OPTIMAL USER-LEVEL PRIVATE STOCHASTIC CONVEX OPTIMIZATION AND MEAN ESTIMATION WITH HEAVY TAILS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study stochastic convex optimization with heavy-tailed stochastic gradients under (ε, δ)-user-level differential privacy, which protects each user's entire collection of samples. Instead of assuming a uniformly bounded Lipschitz constant, we only assume bounds on the second- and k-th moments of the random sample Lipschitz parameters. This allows for unbounded, heavy-tailed stochastic gradient distributions and can lead to sharper excess risk bounds. We characterize the minimax optimal excess population risk up to logarithmic factors. Our algorithm achieves the optimal rate with only $\widetilde{O}(1/\varepsilon)$ users. As a byproduct of our algorithmic developments, we obtain tight bounds for user-level private mean estimation under norm-moment assumptions, improving over the guarantees of Esfandiari et al. (2022); Zhao et al. (2024). Finally, for mean estimation under the directional-moment assumption of Agarwal et al. (2025), we close their remaining approximate-DP gap by removing the additional d/ε user requirement and attaining the optimal sample complexity up to logarithmic factors.

1 INTRODUCTION

Stochastic convex optimization (SCO) is a fundamental problem in machine learning. Given samples from an unknown distribution P, the goal is to minimize the population loss

$$\min_{w \in \mathcal{W}} F(w), \qquad F(w) := \mathbb{E}_{z \sim P}[f(w, z)], \tag{1}$$

where $\mathcal{W} \subset \mathbb{R}^d$ is a compact convex domain of ℓ_2-diameter D and $f(\cdot, z)$ is convex. We measure the quality of a solution w by its *excess population risk*, $F(w) - F^*$, where $F^* := \min_{w' \in \mathcal{W}} F(w')$.

In many applications, the data used to train models contains sensitive information, and each person contributes multiple training examples: messages typed on a phone, repeated measurements from a patient, or interactions with an online service. *Differential privacy* (DP) provides a rigorous framework for protecting sensitive data (Dwork et al., 2006). In the *user-level* model, this protection extends to each user's entire collection of samples: the algorithm's output distribution must remain stable when all observations of one user are replaced. Motivated by these considerations, user-level private estimation and SCO have received substantial attention (Levy et al., 2021; Esfandiari et al., 2022; Bassily & Sun, 2023; Asi & Liu, 2024; Lowy et al., 2024; Agarwal et al., 2025).

Much of the user-level SCO literature assumes that the stochastic gradients are *uniformly bounded* (Levy et al., 2021; Asi & Liu, 2024; Lowy et al., 2024). The sample Lipschitz constant is $\sup_{w \in \mathcal{W}} \|\nabla f(w, z)\|$, and uniform boundedness requires $\sup_z \sup_{w \in \mathcal{W}} \|\nabla f(w, z)\| \le L$. The resulting excess risk guarantees scale with the worst-case Lipschitz parameter L. In many problems, however, L can be extremely large or infinite, making these guarantees overly pessimistic or vacuous.

Consequently, a recent line of work studies *heavy-tailed* DP SCO under a weaker assumption: bounded *moments* of the random sample Lipschitz parameter. Instead of assuming a bounded worst-case Lipschitz parameter, one assumes that, for public $0 < G_2 \le G_k$ and some $k \ge 2$:

$$\mathbb{E}_{z \sim P}\left[\sup_{w \in \mathcal{W}} \|\nabla f(w, z)\|^2\right] \le G_2^2, \qquad \mathbb{E}_{z \sim P}\left[\sup_{w \in \mathcal{W}} \|\nabla f(w, z)\|^k\right] \le G_k^k. \tag{2}$$

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.

This assumption allows for unbounded, heavy-tailed stochastic gradients and can yield sharper excess risk bounds when G_k is much smaller than L. For instance, for squared-loss regression with $\mathcal{W}$ contained in a ball of radius R, we have $\sup_{w\in\mathcal{W}} \|\nabla f(w,(x,y))\| \le R\|x\|^2 + |y|\|x\|$. Thus, if moments of the features and responses are bounded, then (2) holds even when data is unbounded and $L = \infty$. Heavy-tailed private estimation and SCO have been studied extensively under *item-level privacy*, which protects one data point at a time (Kamath et al., 2020; Wang et al., 2020; Kamath et al., 2022; Lowy & Razaviyayn, 2023; Asi et al., 2024; Lowy, 2026), and sharp rates are known in several of these settings. Much less is understood about the user-level setting, however.

There are n users, each holding m observations, and all nm observations are drawn i.i.d. from P. Our main contribution is to characterize the minimax optimal excess population risk under second- and kth-moment bounds (2) up to logarithmic factors.

1.1 Our contributions

Before addressing SCO, we study the simpler problem of mean estimation with *bounded norm-moments*: for a sample $z \sim P$ with unknown mean μ, assume $\mathbb{E}_{z\sim P}\|z-\mu\|^j \le \widetilde{G}_j^j$ for $j \in \{2,k\}$. Let $\mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$ denote this distribution class. We use $\widetilde{G}_j$ for these centered moments and reserve G_j for the raw SCO moments in (2). Despite recent progress (Esfandiari et al., 2022; Zhao et al., 2024), existing error bounds for user-level private mean estimation with norm-moment bounds are not tight in every parameter regime; see Table 1 for a summary. Furthermore, mean estimation provides a key building block for our SCO algorithm. This leads to:

> **Question 1.** What is the minimax optimal ℓ_2-error for user-level private mean estimation under norm-moment bounds?

Contribution 1: Optimal rates for user-level DP mean estimation with norm-moment bounds. We answer Question 1 by providing novel tight upper and lower bounds under user-level (ε,δ)-DP (up to logarithms). The minimax optimal ℓ_2-error rate scales as

$$\mathcal{E}_{n,m}(\widetilde{G}_2,\widetilde{G}_k) := \frac{\widetilde{G}_2}{\sqrt{nm}} + \frac{\widetilde{G}_2\sqrt{d}}{n\varepsilon\sqrt{m}} + \min\left\{\widetilde{G}_2\left(\frac{\sqrt{d}}{nm\varepsilon}\right)^{1/2},\ \widetilde{G}_k\left(\frac{\sqrt{d}}{nm\varepsilon}\right)^{1-1/k}\right\}. \tag{3}$$

Our estimator attains the optimal error rate up to logarithmic factors whenever the number of users $n \ge \widetilde{\Omega}_k(1/\varepsilon)$ — a user threshold that is information-theoretically necessary (for finite error guarantees) up to logarithmic factors (Esfandiari et al., 2022, Lemma 5.1).

The three terms in $\mathcal{E}_{n,m}(\widetilde{G}_2,\widetilde{G}_k)$ reflect non-private sampling error, the user-level privacy cost of typical (subgaussian) fluctuations of user averages, and the privacy cost of rare large observations. The second term is the optimal privacy term for bounded-data user-level mean estimation (Esfandiari et al., 2022, Corollary 4.2 and Theorem 5.2), whereas the third term has the form of the optimal item-level DP heavy-tailed error with nm observations (Barber & Duchi, 2014; Lowy & Razaviyayn, 2023), while accounting for both second- and kth-moment constraints.

These results improve the guarantees of prior user-level mean estimators: see Table 1 and Appendix H. Using our mean estimation results and techniques, we next tackle stochastic convex optimization.

> **Question 2.** What is the minimax optimal excess population risk for user-level DP SCO under second- and kth-moment bounds (2)?

Contribution 2: Optimal rates for user-level DP SCO with bounded moments. We answer Question 2 by providing tight upper and lower bounds (up to logarithms). With $n \ge \widetilde{\Omega}_k(1/\varepsilon)$ users, the optimal rate under (2) is

$$\widetilde{\Theta}_k\left(D\min\{G_2, \mathcal{E}_{n,m}(G_2,G_k)\}\right), \tag{4}$$

Table 1: Norm-moment mean estimation with user-level (ε, δ)-DP: constant-probability Euclidean errors. Logarithmic factors are suppressed.

Result	Error bound	Restrictions
Esfandiari et al. (2022) upper bound	$\frac{\widetilde{G}_2\sqrt{d}}{n\varepsilon\sqrt{m}}$	$1/\varepsilon \lesssim n \lesssim \sqrt{d}/\varepsilon$
Zhao et al. (2024) upper bound	$\mathcal{E}_{n,m}(\widetilde{G}_k, \widetilde{G}_k)$	$n \gtrsim d/\varepsilon$; known $R_{\text{public}} \geq \|\mu\|$
Our matching bounds	$\mathcal{E}_{n,m}(\widetilde{G}_2, \widetilde{G}_k)$	$n \gtrsim 1/\varepsilon$
Too few users (Esfandiari et al., 2022, Lemma 5.1)	No finite uniform error	$n \leq \log(1/\delta)/(3\varepsilon)$

Table 2: Summary of SCO rates: excess population risk. Logarithmic factors are suppressed. The prior rows assume a bound $\sup_{w,z} \|\nabla f(w,z)\| \leq L$; our row requires only moment bounds (2).

Result	Excess population risk	Restrictions
Levy et al. (2021)	$LD\min\left\{1,\ \frac{1}{\sqrt{nm}} + \frac{d}{n\varepsilon\sqrt{m}}\right\}$	$n \gtrsim \sqrt{d}/\varepsilon$; additional smoothness restriction
Esfandiari et al. (2022)	$LD\min\left\{1,\ \frac{1}{\sqrt{nm}} + \frac{d}{n\varepsilon\sqrt{m}}\right\}$	$n \gtrsim 1/\varepsilon$; additional smoothness restriction
Bassily & Sun (2023)	$LD\min\left\{1,\ \frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\varepsilon\sqrt{m}}\right\}$	$n \gtrsim \sqrt{d}/\varepsilon$; restricted m, δ, and smoothness parameter
Asi & Liu (2024); Lowy et al. (2024)	$LD\min\left\{1,\ \frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\varepsilon\sqrt{m}}\right\}$	$n \gtrsim 1/\varepsilon$
Our matching bounds	$D\min\{G_2, \mathcal{E}_{n,m}(G_2, G_k)\}$	$n \gtrsim 1/\varepsilon$

where $\mathcal{E}_{n,m}(G_2, G_k)$ is the expression in (3) evaluated at the raw SCO moment bounds. Table 2 compares our result with prior SCO guarantees.

Finally, we revisit mean estimation under an alternative bounded *directional moment* condition that was studied by Agarwal et al. (2025):

$$\sup_{\|v\|=1} \mathbb{E}|\langle z - \mu, v\rangle|^k \leq 1. \tag{5}$$

Agarwal et al. (2025) proves a lower bound on the number of users necessary to achieve ℓ_2-error α:

$$\widetilde{\Omega}_k\left(\frac{d}{m\alpha^2} + \frac{d}{\alpha\varepsilon\sqrt{m}} + \frac{d}{m\varepsilon\alpha^{k/(k-1)}} + \frac{1}{\varepsilon}\right) \tag{6}$$

By contrast, the upper bound of Agarwal et al. (2025) has d/ε in place of the $1/\varepsilon$ term in (6). The following question remained open:

> **Question 3.** Does there exist a user-level DP algorithm for mean estimation under (5) matching the lower bound (6)?

Contribution 3: Optimal directional-moment mean estimation. We give a positive answer to Question 3, providing an optimal estimator whose user complexity matches (6) up to logarithms.

Complete proofs are deferred to the appendices.

1.2 Challenges and techniques

We provide an overview of the key technical challenges and tools that we develop.

1.2.1 Mean estimation with bounded norm-moments.

It is unclear how to achieve the optimal rate throughout the $n \geq \widetilde{\Omega}(1/\varepsilon)$ regime by directly modifying the Huber-loss algorithm of Zhao et al. (2024) or the private ball-finding method of Esfandiari et al.

(2022). Therefore, we depart from these prior approaches for heavy-tailed estimation and instead build on FriendlyCore, which is optimal for *bounded-data* user-level DP estimation with $n \geq \widetilde{\Omega}(1/\varepsilon)$ users (Tsfadia et al., 2022; Asi & Liu, 2024). Applying FriendlyCore verbatim appears suboptimal in the heavy-tailed case: some outlier users can be removed during the filtering stage, introducing bias that can exceed the optimal error. To deal with this, we develop a novel **two-stage algorithm**: first, we use FriendlyCore to provide a coarse estimate of the mean using half the users; we then use the coarse estimate as a clipping center and do a clip-average-and-noise step on the rest of the users.

The clipping center is only coarsely accurate and it is challenging to bound the clipping bias. By **conditioning on a "good clipping center" event**, we show that clipping still leaves user averages near μ unchanged, allowing a Fuk–Nagaev bound on the clipping bias (Mollenhauer & Fiedler, 2025).

Lower bound. For the heavy-tailed lower bound term, we build on the rare-outlier approach of Barber & Duchi (2014); Lowy & Razaviyayn (2023), which gave a loose approximate DP lower bound and a tight zCDP lower bound, respectively. The challenge is obtaining the sharp dependence on dimension under approximate DP. Prior constructions were impeded by using a fixed outlier vector. To overcome this, we **draw outliers randomly**; observing an outlier no longer reveals the mean exactly. We can then apply the fingerprinting lemma of Kamath et al. (2019) to get our tight bound.

1.2.2 SCO.

Our SCO algorithm builds on the population-level localization framework of Asi et al. (2024), reducing heavy-tailed SCO to a sequence of regularized ERM problems. We privately solve each regularized ERM problem by running a user-level DP variant of (stochastic) gradient descent. The key question is *how to estimate the gradient of the empirical loss function in each iteration when user-level DP is required for T adaptive iterates*.

The central challenge is that increasing the privacy noise in the FriendlyCore stage of our algorithm to account for T releases would destroy the accuracy of clipping centers, leading to suboptimal excess risk. Instead of adding noise T times, we develop a novel **fused adaptive variant** of Algorithm 1: each time we run the FriendlyCore step, we keep the coarse center that it outputs internal and don't add any noise to it; we only release the noisy clipped gradient. We combine this technique with *AboveThreshold*.

Another challenge is that *reusing users makes the gradient queries adaptive*. Rather than standard worst-case bias analysis, we control the **average squared empirical clipping bias** via DP stability, avoiding spurious poly(T) factor.

1.2.3 Mean estimation with bounded directional moments.

The key challenge is *removing the d/ε user requirement* of Agarwal et al. (2025) in the intermediate-accuracy regime. We **modify the refinement step** of their algorithm by relaxing their per-round accuracy requirement and proving that the error still decreases toward the requested accuracy.

2 Preliminaries

Write $B(u, r) = \{x : \|x - u\| \leq r\}$ and Π_K for Euclidean projection onto a closed convex set. User i holds $Z_i = (z_{i,1}, \ldots, z_{i,m})$, and $\mathsf{Z} = (Z_1, \ldots, Z_n) \sim P^{nm}$. All norms are Euclidean; I_d is the identity matrix. Assume $0 < \varepsilon \leq 1$, $0 < \delta < 1/2$, fixed $k > 2$, and failure probability $\beta \in (0, 1/4)$. $\widetilde{O}_k$ hides logarithms and constants depending on k, and $C_{(\cdot)}$ and $c_{(\cdot)}$ denote constants (possibly depending on k). Throughout, $\Lambda := \log(enmd/(\varepsilon\delta\beta))$.

Definition 2.1 (User-level differential privacy)**.** An algorithm A is (ε, δ)-user-level DP if, for all datasets Z, Z' differing in one entire user block and all measurable output events S,

$$\Pr(A(\mathsf{Z}) \in S) \leq e^{\varepsilon} \Pr(A(\mathsf{Z}') \in S) + \delta. \tag{7}$$

SCO assumptions. In addition to the SCO assumptions stated in the introduction, sample losses are differentiable on a neighborhood of $\mathcal{W}$. For differentiable h, *convexity* means $h(v) \geq h(w) + \langle \nabla h(w), v - w \rangle$; *$\lambda$-strong convexity* adds $\lambda\|v - w\|^2/2$ to the right-hand side. A function is *L-Lipschitz* if $|h(v) - h(w)| \leq L\|v - w\|$. By (2), F is G_2-Lipschitz, so $F(w) - F^* \leq DG_2 \; \forall w \in \mathcal{W}$.

Assume without loss of generality that $\widetilde{G}_2 \le \widetilde{G}_k$; the SCO parameters already satisfy $G_2 \le G_k$.

A common sufficient condition for our norm-moment *lower bounds* is

$$\delta \le c \min\{1/(nm), \varepsilon^2/d\}. \tag{8}$$

3 Mean estimation under norm-moment bounds

In this section, we provide tight upper and lower bounds for mean estimation over $P \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$.

3.1 Upper bound: FriendlyCore Coarse Estimation, Clip-Average-Noise

For user averages $Y_i = m^{-1} \sum_a z_{i,a}$, our Algorithm 1 uses half the users to obtain a coarse estimate c of μ via FriendlyCore (Algorithm 2) (Asi & Liu, 2024; Tsfadia et al., 2022). It then clips the remaining users' averages around c, averages them, and adds Gaussian noise. The clipping radius R balances clipping bias and privacy noise.

FriendlyCore first tests whether the data set is sufficiently concentrated — i.e. whether most user averages have many nearby neighbors. If not, the algorithm aborts; if yes, it then filters out outlier users (i.e. those with few neighbors). Finally, it adds Gaussian noise u to the average Q of the surviving inlier user averages $\mathcal{S}$ and outputs this vector $\widetilde{Q} = Q + u$ as the coarse estimate of μ.

Algorithm 1 Norm-moment mean estimation

Input: User averages $Y_1, \dots, Y_n$ and public $\widetilde{G}_2, \widetilde{G}_k, m, d, \varepsilon, \delta, \beta$.

1. Split users into two disjoint groups S_1, S_2 of sizes $n_1 = \lfloor n/2 \rfloor$ and $n_2 = n - n_1$.
2. Compute the coarse center using Algorithm 2:
$$c \leftarrow \mathsf{FriendlyCore}\Big((Y_i)_{i\in S_1};\, 2\widetilde{G}_2\sqrt{40/m},\, \varepsilon,\, \delta\Big), \qquad n_1 = |S_1|.$$
3. If $\sqrt{d}/(n\varepsilon) > 1$, return $\hat{\mu} := c$.
4. Set $a = \sqrt{d\log(e/\delta)}/(n\varepsilon)$ and choose $j \in \arg\min_{\ell\in\{2,k\}} \widetilde{G}_\ell (a/m)^{1-1/\ell}$.
5. Set $R = \max\{C_k \Lambda^2 \widetilde{G}_2/\sqrt{m},\, \widetilde{G}_j m^{-(1-1/j)} a^{-1/j}\}$.
6. Return $\hat{\mu} := n_2^{-1} \sum_{i\in S_2} \Pi_{B(c,R)}(Y_i) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma_2^2 I_d)$, $\sigma_2 = 8R\sqrt{\log(e/\delta)}/(n_2\varepsilon)$.

Theorem 3.1 (Norm-moment estimation upper bound)**.** *For every $P \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$ and $n \ge C_k\Lambda^4/\varepsilon$, there exist algorithmic parameters such that Algorithm 1 is user-level (ε,δ)-DP and satisfies $\|\widehat{\mu} - \mu\| \le \widetilde{O}_k(\mathcal{E}_{n,m}(\widetilde{G}_2, \widetilde{G}_k))$ with probability at least $1-\beta$.*

Privacy proof sketch. FriendlyCore is (ε,δ)-DP by Asi & Liu (2024, Lemma 3.6). For fixed c, the second-stage average has sensitivity $2R/n_2$, so the Gaussian mechanism is also (ε,δ)-DP. Since the stages use disjoint users, parallel composition gives the claim.

Utility proof sketch. At a high level, the argument goes as follows: We first show that the coarse center provided by FriendlyCore is accurate with high probability (Lemmas 3.2 and 3.3). We then argue that conditional on this "good center" event, the clipping bias in the second stage of the algorithm is sufficiently small if the clip threshold R is sufficiently large (Lemma 3.4).

Step 1: Accuracy of FriendlyCore. With bounded or subexponential data, no users get filtered out with high probability and accuracy follows. With heavy tails, some user averages can be far from μ and hence not every user survives the FriendlyCore filter. However, a large majority of user averages are near μ and hence survive the filter:

Lemma 3.2 (Large majority core)**.** *Define* $\mathrm{GOOD} := \{i \in S_1 : \|Y_i - \mu\| \le \tau/2\}$ *and* $\tau = 2\widetilde{G}_2\sqrt{40/m}$*. With probability at least* $1 - e^{-n_1/800}$*, we have* $|\mathrm{GOOD}| \ge 0.95 n_1$.

Chebyshev and Hoeffding prove Lemma 3.2. The large majority core does not give optimal accuracy from FriendlyCore alone, but it does provide a fairly accurate coarse estimate of μ:

Lemma 3.3 (Accuracy of the coarse center)**.** *Suppose* $n_1 \geq \frac{1000}{\varepsilon}\log\frac{16}{\delta\beta}$. *Conditional on* $|\mathrm{GOOD}| \geq 0.95n_1$*, with probability at least* $1-2\beta$*, the output* $\widetilde{Q}$ *of Algorithm 2 satisfies*

$$\|\widetilde{Q}-\mu\| \leq \widetilde{O}\left(\frac{\widetilde{G}_2}{\sqrt{m}} + \frac{\widetilde{G}_2\sqrt{d+\log(1/\beta)}}{n_1\varepsilon\sqrt{m}}\right). \tag{9}$$

Proof idea. With high probability, every GOOD user survives (i.e. GOOD $\subseteq \mathcal{S}$), and every surviving user has a GOOD neighbor within distance 2τ. Thus the average Q of the surviving user-averages is $O(\tau)$-close to μ; $\tau = 2\widetilde{G}_2\sqrt{40/m}$. Gaussian noise contributes the second term in (9). □

If $\sqrt{d}/(n\varepsilon) > 1$, returning $c = \widetilde{Q}$ already attains (3), up to logarithms, by Lemma 3.3. Otherwise, stage two refines the estimate.

Step 2: Bounding clipping bias on "good center" event. On the "good clipping center" event $\|c-\mu\| \leq R/2$, clipping around $c = \widetilde{Q}$ is almost as good as clipping around μ because clipping only affects user averages who are at least $R/2$-far from μ. Only these outlier users contribute to the bias. Thus the bias can be controlled by the tail of the random variable $\frac{1}{m}\sum_{j=1}^{m} z_j - \mu$, which can be tightly bounded via Fuk-Nagaev inequality, yielding:

Lemma 3.4 (Clipping bias around a good center)**.** *Let* Y *average* m *independent observations from* $\mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$. *If* $\|c-\mu\| \leq R/2$ *and* $R \gtrsim_k \widetilde{G}_2/\sqrt{m}$*, then*

$$\begin{aligned}
&\|\mathbb{E}\Pi_{B(c,R)}Y - \mu\| \leq C\widetilde{G}_2^2/(mR),\\
&\|\mathbb{E}\Pi_{B(c,R)}Y - \mu\| \leq C_k\left[\frac{\widetilde{G}_2}{\sqrt{m}}e^{-c_k mR^2/\widetilde{G}_2^2} + \frac{\widetilde{G}_k^k}{m^{k-1}R^{k-1}}\right].
\end{aligned} \tag{10}$$

Finally, to obtain the error bound in Theorem 3.1, we choose the clipping threshold R optimally to balance the bias term (Lemma 3.4) and the Gaussian noise term $\sigma_2\sqrt{d}$, subject to the constraint that the good clipping center event occurs with high probability. By Lemma 3.3, this constraint can be satisfied by choosing R at least as large the right-hand-side of (9).

3.2 Matching lower bound

The following tight lower bound establishes the optimality of Algorithm 1 up to logarithmic factors in the regime $n \geq \widetilde{\Omega}_k(1/\varepsilon)$.

Theorem 3.5 (Norm-moment estimation lower bound)**.** *Let* $\delta \leq c\min\{\frac{1}{nm}, \frac{\varepsilon^2}{d}\}$ *for sufficiently small* $c > 0$. *For every user-level* (ε,δ)*-DP estimator, there exists a distribution* $P \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$ *such that* $\Pr(\|\widehat{\mu}-\mu\| \geq c_k\mathcal{E}_{n,m}(\widetilde{G}_2, \widetilde{G}_k)) \geq 1/3$.

Separately, Esfandiari et al. (2022, Lemma 5.1) show that if $\delta \leq \varepsilon^2$ and $n \leq \log(1/\delta)/(3\varepsilon)$ then no user-level DP algorithm can achieve a finite uniform error bound for unrestricted mean location. Thus, the user count $n \geq \widetilde{\Omega}_k(1/\varepsilon)$ of our algorithm is tight up to logarithms.

Proof overview. Recall the three terms in (3). The nonprivate term is classical Duchi (2023). For the Gaussian privacy term, we adapt the fingerprinting argument of Agarwal et al. (2025, Appendix D), scaling the covariance to satisfy both norm-moment bounds; translation lets the mean range be larger without changing the centered moments. For the heavy-tailed privacy term, we modify the rare-outlier construction of Barber & Duchi (2014); Lowy & Razaviyayn (2023): we calibrate the magnitude of the outliers to satisfy both moment bounds, and randomly sample its direction from a product distribution on $\{-1,1\}^d$. This randomization permits the approximate-DP fingerprinting lemma of Kamath et al. (2019, Lemma 53); grouping the nm observations into n users then gives the user-level bound via Levy et al. (2021).

4 STOCHASTIC CONVEX OPTIMIZATION UNDER NORM-MOMENT BOUNDS

In this section, we address the SCO problem (1) with bounded moments (2) under user-level DP.

4.1 UPPER BOUND: LOCALIZED SGD WITH FUSED ADAPTIVE PRIVATE GRADIENT ESTIMATOR

Our SCO algorithm builds on the population-level localization framework of Asi et al. (2024), reducing heavy-tailed SCO to a sequence of regularized ERM problems: see Algorithm 3. We privately solve each regularized ERM problem by running a user-level DP variant of gradient descent.

4.1.1 POPULATION LOCALIZATION REDUCES SCO TO REGULARIZED ERM

For a fresh dataset Z of s users with m observations each, write $\widehat{F}_Z$ for the empirical loss. For a regularization center $w_0 \in \mathcal{W}$ and $\lambda > 0$, define

$$\widehat{F}_{\lambda,Z}(w) = \frac{1}{sm}\sum_{i,a} f(w, z_{i,a}) + \frac{\lambda}{2}\|w - w_0\|^2, \qquad \widehat{w}_{\lambda,Z} = \arg\min_{w\in\mathcal{W}} \widehat{F}_{\lambda,Z}(w). \tag{11}$$

The population-localization framework of Asi et al. (2024), as abstracted in Lowy (2026, Theorem 2.1), reduces the problem to a private solver A_{ERM} for (11) satisfying

$$\Pr\big(\|A_{\mathrm{ERM}}(Z; w_0, \lambda) - \widehat{w}_{\lambda,Z}\| \le C_k \Lambda^4 \mathcal{E}_{s,m}(G_2, G_k)/\lambda\big) \ge 0.8. \tag{12}$$

Here $\mathcal{E}_{s,m}(G_2, G_k)$ denotes the rate in (3) with n replaced by s and moment bounds G_2, G_k.

The population localization meta-Algorithm 3 partitions the n users into disjoint blocks. At phase i, it runs A_{ERM} on fresh blocks centered at the preceding output, aggregates the estimates, and increases the regularization parameter.

Theorem 4.1 (Population localization (Asi et al., 2024; Lowy, 2026)). *Let A_{ERM} be (ε, δ)-user-level DP and suppose it satisfies* (12) *for every $w_0 \in \mathcal{W}$ and every regularization parameter used by Algorithm 3 on each fresh s-user block. If $n \ge C_k\Lambda^6/\varepsilon$, then Algorithm 3 is (ε, δ)-user-level DP and, with probability at least $1 - \beta$,*

$$F(\widehat{w}) - F^* \le D \min\{G_2, C_k \Lambda^8 \mathcal{E}_{n,m}(G_2, G_k)\}. \tag{13}$$

Thus, it only remains to construct a user-level DP regularized ERM solver A_{ERM} satisfying (12).

4.1.2 SGD WITH NOISY BIASED GRADIENTS

We will instantiate Algorithm 3 with a user-level DP variant of SGD for solving (11). Our SGD method will involve biased, noisy estimates of the gradient $\nabla\widehat{F}_Z(w_t)$. Unlike the worst-case-bias analyses of Asi et al. (2024); Lowy & Razaviyayn (2023), the following lemma controls distance to the empirical minimizer using the *average squared bias*. This avoids requiring every gradient estimate to be accurate.

Lemma 4.2 (Excess empirical risk of biased noisy SGD). *Let $\mathcal{F}_t$ be a filtration such that Z, w_0, w_t are $\mathcal{F}_t$-measurable. For gradient estimates g_t and updates*

$$w_{t+1} = \Pi_{\mathcal{W}}(w_t - \eta_t[g_t + \lambda(w_t - w_0)])\,,$$

define the bias $b_t := \mathbb{E}[g_t \mid \mathcal{F}_t] - \nabla\widehat{F}_Z(w_t)$ and noise $v_t := g_t - \mathbb{E}[g_t \mid \mathcal{F}_t]$. Assume $\mathbb{E}\|\nabla\widehat{F}_Z(w_t)\|^2 \le G_2^2$ and $\mathbb{E}[\|v_t\|^2 \mid \mathcal{F}_t] \le \Sigma^2$ for every t. Let $B^2 = \frac{1}{T}\sum_{t=1}^T \mathbb{E}\|b_t\|^2$, $\eta_t = 4/(\lambda(t+1))$, and $\bar{w} := 2\sum_{t=1}^T t w_t/[T(T+1)]$. Then with probability at least 0.9,

$$\|\bar{w} - \widehat{w}_{\lambda,Z}\| \lesssim \frac{1}{\lambda}\left(B + \frac{G_2 + \Sigma + \lambda D}{\sqrt{T}}\right). \tag{14}$$

The main remaining task is to provide a private estimator g_t of $\nabla\widehat{F}_Z(w_t)$ at every iteration with bias B and noise Σ such that (14) implies (12) for suitable T.

4.1.3 SOLVING REGULARIZED ERM: SGD WITH A FUSED DP GRADIENT ESTIMATOR

To extend our mean estimation procedure to the adaptive setting—where w_t depends on the private data—without incurring excessive clipping bias or privacy noise, we develop a novel *fused adaptive* variant of Algorithm 1.

Fix a block S of s users and an initial w_0. At iteration t, form $\nabla \widehat{F}_{Z_i}(w_t) := m^{-1} \sum_a \nabla f(w_t, z_{i,a})$ for each user $i \in S$. An AboveThreshold test checks whether users are τ-concentrated with $\tau \approx G_2/\sqrt{m}$. If accepted, apply the FriendlyCore filter, let $\mathcal{S}_t$ be its selected set of users, and set $Q_t = |\mathcal{S}_t|^{-1} \sum_{i \in \mathcal{S}_t} \nabla \widehat{F}_{Z_i}(w_t)$ when $\mathcal{S}_t \neq \varnothing$. Keep Q_t noiseless and internal; release only

$$g_t := A_t + v_t, \qquad A_t := \frac{1}{s} \sum_{i=1}^{s} \Pi_{B(Q_t,R)} \nabla \widehat{F}_{Z_i}(w_t), \qquad v_t \sim \mathcal{N}(0, \sigma^2 I_d). \tag{15}$$

Then use the update, step sizes, and returned weighted average of Lemma 4.2. If the AboveThreshold test is rejected (user averages insufficiently concentrated) or the user core $\mathcal{S}_t$ is empty, the algorithm sets $g_t = 0$ from that point onward. Algorithm 4 in Appendix E.2 gives the full user-level DP SGD algorithm that we use as our regularized ERM solver A_{ERM}.

Proposition 4.3 (Private regularized ERM solver—informal)**.** *For $s \geq C_k \Lambda/\varepsilon$, there are choices of parameters in Algorithm 4 for which it is (ε, δ)-user-level DP and satisfies* (12)*.*

We now discuss the main challenges and ideas of the proof of Proposition 4.3.

Privacy. The main challenge is bounding the sensitivity of the fused gradient estimator $s^{-1} \sum_{i=1}^{s} \Pi_{B(Q_t,R)} \nabla \widehat{F}_{Z_i}(w_t)$: replacing one user changes the clipping center Q_t—which can in turn change every clipped gradient—while also shifting the final average of clipped gradients.

The FriendlyCore coupling of Asi & Liu (2024) controls the movement of Q_t: with high probability (accounted for in δ), its movement is bounded by $\widetilde{O}(\tau/s)$. Combining this with the $O(R/s)$ bound contribution of the changed user to the final average of the clipped user-gradients yields a sensitivity bound of the fused estimator by $\widetilde{O}((\tau + R)/s)$.

Utility. The key challenge is bounding the empirical *bias*

$$b_t = \frac{1}{s} \sum_{i=1}^{s} \left[\Pi_{B(Q_t,R)} \left(\nabla \widehat{F}_{Z_i}(w_t) \right) - \nabla \widehat{F}_{Z_i}(w_t) \right] = A_t - \nabla \widehat{F}_Z(w_t) \tag{16}$$

of adaptive iterates (before any abort): we reuse the same users at every iteration, so w_t depends on their data. Conditioning on w_t therefore breaks independence and we cannot directly apply the fixed-query analysis from Section 3.

Lemma 4.4 (Adaptive empirical clipping bias – informal)**.** *Recall the definition of b_t in* (16)*. Let $\psi_2(R) := \frac{CG_2^2}{mR}$ and $\psi_k(R) := C_k \left[\left(R + \frac{G_2^2}{mR} \right) e^{-c_k m R^2/G_2^2} + \frac{G_k^k}{m^{k-1} R^{k-1}} \right]$. There exist algorithmic parameters such that for either fixed choice of $j \in \{2, k\}$, with probability at least* 0.97*,*

$$\frac{1}{T} \sum_{t=1}^{T} \|b_t\|^2 \leq C_k \left[\psi_j(R)^2 + \frac{G_2^2}{sm} \right]. \tag{17}$$

Proof sketch. With an accurate coarse center, clipping bias is bounded by the average contribution of outlier users. Expanding the square of this average gives one- and two-user terms. Independence controls these terms at fixed queries; DP stability lets us control them at adaptive queries by comparison with independent replacement users. Averaging over iterations and applying Markov's inequality yields (17), without a union bound over iterations. □

An obstacle in the proof of Proposition 4.3 is that Lemma 4.4 gives a constant-probability guarantee on the bias, whereas Lemma 4.2 requires an in-expectation bias bound. Thus, we analyze a variation of Algorithm 4 that switches to unbiased noisy empirical gradients before its average squared bias exceeds the bound in (17). Its average squared bias is then bounded on every run, hence also in expectation, so Lemma 4.2 applies. Whenever (17) holds, no switch occurs and the two outputs agree. Thus the original algorithm satisfies (12) with constant probability.

4.1.4 THE FINAL UPPER BOUND

Combining Theorem 4.1 with Proposition 4.3 yields our SCO upper bound.

Theorem 4.5 (Norm-moment SCO upper bound)**.** *Under* (2)*, if $n \geq C_k \Lambda^6/\varepsilon$, Algorithm 3 instantiated with Algorithm 4 as the private ERM solver is (ε, δ)-user-level-DP and, with probability at least $1-\beta$, has output satisfying*

$$F(\widehat{w}) - F^* \leq D \min\{G_2, C_k \Lambda^8 \mathcal{E}_{n,m}(G_2, G_k)\}. \tag{18}$$

4.2 LOWER BOUND

Theorem 4.6 (SCO lower bound)**.** *Under* (8)*, on a ball of diameter D, every user-level (ε, δ)-DP algorithm $\mathcal{A}$ has a distribution of linear losses satisfying* (2) *such that*

$$\mathbb{E}[F(\mathcal{A}(\mathsf{Z})) - F^*] \geq c_k D \min\{G_2, \mathcal{E}_{n,m}(G_2, G_k)\}. \tag{19}$$

Since the trivial algorithm $\mathcal{A}(\mathsf{Z}) = w_0$ achieves excess risk DG_2, we conclude by Theorem 4.5 that our algorithm is optimal up to logarithmic factors.

Proof overview. The nonprivate term is classical (Nemirovski et al., 2009). For the private terms, use linear losses: a restricted, rescaled Gaussian family satisfies the gradient moment bounds and yields the user-level term by fingerprinting. For the heavy-tailed term, we make nonzero Gaussian gradients rare and increase their magnitude, reducing the number of informative observations.

5 CLOSING THE DIRECTIONAL-MOMENT MEAN-ESTIMATION GAP

Assume (5) and let $\alpha > 0$ denote target Euclidean accuracy.

Theorem 5.1 (Optimal directional-moment estimation)**.** *There is a computationally efficient user-level (ε, δ)-DP estimator with constant-probability ℓ_2-error at most α using*

$$n = \widetilde{O}_k \left[\frac{d}{m\alpha^2} + \frac{d\sqrt{\log(1/\delta)}}{\alpha\varepsilon\sqrt{m}} + \frac{d\sqrt{\log(1/\delta)}}{m\varepsilon\alpha^{k/(k-1)}} + \frac{\log(1/\delta)}{\varepsilon} \right]. \tag{20}$$

This is tight up to logarithms if (8) *holds and $\delta \leq c_k(\varepsilon \min\{1, \alpha\}/d)^{C_k}$ for sufficiently large C_k.*

Proof idea. For $\alpha \leq 1/\sqrt{m}$, the estimator of Agarwal et al. (2025) is already optimal. For $\alpha \geq \sqrt{d/m}$, we apply our norm-moment estimator with $\widetilde{G}_2 = \widetilde{G}_k = \sqrt{d}$. For $1/\sqrt{m} \leq \alpha \leq \sqrt{d/m}$, we refine a private coarse estimate by repeatedly clipping fresh user averages around the current estimate, averaging, and adding noise. The refinement approach of Agarwal et al. (2025) requires error at most $\widetilde{O}(R_t/\sqrt{d})$ for clipping radius R_t, forcing $n \geq \widetilde{\Omega}(d/\varepsilon)$. We relax this requirement: we choose the radius from a known upper bound on the current center error and prove that this bound decreases toward the requested accuracy α, reaching it in log-log rounds. Appendix G gives the algorithms and the upper and lower bound proofs.

6 CONCLUSION

We characterized the minimax optimal rates (up to logarithms) for three fundamental problems in (ε, δ)-user-level DP learning with heavy tails: stochastic convex optimization, mean estimation under bounded norm-moments, and mean estimation under bounded directional moments. Our bounds are tight throughout the $n \geq \widetilde{\Omega}(1/\varepsilon)$ regime, and our algorithms are efficient. Our work paves the way for several interesting future directions: What are the optimal rates under pure user-level DP? Can the runtime of our algorithms be improved? Can we develop practical algorithms for training large-scale ML models with user-level DP?

AI USE STATEMENT

Generative AI tools (ChatGPT) were used to assist with literature searches and comparisons, develop and revise mathematical proof arguments, and identify potential gaps. They also assisted with LaTeX organization, equation formatting, and compilation checks. This assistance included some mathematical suggestions, not only copyediting. The authors are responsible for the final text, mathematical claims, and citations.

ETHICS STATEMENT

This paper is theoretical and uses no human-subject experiments or private empirical datasets. Its guarantees apply under the stated adjacency relation and statistical assumptions.

REPRODUCIBILITY STATEMENT

Appendices A–G provide algorithmic definitions, parameter choices, theorem statements, and proofs, including the small-δ and user-count conditions. The source package includes the official ICLR style files and all LaTeX and bibliography sources needed to reproduce this document. No experimental results are reported.

REFERENCES

Sushant Agarwal, Gautam Kamath, Mahbod Majid, Argyris Mouzakis, Rose Silver, and Jonathan Ullman. Private mean estimation with person-level differential privacy. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2819–2880, 2025. doi: 10.1137/1.9781611978322.92. URL https://epubs.siam.org/doi/10.1137/1.9781611978322.92.

Hilal Asi and Daogao Liu. User-level differentially private stochastic convex optimization: Efficient algorithms with optimal rates. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4240–4248, 2024. URL https://proceedings.mlr.press/v238/liu24g.html.

Hilal Asi, Daogao Liu, and Kevin Tian. Private stochastic convex optimization with heavy tails: Near-optimality from simple reductions. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6cc066e48b391a749c2161f271fda80b-Abstract-Conference.html.

Rina Foygel Barber and John C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

Raef Bassily and Ziteng Sun. User-level private stochastic convex optimization with optimal rates. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1838–1851, 2023. URL https://proceedings.mlr.press/v202/bassily23a.html.

John Duchi. Lecture notes on statistics and information theory, 2023. URL https://web.stanford.edu/class/stats311/lecture-notes.pdf.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042. URL https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf.

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, 2006.

Hossein Esfandiari, Vahab Mirrokni, and Shyam Narayanan. Tight and robust private mean estimation with few users. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16383–16412, 2022. URL https://proceedings.mlr.press/v162/narayanan22a.html.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1853–1902, 2019. URL https://proceedings.mlr.press/v99/kamath19a.html.

Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2204–2235, 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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10633–10660, 2022.

Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12466–12479, 2021.

Andrew Lowy. Optimal rates for pure ε-differentially private stochastic convex optimization with heavy tails. *arXiv preprint arXiv:2604.06492*, 2026.

Andrew Lowy and Meisam Razaviyayn. Private stochastic optimization with large worst-case lipschitz parameter: Optimal rates for (non-smooth) convex losses and extension to non-convex losses. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 986–1054, 2023.

Andrew Lowy, Daogao Liu, and Hilal Asi. Faster algorithms for user-level private stochastic convex optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ae338b3fe6624798efa80b6581924028-Abstract-Conference.html.

Mattes Mollenhauer and Christian Fiedler. Fuk–nagaev inequality in smooth banach spaces: Optimum bounds for distributions of heavy-tailed martingales. *arXiv preprint arXiv:2512.10012*, 2025.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21828–21863, 2022. URL https://proceedings.mlr.press/v162/tsfadia22a.html.

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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10081–10091, 2020.

Puning Zhao, Lifeng Lai, Li Shen, Qingming Li, Jiafei Wu, and Zhe Liu. A huber loss minimization approach to mean estimation under user-level differential privacy. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ead8e195df586b9e7c10fcd114a6b9d1-Abstract-Conference.html.

Notation. We retain the main-body conventions: $\mathsf{Z} = (Z_1, \dots, Z_n)$ denotes the full user dataset, while Z denotes the fresh s-user block in a regularized ERM problem. The centered norm-moment bounds are $\widetilde{G}_2, \widetilde{G}_k$; the raw SCO bounds are G_2, G_k. The rate $\mathcal{E}_{s,m}$ is (3) with n replaced by s, and $\Lambda = \log(enmd/(\varepsilon\delta\beta))$ retains the full user count n. For a positive integer r, write $[r] = \{1, \dots, r\}$, and write $(x)_+ = \max\{x, 0\}$. Freshly drawn noises and filter coins are independent of previous draws unless stated otherwise.

A Norm-moment mean estimation: upper-bound proofs

Theorem A.1 (Restatement of Theorem 3.1)**.** *For every $P \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$ and $n \geq C_k\Lambda^4/\varepsilon$, there exist algorithmic parameters such that Algorithm 1 is user-level (ε, δ)-DP and satisfies $\|\widehat{\mu} - \mu\| \leq \widetilde{O}_k(\mathcal{E}_{n,m}(\widetilde{G}_2, \widetilde{G}_k))$ with probability at least $1 - \beta$.*

We prove Theorem A.1 in two steps. Lemmas A.3 and A.4 give a sufficiently accurate coarse center. Lemma A.6 then bounds the bias, sampling error, and privacy noise of clipping an independent group of users around that center. The final proof chooses the radius to balance these terms. The FriendlyCore privacy guarantee is inherited from prior work.

A.1 FriendlyCore subroutine

Algorithm 2 Coarse FriendlyCore estimator (Tsfadia et al., 2022; Asi & Liu, 2024)

Procedure: $\mathsf{FriendlyCore}((Y_i)_{i=1}^{n_1}; \tau, \varepsilon, \delta)$; n_1 is the number of supplied user averages and τ is the radius.

1. *Public size check.* Set $C_{\mathrm{FC}} = 1000$ and $L_{\mathrm{FC}} = \log(16/\delta)$. If $n_1 < C_{\mathrm{FC}}L_{\mathrm{FC}}/\varepsilon$, return 0.
2. *Concentration test.* Compute $s_{\mathrm{conc}} = n_1^{-1}\sum_{i,j=1}^{n_1} \mathbf{1}\{\|Y_i - Y_j\| \leq \tau\}$. Add $\mathrm{Lap}(20/\varepsilon)$ noise; if the noisy score is below $4n_1/5$, return 0.
3. *Outlier removal.* For each $i \in [n_1]$, set $h_i = \sum_{j=1}^{n_1} \mathbf{1}\{\|Y_i - Y_j\| \leq 2\tau\}$ and
$$p_i = \begin{cases} 0, & h_i < n_1/2, \\ (h_i - n_1/2)/(n_1/6), & n_1/2 \leq h_i < 2n_1/3, \\ 1, & h_i \geq 2n_1/3. \end{cases}$$
4. Starting with $\mathcal{S} = \varnothing$, independently add each i to $\mathcal{S}$ with probability p_i.
5. If $\mathcal{S} = \varnothing$, return 0. Otherwise let $Q = |\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}} Y_i$.
6. Return $\widetilde{Q} = Q + u$, where $u \sim \mathcal{N}(0, \sigma_1^2 I_d)$ is independent and
$$\sigma_1 = \frac{C_{\mathrm{FC}}\tau L_{\mathrm{FC}}^{3/2}}{n_1\varepsilon}.$$

Lemma A.2 (Privacy of Algorithm 2; Asi & Liu, 2024, Lemma 3.6)**.** *Algorithm 2, with n_1 input user averages, is (ε, δ)-user-level DP on every dataset.*

This follows from the single-query privacy argument of Asi & Liu (2024, Lemma 3.6). Their paper covers bounded data but the privacy proof is distribution-free.

The two-stage estimator. For user i, let $Y_i = m^{-1}\sum_a z_{i,a}$, so $\mathbb{E}Y_i = \mu$ and $\mathbb{E}\|Y_i - \mu\|^2 \leq \widetilde{G}_2^2/m$. Split the users into S_1, S_2 of sizes $n_1 = \lfloor n/2 \rfloor$, $n_2 = n - n_1$. Set $\tau = 2\widetilde{G}_2\sqrt{40/m}$. Run Algorithm 2 on S_1 and call its output $c = \widetilde{Q}$. If $\sqrt{d}/(n\varepsilon) > 1$, return c. Otherwise return
$$\widehat{\mu} = \frac{1}{n_2}\sum_{i\in S_2} \Pi_{B(c,R)}Y_i + \xi, \quad \xi \sim \mathcal{N}(0, \sigma_2^2 I_d), \quad \sigma_2 = \frac{8R\sqrt{\log(e/\delta)}}{n_2\varepsilon}. \tag{21}$$

A.2 COARSE LOCALIZATION

We need a large majority of user averages near the mean, not a bound on every user average. The following elementary consequence of Chebyshev's and Hoeffding's inequalities supplies that majority.

Define

$$\mathrm{GOOD} := \{i \in S_1 : \|Y_i - \mu\| \le \tau/2\}.$$

Lemma A.3 (Restatement of Lemma 3.2)**.** *With probability at least* $1 - e^{-n_1/800}$*,*

$$|\mathrm{GOOD}| \ge 0.95 n_1.$$

Proof. Chebyshev gives $\mathbb{P}(\|Y_i - \mu\| > \tau/2) \le 1/40$. Apply Hoeffding to the independent Bernoulli indicators $\mathbf{1}\{i \notin \mathrm{GOOD}\}$. □

Once that majority exists, every selected point is near a good point, even though a few user averages may be arbitrarily far away. This gives the coarse-center guarantee needed for the second stage.

Lemma A.4 (Restatement of Lemma 3.3)**.** *Suppose*

$$n_1 \ge \frac{1000}{\varepsilon} \log \frac{16}{\delta\beta}.$$

Conditional on $|\mathrm{GOOD}| \ge 0.95 n_1$*, with probability at least* $1 - 2\beta$*, the output* $\widetilde{Q}$ *of Algorithm 2 satisfies*

$$\|\widetilde{Q} - \mu\| \le \widetilde{O}\left(\frac{\widetilde{G}_2}{\sqrt{m}} + \frac{\widetilde{G}_2 \sqrt{d + \log(1/\beta)}}{n_1 \varepsilon \sqrt{m}} \right). \tag{22}$$

Proof. Every pair of good users is within distance τ, so the concentration score is at least

$$\frac{|\mathrm{GOOD}|^2}{n_1} > 0.9 n_1.$$

The Laplace tail bound therefore implies that the test accepts with probability at least $1 - \beta$.

Every good user has $h_i \ge |\mathrm{GOOD}| > 2n_1/3$, hence every good user is selected. Conversely, a selected user has $h_i \ge n_1/2$. Since fewer than $0.05 n_1$ users are bad, every selected user has a good neighbor within distance 2τ. Thus, for every $i \in \mathcal{S}$,

$$\|Y_i - \mu\| \le 5\tau/2.$$

Hence $\|Q - \mu\| \le 5\tau/2$. The stated result follows from the standard Gaussian norm bound for the final FriendlyCore noise. □

A.3 CLIPPING AROUND A GOOD CENTER

The center need not have the final estimation accuracy. Its role is to ensure that clipping leaves user averages near the mean unchanged. We first restate the bias bound from the main body, then include sampling and Gaussian noise.

Lemma A.5 (Restatement of Lemma 3.4)**.** *Let* Y *average* m *independent observations from* $P \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$*. If* $\|c - \mu\| \le R/2$ *and* $R \ge C_k \widetilde{G}_2/\sqrt{m}$*, then*

$$\|\mathbb{E}\Pi_{B(c,R)} Y - \mu\| \le C \widetilde{G}_2^2/(mR)$$

and

$$\|\mathbb{E}\Pi_{B(c,R)} Y - \mu\| \le C_k \left[\frac{\widetilde{G}_2}{\sqrt{m}} e^{-c_k m R^2/\widetilde{G}_2^2} + \frac{\widetilde{G}_k^k}{m^{k-1} R^{k-1}} \right].$$

Proof. Since $\|c - \mu\| \le R/2$, clipping can change Y only if $\|Y - \mu\| > R/2$. Its displacement is at most $(\|Y - \mu\| + R/2)\mathbf{1}\{\|Y - \mu\| > R/2\}$. Consequently,

$$\|\mathbb{E}\Pi_{B(c,R)} Y - \mu\| \le 2\mathbb{E}[\|Y - \mu\| \mathbf{1}\{\|Y - \mu\| > R/2\}]. \tag{23}$$

The bound $\mathbb{E}\|Y-\mu\|^2 \le \widetilde{G}_2^2/m$ gives the first claim. For the higher-moment bound, Corollary 2.3 of Mollenhauer & Fiedler (2025), applied to the sum of m independent centered vectors, gives

$$\Pr(\|Y-\mu\| > t) \le C_k \left[e^{-c_k m t^2/\widetilde{G}_2^2} + \frac{\widetilde{G}_k^k}{m^{k-1}t^k} \right].$$

Integrate this tail using $\mathbb{E}[V\mathbf{1}\{V>a\}] = a\Pr(V>a) + \int_a^\infty \Pr(V>t)\,dt$ for nonnegative V. Substitution in (23) yields

$$\|\mathbb{E}\Pi_{B(c,R)}Y - \mu\| \le C_k \left[\left(R + \frac{\widetilde{G}_2^2}{mR} \right) e^{-c_k m R^2/\widetilde{G}_2^2} + \frac{\widetilde{G}_k^k}{m^{k-1}R^{k-1}} \right]. \tag{24}$$

The bound (24) is valid for every $R > 0$. To obtain the lemma's simpler prefactor, put $x = R\sqrt{m}/\widetilde{G}_2 \ge 1$ and use $(x + x^{-1})e^{-cx^2} \le Ce^{-c'x^2}$. This gives the stated exponential prefactor $\widetilde{G}_2/\sqrt{m}$, after adjusting constants. □

We now add sampling error and Gaussian noise. Conditioning on the first-stage center is legitimate because the second group of users is independent of it.

Lemma A.6 (Conditional clipped-mean bound)**.** *Let c be independent of $Y_1, \dots, Y_{n_2}$ and satisfy $\|c-\mu\| \le R/2$. For the estimator in* (21)*, with probability at least $1-2\beta$,*

$$\begin{aligned}\|\widehat{\mu}-\mu\| \lesssim_k \widetilde{G}_2\sqrt{\frac{\log(1/\beta)}{mn_2}} + \frac{R\log(1/\beta)}{n_2} \\ + \left(R + \frac{\widetilde{G}_2^2}{mR} \right) e^{-c_k m R^2/\widetilde{G}_2^2} + \frac{\widetilde{G}_k^k}{m^{k-1}R^{k-1}} \\ + \frac{R\sqrt{\log(1/\delta)}(\sqrt{d}+\sqrt{\log(1/\beta)})}{n_2\varepsilon}.\end{aligned} \tag{25}$$

The exponential and polynomial bias terms in the second line may instead be replaced by $C\widetilde{G}_2^2/(mR)$, using only the second-moment bound.

Proof. For this proof, write $X_i = \Pi_{B(c,R)}Y_i$ and condition on c. The X_i are independent, and

$$\widehat{\mu} - \mu = \left(\frac{1}{n_2}\sum_i X_i - \mathbb{E}X_i \right) + (\mathbb{E}X_i - \mu) + \xi.$$

Because $\mu \in B(c,R)$, nonexpansiveness gives

$$\mathbb{E}\|X_i-\mu\|^2 \le \mathbb{E}\|Y_i-\mu\|^2 \le \frac{\widetilde{G}_2^2}{m}, \qquad \|X_i - \mathbb{E}X_i\| \le 2R.$$

The Hilbert-space Bernstein inequality of Pinelis (1994) gives the first line of (25) with failure probability at most β. Lemma A.5, specifically (24), supplies the bias terms. The Gaussian norm bound supplies the last line with failure probability at most β. A union bound completes the proof. □

A.4 Two public choices of moment order

Write $a = \sqrt{d\log(e/\delta)}/(n\varepsilon)$ for the privacy-noise coefficient in this radius calculation. Select $j \in \{2,k\}$ to minimize $\widetilde{G}_j(a/m)^{1-1/j}$ using only public parameters; only the selected algorithm is run. When the second stage is used, set

$$R = \max\left\{ C_k\Lambda^2\frac{\widetilde{G}_2}{\sqrt{m}}, \quad \frac{\widetilde{G}_j}{m^{1-1/j}}a^{-1/j} \right\}. \tag{26}$$

Proof of Theorem A.1. Allocate the failure probability among the majority-core, coarse-center, sampling, and Gaussian-noise events. The stated sample condition absorbs the exponential majority-core failure and all FriendlyCore requirements. If $\sqrt{d}/(n\varepsilon) > 1$, the coarse estimate already has the displayed Gaussian privacy error and the theorem follows. Otherwise the radius floor ensures the independent center is within $R/2$ and makes the exponential bias negligible. Lemma A.6 with moment order j bounds the error by

$$\widetilde{O}_k\left[\frac{\widetilde{G}_2}{\sqrt{nm}} + aR + \frac{\widetilde{G}_j^j}{m^{j-1}R^{j-1}}\right].$$

Here $R\log(1/\beta)/n$ is absorbed by the privacy term up to logarithms because $\varepsilon \le 1$ and $d \ge 1$. Substituting (26) gives the common sampling and Gaussian terms plus $\widetilde{G}_j(a/m)^{1-1/j}$. Minimizing over $j \in \{2, k\}$ gives the minimum in (3). The second stage has sensitivity $2R/n_2$ conditional on c; Lemma A.2 and adaptive parallel composition across the two disjoint groups prove privacy. □

B Norm-moment mean estimation: lower bounds

Throughout this section, n is the number of users, m is the number of samples held by each user, and

$$N := nm$$

is the total number of samples. We use lower-case z for one sample and Z for the full user dataset; auxiliary item-level datasets are defined locally. We impose the convenient sufficient condition

$$\delta \le c\min\left\{\frac{1}{nm}, \frac{\varepsilon^2}{d}\right\} \tag{27}$$

for a sufficiently small absolute constant $c > 0$. This is an unoptimized common sufficient condition: the sparse-mixture ingredient only needs $\delta \lesssim 1/N$, and the Gaussian argument below specifies its own condition. The stronger shared choice makes every comparison uniform.

We use the centered norm-moment class $\mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$ from the introduction, with $\widetilde{G}_2 \le \widetilde{G}_k$.

Recall the rate in (3):

$$\begin{aligned}\mathcal{E}_{n,m}(\widetilde{G}_2, \widetilde{G}_k) = &\frac{\widetilde{G}_2}{\sqrt{nm}} + \frac{\widetilde{G}_2\sqrt{d}}{n\varepsilon\sqrt{m}} \\ &+ \min\left\{\widetilde{G}_2\left(\frac{\sqrt{d}}{nm\varepsilon}\right)^{1/2}, \quad \widetilde{G}_k\left(\frac{\sqrt{d}}{nm\varepsilon}\right)^{1-1/k}\right\}.\end{aligned} \tag{28}$$

Theorem B.1 (Restatement of Theorem 3.5)**.** *Let $\delta \le c\min\{1/(nm), \varepsilon^2/d\}$ for sufficiently small $c > 0$. For every user-level (ε, δ)-DP estimator, there exists $P \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$ such that*

$$\Pr\big(\|\widehat{\mu} - \mu_P\| \ge c_k\mathcal{E}_{n,m}(\widetilde{G}_2, \widetilde{G}_k)\big) \ge 1/3. \tag{29}$$

In particular the worst-case root-mean-square error is at least $c_k'\mathcal{E}_{n,m}(\widetilde{G}_2, \widetilde{G}_k)$. Thus Theorem A.1 and this lower bound match up to logarithms for every ratio $\widetilde{G}_k/\widetilde{G}_2$, throughout the upper theorem's user-count range. The lower bound itself has no user-count floor.

Separately, Esfandiari et al. (2022, Lemma 5.1) prove the following unbounded-location obstruction. If $\delta \le \varepsilon^2$ and $n \le \log(1/\delta)/(3\varepsilon)$, then for every finite $R > 0$ there is a point-mass distribution in the class such that

$$\Pr(\|\widehat{\mu} - \mu_P\| > R) > 1/3. \tag{30}$$

The proof has three ingredients. The classical nonprivate bound (Lemma B.5) supplies the sampling term. Lemma B.7 supplies the user-level Gaussian term by adapting Gaussian fingerprinting to whole user blocks. The randomized rare-outlier construction in Lemma B.4 supplies the heavy-tail term, using the bounded product-family lower bound of Kamath et al. (2019). Lemma B.2 transfers this last item-level bound to users. We combine the three bounds at the end of the section.

B.1 LOWER-BOUND INGREDIENTS

The first observation is the standard lower-bound reduction used in user-level privacy, including Levy et al. (2021): an algorithm that protects a user's entire block also protects a single item in that block.

Lemma B.2 (Grouping item-level samples into users)**.** *Let $\mathcal{A}$ be a user-level (ε, δ)-DP algorithm for n users with m samples per user. Given any dataset $Z = (z_1, \ldots, z_N)$ of $N = nm$ items, partition Z into n consecutive blocks of size m and run $\mathcal{A}$ on those blocks. The resulting algorithm is item-level (ε, δ)-DP.*

Consequently, an item-level lower bound for N samples applies verbatim to every user-level algorithm with n users and m samples per user after substituting $N = nm$.

Proof. Changing one item z_i changes at most one of the n blocks. The two blocked datasets are therefore neighboring in the user-level sense, so the privacy guarantee of $\mathcal{A}$ gives item-level (ε, δ)-DP. If the items are i.i.d. from P, the resulting blocks have exactly the distribution of n users each holding m i.i.d. samples from P. □

The rare-outlier construction needs a constant-probability lower bound for a bounded product family. We use the following consequence of the classical fingerprinting lower bound, with a standard constant-size repetition step to convert its mean-squared-error formulation.

Lemma B.3 (Consequence of Kamath et al. (2019, Lemma 53))**.** *Suppose $0 < \varepsilon \leq 1$,*

$$s \leq \frac{\sqrt{d}}{10^4 \varepsilon} \qquad \textit{and} \qquad \delta \leq \frac{1}{10^5 s}.$$

For every item-level (ε, δ)-DP estimator $\mathcal{B} : (\mathbb{R}^d)^s \to \mathbb{R}^d$, there is a product distribution Q supported on $\{\pm 1\}^d$, with $\mu_Q \in [-1/3, 1/3]^d$, such that

$$\Pr_{Z \sim Q^{\otimes s},\, \mathcal{B}} \left[\|\mathcal{B}(Z) - \mu_Q\|_2 \geq \frac{\sqrt{d}}{100} \right] \geq \frac{2}{5}. \tag{31}$$

Proof. We use Lemma 53 of Kamath et al. (2019): an $(\varepsilon, 3/(64S))$-DP estimator with mean-squared error at most $\alpha^2 \leq d/54$ over this product family requires $S \geq d/(72\alpha\varepsilon)$. Only the conversion from constant success probability to mean-squared error is needed here.

Suppose every distribution in the family admitted error below $\rho = \sqrt{d}/100$ with probability greater than $3/5$. Take $S = 1000s$ fresh samples and run the estimator independently on 1000 disjoint blocks. Project each output onto the known mean cube $[-1/3, 1/3]^d$; this preserves privacy and cannot increase its error. By Hoeffding's inequality, more than half of the outputs are within ρ of the mean, except with probability e^{-20}.

Apply the standard geometric majority rule (Asi et al., 2024, Fact 2): return the first output with more than 500 outputs within 2ρ, or zero if none qualifies. On the majority event, a qualifying point exists and is within 3ρ of the mean, since its neighborhood intersects the good majority. Denote the result by $\widehat{\mu}_S$. It already belongs to the cube, whose diameter is $2\sqrt{d}/3$. Thus

$$\mathbb{E}\|\widehat{\mu}_S - \mu_Q\|^2 \leq 9\rho^2 + \frac{4d}{9} e^{-20} < \frac{d}{1000}.$$

Parallel composition preserves (ε, δ)-DP, and $\delta \leq 1/(10^5 s) < 3/(64S)$. Applying the cited lemma with $\alpha^2 = d/1000$ gives

$$S \geq \frac{\sqrt{1000}}{72} \frac{\sqrt{d}}{\varepsilon} > \frac{1}{10} \frac{\sqrt{d}}{\varepsilon} \geq 1000s = S,$$

a contradiction. This proves the constant-probability formulation without reproving the fingerprinting bound. □

We now place rare, scaled draws from that bounded family among zero observations. This follows the rare-outlier approach of Barber & Duchi (2014); Lowy & Razaviyayn (2023), but a nonzero observation is itself random rather than revealing the unknown mean.

Lemma B.4 (Sparse-mixture norm-moment item-level DP lower bound)**.** *Fix $k > 2$ and $\widetilde{G}_2, \widetilde{G}_k > 0$, and let*

$$q_N := \min\left\{1, \frac{\sqrt{d}}{N\varepsilon}\right\}.$$

Suppose $0 < \varepsilon \le 1$, $\delta \le 1/(10^5 N)$, and $\sqrt{d}/\varepsilon$ is larger than a sufficiently large absolute constant. Then, for every item-level (ε, δ)-DP estimator $\mathcal{A} : (\mathbb{R}^d)^N \to \mathbb{R}^d$, there exists $P \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$ such that

$$\Pr_{Z\sim P^{\otimes N},\,\mathcal{A}}\left[\|\mathcal{A}(Z) - \mu_P\|_2 \gtrsim_k \min\left\{\widetilde{G}_2 q_N^{1/2}, \widetilde{G}_k q_N^{1-1/k}\right\}\right] \ge \frac{1}{3}. \tag{32}$$

Proof. Fix a sufficiently small absolute constant $c_0 > 0$, write $p := c_0 q_N$, and set

$$a := \frac{1}{2\sqrt{d}} \min\left\{\widetilde{G}_2 p^{-1/2}, \widetilde{G}_k p^{-1/k}\right\}. \tag{33}$$

For every product distribution Q appearing in Lemma B.3, define P_Q by

$$z \sim P_Q \quad \Longleftrightarrow \quad z = \begin{cases} 0, & \text{with probability } 1-p, \\ a\widetilde{z}, \quad \widetilde{z} \sim Q, & \text{with probability } p. \end{cases}$$

Then

$$\mu_{P_Q} = pa\mu_Q. \tag{34}$$

We first verify the moment conditions. Since $\|\widetilde{z}\|_2 = \sqrt{d}$ almost surely, for $j \in \{2, k\}$,

$$\begin{aligned} \mathbb{E}_{z\sim P_Q}\|z - \mu_{P_Q}\|_2^j &\le 2^{j-1}\left(\mathbb{E}\|z\|_2^j + \|\mu_{P_Q}\|_2^j\right) \\ &\le 2^{j-1}\left(pa^j d^{j/2} + p^j a^j d^{j/2}\right) \\ &\le 2^j pa^j d^{j/2}. \end{aligned}$$

By (33), this is at most $\widetilde{G}_2^2$ when $j = 2$ and at most $\widetilde{G}_k^k$ when $j = k$. Therefore $P_Q \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$.

It remains to prove the privacy lower bound. Suppose, toward a contradiction, that $\mathcal{A}$ estimates the mean of every P_Q with error less than $cpa\sqrt{d}$ and failure probability less than $1/3$, where $c > 0$ will be chosen sufficiently small.

Let

$$s := \lceil 100Np \rceil.$$

We construct an estimator $\mathcal{B}$ for the bounded product family. Given an auxiliary dataset

$$Z = (z_1, \dots, z_s) \sim Q^{\otimes s},$$

draw independent indicators $B_1, \dots, B_N \sim \mathrm{Bernoulli}(p)$ and let $K := \sum_{i=1}^N B_i$. If $K > s$, output zero. Otherwise, put $az_1, \dots, az_K$ into the K locations for which $B_i = 1$, in their natural order, and put zero in every other location. Denote the resulting dataset by $\widetilde{Z}$, and output

$$\mathcal{B}(Z) := \frac{\mathcal{A}(\widetilde{Z})}{pa}. \tag{35}$$

This is the padding step. Conditional on the indicators, each auxiliary point z_j is used either once or not at all. Consequently, changing one point of Z changes at most one point of $\widetilde{Z}$. Thus $\mathcal{B}$ is item-level (ε, δ)-DP by the privacy of $\mathcal{A}$ and post-processing.

To check accuracy, couple this construction to N independent draws from P_Q. Before the possible abort, the indicators choose the nonzero locations independently with probability p, and the values placed in those locations are independent draws from aQ. Consequently, whenever $K \le s$, $\widetilde{Z}$ is exactly the dataset obtained from this coupling. Moreover,

$$\mathbb{E}K = Np, \qquad \Pr(K > s) \le \frac{1}{100}$$

by Markov's inequality. Hence the abort changes the success probability by at most $1/100$.

Using (34), an error of at most $cpa\sqrt{d}$ for estimating μ_{P_Q} becomes an error of at most $c\sqrt{d}$ for estimating μ_Q after the division in (35). Therefore $\mathcal{B}$ has failure probability at most

$$\frac{1}{3} + \frac{1}{100} < \frac{2}{5}.$$

Finally,

$$s \le 100 c_0 \min\left\{N, \frac{\sqrt{d}}{\varepsilon}\right\} + 1.$$

For example, take $c_0 = 1/(2 \cdot 10^6)$ and assume $\sqrt{d}/\varepsilon \ge 2 \cdot 10^4$. Then $s \le \lceil N/20000 \rceil \le N$ and $s \le \sqrt{d}/(20000\varepsilon) + 1 \le \sqrt{d}/(10^4\varepsilon)$. Thus integer rounding causes no exception. Since $s \le N$, the assumption $\delta \le 1/(10^5 N)$ also gives $\delta \le 1/(10^5 s)$. Thus Lemma B.3 applies. Taking $c < 1/100$ contradicts that lemma.

It remains only to compute the separation. By (33),

$$pa\sqrt{d} = \frac{1}{2} \min\left\{\widetilde{G}_2 p^{1/2}, \widetilde{G}_k p^{1-1/k}\right\},$$

which proves (32) after adjusting the k-dependent constant, because $p = c_0 q_N$. □

The sampling term does not involve privacy. We record the usual Gaussian-location specialization of Le Cam's two-point lower bound, rather than reprove the testing argument.

Lemma B.5 (Classical nonprivate bound; Duchi, 2023, Section 8.3)**.** *Fix $k > 2$ and $\widetilde{G}_2 \le \widetilde{G}_k$. For every estimator $\mathcal{A}$ based on N i.i.d. samples, whether private or not, there is a Gaussian $P \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$ such that*

$$\Pr_{Z \sim P^{\otimes N},\, \mathcal{A}}\left[\|\mathcal{A}(Z) - \mu_P\|_2 \gtrsim_k \frac{\widetilde{G}_2}{\sqrt{N}}\right] \ge \frac{1}{3}.$$

In particular, the root-mean-square error is $\gtrsim_k \widetilde{G}_2/\sqrt{N}$.

This is the one-dimensional Gaussian specialization of Duchi (2023, Proposition 8.2.1 and Eq. (8.3.2)), embedded in one coordinate and scaled so that its kth centered moment is at most $\widetilde{G}_2^k \le \widetilde{G}_k^k$. The scaling changes only the k-dependent constant.

For the privacy term, the mean may be anywhere in $\mathbb{R}^d$. The next lemma makes explicit the whole-user and arbitrary-location form of the Gaussian fingerprinting argument of Kamath et al. (2019); Agarwal et al. (2025); we give the adaptation because a lower bound restricted to small mean error would not cover all parameter regimes used here.

Lemma B.6 (Gaussian lower bound with unbounded mean location)**.** *Let all nm samples be independent, with each of n users holding m samples from $\mathcal{N}(\theta, \sigma^2 I_d)$, where $\sigma > 0$ is known and $\theta \in \mathbb{R}^d$ is unrestricted. Suppose $0 < \varepsilon \le 1$, $0 < \delta < 1/2$, and*

$$\delta\big(\sqrt{d} + \sqrt{\log(e/\delta)}\big) \le c\varepsilon. \tag{36}$$

For every (ε, δ)-user-level DP estimator $\widehat{\theta}$, there is a $\theta \in \mathbb{R}^d$ such that

$$\Pr_{Z \sim (\mathcal{N}(\theta, \sigma^2 I_d)^{\otimes m})^{\otimes n},\, \widehat{\theta}}\left(\|\widehat{\theta}(Z) - \theta\| \ge c' \frac{\sigma d}{n\varepsilon\sqrt{m}}\right) \ge \frac{1}{3}. \tag{37}$$

The worst-case root-mean-square error has the same lower bound, up to a universal constant. No restriction on the size of this error relative to $\sigma\sqrt{d}$ is imposed.

Proof. We follow the arbitrary-range Gaussian fingerprinting argument of Kamath et al. (2019, Lemma 58); see also the user-block calculation in Agarwal et al. (2025, Appendix D of the full version). All expectations below include the prior, the samples, and the algorithm's randomness.

Step 1: accuracy forces correlation. Suppose the root-mean-square error is at most α uniformly over the cube $[-a, a]^d$, and project the output onto this cube. Give θ the uniform prior on that cube. Let $\widehat{\theta}$ denote the projected output, let Y_i be user i's sample mean, and put

$$g_i := \frac{\sqrt{m}}{\sigma}(Y_i - \theta), \qquad V_i := \sum_{j=1}^{d}(a^2 - \theta_j^2)(\widehat{\theta}_j - \theta_j)g_{i,j}.$$

Conditional on θ, the g_i are independent standard Gaussian vectors. Write $h_j(\theta) := \mathbb{E}[\widehat{\theta}_j \mid \theta]$. The derivative of the log density of the full dataset with respect to θ_j is $(\sqrt{m}/\sigma)\sum_i g_{i,j}$, so

$$\partial_j h_j(\theta) = \frac{\sqrt{m}}{\sigma}\sum_i \mathbb{E}[(\widehat{\theta}_j - \theta_j)g_{i,j} \mid \theta].$$

Subtracting θ_j does not change this expectation because $\mathbb{E}[g_{i,j} \mid \theta] = 0$. The output is bounded, so differentiation under the Gaussian integral is justified. Integration by parts in θ_j gives

$$\mathbb{E}[(a^2 - \theta_j^2)\partial_j h_j(\theta)] = 2\mathbb{E}[\theta_j h_j(\theta)],$$

with zero boundary term because $a^2 - \theta_j^2$ vanishes at the endpoints. Summing over coordinates and applying Cauchy–Schwarz gives

$$\frac{\sqrt{m}}{\sigma}\sum_i \mathbb{E}V_i = 2\mathbb{E}\langle\theta, \widehat{\theta}\rangle \geq 2\left(\frac{da^2}{3} - a\sqrt{d/3}\,\alpha\right). \tag{38}$$

In particular, if $a \geq 4\alpha/\sqrt{d}$, the right-hand side is at least a universal positive constant times da^2.

Step 2: privacy limits each user's correlation. Replace user i by an independent fresh user and denote the new output by $\widehat{\theta}^{(i)}$. Define V_i' by using this output but the original g_i in the preceding display. Conditional on θ and the replaced dataset, V_i' is a centered Gaussian, since the original g_i is independent of that output. Therefore

$$\mathbb{E}V_i' = 0, \qquad \mathbb{E}|V_i'| \leq a^2\mathbb{E}\|\widehat{\theta}^{(i)} - \theta\| \leq a^2\alpha.$$

To use approximate DP without an unbounded test, set

$$u = \sqrt{d} + \sqrt{2\log(1/\delta)}, \qquad B = 2a^3\sqrt{d}\,u, \quad [x]_{-B}^{B} := \max\{-B, \min\{x, B\}\}.$$

Let $\phi_+(v) := \min\{\max\{v, 0\}, B\}$ and $\phi_-(v) := \min\{\max\{-v, 0\}, B\}$. Fixing θ, the original dataset, and its replacement makes these fixed tests of the output. The datasets differ in one user. Apply DP to ϕ_+ and reverse DP to ϕ_-, then average:

$$\mathbb{E}\phi_+(V_i) \leq e^{\varepsilon}\mathbb{E}\phi_+(V_i') + \delta B, \qquad \mathbb{E}\phi_-(V_i) \geq e^{-\varepsilon}\mathbb{E}\phi_-(V_i') - e^{-\varepsilon}\delta B.$$

Subtract the second inequality from the first. The symmetry of V_i' implies that its clipped positive and negative parts have equal expectations. Consequently

$$\mathbb{E}[V_i]_{-B}^{B} \leq (e^{\varepsilon} - e^{-\varepsilon})\mathbb{E}([V_i']_{-B}^{B})_+ + (1 + e^{-\varepsilon})\delta B \leq C\varepsilon a^2\alpha + 2\delta B.$$

Pointwise, $|V_i| \leq 2a^3\sqrt{d}\,\|g_i\|$, even though the output depends on g_i. Put $t_0 = \sqrt{2\log(1/\delta)}$. The Gaussian norm tail $\Pr(\|g_i\| > \sqrt{d} + t) \leq e^{-t^2/2}$ implies

$$\mathbb{E}[\|g_i\|\mathbf{1}_{\{\|g_i\| > u\}}] \leq u\delta + \int_{t_0}^{\infty} e^{-t^2/2}\,dt \leq \delta(u + 1/t_0).$$

Consequently

$$\mathbb{E}|V_i|\mathbf{1}_{\{|V_i| > B\}} \leq C\delta a^3\sqrt{d}(\sqrt{d} + \sqrt{\log(e/\delta)}).$$

Step 3: choose the mean range to match the target error. Combining with (38), and taking $a = 4\alpha/\sqrt{d}$, yields

$$d \leq C\frac{n\sqrt{m}}{\sigma}\alpha\left[\varepsilon + \delta(\sqrt{d} + \sqrt{\log(e/\delta)})\right].$$

Under (36) this proves the root-mean-square lower bound. The cube grows with α while σ stays fixed, so this bound does not saturate. The same calculation holds if $a\sqrt{d}/\alpha$ is any fixed constant at least four.

Step 4: the median trick gives a probability lower bound. A mean-squared-error lower bound alone does not give a probability lower bound. We argue by contradiction: amplify a hypothetical constant-success estimator and project it to bound the squared loss. Only a constant number of repetitions is needed because the cube's diameter is a constant multiple of the target accuracy. Suppose an estimator achieved error e_0 with probability greater than $2/3$ uniformly over all means. Restrict to the cube with $a = 16e_0/\sqrt{d}$, project each output onto it, and run the estimator on $r = 1000$ independent groups of n users. Use the majority-neighborhood rule from Lemma B.3: select an output with more than $r/2$ outputs within distance $2e_0$, if one exists, and otherwise return zero. Except with probability $e^{-r/18}$, the aggregate has error at most $3e_0$. Every candidate, the selected output, and the zero fallback already lie in the cube, so its squared error is always at most $4a^2d$. Explicitly, the mean-squared error is at most

$$9e_0^2 + 4a^2 d\, e^{-r/18} = (9 + 1024e^{-1000/18})e_0^2 < 10e_0^2.$$

Parallel composition preserves (ε, δ)-DP. Apply the calculation above with rn users, $\alpha = \sqrt{10}e_0$, and $a = 16e_0/\sqrt{d}$ to obtain $e_0 \geq c\sigma d/(rn\varepsilon\sqrt{m})$. Since r is an absolute constant, this proves the claimed probability lower bound by contraposition. Condition (36) is independent of the number of repetitions. □

Lemma B.7 (User-level Gaussian norm-moment lower bound)**.** *Under the hypotheses of Theorem B.1, every user-level private estimator has, for some Gaussian $P \in \mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$,*

$$\Pr\left(\|\widehat{\mu} - \mu_P\| \geq c_k \frac{\widetilde{G}_2\sqrt{d}}{n\varepsilon\sqrt{m}}\right) \geq 1/3. \tag{39}$$

In particular its mean-squared error is at least

$$c_k' \frac{\widetilde{G}_2^2 d}{n^2 m\varepsilon^2}. \tag{40}$$

Proof. Let $K_k = (\mathbb{E}|g|^k)^{1/k} = [2^{k/2}\Gamma((k+1)/2)/\sqrt{\pi}]^{1/k}$ for $g \sim \mathcal{N}(0,1)$, and set $\sigma = \widetilde{G}_2/(K_k\sqrt{d})$. For a standard Gaussian vector g, Minkowski's inequality in $L_{k/2}$ gives

$$(\mathbb{E}\|g\|^k)^{2/k} = \left\|\sum_{j=1}^d g_j^2\right\|_{k/2} \leq \sum_{j=1}^d \|g_j^2\|_{k/2} = dK_k^2.$$

Hence every $\mathcal{N}(\theta, \sigma^2 I_d)$, with arbitrary $\theta \in \mathbb{R}^d$, has centered second moment at most $\widetilde{G}_2^2$ and centered kth moment at most $\widetilde{G}_2^k \leq \widetilde{G}_k^k$. Condition (27) implies (36): indeed, $\delta\sqrt{d} \leq c\varepsilon$ and $\delta\sqrt{\log(e/\delta)} \leq \sqrt{\delta} \leq \sqrt{c}\varepsilon$. Taking the shared constant sufficiently small, apply Lemma B.6 and substitute σ. □

Lemma B.8 (Restatement of Esfandiari et al. (2022, Lemma 5.1))**.** *Suppose $0 < \varepsilon \leq 1$, $0 < \delta < 1/2$, $\delta \leq \varepsilon^2$, and*

$$n \leq \frac{\log(1/\delta)}{3\varepsilon}.$$

For every user-level (ε, δ)-DP mean estimator based on n users with m samples each, and every finite $R > 0$, there exists a point-mass distribution P such that

$$\Pr(\|\widehat{\mu} - \mu_P\| > R) > \frac{1}{3}.$$

Consequently, no finite error bound holds uniformly over $\mathcal{P}_k(\widetilde{G}_2, \widetilde{G}_k)$, since every point mass belongs to this class.

B.2 Proof of the mean-estimation lower bound

Proof of Theorem B.1. The nonprivate lemma with $N = nm$ gives $\widetilde{G}_2/\sqrt{nm}$, and Lemma B.7 gives the uncapped Gaussian term. For the heavy-tail term put $x = \sqrt{d}/(nm\varepsilon)$. If $x \le 1$ and $\sqrt{d}/\varepsilon$ is sufficiently large, grouping followed by Lemma B.4 gives $\min\{\widetilde{G}_2 x^{1/2}, \widetilde{G}_k x^{1-1/k}\}$. If $\sqrt{d}/\varepsilon = O(1)$, this minimum is at most a constant multiple of $\widetilde{G}_2/\sqrt{nm}$, so the nonprivate bound suffices. If $x > 1$, then $\sqrt{d}/(n\varepsilon) = mx$ and

$$\min\{\widetilde{G}_2\sqrt{x}, \widetilde{G}_k x^{1-1/k}\} \le \widetilde{G}_2\sqrt{x} \le \widetilde{G}_2\sqrt{m}\,x = \frac{\widetilde{G}_2\sqrt{d}}{n\varepsilon\sqrt{m}}.$$

Thus the uncapped Gaussian lower bound covers this entire remaining regime. Select the hard distribution for the largest of the three numerical terms; its lower bound is at least a constant times their sum. Every ingredient holds with probability at least $1/3$, which also implies the root-mean-square claim. Lemma B.8 proves the unbounded-location obstruction, independently of the finite-error lower bound. □

C Population and empirical regularized minimizers

Let $\mathcal{W} \subset \mathbb{R}^d$ be nonempty, compact, and convex, with diameter $D > 0$, and define $F(w) = \mathbb{E}_{z\sim P} f(w, z)$. The following restates the SCO assumptions and ERM target from Section 4. Assume sample losses are convex and differentiable on a neighborhood of $\mathcal{W}$, gradients are measurable, and F is finite at one point. For the appendix proofs, define the sample Lipschitz constant

$$\mathcal{L}(z) := \sup_{w\in\mathcal{W}} \|\nabla f(w, z)\|. \tag{41}$$

Then the main-body assumption (2) reads

$$\mathbb{E}_{z\sim P}\mathcal{L}(z)^2 \le G_2^2, \qquad \mathbb{E}_{z\sim P}\mathcal{L}(z)^k \le G_k^k, \qquad 0 < G_2 \le G_k. \tag{42}$$

For a fresh phase dataset Z of s users, a center $w_0 \in \mathcal{W}$, and $\lambda > 0$, define

$$\widehat{F}_Z(w) = \frac{1}{sm}\sum_{u=1}^{s}\sum_{j=1}^{m} f(w, z_{u,j}),$$

$$\widehat{F}_{\lambda,Z}(w) = \widehat{F}_Z(w) + \frac{\lambda}{2}\|w - w_0\|^2, \qquad \widehat{w}_{\lambda,Z} = \arg\min_{w\in\mathcal{W}} \widehat{F}_{\lambda,Z}(w).$$

The logarithmic factor Λ is the one defined in the preliminaries, using the full user count n, not the phase size s. For any phase size $s \le n$, $\mathcal{E}_{s,m}(G_2, G_k)$ is the rate in (3) with n replaced by s and raw moment bounds G_2, G_k. The constant C_k is fixed sufficiently large in the proofs below. The target (12) is

$$\Pr\left(\|A_{\mathrm{ERM}}(Z; w_0, \lambda) - \widehat{w}_{\lambda,Z}\| \le C_k\Lambda^4\mathcal{E}_{s,m}(G_2, G_k)/\lambda\right) \ge 0.8. \tag{43}$$

The probability is over the fresh i.i.d. phase dataset and the algorithm. It holds uniformly in the incoming center; the center may depend on earlier disjoint phases. The solver may be taken to return a point in $\mathcal{W}$: projecting its output onto $\mathcal{W}$ preserves privacy and cannot increase its distance to $\widehat{w}_{\lambda,Z}$.

To amplify the ERM solver, independent runs must approximate the same target. Their empirical minimizers differ because they use fresh data; the following elementary strong-convexity argument compares each one with the population minimizer. We use it in the localization proof in Appendix D.

Lemma C.1 (Distance between empirical and population minimizers)**.** *Fix w_0 independently of the fresh block Z. Let w_λ^{pop} minimize $F(w) + \lambda\|w - w_0\|^2/2$ on $\mathcal{W}$. Then $\mathbb{E}\|\widehat{w}_{\lambda,Z} - w_\lambda^{\mathrm{pop}}\|^2 \le G_2^2/(\lambda^2 sm)$.*

Proof. Strong monotonicity and the first-order optimality inequalities at the two constrained minimizers give, pointwise in Z,

$$\lambda\|\widehat{w}_{\lambda,Z} - w_\lambda^{\mathrm{pop}}\| \le \|\nabla\widehat{F}_Z(w_\lambda^{\mathrm{pop}}) - \nabla F(w_\lambda^{\mathrm{pop}})\|.$$

The query w_λ^{pop} is independent of the fresh phase dataset, so

$$\mathbb{E}\|\widehat{w}_{\lambda,Z} - w_\lambda^{\mathrm{pop}}\|^2 \le \frac{G_2^2}{\lambda^2 sm}. \tag{44}$$

□

By Markov's inequality, with probability at least 0.99 this distance is at most $10G_2/(\lambda\sqrt{sm})$. Combining with (43) makes every independent repetition accurate around the same population minimizer with probability at least 0.79.

D Population localization

Theorem D.1 (Restatement of Theorem 4.1; adapted from Asi et al., 2024; Lowy, 2026)**.** *Assume* (42). *Let A_{ERM} be (ε,δ)-user-level DP and suppose that for every center $w_0 \in \mathcal{W}$ and each regularization parameter used by Algorithm 3, on every fresh s-user block it satisfies* (12). *If $n \ge C_k\Lambda^6/\varepsilon$, then Algorithm 3 is (ε,δ)-user-level DP and, with probability at least $1-\beta$,*

$$F(\widehat{w}) - F^* \le D\min\{G_2,\ C_k\Lambda^8\mathcal{E}_{n,m}(G_2,G_k)\}. \tag{45}$$

Theorem D.1 is the user-level approximate-DP adaptation of the population-localization reduction of Asi et al. (2024), presented as Theorem 2.1 of Lowy (2026). We use the same population-target, repetition, and telescoping argument. Here each user is kept intact and the calls use disjoint user blocks; the equal block sizes in Algorithm 3 affect only logarithmic factors.

Lemma C.1 first puts independent ERM calls near a common population minimizer. The geometric majority rule of Asi et al. (2024, Fact 2) amplifies this to high probability. The increasing regularization then makes the population-risk comparisons telescope. For completeness, we give a self-contained proof with the present block sizes and privacy parameters.

Algorithm 3 Population localization meta-algorithm

Input: Data Z from n users with m observations each; moment bounds G_2, G_k; privacy parameters ε,δ; failure probability β; private regularized ERM solver A_{ERM}. Assume $n \ge C_k\Lambda^6/\varepsilon$.

1. Set $I = 1 + \lceil \log_{32}\sqrt{nm}\rceil$, $J = \lceil 200\log(2I/\beta)\rceil$, and $s = \lfloor n/(IJ)\rfloor$. Initialize $w_0 \in \mathcal{W}$.
2. Partition the users into IJ disjoint datasets $Z^{i,j}$, each containing s users and all m observations of each user. Discard unused users.
3. Set $\Delta = 6C_k\Lambda^4\mathcal{E}_{s,m}(G_2,G_k)$ and $\lambda_1 = \Delta/D$. If $\Delta \ge G_2$, return w_0.
4. For $i = 1,\dots,I$:
   (a) Set $\lambda_i = 32^{i-1}\lambda_1$ and $\rho_i = \Delta/(3\lambda_i)$.
   (b) For each $j \in [J]$, independently compute
   $$v_{i,j} \leftarrow A_{\mathrm{ERM}}(Z^{i,j}; w_{i-1}, \lambda_i, \varepsilon, \delta).$$
   (c) *Geometric aggregation* (Asi et al., 2024, Fact 2): Set w_i to the first $v_{i,j}$ with more than $J/2$ candidates within distance $2\rho_i$. If no such candidate exists, set $w_i = w_0$.
5. Return w_I.

Proof of Theorem D.1. Use I, J, s, Δ and the 32-fold regularization schedule of Algorithm 3. Since $I, J = O(\Lambda)$ and $n \ge C_k\Lambda^6/\varepsilon$, we have $n \ge 2IJ$ and $s \ge n/(2IJ) \ge 1$. Each exponent of $1/s$ in $\mathcal{E}_{s,m}$ is at most one, so

$$\mathcal{E}_{s,m}(G_2,G_k) \le \frac{n}{s}\mathcal{E}_{n,m}(G_2,G_k), \qquad \Delta \le C_k\Lambda^6\mathcal{E}_{n,m}(G_2,G_k). \tag{46}$$

If $\Delta \ge G_2$, the fixed output w_0 has risk at most $DG_2 \le D\Delta$, proving the claimed bound in that branch. We therefore consider $\Delta < G_2$.

Amplifying the ERM guarantee. At phase i, the J fresh user blocks are independent of the incoming center w_{i-1}. Each solver call uses this center and λ_i. Let

$$w_i^* = \arg\min_{w \in \mathcal{W}} \{F(w) + \lambda_i \|w - w_{i-1}\|^2/2\}.$$

Conditional on the preceding phases, each candidate is within ρ_i of the same w_i^* with probability at least 0.79, by Lemma C.1 and (12), after fixing the constant in C_k. The geometric aggregation in Algorithm 3 therefore returns a point satisfying

$$\|w_i - w_i^*\| \le 3\rho_i = \Delta/\lambda_i \quad (1 \le i \le I) \tag{47}$$

whenever a majority of candidates in every phase are good. Hoeffding's inequality and a union bound over the phases make this event hold with probability at least $1 - \beta$.

From the phase invariant to population risk. Let w^* minimize F. The variational comparison for w_1^* gives $F(w_1^*) - F(w^*) \le \lambda_1 D^2/2$. For $i \ge 2$, optimality of w_i^* and (47) give

$$F(w_i^*) - F(w_{i-1}^*) \le \frac{\lambda_i}{2}\|w_{i-1}^* - w_{i-1}\|^2 \le \frac{16\Delta^2}{\lambda_{i-1}}.$$

The factor 16 is exactly the contribution of the 32-fold schedule $\lambda_i = 32\lambda_{i-1}$. Since F is G_2-Lipschitz, summing yields

$$\begin{aligned} F(w_I) - F(w^*) &\le \frac{G_2\Delta}{\lambda_I} + \frac{\lambda_1 D^2}{2} + 16\sum_{i=2}^{I}\frac{\Delta^2}{\lambda_{i-1}} \\ &\le \frac{DG_2}{\sqrt{nm}} + \left(\frac{1}{2} + 16\sum_{r\ge 0} 32^{-r}\right) D\Delta \le \frac{DG_2}{\sqrt{nm}} + 18D\Delta. \end{aligned}$$

All outputs are in $\mathcal{W}$, so their risk is also at most DG_2. Combine the last display with (46) and $G_2/\sqrt{nm} \le \mathcal{E}_{n,m}(G_2, G_k)$ to obtain the stated $D\min\{G_2, C_k\Lambda^8\mathcal{E}_{n,m}(G_2, G_k)\}$ bound.

Privacy. Finally, each user occurs in exactly one block. Conditional on preceding outputs, its block is processed by one (ε, δ)-DP ERM call; all later computation involving that user's influence is post-processing of that call together with disjoint fresh user blocks. Adaptive parallel composition therefore makes Algorithm 3 (ε, δ)-user-level DP. □

E FUSED ADAPTIVE ORACLE AND PRIVATE ERM: COMPLETE PROOFS

Theorem E.1 (Restatement of Theorem 4.5)**.** *Assume* (42) *and* $n \ge C_k\Lambda^6/\varepsilon$. *Algorithm 3, instantiated with Algorithm 4, is* (ε, δ)*-user-level DP and, with probability at least* $1 - \beta$*, satisfies*

$$F(\widehat{w}) - F^* \le D\min\big\{G_2,\ C_k\Lambda^8\mathcal{E}_{n,m}(G_2, G_k)\big\}. \tag{48}$$

We prove Theorem E.1 by supplying the private ERM solver required by Theorem D.1. Lemma E.2 reduces its utility to noise variance and average squared empirical bias. Lemma E.3 controls privacy and noise. Lemmas E.4–E.9 control clipping bias at the adaptive iterates. Proposition E.10 chooses the parameters and proves the ERM guarantee; the final proof applies population localization.

E.1 A GENERIC NOISY BIASED-GRADIENT LEMMA

We use the usual projected-SGD recursion from the biased-gradient analyses of Lowy & Razaviyayn (2023); Asi et al. (2024). We retain its dependence on the average squared bias, rather than replace each bias by a worst-case bound. The proof is included to justify this precise version and its use with the conditional bias in the main body.

Lemma E.2 (Restatement of Lemma 4.2)**.** *Let* $(\mathcal{F}_t)$ *be a filtration such that* Z, w_0, w_t *are* $\mathcal{F}_t$*-measurable. For gradient estimates* g_t *and updates*

$$w_{t+1} = \Pi_{\mathcal{W}}(w_t - \eta_t[g_t + \lambda(w_t - w_0)])\,,$$

define the bias $b_t := \mathbb{E}[g_t \mid \mathcal{F}_t] - \nabla \widehat{F}_Z(w_t)$ *and noise* $v_t := g_t - \mathbb{E}[g_t \mid \mathcal{F}_t]$. *Assume* $\mathbb{E}\|\nabla \widehat{F}_Z(w_t)\|^2 \le G_2^2$ *and* $\mathbb{E}[\|v_t\|^2 \mid \mathcal{F}_t] \le \Sigma^2$ *for every* t. *Let* $B^2 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|b_t\|^2$, $\eta_t = 4/(\lambda(t+1))$, *and* $\bar{w} := 2\sum_{t=1}^{T} t w_t/[T(T+1)]$. *Then, with probability at least* 0.9,

$$\|\bar{w} - \widehat{w}_{\lambda,Z}\| \lesssim \frac{1}{\lambda}\left(B + \frac{G_2 + \Sigma + \lambda D}{\sqrt{T}}\right). \tag{49}$$

Proof. By definition, b_t is $\mathcal{F}_t$-measurable and $\mathbb{E}[v_t \mid \mathcal{F}_t] = 0$. Nonexpansiveness of projection, strong convexity of $\widehat{F}_{\lambda,Z}$, and Young's inequality give

$$\begin{aligned}\widehat{F}_{\lambda,Z}(w_t) - \widehat{F}_{\lambda,Z}(\widehat{w}_{\lambda,Z}) \le& \left(\frac{1}{2\eta_t} - \frac{\lambda}{4}\right)\|w_t - \widehat{w}_{\lambda,Z}\|^2 - \frac{\|w_{t+1} - \widehat{w}_{\lambda,Z}\|^2}{2\eta_t} \\ &+ \frac{\|b_t\|^2}{\lambda} - \langle v_t, w_t - \widehat{w}_{\lambda,Z}\rangle + \frac{\eta_t}{2}\|g_t + \lambda(w_t - w_0)\|^2. \end{aligned} \tag{50}$$

Both w_t and the empirical minimizer are $\mathcal{F}_t$-measurable, so the inner product with v_t has expectation zero. Also, using the gradient-moment assumption and $\|w_t - w_0\| \le D$,

$$\mathbb{E}\|g_t + \lambda(w_t - w_0)\|^2 \lesssim G_2^2 + \lambda^2 D^2 + \mathbb{E}\|b_t\|^2 + \Sigma^2. \tag{51}$$

Multiply (50) by t, take expectations, and sum. For $\eta_t = 4/(\lambda(t+1))$, the squared-distance terms telescope: the positive coefficient is $\lambda t(t-1)/8$ and the preceding negative coefficient is its opposite. Dropping the final nonpositive term and using (51) gives

$$\begin{aligned}\sum_{t=1}^{T} t\,\mathbb{E}[\widehat{F}_{\lambda,Z}(w_t) - \widehat{F}_{\lambda,Z}(\widehat{w}_{\lambda,Z})] \lesssim& \frac{T}{\lambda}\sum_{t=1}^{T}\mathbb{E}\|b_t\|^2 \\ &+ \frac{T}{\lambda}(G_2^2 + \Sigma^2 + \lambda^2 D^2). \end{aligned} \tag{52}$$

Convexity and the weights defining $\bar{w}$ now imply

$$\mathbb{E}[\widehat{F}_{\lambda,Z}(\bar{w}) - \widehat{F}_{\lambda,Z}(\widehat{w}_{\lambda,Z})] \lesssim \frac{B^2}{\lambda} + \frac{G_2^2 + \Sigma^2}{\lambda T} + \frac{\lambda D^2}{T}. \tag{53}$$

Strong convexity bounds the expected squared distance by $2/\lambda$ times (53). Markov's inequality gives (49) with probability at least 0.9. □

E.2 Privacy and empirical-bias bounds for Algorithm 4

Algorithm 4 User-level DP SGD with the fused adaptive gradient estimator

Procedure: $\mathsf{FusedERM}(Z; w_0, \lambda, G_2, G_k, \varepsilon, \delta, \Lambda)$, where Z contains s users with m observations each.

1. Choose T, R, δ_0 as in Proposition E.10. Set $\tau = 2\sqrt{160}e^{\varepsilon/2}G_2/\sqrt{m}$, $L_0 = \log(16T/\delta_0)$, and $\sigma = C(R + \tau L_0)\sqrt{T}L_0/(s\varepsilon)$.
2. Set $w_1 = w_0$ and $\mathsf{aborted} = \mathsf{false}$. Draw one noisy threshold $\widehat{\theta} = 4s/5 + \mathrm{Lap}(8/\varepsilon)$.
3. For $t = 1, \dots, T$:
   (a) If $\mathsf{aborted} = \mathsf{true}$, set $g_t = 0$ and go to step (f).
   (b) Compute $\nabla \widehat{F}_{Z_i}(w_t) = m^{-1}\sum_a \nabla f(w_t, z_{i,a})$ for each user and
   $$\widehat{s}_t = \frac{1}{s}\sum_{i,j=1}^{s} \mathbf{1}\{\|\nabla \widehat{F}_{Z_i}(w_t) - \nabla \widehat{F}_{Z_j}(w_t)\| \le \tau\} + \mathrm{Lap}(16/\varepsilon).$$
   If $\widehat{s}_t < \widehat{\theta}$, set $\mathsf{aborted} = \mathsf{true}$, $g_t = 0$, and go to step (f).
   (c) For each $i \in [s]$, compute $h_{t,i} = \sum_{j=1}^{s} \mathbf{1}\{\|\nabla \widehat{F}_{Z_i}(w_t) - \nabla \widehat{F}_{Z_j}(w_t)\| \le 2\tau\}$ and
   $$p_{t,i} = \begin{cases} 0, & h_{t,i} < s/2, \\ (h_{t,i} - s/2)/(s/6), & s/2 \le h_{t,i} < 2s/3, \\ 1, & h_{t,i} \ge 2s/3. \end{cases}$$
   Starting with $\mathcal{S}_t = \varnothing$, independently add each i with probability $p_{t,i}$.
   (d) If $\mathcal{S}_t = \varnothing$, set $\mathsf{aborted} = \mathsf{true}$, $g_t = 0$, and go to step (f). Otherwise compute $Q_t = |\mathcal{S}_t|^{-1}\sum_{i\in\mathcal{S}_t} \nabla \widehat{F}_{Z_i}(w_t)$.
   (e) Set $g_t = s^{-1}\sum_{i=1}^{s} \Pi_{B(Q_t,R)} \nabla \widehat{F}_{Z_i}(w_t) + v_t$, where $v_t \sim \mathcal{N}(0, \sigma^2 I_d)$ is fresh noise.
   (f) Update $w_{t+1} = \Pi_{\mathcal{W}}(w_t - \frac{4}{\lambda(t+1)}[g_t + \lambda(w_t - w_0)])$.
4. Return $2\sum_{t=1}^{T} t w_t/[T(T+1)]$.

For Algorithm 4, take $\mathcal{F}_t$ to include Z, w_0, all earlier randomness, and the current test and filter randomness, but not the current Gaussian noise. Thus $\mathbb{E}[g_t \mid \mathcal{F}_t] = A_t$ on accepted, nonempty-core rounds, giving $b_t = A_t - \nabla \widehat{F}_Z(w_t)$ as in (16). On and after an abort, $g_t = 0$, so the same definitions give $b_t = -\nabla \widehat{F}_Z(w_t)$ and $v_t = 0$. The decomposition and conditional-noise bound therefore hold on every round. This analysis filtration need not itself be private; the privacy guarantee excludes the internal centers and filter randomness.

We next verify that keeping the center internal is compatible with privacy. We use the FriendlyCore coupling of Asi & Liu (2024, Lemmas 3.5 and 3.6); the additional calculation is the effect of the moving center on the average of all clipped user gradients. The AboveThreshold and Gaussian-composition tools are standard (Dwork & Roth, 2014, Sections 3.5 and 3.6).

Lemma E.3 (Privacy and noise of the fused oracle)**.** *Let $0 < \delta_0 < 1/2$, $0 < \varepsilon \le 1$, and $L_0 := \log(16T/\delta_0)$. If $s \ge CL_0/\varepsilon$, the transcript of noisy gradients and abort flags is (ε, δ_0)-user-level DP with independent Gaussian noise on accepted rounds of standard deviation*

$$\sigma = C\frac{(R + \tau L_0)\sqrt{T}L_0}{s\varepsilon}$$

in each coordinate. Consequently the conditional second moment of the noise is at most

$$\Sigma^2 = C^2\frac{d(R + \tau L_0)^2 T L_0^2}{s^2\varepsilon^2}. \tag{54}$$

If

$$R \geq C\tau L_0, \tag{55}$$

this becomes

$$\Sigma^2 \leq C' \frac{dR^2 T L_0^2}{s^2 \varepsilon^2}. \tag{56}$$

After an abort the noise is identically zero, so both its conditional mean-zero property and this variance upper bound remain valid.

Proof. Use $\varepsilon/2$ for the AboveThreshold comparisons, stopping permanently at the first rejection. We follow the coupling proof of Asi–Liu (Asi & Liu, 2024, Lemmas 3.5 and 3.6). Their coupling applies on the event that the AboveThreshold noises are bounded, and to a common accepted comparison history; it is not an unconditional global-sensitivity claim. The sample condition makes the probability of the excluded test-noise event at most $\delta_0/16$ in each of two neighboring executions. On this event an accepted test certifies sufficient concentration for their selected-mean coupling. For each round set $\zeta = \delta_0/(16T)$. Conditional on a common preceding private history, the query is the same in the two executions, and the selected means can be coupled so that

$$\|Q_t - Q_t'\| \leq C\tau \log(1/\zeta)/s$$

except with probability ζ. In particular the coupled selected sets are nonempty on these good events. Empty-set aborts outside them are included in the excluded failure probabilities.

For every unchanged vector y, translation of a Euclidean projection and nonexpansiveness give

$$\|\Pi_{B(Q_t,R)}(y) - \Pi_{B(Q_t',R)}(y)\| \leq 2\|Q_t - Q_t'\|.$$

The changed user's two projected vectors are at distance at most $2R + \|Q_t - Q_t'\|$. Hence the fused averages obey

$$\|A_t - A_t'\| \leq 2\|Q_t - Q_t'\| + (2R + \|Q_t - Q_t'\|)/s \leq C(R + \tau L_0)/s.$$

On the good coupling, Gaussian distributions with this mean separation are $(\varepsilon_t, \delta_t)$-indistinguishable for

$$\varepsilon_t = \frac{\varepsilon}{C\sqrt{T\log(16/\delta_0)}}, \qquad \delta_t = \delta_0/(16T),$$

with the displayed choice of σ and sufficiently large C. Averaging the Gaussian privacy inequality over the coupling preserves that inequality, with an additional failure probability ζ for its bad pairs. Composition then charges the coupling failures, the two excluded test-noise events, and the Gaussian composition failure to the remaining δ_0 budget. Advanced composition spends at most $\varepsilon/2$ on the Gaussian releases; the comparison transcript spends the other $\varepsilon/2$. All releases following an abort are fixed post-processing. This follows the cited privacy argument with the selected-mean sensitivity replaced by the fused-average bound derived above. No stochastic moment assumption is used. The variance claims follow from the Gaussian covariance. □

Clipping bias and fixed queries. The next deterministic inequality connects the two probabilistic estimates we will prove: the centers are accurate, and the average contribution of outlier users is small. This is the same projection argument used for the mean estimator in Appendix A; here it bounds empirical rather than population bias.

Lemma E.4 (Clipping bias under an accurate center)**.** *For any dataset Z, query $w \in \mathcal{W}$, and center Q satisfying $\|Q - \nabla F(w)\| \leq R/2$,*

$$\left\| \frac{1}{s}\sum_{i=1}^{s} \Pi_{B(Q,R)} \nabla \widehat{F}_{Z_i}(w) - \nabla \widehat{F}_Z(w) \right\| \leq \frac{1}{s}\sum_{i=1}^{s} \left(\|\nabla \widehat{F}_{Z_i}(w) - \nabla F(w)\| - R/2 \right)_+ . \tag{57}$$

Proof. For each user the length of the clipping displacement is $(\|\nabla \widehat{F}_{Z_i}(w) - Q\| - R)_+$, which is at most $(\|\nabla \widehat{F}_{Z_i}(w) - \nabla F(w)\| - R/2)_+$ by the center assumption. Average these inequalities and apply the triangle inequality. □

At a fixed query, independence controls the square of the right-hand side of (57). The second moment contributes the diagonal terms; the integrated tail bound from Appendix A contributes the cross terms.

Lemma E.5 (Average squared tail contribution at a fixed query)**.** *For every fixed $w \in \mathcal{W}$ and $j \in \{2, k\}$,*

$$\mathbb{E}\left[\left(\frac{1}{s}\sum_{i=1}^{s}(\|\nabla \widehat{F}_{Z_i}(w) - \nabla F(w)\| - R/2)_+\right)^2\right] \le \psi_j(R)^2 + \frac{G_2^2}{ms},$$

where ψ_j is the tail bound in Lemma 4.4; explicitly,

$$\psi_2(R) = \frac{CG_2^2}{mR}, \qquad \psi_k(R) = C_k\left[\left(R + \frac{G_2^2}{mR}\right)e^{-c_k mR^2/G_2^2} + \frac{G_k^k}{m^{k-1}R^{k-1}}\right].$$

Proof. The centered user gradient has second moment at most G_2^2/m. Its expected tail contribution is at most $\psi_j(R)$, by the integrated Fuk–Nagaev calculation in the proof of Lemma A.5 for $j = k$, or $(x - R/2)_+ \le 2x^2/R$ for $j = 2$. The raw gradient-moment bounds imply centered moment bounds with a factor at most two, absorbed into C_k. Expanding the square gives s diagonal terms of expectation at most G_2^2/m and $s(s-1)$ cross terms of expectation at most $\psi_j(R)^2$, since distinct users are independent. Division by s^2 proves the claim. □

From fixed queries to private adaptive queries. We follow the DP generalization approach of Bassily & Sun (2023, Theorem 3.4) and Asi & Liu (2024, Lemma 4.6): compare a private query's behavior on its input users with its behavior on independent users. The next lemma is just the standard bounded-expectation consequence of DP and group privacy (Dwork & Roth, 2014, Section 2.3), written for the one- and two-user terms that occur when we square an average. We also use its general form for products of bad-user indicators below.

Lemma E.6 (DP comparison with independent users)**.** *Let $\mathcal{A}$ be $(\varepsilon_0, \delta_0)$-user-level DP and let $Z = (Z_1, \dots, Z_s)$ consist of independent user blocks, each distributed as P^m. Fix a set J of ℓ indices, and let Z'_J be independent copies of those blocks, independent also of Z and the randomness of $\mathcal{A}$. For any measurable $h(w, z_J) \in [0, M]$,*

$$\mathbb{E}h(\mathcal{A}(Z), Z_J) \le e^{\ell\varepsilon_0}\mathbb{E}h(\mathcal{A}(Z), Z'_J) + M\delta_0\sum_{a=0}^{\ell-1}e^{a\varepsilon_0}. \tag{58}$$

Proof. Let $Z^{(J)}$ be Z with the blocks in J replaced by Z'_J. First hold all these blocks fixed. The two mechanism inputs differ in ℓ users. Applying group privacy to the bounded function $w \mapsto h(w, Z_J)$ gives

$$\mathbb{E}_{\mathcal{A}}h(\mathcal{A}(Z), Z_J) \le e^{\ell\varepsilon_0}\mathbb{E}_{\mathcal{A}}h(\mathcal{A}(Z^{(J)}), Z_J) + M\delta_0\sum_{a=0}^{\ell-1}e^{a\varepsilon_0}.$$

For clarity, this bounded-function form follows by integrating the DP inequality for the events $\{h(w, Z_J) > u\}$ over $0 \le u \le M$; the additive privacy error is therefore multiplied by M. Now average over the data. The removed blocks Z_J are independent of $\mathcal{A}(Z^{(J)})$, and $Z^{(J)}$ has the same distribution as Z. Thus the expectation on the right equals $\mathbb{E}h(\mathcal{A}(Z), Z'_J)$, proving (58). □

The departure from the bounded-data analyses is in what we transfer, not in the DP comparison itself. Under heavy tails, requiring every user's gradient to be accurate at every round is too costly. We instead transfer the first and second moments of the users' tail contributions and then average over rounds. Because approximate DP has an additive error, we first cap these contributions at a deterministic threshold; the moment assumption shows that the cap is inactive simultaneously for every user and query with high probability.

Lemma E.7 (Average squared tail contribution at private queries)**.** *Suppose $(w_1, \dots, w_T)$ is $(\varepsilon_0, \delta_0)$-user-level DP, with $0 < \varepsilon_0 \le 1$. Let $0 < \gamma < 1/10$ and $\delta_0 \le \gamma/[4(1+e)ms^2]$. For $j \in \{2, k\}$ and $0 < \nu < 1$, with probability at least $1 - \gamma - \nu$,*

$$\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{s}\sum_{i=1}^{s}(\|\nabla \widehat{F}_{Z_i}(w_t) - \nabla F(w_t)\| - R/2)_+\right]^2 \le \frac{C}{\nu}\left[\psi_j(R)^2 + \frac{G_2^2}{ms}\right]. \tag{59}$$

Proof. Set $H = G_2(1 + \sqrt{s/\gamma})$, and recall $\mathcal{L}$ from (41). Jensen's inequality gives $\mathbb{E}[(m^{-1}\sum_a \mathcal{L}(z_{i,a}))^2] \le G_2^2$. Markov's inequality and a union bound over users therefore give

$$\Pr\left(\max_{i\in[s]} \frac{1}{m}\sum_{a=1}^{m} \mathcal{L}(z_{i,a}) > G_2\sqrt{s/\gamma}\right) \le \gamma.$$

Outside this event, for every user and every $w \in \mathcal{W}$, the tail contribution $(\|\nabla \widehat{F}_{Z_i}(w) - \nabla F(w)\| - R/2)_+$ is at most H, since $\|\nabla F(w)\| \le G_2$. In particular, replacing each tail contribution by its minimum with H changes nothing at any of the adaptive iterates.

We bound the capped quantities first. At a fixed query, Lemma E.5's proof bounds each capped contribution's mean by $\psi_j(R)$ and its second moment by G_2^2/m. At w_t, expand the square of their average. Apply Lemma E.6 with $\ell = 1$ to each squared contribution, and with $\ell = 2$ to each product from distinct users. Both tests lie in $[0, H^2]$. Their independent-user expectations give

$$\mathbb{E}\left[\left(\frac{1}{s}\sum_{i=1}^{s} \min\Big\{(\|\nabla \widehat{F}_{Z_i}(w_t) - \nabla F(w_t)\| - R/2)_+, H\Big\}\right)^2\right] \le e^{2\varepsilon_0}\psi_j(R)^2 + \frac{e^{\varepsilon_0}G_2^2}{ms} + (1+e^{\varepsilon_0})\delta_0 H^2.$$

Since $H^2 \le 4G_2^2 s/\gamma$, the assumed bound on δ_0 makes the last term at most $G_2^2/(ms)$. Average over t and apply Markov's inequality once, with failure probability ν. Intersect with the event that all caps are inactive, whose failure probability is at most γ. This proves (59), with no union bound over the T heavy-tailed quantities. □

Accurate centers at all iterations. The remaining requirement in Lemma E.4 is an accurate clipping center. We show that fewer than 5% of users are far from the population gradient at every iterate. Rather than rule out every individual outlier, we apply the same DP comparison to products of bad-user indicators. This controls the probability of having too many outliers and allows the coarse-center argument of Appendix A to apply.

Lemma E.8 (Accurate centers along the adaptive trajectory)**.** *Run Algorithm 4 with privacy calibration $(\varepsilon_0, \delta_0)$, where $0 < \varepsilon_0 \le 1$, and suppose its iterate transcript has this privacy guarantee. For $0 < \beta_c < 1/10$, use*

$$r = \sqrt{160}\, e^{\varepsilon_0/2} G_2/\sqrt{m}, \qquad \tau = 2r, \qquad \ell = \lceil \log_4(4T/\beta_c) \rceil .$$

Assume $R \ge 10r$, $s \ge 40\ell$, $s \ge C\log(T/\beta_c)/\varepsilon_0$, and

$$\delta_0 \le \frac{\beta_c}{4T\ell(40e)^\ell}. \tag{60}$$

Then with probability at least $1 - \beta_c$, Algorithm 4 never aborts and $\|Q_t - \nabla F(w_t)\| \le R/2$ for every t.

Proof. For a fixed w, Chebyshev gives $\Pr(\|\nabla \widehat{F}_{Z_i}(w) - \nabla F(w)\| > r) \le e^{-\varepsilon_0}/160$. For any set J of ℓ users, independence at a fixed query and Lemma E.6 with $M = 1$ therefore imply

$$\mathbb{E}\prod_{i\in J} \mathbf{1}\{\|\nabla \widehat{F}_{Z_i}(w_t) - \nabla F(w_t)\| > r\} \le 160^{-\ell} + \delta_0 \sum_{a=0}^{\ell-1} e^{a\varepsilon_0}.$$

Sum this inequality over the $\binom{s}{\ell}$ choices of J. If at least $\lceil s/20 \rceil$ users violate the radius-r bound, at least $\binom{\lceil s/20 \rceil}{\ell}$ of these products equal one. Since $s \ge 40\ell$, the ratio of the two binomial coefficients is at most 40^ℓ. Markov's inequality thus gives

$$\Pr\Big(\#\{i : \|\nabla \widehat{F}_{Z_i}(w_t) - \nabla F(w_t)\| > r\} \ge \lceil s/20 \rceil\Big) \le 4^{-\ell} + \ell(40e)^\ell \delta_0 \le \beta_c/(2T).$$

A union bound over rounds leaves at least $0.95s$ good users at every iterate, except with probability $\beta_c/2$. The sample-size assumption also makes every AboveThreshold Laplace noise at most $s/40$ in absolute value, except with probability $\beta_c/2$. On their intersection the concentration score exceeds $0.95^2 s > 0.9s$, so every test accepts. Every good user is selected, and every selected user is within 2τ of a good one, hence within $2\tau + r = 5r$ of the population gradient. Their average is therefore within $5r \le R/2$, and the selected set is never empty. □

Choice of the privacy parameter used by the solver. For an advertised guarantee of (ε, δ)-DP, calibrate the solver to (ε, δ_0) instead. Set $\varepsilon_0 = \varepsilon$, $\gamma = \nu = \beta_c = 1/100$, $\ell = \lceil \log_4(4T/\beta_c) \rceil$, and

$$\delta_0 := \min\left\{ \frac{\delta}{10},\ \frac{\gamma}{4(1+e)ms^2},\ \frac{\beta_c}{4T\ell(40e)^\ell} \right\}. \tag{61}$$

Since $(40e)^\ell \le 40e(4T/\beta_c)^{\log_4(40e)}$, this is inverse polynomial in m, s, T in addition to the external δ; in particular $\log(1/\delta_0) = O(\log(msT/\delta))$. This same δ_0 is used for the oracle's privacy calibration and for both transfer lemmas. The two extra upper bounds control the additive approximate-DP terms in (58). Setting $\delta_0 = \delta$ would not satisfy these bounds for arbitrary advertised δ. Since $\delta_0 \le \delta$, this calibration still gives the advertised privacy guarantee.

Bias of the fused adaptive oracle.

Lemma E.9 (Precise statement of Lemma 4.4)**.** *Assume* (2)*. Run Algorithm 4 for $T \ge 1$ iterations with δ_0 from* (61)*, $r = \sqrt{160}e^{\varepsilon/2}G_2/\sqrt{m}$, $\tau = 2r$, and $L_0 = \log(16T/\delta_0)$. If $s \ge CL_0/\varepsilon$ and $R \ge 10r$, then for either fixed $j \in \{2, k\}$, with probability at least* 0.97*,*

$$\frac{1}{T}\sum_{t=1}^{T} \|b_t\|^2 \le C_k \left[\psi_j(R)^2 + \frac{G_2^2}{ms} \right]. \tag{62}$$

Proof. Lemma E.3 makes the noisy-gradient transcript (ε, δ_0)-DP; the iterates are its post-processing. Lemma E.7 therefore bounds the average squared tail contributions except with probability $\gamma + \nu = 0.02$. Lemma E.8 ensures no abort and accurate centers except with probability $\beta_c = 0.01$. On their intersection, Lemma E.4 bounds $\|b_t\|$ by the average tail contribution in (59) at every round. Substitution proves (62), with failure probability at most 0.03. □

E.3 Regularized ERM guarantee

We now choose the clipping radius, number of iterations, and internal privacy parameter. The only additional step is connecting the constant-probability realized-bias bound to Lemma E.2's expected-bias assumption. The proof uses an auxiliary process for that connection.

Proposition E.10 (Precise statement of Proposition 4.3)**.** *Assume* (2)*, $s \le n$, $s \ge C_k\Lambda/\varepsilon$, and $0 < \lambda D \le 32G_2\sqrt{nm}$. The last condition includes every regularization parameter used in the nontrivial branch of Algorithm 3. Choose*

$$T = \left\lceil C_k \frac{G_2^2 + \lambda^2 D^2}{\mathcal{E}_{s,m}(G_2, G_k)^2} \right\rceil, \qquad a = \frac{C_k\Lambda\sqrt{d}}{s\varepsilon}.$$

Choose δ_0 by (61)*. Select $j \in \{2, k\}$ to minimize $G_j(a/m)^{1-1/j}$ and take*

$$R = \max\left\{ C_k\Lambda G_2/\sqrt{m},\quad G_j m^{-(1-1/j)} a^{-1/j} \right\}. \tag{63}$$

For sufficiently large fixed constants, Algorithm 4 with these choices is (ε, δ)-user-level DP and satisfies (12)*.*

Proof. Since $\mathcal{E}_{s,m}(G_2, G_k) \ge G_2/\sqrt{sm}$ and $\lambda D \le 32G_2\sqrt{nm}$, the displayed choice gives $T \le 1 + C_k(1 + 1024nm)sm \le C_k'(nm)^2$. Equation (61) therefore implies $L_0 = \log(16T/\delta_0) = O_k(\Lambda)$. Thus the stated user count implies the requirements of Lemmas E.3 and E.8, and the radius floor ensures $R \ge C\tau L_0$ and $R \ge 10r$. These are consequences of the chosen parameters, not extra

hypotheses on T. By Lemma E.9, with probability at least 0.97 the realized average squared bias is bounded by

$$B_0^2 := C_k[\psi_j(R)^2 + G_2^2/(ms)].$$

For the proof only, couple a second run to the original using the same data and random draws. Before each update, include the current test and filter outcomes in $\mathcal{F}_t$ and compute the current bias, but do not yet draw the Gaussian noise. If using this bias would make its accumulated squared bias exceed TB_0^2, replace the current and all subsequent gradient estimates by exact empirical gradients plus fresh Gaussian noise. Otherwise take the same update as the original algorithm. This is an analysis device, not a test implemented by Algorithm 4.

The bias that would exceed the threshold is never used, and the biases after the switch are zero. Thus, writing $\widetilde{b}_t$ for this run's biases, on every outcome

$$\sum_{t=1}^{T} \|\widetilde{b}_t\|^2 \le TB_0^2, \qquad \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\widetilde{b}_t\|^2 \le B_0^2.$$

The switch decision precedes the current Gaussian draw, so the noise remains conditionally mean zero with second moment at most Σ^2. This second run need not be private: DP stability was applied only to the original run. The gradient-moment assumption also holds for its possibly different iterates, because (41) gives the following bound uniformly in w:

$$\|\nabla \widehat{F}_Z(w)\| \le \frac{1}{sm}\sum_{i,a} \mathcal{L}(z_{i,a}), \quad \mathbb{E}\left(\frac{1}{sm}\sum_{i,a} \mathcal{L}(z_{i,a})\right)^2 \le G_2^2.$$

This verifies the gradient second-moment hypothesis of Lemma E.2 without any independence of the iterate. The distance conclusion of that lemma gives, with probability at least 0.9, an auxiliary output within

$$\frac{C}{\lambda}\left[B_0 + \frac{G_2 + \lambda D}{\sqrt{T}} + \frac{\Sigma}{\sqrt{T}}\right]$$

of $\widehat{w}_{\lambda,Z}$. On the original run's event $\sum_t \|b_t\|^2 \le TB_0^2$, every partial sum is below the same threshold, so no switch occurs and both outputs agree. A union bound with the second run's distance guarantee therefore gives success probability at least $0.97 + 0.9 - 1 = 0.87$ for the original algorithm.

The oracle calibration gives $\Sigma/\sqrt{T} \le aR$, after fixing the constant in a. For $j = 2$ the bias is at most $CG_2^2/(mR)$; for $j = k$ the exponential term is at most $G_2/\sqrt{ms}$ because of the radius floor. Substitution of (63) therefore gives

$$B_0 + aR \le C_k\left[\frac{G_2}{\sqrt{ms}} + \frac{G_2\sqrt{d}\Lambda^2}{s\varepsilon\sqrt{m}} + \min\{G_2(a/m)^{1/2}, G_k(a/m)^{1-1/k}\}\right].$$

This is at most a fixed multiple of $\Lambda^2\mathcal{E}_{s,m}(G_2, G_k)$. The choice of T bounds the optimization remainder by $\mathcal{E}_{s,m}(G_2, G_k)$. The fixed constant in $C_k\Lambda^4\mathcal{E}_{s,m}(G_2, G_k)$ absorbs all these terms, giving the target success probability at least 0.8. Privacy follows from Lemma E.3, since $\delta_0 \le \delta$. □

E.4 The final SCO upper bound

Proof of Theorem E.1. In the branch $\Delta \ge G_2$, Algorithm 3 returns w_0 and the risk bound follows directly from the definition of Δ. Otherwise, $32^{I-1} < 32\sqrt{nm}$ and $\lambda_1 D = \Delta < G_2$ give $\lambda_i D \le 32G_2\sqrt{nm}$ for every phase. Moreover, $IJ = O(\Lambda^2)$ and $s = \lfloor n/(IJ)\rfloor$, so $n \ge C_k\Lambda^6/\varepsilon$ ensures $s \ge C_k'\Lambda/\varepsilon$. Proposition E.10 therefore gives (12) for every incoming center and every scheduled λ_i. Theorem D.1 completes the proof. □

F Norm-moment SCO: lower bounds

Theorem F.1 (Restatement of Theorem 4.6)**.** *Under* (8)*, on a ball of diameter D, every user-level (ε, δ)-DP algorithm $\mathcal{A}$ has a distribution of linear losses satisfying* (2) *such that*

$$\mathbb{E}[F(\mathcal{A}(\mathsf{Z})) - F^*] \ge c_k D \min\{G_2, \mathcal{E}_{n,m}(G_2, G_k)\}. \tag{64}$$

We prove Theorem F.1 by combining the classical nonprivate term with two private lower bounds. Lemma F.2 gives the user-level Gaussian term under raw gradient-moment bounds. Lemma F.3 inserts rare scaled losses from that family among zero losses to obtain the heavy-tail term. The final proof groups items into users and combines these terms.

F.1 Gaussian linear losses under raw moment bounds

We adapt the Gaussian fingerprinting calculation of Kamath et al. (2019); Agarwal et al. (2025), already used in Lemma B.6. The means must now be restricted, not just translated, because SCO bounds raw moments. Small excess risk forces correlation between the optimization output and the mean; the same integration-by-parts and DP comparison then bound that correlation.

Lemma F.2 (Gaussian linear-SCO lower bound)**.** *Fix $k > 2$, $G_2 \le G_k$, and assume* (36)*. On a Euclidean ball of diameter D, every (ε, δ)-user-level DP algorithm has, for some Gaussian linear-loss distribution satisfying* (42)*, expected excess risk at least*

$$c_k D G_2 \min\left\{1, \frac{\sqrt{d}}{n\varepsilon\sqrt{m}}\right\}.$$

The hard family's population excess risk is uniformly bounded by a constant depending on k times this scale. In particular, for $m = 1$, $G_2 = G_k = 1$, and $n \le \sqrt{d}/\varepsilon$, this gives a family with constant-order expected risk D and raw moments at most one.

Proof. Center the feasible ball at zero and give it radius $D/2$. Let $K_k = (\mathbb{E}|g|^k)^{1/k}$ for a standard scalar Gaussian g, and choose

$$\sigma = \frac{G_2}{2(K_k+1)\sqrt{d}}, \qquad a = c_0\sigma \min\left\{1, \frac{\sqrt{d}}{n\varepsilon\sqrt{m}}\right\}.$$

For $\theta \in [-a, a]^d$ let $z \sim \mathcal{N}(\theta, \sigma^2 I_d)$ and $f(w, z) = -\langle w, z\rangle$. Minkowski's inequality gives $(\mathbb{E}\|z\|^j)^{1/j} \le a\sqrt{d} + \sigma K_k\sqrt{d} \le G_2$ for $j = 2, k$, if $c_0 \le 1$. Thus both raw moment constraints hold. The population objective is $-\langle w, \theta\rangle$, with optimum $-(D/2)\|\theta\|$.

Give θ the uniform prior on its cube. If an algorithm's output w had expected excess risk below $Da\sqrt{d}/8$ for every θ, then $\mathbb{E}\|\theta\| \ge \mathbb{E}\|\theta\|_1/\sqrt{d} = a\sqrt{d}/2$ implies

$$\mathbb{E}\langle\theta, w\rangle \ge Da\sqrt{d}/8.$$

Let $g_i = \sqrt{m}(Y_i - \theta)/\sigma$ as in Lemma B.6, and define $V_i = \sum_j (a^2 - \theta_j^2) w_j g_{i,j}$. The same integration by parts gives

$$\frac{\sqrt{m}}{\sigma}\sum_i \mathbb{E}V_i = 2\mathbb{E}\langle\theta, w\rangle \ge Da\sqrt{d}/4.$$

After replacing user i, the corresponding statistic computed using the independent replacement output is conditionally centered Gaussian with mean absolute value at most $a^2 D/2$. Also $|V_i| \le (a^2 D/2)\|g_i\|$ pointwise. Apply the signed clipped-test comparison and Gaussian tail integration from Lemma B.6, now with clipping threshold $(a^2 D/2)(\sqrt{d} + \sqrt{2\log(1/\delta)})$. This yields

$$\mathbb{E}V_i \le Ca^2 D[\varepsilon + \delta(\sqrt{d} + \sqrt{\log(e/\delta)})] \le C' a^2 D\varepsilon.$$

The two bounds imply $a\sqrt{d} \le C'' n\sqrt{m}\varepsilon a^2/\sigma$, which contradicts the definition of a for sufficiently small c_0. Thus some mean gives risk at least $Da\sqrt{d}/8$, the claimed scale. For every mean in the cube and every feasible output, the excess risk is at most $D\|\theta\| \le Da\sqrt{d}$, proving the range assertion. □

F.2 Rare scaled losses

Following the rare-outlier constructions of Barber & Duchi (2014); Lowy & Razaviyayn (2023), we scale a hard bounded-moment family and use it with small probability. As in the mean-estimation proof, the nonzero observations are random. We need the uniform risk bound in Lemma F.2 to control the reduction's possible abort.

Lemma F.3 (Sparse-mixture SCO lower bound)**.** *Fix $k > 2$, $G_2 \le G_k$, and an item sample size N. Suppose $\delta(\sqrt{d}+\sqrt{\log(e/\delta)}) \le c\varepsilon$ and $\sqrt{d}/\varepsilon$ is larger than a sufficiently large absolute constant. For every item-level (ε, δ)-DP algorithm there is a distribution over convex linear losses on a Euclidean ball of diameter D satisfying, for $j \in \{2, k\}$,*

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}} \|\nabla f(w, z)\|_2^j\right] \le G_j^j, \tag{65}$$

for which that algorithm has expected excess population risk

$$\Omega_k\left(D \min\left\{G_2 \min\{1, \sqrt{d}/(N\varepsilon)\}^{1/2}, \quad G_k \min\{1, \sqrt{d}/(N\varepsilon)\}^{1-1/k}\right\}\right). \tag{66}$$

Proof. Set $p = c_0 \min\{1, \sqrt{d}/(N\varepsilon)\}$, let C be a sufficiently large constant chosen below, and put $s = \lceil CNp \rceil$. Set

$$b := \frac{1}{2}\min\{G_2 p^{-1/2}, G_k p^{-1/k}\}.$$

Use Lemma F.2 with $n = s$, $m = 1$, and $G_2 = G_k = 1$ to obtain a Gaussian family $\mathcal{Q}$. For $X \sim Q \in \mathcal{Q}$, define a sample and its linear loss by

$$z = \begin{cases} 0, & \text{with probability } 1-p, \\ bX, & \text{with probability } p, \end{cases} \qquad f(w, z) = -\langle w, z\rangle.$$

Call this sample distribution P_Q. Zero samples contribute no loss or gradient; the rare nonzero samples carry the unknown mean. For $j \in \{2, k\}$, $\mathbb{E}_{P_Q}\|z\|^j = pb^j \mathbb{E}_Q\|X\|^j \le G_j^j$. Every normalized Q has population loss with Lipschitz constant at most one, so its excess risk is at most D on $\mathcal{W}$.

Suppose a private algorithm $\mathcal{A}$ had expected risk less than $cDpb$ for every P_Q. Given s auxiliary losses from Q, draw N independent Bernoulli-p indicators. If their sum K is at most s, place the first K auxiliary losses multiplied by b into the active positions and put the zero loss in the rest; run $\mathcal{A}$ on this synthetic dataset. If $K > s$, return a fixed feasible point. For fixed indicators each auxiliary loss is used at most once, so this defines an item-level (ε, δ)-DP algorithm $\mathcal{B}$ on the s auxiliary samples.

Let Z be the auxiliary s-sample dataset from Q, and couple the synthetic input to $\widetilde{Z} \sim P_Q^{\otimes N}$. The coupling fails only if $K > s$, an event of probability at most $1/C$ by Markov's inequality. Writing $F_{P_Q} = pbF_Q$ and $F_Q^* = \min_{w\in\mathcal{W}} F_Q(w)$, the bounded risk of the normalized family gives

$$\mathbb{E}[F_Q(\mathcal{B}(Z)) - F_Q^*] \le \frac{1}{pb}\mathbb{E}[F_{P_Q}(\mathcal{A}(\widetilde{Z})) - F_{P_Q}^*] + D/C.$$

Choose C sufficiently large, then c_0 sufficiently small, so that $s \le \sqrt{d}/\varepsilon$; the additive one is absorbed by the lower bound on $\sqrt{d}/\varepsilon$. The Gaussian SCO lemma supplies a Q for which $\mathcal{B}$ has expected risk at least a fixed k-dependent constant times D. The preceding display is a contradiction for sufficiently small c and large C. Since $pb = \frac{1}{2}\min\{G_2 p^{1/2}, G_k p^{1-1/k}\}$ and $p = c_0 \min\{1, \sqrt{d}/(N\varepsilon)\}$, the asserted lower bound follows with constants depending only on k. □

F.3 Combining the lower bounds

Here $\mathcal{A}$ has output in $\mathcal{W}$, and the expectation includes both the nm i.i.d. observations and the algorithm's randomness. The rate $\mathcal{E}_{n,m}(G_2, G_k)$ is (3) with the raw SCO moment bounds. This matches Theorem 4.5 for every G_k/G_2 in that theorem's user-count range.

Proof of Theorem F.1. The classical two-point stochastic-optimization bound supplies $cDG_2/\sqrt{nm}$. One may use the scaled Bernoulli testing family in Duchi (2023, Example 8.3.1) with linear losses $f(w, z) = -\langle w, z\rangle$. It satisfies both raw moment bounds because $\|z\| = G_2$ almost surely; we do not repeat the testing proof.

For the heavy-tail privacy term, group the $N = nm$ loss samples into n users exactly as in Lemma B.2 and apply Lemma F.3. When $\sqrt{d}/\varepsilon = O(1)$, the same calculation as in the proof of Theorem B.1 shows that this term is already dominated by the nonprivate term.

Finally, Lemma F.2 gives the capped Gaussian user-level privacy term directly under both raw moment constraints. The common assumption (27) implies its privacy-parameter condition.

These three constructions may be different. Select the construction corresponding to the largest numerical term and use that the maximum of three nonnegative quantities is at least one third of their sum. This proves (64). To see the equivalence to $D\min\{G_2, \mathcal{E}_{n,m}(G_2, G_k)\}$, let $x = \sqrt{d}/(nm\varepsilon)$. For $x \le 1$ the heavy-tail branch already equals that in $\mathcal{E}_{n,m}(G_2, G_k)$ and is at most G_2. For $x > 1$ both the capped heavy-tail term and the capped Gaussian term are of order G_2. A sum of three individually capped nonnegative terms is within a factor of three of capping their sum. □

G Directional-moment mean estimation: proofs

Theorem G.1 (Restatement of Theorem 5.1)**.** *Under* (5)*, there is a computationally efficient user-level* (ε, δ)*-DP estimator with constant-probability* ℓ_2*-error at most* $\alpha > 0$ *using*

$$n = \widetilde{O}_k\left[\frac{d}{m\alpha^2} + \frac{d\sqrt{\log(1/\delta)}}{\alpha\sqrt{m}\,\varepsilon} + \frac{d\sqrt{\log(1/\delta)}}{m\alpha^{k/(k-1)}\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right]. \tag{67}$$

This is tight up to logarithms if (8) *holds and* $\delta \le c_k(\varepsilon\min\{1,\alpha\}/d)^{C_k}$ *for sufficiently large* C_k*.*

Assume $k > 2$ and $\sup_{\|v\|=1}\mathbb{E}|\langle z-\mu, v\rangle|^k \le 1$. For each user let $Y_i = m^{-1}\sum_{a=1}^m z_{i,a}$.

For Theorem G.1, the estimator of Agarwal et al. (2025) already suffices when $\alpha \le 1/\sqrt{m}$, and our norm-moment estimator suffices when $\alpha \ge \sqrt{d/m}$. The work below concerns the intermediate range: Lemma G.2 recalls the clipping bound, Lemma G.3 chooses the radius, and Lemma G.4 iterates the resulting bound with fresh users. We then combine the three accuracy ranges and the lower bounds. For precise lookup, numbered references to Agarwal et al. (2025) use the authors' full version, arXiv:2405.20405v3; the bibliography cites its SODA publication.

G.1 Refinement for intermediate accuracy

The following one-step guarantee is an application of the clipping-bias and sampling bounds of Agarwal et al. (2025), not a new concentration inequality. It shows explicitly how the current center error enters the bias.

Lemma G.2 (One Clip-and-Noise step)**.** *Let* u *be independent of* s *fresh users and satisfy* $\|u-\mu\| \le \Gamma \le \sqrt{d/m}$*. Clip their averages to* $B(u,R)$*, average, and add Gaussian noise with standard deviation* $CR\sqrt{\log(1/\delta)}/(s\varepsilon)$ *per coordinate. The result* $\widehat{\mu}^+$ *is* (ε,δ)*-user-level DP and, with constant probability, its error* $\Gamma^+ = \|\widehat{\mu}^+ - \mu\|$ *satisfies*

$$\Gamma^+ \le \widetilde{O}_k\left[\sqrt{\frac{d}{ms}} + \frac{R\sqrt{d\log(1/\delta)}}{s\varepsilon} + \frac{d^{(k-1)/2}}{m^{k-1}R^{k-1}} + \frac{d^{k/2}\Gamma}{m^{k-1}R^k}\right], \tag{68}$$

provided

$$R \gtrsim \sqrt{\frac{d\log(em)}{m}}.$$

Proof. Apply Agarwal et al. (2025, Theorem 4.2 and Lemma A.9) to the fixed center u. The Gaussian mechanism applies because the clipped average has replacement sensitivity $2R/s$. The same bounds apply with a prescribed smaller failure probability: the sampling term increases by its inverse square root, while the Gaussian term changes logarithmically. We account for these factors explicitly in the proof of Lemma G.4. □

We balance the two bias terms in Lemma G.2 against its privacy-noise term, subject to the minimum admissible radius.

Lemma G.3 (Radius choice and contraction)**.** *Let* $L_{\mathrm{dir}} = \log(esmd/(\varepsilon\delta))$*. We seek a constant overall success probability; the allocation of per-round failure probabilities is specified below.*

Choose

$$R = C_k L_{\mathrm{dir}}^{C_k} \max\left\{ \sqrt{\frac{d\log(em)}{m}}, \quad \left(\frac{d^{(k-1)/2} s\varepsilon}{m^{k-1}\sqrt{d\log(1/\delta)}}\right)^{1/k}, \right. \\ \left. \left(\frac{d^{k/2}\Gamma s\varepsilon}{m^{k-1}\sqrt{d\log(1/\delta)}}\right)^{1/(k+1)} \right\}. \tag{69}$$

The error scale used throughout the refinement is

$$A_s = C_k' L_{\mathrm{dir}}^{C_k'} \left[\sqrt{\frac{d}{ms}} + \frac{d\sqrt{\log(1/\delta)}}{s\varepsilon\sqrt{m}} + \left(\frac{d\sqrt{\log(1/\delta)}}{ms\varepsilon}\right)^{1-1/k} \right]. \tag{70}$$

For sufficiently large constants, a successful step satisfies

$$\Gamma^+ \le C_k A_s \left[1 + (\Gamma/A_s)^{1/(k+1)}\right]. \tag{71}$$

Proof. Substituting (69) into Lemma G.2 bounds the error, up to its logarithmic factor, by

$$\sqrt{\frac{d}{ms}} + \frac{d\sqrt{\log(1/\delta)\log(em)}}{s\varepsilon\sqrt{m}} + \left(\frac{d\sqrt{\log(1/\delta)}}{ms\varepsilon}\right)^{1-1/k} \\ + \left[\Gamma\left(\frac{d\sqrt{\log(1/\delta)}}{s\varepsilon\, m^{1-1/k}}\right)^{k}\right]^{1/(k+1)}. \tag{72}$$

The first three terms, with their logarithmic factors, are bounded by A_s from (70). Since $m^{1-1/k} \ge \sqrt{m}$, the coefficient of Γ in the last term is at most $[d\sqrt{\log(1/\delta)}/(s\varepsilon\sqrt{m})]^k$. Choosing the logarithmic power in A_s large enough therefore bounds that term by $C_k A_s^{k/(k+1)}\Gamma^{1/(k+1)}$, proving (71). □

The radius must depend on a known upper bound for the center error, not on the unknown mean. The next lemma implements this using a deterministic sequence of bounds and disjoint user blocks.

Lemma G.4 (Private initialization and public error bounds)**.** *There is a private refinement procedure using $\widetilde{O}_k(\sqrt{d}/\varepsilon)$ initialization users and $T = O_k(1 + \log\log(e\max\{1, B_0/A_s\}))$ fresh blocks of s users, where $B_0 \le \sqrt{d/m}/4$ is public. With constant probability its error is $O_k(A_s)$, up to the logarithmic factors in* (70)*. It is (ε, δ)-user-level DP and uses no bound on the mean location.*

Proof. To obtain this initialization explicitly, group L_k actual users into each synthetic user, giving it $L_k m$ independent samples. Run the coarse estimator of (Agarwal et al., 2025, Theorem 4.19) on these synthetic users. If its accuracy constant is C_k, take a fixed integer $L_k \ge 16C_k^2$. Its data-independent error bound is then $B_0 = C_k\sqrt{d/(L_k m)} \le \sqrt{d/m}/4$. Changing one actual user changes only one synthetic user, so privacy transfers directly. The number of actual initialization users is $\widetilde{O}_k(\sqrt{d}/\varepsilon)$, a constant-factor change. For the approximate-DP histogram, use all integer-indexed bins of the specified width and a stability-based histogram that only processes occupied bins; this is the countable-domain version of the cited approximate-DP histogram. Its threshold suppresses bins supported only by the changed user. The accuracy proof is unchanged and no public bound on the location of μ is needed. Let $K = B(u_0, B_0)$; on its good event, $\mu \in K$.

At round t, use a fresh block of s users and set R_t to the radius in (69) with $\Gamma = B_t$. Here B_t is the known error bound updated below, and A_s retains its definition (70).

$$R_t = C_k L_{\mathrm{dir}}^{C_k} \max\left\{ \sqrt{\frac{d\log(em)}{m}}, \quad \left(\frac{d^{(k-1)/2} s\varepsilon}{m^{k-1}\sqrt{d\log(1/\delta)}}\right)^{1/k}, \right. \\ \left. \left(\frac{d^{k/2} B_t s\varepsilon}{m^{k-1}\sqrt{d\log(1/\delta)}}\right)^{1/(k+1)} \right\}. \tag{73}$$

Run Clip-and-Noise centered at u_t, obtaining v_{t+1}, and set $u_{t+1} = \Pi_K(v_{t+1})$. Define the deterministic error bound

$$B_{t+1} := \min\left\{2B_0, C_k A_s\left[1 + \left(\frac{B_t}{A_s}\right)^{1/(k+1)}\right]\right\}, \tag{74}$$

where C_k dominates the constant in (71).

Error bound at each round. Conditional on $\|u_0 - \mu\| \le B_0$ and $2B_0 \le \sqrt{d/m}$, whenever every completed round satisfies its one-step accuracy event,

$$\|u_t - \mu\| \le B_t \qquad \text{and} \qquad B_t \le \sqrt{d/m}$$

for all t.

Induct on t. The miscentering term in (68) is monotone in the center error, so (73) and the one-step calculation give

$$\|v_{t+1} - \mu\| \le C_k A_s\left[1 + \left(\frac{B_t}{A_s}\right)^{1/(k+1)}\right].$$

Projection onto K cannot increase distance to $\mu \in K$, and both $u_{t+1}, \mu \in K$ are also at distance at most $2B_0$. Taking the better bound gives (74); the final inequality follows from $\|u_0 - \mu\| \le B_0$ and $2B_0 \le \sqrt{d/m}$.

Dividing (74) by A_s gives

$$\max\{1, B_{t+1}/A_s\} \lesssim_k \max\{1, B_t/A_s\}^{1/(k+1)}. \tag{75}$$

Iteration yields $\max\{1, B_t/A_s\} \lesssim_k \max\{1, B_0/A_s\}^{1/(k+1)^t}$; the accumulated constants form a bounded geometric series. Thus after $T = O_k(1 + \log\log(e\max\{1, B_0/A_s\}))$ rounds, $B_T = O_k(A_s)$ and $\|u_T - \mu\| \le B_T$.

Success probability and user count. Assign failure probability $1/[10(T+1)]$ to initialization and to each of the T refinement rounds. Conditional on any successful preceding rounds, the next block is independent of its clipping center, so the one-step guarantee applies. A union bound gives overall failure probability at most $1/10$.

Lemma A.9 of Agarwal et al. (2025) increases the sampling term by $\sqrt{10(T+1)}$, not merely a logarithm. This factor is allowed in (70): since $A_s \ge \sqrt{d/(ms)}$ and $B_0 \le \sqrt{d/m}/4$, we have $B_0/A_s \le \sqrt{s}/4$ and hence $T = O_k(1 + \log\log(es))$. The fixed logarithmic power in A_s absorbs this factor and the smaller Gaussian-noise and initialization penalties. The procedure uses the initialization users plus Ts refinement users and returns u_T.

Every block uses the full privacy parameters. If a changed user belongs to an earlier block, later steps use only that block's private output and fixed fresh data; if the changed user belongs to the current block, its incoming center is fixed. Thus adaptive parallel composition proves privacy. □

G.2 Coarse accuracy from the norm-moment estimator

For large target error no directional refinement is needed. The directional moment assumption implies centered norm bounds $\widetilde{G}_2 = \widetilde{G}_k = \sqrt{d}$, so we can use Theorem A.1 directly.

Lemma G.5 (Norm-moment corollary for $\alpha \ge \sqrt{d/m}$)**.** *For $\alpha \ge \sqrt{d/m}$, Algorithm 1 with $\widetilde{G}_2 = \widetilde{G}_k = \sqrt{d}$ achieves error at most α with constant probability using*

$$n = \widetilde{O}_k\left[\frac{d}{\alpha\varepsilon\sqrt{m}} + \frac{1}{\varepsilon}\right]$$

users. No upper restriction on n is needed.

Proof. Minkowski's inequality in $L_{k/2}$ gives

$$\left(\mathbb{E}\|z - \mu\|^k\right)^{2/k} = \left\|\sum_{j=1}^d (z_j - \mu_j)^2\right\|_{L_{k/2}} \le \sum_{j=1}^d \left(\mathbb{E}|z_j - \mu_j|^k\right)^{2/k} \le d.$$

Also $\mathbb{E}\|z-\mu\|^2 \le d$. The second-moment branch of Theorem A.1 therefore has error

$$\widetilde{O}_k\left[\sqrt{\frac{d}{nm}}+\frac{d}{n\varepsilon\sqrt{m}}+\frac{d^{3/4}}{\sqrt{nm\varepsilon}}\right]$$

when $n \ge C_k\Lambda^4/\varepsilon$. Consequently it suffices to take

$$n=\widetilde{O}_k\left[\frac{d}{m\alpha^2}+\frac{d}{\alpha\varepsilon\sqrt{m}}+\frac{d^{3/2}}{m\varepsilon\alpha^2}+\frac{1}{\varepsilon}\right].$$

Since $\alpha \ge \sqrt{d/m}$, the first term is at most one and the third is at most the second. This gives the claim, including the logarithmic user requirement of the norm-moment theorem. □

G.3 Complete guarantee

Proof of Theorem G.1. If $d=1$, Theorem A.1 gives the claimed upper bound directly. If a cited high-dimensional statement requires $m \ge 2$ and the actual $m=1$, first group pairs of actual users into synthetic users with two samples; changing one actual user changes one synthetic block, and all displayed rates change by constant factors only. For $1/\sqrt{m} < \alpha < \sqrt{d/m}$, choose

$$s=\widetilde{O}_k\left[\frac{d}{m\alpha^2}+\frac{d\sqrt{\log(1/\delta)}}{\alpha\sqrt{m}\,\varepsilon}+\frac{d\sqrt{\log(1/\delta)}}{m\alpha^{k/(k-1)}\varepsilon}\right] \tag{76}$$

so that $C_kA_s \le \alpha$ for the constant in Lemma G.4. Since $\alpha \le \sqrt{d/m}$, the second term pays for the initialization users. For $\alpha \le 1/\sqrt{m}$, Agarwal et al.'s $\widetilde{O}(d/\varepsilon)$ refinement cost is dominated by the second term of (67). In the intermediate regime, use their coarse estimator and the implementable recursion (73) on fresh disjoint blocks. Lemma G.4 verifies every invocation of the clipping-bias theorem, and choosing s as in (76) yields the first three terms.

For $\alpha \ge \sqrt{d/m}$, Lemma G.5 applies. Here $d/(m\alpha^2) \le 1$, and

$$\frac{d/(m\alpha^{k/(k-1)}\varepsilon)}{d/(\alpha\sqrt{m}\,\varepsilon)}=\frac{1}{\sqrt{m}\,\alpha^{1/(k-1)}} \le 1.$$

Thus the target rate reduces to $d/(\alpha\sqrt{m}\,\varepsilon)+1/\varepsilon$, as supplied by Lemma G.5.

Privacy and the total cost of Ts refinement users follow from Lemma G.4. Since $T = O_k(1+\log\log(es))$, these repetitions add only logarithmic factors to (76).

For $0<\alpha\le 1$ and the sufficiently small-δ range of (Agarwal et al., 2025, Theorem 6.1), the first three terms match that lower bound after constant rescaling of its moment bound. For an unbounded mean location the last term is necessary when $\delta \le \varepsilon^2$ by (Esfandiari et al., 2022, Lemma 5.1). For $\alpha > 1$, the first and third terms are dominated by the second; Lemma B.6, with $\sigma = (\mathbb{E}|g|^k)^{-1/k}$ for $g \sim \mathcal{N}(0,1)$, supplies that Gaussian term for arbitrary target accuracy under (36). Together with the unbounded-location lower bound, this also proves tightness for $\alpha > 1$. A convenient common sufficient comparison range is (27) together with $\delta \le c_k(\varepsilon\min\{1,\alpha\}/d)^{C_k}$ for a sufficiently large constant C_k. Thus the complete displayed sample complexity is tight up to logarithms for all target accuracies in this small-δ range; no extension of the cited heavy-tail theorem beyond its hypotheses is assumed. □

H Additional comparisons

Bounded-gradient SCO. Bassily & Sun (2023, Theorem 3.1) obtained the optimal rate under smoothness assumptions when $n > \sqrt{d}/\varepsilon$. Their theorem also assumes

$$m < \max\{\sqrt{d}, n\varepsilon^2/\sqrt{d}\}, \qquad \delta < \frac{d}{4n^{5/2}\sqrt{m}\,\varepsilon^2},$$

and a parameter-dependent upper bound on the smoothness constant. Thus this result does not provide the nearly dimension-independent user threshold of Asi & Liu (2024); Lowy et al. (2024). All these results assume uniformly bounded sample gradients.

Grouping the few-user mean estimator. Consider the estimator of Esfandiari et al. (2022, Theorem 4.1). For independent, uncorrupted observations, a grouping corollary makes the large-n comparison explicit. Form $n_0 = \widetilde{\Theta}(\sqrt{d}/\varepsilon)$ groups of $b = \lfloor n/n_0 \rfloor$ users, discarding any remainder. Treat each group as one user with bm independent samples. Changing an original user changes only one group, so privacy is preserved. Their endpoint guarantee gives $\widetilde{O}(\widetilde{G}_2/\sqrt{bm}) = \widetilde{O}(\widetilde{G}_2\sqrt{\sqrt{d}/(nm\varepsilon)})$. Together with their few-user bound, this yields

$$\widetilde{O}\left(\max\left\{\frac{\widetilde{G}_2\sqrt{d}}{n\varepsilon\sqrt{m}}, \widetilde{G}_2\sqrt{\frac{\sqrt{d}}{nm\varepsilon}}\right\}\right), \qquad n \geq \widetilde{\Omega}(1/\varepsilon).$$

This is a grouping corollary, not the theorem's stated all-n guarantee, and does not preserve its original corrupted-user robustness. The large-n branch can match our second-moment branch; it is not universally suboptimal, but does not exploit $\widetilde{G}_k$ when our higher-moment branch is smaller. Their Corollary 4.2 gives an all-count result for bounded support, a stronger assumption than bounded second norm moment.

Huber-loss norm-moment estimation. Zhao et al. (2024) study user-level mean estimation through Huber-loss minimization, including imbalanced user sample sizes. Their Theorem 3 gives a normalized mean-squared-error rate whose square root yields the rate $\mathcal{E}_{n,m}(\widetilde{G}_k, \widetilde{G}_k)$ in Table 1, up to logarithms, after restoring one public moment scale. The theorem assumes a public bound on the mean location. Its sufficient user condition is expressed through a smooth-sensitivity parameter of order $\varepsilon/(d + \log(1/\delta))$, hence the dimension-dependent count in Table 1. The printed Assumption 2 describes a centered moment in prose but writes a raw moment in its display; we use its stated single-scale rate for comparison and do not silently strengthen that theorem to an unrestricted-location, two-parameter result. Our simultaneous-moment improvement refers to the published single-scale guarantee; recalibrating a prior estimator with additional moment information would need a separate analysis.