D/\delta)}$. Since $\epsilon \le \mathcal{O}(1)$, the third term of equation 8, with a large enough constant, then gives

$$pC_F \ge \sqrt{D}\max\{1,\epsilon\}, \qquad \text{so} \qquad \sqrt{D}\,\epsilon \le \Delta_2^{(p)} \le 3pC_F. \tag{29}$$

We first prove privacy. The rounding step of Algorithm 2 has the same distribution as $\bar{U}_{i,j} = \lfloor p\bar{F}_{i,j} + u_{i,j}\rfloor$ with $u_{i,j}$ independent and uniform on $(0,1]$, and we use this form. Fix two neighboring datasets that differ in the sample of user i_0, and fix $u = (u_{i,j})$. The rounded vectors of all users other than i_0 are the same for the two datasets. For user i_0, since $|\lfloor a+u\rfloor - \lfloor b+u\rfloor| \le |a-b|+1$ and $\|\bar{F}_{i_0} - \bar{F}'_{i_0}\|_2 \le 2C_F$,

$$\|\bar{U}_{i_0} - \bar{U}'_{i_0}\|_2 \le p\|\bar{F}_{i_0} - \bar{F}'_{i_0}\|_2 + \sqrt{D} \le 2pC_F + \sqrt{D} = \Delta_2^{(p)}.$$

Let $T_j = \sum_{i=1}^n (\bar{U}_{i,j} + \zeta_{i,j})$ for $j \in [D]$. Then $\sum_{i=1}^n \zeta_{i,j} = B_j - nN_0/2$ with $B_j \sim \mathrm{Bin}(nN_0, 1/2)$ independent across j, and $nN_0 \ge M \ge 32(\Delta_2^{(p)})^2\log(4D/\delta)/\epsilon^2$. By equation 29 and $\epsilon \le \log(4D/\delta)$, Lemma 1 applies with $s = D$, $\Delta = \Delta_2^{(p)}$ and $\delta/2$ in place of δ. Hence, for fixed u, the vector $T = (T_1, \dots, T_D)$ is $(\epsilon, \delta/2)$-DP as a function of the dataset. The variables u have the same distribution for both datasets, so T is also $(\epsilon, \delta/2)$-DP without conditioning on u.

We next relate the view of the aggregator to T. Since $|\bar{U}_{i,j}| \le pC_F + 1$, $|\zeta_{i,j}| \le N_0/2$ and $N_0 \ge 2$, we have $|T_j| \le n(pC_F + 1 + N_0/2) < q/2$. So no wrap-around occurs, and T_j is the value S_j computed by the aggregator. On coordinate j, the users run the summation protocol $\widetilde{V}_{r,n}$ of Balle et al. (2020) on the inputs $W_{1,j}, \dots, W_{n,j} \in \mathbb{Z}_q$, with r shuffled shares and one more share. Since $q > 4n$, the setting of r gives $r \ge 3$, and $n \ge 20$. By Balle et al. (2020, Corollary 6.1), the protocol has worst-case statistical security with parameter $\sigma_{\mathrm{sec}} = \log_2(8De^\epsilon/\delta)$: for two input vectors with the same sum in $\mathbb{Z}_q$, the total variation distance between the messages of coordinate j is at most $2^{-\sigma_{\mathrm{sec}}} = e^{-\epsilon}\delta/(8D)$. Let $\mathcal{V}^{\mathrm{sim}}$ be the view obtained when, on every coordinate j, the inputs are replaced by $(T_j \bmod q, 0, \dots, 0)$. It is a post-processing of T, hence $(\epsilon, \delta/2)$-DP. The D coordinates use independent shares and separate shufflers, so coupling them one at a time gives, for every dataset,

$$d_{TV}\big(\mathcal{V}_{\mathcal{P}}, \mathcal{V}^{\mathrm{sim}}\big) \le D \cdot \frac{e^{-\epsilon}\delta}{8D} = \frac{e^{-\epsilon}\delta}{8}.$$

By Balle et al. (2020, Lemma F.2), $\mathcal{V}_{\mathcal{P}}$ is $(\epsilon, \delta/2 + (1+e^\epsilon)e^{-\epsilon}\delta/8)$-DP. Since $(1+e^\epsilon)e^{-\epsilon}/8 \le 1/4$, Algorithm 2 satisfies (ϵ, δ)-shuffled DP.

We now bound the error. Under Assumption 3, $\|v(H_i)\|_2 \le \|x_i x_i^\top\|_F = \|x_i\|_2^2$, so

$$\|F_i\|_2 \le \|x_i\|_2^2 + \|x_i\|_2|y_i| \le c_x^2 d + c_x c_y\sqrt{d} \le C_F$$

for $c_0 \ge c_x^2 + c_x c_y$ and $n \ge 3$. Hence $\bar{F}_i = F_i$ for every i. The aggregator computes $\hat{F} = T/(np)$, and

$$\hat{F} = \frac{1}{n}\sum_{i=1}^n F_i + E^{\mathrm{rd}} + E^{\mathrm{bin}}, \qquad E_j^{\mathrm{rd}} = \frac{1}{np}\sum_{i=1}^n \big(\bar{U}_{i,j} - pF_{i,j}\big), \qquad E_j^{\mathrm{bin}} = \frac{B_j - nN_0/2}{np}.$$

Let

$$\eta = \frac{C_F\sqrt{\log(4D/\delta)\log D}}{n\epsilon}.$$

The terms $\bar{U}_{i,j} - pF_{i,j}$ are independent, have mean zero and lie in $(-1, 1]$. By Hoeffding's inequality and a union bound over $j \in [D]$, with probability at least $1 - \mathcal{O}(d^{-2})$,

$$\|E^{\mathrm{rd}}\|_\infty \le \mathcal{O}\Big(\frac{\sqrt{\log D}}{\sqrt{n}\,p}\Big) \le \mathcal{O}(\eta),$$

where the second step uses $\frac{1}{\sqrt{n}\,p} \le \frac{d \log n \sqrt{\log(4D/\delta)}}{n\epsilon} = \frac{C_F \sqrt{\log(4D/\delta)}}{c_0 n \epsilon}$. Each $B_j - nN_0/2$ is a sum of nN_0 independent terms in $[-1/2, 1/2]$, and $nN_0 \le M + 2n$. By equation 29, $\sqrt{M} \le \mathcal{O}\big(pC_F\sqrt{\log(4D/\delta)}/\epsilon\big)$, so with probability at least $1 - \mathcal{O}(d^{-2})$,

$$\|E^{\mathrm{bin}}\|_\infty \le \mathcal{O}\Big(\frac{\sqrt{(M+2n)\log D}}{np}\Big) \le \mathcal{O}\Big(\eta + \frac{\sqrt{\log D}}{\sqrt{n}\,p}\Big) = \mathcal{O}(\eta).$$

Let $\bar{H} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and $\bar{g} = \frac{1}{n}\sum_{i=1}^n x_i y_i$. Since $\hat{H}$ and $\hat{g}$ are read from $\hat{F}$, we obtain $\|\hat{H} - \bar{H}\|_{\infty,\infty} \le \mathcal{O}(\eta)$ and $\|\hat{g} - \bar{g}\|_\infty \le \mathcal{O}(\eta)$.

For the statistical error, the entries of $x_i x_i^\top$ are bounded by c_x^2, and $\bar{g} - \bar{H}\theta^* = \frac{1}{n}\sum_{i=1}^n x_i \zeta_i$, where, given x_i, each entry $x_{i\ell}\zeta_i$ is sub-Gaussian with mean zero and variance at most $c_x^2\sigma^2$ by Assumption 1. By Hoeffding's inequality and a union bound, with probability at least $1 - \mathcal{O}(d^{-2})$,

$$\|\bar{H} - \Sigma\|_{\infty,\infty} \le \mathcal{O}\Big(c_x^2\sqrt{\frac{\log d}{n}}\Big), \qquad \|\bar{g} - \bar{H}\theta^*\|_\infty \le \mathcal{O}\Big(c_x\sigma\sqrt{\frac{\log d}{n}}\Big).$$

For any $w \in \mathbb{R}^d$, $\|(\hat{H} - \Sigma)w\|_\infty \le d\,\|\hat{H} - \Sigma\|_{\infty,\infty}\|w\|_\infty$. The first two terms of equation 8, with large enough constants, give $d\,c_x^2\sqrt{\log d/n} \le \kappa_\infty/4$ and $d\eta \le \kappa_\infty/4$, so by Assumption 3,

$$\|\hat{H}w\|_\infty \ge \|\Sigma w\|_\infty - \|(\hat{H} - \Sigma)w\|_\infty \ge \frac{\kappa_\infty}{2}\|w\|_\infty.$$

Hence $\hat{H}$ is invertible and $\|\hat{H}^{-1}z\|_\infty \le \frac{2}{\kappa_\infty}\|z\|_\infty$ for all z. Since

$$\hat{g} - \hat{H}\theta^* = (\hat{g} - \bar{g}) - (\hat{H} - \bar{H})\theta^* + (\bar{g} - \bar{H}\theta^*)$$

and $\|(\hat{H} - \bar{H})\theta^*\|_\infty \le \|\hat{H} - \bar{H}\|_{\infty,\infty}\|\theta^*\|_1 \le \mathcal{O}(\eta)$, we get

$$\|\tilde{\theta} - \theta^*\|_\infty = \|\hat{H}^{-1}(\hat{g} - \hat{H}\theta^*)\|_\infty \le \frac{2}{\kappa_\infty}\,\mathcal{O}\Big(c_x\sigma\sqrt{\frac{\log d}{n}} + \eta\Big) \le \lambda_n$$

for a large enough c_1. By Zhu et al. (2024, Lemma 29), $\|\mathcal{S}_{\lambda_n}(\tilde{\theta}) - \theta^*\|_2 \le 4\sqrt{k}\,\lambda_n$. The projection onto $\{\theta : \|\theta\|_1 \le 1\}$ does not increase the distance to θ^*, which lies in this set. Hence, with κ_∞ a constant,

$$\|\hat{\theta} - \theta^*\|_2 \le 4\sqrt{k}\,\lambda_n \le \mathcal{O}\Big(\sigma\sqrt{\frac{k\log d}{n}} + \frac{d\sqrt{k}\log n\sqrt{\log(4D/\delta)\log D}}{n\epsilon}\Big),$$

which gives the rate of Theorem 2. The events above hold together with probability at least $1 - \mathcal{O}(d^{-2})$. □

E Omitted Proofs in Section 5

This section provides the rigorous proofs for the theoretical bounds in Section 5.

E.1 Proof of Proposition 1

Proof of Proposition 1. Throughout the proof we take $k = |S^*|$. The parameters of Algorithm 3 are set as in Table 4, in particular $\gamma = C_\gamma \log(1/\delta)/(n_1\epsilon^2)$. We use $\epsilon \le \epsilon_0$ for an absolute constant ϵ_0, and take $\delta \le 1/8$ small enough that $\epsilon_0/2 \le \log(8/\delta)$. By Assumption 4 with $v = e_\ell$ for $\ell \in S^*$, we have $\kappa \le \Sigma_{\ell\ell} \le 1$. Since $k\log(ed/k) \ge \log d$, the second term of equation 10 gives $m \ge \Omega(\log d)$. The proof has three parts: the local Lasso of one user of Stage 1 selects S^* with probability at least $1 - 1/(32k)$; Stage 1 is (ϵ, δ)-shuffled DP; and the debiased counts separate S^* from its complement.

The first part is about one user, so it needs no union bound over the users, and a user with $\hat{S}_i \neq S^*$ is allowed. The privacy part is not needed for the statement of the proposition, and is used in the proof of Theorem 3.

We first show that $\mathbb{P}[G_i^c] \leq 1/(32k)$ for every $i \in [n_1]$, where G_i is the event $\hat{S}_i = S^*$. Fix $i \in [n_1]$ and let $\zeta_i \in \mathbb{R}^m$ be the noise of user i. For $\ell \in [d]$, write $x_\ell = \langle \Sigma^{1/2} e_\ell, \Sigma^{-1/2} x \rangle$. By Assumption 2 and $\Sigma_{\ell\ell} \leq 1$, we have $\|x_\ell\|_{\psi_2} \leq \kappa_x$. By Assumption 1, $\|\zeta\|_{\psi_2} \leq \mathcal{O}(\sigma)$ and $\mathbb{E}[x_\ell \zeta] = \mathbb{E}[x_\ell \mathbb{E}[\zeta \mid x]] = 0$. So each entry of $\frac{1}{m} X_i^\top \zeta_i$ is an average of m independent sub-exponential variables with mean zero and ψ_1-norm at most $\mathcal{O}(\kappa_x \sigma)$. Since $k \leq d$, we have $\log(128dk) \leq 2\log(128d)$. Bernstein's inequality, $m \geq \Omega(\log d)$ and a union bound over $\ell \in [d]$ give, with probability at least $1 - 1/(64k)$,

$$\Big\| \frac{1}{m} X_i^\top \zeta_i \Big\|_\infty \leq \mathcal{O}\Big(\kappa_x \sigma \sqrt{\frac{\log d}{m}} \Big) \leq \frac{\lambda_{\mathrm{loc}}}{2}, \tag{30}$$

where the last step holds for c_λ large enough. Let $\mathcal{C} = \{ v \in \mathbb{R}^d : \|v_{(S^*)^c}\|_1 \leq 3\|v_{S^*}\|_1 \}$. The rows of X_i are $x = \Sigma^{1/2}(\Sigma^{-1/2} x)$ with $\Sigma^{-1/2} x$ isotropic and sub-Gaussian. By Assumption 4, $\max_\ell \Sigma_{\ell\ell} \leq 1$ and the result of Rudelson & Zhou (2013) for such designs, there are constants such that, when $m \geq \Omega(k \log(ed/k)/\kappa)$, with probability at least $1 - 2e^{-cm}$,

$$\frac{1}{m} \|X_i v\|_2^2 \geq \frac{1}{2} v^\top \Sigma v \geq \frac{\kappa}{2} \|v\|_2^2 \qquad \text{for all } v \in \mathcal{C}.$$

Since $\kappa \leq 1$ and $k \log(ed/k) \geq k$, the second term of equation 10 gives the required m and $2e^{-cm} \leq 1/(64k)$. So the two events hold together with probability at least $1 - 1/(32k)$.

On these two events, let $\Delta = \hat{\theta}^{(i)} - \theta^*$. Since $\hat{\theta}^{(i)}$ minimizes the Lasso objective and $y_i = X_i \theta^* + \zeta_i$,

$$\begin{aligned} \frac{1}{2m} \|X_i \Delta\|_2^2 &\leq \frac{1}{m} \langle X_i^\top \zeta_i, \Delta \rangle + \lambda_{\mathrm{loc}} \big(\|\theta^*\|_1 - \|\theta^* + \Delta\|_1 \big) \\ &\leq \frac{3\lambda_{\mathrm{loc}}}{2} \|\Delta_{S^*}\|_1 - \frac{\lambda_{\mathrm{loc}}}{2} \|\Delta_{(S^*)^c}\|_1, \end{aligned}$$

where we used equation 30 and $\|\theta^*\|_1 - \|\theta^* + \Delta\|_1 \leq \|\Delta_{S^*}\|_1 - \|\Delta_{(S^*)^c}\|_1$. The left-hand side is non-negative, so $\Delta \in \mathcal{C}$, and then

$$\frac{\kappa}{4} \|\Delta\|_2^2 \leq \frac{1}{2m} \|X_i \Delta\|_2^2 \leq \frac{3\lambda_{\mathrm{loc}}}{2} \sqrt{k}\, \|\Delta\|_2.$$

Hence $\|\Delta\|_\infty \leq \|\Delta\|_2 \leq 6\lambda_{\mathrm{loc}} \sqrt{k}/\kappa \leq 3\rho/8$, where the last step uses $\lambda_{\mathrm{loc}} \leq \kappa\rho/(16\sqrt{k})$, which holds under the first term of equation 10. For $\ell \in S^*$, $|\hat{\theta}_\ell^{(i)}| \geq \rho - 3\rho/8 > \rho/2$, and for $\ell \notin S^*$, $|\hat{\theta}_\ell^{(i)}| \leq 3\rho/8 < \rho/2$. So $\hat{S}_i = S^*$, and therefore $\mathbb{P}[G_i^c] \leq 1/(32k)$.

We next show that Stage 1 satisfies (ϵ, δ)-shuffled DP. For a pair of distributions, let $H_{e^\epsilon}(P \,\|\, Q) = \sup_E \left(P(E) - e^\epsilon Q(E) \right)$; a mechanism is (ϵ, δ)-DP if H_{e^ϵ} between its outputs on any two neighboring datasets, in both orders, is at most δ. We use that H_{e^ϵ} is jointly convex, and that for product distributions $H_{e^{\epsilon_1+\epsilon_2}}(P_1 \times P_2 \,\|\, Q_1 \times Q_2) \leq \delta_1 + \delta_2$ whenever $H_{e^{\epsilon_j}}(P_j \,\|\, Q_j) \leq \delta_j$ for $j = 1, 2$ (Dwork & Roth, 2014). Fix two datasets that differ in the local data of user i. We proceed in five steps.

1. The ℓ-th shuffler outputs a multiset of bits, which is determined by the count $C_\ell = \sum_{j=1}^{n_1} r_{j\ell}$. So the view is $(C_1, \ldots, C_d)$. Given $v = (v_1, \ldots, v_{n_1})$, the bits $r_{j\ell}$ are independent, and so are $C_1, \ldots, C_d$.

2. The indices v_j with $j \neq i$ have the same distribution under the two datasets, while $v_i \sim \mathrm{Unif}(\hat{S}_i)$ and $v_i' \sim \mathrm{Unif}(\hat{S}_i')$ may differ. By joint convexity, it suffices to bound H_{e^ϵ} between the views given $(v_{-i}, v_i = a)$ and $(v_{-i}, v_i = b)$, for any fixed v_{-i} and $a, b \in [d]$. If $a = b$ the two views are equal, so we take $a \neq b$.

3. Given (v_{-i}, a) and (v_{-i}, b), the counts C_ℓ with $\ell \notin \{a, b\}$ have the same distribution, and all counts are independent. By the composition bound above, it suffices to show that C_a, and in the same way C_b, is $(\epsilon/2, \delta/2)$-indistinguishable between the two cases.

4. Consider C_a. Write $r_{ia} = \mathbb{I}[v_i = a] \oplus \beta$ with $\beta \sim \mathrm{Ber}(\gamma)$, which has the same distribution in both cases; by joint convexity we fix β. For $j \neq i$, since $\gamma \leq 1/2$, the bit $r_{ja} = \mathbb{I}[v_j = a] \oplus \mathrm{Ber}(\gamma)$ has the same distribution as a uniform bit with probability 2γ and $\mathbb{I}[v_j = a]$ otherwise. Let $J \subseteq [n_1] \setminus \{i\}$ be the users that output a uniform bit, so that $K = |J| \sim \mathrm{Bin}(n_1 - 1, 2\gamma)$ does not depend on the data. Given J,

$$C_a = \sum_{j \notin J,\, j \neq i} \mathbb{I}[v_j = a] + r_{ia} + \mathrm{Bin}(K, 1/2),$$

and the integer part before $\mathrm{Bin}(K, 1/2)$ changes by one between the two cases. Lemma 1 with $s = 1$, $\Delta = \max\{1, \epsilon/2\}$, $\epsilon/2$ in place of ϵ and $\delta/4$ in place of δ applies, since $\epsilon/2 \leq \log(8/\delta)$. It shows that, given J, C_a is $(\epsilon/2, \delta/4)$-indistinguishable as long as

$$K \geq \frac{128 \max\{1, \epsilon^2/4\} \log(8/\delta)}{\epsilon^2}. \tag{31}$$

By joint convexity over J, $H_{e^{\epsilon/2}}$ for C_a is at most $\mathbb{P}[\text{equation 31 fails}] + \delta/4$.

5. Since $n_1 \geq 8$ by equation 11, $\mathbb{E}[K] = 2\gamma(n_1 - 1) \geq \frac{7}{4}\gamma n_1$, and a Chernoff bound gives $\mathbb{P}[K < \mathbb{E}[K]/2] \leq e^{-\mathbb{E}[K]/8} \leq e^{-7\gamma n_1/32}$. We take $C_\gamma = 256 \max\{1, \epsilon_0^2/4\}$. With $\delta \leq 1/8$ we have $\log(8/\delta) \leq \frac{3}{2}\log(1/\delta)$ and $\log(4/\delta) \leq \frac{4}{3}\log(1/\delta)$, so $\gamma n_1 = C_\gamma \log(1/\delta)/\epsilon^2$ gives both $\mathbb{E}[K]/2 \geq \frac{7}{8}\gamma n_1$, which is at least the right-hand side of equation 31, and $e^{-7\gamma n_1/32} \leq \delta/4$. Hence C_a is $(\epsilon/2, \delta/2)$-indistinguishable, and the same holds for C_b.

By steps 2 and 3, Stage 1 satisfies (ϵ, δ)-shuffled DP.

We finally show that $\hat{S} = S^*$. For $\ell \in [d]$, let $q_\ell = \mathbb{P}[v_i = \ell]$ and $w_\ell = |\{i \in [n_1] : v_i = \ell\}|$. The local datasets of the users of Stage 1 are i.i.d., and each user draws v_i with its own randomness, so $v_1, \ldots, v_{n_1}$ are i.i.d. and $w_\ell \sim \mathrm{Bin}(n_1, q_\ell)$. On G_i, v_i is uniform on S^*; off G_i, v_i may be any index. So for $\ell \in S^*$ and $\ell' \notin S^*$,

$$q_\ell \geq \frac{\mathbb{P}[G_i]}{k} \geq \frac{31}{32k}, \qquad q_{\ell'} = \mathbb{P}\big[v_i = \ell',\, G_i^c\big] \leq \mathbb{P}[G_i^c] \leq \frac{1}{32k}.$$

Let $L = \log(dn)$. For $\ell \in S^*$, w_ℓ dominates $\mathrm{Bin}(n_1, 31/(32k))$, and for $\ell' \notin S^*$, $w_{\ell'}$ is dominated by $\mathrm{Bin}(n_1, 1/(32k))$. Chernoff bounds give

$$\mathbb{P}\Big[w_\ell < \frac{n_1}{2k}\Big] \leq e^{-n_1/(9k)}, \qquad \mathbb{P}\Big[w_{\ell'} > \frac{n_1}{8k}\Big] \leq e^{-n_1/(18k)}.$$

By a union bound over $[d]$ and $n_1 \geq \Omega(kL)$ from equation 11, with probability at least $1 - \mathcal{O}(n^{-2})$,

$$w_\ell \geq \frac{n_1}{2k} \quad \text{for all } \ell \in S^*, \qquad w_{\ell'} \leq \frac{n_1}{8k} \quad \text{for all } \ell' \notin S^*.$$

Given v, $\mathbb{E}[C_\ell \mid v] = n_1\gamma + (1 - 2\gamma)w_\ell$, so $\hat{W}_\ell - w_\ell = (C_\ell - \mathbb{E}[C_\ell \mid v])/(1 - 2\gamma)$. The count C_ℓ is a sum of n_1 independent bits, each with variance at most γ. By Bernstein's inequality and a union bound over $\ell \in [d]$, with probability at least $1 - \mathcal{O}(n^{-2})$,

$$\big|C_\ell - \mathbb{E}[C_\ell \mid v]\big| \leq \sqrt{6n_1\gamma L} + 2L \qquad \text{for all } \ell \in [d].$$

We have $\gamma \leq 1/4$: this is $n_1 \geq 4C_\gamma \log(1/\delta)/\epsilon^2$, which follows from equation 11 when $\delta \geq (dn)^{-Ck^2\epsilon^2}$, as in Appendix B.2. So $1/(1 - 2\gamma) \leq 2$, and

$$\max_{\ell \in [d]} \big|\hat{W}_\ell - w_\ell\big| \leq 2\sqrt{6n_1\gamma L} + 4L \leq \frac{n_1}{32k} + \frac{n_1}{32k} = \frac{n_1}{16k}.$$

For the first term we used $n_1\gamma = C_\gamma \log(1/\delta)/\epsilon^2$ and $n_1 \geq \Omega(k\sqrt{L \log(1/\delta)}/\epsilon)$; for the second, $n_1 \geq \Omega(kL)$; both are in equation 11. Then for $\ell \in S^*$, $\hat{W}_\ell \geq n_1/(2k) - n_1/(16k) = 7n_1/(16k) > \tau_{HH}$, and for $\ell' \notin S^*$, $\hat{W}_{\ell'} \leq n_1/(8k) + n_1/(16k) = 3n_1/(16k) < \tau_{HH}$, where $\tau_{HH} = n_1/(4k)$. Therefore $\hat{S} = S^*$ and $s = |\hat{S}| = k$. Combining the events, the statement holds with probability at least $1 - \mathcal{O}(n^{-2})$. □

E.2 Proof of Theorem 3

Proof of Theorem 3. The parameters are set as in Table 4 and Table 5. We use $n_3, n_4 \geq 19$, which the summation step requires, and $\epsilon \leq \log(4s/\delta)$. The constants κ of Assumption 4 and ρ of Proposition 1 are constants of the model and are absorbed into the constants of the proof. Throughout, $\Pi_{B(0,R)}$ does not move a point of $B(0,R)$, and we use the following fact: if $v \in \mathbb{R}^s$ has independent, mean-zero coordinates with sub-Gaussian norm at most ν, then

$$\|v\|_2 \leq \mathcal{O}\big(\nu\sqrt{s+\log n}\big) \qquad \text{with probability at least } 1-n^{-3}, \tag{32}$$

which follows from Bernstein's inequality applied to $\sum_j v_j^2$. We write $S = \hat{S}$ and $s = |\hat{S}|$, and set $\tilde{\theta}_i = 0$ when $X_{i,S}^\top X_{i,S}$ is singular, so that the local step is defined on every dataset.

We first prove privacy, which holds for every value of $\hat{S}$ and every dataset. The three groups of users, of sizes n_1, n_3 and n_4, are disjoint, and a neighboring pair of user-level datasets changes the local data of one user, hence the messages of one group only. The randomizer of Stage 2a depends on the earlier rounds through $\hat{S}$, and the randomizer of Stage 2b through $(\hat{S}, \tau, c)$, which are computed from the messages of the other groups. So it is enough to show that each group, conditionally on the released quantities of the earlier groups, satisfies (ϵ, δ)-shuffled DP; the view of the data collector is then (ϵ, δ)-DP by parallel composition. Stage 1 was treated in the proof of Proposition 1. We give the argument for Stage 2b; Stage 2a is the same with $(n_3, p_a, R_0, \Delta_a, M_a, q_a)$ in place of $(n_4, p_b, \tau, \Delta_b, M_b, q_b)$, and with $\bar{\theta}_i$ in place of Δ_i.

Condition on $(\hat{S}, \tau, c)$ and fix two datasets that differ in the local data of one user i_0 of Stage 2b. As in the proof of Theorem 2, write the rounding step as $\bar{U}_{i,j} = \lfloor p_b \Delta_{i,j} + u_{i,j} \rfloor$ with $u_{i,j}$ independent and uniform on $(0,1]$, and fix u. Since $\|\Delta_{i_0}\|_2 \leq \tau$ and $\|\Delta'_{i_0}\|_2 \leq \tau$,

$$\|\bar{U}_{i_0} - \bar{U}'_{i_0}\|_2 \leq p_b \|\Delta_{i_0} - \Delta'_{i_0}\|_2 + \sqrt{s} \leq 2p_b\tau + \sqrt{s} + 2 = \Delta_b.$$

The setting of p_b gives $p_b\tau \geq (\sqrt{s}+2)\max\{1,\epsilon\}$, so

$$\sqrt{s}\max\{1,\epsilon\} \leq \Delta_b \leq 3p_b\tau. \tag{33}$$

Let $T_j = \sum_i (\bar{U}_{i,j} + \zeta_{i,j})$ over the n_4 users of Stage 2b, so that $\sum_i \zeta_{i,j} = B_j - n_4 N_0^{(b)}/2$ with $B_j \sim \mathrm{Bin}(n_4 N_0^{(b)}, 1/2)$ independent across j, and $n_4 N_0^{(b)} \geq M_b \geq 32\Delta_b^2 \log(4s/\delta)/\epsilon^2$. By equation 33 and $\epsilon \leq \log(4s/\delta)$, Lemma 1 applies with this Δ_b and with $\delta/2$ in place of δ, so $T = (T_1, \ldots, T_s)$ is $(\epsilon, \delta/2)$-DP for fixed u, and therefore also without conditioning on u.

Since $|\bar{U}_{i,j}| \leq p_b\tau + 1$, $|\zeta_{i,j}| \leq N_0^{(b)}/2$ and $N_0^{(b)} \geq 2$, we have $|T_j| \leq n_4(p_b\tau + 1 + N_0^{(b)}/2) < q_b/2$, so no wrap-around occurs and T_j is the value decoded by the aggregator. On coordinate j the users run the summation protocol $\widetilde{V}_{r_b,n_4}$ of Balle et al. (2020) over $\mathbb{Z}_{q_b}$; the setting of r_b gives $r_b \geq 3$, and $n_4 \geq 19$. By Balle et al. (2020, Corollary 6.1) the protocol has worst-case statistical security with parameter $\sigma_{\mathrm{sec}} = \log_2(8se^\epsilon/\delta)$. Let $\mathcal{V}^{\mathrm{sim}}$ be the view obtained when, on every coordinate, the inputs are replaced by $(T_j \bmod q_b, 0, \ldots, 0)$; it is a post-processing of T, hence $(\epsilon, \delta/2)$-DP. Coupling the s coordinates one at a time gives $d_{TV}(\mathcal{V}_{\mathcal{P}}, \mathcal{V}^{\mathrm{sim}}) \leq e^{-\epsilon}\delta/8$ for every dataset, so by Balle et al. (2020, Lemma F.2) the view of Stage 2b is $(\epsilon, \delta/2 + (1+e^\epsilon)e^{-\epsilon}\delta/8)$-DP, hence (ϵ, δ)-DP. Therefore Algorithm 1 satisfies user-level (ϵ, δ)-shuffled DP.

We now bound the error. By Proposition 1, with probability at least $1 - \mathcal{O}(n^{-2})$ we have $\hat{S} = S^*$ and $s = k$, and we work on this event. In particular $\theta^*_{\hat{S}} = \theta^*_{S^*}$ carries all of θ^*, so $\|\hat{\theta} - \theta^*\|_2 = \|\hat{\theta}_{\hat{S}} - \theta^*_{\hat{S}}\|_2$.

Consider the local least squares estimates. The rows of X_{i,S^*} are sub-Gaussian with covariance Σ_{S^*}, and $\lambda_{\min}(\Sigma_{S^*}) \geq \kappa$ by Assumption 4. By the concentration of the smallest singular value of a sub-Gaussian design, for a fixed user $\lambda_{\min}(X_{i,S^*}^\top X_{i,S^*}) \geq \kappa m/2$ with probability at least $1 - 2e^{-cm}$ once $m \geq \Omega(k)$, so by the third term of equation 10 and a union bound this holds for all users of Stage 2 with probability at least $1 - \mathcal{O}(n^{-2})$. On this event, for every user of Stage 2,

$$\tilde{\theta}_i - \theta^*_{S^*} = (X_{i,S^*}^\top X_{i,S^*})^{-1} X_{i,S^*}^\top \zeta_i,$$

which, conditionally on X_i, has mean zero and is sub-Gaussian with covariance proxy $\sigma^2(X_{i,S^*}^\top X_{i,S^*})^{-1} \preceq \frac{2\sigma^2}{\kappa m} I_k$ by Assumption 1. By equation 32 and a union bound over the users of Stage 2, with probability at least $1 - \mathcal{O}(n^{-2})$,

$$\big\|\tilde\theta_i - \theta^*_{S^*}\big\|_2 \le \mathcal{O}\Big(\sigma\sqrt{\frac{k+\log n}{m}}\Big) \le \tau_\beta \qquad \text{for every user } i \text{ of Stage 2,} \tag{34}$$

where we used $k + \log n \le 2k \log n$ and took C_β large enough. We work on this event as well.

We check that the two clipping steps do not change the local statistics. For Stage 2a, $\|\tilde\theta_i\|_2 \le \|\theta^*\|_2 + \tau_\beta = R_0$ by equation 34, so $\bar\theta_i = \tilde\theta_i$. The aggregator of Stage 2a decodes

$$c = \frac{1}{n_3}\sum_i \tilde\theta_i + E_a^{\mathrm{rd}} + E_a^{\mathrm{bin}},$$

where the sum is over the n_3 users of Stage 2a, the coordinates of E_a^{rd} are averages of n_3 independent terms in $(-1, 1]$ divided by p_a, and the coordinates of E_a^{bin} are centered $\mathrm{Bin}(n_3 N_0^{(a)}, 1/2)$ variables divided by $n_3 p_a$. Applying equation 32 to each of the three terms, together with $\sqrt{n_3 N_0^{(a)}} \le \mathcal{O}(\sqrt{M_a} + \sqrt{n_3})$, $\sqrt{M_a} \le \mathcal{O}(p_a R_0 \sqrt{\log(4s/\delta)}/\epsilon)$ from equation 33 and $\frac{1}{\sqrt{n_3}\, p_a} \le \frac{R_0\sqrt{\log(1/\delta)}}{c_p n_3 \sqrt{\log n}\,\epsilon}$ from the setting of p_a, we get, with probability at least $1 - \mathcal{O}(n^{-2})$,

$$\big\|c - \theta^*_{S^*}\big\|_2 \le \mathcal{O}\Big(\sigma\sqrt{\frac{k+\log n}{n_3 m}} + \frac{R_0\sqrt{(k+\log n)\log(4k/\delta)}}{n_3\epsilon}\Big) \le \hat A,$$

using $k + \log n \le 2k \log n$ and taking c_A, c'_A large enough. The first term of $\hat A$ is at most $\tau_\beta/2$ once $C_\beta \ge 2c_A$. Since $n_3 = \Theta(n)$, $R_0 = \Theta(1)$ and $\log(4k/\delta) = \Theta(\log(k/\delta))$, the second term of $\hat A$ is at most $\tau_\beta/2$ under equation 13. So $\hat A \le \tau_\beta$, hence

$$\tau = \tau_\beta + \hat A \le 2\tau_\beta. \tag{35}$$

For Stage 2b, $\|\tilde\theta_i - c\|_2 \le \|\tilde\theta_i - \theta^*_{S^*}\|_2 + \|\theta^*_{S^*} - c\|_2 \le \tau_\beta + \hat A = \tau$, so $\Delta_i = \tilde\theta_i - c$ and the clipping does not change it either.

The output of Stage 2b is therefore

$$\hat\theta_{\hat S} - \theta^*_{S^*} = \frac{1}{n_4}\sum_i \big(\tilde\theta_i - \theta^*_{S^*}\big) + E_b^{\mathrm{rd}} + E_b^{\mathrm{bin}},$$

where the sum is over the n_4 users of Stage 2b. Conditionally on the designs, the summands are independent, have mean zero and are sub-Gaussian with proxy $2\sigma^2/(\kappa m)$, so their average has proxy $2\sigma^2/(\kappa n_4 m)$, and equation 32 gives

$$\Big\|\frac{1}{n_4}\sum_i \big(\tilde\theta_i - \theta^*_{S^*}\big)\Big\|_2 \le \mathcal{O}\Big(\sigma\sqrt{\frac{k+\log n}{n_4 m}}\Big)$$

with probability at least $1 - \mathcal{O}(n^{-2})$. The two error terms are bounded as in Stage 2a, with τ in place of R_0 and n_4 in place of n_3:

$$\big\|E_b^{\mathrm{bin}}\big\|_2 \le \mathcal{O}\Big(\frac{\tau\sqrt{(k+\log n)\log(4k/\delta)}}{n_4\epsilon}\Big), \qquad \big\|E_b^{\mathrm{rd}}\big\|_2 \le \mathcal{O}\Big(\frac{\sqrt{k+\log n}}{\sqrt{n_4}\, p_b}\Big) \le \frac{1}{c_p}\big\|E_b^{\mathrm{bin}}\big\|_2,$$

where the last step uses the setting of p_b and $\log(1/\delta) \le \log(4k/\delta)$. By equation 35 and $\tau_\beta = C_\beta\sigma\sqrt{k\log n/m}$,

$$\big\|E_b^{\mathrm{bin}}\big\|_2 \le \mathcal{O}\Big(\frac{\sigma\sqrt{k\log n}\,\sqrt{(k+\log n)\log(4k/\delta)}}{n_4\sqrt{m}\,\epsilon}\Big) \le \mathcal{O}\Big(\frac{\sigma k\log n\sqrt{\log(4k/\delta)}}{n_4\sqrt{m}\,\epsilon}\Big),$$

using $\sqrt{k}\sqrt{k+\log n} \le k + \sqrt{k\log n} \le 2k\sqrt{\log n}$. Since $n_1 \le n/2$, we have $n_3, n_4 = \Theta(n)$. Collecting the three terms,

$$\big\|\hat\theta - \theta^*\big\|_2 \le \mathcal{O}\Big(\sigma\sqrt{\frac{k\log n}{nm}} + \frac{\sigma k\log n\sqrt{\log(4k/\delta)}}{n\sqrt{m}\,\epsilon}\Big) \le \widetilde{\mathcal{O}}\Big(\sqrt{\frac{k}{nm}} + \frac{k}{n\sqrt{m}\,\epsilon}\Big),$$

since $\log(4k/\delta) \le \mathcal{O}(\log n + \log(1/\delta))$ for $k \le n$. All the events above hold together with probability at least $1 - \mathcal{O}(n^{-2})$. □

E.3 Proof of Proposition 2

Proof of Proposition 2. Throughout, the expectation is over the data and the randomness of $\mathcal{A}$, which is independent of the data. For one sample $z = (x, y)$, the density is $f_\theta(y \mid x) f_x(x)$, and f_x does not depend on θ, so

$$S_\theta(z) = \nabla_\theta \log f_\theta(y \mid x) = \frac{\big(y - \mu(x^\top \theta)\big) x}{c(\sigma)}, \qquad \mu = \psi'.$$

We use two standard properties of the score. First,

$$\mathbb{E}[S_\theta(z)] = \int \frac{\nabla_\theta f_\theta(z)}{f_\theta(z)} f_\theta(z)\, dz = \nabla_\theta \int f_\theta(z)\, dz = 0. \tag{36}$$

Second, $\mathrm{Var}(y \mid x) = c(\sigma)\psi''(x^\top \theta)$ for the exponential family, so

$$\mathrm{Cov}\big(S_\theta(z)\big) = \frac{\mathbb{E}\big[\mathrm{Var}(y \mid x)\, x x^\top\big]}{c(\sigma)^2} = \frac{1}{c(\sigma)}\, \mathbb{E}\big[\psi''(x^\top \theta)\, x x^\top\big].$$

Write $S_\theta(U_i) = \sum_{j=1}^m S_\theta(z_{i,j})$ for the score of the local dataset of user i. Since the m samples of U_i are i.i.d., $\mathbb{E}[S_\theta(U_i)] = 0$ and

$$\mathrm{Cov}\big(S_\theta(U_i)\big) = \frac{m}{c(\sigma)}\, \mathbb{E}\big[\psi''(x^\top \theta)\, x x^\top\big]. \tag{37}$$

We first prove soundness. The dataset $Z'_{(i)}$ does not contain U_i, so $\mathcal{A}(Z'_{(i)})$ is independent of U_i, and by equation 36,

$$\mathbb{E}[A'_i] = \big\langle \mathbb{E}[\mathcal{A}(Z'_{(i)})] - \theta,\, \mathbb{E}[S_\theta(U_i)] \big\rangle = 0.$$

For the first absolute moment, we condition on $\mathcal{A}(Z'_{(i)})$ and use equation 37:

$$\begin{aligned}
\mathbb{E}|A'_i| \le \sqrt{\mathbb{E}[(A'_i)^2]} &= \sqrt{\mathbb{E}\Big[\big(\mathcal{A}(Z'_{(i)}) - \theta\big)^\top \mathrm{Cov}\big(S_\theta(U_i)\big)\big(\mathcal{A}(Z'_{(i)}) - \theta\big)\Big]} \\
&\le \sqrt{\mathbb{E}\|\mathcal{A}(Z'_{(i)}) - \theta\|_2^2} \cdot \sqrt{\frac{m}{c(\sigma)} \lambda_{\max}\big(\mathbb{E}[\psi''(x^\top\theta) x x^\top]\big)}.
\end{aligned}$$

Since U_i and $\widetilde{U}_i$ are i.i.d., $Z'_{(i)}$ and Z have the same distribution, so $\mathbb{E}\|\mathcal{A}(Z'_{(i)}) - \theta\|_2^2 = \mathbb{E}\|\mathcal{A}(Z) - \theta\|_2^2$, which gives the bound in (i).

We next prove completeness. The samples of all users are i.i.d., so $\sum_{i=1}^n S_\theta(U_i) = \nabla_\theta \log f_\theta(Z)$, where $f_\theta(Z)$ is the joint density of the nm samples. By equation 36 applied to Z, the term θ does not contribute:

$$\sum_{i=1}^n \mathbb{E}[A_i] = \mathbb{E}\big\langle \mathcal{A}(Z) - \theta,\, \nabla_\theta \log f_\theta(Z) \big\rangle = \mathbb{E}\big\langle \mathcal{A}(Z),\, \nabla_\theta \log f_\theta(Z) \big\rangle.$$

Writing the inner product over coordinates and exchanging differentiation and integration under the regularity conditions,

$$\begin{aligned}
\sum_{j=1}^d \mathbb{E}\Big[\mathcal{A}(Z)_j\, \frac{\partial}{\partial \theta_j} \log f_\theta(Z)\Big] &= \sum_{j=1}^d \int \mathbb{E}\big[\mathcal{A}(z)_j\big]\, \frac{\partial}{\partial \theta_j} f_\theta(z)\, dz \\
&= \sum_{j=1}^d \frac{\partial}{\partial \theta_j} \mathbb{E}\big[\mathcal{A}(Z)_j\big],
\end{aligned}$$

where the inner expectation is over the randomness of $\mathcal{A}$. This proves (ii). □

The proof of Theorem 4 uses a sparse version of the user-level score attack, in which the score is kept only on the support of θ. For $T = \mathrm{supp}(\theta)$, define

$$A_i^T = \big\langle \big(\mathcal{A}(Z) - \theta\big)_T,\, S_\theta(U_i)_T \big\rangle, \qquad A_i^{T\prime} = \big\langle \big(\mathcal{A}(Z'_{(i)}) - \theta\big)_T,\, S_\theta(U_i)_T \big\rangle. \tag{38}$$

The proof above applies to A_i^T with every vector restricted to the coordinates in T. For soundness, $\mathbb{E}[A_i^{T\prime}] = 0$ and

$$\mathbb{E}|A_i^{T\prime}| \le \sqrt{\mathbb{E}\|(\mathcal{A}(Z)-\theta)_T\|_2^2} \cdot \sqrt{\frac{m}{c(\sigma)}\lambda_{\max}\big(\mathbb{E}[\psi''(x^\top\theta)x_T x_T^\top]\big)}.$$

For completeness, viewing f_θ as a function of θ_T with $\theta_{T^c} = 0$ held fixed,

$$\sum_{i=1}^n \mathbb{E}[A_i^T] = \sum_{j\in T} \frac{\partial}{\partial\theta_j}\mathbb{E}\big[\mathcal{A}(Z)_j\big].$$

E.4 Proof of Theorem 4

Throughout this subsection we work on the parameter space

$$\Theta = \big\{\theta \in \mathbb{R}^d : \|\theta\|_0 \le k,\ \|\theta\|_\infty \le 1\big\} \subseteq \Theta_k,$$

we write $T = \mathrm{supp}(\theta)$, and A_i^T, $A_i^{T\prime}$ are the sparse user-level score attack of equation 38. The first lemma turns the differential privacy of the estimator into an upper bound on the score attack of one user; it is the user-level form of Proposition 2.1 of Cai et al. (2023).

Lemma 2. *Let $\mathcal{A}$ satisfy user-level (ϵ,δ)-DP with $0 < \epsilon < 1$, and let $\theta \in \Theta$. Then for every $t_0 > 0$ and every $i \in [n]$,*

$$\mathbb{E}[A_i^T] \le 2\epsilon\,\mathbb{E}\big|A_i^{T\prime}\big| + 2\delta t_0 + \int_{t_0}^\infty \mathbb{P}\big[|A_i^T| > t\big]\,dt.$$

Proof of Lemma 2. Write $A = A_i^T$ and $A' = A_i^{T\prime}$, and let $Z^+ = \max(Z,0)$ and $Z^- = -\min(Z,0)$. The datasets Z and $Z'_{(i)}$ are user-level neighbors, and A and A' are the same function of $(\mathcal{A}(Z), U_i)$ and of $(\mathcal{A}(Z'_{(i)}), U_i)$, so conditionally on $(U_i, \widetilde{U}_i)$ the privacy of $\mathcal{A}$ gives $\mathbb{P}[A^+ > t] \le e^\epsilon \mathbb{P}[A'^+ > t] + \delta$ for every t, and the same with the two sides exchanged. Using $\mathbb{E}[A] = \int_0^\infty \mathbb{P}[A^+ > t]\,dt - \int_0^\infty \mathbb{P}[A^- > t]\,dt$ and $e^\epsilon \le 1 + 2\epsilon$ for $0 < \epsilon < 1$,

$$\begin{aligned}\int_0^\infty \mathbb{P}[A^+ > t]\,dt &\le \int_0^{t_0} \big(e^\epsilon \mathbb{P}[A'^+ > t] + \delta\big)\,dt + \int_{t_0}^\infty \mathbb{P}[A^+ > t]\,dt \\ &\le \int_0^\infty \mathbb{P}[A'^+ > t]\,dt + 2\epsilon\int_0^\infty \mathbb{P}[A'^+ > t]\,dt + \delta t_0 + \int_{t_0}^\infty \mathbb{P}[|A| > t]\,dt.\end{aligned}$$

In the same way, from $\mathbb{P}[A^- > t] \ge e^{-\epsilon}(\mathbb{P}[A'^- > t] - \delta)$ and $e^{-\epsilon} \ge 1 - 2\epsilon$,

$$\int_0^\infty \mathbb{P}[A^- > t]\,dt \ge \int_0^\infty \mathbb{P}[A'^- > t]\,dt - 2\epsilon\int_0^\infty \mathbb{P}[A'^- > t]\,dt - \delta t_0.$$

Subtracting the second display from the first gives $\mathbb{E}[A] \le \mathbb{E}[A'] + 2\epsilon\,\mathbb{E}|A'| + 2\delta t_0 + \int_{t_0}^\infty \mathbb{P}[|A| > t]\,dt$, and $\mathbb{E}[A'] = 0$ by the soundness part of Proposition 2. □

The next lemma sums Lemma 2 over the n users. The tail term is where the user-level regime differs from the sample-level one: the score of a user is a sum of m terms, so its size is governed by the empirical second moments of that user's m design vectors.

Lemma 3. *Let Assumptions 2 and 4 hold, let $\|\psi''\|_\infty \le c_2$, and let $\theta \in \Theta$. Let $\mathcal{A}$ satisfy user-level (ϵ,δ)-DP with $0 < \epsilon < 1$ and take values in Θ, and let $L = \log(2k/\delta) \le m$. Then*

$$\sum_{i=1}^n \mathbb{E}\big[A_i^T\big] \le 2n\epsilon\sqrt{\mathbb{E}\|\mathcal{A}(Z)-\theta\|_2^2}\sqrt{\frac{mCc_2}{c(\sigma)}} + C_1\, n\delta k\kappa_x\sqrt{\frac{c_2}{c(\sigma)}}\sqrt{mL},$$

where C_1 is an absolute constant.

Proof of Lemma 3. By the soundness part of Proposition 2 applied to $A_i^{T\prime}$, together with $\|\psi''\|_\infty \le c_2$ and $\lambda_{\max}(\Sigma_{T,T}) \le C$ from Assumption 4,

$$\mathbb{E}\big|A_i^{T\prime}\big| \le \sqrt{\mathbb{E}\|\mathcal{A}(Z)-\theta\|_2^2}\sqrt{\frac{mCc_2}{c(\sigma)}}. \tag{39}$$

It remains to choose t_0 in Lemma 2 and to bound the tail. Write $w_{i,j} = y_{i,j} - \psi'(x_{i,j}^\top\theta)$ and $U_{i,\ell} = \sum_{j=1}^m w_{i,j}x_{i,j,\ell}$ for $\ell \in T$, so that

$$A_i^T = \frac{1}{c(\sigma)}\big\langle \big(\mathcal{A}(Z)-\theta\big)_T,\, U_{i,T}\big\rangle, \qquad \big|A_i^T\big| \le \frac{\big\|(\mathcal{A}(Z)-\theta)_T\big\|_1}{c(\sigma)}\max_{\ell\in T}\big|U_{i,\ell}\big| \le \frac{2k}{c(\sigma)}\max_{\ell\in T}\big|U_{i,\ell}\big|,$$

where the last step uses $|T| \le k$ and $\|\mathcal{A}(Z)\|_\infty, \|\theta\|_\infty \le 1$. For the exponential family, $\mathbb{E}[e^{\lambda w/c(\sigma)} \mid x] \le e^{c_2\lambda^2/(2c(\sigma))}$, so conditionally on x the variable w has mean zero and is sub-Gaussian with variance proxy $c(\sigma)c_2$. By Assumption 2 and $\Sigma_{\ell\ell} \le 1$ we have $\|x_\ell\|_{\psi_2} \le \kappa_x$, so each product $w_{i,j}x_{i,j,\ell}$ has mean zero and sub-exponential norm at most $K = \mathcal{O}(\kappa_x\sqrt{c(\sigma)c_2})$. Bernstein's inequality for the sum of the m independent products, followed by a union bound over $\ell \in T$, gives

$$\mathbb{P}\Big[\max_{\ell\in T}|U_{i,\ell}| > u\Big] \le 2k\exp\Big(-c\min\Big\{\frac{u^2}{mK^2},\, \frac{u}{K}\Big\}\Big).$$

Set $a = 2kK/c(\sigma)$ and $t_0 = a(\sqrt{mL}+L)/\sqrt{c}$, so that $\mathbb{P}[|A_i^T| > t_0] \le 2ke^{-L} = \delta$ and, integrating the same bound,

$$\int_{t_0}^\infty \mathbb{P}\big[|A_i^T| > t\big]\,dt \le \mathcal{O}\big(a\sqrt{m/L}+a\big)\delta \le \mathcal{O}\big(a\sqrt{mL}\big)\delta.$$

Since $L \le m$, we have $\sqrt{mL}+L \le 2\sqrt{mL}$, so $2\delta t_0 + \int_{t_0}^\infty \mathbb{P}[|A_i^T| > t]dt \le \mathcal{O}(a\delta\sqrt{mL})$. Substituting $a = 2kK/c(\sigma)$ and $K = \mathcal{O}(\kappa_x\sqrt{c(\sigma)c_2})$ bounds this by $\mathcal{O}\big(\delta k\kappa_x\sqrt{c_2/c(\sigma)}\sqrt{mL}\big)$. Combining with equation 39 in Lemma 2 and summing over the n users gives the claim. □

The last lemma bounds the same sum from below, on average over a prior supported on sparse vectors. The prior is the one of Cai et al. (2023): let $\widetilde{\theta}_1, \dots, \widetilde{\theta}_d$ be i.i.d. from the $N(0,\gamma^2)$ distribution truncated to $[-1,1]$ with

$$\gamma^2 = \frac{1}{4\log(d/4k)} \asymp \frac{1}{\log(d/k)}, \tag{40}$$

let S be the set of indices of the k largest $|\widetilde{\theta}_j|$, and set $\theta_j = \widetilde{\theta}_j\mathbb{I}[j \in S]$. We write π for the law of θ, and p_{S_l} for its density conditioned on $S = S_l$, where $S_1, \dots, S_{\binom{d}{k}}$ enumerate the subsets of $[d]$ of size k.

Lemma 4. *Let $k \le \sqrt{d}$ and $k \ge C_0\log^2(d/k)$ for a large enough constant C_0. Let $\mathcal{A}$ take values in Θ and satisfy $\mathbb{E}_{Z|\theta}\|\mathcal{A}(Z)-\theta\|_2^2 \le R$ for every $\theta \in \Theta$, where*

$$R \le \frac{c_1\sqrt{k}}{\log(d/k)} \tag{41}$$

for a small enough constant c_1. Then

$$\sum_{i=1}^n \mathbb{E}_\pi\mathbb{E}_{Z|\theta}\big[A_i^T\big] \ge c\,k\log(d/k)$$

for an absolute constant $c > 0$.

Proof of Lemma 4. Let $g(\theta) = \mathbb{E}_{Z|\theta}[\mathcal{A}(Z)]$. By the completeness part of equation 38, for every θ with $\mathrm{supp}(\theta) = T$, $\sum_{i=1}^n \mathbb{E}_{Z|\theta}[A_i^T] = \sum_{j\in T}\partial g_j(\theta)/\partial\theta_j$. The prior π is not absolutely continuous on $\mathbb{R}^d$, so we condition on the support: $\pi = \binom{d}{k}^{-1}\sum_l p_{S_l}$, and

$$\sum_{i=1}^n \mathbb{E}_\pi\mathbb{E}_{Z|\theta}\big[A_i^T\big] = \frac{1}{\binom{d}{k}}\sum_l \mathbb{E}_{p_{S_l}}\sum_{j\in S_l}\frac{\partial g_j(\theta)}{\partial\theta_j}.$$

Within a fixed S_l the law of θ has a density, so Stein's lemma applies coordinatewise and gives

$$\mathbb{E}_{p_{S_l}} \sum_{j\in S_l} \frac{\partial g_j(\theta)}{\partial \theta_j} \geq \mathbb{E}_{p_{S_l}}\Big[\sum_{j\in S_l} -\theta_j \frac{p'_{S_l,j}(\theta_j)}{p_{S_l,j}(\theta_j)}\Big] - \mathbb{E}_{p_{S_l}}\Big[\sum_{j\in S_l} |g_j(\theta)-\theta_j|\Big|\frac{p'_{S_l,j}(\theta_j)}{p_{S_l,j}(\theta_j)}\Big|\Big] + B,$$

where B collects the boundary terms at ± 1. Each θ_j is an order statistic of d truncated normals, so its density vanishes at -1, and at $+1$ it is bounded by $d\phi(1) = \mathcal{O}(k^2\sqrt{\log(d/k)}/d)$, which is $\mathcal{O}(\sqrt{\log(d/k)})$ since $k \leq \sqrt{d}$. As $\mathcal{A}$ takes values in Θ, $|g_j(\theta)-\theta_j| \leq 2$, so $|B| = \mathcal{O}(k\sqrt{\log(d/k)}) \ll k\log(d/k)$.

The two remaining terms depend on the prior only. By Lemma C.1 of Cai et al. (2023), the first term is at least $c'k\log(d/k)$ for an absolute constant c', using equation 40 and $k \leq \sqrt{d}$. For the second term, the Cauchy–Schwarz inequality and the bound on the score of the order statistics in the proof of Lemma C.2 of Cai et al. (2023) give

$$\mathbb{E}_{p_{S_l}}\Big[\sum_{j\in S_l} |g_j(\theta)-\theta_j|\Big|\frac{p'_{S_l,j}(\theta_j)}{p_{S_l,j}(\theta_j)}\Big|\Big] \leq \sqrt{R}\sqrt{\mathbb{E}_{p_{S_l}}\sum_{j\in S_l}\Big(\frac{p'_{S_l,j}(\theta_j)}{p_{S_l,j}(\theta_j)}\Big)^2} \leq \sqrt{R}\,\mathcal{O}\big(k^{3/4}\log^{3/2}(d/k)\big),$$

where we used $\mathbb{E}_{Z|\theta}\|\mathcal{A}(Z)-\theta\|_2^2 \leq R$ and $\sum_{j\leq k}\sqrt{j}\log^3(d/k) = \mathcal{O}(k^{3/2}\log^3(d/k))$. By equation 41 this is $\mathcal{O}(\sqrt{c_1}\,k\log(d/k))$, which is at most $c'k\log(d/k)/4$ for c_1 small enough. The same bounds hold for every S_l, so averaging over l gives the claim with $c = c'/2$. □

Proof of Theorem 4. It is enough to bound $\inf_{\mathcal{A}}\sup_{\theta\in\Theta}\mathbb{E}\|\mathcal{A}(Z)-\theta\|_2^2$ from below, since $\Theta \subseteq \Theta_k$. We may also assume that $\mathcal{A}$ takes values in Θ: replacing $\mathcal{A}(Z)$ by a closest point of Θ is a post-processing, hence preserves privacy, and at most doubles the error at every $\theta \in \Theta$. Let $R = \sup_{\theta\in\Theta}\mathbb{E}\|\mathcal{A}(Z)-\theta\|_2^2$ and let $L = \log(2k/\delta)$.

We first check that the conditions of the theorem control the tail term of Lemma 3. The map $x \mapsto x\sqrt{\log(2k/x)}$ is increasing on $(0, 2k/e)$, so the condition $\delta = o\big(\log(d/k)/(n\sqrt{m\log(nm)})\big)$ together with $k \leq nm$ gives

$$n\delta\sqrt{mL} \leq c'_0\log(d/k)\sqrt{\frac{c(\sigma)}{c_2}}\frac{1}{\kappa_x} \tag{42}$$

for a constant c'_0 as small as we wish, and $m \geq \log(nm)$ gives $L \leq m$ once $\delta \geq 2k/(nm)^C$; for smaller δ the left-hand side of equation 42 only decreases, and $L \leq m$ can be assumed by replacing δ by the larger value. Hence the tail term of Lemma 3 is at most $\frac{c}{2}k\log(d/k)$ with c the constant of Lemma 4.

Suppose first that $R > c_1\sqrt{k}/\log(d/k)$. The condition $n\sqrt{m}\,\epsilon \geq \Omega\big(k^{3/4}\log^{3/2}(d/k)\big)$ gives

$$\frac{k^2\log^2(d/k)}{n^2m\epsilon^2} \leq \mathcal{O}\Big(\frac{\sqrt{k}}{\log(d/k)}\Big),$$

so R is already of the order claimed by the theorem and there is nothing to prove. We may therefore assume $R \leq c_1\sqrt{k}/\log(d/k)$, so that Lemma 4 applies. Combining it with Lemma 3, averaged over π and using $\mathbb{E}_\pi\sqrt{\mathbb{E}_{Z|\theta}\|\mathcal{A}(Z)-\theta\|_2^2} \leq \sqrt{R}$,

$$c\,k\log(d/k) \leq 2n\epsilon\sqrt{R}\sqrt{\frac{mCc_2}{c(\sigma)}} + \frac{c}{2}k\log(d/k),$$

and therefore

$$R \geq \Omega\Big(\frac{c(\sigma)}{Cc_2}\cdot\frac{k^2\log^2(d/k)}{n^2m\epsilon^2}\Big).$$

The privacy constraint is not the only one: any estimator, private or not, is bounded below by the non-private minimax risk over Θ, which is $\Omega\big(c(\sigma)k\log(d/k)/(nm)\big)$ for the $N = nm$ samples; for

the linear model $\psi(u) = u^2/2$ this is the rate of Raskutti et al. (2011), whose design conditions are supplied by Assumption 4. Adding the two bounds and absorbing C, c_2 and $c(\sigma)$ into the constants,

$$\inf_{\mathcal{A}\in\mathcal{M}_{\epsilon,\delta}} \sup_{\theta^*\in\Theta_k} \mathbb{E}\big\|\mathcal{A}(Z) - \theta^*\big\|_2^2 \geq \Omega\Big(\frac{k\log(d/k)}{nm} + \frac{k^2\log^2(d/k)}{n^2m\epsilon^2}\Big),$$

and taking square roots gives the statement of Theorem 4. □

F Additional Results

F.1 Non-interactive User-level Shuffled DP

Theorem 3 removes the dimension from the rate by using two rounds. We show here that the second ingredient, interaction, is also needed: under a non-interactive protocol, user-level data alone leaves a polynomial dependence on d in the privacy term, whatever the local sample size m is.

Theorem 5. *Let $\mathcal{M}^{\mathrm{SDP}}_{\epsilon,\delta}$ be the family of all non-interactive protocols that satisfy robust user-level (ϵ,δ)-shuffled DP with $0 < \epsilon \leq \mathcal{O}(1)$ and $\delta = o(1/n)$. Let $Z = \{U_i\}_{i=1}^n$ be a dataset in which each user holds m i.i.d. samples from the sparse linear model of Assumption 1 with parameter θ^*. If*

$$n^{3/2}\epsilon^2 \geq \Omega(dk), \tag{43}$$

then

$$\inf_{\mathcal{A}\in\mathcal{M}^{\mathrm{SDP}}_{\epsilon,\delta}} \sup_{\theta^*\in\Theta_k} \mathbb{E}\big[\|\mathcal{A}(Z) - \theta^*\|_2\big] \geq \Omega\left(\sqrt{\frac{k\log(d/k)}{nm}} + \frac{\sqrt{dk}}{n\sqrt{m}\,\epsilon}\right).$$

Comparing with Theorem 3, the privacy term of the non-interactive protocol carries $\sqrt{d}$, while the interactive one carries $\sqrt{k}$ in its place. The first term is the non-private minimax rate for $N = nm$ samples, which is $\Omega(\sqrt{k\log(d/k)/(nm)})$ (Raskutti et al., 2011); we only prove the privacy term below.

Proof of Theorem 5. We use the notations of Appendix D.1: $\mathcal{X} = \{\pm1\}^{d+1}$, U is uniform on $\mathcal{X}$, $\chi_i(x,y) = yx_i$, and $Q_\theta = U(1 + y\langle x,\theta\rangle)$. For $z \in \{-1,0,1\}^d$ and a scale $\beta > 0$ we write

$$D_z^{(\beta)} = Q_{\beta z}^{\otimes m},$$

the law of the m samples of one user at parameter βz, and $Z_0(z)$, $z^{i,b}$ are as in Appendix D.1. All densities below are taken with respect to $U^{\otimes m}$, and $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ are the inner product and the norm of $L_2(U^{\otimes m})$. Fix a scale γ with

$$mk\gamma^2 \leq 1, \tag{44}$$

and let $\gamma' = 2\gamma/9$. Let $\mathcal{P}$ be a protocol in $\mathcal{M}^{\mathrm{SDP}}_{\epsilon,\delta}$, which we take to be $(\epsilon,\delta,1/3)$-robust, and let $n' = n/3$.

By Cheu & Ullman (2021, Theorem 2.6) applied with the record universe $\mathcal{X}^m$ and the fixed distribution $U^{\otimes m} = D_z^{(0)}$, there is an (ϵ,δ)-pan-private algorithm M that processes n' records and satisfies, for every z,

$$d_{TV}\Big(M\big((D_z^{(\gamma)})^{\otimes n'}\big),\, \mathcal{M}_{\mathcal{P}}\big(Q_z^{\otimes n}\big)\Big) \leq e^{-c_1 n}, \qquad Q_z = \frac{2}{9}D_z^{(\gamma)} + \frac{7}{9}U^{\otimes m}. \tag{45}$$

The mixture Q_z is not a member of the model, and this is the one point where the user-level regime differs from Appendix D.1: the law of a record is $Q_{\beta z}^{\otimes m}$, which is not linear in the parameter. We therefore compare Q_z with the member of the model that agrees with it to first order, namely $D_z^{(\gamma')}$. Writing

$$e_j = \sum_{T\subseteq[m],\,|T|=j} \prod_{t\in T} \chi_z(w_t), \qquad \chi_z = \sum_{l=1}^{d} z_l\chi_l,$$

we have $D_z^{(\beta)} = \sum_{j=0}^m \beta^j e_j$ and $Q_z = 1 + \frac{2}{9}\sum_{j\geq1}\gamma^j e_j$, so the two first-order terms agree and

$$R_z := Q_z - D_z^{(\gamma')} = \sum_{j\geq2}\Big(\frac{2}{9} - \Big(\frac{2}{9}\Big)^j\Big)\gamma^j e_j.$$

The characters appearing in e_j for different j are distinct and $\|e_j\|^2 = \binom{m}{j}\|z\|_0^j \le (mk)^j/j!$, so with $u = \gamma\sqrt{mk} \le 1$ from equation 44,

$$\|R_z\| \le \sum_{j\ge 2} \frac{u^j}{\sqrt{j!}} \le u^2 \sum_{j\ge 0} \frac{1}{\sqrt{j!}} \le 3mk\gamma^2.$$

Since $Q_z \ge \frac{7}{9}$ pointwise, the squared Hellinger distance satisfies

$$d_H^2\big(D_z^{(\gamma')}, Q_z\big) \le \chi^2\big(D_z^{(\gamma')} \,\|\, Q_z\big) = \int \frac{R_z^2}{Q_z}\, dU^{\otimes m} \le \frac{9}{7}\|R_z\|^2 \le 12\big(mk\gamma^2\big)^2,$$

and by the subadditivity of d_H^2 over product measures and $d_{TV} \le \sqrt{2}\, d_H$,

$$d_{TV}\Big(Q_z^{\otimes n},\ \big(D_z^{(\gamma')}\big)^{\otimes n}\Big) \le 6\sqrt{n}\, mk\gamma^2. \tag{46}$$

Combining equation 45 and equation 46, and writing $\sigma_n = e^{-c_1 n} + 6\sqrt{n}\, mk\gamma^2$, we get for every z

$$d_{TV}\Big(M\big((D_z^{(\gamma)})^{\otimes n'}\big),\ \mathcal{M}_{\mathcal{P}}\big((D_z^{(\gamma')})^{\otimes n}\big)\Big) \le \sigma_n. \tag{47}$$

So it is enough to bound the total variation distances between the outputs of the pan-private algorithm M, and to pay σ_n per comparison.

We bound these distances as in Appendix D.1. Fix z with $\|z\|_0 \le k-1$ and let $i \in Z_0(z)$, $b \in \{\pm 1\}$. For $j \in [n']$, the state S_j of M is, for every fixed prefix, an (ϵ, δ)-DP function of the j-th record, so by Cheu & Ullman (2021, Lemma 3.7) there is a $(2\epsilon, 0)$-DP kernel q_j' within total variation distance δ of it. Running the hybrid in which the first j records are drawn from $D_z^{(\gamma)}$ and the rest from $D_{z^{i,b}}^{(\gamma)}$, the output is a post-processing of $(S_j, X_{>j})$, the prefix has the same law in the two hybrids, and the suffix is independent of the j-th record, so

$$d_{TV}\big(H_{j-1}, H_j\big) \le \frac{1}{2}\,\mathbb{E}_{X_{<j}}\mathbb{E}_{s\sim\bar{q}_j'}\big|\langle h_{j,s},\, D_i^b\rangle\big| + 2\delta, \qquad D_i^b = D_{z^{i,b}}^{(\gamma)} - D_z^{(\gamma)}, \tag{48}$$

where $\bar{q}_j'(s) = \mathbb{E}_{U^{\otimes m}}[q_j'(s \mid W)]$ and $h_{j,s} = q_j'(s \mid \cdot)/\bar{q}_j'(s) - 1$ satisfies $\|h_{j,s}\|_\infty \le e^{2\epsilon} - 1$. Neither the law of the prefix nor the kernel depends on i.

Two facts about D_i^b are used. First, expanding the two products, every character appearing in D_i^b is of the form $\prod_{t\in T}\chi_i(w_t)\prod_{t\in T'}\chi_{l_t}(w_t)$ with $T \ne \emptyset$, $T \cap T' = \emptyset$ and $l_t \in \mathrm{supp}(z)$. Since $i \notin \mathrm{supp}(z)$, the characters of D_i^b and $D_{i'}^b$ are distinct for $i \ne i'$ in $Z_0(z)$. Second, writing $k' = \|z\|_0$,

$$\big\|D_i^b\big\|^2 = \big(1+\gamma^2(k'+1)\big)^m - \big(1+\gamma^2 k'\big)^m \le e^{\gamma^2 m(k'+1)}\, m\gamma^2 \le e^2 m\gamma^2,$$

where the last step uses equation 44. By the first fact and the Cauchy–Schwarz inequality, for every h with $\|h\|_\infty \le e^{2\epsilon} - 1$,

$$\sum_{i\in Z_0(z)} \big|\langle h, D_i^b\rangle\big| \le \sqrt{d}\sqrt{\sum_{i\in Z_0(z)} \langle h, D_i^b\rangle^2} \le \sqrt{d}\,\|h\| \max_i \big\|D_i^b\big\| \le e\sqrt{dm}\,\gamma\,\big(e^{2\epsilon}-1\big).$$

Summing equation 48 over $j \in [n']$, over $b \in \{\pm 1\}$ and over $i \in Z_0(z)$, and then adding $4\sigma_n$ per coordinate from equation 47,

$$\sum_{i\in Z_0(z)} T_i(z) \le e\,\gamma\, n'\sqrt{dm}\,\big(e^{2\epsilon}-1\big) + 4dn'\delta + 4d\sigma_n, \tag{49}$$

where $T_i(z) = \sum_b d_{TV}\big(\mathcal{M}_{\mathcal{P}}((D_{z^{i,b}}^{(\gamma')})^{\otimes n}), \mathcal{M}_{\mathcal{P}}((D_z^{(\gamma')})^{\otimes n})\big)$ is stated for the protocol.

The rest is Assouad's method, exactly as in Appendix D.1. Draw Z as in Definition 6, let $\theta_Z = \gamma' Z$, draw the n local datasets from $D_Z^{(\gamma')}$ given Z, and let $\hat{\theta} = \mathcal{M}_{\mathcal{P}}(Z)$. Let $G_i = \{\|Z_{-i}\|_0 \le k-1\}$, let $\psi_i = \mathrm{sign}(\hat{\theta}_i)$ and let $N = \sum_i \mathbb{I}[Z_i \ne 0,\, G_i,\, \psi_i \ne Z_i]$. Then $N \le k$ whenever $N \ge 1$, and $\|\hat{\theta} - \theta_Z\|_2 \ge \gamma'\sqrt{N} \ge \gamma' N/\sqrt{k}$. Le Cam's method, $\mathbb{P}(Z_i \ne 0) = k/(2d)$, $\sum_i \mathbb{P}(G_i) \ge d/2$ by Markov's inequality, and equation 49 together give

$$\mathbb{E}[N] \ge \frac{k}{4d}\Big(\frac{d}{2} - 2e\gamma n'\sqrt{dm}\big(e^{2\epsilon}-1\big) - 8dn'\delta - 8d\sigma_n\Big).$$

We now fix γ. Take $\gamma = \sqrt{d}/\big(C_2 n' \sqrt{m}(e^{2\epsilon}-1)\big)$ with C_2 a large enough constant, so that the second term is at most $d/16$; the third term is at most $d/16$ because $\delta = o(1/n)$; and the fourth is at most $d/16$ once $\sigma_n \le 1/128$, which holds for n larger than an absolute constant together with

$$\sqrt{n}\, mk\gamma^2 = \frac{dk}{C_2^2\, n'^2\, \sqrt{n}\, (e^{2\epsilon}-1)^2} \le \frac{1}{768},$$

which is equation 43 for $\epsilon \le \mathcal{O}(1)$. For this choice of γ, the small signal condition equation 44 reads $n\epsilon \ge \Omega(\sqrt{dk})$, which equation 43 implies since $n^2\epsilon^2 \ge n^{3/2}\epsilon^2$. Hence $\mathbb{E}[N] \ge k/32$, and

$$\sup_{\theta^* \in \Theta_k} \mathbb{E}\big[\|\hat\theta - \theta^*\|_2\big] \ge \mathbb{E}\big[\|\hat\theta - \theta_Z\|_2\big] \ge \frac{\gamma'\sqrt{k}}{32} = \Omega\left(\frac{\sqrt{dk}}{n\sqrt{m}\,\epsilon}\right),$$

using $\gamma' = 2\gamma/9$, $n' = n/3$ and $e^{2\epsilon}-1 \le \mathcal{O}(\epsilon)$. □

F.2 User-level DP in Dense Regression

The user-level score attack is not tied to sparsity. We record here what it gives for the dense problem, where the parameter is only assumed to lie in the box $\Theta_{\mathrm{box}} = \{\theta \in \mathbb{R}^d : \|\theta\|_\infty \le 1\}$. The rate is the one of Theorem 4 with $k\log(d/k)$ replaced by d.

Theorem 6. *Let $Z = \{U_i\}_{i=1}^n$ be a dataset in which each user holds m i.i.d. samples from the exponential family of Theorem 4. Let Assumption 2 hold with $\lambda_{\max}(\Sigma) \le C$, and let $\|\psi''\|_\infty \le c_2$. Let $\mathcal{M}_{\epsilon,\delta}$ be the family of all estimators satisfying user-level (ϵ,δ)-DP with $0 < \epsilon < 1$ and $\delta = o\big(1/(n\sqrt{m\log(nm)})\big)$. If $m \ge \log(nm)$ and $n\sqrt{m}\,\epsilon \ge \Omega(\sqrt{d})$, then*

$$\inf_{\mathcal{A}\in\mathcal{M}_{\epsilon,\delta}} \sup_{\theta^*\in\Theta_{\mathrm{box}}} \mathbb{E}\big[\|\mathcal{A}(Z) - \theta^*\|_2\big] \ge \Omega\left(\sqrt{\frac{d}{nm}} + \frac{d}{n\sqrt{m}\,\epsilon}\right).$$

Throughout this subsection A_i is the user-level score attack of Definition 7, which is equation 38 with $T = [d]$. The first lemma bounds $\sum_i \mathbb{E}[A_i]$ from below through Stein's lemma, under a prior that is supported on Θ_{box} and vanishes at its boundary.

Lemma 5. *Let π be the product prior on Θ_{box} with marginal densities $\pi_j(t) = \frac{15}{16}(1-t^2)^2$ on $[-1,1]$. Let $\mathcal{A}$ take values in Θ_{box} and satisfy $\mathbb{E}_{Z|\theta}\|\mathcal{A}(Z) - \theta\|_2^2 \le R$ for every $\theta \in \Theta_{\mathrm{box}}$, with $R \le d/40$. Then*

$$\sum_{i=1}^n \mathbb{E}_\pi \mathbb{E}_{Z|\theta}\big[A_i\big] \ge \frac{d}{2}.$$

Proof of Lemma 5. Let $g(\theta) = \mathbb{E}_{Z|\theta}[\mathcal{A}(Z)]$. By the completeness part of Proposition 2, $\sum_{i=1}^n \mathbb{E}_{Z|\theta}[A_i] = \sum_{j=1}^d \partial g_j(\theta)/\partial\theta_j$. Each π_j is continuously differentiable on $(-1,1)$ and vanishes at ± 1, so Stein's lemma applies coordinatewise with no boundary term, and, as in Proposition 2.2 of Cai et al. (2023),

$$\mathbb{E}_\pi \sum_{j=1}^d \frac{\partial g_j(\theta)}{\partial\theta_j} \ge \mathbb{E}_\pi\Big[\sum_{j=1}^d -\theta_j \frac{\pi_j'(\theta_j)}{\pi_j(\theta_j)}\Big] - \sqrt{R\,\mathbb{E}_\pi\Big[\sum_{j=1}^d \Big(\frac{\pi_j'(\theta_j)}{\pi_j(\theta_j)}\Big)^2\Big]}.$$

For this prior, $\pi_j'(t)/\pi_j(t) = -4t/(1-t^2)$, so

$$\mathbb{E}_{\pi_j}\Big[-t\frac{\pi_j'(t)}{\pi_j(t)}\Big] = \frac{15}{4}\int_{-1}^1 t^2(1-t^2)\,dt = 1, \qquad \mathbb{E}_{\pi_j}\Big[\Big(\frac{\pi_j'(t)}{\pi_j(t)}\Big)^2\Big] = 15\int_{-1}^1 t^2\,dt = 10.$$

The first term is therefore d and the second is $\sqrt{10Rd} \le d/2$ for $R \le d/40$. □

The second lemma bounds the same sum from above through privacy. It is the dense form of Lemma 3, with $\|(\mathcal{A}(Z) - \theta)\|_1 \le 2d$ in place of $2k$ and the union bound taken over all d coordinates.

Lemma 6. *Under the assumptions of Theorem 6, let $\mathcal{A}$ take values in Θ_{box} and satisfy user-level (ϵ, δ)-DP with $0 < \epsilon < 1$, and let $L = \log(2d/\delta) \le m$. Then*

$$\sum_{i=1}^{n} \mathbb{E}\big[A_i\big] \le 2n\epsilon \sqrt{\mathbb{E}\|\mathcal{A}(Z) - \theta\|_2^2} \sqrt{\frac{mCc_2}{c(\sigma)}} + C_1\, n\delta d\kappa_x \sqrt{\frac{c_2}{c(\sigma)}} \sqrt{mL},$$

where C_1 is an absolute constant.

Proof of Lemma 6. The proof of Lemma 3 applies verbatim with $T = [d]$. The soundness part of Proposition 2 and $\lambda_{\max}(\Sigma) \le C$ give $\mathbb{E}|A_i'| \le \sqrt{\mathbb{E}\|\mathcal{A}(Z) - \theta\|_2^2} \sqrt{mCc_2/c(\sigma)}$. For the tail, $\|(\mathcal{A}(Z) - \theta)\|_1 \le 2d$ because both vectors lie in Θ_{box}, so $|A_i| \le 2d \max_{\ell \in [d]} |U_{i,\ell}|/c(\sigma)$ with $U_{i,\ell} = \sum_{j=1}^{m} (y_{i,j} - \psi'(x_{i,j}^\top \theta)) x_{i,j,\ell}$. Bernstein's inequality and a union bound over $\ell \in [d]$ give the same tail bound as there with k replaced by d, and the choice $t_0 \asymp d\kappa_x \sqrt{c_2/c(\sigma)}\, \sqrt{mL}$ in Lemma 2 gives the claim. □

Proof of Theorem 6. We may assume that $\mathcal{A}$ takes values in Θ_{box}, since projecting onto this set is a post-processing and does not increase the error at any $\theta \in \Theta_{\mathrm{box}}$. Let $R = \sup_{\theta \in \Theta_{\mathrm{box}}} \mathbb{E}\|\mathcal{A}(Z) - \theta\|_2^2$ and $L = \log(2d/\delta)$.

Suppose first that $R > d/40$. The condition $n\sqrt{m}\,\epsilon \ge \Omega(\sqrt{d})$ gives $d^2/(n^2 m\epsilon^2) \le \mathcal{O}(d)$, so R already exceeds the bound claimed by the theorem and there is nothing to prove. We may therefore assume $R \le d/40$, so that Lemma 5 applies. As in the proof of Theorem 4, the condition on δ together with $m \ge \log(nm)$ gives $L \le m$ and

$$C_1\, n\delta d\kappa_x \sqrt{\frac{c_2}{c(\sigma)}} \sqrt{mL} \le \frac{d}{4}.$$

Combining Lemma 5 with Lemma 6, averaged over π,

$$\frac{d}{2} \le 2n\epsilon\sqrt{R} \sqrt{\frac{mCc_2}{c(\sigma)}} + \frac{d}{4}, \qquad \text{hence} \qquad R \ge \Omega\left(\frac{c(\sigma)}{Cc_2} \cdot \frac{d^2}{n^2 m\epsilon^2} \right).$$

Adding the non-private minimax rate $\Omega\big(c(\sigma)d/(nm)\big)$ for $N = nm$ samples, absorbing C, c_2 and $c(\sigma)$ into the constants and taking square roots gives the statement. □