Nearly Linear-Time User-Level DP-SCO with Optimal Rates

Badih Ghazi BADIHGHAZI@GMAIL.COM
Ravi Kumar RAVI.K53@GMAIL.COM
Daogao Liu LIUDAOGAO@GMAIL.COM
Pasin Manurangsi PASIN@GOOGLE.COM
Google

Editors: Steve Hanneke and Tor Lattimore

Abstract

User-level differentially private (DP) stochastic convex optimization has garnered significant attention due to the paramount importance of safeguarding user privacy in modern large-scale ML applications. Current methods, such as those based on DP stochastic gradient descent (SGD), often struggle with high gradient computation complexity or suboptimal utility due to the need to privatize every intermediate iterate. In this work, we introduce a new nearly linear-time algorithm that resolves this trade-off and achieves the optimal excess rates via an adaptive outlier removal framework. Our key innovation is integrating the sparse vector technique directly into the SGD loop, supported by a novel robust divergence analysis. This approach naturally bounds the sensitivity of gradient estimates without requiring privatization of all intermediate steps. Specifically, our mechanism computes a local concentration score to adaptively filter out users whose updates diverge from the population geometry. Crucially, this approach preserves the unbiasedness of the gradient estimate in well-concentrated regimes while strictly bounding sensitivity in the presence of outliers. We also explore extensions to the ℓ_∞ setting demonstrating the generality of our analysis.
Keywords: Stochastic Convex Optimization, Differential Privacy, Linear Time

1. Introduction

With the rapid development and widespread applications of modern ML and AI, particularly driven by advancements in large language models (LLMs), privacy concerns have come to the forefront. For example, recent studies have highlighted significant privacy risks associated with LLMs, including well-documented instances of training data leakage (Carlini et al., 2021; Lukas et al., 2023). These challenges underscore the urgent need for privacy-preserving mechanisms in ML systems.

Differential Privacy (DP) (Dwork et al., 2006) has emerged as a rigorous mathematical framework for ensuring privacy and is now the gold standard for addressing privacy concerns in ML. The classic definition of DP, referred to as item-level DP, guarantees that the replacement of any single training example has a negligible impact on the model's output. Formally, a mechanism $\mathcal{M}$ is said to satisfy (ε, δ)-*item-level DP* if, for any pair $\mathcal{D}$ and $\mathcal{D}'$ of datasets that differ by a single item, and for any event $\mathcal{O} \in \mathrm{Range}(\mathcal{M})$, the following condition holds:

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{O}] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{O}] + \delta. \tag{1}$$

Item-level DP protects the information of a single item in a dataset. However, many applications involve users who contribute multiple items, creating a potential privacy vulnerability (Xu and Zhang, 2024). For example, in scenarios such as federated learning or on-device LLM fine-tuning, a single user contributes a large batch of data (e.g., chat history and messages). Protecting privacy

Table 1: User-Level DP-SCO: Error Rates and Gradient Evaluations.

Reference	Error Rate	Gradient Evaluations
Group Privacy	$O(\frac{1}{nm} + \frac{\sqrt{d}}{n\varepsilon})$	nm based on linear-time item-level Algorithm
Levy et al. (2021)	$O\left(\frac{1}{\sqrt{nm}} + \frac{d}{n\sqrt{m}\varepsilon}\right)$	$\frac{n^3m}{d}$
Bassily and Sun (2023)	$O\left(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\right)$	$\frac{n^2m}{\sqrt{d}}$ (parameter restriction)
Ghazi et al. (2023)	$O\left(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\sqrt{m}\varepsilon^{2.5}}\right)$	Super-polynomial
Asi and Liu (2024)	$O\left(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\right)$	$\beta(nm)^{3/2}$ or $(nm)^3$
Lowy et al. (2024) (Polynomial)	$O\left(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\right)$	$\max\{\beta^{1/4}(nm)^{9/8}, \beta^{1/2}n^{1/4}m^{5/4}\}$ or $n^{11/8}m^{5/4}$
Lowy et al. (2024) (Linear)	$O\left(\frac{\sqrt{d}}{\sqrt{nm}\varepsilon}\right)$	nm (Bassily and Sun (2023) in local DP)
This work	$\tilde{O}\left(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\right)$	nm

at the item-level is insufficient, as it fails to conceal the user's participation or distinctive characteristics. To address this, the stronger notion of *user-level DP* was introduced to protect the privacy of each user as a whole. User-level DP ensures that adding or removing one user, who may contribute up to m items, has a negligible effect on the model's output. Formally, this means that (1) holds for datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ by the contributions of a single user (up to m items). When $m = 1$, user-level DP becomes equivalent to item-level DP.

DP-SCO. As one of the central problems in privacy-preserving ML and statistical learning, DP stochastic convex optimization (DP-SCO) has garnered significant attention in recent years (e.g., Bassily et al. (2014, 2019); Feldman et al. (2020); Bassily et al. (2020, 2021); Su et al. (2022); Gopi et al. (2022); Asi et al. (2024)). In DP-SCO, we are provided with a dataset $\{Z_i\}_{i\in[n]}$ of users, where the ith user contributes m samples $Z_i \in \mathcal{Z}^m$ drawn i.i.d. from an underlying distribution $\mathcal{P}$. The goal is to minimize the population objective $F(x) := \mathbb{E}_{z\sim\mathcal{P}} f(x; z)$, under DP constraint, where $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is a convex[1] function, defined on the convex domain $\mathcal{X} \subset \mathbb{R}^d$.

User-Level DP-SCO. While DP-SCO has been extensively studied in the item-level setting, extending these methods using group privacy yields suboptimal rates. This highlights the clear need for specialized user-level algorithms.

The study of user-level DP-SCO was initiated by Levy et al. (2021), who achieved an error rate of $O(\frac{1}{\sqrt{mn}} + \frac{d}{n\varepsilon\sqrt{m}})$ for smooth functions and a lower bound of $\Omega(\frac{1}{\sqrt{mn}} + \frac{\sqrt{d}}{n\varepsilon\sqrt{m}})$. Bassily and Sun (2023) later achieved the asymptotically optimal rate of $O(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\sqrt{m}\varepsilon})$ using improved mean-estimation; however, they rely on the smoothness of the loss function and imposes parameter restrictions, including $n \geq \sqrt{d}/\varepsilon$ and $m \leq \max\{\sqrt{d}, n\varepsilon^2/\sqrt{d}\}$. On the other hand, Ghazi et al. (2023) observed that user-level DP-SCO has low local sensitivity to user deletions. Using the propose-test-release mechanism, they developed algorithms applicable even to non-smooth functions and requiring only $n \geq (\log d)/\varepsilon$ users. However, their algorithms run in super-polynomial time besides achieving a sub-optimal error rate of $O(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\sqrt{m}\varepsilon^{2.5}})$. Asi and Liu (2024) proposed a polynomial-time algorithm that achieves optimal excess risk, requiring only $n \geq (\log(md))/\varepsilon$

1. Throughout this work, when we say that a function f is convex or smooth, we only refer to its first argument x.

and accommodating non-smooth losses. However, their algorithm is also computationally expensive, requiring $\beta(nm)^{3/2}$ gradient evaluations for β-smooth losses and $(nm)^3$ gradient evaluations for non-smooth losses. Lowy et al. (2024) focused on improving the computational cost while maintaining optimal excess risk. For β-smooth losses, they designed an algorithm requiring $\max\{\beta^{1/4}(nm)^{9/8}, \beta^{1/2}n^{1/4}m^{5/4}\}$ gradient evaluations; for non-smooth losses, they achieved the same optimal excess risk using $n^{11/8}m^{5/4}$ evaluations.

Linear-time algorithms[2] have also been explored in the user-level DP-SCO setting. Bassily and Sun (2023) proposed a linear-time algorithm in the local DP model, achieving an error rate of $O(\sqrt{d}/\sqrt{nm}\varepsilon)$ under the constraints $m < d/\varepsilon^2$, $n > d/\varepsilon^2$, and $\beta \le \sqrt{n^3/md^3}$. Similarly, Lowy et al. (2024) achieved the same rate under slightly relaxed conditions, requiring $n \ge \log(ndm)/\varepsilon$ and $\beta \le \sqrt{nmd}$, but in the central DP model. See Table 1 for a summary of previous results.

1.1. Previous Techniques

A key challenge in solving user-level DP-SCO using DP-SGD lies in obtaining more accurate gradient estimates while maintaining privacy. Consider a simple scenario where we perform gradient descent for t steps and seek an estimate of $\nabla F(x_t)$. To achieve this, we sample B users $Z_1, \dots, Z_B$ and compute $q_t(Z_i) := \frac{1}{m}\sum_{z\in Z_i} \nabla f(x_t; z)$, the average of the m gradients from user Z_i at point x_t. If each user's m inputs are i.i.d. from $\mathcal{P}$, then whp., $\|q_t(Z_i) - \nabla F(x_t)\| \le \tilde{O}(1/\sqrt{m})$.

This naturally leads to the following mean-estimation problem: Given points $q_t(Z_1), \dots, q_t(Z_B)$ in the unit ball, with most of them likely to be within a distance of $1/\sqrt{m}$ from each other (under the i.i.d. assumption for utility guarantees), how well can we privately estimate their mean?

A straightforward approach to recover the item-level rate is to apply the Gaussian mechanism:

$$\left(\frac{1}{B}\sum_{i\in[B]} q_t(Z_i)\right) + \mathcal{N}(0, \sigma_1^2 I_d), \tag{2}$$

where to privatize the gradient, the noise level is set as $\sigma_1 \propto 1/B$. To improve upon this, Asi and Liu (2024); Lowy et al. (2024) designed mean-estimation methods with the following properties.

- *Outlier detection:* The procedure tests whether the number of "bad" users (whose gradients significantly deviate from the majority) exceeds a predefined threshold (or "break point").
- *Outlier removal and sensitivity reduction:* If the number of "bad" users is below the threshold, the procedure removes outliers and produces a gradient estimate with sensitivity $\tilde{O}(\frac{1}{B\sqrt{m}})$. The privatized gradient is then achieved by adding $\mathcal{N}(0, \sigma_2^2 I_d)$, where $\sigma_2 \propto \frac{1}{B\sqrt{m}}$.
- *Better variance control:* When all users provide consistent estimates, the output follows (2) but with σ_2 instead of σ_1, resulting in significantly smaller noise.

By leveraging such ideas, prior works have achieved the optimal excess risk rate in polynomial time. However, extending these to obtain *linear-time* algorithms poses new challenges.

1.2. Towards Linear-Time Algorithms: Challenges

Prior linear-time approaches for item-level DP-SCO (Feldman et al., 2020) work with a notably mild smoothness requirement ($\beta \le \sqrt{n} + \sqrt{d}/\varepsilon$; see Definition 4) and proceed by analyzing the *stability*

2. Following Feldman et al. (2020), "linear-time" is measured against gradient complexity: if computing or processing a single gradient takes $O(d)$ time, then they need n gradients for a total of $O(nd)$ running time.

of non-private SGD. Their core insight is demonstrating that, for any two neighboring datasets, the corresponding SGD trajectories (represented by $\{x_t\}_{t\in[T]}$ and $\{x'_t\}_{t\in[T]}$) stay remarkably close. This proximity guarantees a low sensitivity for the average iterate $\frac{1}{T}\sum_{t\in[T]} x_t$, which allows them to apply the Gaussian mechanism directly to privatize the average iterate in the end.

Motivated by this stability-based analysis, Lowy et al. (2024) attempted to generalize the linear-time approach of Feldman et al. (2020) to the user-level setting. However, a key difficulty arises when incorporating the mean-estimation sub-procedure. Specifically, even if one can bound $\|x_t - x'_t\|$, there is no clear understanding of how applying the sub-procedure impacts stability in subsequent iterations. In particular, after performing one gradient descent step using gradient estimations from the sub-procedure, we do not have guarantees on how well $\|x_{t+1} - x'_{t+1}\|$ remains bounded.

Due to this lack of stability analysis for the sub-procedure, Lowy et al. (2024) resorted to privatizing all iterations, resulting in excessive Gaussian noise accumulation. Consequently, their algorithm achieved only a suboptimal error rate of $O(\frac{\sqrt{d}}{\sqrt{nm\varepsilon}})$, highlighting a fundamental challenge of designing a linear-time algorithm by controlling the stability of the sub-procedures.

1.3. Our Contributions and Techniques

We develop the first linear-time algorithm[3] that achieves near-optimal privacy-utility trade-offs for user-level DP-SCO, by introducing an adaptive stability framework. Our algorithm requires only mn gradient computations while achieving the optimal excess risk rate of $\tilde{O}(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\sqrt{m\varepsilon}})$ for convex functions with smoothness $\beta \le \tilde{O}(\sqrt{(n+d)/m})$. When m is constant, this recovers the smoothness requirement from Feldman et al. (2020), demonstrating that our approach naturally generalizes item-level techniques to the user-level setting.

Our key algorithmic innovation is a concentration-based robust mean-estimation method that resolves the stability issue plaguing prior linear-time approaches. Unlike static robust estimators that always discard data, our method is adaptive: it reduces to the unbiased empirical mean when gradients are well-concentrated, and only triggers privacy-consuming filtering when necessary. At each step and for a user, we compute a *concentration score* that measures how well the user's gradient aligns with that of the overall population. Users with low concentration scores (outliers) are filtered out using a noisy threshold mechanism via the sparse vector technique (SVT) (Dwork and Roth, 2014, Section 3.6). A major advantage of our method is that when all user gradients are well-concentrated (say, for example, data is drawn i.i.d.), no user is filtered out and the method reduces to standard empirical mean estimation. This ensures our gradient estimates is unbiased in the "good" case, which is essential for obtaining optimal utility. This adaptive nature allows us to achieve both robustness against adversarial users and optimality for well-behaved data.

From an analysis perspective, one challenge is that our method combines SVT and SGD, which in turn requires combining the stability guarantee of SGD with a "DP-style" guarantee of SVT. To accomplish this, we introduce a novel robust shifted approximate max divergence that enables us to combine these two guarantees. This framework allows us to track stability of all intermediate iterates without privatizing each step, and apply the Gaussian mechanism directly to the final averaged iterate to obtain DP guarantees. This significantly departs from prior approaches that either sacrifice optimality by privatizing all iterates or sacrifice efficiency by using super-linear-time procedures.

3. We use $\tilde{O}(nm)$ gradients, resulting in a total running time of $\tilde{O}(nmd/\varepsilon^3)$. The bottleneck is in computing the concentration score, which takes $\tilde{O}(md/\varepsilon^3)$ time for each step by our current parameter setting. With standard (and tedious) subsampling tricks (Kumar et al., 2025), the total running time can be reduced to $\tilde{O}(nmd)$.

Connection to Robust Statistics. Our approach of filtering users based on concentration draws inspiration from the rich literature on high-dimensional robust statistics. The foundational work of Dwork and Lei (2009) first established the connection between robust statistics and DP using the propose-test-release framework. In the context of robust mean estimation, outlier removal techniques—such as that of Tsfadia et al. (2022); Ashtiani and Liaw (2022)—have been highly effective in identifying subsets of clean data with bounded covariance. There has also been significant progress on (user-level) private robust mean estimation, e.g., (Liu et al., 2021; Hopkins et al., 2022; Narayanan et al., 2022; Agarwal et al., 2025; Zhao et al., 2024).

However, directly applying these high-dimensional robust estimators as "plug-in" subroutines in the inner loop of user-level DP-SGD is challenging. First, sophisticated robust estimators (e.g., those requiring spectral analysis or iterative filtering) often incur super-linear computational costs, which is prohibitive when repeated at every SGD step. Second, and more importantly, maintaining the iterative stability of the optimization trajectory is difficult with standard robust estimators without privatizing every step, which leads to suboptimal utility due to noise accumulation. Instead, our concentration-based filtering can be viewed as a lightweight, scalar-adaptive variant of these principles. By integrating SVT, we achieve a similar goal—removing data that violates local concentration geometry—but we do it in a computationally efficient way (linear-time), while rigorously controlling the sensitivity of the entire optimization path.

Finally, we also explore how classical robust statistics can be applied directly in the ℓ_∞-setting through coordinate-wise median and trimmed mean steps, and provide information-theoretic lower bounds (see Appendix C). While these results require stronger assumptions—specifically, diagonal dominance of Hessians to ensure contractivity of gradient descent in the ℓ_∞-norm—they provide more insights into the fundamental connections between robust statistics and DP. We hope this motivates further research into leveraging such tools for private optimization in other settings.

2. Preliminaries

Let $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a loss function, where $\mathcal{X}$ is the parameter space and $\mathcal{Z}$ is the domain of the dataset. In user-level DP-SCO, we are given a dataset $\mathcal{D} = \{Z_i\}_{i\in[n]}$ of n users, where $Z_i \in \mathcal{Z}^m$ is the user i's data consisting of m functions of the form $f(\cdot; z)$ where $z \in \mathcal{Z}$ is drawn i.i.d. from an (unknown) underlying distribution $\mathcal{P}$. The objective is to minimize the following expected population loss under the user-level DP constraint: $F(x) := \mathbb{E}_{z\sim\mathcal{P}} f(x; z)$. Let $x^* = \arg\min_{x\in\mathcal{X}} F(x)$. We now present some key definitions that we use (more in Appendix A).

Definition 1 (Divergence Measures) *For simplicity, let μ and ν be distributions on $\mathbb{R}^d$, let $\Gamma(\mu, \nu)$ be the collection of couplings of μ and ν, and let $\mu * \nu$ denote the convolution of μ and ν.*

- ***TV Distance:*** *The* total variation (TV) distance *between μ and ν is defined as* $d_{\mathrm{tv}}(\mu, \nu) = \sup_{S\subseteq\mathbb{R}^d} |\mu(S) - \nu(S)|$.
- ***∞-Wasserstein Distance:*** *The* ∞-Wasserstein distance *between μ and ν is defined as* $W_\infty(\mu, \nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \operatorname*{ess\,sup}_{(x,y)\sim\gamma} \|x - y\|_2$.
- ***Approximate Max Divergence:*** *The* δ-approximate max divergence *between μ and ν is defined as* $D^\delta_\infty(\mu \parallel \nu) = \max_{S\subseteq\mathrm{supp}(\mu),\mu(S)\geq\delta} \left\{ \log \frac{\mu(S)-\delta}{\nu(S)} \right\}$.

We use $D^\delta_\infty(\mu, \nu)$ to denote $\max\{D^\delta_\infty(\mu \parallel \nu), D^\delta_\infty(\nu \parallel \mu)\}$. Note that a mechanism $\mathcal{M}$ is (ε, δ)-DP if and only if $D^\delta_\infty(\mathcal{M}(\mathcal{D}), \mathcal{M}(\mathcal{D}')) \leq \varepsilon$ for every pair $\mathcal{D}, \mathcal{D}'$ of neighboring datasets.

We also introduce robust versions of these measures that will be crucial in our framework. The robust Wasserstein distance has recently appeared in various forms, e.g., in distributionally robust estimation (Nietert et al., 2023). We introduce the robust shifted approximate max divergence, which is a novel notion specific for our framework but may be of independent interest.

Definition 2 (Robust Divergence Measures) *Let $\zeta \geq 0$ be a tolerance level.*

- ***Robust Wasserstein Distance:*** *The* ζ-robust ∞-Wasserstein distance *between μ and ν is given by* $W_\infty^\zeta(\mu, \nu) := \inf_{\mu' : d_{\mathrm{tv}}(\mu,\mu') \leq \zeta} W_\infty(\mu', \nu)$.
- ***Robust Shifted Approx. Max Divergence:*** *For $b \geq 0$, the* ζ-robust b-shifted δ-approximate max divergence *between μ and ν is defined as* $D_\infty^{\delta,(b,\zeta)}(\mu, \nu) := \inf_{\mu' : W_\infty^\zeta(\mu,\mu') \leq b} D_\infty^\delta(\mu', \nu)$.

Recall the Gaussian mechanism (Proposition 16): if $W_\infty(\mu, \nu) \leq b$ and if $\sigma \geq \frac{b\sqrt{2\log(1.25/\delta)}}{\varepsilon}$, then $D_\infty^\delta(\mu * \rho, \nu * \rho) \leq \varepsilon$ for $\rho = \mathcal{N}(0, \sigma^2 I_d)$. A convenient property is that we can still use the Gaussian mechanism to privatize an algorithm, if its outputs have bounded shifted approximate max divergence with respect to neighboring datasets. This is captured next.

Lemma 3 *Let $\rho = \mathcal{N}(0, \sigma^2 I_d)$ and for $\varepsilon \in (0, 1]$, suppose $D_\infty^{\delta,(b,\zeta)}(\mu, \nu) \leq \varepsilon$. If $\sigma \geq \frac{b\sqrt{2\log(1.25/\delta)}}{\varepsilon}$, then we have $D_\infty^{(1+e^\varepsilon)\delta+\zeta}(\mu * \rho, \nu * \rho) \leq 2\varepsilon$.*

Optimization. For $X \in \mathbb{R}^d$, we use $X[i]$ to denote its ith coordinate. For a convex set $\mathcal{X} \subset \mathbb{R}^d$, let $\Pi_{\mathcal{X}}(X) := \arg\min_{Y \in \mathcal{X}} \|Y - X\|_2$. For $r \in \mathbb{R}_{\geq 0}$, let $\mathbf{B}(X, r)$ (resp., $\mathbf{B}_\infty(X, r)$) denote the ℓ_2-ball (resp., ℓ_∞-ball) centered at X of radius r respectively. We recall standard optimization notions.

Definition 4 (Lipschitzness and Smoothness) *We say a function $f : \mathcal{X} \to \mathbb{R}$ is* G-Lipschitz *if, for any $x, y \in \mathcal{X}$, we have $|f(x) - f(y)| \leq G\|x - y\|_2$; this means $\|\nabla f(x)\|_2 \leq G$ for any $x \in \mathcal{X}$. We say a function f is* β-smooth *if the operator norm of the Hessian, $\|\nabla^2 f(x)\|_2 \leq \beta$ for any $x \in \mathcal{X}$, where we assume that the objective function is twice differentiable with respect to its first argument.*

The following facts are well-known in the convex optimization literature, see, e.g., (Nesterov, 2013).

Fact 5 *Let $f : \mathcal{X} \to \mathbb{R}$ be convex and β-smooth and let $\eta \leq 2/\beta$. For any $x, y \in \mathcal{X}$, we have $\|(x - \eta\nabla f(x)) - (y - \eta\nabla f(y))\|_2 \leq \|x - y\|_2$.*

Fact 6 *If $\mathcal{X}$ is convex and compact, then for any $x, y \in \mathbb{R}^d$, $\|\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(y)\|_2 \leq \|x - y\|_2$.*

3. Algorithms for the Euclidean Norm

In this section, we present a linear-time algorithm for user-level DP-SCO with optimal error rate in the Euclidean norm. We make the following assumption, which is standard in the literature.

Assumption 7 *Each function $f(\cdot; z) : \mathcal{X} \to \mathbb{R}$ is convex, G-Lipschitz and β-smooth (Definition 4), where $\mathcal{X}$ is a convex compact domain of radius D in ℓ_2-norm.*

Throughout this section, we also assume that $\varepsilon \in (0, 1)$ and the number of users (n) and the number of items per user (m) satisfy the following constraints: $n \geq \frac{300\log^3(mn/\delta)}{\varepsilon^{3/2}}$ and $m \leq n^{O(\log\log n)}$. Our linear-time algorithm is based on a localization framework (Algorithm 1), which in turn uses a modification of SGD with our novel gradient estimation method (Algorithm 2).

We further assume that the parameters in the algorithms are chosen in the following manner:

- $\beta \le \tilde{O}\left(\frac{G}{D} \cdot \sqrt{\frac{\varepsilon}{m}} \cdot \left(\varepsilon\sqrt{n} + \sqrt{d}\right)\right)$; recall that we assumed $f(\cdot; z)$ is β-smooth.
- $B := 300\frac{\log^2(mn/\delta)}{\varepsilon^{3/2}}$; B will be the user *batch size* in the gradient estimation method.
- $\eta := \frac{D}{G} \cdot B\sqrt{m} \cdot \min\left\{\frac{1}{\sqrt{n}}, \frac{\varepsilon}{\sqrt{d\log(1/\delta)\log(nmd)}}\right\}$; η will be the *starting learning rate* that will be decreased (exponentially) every time the gradient estimation method is invoked.
- $\tau := \frac{\sqrt{m}}{20G\log(nmd)}$; τ will be the *temperature* in our concentration test.

Under these assumptions, our main result is the following.

Theorem 8 *With the above assumptions, Algorithm 1 is (ε, δ)-user-level-DP. When the nm functions in dataset $\mathcal{D}$ are sampled i.i.d. from the underlying distribution $\mathcal{P}$, the algorithm performs nm gradient computations and outputs x_S such that $\mathbb{E}[F(x_S) - F(x^*)] \le \tilde{O}\left(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\varepsilon\sqrt{m}}\right)$.*

3.1. Overview of the Algorithms

Localization. We first describe the localization framework in Algorithm 1, which follows Feldman et al. (2020). The algorithm runs in logarithmically many phases, where the size of the dataset decreases geometrically in each phase. In the first phase, starting from an initial point x_0, it invokes (a non-private) SGD using half of the dataset, and averages the iterates to obtain $\overline{x}_1$. Intuitively, the solution $\overline{x}_1$ already provides a good approximation with a small population loss when the datasets are drawn i.i.d. from the underlying distribution; however it is not private. To address this, we add noise ζ_1 that depends on the sensitivity bound on $\|\overline{x}_1\|$ to obtain the privatized version $x_1 \leftarrow \overline{x}_1 + \zeta_1$.

A naive bound on the excess loss due to the privatization is given by $\mathbb{E}[F(x_1) - F(\overline{x}_1)] \le G\|\zeta_1\|_2$. While the magnitude of the noise $\|\zeta_1\|_2$ is typically too large to achieve a good utility guarantee outright, nevertheless, this process yields a much better initial point x_1 compared to the original starting point x_0. Consequently, a smaller dataset and a smaller step size are sufficient to find the next good solution $\overline{x}_2$ in expectation, with smaller noise $\|\zeta_2\|_2$ added to privatize $\overline{x}_2$.

This process is repeated over $O(\log n)$ phases, where each subsequent solution $\overline{x}_s$ is progressively refined, and the Gaussian noise $\|\zeta_s\|_2$ becomes negligible, ultimately balancing both privacy and utility. The privacy guarantee is in Section 3.2.2 and the utility guarantee is in Section 3.3.2.

Algorithm 1: User-level DP-SCO.

Input: Number of users n, Number of items per user m; Privacy parameters ε, δ; Other parameters η, τ, B; Initial point $x_0 \in \mathbb{R}^d$

$S \leftarrow \lceil \log(n/B) \rceil$

for $s = 1, \ldots, S$ **do**

- $n_s \leftarrow n/2^s$, $\eta_s \leftarrow \eta/\log^s m$
- $\mathcal{D}_s \leftarrow$ dataset of size n_s drawn from the unused users, i.e., $\mathcal{D}_s \in \mathcal{Z}^{n_s \times m}$
- $\overline{x}_s \leftarrow$ Algorithm 2 with inputs $(\mathcal{D}_s; \varepsilon, \delta; \eta_s, \tau, B; x_{s-1})$
- $\zeta_s \sim \mathcal{N}(0, \sigma_s^2 I_d)$, where $\sigma_s = O\left(\frac{\eta_s G\sqrt{\log(1/\delta)\log(nmd)}}{B\sqrt{m}\varepsilon}\right)$
- $x_s \leftarrow \overline{x}_s + \zeta_s$

end

return x_S

Modified Robust SGD. Our main new contribution, Algorithm 2, is a modified version of the vanilla SGD. As already explained in Section 1.3, this algorithm uses the novel gradient estimation sub-procedure. For each batch, this involves computing a low-sensitivity score indicating how close the user's (average) gradient is to the other users' (average) gradients (Line 2) and using the SVT to filter out the users that have scores below a certain threshold (Line 2). The remaining users' gradients are then averaged and used as the gradient estimate (Line 2). Analyzing the algorithm's "stability" is a crucial step in our overall privacy proof (Section 3.2.1).

Algorithm 2: SGD for User-level DP-SCO.

Input: Dataset $\mathcal{D}$; Privacy parameters ε, δ; Other parameters η, τ, B; Initial point $x_0 \in \mathbb{R}^d$

$T \leftarrow |\mathcal{D}|/B, c \leftarrow \frac{10\log(nm/\delta)}{\varepsilon}, \widehat{c} \leftarrow c + \mathrm{Lap}(4/\varepsilon), \sigma \leftarrow \frac{4\sqrt{32\widehat{c}\log(1/\delta)}}{\varepsilon}$, count $\leftarrow 0$

$\digamma \leftarrow 3B/4$

Divide $\mathcal{D}$ into T disjoint subsets of equal size B: $\mathcal{D} = \sqcup_{t\in[T]}\mathcal{D}_t$ and $\mathcal{D}_t = \{Z_{t,i}\}_{i\in[B]}$

for $t = 1, \ldots, T$ **do**

- $\widehat{\mathcal{D}}_t \leftarrow \mathcal{D}_t$
- $q_t(Z) \leftarrow \frac{1}{m}\sum_{z\in Z}\nabla f(x_{t-1}; z)$, for each $Z \in \mathcal{D}_t$
- **for** $i = 1, \ldots, B$ **do**
  - $q_t(Z_{t,i}) \leftarrow \frac{1}{m}\sum_{z\in Z_{t,i}}\nabla f(x_{t-1}; z)$
  - $\kappa_{t,i} \leftarrow \sum_{Z\in\mathcal{D}_t}\exp(-\tau\cdot\|q_t(Z_{t,i}) - q_t(Z)\|_2)$ ▷ Concentration scores
  - $\nu_{t,i} \sim \mathrm{Lap}(2\sigma)$
  - **if** $\kappa_{t,i} + \nu_{t,i} \le \digamma$ **then**
    - $\widehat{\mathcal{D}}_t \leftarrow \widehat{\mathcal{D}}_t \setminus \{Z_{t,i}\}$ ▷ Filter $Z_{t,i}$ out
    - count $\leftarrow$ count $+ 1$
    - $a_{t,i} \leftarrow \perp$ ▷ $a_{t,i}$'s are used only in the analysis
  - **end**
  - **else**
    - $a_{t,i} \leftarrow \top$
  - **end**
  - **if** count $\ge \widehat{c}$ or $\widehat{\mathcal{D}}_t = \emptyset$ **then**
    - **halt**
  - **end**
- **end**
- $g_{t-1} \leftarrow \frac{1}{|\widehat{\mathcal{D}}_t|}\sum_{Z_{t,i}\in\widehat{\mathcal{D}}_t} q_t(Z_{t,i})$
- $x_t \leftarrow \Pi_{\mathcal{X}}(x_{t-1} - \eta g_{t-1})$

end

return $\overline{x} \leftarrow \frac{1}{T}\sum_{t\in[T]} x_t$

In the standard SVT framework, privacy requires obfuscating the threshold in addition to noising the queries. To simplify the presentation of our algorithm and its analysis, we utilize the target-charging technique (TCT) proposed by Cohen and Lyu (2023). By shifting to this framework, the queries become more decoupled, enabling us to maintain strict DP guarantees (as stated in Lemma 18) even with a fixed, unnoised threshold.

3.2. Privacy Analysis

3.2.1. STABILITY OF ALGORITHM 2

We now analyze the "stability" of Algorithm 2. A main challenge is that, due to the use of SVT, the algorithm is not stable in the usual sense (Feldman et al., 2020) in that its outputs on two neighboring datasets are not too different. Instead, we analyze its stability via the robust shifted approximate max divergence notion (Definition 2); the privacy of Algorithm 1 will then follow from Lemma 3.

Let $(\mathcal{D}, \mathcal{D}')$ be neighboring datasets that differ by one user. Without loss of generality, assume that $Z_{1,1} \neq Z'_{1,1}$ represents the differing user between the two datasets. We use primed notation for quantities computed by Algorithm 2 on $\mathcal{D}'$: $\{x'_t\}_{t\in[T]}$ denote the iterates, $\{\widehat{D}'_t\}_{t\in[T]}$ denote the sets of users passing the filter at each step, $\widehat{c}'$ denotes the privatized cutoff point, $\{\kappa'_{t,i}\}_{t\in[T],i\in[B]}$ denote the concentration scores, and $\{a'_{t,i}\}_{t\in[T],i\in[B]}$ denote the filtering decisions.

Conditioned on the $\{a_{t,i}\}$'s, we show that the algorithm is stable from the second step onward.

Lemma 9 *For any $t \geq 2$, suppose $a_{t,i} = a'_{t,i}, \forall i \in [B]$, then $\|x_t - x'_t\|_2 \leq \|x_{t-1} - x'_{t-1}\|_2$.*

Proof Since $a_{t,i} = a'_{t,i}$ for all $i \in [B]$, we have $\widehat{D}_t = \widehat{D}'_t$. Let g'_{t-1} be the gradient estimate corresponding to the dataset $\mathcal{D}'$. Then by Fact 5 and Assumption 7, we have $\|(x_{t-1} - \eta g_{t-1}) - (x'_{t-1} - \eta g'_{t-1})\|_2 \leq \|x_{t-1} - x'_{t-1}\|_2$. Combining it with Fact 6, we obtain that $\|x_t - x'_t\|_2 \leq \|x_{t-1} - x'_{t-1}\|_2$. ■

With Lemma 9, we analyze the similarity between $\overline{x}$ and $\overline{x}'$ by considering the similarity between x_1 and x'_1, and the closeness between $\{a_{t,i}\}_{t\geq 1}$ and $\{a'_{t,i}\}_{t\geq 1}$ respectively. To formalize this analysis, we use the robust divergence measures to capture the similarity between $\overline{x}$ and $\overline{x}'$.

The key technical novelty is that, for any $t \in [T]$, we can control the similarity between $\{x_t\}$ and $\{x'_t\}$ as long as the number of "bad" users filtered out does not exceed the cutoff point, say the privatized cutoff points $\widehat{c}$ and $\widehat{c}'$ in the algorithm. We consider the similarity between x_1 and x'_1 first.

Lemma 10 *If $b = O\left(\frac{\eta}{B\tau}\right)$, then $D_\infty^{\delta/4,(b,\delta/4)}(x_1, x'_1) \leq \varepsilon/4$.*

Proof Recall that we consider neighboring datasets that differ only in the first user's data $Z_{1,1}$. We have control over the sensitivity of queries $\{\kappa_{1,i}\}_{i\geq 2}$ but not for $\kappa_{1,1}$. Specifically, we have $|\kappa_{1,i} - \kappa'_{1,i}| \leq 1$ for any $i \geq 2$. As we add noise to privatize the cutoff point $\widehat{c}$, and by the property of SVT (Dwork and Roth, 2014, Theorem 3.25), we have $D_\infty^{\delta/4}(\{a_{1,i}\}_{i\geq 2}, \{a'_{1,i}\}_{i\geq 2}) \leq \varepsilon/4$. Define the following (high probability) events:

- $\mathcal{E}_1$: "$\widehat{c} \leq B/2$ and $\widehat{c}' \leq B/2$". Concentration of $\mathrm{Lap}(4/\varepsilon)$ yields $\Pr[\mathcal{E}_1] \geq 1 - \delta/24$.
- $\mathcal{E}_2$: "$\forall i \geq 1$, $|\nu_{1,i}| \leq \frac{20\log^2(mn/\delta)}{\varepsilon^{3/2}}$ and $|\nu'_{1,i}| \leq \frac{20\log^2(mn/\delta)}{\varepsilon^{3/2}}$". Again, concentration of $\mathrm{Lap}(2\sigma)$ yields $\Pr[\mathcal{E}_2] \geq 1 - \delta/24$.

The rest of the proof is conditioned on both events occurring. Given $\mathcal{E}_1$, we have $|\widehat{D}_1| \geq B/2$ and $|\widehat{D}'_1| \geq B/2$, as the total number of filtered users is smaller than the cutoff points. Moreover, for any $i \geq 2$ such that $Z_{1,i}$ is not filtered out, given $\mathcal{E}_2$, we know that $\kappa_{1,i} \geq 3B/5$ as $\kappa_{1,i} + \nu_{1,i} \geq$ [illegible].

Let $\tilde{q}_1(Z_{1,1}) = \Pi_{\mathbf{B}(q_1(Z_{1,i}),100/\tau)}(q_1(Z_{1,1}))$ be the projection of the average gradient of user $Z_{1,1}$ towards the ball centered at the average gradient $q_1(Z_{1,i})$. Let $\tilde{x}_1$ be the output if we replace $q_1(Z_{1,1})$ by $\tilde{q}_1(Z_{1,1})$ if $Z_{1,1}$ is not filtered out, and keep the remaining the same.

We next argue that $d_{\mathrm{tv}}(x_1, \tilde{x}_1) \leq \delta/8$. This vacuously holds when $\tilde{q}_1(Z_{1,1}) = q_1(Z_{1,1})$. We henceforth consider only the non-trivial case when $\tilde{q}_1(Z_{1,1}) \neq q_1(Z_{1,1})$, i.e., $\|q_1(Z_{1,1}) -$

$q_1(Z_{1,i})\|_2 > 100/\tau$. Since $\kappa_{1,i} \geq 3B/5$, the neighboring index set $N_i = \{j \in [B] : \|q_1(Z_{1,i}) - q_1(Z_{1,j})\|_2 \leq 4/\tau\}$, has size $|N_i| \geq B/2$. Then we know for any $j \in [B]$ and $j \neq 1$, we have

$$\|q_1(Z_{1,1}) - q_1(Z_{1,j})\|_2 \geq \|q_1(Z_{1,1}) - q_1(Z_{1,i})\|_2 - \|q_1(Z_{1,i}) - q_1(Z_{1,j})\|_2 \geq 100/\tau - 4/\tau = 96/\tau,$$

which implies $\kappa_{1,1} \leq 2B/3$. Similarly, the concentration score with respect to $\tilde{q}_1(Z_{1,1})$ is also smaller than $2B/3$. Hence, conditioned on events $\mathcal{E}_1$ and $\mathcal{E}_2$, $Z_{1,1}$ is filtered out and hence $x_1 = \tilde{x}_1$, which implies

$$d_{\mathrm{tv}}(x_1, \tilde{x}_1) \leq \delta/8. \tag{3}$$

Define $\overline{x}_1$ as the output when we deterministically filter out $Z_{1,1}$ regardless of its concentration score. Fixing the values of $\{a_{1,i}\}_{i\geq 2}$, consider the non-trivial case when $a_{1,1} = \top$,

$$\begin{aligned}\|\tilde{x}_1 - \overline{x}_1\|_2 \leq& \eta \cdot \left\| \frac{1}{|\widehat{D}_t|}(\tilde{q}_1(Z_{1,1}) + \sum_{Z_{1,i}\in\widehat{D}_t, i\geq 2} q_1(Z_{1,i})) - \frac{1}{|\widehat{D}_t| - 1} \sum_{Z_{1,i\in\widehat{D}_t}, i\geq 2} q_1(Z_{1,i}) \right\| \\ =& \eta \cdot \left\| \frac{1}{|\widehat{D}_t|}\tilde{q}_1(Z_{1,1}) - \frac{1}{|\widehat{D}_t|(|\widehat{D}_t| - 1)} \sum_{Z_{1,i\in\widehat{D}_t}, i\geq 2} q_1(Z_{1,i}) \right\| \lesssim \frac{\eta}{B\tau},\end{aligned}$$

where we use the fact that $\|\tilde{q}_1(Z_{1,1}) - q_1(Z_{1,i})\| \lesssim 1/\tau$ and $|\widehat{D}_t| \geq B/2$. Hence we can construct a coupling between $\tilde{x}_1, \overline{x}_1$, which means

$$W_\infty(\tilde{x}_1, \overline{x}_1) \leq O\left(\frac{\eta}{B\tau}\right) =: b. \tag{4}$$

We can do a similar analysis for x'_1 and define $\tilde{x}'_1$ and $\overline{x}'_1$ correspondingly. By definition, $\overline{x}_1$ and $\overline{x}'_1$ are functions of $\{a_{1,i}\}_{i\geq 2}$ and $\{a'_{1,i}\}_{i\geq 2}$. Then we know

$$D_\infty^{\delta/4}(\overline{x}_1, \overline{x}'_1) \leq D_\infty^{\delta/4}(\{a_{1,i}\}_{i\geq 2}, \{a'_{1,i}\}_{i\geq 2}) \leq \varepsilon/4. \tag{5}$$

From Equations (3), (4), and (5), by definition, we have proved that $D_\infty^{\delta/4,(b,\delta/4)}(x_1, x'_1) \leq \varepsilon/4$. ■

As shown in Algorithm 1, we will add Gaussian noise to privatize $\overline{x}$. By Lemma 3 it suffices to work with the shifted approximate max divergence between $\overline{x}$ and $\overline{x}'$ to prove privacy.

Having established the similarity between x_1 and x'_1, we now establish the query sensitivity to show the similarity between queries, which leads to the similarity between $\{a_{t,i}\}$ and $\{a'_{t,i}\}$.

Lemma 11 (Query Sensitivity) *For any $t \geq 2$, suppose $\beta C\eta \leq 1$ and $\|x_{t-1} - x'_{t-1}\| \leq \frac{C\eta}{B\tau}$ for some $C \geq 0$ Then, for all $i \in [B]$, $|\kappa_{t,i} - \kappa'_{t,i}| \leq 2$.*

Proof Since $t \geq 2$, we have $\mathcal{D}_t = \mathcal{D}'_t$. By the β-smoothness assumption, for $Z \in \mathcal{D}_t$, we have

$$\begin{aligned}&\|q_t(Z) - q_t(Z_{t,i})\|_2 - \|q'_t(Z) - q'_t(Z_{t,i})\|_2 \leq \|(q_t(Z_{t,i}) - q'_t(Z_{t,i})) - (q_t(Z) - q'_t(Z))\|_2 \\ \leq& \|q_t(Z_{t,i}) - q'_t(Z_{t,i})\|_2 + \|q_t(Z) - q'_t(Z)\|_2 \leq 2\beta\|x_{t-1} - x'_{t-1}\|_2 \leq 2\beta\frac{C\eta}{B\tau}.\end{aligned}$$

Hence,

$$\kappa_{t,i} = \sum_{Z \in \mathcal{D}_t} e^{-\tau \|q_t(Z_{t,i}) - q_t(Z)\|_2} \geq \sum_{Z \in \mathcal{D}'_t} e^{-\tau \|q'_t(Z_{t,i}) - q'_t(Z)\|_2)} \cdot e^{-2\beta \frac{C\eta}{B}} \geq \kappa'_{t,i} \cdot e^{-2\beta \frac{C\eta}{B}}.$$

As both $\kappa_{t,i}$ and $\kappa'_{t,i}$ are in the range $[0, B]$, using $1 - e^{-x} \leq x$ for $x \geq 0$, we obtain

$$\kappa'_{t,i} - \kappa_{t,i} \leq \kappa_{t,i} \left(1 - e^{-2\beta \frac{C\eta}{B}}\right) \leq 2\beta C\eta \leq 2.$$

Similarly, we can bound $\kappa_{t,i} - \kappa'_{t,i} \leq 2$ and complete the proof. ■

We now complete the proof of robust shifted approximate max divergence between $\overline{x}$ and $\overline{x}'$.

Lemma 12 *Under Assumption 7, if* $b = O\left(\frac{\eta}{B\tau}\right)$ *and* $\beta\eta \leq 2$, *then* $D_\infty^{\delta/2,(b,\delta/4)}(\overline{x}, \overline{x}') \leq \varepsilon/2$.

Proof By Lemma 10, we know $D_\infty^{\delta/4,(b,\delta/4)}(x_1, x'_1) \leq \varepsilon/4$. This means there exists a random vector y_1 such that (i) $W_\infty^{\delta/4}(x_1, y_1) \leq b$ and (ii) $D_\infty^{\delta/4}(x'_1, y_1) \leq \varepsilon/4$.

Let $y_1, \ldots, y_T$ be the sequence of iterates generated by the algorithm when starting from y_1 instead of x_1, and $\{\tilde{a}_{t,i}\}$ be the corresponding filtering decisions. As the neighboring datasets differ in the user $Z_{1,1}$, by the post-processing property of DP, we have $D_\infty^{\delta/4}(\{x'_1, \ldots, x'_T\}, \{y_1, \ldots, y_T\}) \leq D_\infty^{\delta/4}(x'_1, y_1) \leq \varepsilon/4$, from (ii). If $\overline{y}$ is the average over $\{y_1, \ldots, y_T\}$, this immediately leads to

$$D_\infty^{\delta/4}(\overline{x}', \overline{y}) \leq \varepsilon/4, \tag{6}$$

Now we analyze the similarity between $\{y_1, \ldots, y_T\}$ and $\{x_1, \ldots, x_T\}$. Suppose we fix x_1 and y_1 such that $\|x_1 - y_1\|_2 \leq b$, where $b = O(\eta/B\tau)$ satisfies the precondition. Lemma 9 establishes the contractivity of gradient descent (for which we need $\beta\eta \leq 2$), and Lemma 11 controls the query sensitivities when $\|x_t - y_t\|_2 \leq b$. By the properties of SVT (Lemma 18), we have $D_\infty^{\delta/4}(\{a_{t,i}\}_{t\geq 2}, \{\tilde{a}_{t,i}\}_{t\geq 2}) \leq \varepsilon/4$, hence $D_\infty^{\delta/4}(\overline{x}, \overline{y}) \leq \varepsilon/4$. Combining with (i), we thus obtain

$$D_\infty^{\delta/4,(b,\delta/4)}(\overline{x}, \overline{y}) \leq \varepsilon/4. \tag{7}$$

By the basic composition over (6) and (7), we conclude that $D_\infty^{\delta/2,(b,\delta/4)}(\overline{x}, \overline{x}') \leq \varepsilon/2$. ■

3.2.2. PRIVACY OF ALGORITHM 1

From Lemma 12 and Lemma 3, each iteration in the **for** loop of Algorithm 1 is (ε, δ)-user-level-DP. Since each iteration used disjoint subsets of users, the parallel composition theorem (McSherry, 2010) ensures that Algorithm 1 is also (ε, δ)-DP.

3.3. Utility Analysis

We nextanalyze the utility of both algorithms when the dataset $\mathcal{D}$ is drawn i.i.d. from $\mathcal{P}$.

3.3.1. UTILITY OF ALGORITHM 2

Given that the dataset $\mathcal{D}, |\mathcal{D}| = n$ is i.i.d., no user is filtered out with high probability. This ensures that our gradient estimates remain unbiased, allowing us to directly apply the standard convergence analysis of SGD for smooth convex functions with unbiased gradient estimates at each iteration. In particular, we use the following standard bound:

Lemma 13 (Theorem 6.3, Bubeck (2015)) *Consider a β-smooth convex function f over a convex set $\mathcal{X}$. For any $x \in \mathcal{X}$, suppose that the random initial point x_0 satisfies $\mathbb{E}[\|x_0 - x\|_2^2] \leq R^2$. Suppose we have an unbiased stochastic gradient oracle such that $\mathbb{E}\,\|\tilde{g}(x_t) - \nabla f(x_t)\|_2^2 \leq \sigma_t^2$, then running SGD for T steps with fixed step size η satisfies that*

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x)\right] \leq \left(\beta + \frac{1}{\eta}\right)\frac{R^2}{T} + \frac{\eta\sum_t \sigma_t^2}{2T}.$$

Lemma 14 *For any $x \in \mathcal{X}$, the final output $\overline{x}$ of Algorithm 2 satisfies*

$$\mathbb{E}[F(\overline{x}) - F(x)] \lesssim \left(\beta + \frac{1}{\eta}\right)\frac{\mathbb{E}[\|x_0 - x\|^2]}{T} + \frac{\eta G^2}{Bm} + \frac{GD}{(nm)^3}.$$

Proof For simplicity, let $\omega = 1/(nm)^3$. We define the following (high-probability) events:

- $\mathcal{E}_1$: "$\forall t \in [T], i \in [B]$, $\|q_t(Z_{i,t}) - \nabla F(x_{t-1})\|_2 \leq 1/(20\tau)$." By the Hoeffding inequality for norm-subGaussian vectors (Theorem 20), for each $t \in [T]$ and $i \in [B]$, we have

$$\Pr\left[\|q_t(Z_{t,i}) - \nabla F(x_{t-1})\|_2 \geq \frac{1}{20\tau} = G\log(ndm/\omega)/\sqrt{m}\right] \leq \frac{\omega}{nm}.$$

  Hence, $\Pr[\mathcal{E}_1] \geq 1 - \omega$ by a union bound.
- $\mathcal{E}_2$: "$\forall t \in [T], i \in [B], |\nu_{t,i}| \leq 20\log^{3/2}(mn/\delta)/\varepsilon^{3/2}$." Indeed, from the concentration of $\mathrm{Lap}(2\sigma)$ added to $\kappa_{t,i}$, we have $\Pr[\mathcal{E}_2] \geq 1 - \omega$.

Let $f_{t,i}$ denote the number of close neighbors for user i in batch t, that is

$$f_{t,i} = \sum_{Z \in \mathcal{D}_t} \mathbf{1}(\|q_t(Z_{t,i} - q_t(Z)\|_2 \leq 1/20\tau).$$

Conditioned on these events, which all hold by our setting of parameters, we know that $f_{t,i} \geq 0.9B$, and each user passes the concentration tests with $a_{t,i} = \top, \forall t \in [T], i \in [B]$, and that $g_{t-1} = \frac{1}{B}\sum_{i\in[B]} q_t(Z_{t,i})$. This implies $d_{\mathrm{tv}}\left(\{g_{t-1}\}_{t\in[T]}, \{\frac{1}{B}\sum_{i\in[B]} q_t(Z_{t,i})\}_{t\in[T]}\right) \leq \omega$. Note that $\mathbb{E}[\frac{1}{B}\sum_{i\in[B]} q_t(Z_{t,i})] = \nabla F(x_{t-1})$ and $\mathbb{E}[\|\frac{1}{B}\sum_{i\in[B]} q_t(Z_{t,i}) - \nabla F(x_{t-1})\|_2^2] \leq \frac{G^2}{Bm}$ when all functions are drawn i.i.d. from the distribution. Since the TV distance between g_{t-1} and the true gradient estimate $\frac{1}{B}\sum_{i\in[B]} q_t(Z_{t,i})$ is small, we can appeal to Lemma 13 to conclude the proof, where the last term comes from the worst value GD, and the small failure probability ω. ■

3.3.2. UTILITY OF ALGORITHM 1

Now we apply the localization framework to establish the utility claim in Theorem 8; the proof appeals to Lemma 14 and mostly follows from Feldman et al. (2020).

Let $\overline{x}_0 = x^*$ and $\zeta_0 = x_0 - x^*$; by the assumption on $\mathcal{X}$, we know $\|\zeta_0\|_2 \le D$. Since we add Gaussian noise ζ_s to $\overline{x}_s$ in each phase, we analyze the influence of ζ_s first. From our choice of η, for all $s \ge 0$, $\mathbb{E}[\|\zeta_s\|_2^2] = d\sigma_s^2 = \eta_s^2 d\frac{G^2\log(1/\delta)\log(nmd)}{B^2m\varepsilon^2} \le D^2 \cdot \log^{-s} m$.

Since our choice of parameters and the precondition ensure that $\eta\beta \le O(1)$, by Lemma 14,

$$\begin{aligned}
\mathbb{E}[F(x_S)] - F(x^*) &= \sum_{s=1}^{S} \mathbb{E}[F(\overline{x}_s) - F(\overline{x}_{s-1})] + \mathbb{E}[F(x_s) - F(\overline{x}_s)] \\
&\le \sum_{s=1}^{S} \left(\frac{\mathbb{E}[\|\zeta_{s-1}\|_2^2]}{\eta_s T_s} + \frac{\eta_s G^2}{2Bm} \right) + G\,\mathbb{E}[\|\zeta_S\|_2] \\
&\le \sum_{s=1}^{S} \left(\frac{\log m}{2} \right)^{-(s-2)} \left(\frac{D^2}{\eta n/B} + \frac{\eta G^2}{2Bm} \right) + \frac{GD}{(\log m)^{O(\log n)}} \\
&\le \tilde{O}\Big(GD\big(\frac{1}{\sqrt{nm}} + \frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\big) \Big),
\end{aligned}$$

where we use the fact that $\frac{1}{(\log m)^{O(\log n)}} \le \frac{1}{nm}$ when $m \le n^{O(\log\log n)}$.

4. Conclusions

In this work, we present the first nearly linear-time algorithm for user-level DP-SCO for ℓ_2- and ℓ_∞-norms (in the Appendix), achieving optimal excess risk rates, thereby resolving the trade-off between computational efficiency and utility. Our approach leverages a novel adaptive outlier removal framework that integrates the sparse vector technique directly into SGD.

4.1. Open Problems

A natural question is whether one could simply employ robust statistics (combined with privatization) to solve this problem. However, in DP optimization—as opposed to one-shot mean estimation—stepwise stability, estimator bias, and per-step noise/utility trade-offs are critical factors. Naively inserting a robust estimator at each step fails to attain optimal rates: privatizing every iterate overwhelms utility (we conjecture $\sim O(1/\sqrt{n})$ or worse), while skipping per-step privacy leaves trajectory stability undefined. In contrast, our analysis establishes stability for all iterates while privatizing only the final output. Within this framework (SVT plus divergence), other robust or DP mean estimators may be readily adapted.

Future research could focus on extending our framework to related settings, particularly by relaxing the strong diagonal dominance assumption currently required for the ℓ_∞-norm case. Additionally, one might explore utilizing the exponential mechanism with Langevin dynamics. While achieving optimal rates for item-level DP-SCO via the (regularized) exponential mechanism (Gopi et al., 2022) is already non-trivial, obtaining optimal rates for user-level DP-SCO remains open.

Besides, an important and challenging open direction in the user-level DP setting is relaxing the homogeneous data assumption (where all users' data are drawn from the same distribution).

References

Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *NIPS*, 2009.

Sushant Agarwal, Gautam Kamath, Mahbod Majid, Argyris Mouzakis, Rose Silver, and Jonathan Ullman. Private mean estimation with person-level differential privacy. In *SODA*, pages 2819–2880, 2025.

Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning Gaussians and beyond. In *COLT*, pages 1075–1076, 2022.

Hilal Asi and Daogao Liu. User-level differentially private stochastic convex optimization: Efficient algorithms with optimal rates. In *AISTATS*, pages 4240–4248, 2024.

Hilal Asi, Daogao Liu, and Kevin Tian. Private stochastic convex optimization with heavy tails: Near-optimality from simple reductions. In *NeurIPS*, 2024.

Raef Bassily and Ziteng Sun. User-level private stochastic convex optimization with optimal rates. In *ICML*, pages 1838–1851, 2023.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*, pages 464–473, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *NIPS*, pages 11282–11291, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *NeurIPS*, 2020.

Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. In *COLT*, pages 474–499, 2021.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security*, pages 2633–2650, 2021.

Edith Cohen and Xin Lyu. The target-charging technique for privacy analysis across interactive computations. In *NeurIPS*, pages 62139–62168, 2023. Also, arXiv:2302.11044.

Stephane Durocher and David Kirkpatrick. The projection median of a set of points. *Computational Geometry*, 42(5):364–375, 2009.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, pages 371–380, 2009.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *STOC*, pages 439–449, 2020.

Badih Ghazi, Pritish Kamath, Ravi Kumar, Raghu Meka, Pasin Manurangsi, and Chiyuan Zhang. On user-level private convex optimization. In *ICML*, pages 11283–11299, 2023.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *COLT*, pages 1948–1989, 2022.

Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *STOC*, pages 1406–1417, 2022.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subGaussian norm. *arXiv*, 1902.03736, 2019.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *COLT*, pages 1853–1902, 2019.

Syamantak Kumar, Daogao Liu, Kevin Tian, and Chutong Yang. Private geometric median in nearly-linear time. In *NeurIPS*, 2025.

Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *NeurIPS*, pages 12466–12479, 2021.

Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In *NeurIPS*, pages 15954–15964, 2020.

Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. In *NeurIPS*, pages 3887–3901, 2021.

Andrew Lowy, Daogao Liu, and Hilal Asi. Faster algorithms for user-level private stochastic convex optimization. In *NeurIPS*, 2024.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *S & P*, pages 346–363, 2023.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *ICML*, pages 2408–2417, 2015.

Frank McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *C. ACM*, 53(9):89–97, 2010.

Shyam Narayanan, Vahab Mirrokni, and Hossein Esfandiari. Tight and robust private mean estimation with few users. In *ICML*, pages 16383–16412, 2022.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

Sloan Nietert, Rachel Cummings, and Ziv Goldfeld. Robust estimation under the Wasserstein distance. *arXiv*, 2302.01237, 2023.

Jinyan Su, Lijie Hu, and Di Wang. Faster rates of private stochastic convex optimization. In *ALT*, pages 995–1002, 2022.

Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *ICML*, pages 21828–21863, 2022.

Hongjian Wang, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. In *NeurIPS*, pages 18866–18877, 2021.

Zheng Xu and Yanxiang Zhang. Advances in private training for production on-device language models. https://research.google/blog/advances-in-private-training-for-production-on-device-language-models/, 2024. Google Research Blog.

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 *NeurIPS*, pages 130018–130056, 2024.

Appendix A. Preliminaries

A.1. Differential Privacy

Definition 15 (User-level DP) *We say a mechanism $\mathcal{M}$ is* (ε, δ)-user-level DP*, if for any neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ from one user, and any output event set $\mathcal{O}$, we have*

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{O}] \leq e^{\varepsilon} \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{O}] + \delta.$$

Proposition 16 (Gaussian Mechanism,Dwork and Roth (2014)) *Consider a function $f : \mathcal{P}^* \to \mathbb{R}^d$. If $0 < \varepsilon \leq 1, \max_{\mathcal{D}\sim\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2 \leq \Delta$, where $\mathcal{D} \sim \mathcal{D}'$ means $\mathcal{D}$ and $\mathcal{D}'$ are neighboring datasets, then the Gaussian mechanism*

$$\mathcal{M}(\mathcal{D}) := f(\mathcal{D}) + \zeta,$$

where $\zeta \sim \mathcal{N}(0, \sigma^2 I_d)$ with $\sigma^2 \geq \frac{2\Delta^2 \log(1.25/\delta)}{\varepsilon^2}$ is (ε, δ)-DP.

A.1.1. ABOVETHRESHOLD

Our algorithms use the AboveThreshold algorithm (Dwork and Roth, 2014), which is a key tool in DP to identify whether there is a query $q_i : \mathcal{Z} \to \mathbb{R}$ in a stream $q_1, \ldots, q_T$ of queries that is above a certain threshold Δ. The AboveThreshold algorithm (Algorithm 3) has the following guarantees:

Lemma 17 (Dwork and Roth (2014), Theorem 3.24) AboveThreshold *is $(\varepsilon, 0)$-DP. Moreover, let $\alpha = \frac{8\log(2T/\gamma)}{\varepsilon}$ and $\mathcal{D} \in \mathcal{Z}^n$. For any sequence $q_1, \cdots, q_T : \mathcal{Z}^n \to \mathbb{R}$ of T queries each of sensitivity 1,* AboveThreshold *halts at time $k \in [T+1]$ such that with probability at least $1-\gamma$,*

- *For all $t < k$, $a_t = \top$ and $q_t(\mathcal{D}) \geq \Delta - \alpha$;*
- *$a_k = \perp$ and $q_k(\mathcal{D}) \leq \Delta + \alpha$ or $k = T+1$.*

A.1.2. THE SPARSE VECTOR TECHNIQUE

The classic SVT algorithm can be viewed as follows. As queries arrive, it makes repeated calls to AboveThreshold. Each time an above threshold query is reported, the algorithm simply restarts the remaining stream of queries on a new instantiation of AboveThreshold. It halts after it has restarted AboveThreshold c times (i.e., after c above threshold queries have appeared). Each instantiation of AboveThreshold is $(\varepsilon, 0)$-DP, and so the composition theorems apply.

Due to some technical reasons, we use a variant of the standard SVT algorithm: we only add noise to the query but use the same fixed threshold, since this makes the queries more independent and easier to analyze; see Algorithm 4.

Lemma 18 (Cohen and Lyu (2023), Theorem 2.10 in arXiv version) *Assume $\varepsilon \leq 1$. Then Algorithm 4 is $O(\sqrt{c\log(1/\delta)} \cdot \varepsilon, 2^{-\Omega(c)} + \delta)$-DP for every $\delta \in (0,1)$.*

In our application of Lemma 18 to bound the filtering decisions across iterations, we set the targeted cutoff parameter $c = O(\log(\frac{mn}{\delta})/\varepsilon)$, while adding $\mathrm{Lap}(2\sigma)$ noise to the queries where $\sigma = O(\sqrt{\log(nm/\delta)}/\varepsilon^{3/2})$. This guarantees that the cumulative privacy cost scales gracefully as $\tilde{O}(\sqrt{c}/\sigma) = \tilde{O}(\varepsilon)$.

A.2. SubGaussian and Norm-SubGaussian Random Vectors

Definition 19 *Let $\zeta > 0$. We say a random vector X is* subGaussian *($\mathrm{SG}(\zeta)$) with parameter ζ if $\mathbb{E}[e^{\langle v, X - \mathbb{E}\, X\rangle}] \le e^{\|v\|^2\zeta^2/2}$ for any $v \in \mathbb{R}^d$. A random vector $X \in \mathbb{R}^d$ is* norm-subGaussian *with parameter ζ ($\mathrm{nSG}(\zeta)$) if $\mathbb{P}[\|X - \mathbb{E}\, X\|_2 \ge t] \le 2e^{-\frac{t^2}{2\zeta^2}}$ for all $t > 0$.*

Theorem 20 (Hoeffding-type inequality for norm-subGaussian, Jin et al. (2019)) *Let $X_1, \cdots, X_k \in \mathbb{R}^d$ be random vectors, and let $\mathcal{F}_i = \sigma(x_1, \cdot, x_i)$ for $i \in [k]$ be the corresponding filtration. Suppose for each $i \in [k]$, $X_i \mid \mathcal{F}_{i-1}$ is zero-mean $\mathrm{nSG}(\zeta_i)$. Then, there exists an absolute constant $c > 0$, for any $\gamma > 0$,*

$$\mathbb{P}\left[\left\|\sum_{i\in[k]} X_i\right\|_2 \ge c\sqrt{\log(d/\gamma)\sum_{i\in[k]}\zeta_i^2}\right] \le \gamma.$$

Algorithm 3: AboveThreshold

Input: Dataset $\mathcal{D}$, threshold $\Delta \in \mathbb{R}$, privacy parameter ε;
Let $\widehat{\Delta} := \Delta - \mathrm{Lap}(\frac{2}{\varepsilon})$;
for $i = 1$ *to* T **do**
  Receive a new query $q_i : \mathcal{Z}^n \to \mathbb{R}$;
  Sample $\nu_i \sim \mathrm{Lap}(\frac{4}{\varepsilon})$;
  if $q_t(\mathcal{D}) + \nu_i < \widehat{\Delta}$ **then**
    Output: $a_i = \perp$;
    Halt;
  end
  else
    Output: $a_i = \top$;
  end
end

Appendix B. Missing Proofs in Section 3

B.1. Proof of Lemma 3

Proof It suffices to show $D_\infty^{(1+e^\varepsilon)\delta+\zeta}(\mu * \rho \parallel \nu * \rho) \le 2\varepsilon$; a similar bound on $D_\infty^{(1+e^\varepsilon)\delta+\zeta}(\nu * \rho \parallel \mu * \rho)$ follows by symmetry.

From the premise and Definition 2, there exists a distribution μ' such that

$$W_\infty^\zeta(\mu, \mu') \le b. \tag{8}$$
$$D_\infty^\delta(\mu', \nu) \le \varepsilon. \tag{9}$$

Applying Definition 2 to (8), let μ'' be the distribution such that

$$d_{\mathrm{tv}}(\mu, \mu'') \le \zeta. \tag{10}$$

Algorithm 4: SVT with Target Privacy Charging

Input: Dataset $\mathcal{D}$, cutoff point c, privacy parameter ε;
Initialize $c_b = 0$ **for** $i = 1, 2, \cdots$ **do**
  Receive a new query $q_i : \mathcal{Z}^n \to \mathbb{R}$ and the threshold Δ_i;
  Sample ν_i from $\mathrm{Lap}(1/\varepsilon)$;
  if $q_t(\mathcal{D}) + \nu_i \geq \Delta_i$ **then**
    Output: $a_i = \top$;
    $c_b = c_b + 1$;
    if $c_b \geq c$ **then**
      Break
    end
  end
  else
    Output: $a_i = \perp$;
  end
end

$$W_\infty(\mu', \mu'') \leq b. \tag{11}$$

Apply the Gaussian Mechanism (Proposition 16) to (11), we have

$$D_\infty^\delta(\mu' * \rho, \mu'' * \rho) \leq \varepsilon. \tag{12}$$

Applying the DP postprocessing property to (9), we have

$$D_\infty^\delta(\mu' * \rho, \nu * \rho) \leq \varepsilon. \tag{13}$$

Then for any set S, we have

$$\begin{aligned}
(\mu * \rho)(S) &\overset{(10)}{\leq} (\mu'' * \rho)(S) + \zeta \\
&\overset{(12)}{\leq} e^\varepsilon (\mu' * \rho)(S) + \delta + \zeta \\
&\overset{(13)}{\leq} e^\varepsilon (e^\varepsilon (\nu * \rho)(S) + \delta) + \delta + \zeta \\
&= e^{2\varepsilon} (\nu * \rho)(S) + (1 + e^\varepsilon)\delta + \zeta,
\end{aligned}$$

which completes the proof. ■

Appendix C. Extension to ℓ_∞-Norm

Having established our main results for the Euclidean norm, we now explore how classical robust statistics can be applied directly to the ℓ_∞-norm. This extension demonstrates the broader applicability of robust statistical principles to private optimization, though it requires stronger assumptions than in the Euclidean norm case.

C.1. Direct Application of Robust Statistics

The ℓ_∞ geometry allows us to apply classical robust statistics (e.g., median, trimmed mean) in a coordinate-wise manner, exploiting the separability that the ℓ_∞-norm provides. This contrasts with the Euclidean setting, which required novel concentration-based robust mean estimation.

In the ℓ_∞-space, many robust statistics satisfy a *1-Lipschitz property* in each coordinate. If each data point is perturbed by at most ι in the ℓ_∞-norm, coordinate-wise robust statistics shift by at most ι. Combined with the fact that gradient descent can be made contractive in the ℓ_∞-norm under appropriate assumptions, this naturally provides the stability properties needed for our framework.

Specifically, if $x_t - x'_t$ is bounded coordinate-wise, then by smoothness we have $\|q_t(Z_i) - q'_t(Z_i)\|_\infty \le \beta\|x_t - x'_t\|_\infty$. The robust statistic computed from $\{q_t(Z_i)\}_{i\in[B]}$ and $\{q'_t(Z_i)\}_{i\in[B]}$ remains bounded by $\beta\|x_t - x'_t\|_\infty$ from the 1-Lipschitz property, establishing the desired stability.

While robust statistics control sensitivity effectively, they can introduce bias even in well-behaved datasets. We address this through coordinate-wise debiasing: if the empirical mean and the robust statistic are sufficiently close in a given coordinate, we use the mean directly; otherwise, we project the mean onto a ball centered at the robust statistic. This preserves both the 1-Lipschitz property and unbiasedness when the dataset is well-concentrated.

Unfortunately, this approach requires stronger assumptions than the Euclidean case.

Assumption 21 *Let $\mathcal{X} = \mathbf{B}_\infty(0, D)$. Each function $f(;z) : \mathcal{X} \to \mathbb{R}$ in the universe is convex, G-Lipschitz with respect to ℓ_1-norm (Definition 24) and β-smooth in ℓ_∞-norm (Definition 25).*

Assumption 22 *The Hessian of each function $f(\cdot; z)$ in the universe is diagonally dominant (Definition 26).*

The diagonal dominance assumption is primarily needed to ensure that gradient descent remains contractive in the ℓ_∞-norm. While somewhat restrictive, this assumption appears in various contexts, including heavy-tailed optimization (Wang et al., 2021) and neural network analysis where Hessians are often diagonal or block-diagonal (Liu et al., 2020; Martens and Grosse, 2015).

C.2. Algorithm

Our algorithm follows the same localization framework as in the Euclidean case, but replaces the concentration-based robust mean estimation with direct coordinate-wise application of classical robust statistics. The key modification lies in gradient estimation, which can be found in Appendix D.

Under Assumptions 21 and 22, suppose $\beta \le \frac{G}{D}(\frac{\sqrt{n}\varepsilon}{\sqrt{m}\log(nmd/\delta)} + \frac{\sqrt{d\log(1/\delta)\log(nmd)}}{\sqrt{m}\varepsilon})$, $\varepsilon \le O(1), n \ge \log^2(nd/\delta)/\varepsilon$ and $m \le n^{O(\log\log n)}$. Setting $\eta = \frac{D}{G}\cdot\min\{\frac{B\sqrt{m}}{\sqrt{n}}, \frac{\sqrt{m}\varepsilon}{\sqrt{d\log(1/\delta)\log(nmd)}}\}$, $B = 100\log(mnd/\delta)/\varepsilon$, $\tau = O(G\log(nmd)/\sqrt{m})$ and $\upsilon = 0.9B + \frac{2\log(T/\delta)}{\varepsilon}$, Algorithm 5 is (ε, δ)-user-level DP. When the nm functions in dataset $\mathcal{D}$ are i.i.d. drawn from the underlying distribution $\mathcal{P}$, it takes mn gradient computations and outputs x_S such that

$$\mathbb{E}[F(x_S) - F(x^*)] \le \tilde{O}\left(\frac{d}{\sqrt{nm}} + \frac{d^{3/2}}{n\varepsilon^2\sqrt{m}}\right).$$

C.3. Lower Bound

We also provide matching information-theoretic lower bounds showing this rate is nearly optimal (up to the ε and logarithmic factors) in this setting, validating our approach.

Theorem 23 *There exists a distribution $\mathcal{P}$ and a loss function f satisfying Assumption 21 and Assumption 22, such that for any (ε, δ)-user-level DP algorithm $\mathcal{M}$, given i.i.d. dataset $\mathcal{D}$ drawn from $\mathcal{P}$, the output of $\mathcal{M}$ satisfies*

$$\mathbb{E}[F(\mathcal{M}(\mathcal{D})) - F(x^*)] \geq GD \cdot \tilde{\Omega}\Big(\min\Big\{ d, \frac{d}{\sqrt{mn}} + \frac{d^{3/2}}{n\varepsilon\sqrt{m}} \Big\}\Big).$$

The non-private term $GD\frac{d}{\sqrt{mn}}$ represents the information-theoretic lower bound for SCO under these assumptions (see, e.g., (Agarwal et al., 2009, Theorem 1)).

We construct the hard instance as follows: let $\mathcal{X} = [-1, 1]^d$ be the unit ℓ_∞-ball and let $f(x; z) = -\langle x, z\rangle$ for any $x \in \mathcal{X}$ be the linear function. Let $z \in [-\sqrt{m}, \sqrt{m}]^d$ with $\mathbb{E}_{z\sim\mathcal{P}}[z] = \mu$. Then it is easy to verify that f satisfies Assumptions 21 and 22 with $G = \sqrt{m}$, $D = 1$ and $\beta = 0$. We have

$$\begin{aligned} F(\mathcal{M}(\mathcal{D})) - F(x^*) &= \sum_{i=1}^{d} (\operatorname{sign}(\mu[i]) - \mathcal{M}(\mathcal{D})[i]) \cdot \mu[i] \\ &\geq \sum_{i=1}^{d} |\mu[i]| . \mathbf{1}\big(\operatorname{sign}(\mu[i]) \neq \operatorname{sign}(\mathcal{M}(\mathcal{D})[i])\big). \end{aligned} \tag{14}$$

By (14), we have reduced the optimization error to the weighted sign estimation error. Most existing lower bounds rely on the ℓ_2^2-error of mean estimation. We adapt their techniques, especially the fingerprinting lemma, and provide the proof in the Appendix E.

Appendix D. Proof of Upper Bounds in Appendix C

Algorithm 5: Localization for user-level DP-SCO

Input: Dataset $\mathcal{D}$, parameters ε, δ, B, initial point x_0;
Process: Set $S = \lceil \log n/B \rceil$;
for *Phase* $s = 1, \ldots, S$ **do**
  Set $n_s = n/2^s$ and $\eta_s = (\log^{-s} m)\eta$;
  Draw a dataset $\mathcal{D}_s$ of size n_s from the unused users;
  Run Algorithm 6 with inputs $\mathcal{D}_s, \varepsilon, \delta, \eta_s, \tau, \upsilon, B, x_{s-1}$;
  set $x_s = \overline{x}_s + \zeta_s$, where $\zeta_s \sim \mathcal{N}(0, \sigma_s^2 I_d)$ with $\sigma_s = O(\frac{\eta_s G\sqrt{d\log(\exp(\varepsilon)/\delta)\log(nmd)}}{\sqrt{m}\varepsilon})$;
end
Return: the final solution x_s

Definition 24 (Lipschitz in ℓ_∞) *We say a function $f : \mathcal{X} \to \mathbb{R}$ is G-*Lipschitz *with respect to ℓ_1-norm, if for any $x, y \in \mathcal{X}$, we have $|f(x) - f(y)| \leq G\|x - y\|_1$. This means $\|\nabla f(x)\|_\infty \leq G$ for any $x \in \mathcal{X}$.*

Definition 25 (Smooth in ℓ_∞) *In this work, we say a function f is* β-smooth, *if $\|\nabla^2 f(x)\|_\infty \le \beta, \forall x \in \mathcal{X}$, where $\|A\|_\infty := \max_i \sum_j |A_{i,j}|$ for a symmetric matrix A. This implies that $\|\nabla f(x) - \nabla f(y)\|_\infty \le \beta \|x - y\|_\infty$ for any $x, y \in \mathcal{X}$.*

Definition 26 (Diagonal Dominance) *A matrix $A \in \mathbb{R}^{d \times d}$ is* diagonally dominant *if*

$$|A_{i,i}| \ge \sum_{j \ne i} |A_{i,j}|, \qquad \forall i \in [d].$$

We present our main result in this section and explain the algorithm in a top-down manner. The algorithm is aslo based on the localization framework of Feldman et al. (2020); see Algorithm 5 for details. The main result is restated below: Under Assumptions 21 and 22, suppose $\beta \le \frac{G}{D}(\frac{\sqrt{n}\varepsilon}{\sqrt{m}\log(nmd/\delta)} + \frac{\sqrt{d\log(1/\delta)\log(nmd)}}{\sqrt{m}\varepsilon})$, $\varepsilon \le O(1), n \ge \log^2(nd/\delta)/\varepsilon$ and $m \le n^{O(\log\log n)}$. Setting $\eta = \frac{D}{G} \cdot \min\{\frac{B\sqrt{m}}{\sqrt{n}}, \frac{\sqrt{m}\varepsilon}{\sqrt{d\log(1/\delta)\log(nmd)}}\}$, $B = 100\log(mnd/\delta)/\varepsilon$, $\tau = O(G\log(nmd)/\sqrt{m})$ and $\upsilon = 0.9B + \frac{2\log(T/\delta)}{\varepsilon}$, Algorithm 5 is (ε, δ)-user-level DP. When the nm functions in dataset $\mathcal{D}$ are i.i.d. drawn from the underlying distribution $\mathcal{P}$, it takes mn gradient computations and outputs x_S such that

$$\mathbb{E}[F(x_S) - F(x^*)] \le \tilde{O}\left(\frac{d}{\sqrt{nm}} + \frac{d^{3/2}}{n\varepsilon^2\sqrt{m}}\right).$$

Our main contribution is in Algorithm 6, which uses a novel gradient estimation sub-procedure. Note that some notations and subindices differ from those used in the Euclidean norm setting.

Algorithm 6: SGD for User-level DP-SCO

Input: dataset $\mathcal{D}$, privacy parameters ε, δ, other parameters η, τ, υ, B, initial point x_0;
Divide $\mathcal{D}$ into B disjoint subsets of equal size, denoted by $\{\mathcal{D}_i\}_{i\in[B]}$, $\mathcal{D}_i = \{Z_{i,t}\}_{t\in[|\mathcal{D}|/B]}$;
Set $T = |\mathcal{D}|/B$;
for *Step* $t = 1, \dots, T$ **do**
  For each $i \in [B]$, get $q_t(Z_{i,t}) := \frac{1}{m}\sum_{z\in Z_{i,t}} \nabla f(x_{t-1}; z)$;
  Let g_{t-1} be the output of Algorithm 7 with inputs $\{q_t(Z_{i,t})\}_{i\in[B]}$ and threshold $1/\tau$;
  $x_t = \Pi_{\mathcal{X}}(x_{t-1} - \eta g_{t-1})$
end
`/* Concentration Test */`
`/* Recall the query` $q_t(Z_{i,t})$ `for each` $t \in [T], i \in [B]$ `from above */`
Run Algorithm 8 with query $\{q_t\}_{t\in[T]}$ and parameters $\mathcal{D}_t, \varepsilon, \frac{\delta}{2Tmnd}, \tau, \upsilon$ to get answers $\{a_t\}_{t\in[T]}$;
if $a_t = \top, \forall t \in [T]$ **then**
  Return: Average iterate $\overline{x} = \frac{1}{T}\sum_{t\in[T]} x_t$;
end
else
  Output: Initial point x_0;
end

Algorithm 7: Gradient Estimation based on Robust Statistics

Input: a set of d-dimensional vectors $\{X_i\}_{i\in[B]}$, threshold parameter $\varsigma > 0$;
for *Each dimension* $j = 1, \ldots, d$ **do**
  Compute the robust statistics $X_{\mathrm{rs}}[j]$, and the mean $\overline{x}[j]$ over $\{X_i[j]\}_{i\in[B]}$;
  if $|X_{\mathrm{rs}}[j] - \overline{x}[j]| \geq \varsigma$ **then**
    Set $X_{est}[j] = \Pi_{B(Y_j,\varsigma)}(\overline{x}[j])$;
  end
  else
    Set $X_{est}[j] = \overline{x}[j]$;
  end
end
Return X_{est}

Iteration Sensitivity of Algorithm 6: The contractivity of gradient descent plays a crucial role in the sensitivity analysis, for which we need the Hessians to be diagonally dominant (Assumption 22).

Lemma 27 (Contractivity) *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a convex and β-smooth function satisfying Assumption 22, then for any two points $x, y \in \mathcal{X}$, with step size $\eta \leq 2/\beta$, we have*

$$\|(x - \eta\nabla f(x)) - (y - \eta\nabla f(y))\|_\infty \leq \|x - y\|_\infty.$$

Proof By the diagonal dominance assumption and precondition that $\eta \leq 2/\beta$, we know

$$\|I - \eta\nabla^2 f(z)\|_\infty = \max_j \left\{ |1 - \eta\nabla^2 f(z)_{j,j}| + \sum_{i\neq j} \eta|\nabla^2 f(z)_{j,i}| \right\} \leq 1.$$

Note that by the mean-value theorem,

$$(x - \eta\nabla f(x)) - (y - \nabla f(y)) = x - y + \eta(x - y)^\top \nabla^2 f(z) = (x - y)(I - \eta\nabla^2 f(z)),$$

where z is in the segment between x and y. Hence we have

$$\|(x - \eta\nabla f(x)) - (y - \nabla f(y))\|_\infty \leq \|x - y\|_\infty \cdot \|I - \eta\nabla^2 f(z)\|_\infty \leq \|x - y\|_\infty,$$

completing the proof. ■

Now, we discuss Algorithm 6. Given the dataset $\mathcal{D}$, we proceed in $T = |\mathcal{D}|/B$ steps. At the tth step, we draw B users $\{Z_{i,t}\}_{i\in[B]}$ and compute the average gradient of each user. We then apply our gradient estimation algorithm (Algorithm 7) and perform normal gradient descent for T steps.

In the second phase of Algorithm 6, we perform the concentration test (Algorithm 8) on the B gradients at each step based on AboveThreshold (Algorithm 3). If the concentration test passes for all steps (i.e., $a_t = \top$ for all $t \in [T]$), we output the average iterate. Otherwise, the algorithm fails and returns the initial point. As mentioned in the Introduction, the crucial novelty of Algorithm 6 and Algorithm 7 lies in bounding the sensitivity of each solution x_t by incorporating the (coordinate-wise) robust statistics into SGD.

We utilize robust statistics in the gradient estimation sub-procedure. We make the following assumptions regarding the robust statistics used:

Assumption 28 *Given a set $\{X_i\}_{i\in[B]}$ of vectors, let X_{rs} be any robust statistic satisfying the following properties:*

(i) For any $\rho \geq 0$, if there exists a point X' such that more than $B/2$ points lie within $B_\infty(X', \rho)$, then $X_{\mathrm{rs}} \in B_\infty(X', \rho)$.

(ii) If we perturb each point $Y_i = X_i + \iota_i$ such that $\|\iota_i\|_\infty \leq \Delta$ for any $\Delta \geq 0$, and let Y_{rs} be the robust statistic of $\{Y_i\}$, then $\|X_{\mathrm{rs}} - Y_{\mathrm{rs}}\|_\infty \leq \Delta$.

(iii) For any real numbers a and b, if $Z_i = aX_i + b$ for each $i \in [B]$, and Z_{rs} is the corresponding robust statistic of $\{Z_i\}_{i\in[B]}$, then $Z_{\mathrm{rs}} = aX_{\mathrm{rs}} + b$.

Remark 29 *Common robust statistics, such as the (coordinate-wise) median and trimmed mean, satisfy Assumption 28.*

Remark 30 *It is interesting to see whether any robust statistics satisfy the ℓ_2 analog, even with a mild relaxation (e.g., $\|X_{\mathrm{rs}} - Y_{\mathrm{rs}}\|_2 \leq (1 + r)\Delta$ for some $r = o(1)$.*

In Algorithm 7, we output means if they are close to the robust statistics to control the bias, and project the means onto the sphere around the robust statistics in a coordinate-wise manner when they are far apart. However, we still need to ensure that the sensitivity remains bounded when the projection is operated. The following technical lemma plays a crucial role in establishing iteration sensitivity to deal with the sensitivity with potential projection operations.

Lemma 31 *Consider four real numbers a, b, c, d, such that $|a - b| \leq 1$, and $|c - d| \leq 1$. Let $c' = \Pi_{B(a,r)}(c)$ and $d' = \Pi_{B(b,r)}(d)$ for any $r \geq 0$. Then, we have $|c' - d'| \leq 1$.*

Proof Without loss of generality, we assume $a \leq b$. We do case analysis.

Case (I): if no projection happens. Then it is straightforward that $c' = c$ and $d' = d$ and hence $|c' - d'| = |c - d| \leq 1$.

Case (II): if one projection happens. Without loss of generality, assume we project c and get c'. We analyze the following sub-cases:
(i): $c' = a - r$. In this case we know $c \leq c'$. If $d \geq a - r = c'$, then we know $|c' - d'| \leq |c - d| \leq 1$. If $d < a - r$, then $d - b < -r$ which is impossible.
(ii): $c' = a + r$. If $a + r \leq b$, then we know $|c' - d'| \leq |a - b| \leq 1$. Consider the subsubcase when $a + r > b$. If $d \leq a + r$, then $|c' - d'| \leq |c - d| \leq 1$. Else if $d \geq a + r$, as $b + r \geq d \geq c' = a + r$, we have $|c' - d'| \leq |a - b| \leq 1$.

Case (III): if two projections happen.
(i): $c' = a - r, d' = b - r$. This is a trivial case.
(ii): $c' = a + r, d' = b + r$. This is also a trivial case.
(iii): $c' = a - r, d' = b + r$. We can show that $|c' - d'| \leq |c - d| \leq 1$.
(iv): $c' = a + r, d' = b - r$. If $a + r \leq b$, then we know $|c' - d'| \leq |a - b| \leq 1$. Else, if $a + r > b$, then we know $|c' - d'| \leq |c - d| \leq 1$. ■

Unfortunately, we are unaware of any robust statistic satisfying Assumption 28 in high dimensions under the ℓ_2-norm, and Lemma 31 does not hold in high dimensions either. These limitations restrict the applicability of our techniques in high-dimensional Euclidean spaces.

Let $\{x_t\}_{t\in[T]}$ and $\{y_t\}_{t\in[T]}$ be two trajectories corresponding to neighboring datasets that differ by one user. The crucial technical novelty is that, for any $t \in [T]$, we can control $\|x_t - y_t\|_\infty$ as long as the number of "bad" users in each phase (B in total) does not exceed the "break point", say

$2B/3$. Without loss of generality, assume that $Z_{1,1} \neq Z'_{1,1}$ is the differing user in the neighboring dataset pairs $(\mathcal{D}, \mathcal{D}')$ considered in the following proof.

The first property of Assumption 28 ensures that when the number of "bad" users in each phase does not exceed the "break point" $2B/3$, the robust statistic remains close to most of the gradients, allowing us to control $\|x_1 - y_1\|_\infty$. To formalize this, we say that the neighboring dataset pair $(\mathcal{D}, \mathcal{D}')$ is *ρ-aligned* if there exist points X' and Y' such that $|X_{\text{good}}| \geq 2B/3$ and $|Y_{\text{good}}| \geq 2B/3$, where

$$X_{\text{good}} = \{q_1(Z_{i,1}) : q_1(Z_{i,1}) \in B_\infty(X', \rho), i \in [B]\}, \text{ and}$$

$$Y_{\text{good}} = \{q'_1(Z'_{i,1}) : q'_1(Z'_{i,1}) \in B_\infty(Y', \rho), i \in [B]\}.$$

This definition essentially states that the number of "bad" users does not exceed the "break point" in either $\mathcal{D}$ or $\mathcal{D}'$, ensuring that most gradients remain well-aligned within a bounded region.

Lemma 32 *For some (unknown) parameter $\rho > 0$, suppose $(\mathcal{D}, \mathcal{D}')$ is ρ-aligned. Then, by running Algorithm 7 with threshold parameter $\varsigma \geq 0$, we have $\|x_1 - y_1\|_\infty \leq \eta(4\rho + 2\varsigma)$.*

Proof It suffices to show that $\|g_0 - g'_0\|_\infty \leq 4\rho + 2\varsigma$. By the first property of Assumption 28 and the preconditions in the statement, we know $B_\infty(X', \rho) \cap B_\infty(Y', \rho) \neq \emptyset$, which leads to that

$$\|X' - Y'\| \leq 2\rho.$$

Moreover, we have that the robust statistic $\|X_{\text{rs}} - Y_{\text{rs}}\|_\infty \leq 4\rho$ by the triangle inequality as $\|X_{\text{rs}} - X'\|_\infty \leq \rho$ and $\|Y_{\text{rs}} - Y'\|_\infty \leq \rho$.

By the projection operation in Algorithm 7, we know $g_0 \in B_\infty(X_{\text{rs}}, \varsigma)$ and $g'_0 \in B_\infty(Y_{\text{rs}}, \varsigma)$, and hence we know $\|g_0 - g'_0\|_\infty \leq 4\rho + 2\varsigma$. This completes the proof. ■

The sequential sensitivity can be bounded using induction, with the base case $\|x_1 - y_1\|_\infty$ already established. The formal statement is provided in Lemma 33.

Algorithm 8: Concentration Test

Input: Dataset $\mathcal{D} = (Z_1, \ldots, Z_B)$, privacy parameters ε, δ, parameters τ, υ;
for $t = 1, \ldots, T$ **do**
  Receive a new concentration query $q_t : \mathcal{Z} \to \mathbb{R}^d$;
  Define the concentration score

$$s_t^{\mathsf{conc}}(\mathcal{D}, \tau) := \frac{1}{B} \sum_{Z \in \mathcal{D}} \sum_{Z' \in \mathcal{D}} \exp(-\tau \|q_t(Z) - q_t(Z')\|_\infty) \tag{15}$$

  Return $\mathsf{AboveThreshold}(s_t^{\mathsf{conc}}, \varepsilon/2, \upsilon)$
end

Lemma 33 (Iteration Sensitivity) *If we use a robust statistic satisfying Assumption 28 in Algorithm 7, then for all $t \in [T]$, we have $\|x_t - y_t\|_\infty \leq \|x_1 - y_1\|_\infty$.*

Proof Recall that we assume all users $Z_{i,t}$ are equal to $Z'_{i,t}$ but $Z_{1,1}$. We actually show that

$$\|(x_{t-1} - \eta g_{t-1}) - (y_{t-1} - \eta g'_{t-1})\|_\infty \le \|x_{t-1} - y_{t-1}\|_\infty,$$

as the projection operator to the same convex set is contractive in ℓ_2- and ℓ_∞-norm in our case.

Let $X_{i,t} = x_{t-1} - \eta q_t(Z_{i,t})$ and $Y_{i,t} = y_{t-1} - \eta q'_t(Z_{i,t})$. Note that the users used in computing the gradients are the same. Let X_{est} be the output of Algorithm 7 with $\{X_{i,t}\}_{i\in[B]}$ as inputs, and Y_{est} be the corresponding output of $\{Y_{i,t}\}_{i\in[B]}$. By the third property of Assumption 28, it suffices to show that

$$\|X_{\text{est}} - Y_{\text{est}}\|_\infty \le \|x_{t-1} - y_{t-1}\|_\infty. \tag{16}$$

By Lemma 27, we know

$$\|X_{i,t} - Y_{i,t}\|_\infty \le \|x_{t-1} - y_{t-1}\|_\infty.$$

Then by the second property in Assumption 28, we know that $\|X_{\text{rs}} - Y_{\text{rs}}\|_\infty \le \|x_{t-1} - y_{t-1}\|_\infty$. Similarly, by Lemma 27, we have $\|\overline{x} - \overline{y}\|_\infty \le \|x_{t-1} - y_{t-1}\|_\infty$. Then (16) follows from Lemma 31. This completes the proof. ■

Lemmas 32 and 33 together establish the iteration sensitivity of Algorithm 6.

Query Sensitivity of Concentration Test (Algorithm 8): We have established iteration sensitivity for any aligned neighboring dataset pair $(\mathcal{D}, \mathcal{D}')$. Next, we analyze the influence of the concentration test, which we use to check if the number of "bad" users exceed the "break point".

To apply the privacy guarantee of AboveThreshold (Lemma 17), it suffices to bound the sensitivity of each query in the concentration test. Recall that we assume $Z_{1,1} \neq Z'_{1,1}$ in the neighboring datasets. Thus, by the definition (Equation (15)), it is straightforward to observe that

$$|s_1^{\mathsf{conc}}(\mathcal{D}, \tau) - s_1^{\mathsf{conc}}(\mathcal{D}', \tau)| \le 2. \tag{17}$$

Next, we consider the sensitivity of s_t^{conc} for $t \ge 2$. The sensitivity is proportional to $\|x_t - y_t\|_\infty$, which we have already bounded by $\|x_1 - y_1\|_\infty$. Note that we can bound the iteration sensitivity if the neighboring datasets are aligned, meaning the number of "bad" users does not exceed the "break point". We first show that if the number of "bad" users exceeds the "break point", the algorithm is likely to halt after the first step by failing the first test.

Lemma 34 *Suppose* $B \ge \frac{100\log(T/\delta)}{\varepsilon}, \varepsilon \le O(1)$ *and we set* $\upsilon = 0.9B + \frac{2\log(T/\delta)}{\varepsilon}$. *Suppose for any point* Y, *we get* $|X_{\text{good}}| < B/3$ *where* $X_{\text{good}} = \{q_1(Z_{i,1}) : q_1(Z_{i,1}) \in B_\infty(Y, 1/\tau), i \in [B]\}$. *Then with probability at least* $1 - \delta/T\exp(\varepsilon)$, *the* AboveThreshold *returns* $a_1 = \perp$.

Proof The main randomness comes from the Laplacian noise we add to the query and the threshold. Under the precondition that $|X_{\text{good}}| < B/3$ for any Y, then we know the concentration score

$$s_1^{\mathsf{conc}}(\mathcal{D}, \tau) = \sum_{Z_{i,1}} \sum_{Z_{j,1}} \exp(-\tau\|q_1(Z_{i,1}) - q_1(Z_{j,1}\|) \le 2B/3 + \exp(-1)\cdot B/3 < 0.8B.$$

Then by Lemma 17 with a probability of at least $1 - \delta/T\exp(\varepsilon)$, we have

$$s_1^{\mathsf{conc}}(\mathcal{D}, \tau) + \text{Lap}(6/\varepsilon) \le \upsilon,$$

which means $a_1 = \perp$. ■

We now analyze the query sensitivity between the aligned neighboring datasets.

Lemma 35 (Query Sensitivity) *Suppose $6\beta\eta B \le 1$. Suppose $(\mathcal{D}, \mathcal{D}')$ is $(1/\tau)$-aligned and set threshold parameter $\varsigma = 1/\tau$ in Algorithm 3, the sensitivity of the query is bounded by at most 2:*

$$|s_t^{\mathsf{conc}}(\mathcal{D}, \tau) - s_1^{\mathsf{conc}}(\mathcal{D}', \tau)| \le 2, \qquad \forall t \ge 2.$$

Proof Consider the difference between $\|q_t(Z_{j,t}) - q_t(Z_{i,t})\|_\infty - \|q'_t(Z_{i,t}) - q'_t(Z_{j,t})\|_\infty$.

By Lemma 32 and Lemma 33, we know $\|x_t - y_t\|_\infty \le 6\eta/\tau$. By the smoothness assumption,

$$\begin{aligned}
&\|q_t(Z_{j,t}) - q_t(Z_{i,t})\|_\infty - \|q'_t(Z_{i,t}) - q'_t(Z_{j,t})\|_\infty \\
\le\ &\|q_t(Z_{i,t}) - q'_t(Z_{i,t}) - (q_t(Z_{j,t}) - q'_t(Z_{j,t}))\|_\infty \\
\le\ &\|q_t(Z_{i,t} - q'_t(Z_{i,t})\|_\infty + \|q_t(Z_{j,t}) - q'_t(Z_{j,t})\|_\infty \\
\le\ &2\beta\|x_t - y_t\|_\infty.
\end{aligned}$$

Hence we have

$$\begin{aligned}
s_i^{\mathsf{conc}}(\mathcal{D}, \tau) =& \frac{1}{B} \sum_{Z, Z' \in \mathcal{D}} \exp(-\tau \|q_i(Z) - q_i(Z')\|_\infty) \\
\ge& \frac{1}{B} \sum_{Z, Z' \in \mathcal{D}'} \exp(-\tau \|q'_i(Z) - q'_i(Z')\|_\infty) \exp(-12\beta\eta) \\
\ge& \exp(-12\beta\eta) s_i^{\mathsf{conc}}(\mathcal{D}', \tau).
\end{aligned}$$

As both $s_i^{\mathsf{conc}}(\mathcal{D}, \tau)$ and $s_i^{\mathsf{conc}}(\mathcal{D}', \tau)$ are in the range $[0, B]$, we know that

$$s_i^{\mathsf{conc}}(\mathcal{D}', \tau) - s_i^{\mathsf{conc}}(\mathcal{D}, \tau) \le (1 - \exp(-12\beta\eta))B \le 12\beta\eta B,$$

where we use the fact $1 - \exp(-x) \le x$ for any $x \ge 0$.

Similarly, we can bound $s_i^{\mathsf{conc}}(\mathcal{D}, \tau) - s_i^{\mathsf{conc}}(\mathcal{D}', \tau) \le 12\beta\eta B$, and complete the proof. ■

Equation (17) shows that the sensitivity is always bounded for s_1^{conc}. Lemma 34 shows that if the number of "bad" users exceeds the "break point", we obtain $a_1 = \perp$, and the query sensitivities of the subsequent queries do not need to be considered. Lemma 35 establishes the query sensitivity in the concentration test when the neighboring datasets are aligned, and the number of "bad" users is below the threshold.

Privacy proof. The final privacy guarantee–stated formally below–now easily follows from the previous lemmas.

Lemma 36 (Privacy Guarantee) *Under Assumption 21 and Assumption 22, suppose $\varepsilon \le O(1)$, $B \ge \frac{100\log(T/\delta)}{\varepsilon}$, then Algorithm 5 is (ε, δ)-user-level-DP.*

Proof Consider the implementation over two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, and we use the prime notation to denote the corresponding variables with respect to $\mathcal{D}'$. Without loss of generality, we assume the different user is used in the first phase.

To avoid confusion, let $\overline{x}_1$ and $\overline{x}'_1$ be the outputs of Algorithm 6 with neighboring inputs $\mathcal{D}_1$ and $\mathcal{D}'_1$, x_1 and x'_1 be the outputs of Algorithm 5 by privatizing $\overline{x}_1$ and $\overline{x}'_1$ with Gaussian noise respectively.

The outputs $\overline{x}_1$ and $\overline{x}'_1$ of Algorithm 6 depend on the intermediate random variables $\{a_t\}_{t\in[T]}$ and $\{a'_t\}_{t\in[T]}$.

As the query sensitivity for $t = 1$ is always bounded, by our parameter setting and property of AboveThreshold, we know $a_1 \approx_{\varepsilon/2,0} a'_1$.

We do case analysis.

(i) $(\mathcal{D}_1, \mathcal{D}'_1)$ is not $(1/\tau)$-aligned. Then by Lemma 34, either $\Pr[a_1 = \perp] \geq 1 - \delta/2$ or $\Pr[a'_1 = \perp] \geq 1 - \delta/2e^{\varepsilon}$. Then by union bound and $a_1 \approx_{\varepsilon/2,0} a'_1$, we have

$$\Pr[a_1 = a'_1 = \perp] \geq 1 - (1 + \exp(\varepsilon))\delta/2e^{\varepsilon}.$$

If $a_1 = a'_1 = \perp$, then $\overline{x}_1 = \overline{x}'_1$ is the initial point. We have for any event range $\mathcal{O}$,

$$\begin{aligned}\Pr[x_1 \in \mathcal{O}] =& \Pr[x_1 \in \mathcal{O} \mid a_1 = \perp] \Pr[a_1 = \perp] + \Pr[x_1 \in \mathcal{O} \mid a_1 = \top] \Pr[a_1 = \top] \\ \leq& \Pr[x'_1 \in \mathcal{O} \mid a'_1 = \perp] \exp(\varepsilon/2) \Pr[a'_1 = \perp] + e^{\varepsilon}(\delta/2 \exp(\varepsilon)) \\ \leq& e^{\varepsilon/2} \Pr[x'_1 \in \mathcal{O}] + \delta/2.\end{aligned}$$

This completes the privacy guarantee for case(i).

(ii) $(\mathcal{D}_1, \mathcal{D}'_1)$ is $(1/\tau)$-aligned. In this case, by Lemma 32 and Lemma 33, we know $\|x_t - y_t\|_\infty \leq 6\eta/\tau$. Moreover, Lemma 35 suggests that the query sensitivity is always bounded by 1. Then by the property of AboveThreshold (Lemma 17), we have $\{a_t\}_{t\in[T]} \approx_{\varepsilon/2,0} \{a'_t\}_{t\in[T]}$. We have for any event range $\mathcal{O}$,

$$\begin{aligned}\Pr[x_1 \in \mathcal{O}] =& \Pr[x_1 \in \mathcal{O} \mid \exists t, a_t = \perp] \Pr[\exists t, a_t = \perp] + \Pr[x_1 \in \mathcal{O} \mid a_t = \top, \forall t] \Pr[a_t = \top, \forall t] \\ \leq& \Pr[x'_1 \in \mathcal{O} \mid \exists t, a'_t = \perp] \exp(\varepsilon/2) \Pr[\exists t, a'_t = \perp] + \Pr[x_1 \in \mathcal{O} \mid a_t = \top, \forall t] \Pr[a_t = \top, \forall t] \\ \leq& \exp(\varepsilon/2) \Pr[x'_1 \in \mathcal{O} \mid \exists t, a'_t = \perp] \Pr[\exists t, a'_t = \perp] \\ &+ (\exp(\varepsilon/2) \Pr[x'_1 \in \mathcal{O} \mid a'_t = \top, \forall t] + \delta/2 \exp(\varepsilon)) \exp(\varepsilon/2) \Pr[a'_t = \top, \forall t] \\ \leq& \exp(\varepsilon) \Pr[x'_1 \in \mathcal{O}] + \delta,\end{aligned}$$

where we use the Gaussian Mechanism (Proposition 16) to bound $\Pr[x_1 \in \mathcal{O} \mid a_t = \top, \forall t]$ by $\Pr[x'_1 \in \mathcal{O} \mid a'_t = \top, \forall t]$. This completes the proof. ■

Utility proof. We apply the localization framework in private optimization to finish the utility argument. We analyze the utility guarantee of Algorithm 6 based on the classic convergence rate of SGD on smooth convex functions (Lemma 13) as follows:

Lemma 37 *Let $x \in \mathcal{X}$ be any point in the domain. Suppose the data set $\mathcal{D}$ of the users, whose size $|\mathcal{D}|$ is larger than $\frac{100 \log(T/\delta)}{\varepsilon}$, is drawn i.i.d. from the distribution $\mathcal{P}$. Setting $\tau = G \log(nmd/\omega)/\sqrt{m}$ then the final output $\overline{x}$ of Algorithm 6 satisfies that*

$$\mathbb{E}[F(\overline{x}) - F(x)] \lesssim \left(\beta + \frac{1}{\eta}\right) \frac{\mathbb{E}[\|x_0 - x\|^2]}{T} + \frac{\eta G^2 d}{Bm} + GDd\omega.$$

Proof By the Hoeffding inequality for norm-subGaussian vectors (Theorem 20), for each $t \in [T]$ and each $i \in [B]$, we have

$$\Pr\left[\|q_t(Z_{i,t}) - \nabla F(x_{t-1})\|_\infty \geq G \log(ndm/\omega)/\sqrt{m}\right] \leq 1 - \omega/nm.$$

Conditioned on the event that $\|q_t(Z_{i,t}) - \nabla F(x_{t-1})\|_\infty \le \tau$ for all $i \in [B]$ and $t \in [T]$, by our setting of parameters, we know that we pass all the concentration tests with $a_t = \top, \forall t \in [T]$, and

$$g_{t-1} = \frac{1}{B}\sum_{i\in[B]} q_t(Z_{i,t}),$$

which means $d_{\mathrm{tv}}\Big(\{g_{t-1}\}_{t\in[T]}, \{\frac{1}{B}\sum_{i\in[B]} q_t(Z_{i,t})\}_{t\in[T]}\Big) \le \omega$. Note that $\mathbb{E}[\frac{1}{B}\sum_{i\in[B]} q_t(Z_{i,t})] = \nabla F(x_{t-1})$ and $\mathbb{E}[\|\frac{1}{B}\sum_{i\in[B]} q_t(Z_{i,t}) - \nabla F(x_{t-1})\|_2^2] \le G^2 d/Bm$ when all functions are drawn i.i.d. from the distribution. By the small TV distance between g_{t-1} and the good gradient estimation $\frac{1}{B}\sum_{i\in[B]} q_t(Z_{i,t})$, it follows from Lemma 13 that

$$\mathbb{E}[F(\overline{x}) - F(x)] \lesssim (\beta + \frac{1}{\eta})\frac{\mathbb{E}[\|x_0 - x\|_2^2]}{T} + \frac{\eta G^2 d}{Bm} + GDd\omega,$$

where the last term comes from the worst value GDd, and the small failure probability ω. ■

Now we apply the localization framework. We set $\omega = 1/(nmd)^3$ to make the term depending on it negligible. The proof of the following lemma mostly follows from Feldman et al. (2020).

Lemma 38 (Localization) *Under Assumption 21 and Assumption 22, suppose* $\beta \le \frac{G}{D}(\frac{\sqrt{n}\varepsilon}{\sqrt{m\log(nmd/\delta}})$, $n \ge \log^2(nd/\delta)/\varepsilon, \varepsilon \le O(1)$ *and* $m \le n^{O(\log\log n)}$. *Set* $\eta \le \frac{D}{G}\cdot\min\{\frac{B\sqrt{m}}{\sqrt{n}}, \frac{\sqrt{m}\varepsilon}{\sqrt{d\log(1/\delta)\log(nmd)}}\}$, $B = 100\log(mnd/\delta)/\varepsilon$, $\tau = O(G\log(nmd)/\sqrt{m})$ *and* $\upsilon = 0.9B + \frac{2\log(T/\delta)}{\varepsilon}$. *If the dataset is drawn i.i.d. from the distribution* $\mathcal{P}$, *the final output* x_S *for Algorithm 5 satisfies*

$$\mathbb{E}[F(x_S) - F(x^*)] \le \tilde{O}\Big(GD\Big(\frac{d}{\sqrt{mn}} + \frac{d^{3/2}}{n\varepsilon^2\sqrt{m}}\Big)\Big).$$

Proof Let $\overline{x}_0 = x^*$ and $\zeta_0 = x_0 - x^*$. Lemma 37 can be used to analyze the utility concerning $\overline{x}_s$. As we add Gaussian noise ζ_s to $\overline{x}_s$ in each phase, we analyze the influence of ζ_s first.

By the assumption, we know $\|\zeta_0\|_2 \le D\sqrt{d}$. Recall that by the setting that $\eta \le \frac{D}{G}\cdot\frac{\sqrt{m}\varepsilon}{d\log(1/\delta)\log(nmd)}$, for all $s \ge 0$,

$$\mathbb{E}[\|\zeta_s\|_2^2] = d\sigma_s^2 = \eta_s^2 d\frac{Gd\log(1/\delta)\log(nmd)}{m\varepsilon^2} \le D^2 d\cdot\log^{-s} m.$$

Then by Lemma 37, we have

$$\begin{aligned}
\mathbb{E}[F(x_S)] - F(x^*) &= \sum_{s=1}^{S}\mathbb{E}[F(\overline{x}_s - \overline{x}_{s-1})] + \mathbb{E}[F(x_s) - F(\overline{x}_s)] \\
&\le \sum_{s=1}^{S}\left(\frac{\mathbb{E}[\|\zeta_{s-1}\|_2^2]}{\eta_s T_s} + \frac{\eta_s G^2 d}{2Bm}\right) + G\,\mathbb{E}[\|\zeta_S\|_2] \\
&\le \sum_{s=1}^{S}\left(\frac{\log m}{2}\right)^{-(i-2)}\left(\frac{D^2 d}{\eta n/B} + \frac{\eta G^2 d}{2Bm}\right) + \frac{GDd}{(\log m)^{\log n}}
\end{aligned}$$

$$\leq \tilde{O}\Big(GD(\frac{d}{\sqrt{nm}} + \frac{d^{3/2}}{n\sqrt{m}\varepsilon^2})\Big),$$

where we use the fact that $\frac{1}{(\log m)^{\log n}} \leq 1/nm$ when $m \leq n^{\log\log n}$. ■

Main Result: Theorem C.2 directly follows from Lemma 38 and Lemma 36.

D.1. Relationship to Main Results and Limitations

This extension serves several purposes: it demonstrates the broad applicability of robust statistical principles across different geometric settings, provides theoretical validation through tight lower bounds, and offers insights into when coordinate-wise decomposition can simplify robust estimation.

However, the approach has notable limitations compared to our main Euclidean results:

- **Stronger Assumptions.** The diagonal dominance requirement is more restrictive than the standard smoothness assumptions needed for our main results. This heavily limits the generality and potential practicality of the approach.
- **Geometric Constraints.** The techniques are specifically tailored to the ℓ_∞ setting and do not readily extend to other norms, unlike our concentration-based approach which may be more geometrically flexible.
- **Suboptimal ε Dependence.** The ε^{-2} dependence in the privacy term is worse than our main ε^{-1} result, arising from limitations in how robust statistics benefit from larger batch sizes.

Despite these limitations, this extension validates that robust statistical principles can be successfully leveraged across different geometric settings, complementing our main theoretical framework. The techniques may be of independent interest for problems naturally formulated in ℓ_∞ spaces or when diagonal dominance assumptions are satisfied.

D.2. Counterexample of the 1-Lipschitz of Geometric Median

We use the counterexample from Durocher and Kirkpatrick (2009) to show the geometric median is unstable in 2-dimensional space. Recall given a set of points P in $\mathbb{R}^d$, the geometric median of P, denoted by $M(P)$, is

$$M(P) := \arg\min_x \sum_{p\in P} \|x - p\|_2.$$

Let $P = \{(0,0),(0,0),(1,0),(1,\alpha)\}$ and $P' = \{(0,0),(0,\alpha),(1,0),(1,0)\}$ with $\alpha > 0$ as a very small perturbation. But we know $M(P) = (0,0)$ and $M(P') = (1,0)$.

Appendix E. Details of Lower Bound in Appendix C

Now we present the proof of our lower bound (Theorem 23). As discussed in the main text, we reduce the optimization error to the weighted sign estimation error. We construct a lower bound for the weighted sign estimation error.

Weighted sign estimation error. We construct a distribution $\mathcal{P}_1$ as follows: for each coordinate $k \in [d]$, we draw $\mu[k]$ uniformly random from $[-1, 1]$, and $z_{i,j}[k] \sim \mathcal{N}(\mu[k], m)$ i.i.d., for $i \in [n], j \in [m]$. The objective is to minimize weighted sign estimation error with respect to μ.

Lemma 39 *Let $\varepsilon \le 0.1, \delta \le 1/(dnm)$. For any (ε, δ)-user-level-DP algorithm $\mathcal{M}$, there exists a distribution $\mathcal{P}_2$ such that $\|\mathbb{E}_{z\sim\mathcal{P}_2}[z]\|_\infty \le 1$ and $\|z\|_\infty \le \tilde{O}(\sqrt{m})$ almost surely, and, given dataset $\mathcal{D}$ i.i.d. drawn from $\mathcal{P}_2$, we have*

$$\mathop{\mathbb{E}}_{\mathcal{D},\mathcal{M},\mu} \sum_{i=1}^{d} |\mu[i]| \cdot \mathbf{1}\big(\operatorname{sign}(\mu[i]) \neq \operatorname{sign}(\mathcal{M}(\mathcal{D})[i])\big) \ge \tilde{\Omega}(\frac{d^{3/2}}{n\varepsilon}).$$

First, by the previous result, we can reduce the user-level to item-level setting.

Lemma 40 (Levy et al. (2021)) *Suppose each user $Z_i, i \in [n]$ observes $z_{i,1}, \ldots, z_{i,m} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2 I_d)$ with σ known. For any (ε, δ)-User-level-DP algorithm $\mathcal{M}_{\text{user}}$, there exists an (ε, δ)-Item-level-DP algorithm $\mathcal{M}_{\text{item}}$ that takes inputs $(\overline{Z}_1, \ldots, \overline{Z}_n)$ where $\overline{Z}_i = \frac{1}{m}\sum_{j\in[m]} z_{i,j}$ and has the same performance as $\mathcal{M}_{\text{user}}$.*

Hence by Lemma 40, it suffices to consider the item-level lower bound. We prove the following:

Lemma 41 *Let $\{\mu[k]\}_{k\in[d]} \overset{\text{i.i.d.}}{\sim} [-\sigma, \sigma]$. Let $\{\overline{Z}_1, \ldots, \overline{Z}_n\}$ be i.i.d. drawn from $\mathcal{N}(\mu, \sigma^2 I_d)$. If $\mathcal{M} : \mathbb{R}^{n\times d} \to \{-1, 1\}^d$ is (ε, δ)-DP, and*

$$\mathop{\mathbb{E}}_{\mu,\mathcal{D},\mathcal{M}} \sum_{i=1}^{d} |\mu[i]| \cdot \mathbf{1}\big(\operatorname{sign}(\mu[i]) \neq \operatorname{sign}(\mathcal{M}(\mathcal{D})[i])\big) \le \alpha \le \frac{d\sigma}{8},$$

then $n \ge \frac{\sqrt{d}}{32\varepsilon}$.

By the invariant scaling, it suffices to consider the case when $\sigma = 1$. To prove Lemma 41, we need the fingerprinting lemma:

Lemma 42 (Lemma 6.8 in Kamath et al. (2019)) *For every fixed number p and every $f : \mathbb{R}^n \to [-1, 1]$, define $g(p) := \mathbb{E}_{X_{1,\ldots n}\sim\mathcal{N}(p,1)}[f(X)]$. We have*

$$\mathop{\mathbb{E}}_{X_{1,\ldots n}\sim\mathcal{N}(p,1)}[(1-p^2)(f(X)-p)\sum_{i\in[n]}(X_i - p)] = (1-p^2)\frac{\mathrm{d}}{\mathrm{d}p}g(p). \tag{18}$$

By choosing p uniformly from $[-1, 1]$, we have the following observation over the expectation on the RHS of Equation (18).

Lemma 43 *We have*

$$\mathop{\mathbb{E}}_{p\sim[-1,1]}[(1-p^2)\frac{\mathrm{d}}{\mathrm{d}p}g(p)] = \mathop{\mathbb{E}}_{p}[g(p)\cdot p].$$

Proof Using integration by parts, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{p\sim[-1,1]}[(1-p^2)\frac{\mathrm{d}}{\mathrm{d}p}g(p)] =& \frac{1}{2}\int_{-1}^{1}(1-p^2)\frac{\mathrm{d}}{\mathrm{d}p}g(p)\mathrm{d}p\\ =& \frac{1}{2}\int_{-1}^{1}\Big(\frac{\mathrm{d}}{\mathrm{d}p}\big((1-p^2)g(p)\big) - g(p)\frac{\mathrm{d}}{\mathrm{d}p}\big(1-p^2\big)\Big)\mathrm{d}p\\ =& \int_{-1}^{1}g(p)p\mathrm{d}p\\ =& \mathbb{E}[g(p)\cdot p],\end{aligned}$$

completing the proof. ■

Now we use the fingerprinting lemma.

Lemma 44 *One has*

$$\mathbb{E}\left[\sum_{i\in[n],k\in[d]}(1-\mu[k]^2)(\mathcal{M}(\mathcal{D})[k]-\mu[k])\cdot(\overline{Z}_i[k]-\mu[k])\right] = \mathbb{E}\left[\sum_{k\in[d]}\mathcal{M}(\mathcal{D})[k]\cdot\mu[k]\right].$$

Proof Fix a column $k\in[d]$.

Construct the f for our purpose. Define $f:\mathbb{R}^n\to[-1,1]$ to be

$$f(X) := \mathop{\mathbb{E}}_{\mathcal{D},\mathcal{M}}[\mathcal{M}(\mathcal{D}^{-k}\|X)[k]].$$

That is, $f(X)$ is the expectation of $\mathcal{M}(\mathcal{D})[k]$ conditioned on $\overline{Z}_i[k]=X_i,\forall i\in[n]$. And define $g:[-1,1]\to[-1,1]$ to be

$$g(p) := \mathop{\mathbb{E}}_{\mu^{-k},X_{1,\dots n}\sim\mathcal{N}(p,1)}[f(X)].$$

That is $g(p)$ is the expectation of $\mathcal{M}(\mathcal{D})[k]$ conditional on $\mu[k]=p$.

Now we can calculate

$$\begin{aligned}&\mathbb{E}\left[(1-\mu[k]^2)(\mathcal{M}(\mathcal{D})[k]-\mu[k])\sum_{i\in[n]}(\overline{Z}_i[k]-\mu[k])\right]\\ =&\mathop{\mathbb{E}}_{\mu[k]\sim[-1,1],X_{1,\dots n}\sim\mathcal{N}(\mu[k],1)}\left[(1-\mu[k]^2)(f(X)-\mu[k])\sum_{i\in[n]}(\overline{Z}_i[k]-\mu[k])\right]\\ \overset{(i)}{=}&\mathop{\mathbb{E}}_{\mu[k]}[g(\mu[k])\cdot\mu[k]]\\ =&\mathbb{E}[\mathcal{M}(\mathcal{D})[k]\mu[k]],\end{aligned}$$

where (i) follows from Lemma 42 and Lemma 43. The proofs follows by summing over $k\in[d]$. ■

A small weighted sign error means a large $\mathbb{E}\left[\sum_{k\in[d]}\mathcal{M}(\mathcal{D})[k]\cdot\mu[k]\right]$, as demonstrated by the following lemma:

Lemma 45 *Let $\mathcal{M} : \mathbb{R}^{n\times d} \to [-1,1]^d$. Suppose $\{\mu[k]\} \overset{i.i.d.}{\sim} [-1,1]$, and*

$$\mathop{\mathbb{E}}_{\mu,\mathcal{D},\mathcal{M}}\left[\sum_{k=1}^{d} |\mu[k]| \cdot \mathbf{1}\big(\operatorname{sign}(\mu[k]) \neq \operatorname{sign}(\mathcal{M}(\mathcal{D})[k])\big)\right] \leq \alpha,$$

then we have

$$\mathbb{E}\left[\sum_{k\in[d]} \mathcal{M}(\mathcal{D})[k]\cdot\mu[k]\right] \geq \frac{d}{2} - 2\alpha.$$

Proof Note that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{d} |\mu[k]| \cdot \Big(\mathbf{1}(\operatorname{sign}(\mu[k]) \neq \operatorname{sign}(\mathcal{M}(\mathcal{D})[k])) + \mathbf{1}(\operatorname{sign}(\mu[k]) = \operatorname{sign}(\mathcal{M}(\mathcal{D})[k]))\Big)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{d} |\mu[k]|\right] = d/2.
\end{aligned}$$

Moreover, one has that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k\in[d]} \mathcal{M}(\mathcal{D})[k]\cdot\mu[k]\right] \\
=&\mathbb{E}\left[\sum_{k\in[d]} |\mu[k]| \cdot \Big(\mathbf{1}(\operatorname{sign}(\mu[k]) = \operatorname{sign}(\mathcal{M}(\mathcal{D})[k])) - \mathbf{1}(\operatorname{sign}(\mu[k]) \neq \operatorname{sign}(\mathcal{M}(\mathcal{D})[k]))\Big)\right] \\
\geq& d/2 - 2\alpha.
\end{aligned}$$

This completes the proof. ■

It remains to show the sample complexity to achieve a large value of

$$\mathbb{E}\left[\sum_{i\in[n],k\in[d]} (1-\mu[k]^2)(\mathcal{M}(\mathcal{D})[k]-\mu[k])\cdot(\overline{Z}_i[k]-\mu[k])\right] \geq d/2 - 2\alpha. \tag{19}$$

Now we complete the proof of Lemma 41.

Proof [of Lemma 41]

Let $A_i = \sum_{k\in[d]}(1-\mu[k]^2)(\mathcal{M}(\mathcal{D})[k]-\mu[k])(\overline{Z}_i[k]-\mu[k])$. We use the DP constraint of $\mathcal{M}$ to upper bound $\mathbb{E}[A_i]$.

Let $\mathcal{D}_{\sim i}$ denote $\mathcal{D}$ with $\overline{Z}_i$ replaced with an independent draw from $\mathcal{P}_1$. Define

$$\tilde{A}_i = \sum_{k\in[d]} (1-\mu[k]^2)(\mathcal{M}(\mathcal{D}_{\sim i})[k]-\mu[k])(\overline{Z}_i[k]-\mu[k]).$$

Due to the independence between $\mathcal{M}(\mathcal{D}_{\sim i})$ and $\overline{Z}_i$, we have $\mathbb{E}[\tilde{A}_i] = 0$. Moreover, as $|1-\mu[k]^2| \leq 1$ and $|\mathcal{M}(\mathcal{D}_{\sim i}) - \mu[k]| \leq 2$, we have $\mathbb{E}[|\tilde{A}_i|] \leq 2\sqrt{\sum_{k\in[d]} \mathbf{Var}(\overline{Z}_i[k])} \leq 2\sqrt{d}$.

Split A_i with $A_{i,+} = \max\{0, A_i\}$ and $A_{i,-} = \min\{0, A_i\}$ and split $\tilde{A}_i$ similarly. By the property of DP, we know

$$\Pr[|A_{i,+}| \geq t] \leq \exp(\varepsilon) \Pr[|\tilde{A}_{i,+}| \geq t] + \delta, \forall t \geq 0,$$
$$\Pr[|A_{i,-}| \geq t] \geq \exp(-\varepsilon) \Pr[|\tilde{A}_{i,1}| \geq t] - \delta, \forall t \geq 0.$$

Then we have

$$\begin{aligned}
\mathbb{E}[A_i] &= \int_0^\infty \Pr[|A_{i,+}| \geq t]\mathrm{d}t - \int_0^\infty \Pr[|A_{i,-}| \leq t]\mathrm{d}t \\
&\leq \exp(\varepsilon)\, \mathbb{E}[|\tilde{A}_{i,+}|] - \exp(-\varepsilon)\, \mathbb{E}[|\tilde{A}_{i,-}|] + 2\delta T + \int_T^\infty \Pr[|A_i| \geq t]\mathrm{d}t \\
&\leq \mathbb{E}[\tilde{A}_i] + (\exp(\varepsilon) - 1)\, \mathbb{E}[|\tilde{A}_{i,+}|] + (1 - \exp(-\varepsilon))\, \mathbb{E}[|\tilde{A}_{i,-}|] + 2\delta T + \int_T^\infty \Pr[|A_i| \geq t]\mathrm{d}t \\
&\leq \mathbb{E}[\tilde{A}_i] + 2\varepsilon\, \mathbb{E}[|\tilde{A}_i|] + 2\delta T + \int_T^\infty \Pr[|A_i| \geq t]\mathrm{d}t,
\end{aligned}$$

where the last inequality used the fact that $\exp(\varepsilon) - 1 \leq 2\varepsilon$ when $\varepsilon \leq 1/10$.

When $\delta \leq 1/dn^2$ and set $T = O(\sqrt{d \log(1/\delta)})$, we have

$$\mathbb{E}[A_i] \leq 4\varepsilon\sqrt{d} + d/8n.$$

When $\alpha \leq d/8$, Equation (19) implies that

$$n(4\varepsilon\sqrt{d} + d/8n) \geq d/4,$$

which leads to

$$n \geq \frac{\sqrt{d}}{32\varepsilon}.$$

This completes the proof. ■

It is standard to translate the sample complexity lower bound (Lemma 41) to the error lower bound (Lemma 39). We present a proof below.

Proof [of Lemma 39]

Let $\mathcal{A}^{\text{item}}_{\varepsilon,\delta}$ be the set of item-level DP mechanisms, and let $\mathcal{A}^{\text{user}}_{\varepsilon,\delta}$ be the set of user-level DP mechanisms.

Define the error term:

$$\text{Error}[\mathcal{P}, \mathcal{M}, n] = \mathop{\mathbb{E}}_{\mathcal{D}\sim\mathcal{P}^n, \mathcal{M}} \sum_{i=1}^{d} |\mu[i]| \mathbf{1}(\text{sign}(\mu[i]) \neq \text{sign}(\mathcal{M}(\mathcal{D})[i])),$$

where $\mu = \mathbb{E}_{z\sim\mathcal{P}}\, z$.

Recall that we construct the distribution $\mathcal{P}_1$ as follows: for each coordinate $k \in [d]$, we draw $\mu[k]$ uniformly random from $[-1, 1]$, and $z_{i,j}[k] \sim \mathcal{N}(\mu[k], m)$ i.i.d., for $i \in [n], j \in [m]$.

Let $\overline{\mathcal{P}}_1$ be the following: for each coordinate $k \in [d]$, we draw $\mu[k]$ uniformly random from $[-1, 1]$, and $\overline{Z}_i[k] \sim \mathcal{N}(\mu[k], 1)$ i.i.d., for $i \in [n]$. $\overline{\mathcal{P}}_1$ is corresponding to averaging the m samples for each user. By Lemma 40, we have

$$\inf_{\mathcal{M} \in \mathcal{A}^{\text{user}}_{\varepsilon,\delta}} \text{Error}[\mathcal{P}_1, \mathcal{M}, nm] \geq \inf_{\mathcal{M} \in \mathcal{A}^{\text{item}}_{\varepsilon,\delta}} \text{Error}[\overline{\mathcal{P}}_1, \mathcal{M}, n].$$

By Lemma 41,

$$\inf_{\mathcal{M} \in \mathcal{A}^{\text{item}}_{\varepsilon,\delta}} \text{Error}[\overline{\mathcal{P}}_1, \mathcal{M}, \sqrt{d}/32\varepsilon] \geq \Omega(d).$$

Let $n^* = \tilde{O}(\sqrt{d}/\varepsilon)$. When we have a large data set of size $n \gg n^*$, construct $\overline{\mathcal{P}}_2 = \frac{n^*}{n}\overline{\mathcal{P}}_1 + (1 - \frac{n^*}{n})\mathcal{P}_3$, where $\mathcal{P}_3$ is a Dirac distribution at $\mathbf{0} \in \mathbb{R}^d$.

Hence, by a Chernoff bound, with high probability, the number of samples drawn from $\overline{\mathcal{P}}_1$ in the dataset $\mathcal{D}$ is no more than $O(n^* \cdot \log(nd)) = \frac{\sqrt{d}}{32\varepsilon}$. Then we have

$$\begin{aligned}\inf_{\mathcal{M} \in \mathcal{A}^{\text{user}}_{\varepsilon,\delta}} \text{Error}[\mathcal{P}_1, \mathcal{M}, nm] \geq & \inf_{\mathcal{M} \in \mathcal{A}^{\text{item}}_{\varepsilon,\delta}} \text{Error}[\overline{\mathcal{P}}_2, \mathcal{M}, n] \\ \geq & \frac{n^*}{n} \inf_{\mathcal{M} \in \mathcal{A}^{\text{item}}_{\varepsilon,\delta}} \text{Error}[\overline{\mathcal{P}}_1, \mathcal{M}, \sqrt{d}/32\varepsilon] \geq \tilde{\Omega}(\frac{d^{3/2}}{\varepsilon n}).\end{aligned}$$

In the precondition of $\mathcal{P}_2$, we need $\|z\|_\infty \leq \tilde{O}(\sqrt{m})$ almost surely for $z \sim \mathcal{P}_2$. But $\mathcal{P}_1$ involves some Gaussian distributions. We construct $\mathcal{P}_2$ by truncating the Gaussian distributions.

More specifically, for each item $z_{i,j}$ drawn from $\mathcal{N}(\mu, mI_d)$, we set $z'_{i,j}[k] = \frac{z_{i,j}[k]}{\max\{1, |z_{i,j}[k]|/G\}}$ with $G = \Theta(\sqrt{m \log(mnd)})$. In other words, we get $z'_{i,j}$ by projecting $z_{i,j}$ to $B_\infty(0, G)$. Fixing μ, we first show

$$\| \mathop{\mathbb{E}}_{z \sim \mathcal{P}_2}[z] - \mu\|_\infty \leq O(1/dn^2). \tag{20}$$

It suffices to consider any coordinate $k \in [d]$, and prove

$$| \mathop{\mathbb{E}}_{z \sim \mathcal{P}_2}[z[k]] - \mu[k]| \leq O(1/dn^2).$$

Letting $\beta = \frac{-\mu[k]+G}{\sqrt{m}}$ and $\alpha = \frac{-\mu[k]-G}{\sqrt{m}}$, we know

$$| \mathop{\mathbb{E}}_{z \sim \mathcal{P}_2}[z[k]] - \mu[k]| \leq \frac{n^*}{n} \cdot \sqrt{m} \cdot \frac{\varphi(\alpha) - \varphi(\beta)}{\int_\alpha^\beta \varphi(x)\mathrm{d}x},$$

where φ is density function of standard Gaussian $\mathcal{N}(0, 1)$. As $\mu[k] \in [-1, 1]$ and $G = \Theta(\sqrt{m \log(mnd)})$, we establish Equation (20). Denote $\mu' = \mathbb{E}_{z \sim \mathcal{P}_2}[z \mid \mu]$, that is the mean of $\mathcal{P}_2$ conditional on μ.

Moreover, we have

$$\inf_{\mathcal{M} \in \mathcal{A}^{\text{user}}_{\varepsilon,\delta}} \mathop{\mathbb{E}}_{\mu, \mathcal{D} \sim \mathcal{P}_2^{mn}, \mathcal{M}} \sum_{i=1}^d |\mu'[i]| \mathbf{1}(\text{sign}(\mu'[i]) \neq \text{sign}(\mathcal{M}(\mathcal{D})))$$

$$
\begin{aligned}
&\geq \inf_{\mathcal{M}\in\mathcal{A}^{\text{user}}_{\varepsilon,\delta}} \mathop{\mathbb{E}}_{\mu,\mathcal{D}\sim\mathcal{P}_2^{mn},\mathcal{M}} \Big(\sum_{i=1}^{d} |\mu[i]| \mathbf{1}(\text{sign}(\mu[i]) \neq \text{sign}(\mathcal{M}(\mathcal{D}))) - d\|\mu-\mu'\|_\infty \Big) \\
&\geq \inf_{\mathcal{M}\in\mathcal{A}^{\text{user}}_{\varepsilon,\delta}} \mathop{\mathbb{E}}_{\mu,\mathcal{D}\sim\mathcal{P}_2^{mn},\mathcal{M}} \sum_{i=1}^{d} |\mu[i]| \mathbf{1}(\text{sign}(\mu[i]) \neq \text{sign}(\mathcal{M}(\mathcal{D}))) - O(1/n^2) \\
&\overset{(i)}{\geq} \inf_{\mathcal{M}\in\mathcal{A}^{\text{user}}_{\varepsilon,\delta}} \mathop{\mathbb{E}}_{\mu,\mathcal{D}\sim\mathcal{P}_1^{mn},\mathcal{M}} \sum_{i=1}^{d} |\mu[i]| \mathbf{1}(\text{sign}(\mu[i]) \neq \text{sign}(\mathcal{M}(\mathcal{D}))) - O(1/n^2) \\
&\geq \inf_{\mathcal{M}\in\mathcal{A}^{\text{user}}_{\varepsilon,\delta}} \text{Error}(\mathcal{P}_1,\mathcal{M},nm) - O(1/n^2) \\
&\geq \tilde{\Omega}\left(\frac{d^{3/2}}{\varepsilon n}\right),
\end{aligned}
$$

where the inequality (i) follows from that we can always sample from $\mathcal{P}_2$ with samples from $\mathcal{P}_1$, which means the problem over $\mathcal{P}_2$ is harder than $\mathcal{P}_1$. This completes the proof. ■

The lower bound Theorem 23 follows from Lemma 39 and our construction of the function class, that is

$$
\begin{aligned}
\mathbb{E}[F(\mathcal{M}(\mathcal{D})) - F(x^*)] &\geq \mathop{\mathbb{E}}_{\mathcal{D},\mathcal{M},\mu} \sum_{i=1}^{d} |\mu[i]| \cdot \mathbf{1}\big(\text{sign}(\mu[i]) \neq \text{sign}(\mathcal{M}(\mathcal{D})[i])\big) \\
&\geq \tilde{\Omega}\left(\frac{d^{3/2}}{n\varepsilon}\right) = GD \cdot \tilde{\Omega}\left(\frac{d^{3/2}}{n\varepsilon\sqrt{m}}\right),
\end{aligned}
$$

given $G = \tilde{O}(\sqrt{m})$.