Learning with User-Level Privacy

Daniel Levy[*,1] **Ziteng Sun**[*,2] **Kareem Amin**[3] **Satyen Kale**[3]
Alex Kulesza[3] **Mehryar Mohri**[3,4] **Ananda Theertha Suresh**[3]

[1]Stanford University [2]Cornell University [3]Google Research [4]Courant Institute
danilevy@stanford.edu, zs335@cornell.edu,
{kamin, satyenkale, kulesza, mohri, theertha}@google.com

Abstract

We propose and analyze algorithms to solve a range of learning tasks under user-level differential privacy constraints. Rather than guaranteeing only the privacy of individual samples, user-level DP protects a user's entire contribution ($m \geq 1$ samples), providing more stringent but more realistic protection against information leaks. We show that for high-dimensional mean estimation, empirical risk minimization with smooth losses, stochastic convex optimization, and learning hypothesis classes with finite metric entropy, the privacy cost decreases as $O(1/\sqrt{m})$ as users provide more samples. In contrast, when increasing the number of users n, the privacy cost decreases at a faster $O(1/n)$ rate. We complement these results with lower bounds showing the minimax optimality of our algorithms for mean estimation and stochastic convex optimization. Our algorithms rely on novel techniques for private mean estimation in arbitrary dimension with error scaling as the concentration radius τ of the distribution rather than the entire range.

1 Introduction

Releasing seemingly innocuous functions of a data set can easily compromise the privacy of individuals, whether the functions are simple counts [35] or complex machine learning models like deep neural networks [52, 30]. To protect against such leaks, Dwork et al. proposed the notion of *differential privacy* (DP). Given some data from n participants in a study, we say that a statistic of the data is differentially private if an attacker who already knows the data of $n-1$ participants cannot reliably determine from the statistic whether the n-th remaining participant is Alice or Bob. With the recent explosion of publicly available data, progress in machine learning, and widespread public release of machine learning models and other statistical inferences, differential privacy has become an important standard and is widely adopted by both industry and government [32, 5, 21, 55].

The standard setting of DP described in [22] assumes that each participant contributes a *single* data point to the dataset, and preserves privacy by "noising" the output in a way that is commensurate with the maximum contribution of a single example. This is not the situation faced in many applications of machine learning models, where users often contribute *multiple* samples to the model—for example, when language and image recognition models are trained on the users' own data, or in federated learning settings [37]. As a result, current techniques either provide privacy guarantees that degrade with a user's increased participation or naively add a substantial amount of noise, relying on the group property of differential privacy, which significantly harms the performance of the deployed model.

To remedy this issue, we consider *user-level* DP, which instead of guaranteeing privacy for individual samples, protects a user's *entire contribution* ($m \geq 1$ samples). This is a more stringent but more realistic privacy desideratum. To hold, it requires that the output of our algorithm does not significantly

[*]Equal contribution. Work was done during an internship at Google Research.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

change when changing user's entire contribution—i.e. possibly swapping up to m samples in total. We make this formal in Definition 1. Very recently, for the reasons outlined above, there has been increasing interest in user-level DP for applications such as estimating discrete distributions under user-level privacy constraints [46], PAC learning with user-level privacy [31], and bounding user contributions in ML models [4, 26]. Differentially private SQL with bounded user contributions was proposed in [59]. User-level privacy has been also studied in the context of learning models via federated learning [49, 48, 58, 6].

In this paper, we tackle the problem of *learning* with user-level privacy in the central model of DP. In particular, we provide algorithms and analyses for the tasks of mean estimation, empirical risk minimization (ERM), stochastic convex optimization (SCO), and learning hypothesis classes with finite metric entropy. Our utility analyses assume that all users draw their samples i.i.d. from related distributions, a setting we refer to as *limited heterogeneity*. On these tasks, naively applying standard mechanisms, such as Laplace or Gaussian, or using the group property with item-level DP estimators, both yield a privacy error independent of m. We first develop novel private mean estimators in high dimension with statistical and privacy error scaling with the (arbitrary) concentration radius rather than the range, and apply these to the statistical query setting [SQ; 41]. Our algorithms then rely on (privately) answering a sequence of adaptively chosen queries using users' samples, e.g., gradient queries in stochastic gradient descent algorithms. We show that for these tasks, the additional error due to privacy constraints decreases as $O(1/\sqrt{m})$, contrasting with the naive rate—independent of m. Interestingly, increasing n, the number of users, decreases the privacy cost at a faster $O(1/n)$ rate.

Importantly, our results imply concrete practical recommendations on sample collection, *regardless of the level of heterogeneity*. Indeed, increasing m will yield the most value in the i.i.d. setting and will yield no improvement when the users' distributions are arbitrary. As the real-world will lie somewhere in between, our results exhibit a regime where, for any heterogeneity, it is strictly better to collect more users (increasing n) than more samples per user (increasing m).

1.1 Our Contributions and Related Work

We provide a theoretical tool to construct estimators for tasks with user-level privacy constraints and apply it to a range of learning problems.

Optimal private mean estimation and uniformly concentrated queries (Section 3) We show that for a random variable in $[-B, B]$ concentrated in an unknown interval of radius τ (made precise in Definition 2), we can privately estimate its mean with error proportional to τ rather than B, as we would obtain using standard private mean estimation techniques such as Laplace mechanism [24]. When data is concentrated in ℓ_∞-norm, several papers show that one can achieve an error scaling with τ rather than B, either asymptotically [53], for Gaussian mean-estimation [40, 38], for sub-Gaussian symmetric distributions [18, 17] or for distributions with bounded p-th moment [39]. We propose a private mean estimator (Algorithm 2) with error scaling with τ that works in arbitrary dimension when data is concentrated in ℓ_2-norm (Theorem 2). In Corollary 1, we show it (optimally) solves mean estimation under user-level privacy constraints for random vectors bounded in ℓ_2-norm. In Appendix D.6, we show that for uniformly concentrated queries (see Definition 3), sequentially applying Algorithm 2 privately answers K adaptively chosen queries with privacy cost $\tilde{O}(\tau\sqrt{K}/n\varepsilon)$.

Our conclusions relate to the growing literature in adaptive data analysis. While a sequence of work [25, 9, 27, 28] use techniques from differential privacy and their answers are (ε, δ)-DP with $\varepsilon = \Theta(1)$, our work guarantees privacy for arbitrary ε with the additional assumption of uniform concentration.

Empirical risk minimization (Section 4) An influential line of papers studies ERM under item-level privacy constraints [19, 42, 8]. Importantly, these papers assume *arbitrary* data, i.e., not necessarily samples from users' distributions. The exact analog of ERM in the user-level setting is consequently less interesting as, for n data points $\{z_1, \ldots, z_n\}$, in the worst case, each user $u \in [n]$ contributes m copies of z_u and the problem reduces to the item-level setting. Instead, we consider the (related) problem of ERM when users contribute points sampled i.i.d. Assuming some regularity (A3 and A4), we develop and analyze algorithms for ERM under user-level DP constraints for convex, strongly-convex, and non-convex losses (Theorem 3).

Optimal stochastic convex optimization (Section 5) Under item-level DP (or equivalently, user-level DP with $m = 1$), a sequence of work [19, 8, 10, 11, 29] establishes the constrained minimax risk as $\tilde{\Theta}(1/\sqrt{n} + \sqrt{d}/(n\varepsilon))$. In this paper, with the additional assumptions that the losses are

individually smooth[2] and the gradients are sub-Gaussian random vectors, we prove matching upper (Theorem 4) and lower bounds (Theorem 5) of order $\tilde{\Theta}(1/\sqrt{nm} + \sqrt{d}/(n\sqrt{m}\varepsilon))$ in a regime we make precise. We leave closing the gap outside of this regime to future work.

Limit of learning with a fixed number of users (Appendix B) Finally, we resolve a conjecture of [4] and prove that with a fixed number of users, even in the limit $m \to \infty$ (i.e., each user has an infinite number of samples), we cannot reach zero error. In particular, we prove that for all the learning tasks we consider, the risk under user-level privacy constraints is at least $\Omega(e^{-\varepsilon n})$ regardless of m. Note that this does not contradict the results above since they require $n = \Omega((\log m)/\varepsilon)$.

Finally, we provide results in Appendix A for learning under *pure* user-level DP for function classes with finite metric entropy. We apply these to SCO with ℓ_∞ constraints (Remark 1) and achieve (near)-optimal rates.

2 Preliminaries

Notation. Throughout this work, d denotes the dimension, n the number of users, and m the number of samples per user. Generically, σ will denote the sub-Gaussian parameter, τ the concentration radius, ν the variance of a random vector and P a data distribution. We denote the optimization variable with $\theta \in \Theta \subset \mathbb{R}^d$, use z (or Z when random) to denote the data sample supported on a space $\mathcal{Z}$, and $\ell\colon \Theta \times \mathcal{Z} \to \mathbb{R}$ for the loss function. Gradients (denoted ∇) are always taken with respect to the optimization variable θ. For a convex set $\mathcal{C}$, $\Pi_{\mathcal{C}}$ denotes the euclidean projection on $\mathcal{C}$, i.e. $\Pi_{\mathcal{C}}(y) \coloneqq \operatorname{argmin}_{z\in\mathcal{C}} \|y - z\|_2$. We use A to refer to (possibly random) private mechanisms and X^n as a shorthand for the dataset $(X_1, \dots, X_n)$. For two distributions P and Q, we denote by $\|P - Q\|_{\mathsf{TV}}$ their total variation distance and $D_{\mathrm{kl}}(P\|Q)$ their Kullback-Leibler divergence. For a random vector $X \sim P$ supported on $\mathbb{R}^d$, we use $\mathrm{Var}(P)$ or $\mathrm{Var}(X)$ to denote $\mathbb{E}\left[\|X - \mathbb{E}[X]\|_2^2\right]$, which is equal to the trace of the covariance matrix of X.

Next, we consider differential privacy in the most general way, which only requires specifying a dataset space $\mathbb{S}$ and a distance d on $\mathbb{S}$.

Definition 1 (Differential Privacy)**.** *Let $\varepsilon, \delta \geq 0$. Let $\mathsf{A}\colon \mathbb{S} \to \Theta$ be a (potentially randomized) mechanism. We say that A is (ε, δ)-DP with respect to d if for any measurable subset $O \subset \Theta$ and all $S, S' \in \mathbb{S}$ satisfying $\mathrm{d}(S, S') \leq 1$,*

$$\mathbb{P}(\mathsf{A}(S) \in O) \leq e^{\varepsilon}\mathbb{P}(\mathsf{A}(S') \in O) + \delta. \tag{1}$$

If $\delta = 0$, we refer to this guarantee as pure differential privacy.

For a data space $\mathcal{Z}$, choosing $\mathbb{S} = \mathcal{Z}^n$ and $\mathrm{d}(S, S') = \mathrm{d}_{\mathsf{Ham}}(S, S') = \sum_{i=1}^n 1\{z_i \neq z_i'\}$ recovers the canonical setting considered in most of the literature—we refer to this as *item-level* differential privacy. When we wish to guarantee privacy for *users* rather than individual samples, we instead assume a *structured* dataset into which each of n users contributes $m > 1$ samples. This corresponds to $\mathbb{S} = (\mathcal{Z}^m)^n$ such that for $\mathcal{S} \in \mathbb{S}$, we have

$$\mathcal{S} = (S_1, \dots, S_n), \text{ where } S_u = \left\{z_1^{(u)}, \dots, z_m^{(u)}\right\} \text{ and } \mathrm{d}_{\mathsf{user}}(\mathcal{S}, \mathcal{S}') \coloneqq \sum_{u=1}^n 1\{S_u \neq S_u'\},$$

which means that, in this setting, two datasets are neighboring if at most one of the user's contributions differ. We henceforth refer to this setting as *user-level* differential privacy.

Distributional assumptions. In the case of user-level privacy with n users each providing m samples, we assume existence of a collection of distributions $\{P_u\}_{u\in[n]}$ over $\mathcal{Z}$. One then observes the following user-level dataset[3]

$$\mathcal{S} = (S_1, \dots, S_n) \text{ where } S_u \overset{\text{iid}}{\sim} P_u. \tag{2}$$

[2]We note that the results only require $\tilde{O}(n^{3/2})$-smooth losses. For large n—keeping all other problem parameters fixed—this is a very weak assumption. More precisely, when $n > \mathrm{poly}(d, m, 1/\varepsilon)$, our algorithm on a smoothed version $\tilde{\ell}$ of ℓ (e.g., using the Moreau envelope [33]) yields optimal rates for non-smooth losses. Whether the smoothness assumption can be removed altogether is an open question.

[3]For simplicity, we assume that $|S_u| = m$ but our guarantees directly extend to the setting where users have different number of samples with m replaced by $\mathsf{median}(m_1, \dots, m_n)$ using techniques from [46]. We leave eliciting the optimal rates in settings when m_u is an arbitrary random variable to future work.

In this paper, we consider the *limited heterogeneity* setting, i.e. when the users have related distributions. This setting is more reflective of practice, especially in light of growing interest towards federated learning applications [37, 60].

Assumption A1 (Limited heterogeneity setting)**.** *There exists a distribution P_0 over $\mathcal{Z}$ such that all the user distributions are close to P_0 in total variation distance, i.e.*

$$\max_{u\in[n]} \|P_u - P_0\|_{\mathsf{TV}} \le \Delta,$$

where $\Delta \ge 0$ quantifies the level of heterogeneity. Note that $\Delta = 0$ corresponds to assumption A2.

Note that our TV-based definition is natural in this setting as it is closely related to the notion of *discrepancy* (or d_A *distance*) which plays a key role in domain adaption scenarios [47, 12]. Lower bound results have been given in terms of the discrepancy measure (see [13]), which further justify the adoption of this definition in the presence of multiple distributions.

In the case that $\Delta = 0$, A1 reduces to the standard *homogeneous setting*. Many fundamental papers choose this setting when explicating minimax rates under constraints (e.g. in distributed optimization and federated learning [61] or under communication constraints [63, 15]).

Assumption A2 (Homogeneous setting)**.** *The distributions of individual users are equal, meaning there exists P_0 such that for all $u \in [n]$, $P_u = P_0$.*

In this paper, we develop techniques and provide matching upper and lower bounds for solving learning tasks in the homogeneous setting. In Appendix C, we prove that our techniques naturally apply to the heterogeneous setting in a black-box fashion, and for all considered problems provide meaningful guarantees under Assumption A1. Moreover, the algorithm achieves almost optimal rate whenever Δ is (polynomially) small. See the detailed statement in Theorem 9.

2.1 ERM and stochastic convex optimization

Assumptions on the loss. Throughout this work, we assume that the parameter space Θ is closed, convex, and satisfies $\|\theta - \vartheta\|_2 \le R$ for all $\theta, \vartheta \in \Theta$. We also assume that the loss $\ell\colon \Theta \times \mathcal{Z} \to \mathbb{R}$ is G-Lipschitz w.r.t. the ℓ_2-norm[4], meaning that for all $z \in \mathcal{Z}$, for all $\theta \in \Theta$, $\|\nabla \ell(\theta; z)\|_2 \le G$. We further consider the following assumptions.

Assumption A3. *The function $\ell(\cdot; z)$ is H-smooth. In other words, the gradient $\nabla\ell(\theta; z)$ is H-Lipschitz in the variable θ for all $z \in \mathcal{Z}$.*

Assumption A4. *The random vector $\nabla\ell(\theta; Z)$ is σ^2-sub-Gaussian for all $\theta \in \Theta$ and $Z \sim P_0$. Equivalently, for all $v \in \mathbb{R}^d$, $\langle v, \nabla\ell(\theta; Z)\rangle$ is a σ^2-sub-Gaussian random variable, i.e.,*

$$\mathbb{E}\left[\exp(\langle v, \nabla\ell(\theta; Z) - \mathbb{E}[\nabla\ell(\theta; Z)]\rangle)\right] \le \exp\big(\|v\|_2^2\sigma^2/2\big).$$

In this work, our rates often depend on the sub-Gaussianity and Lipschitz parameters σ and G, and thus we define the shorthands $\widetilde{G} \coloneqq \sigma\sqrt{d}$ and $\underline{G} \coloneqq \min\{G, \widetilde{G}\}$. Intuitively, the G-Lipschitzness assumption bounds the gradient in a ball around 0 (independently of θ), while sub-Gaussianity implies that, for each θ, $\nabla\ell(\theta; Z)$ likely lies in $\mathbb{B}_2^d(\nabla\mathcal{L}(\theta; P_0), \widetilde{G})$. Generically, there is no ordering between G and $\widetilde{G}$: for linear loss $\ell(\theta; z) = \langle\theta, z\rangle$, depending on P_0, it can hold that $G \ll \widetilde{G}$ (e.g., $P_0 = \mathsf{Unif}\{-v, v\}$ for $v \in \mathbb{R}^d$), $\widetilde{G} \ll G$ (e.g., P_0 is $\mathsf{N}(\mu, \sigma^2 I_d)$ truncated in a ball around μ, with $\|\mu\|_2 \gg \sigma\sqrt{d}$) or $G \approx \widetilde{G}$ (e.g., $P_0 = \mathsf{Unif}\{-1, +1\}^d$).

We introduce the tasks we consider in this work, namely empirical risk minimization (ERM) and stochastic convex optimization (SCO). For a collection of samples from n users $\mathcal{S} = (S_1, \ldots, S_n)$, where each $S_u = \{z_1^{(u)}, \ldots, z_m^{(u)}\} \in \mathcal{Z}^m$, we define the empirical risk objectives

$$\mathcal{L}(\theta; S_u) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \ell(\theta; z_i^{(u)}) \quad \text{and} \quad \mathcal{L}(\theta; \mathcal{S}) \coloneqq \frac{1}{n}\sum_{u=1}^{n} \mathcal{L}(\theta; S_u) = \frac{1}{mn}\sum_{u=1}^{n}\sum_{i=1}^{m} \ell(\theta; z_i^{(u)}). \tag{3}$$

In the user-level setting we wish to minimize $\mathcal{L}(\theta; \mathcal{S})$ under user-level privacy constraints. Going beyond the empirical risk, we also solve SCO [51], i.e. minimizing a convex population objective

[4] It is straightforward to develop analogs of the results of Sections 3 and 4 for arbitrary norms, but we restrict our attention to the ℓ_2 norm in this work for clarity.

when provided with samples from each users' distributions. In the user-level setting, for a convex loss ℓ and a convex constraint set Θ, we observe $\mathcal{S} = (S_1, \ldots, S_n) \sim \otimes_{u\in[n]}(P_u)^m$ and wish to

$$\underset{\theta\in\Theta}{\text{minimize}} \frac{1}{n}\sum_{u\in[n]} \mathcal{L}(\theta; P_u) \coloneqq \frac{1}{n}\sum_{u\in[n]} \mathbb{E}_{P_u}[\ell(\theta; Z)]. \tag{4}$$

In the homogeneous case (Assumption A2), this reduces to the classic SCO setting:

$$\underset{\theta\in\Theta}{\text{minimize}} \mathcal{L}(\theta; P_0) \coloneqq \mathbb{E}_{P_0}[\ell(\theta; Z)]. \tag{5}$$

2.2 Uniform concentration of queries

Let $\phi : \mathcal{Z} \to \mathbb{R}^d$ be a d-dimensional query function. We define concentration of random variables and uniform concentration of multiple queries as follows.

Definition 2. *A (random) sample X^n supported on $[-B, B]^d$ is (τ, γ)-concentrated (and we call τ the "concentration radius") if there exists $x_0 \in [-B, B]^d$ such that with probability at least $1 - \gamma$,*

$$\max_{i\in[n]} \|X_i - x_0\|_2 \le \tau.$$

Definition 3 (Uniform concentration of vector queries)**.** *Let $\mathcal{Q}_B^d = \{\phi\colon \mathcal{Z} \to [-B, B]^d\}$ be a family of queries with bounded range. For $Z^n = (Z_1, \ldots, Z_n) \overset{\text{iid}}{\sim} P$, we say that $(Z^n, \mathcal{Q}_B^d)$ is (τ, γ)-uniformly-concentrated if with probability at least $1 - \gamma$, we have*

$$\max_{i\in[n]} \sup_{\phi\in\mathcal{Q}_B^d} \left\|\phi(Z_i) - \mathbb{E}_{Z\sim P}[\phi(Z)]\right\|_2 \le \tau.$$

In this work, we will often consider σ^2-sub-Gaussian random variables (or vectors), which are concentrated according to Definition 2. For example, if X^n is drawn i.i.d. from a σ^2-sub-Gaussian random vector supported on $[-B, B]^d$, then it is $(\sigma\sqrt{d\log(2n/\gamma)}, \gamma)$-concentrated around its mean (see, e.g., [56]). Finally, we define a distance between random variables (and estimators).

Definition 4 (β-close Random Variables)**.** *For any two random variables $X_1 \sim P_1$ and $X_2 \sim P_2$, we say X_1 and X_2 are β-close, if $\|P_1 - P_2\|_{\mathsf{TV}} \le \beta$. We use the notation $X_1 \sim_\beta X_2$ if X_1 and X_2 are β-close.*

β-closeness is useful as, in many of our results, the private estimator we propose returns a simple unbiased estimate with high probability and is bounded otherwise. Thus, it suffices to do the analysis in the "nice" case and crudely bound the error otherwise.

3 High Dimensional Mean Estimation and Uniformly Concentrated Queries

In this section, we present a private mean estimator with privacy cost proportional to the concentration radius. Using these techniques, we show that, under uniform concentration, we answer adaptively-chosen queries with privacy cost proportional to the concentration radius instead of the whole range. Our theorems guarantee that the estimator is β-close (with β exponentially small in n) to a simple unbiased estimator with small noise. We further show how to directly translate these results into bounds on the estimator error, which we demonstrate by providing tight bounds on estimating the mean of ℓ_2-bounded random vectors under user-level DP constraints (Corollary 1).

Given i.i.d samples X^n from a distribution P supported on $\mathbb{R}^d$ with mean μ, the goal of mean estimation is to design a private estimator that minimizes the $\mathbb{E}\left[\|\mathsf{A}(X^n) - \mu\|_2^2\right]$. We focus on distributions with bounded supprot $[-B, B]^d$. However, our algorithm also generalize to the case when the mean is guaranteed to be in $[-B, B]^d$. In the user-level setting (in the homogeneous case), one observes a dataset $\mathcal{S}$ sampled as in (2) and wishes to minimize $\mathbb{E}[\|\mathsf{A}(\mathcal{S}) - \mathbb{E}P_0\|_2^2]$ under user-level privacy constraints. We first focus on the scalar case.

Mean estimation in one dimension. The algorithm uses a two-stage procedure, similar in spirit to those of [53], [40], and [39]. In the first stage of this procedure, we use the approximate median estimation in [27], detailed in Algorithm 6 in Appendix D.1, to privately estimate a crude interval

Algorithm 1 WinsorizedMean1D$(X^n, \varepsilon, \tau, B)$: Winsorized Mean Estimator (WME)

Require: $X^n := (X_1, X_2, ..., X_n) \in [-B, B]^n$, τ : concentration radius, privacy parameter $\varepsilon > 0$.

1: $[a, b] =$ **PrivateRange**$(X^n, \varepsilon/2, \tau, B)$ with $|b - a| = 4\tau$. {Algorithm 6 in Appendix D.1. }

2: Sample $\xi \sim \text{Lap}\big(0, \frac{8\tau}{\varepsilon n}\big)$ and return

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \Pi_{[a,b]}(X_i) + \xi,$$

where $\Pi_{[a,b]}(x) = \max\{a, \min\{x, b\}\}$.

in which the means lie, with accuracy $\Theta(\tau)$. The second stage clips the mean around this interval, reducing the sensitivity from $O(B)$ to $O(\tau)$, and adds the appropriate Laplace noise. With high probability, we can recover the guarantee of the Laplace mechanism with smaller sensitivity since the samples are concentrated in a radius τ. We present the formal guarantees of Algorithm 1 in Theorem 1 and defer its proof to Appendix D.2.

Theorem 1. *Let X^n be a dataset supported on $[-B, B]$. The output of Algorithm 1, denoted by $\mathsf{A}(X^n)$, is ε-DP. Furthermore, if X^n is (τ, γ)-concentrated, it holds that*

$$\mathsf{A}(X^n) \sim_\beta \frac{1}{n}\sum_{i=1}^{n} X_i + \mathit{Lap}\left(\frac{8\tau}{n\varepsilon}\right),$$

where $\beta = \min\left\{1, \gamma + \frac{B}{\tau}\exp\left(-\frac{n\varepsilon}{8}\right)\right\}$. Moreover, Algorithm 1 runs in time $\tilde{O}(n + \log(B/\tau))$.

Compared to [40, 38, 39], our algorithm runs in time $\tilde{O}(n + \log(B/\tau))$ instead of $\tilde{O}(n + B/\tau)$ owing to the approximate median estimation algorithm in [27], which is faster when $\tau \ll B$.

Mean estimation in arbitrary dimension. In the general d-dimensional case, if X^n is concentrated in ℓ_∞-norm, one simply applies Algorithm 1 to each dimension. However, when X^n is concentrated in ℓ_2-norm, naively upper bounding ℓ_∞-norm by the ℓ_2-norm will incur a superfluous $\sqrt{d}$ factor: if $\|v\|_2 \le \rho$, each $|v_j|$ is possibly as large as ρ. To remedy this issue, we use the random rotation trick in [3, 54]. This guarantees that all coordinates have roughly the same range: for $v \in \mathbb{R}^d$, with high probability, $\|Rv\|_\infty \le \tilde{O}(\|v\|_2/\sqrt{d})$, where R is the random rotation. We present this procedure in Algorithm 2 and its performance in Theorem 2.

Algorithm 2 WinsorizedMeanHighD$(X^n, \varepsilon, \delta, \tau, B, \gamma)$: WME - High Dimension

Require: $X^n := (X_1, X_2, ..., X_n), X_i \in [-B, B]^d$, τ, γ: concentration radius and probability, privacy parameter $\varepsilon, \delta > 0$.

1: Let $D = \mathsf{Diag}(\omega)$ where ω is sampled uniformly from $\{\pm 1\}^d$.

2: Set $U = d^{-1/2}\mathbf{H}D$, where $\mathbf{H}$ is a d-dimensional Hadamard matrix. For all $i \in [n]$, compute

$$Y_i = UX_i.$$

3: Let $\varepsilon' = \frac{\varepsilon}{\sqrt{8d\log(1/\delta)}}, \tau' = 10\tau\sqrt{\frac{\log(dn/\gamma)}{d}}$. For $j \in [d]$, compute

$$\bar{Y}(j) = \textbf{WinsorizedMean1D}\Big(\{Y_i(j)\}_{i\in[n]}, \varepsilon', \tau', \sqrt{d}B\Big).$$

4: **return** $\bar{X} = U^{-1}\bar{Y}$.

Theorem 2. *Let $\mathsf{A}(X^n) =$ **WinsorizedMeanHighD**$(X^n, \varepsilon, \delta, \tau, B, \gamma)$ be the output of Algorithm 2. $\mathsf{A}(X^n)$ is (ε, δ)-DP. Furthermore, if X^n is (τ, γ)-concentrated in ℓ_2-norm, there exists an estimator $\mathsf{A}'(X^n)$ such that $\mathsf{A}(X^n) \sim_\beta \mathsf{A}'(X^n)$ and*

$$\mathbb{E}[\mathsf{A}'(X^n)|X^n] = \frac{1}{n}\sum_{i=1}^{n} X_i \;\; \textit{and} \;\; \mathrm{Var}(\mathsf{A}'(X^n)|X^n) \le c_0\frac{d\tau^2\log(dn/\alpha)\log(1/\delta)}{n^2\varepsilon^2}, \tag{6}$$

where $c_0 = 102,400$ and $\beta = \min\left\{1, 2\gamma + \frac{d^2B\sqrt{\log(dn/\gamma)}}{\tau}\exp\left(-\frac{n\varepsilon}{24\sqrt{d\log(1/\delta)}}\right)\right\}$.

We present the proof of Theorem 2 in Appendix D.3. We are able to transfer both Theorem 1 and Theorem 2 into finite-sample estimation error bounds for various types of concentrated distributions

and obtain near optimal guarantees (see Appendix D.5 for an example in mean estimation of sub-Gaussian distributions). The next corollary characterizes the risk of mean estimation for distributions supported on an ℓ_2-bounded domain with user-level DP guarantees (see Appendix D.4 for the proof).

Corollary 1. *Assume A2 holds with P_0 supported on $\mathbb{B}_2^d(0, B)$ with mean μ. Given $\mathcal{S} = (S_1, S_2, ..., S_n)$, $|S_u| = m$, consisting of m i.i.d. samples from P_u. There exists an (ε, δ)-user-level DP algorithm $\mathsf{A}(\mathcal{S})$ such that, if $n \geq (c_1\sqrt{d\log(1/\delta)}/\varepsilon)\log(m(dn + n^2\varepsilon^2))$ for a numerical constant c_1, we have*[5]

$$\mathbb{E}\left[\|\mathsf{A}(\mathcal{S}) - \mu\|_2^2\right] = \frac{\mathrm{Var}(P_0)}{mn} + \tilde{O}\left(\frac{dB^2}{mn^2\varepsilon^2}\right).$$

Note that $\mathrm{Var}(P_0) \leq B^2$ for any P_0 supported on $\mathbb{B}_2^d(0, B)$. Replacing $\mathrm{Var}(P_0)$ by B^2, the bound is minimax optimal up to logarithmic factors. When only A1 holds with $\Delta \leq \mathsf{poly}(d, \frac{1}{n}, \frac{1}{m}, \frac{1}{\varepsilon})$, the same error bounds holds (up to constant) for estimating $\mathbb{E}_{Z\sim P_u}[Z]$ for any $u \in [n]$.

Note that algorithms in [38, 39], which focus on estimating the mean of d-dimensional subGaussian distributions, can also be used to estimate the mean of ℓ_2-bounded distributions since bounded random variables are also subGaussian. However, applying these algorithms directly will incur a superfluous d factor in the mean square error. We void this using the random rotation trick in Algorithm 2.

Answering multiple queries. We end this section by noting that, when a family of queries $\mathcal{Q}$ is uniformly concentrated (as made precise in Definition 3), we answer sequences of K d-dimensional, adaptively chosen queries with error scaling as $\tilde{O}(\sqrt{dK}\tau/(n\varepsilon))$ by applying Algorithm 2 to $\{\phi_k(Z_i)\}_{i\in[n]}$ with the right $(\varepsilon_0, \delta_0)$. We make this formal in Theorem 10 in Appendix D.6.

4 Empirical Risk Minimization with User-Level Differential Privacy

In this section, we present an algorithm to solve the ERM objective of (3) under user-level DP constraints. We apply the results of Section 3 by noting that the SQ framework encompasses stochastic gradient methods. Informally, one can sequentially choose queries $\phi_k(z) = \nabla\ell(\theta_k; z)$ and, for a stepsize η, update $\theta_{k+1} = \Pi_\Theta(\theta_k - \eta v_k)$, where v_k is the answer to the k-th query. For the results to hold, we require a uniform concentration result over the appropriate class of queries.

Uniform concentration of stochastic gradients The class of queries for stochastic gradient methods is $\mathcal{Q}_{\mathsf{erm}} := \{\nabla\ell(\theta; \cdot) : \theta \in \Theta\}$. We prove that when assumptions A3 and A4 hold, $(\{\nabla\ell(\cdot; S_u)\}_{u\in[n]}, \mathcal{Q}_{\mathsf{erm}})$ is $(\tilde{O}(\sigma\sqrt{d/m}), \alpha)$-uniformly concentrated. The next proposition is a simplification of the result of [50] under the (stronger) assumption A3 that ℓ is uniformly H-smooth. The proof, which we defer to Appendix E.1, hinges on a covering number argument.

Proposition 1 (Concentration of random gradients)**.** *Let $S_u \overset{\text{iid}}{\sim} P_u$, $|S_u| = m$ for $u \in [n]$ and $\alpha \geq 0$. Under Assumptions A3 and A4, with probability greater than $1 - \alpha$ it holds that*

$$\max_{u\in[n]} \sup_{\theta\in\Theta} \|\nabla\mathcal{L}(\theta; S_u) - \nabla\mathcal{L}(\theta; P_u)\|_2 = O\left(\sigma\sqrt{\frac{d\log\left(\frac{RHm}{d\sigma}\right)}{m} + \frac{\log\left(\frac{n}{\alpha}\right)}{m}}\right).$$

Stochastic gradient methods We state classical convergence results for stochastic gradient methods for both convex and non-convex losses under smoothness. For a function $F : \Theta \to \mathbb{R}$, we assume access to a first-order stochastic oracle O_{F,ν^2}, i.e., a random mapping such that for all $\theta \in \Theta$,

$$\mathsf{O}_{F,\nu^2}(\theta) = \nabla\widehat{F}(\theta) \text{ with } \mathbb{E}\left[\nabla\widehat{F}(\theta)\right] = \nabla F(\theta) \text{ and } \mathrm{Var}\left(\nabla\widehat{F}(\theta)\right) \leq \nu^2.$$

We abstract optimization algorithms in the following way: an algorithm consists of an output set $\mathcal{O}$, a sub-routine $\mathsf{Query} : \mathcal{O} \to \Theta$ that takes the last output and indicates the next point to query and a sub-routine $\mathsf{Update} : \mathcal{O} \times \mathbb{R}^d \to \mathcal{O}$ that takes the previous output and a stochastic gradient and returns the next output. After T steps, we call $\mathsf{Aggregate} : \mathcal{O}^* \to \Theta$, which takes all the previous outputs and returns the final point. (See Algorithm 7 in Appendix E.2 for how to instantiate generic first-order optimization in this framework.) We detail in Proposition 4 in Appendix E.2 standard convergence results for variations of (projected) stochastic gradient descent (SGD). We introduce this abstraction to forego the details of each specific algorithm and instead focus on the privacy and utility guarantees.

[5] For precise log factors, see Appendix D.4.

Algorithm We recall the ERM setting with user-level DP. We observe $\mathcal{S} = (S_1, \dots, S_n)$ with $S_u \in \mathcal{Z}^m$ for $u \in [n]$ and wish to solve the constrained optimization problem with objective in (3). We present our method in Algorithm 3 and provide utility and privacy guarantees in Theorem 3.

Algorithm 3 Winsorized First-Order Optimization

Input: Number of iterations T, optimization algorithm $\{\mathcal{O}, \mathsf{Query}, \mathsf{Update}, \mathsf{Aggregate}\}$, privacy parameters (ε, δ), data $\mathcal{S} = (S_1, \dots, S_n)$, initial output o_0, parameter set Θ, concentration radius τ, probability γ.
Set $\varepsilon' = \frac{\varepsilon}{2\sqrt{2T\log(2/\delta)}}$ and $\delta' = \frac{\delta}{2T}$
for $t = 0, \dots, T-1$ **do**
  $\theta_t \leftarrow \mathsf{Query}(o_t)$.
  For each user $u \in [n]$, compute
$$g_t^{(u)} = \nabla \mathcal{L}(\theta_t; S_u) = \frac{1}{m} \sum_{j \in [m]} \nabla \ell(\theta_t; z_j^{(u)}).$$
  Compute $\bar{g}_t = \textbf{WinsorizedMeanHighD}(\{g_t^{(u)}\}_{u \in [n]}, \varepsilon', \delta', \tau, G, \gamma)$.
  $o_{t+1} \leftarrow \mathsf{Update}(o_t, \bar{g}_t)$.
end for
return $\bar{\theta} \leftarrow \mathsf{Aggregate}(o_0, \dots, o_T)$.

Theorem 3 (Privacy and utility guarantees for ERM)**.** *Assume A2 holds and recall that $\widetilde{G} = \sigma\sqrt{d}$, assume[6] $n = \tilde{\Omega}(\sqrt{dT}/\varepsilon)$ and let $\widehat{\theta}$ be the output of Algorithm 3. There exists variants of projected SGD (e.g. the ones we present in Proposition 4) such that, with probability greater than $1-\gamma$:*

(i) If for all $z \in \mathcal{Z}$, $\ell(\cdot; z)$ is convex, then
$$\mathbb{E}\left[\mathcal{L}(\widehat{\theta}; \mathcal{S}) - \inf_{\theta' \in \Theta} \mathcal{L}(\theta'; \mathcal{S}) \,\middle|\, \mathcal{S}\right] = \tilde{O}\left(\frac{R^2 H}{T} + R\widetilde{G}\frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\right).$$

(ii) If for all $z \in \mathcal{Z}$, $\ell(\cdot; z)$ is μ-strongly-convex, then
$$\mathbb{E}\left[\mathcal{L}(\widehat{\theta}; \mathcal{S}) - \inf_{\theta' \in \Theta} \mathcal{L}(\theta'; \mathcal{S}) \,\middle|\, \mathcal{S}\right] = \tilde{O}\left(GR\exp\left(-\tfrac{\mu}{H}T\right) + \widetilde{G}^2\frac{d}{\mu n^2 m \varepsilon^2}\right).$$

(iii) Otherwise, defining the gradient mapping[7] $\mathsf{G}_{F,\gamma}(\theta) := \frac{1}{\gamma}[\theta - \Pi_\Theta(\theta - \gamma\nabla F(\theta))]$*, we have*
$$\mathbb{E}\left[\|\mathsf{G}_{\mathcal{L}(\cdot;\mathcal{S}),1/H}(\widehat{\theta})\|_2^2 | \mathcal{S}\right] = \tilde{O}\left(\frac{H^2 R}{T} + HR\widetilde{G}\frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\right).$$

For $\varepsilon \le 1, \delta > 0$, Algorithm 3 instantiated with any first-order gradient algorithm is (ε, δ)-user-level DP. In the case that only A1 holds, the same guarantees hold whenever $\Delta \le \mathsf{poly}(d, \frac{1}{n}, \frac{1}{m}, \frac{1}{\varepsilon})$.

We present the proof in Appendix E.3. For the utility guarantees, the crux of the proof resides in Theorem 10: as well as ensuring small excess loss in expectation, the SQ algorithm produces with high probability a sample from the stochastic gradient oracle $\mathsf{O}_{\mathcal{L}(\cdot;\mathcal{S}),\nu^2}$ where $\nu^2 = \tilde{O}(T\widetilde{G}^2\frac{d}{n^2 m \varepsilon^2})$. When this happens for all T steps, the analysis of stochastic gradient methods provide the desired regret. The privacy guarantees follow from the strong composition theorem of [23].

Importantly, when the function exhibits (some) strong-convexity (which will be the case for any regularized objective), we are able to *localize* the optimal parameter—up to the privacy cost—in $\tilde{O}(H/\mu)$ steps. This will be particularly important in Section 5.

Corollary 2 (Localization)**.** *Let $\widehat{\theta}$ be the output of Algorithm 3 on the ERM problem of (3). Assume that $\ell(\cdot; z)$ is μ-strongly-convex for all $z \in \mathcal{Z}$, that $n = \tilde{\Omega}(\sqrt{dH/\mu})$ and set $T = \frac{H}{\mu}\log\left(n^2 m(\underline{G}/\widetilde{G}^2)\frac{\mu R \varepsilon^2}{d}\right)$ and $\gamma = \frac{\sigma^2 d^2}{\mu^2 n^2 m \varepsilon^2 R^2}$. For $\theta^*_{\mathcal{S}} \in \operatorname{argmin}_{\theta' \in \Theta} \mathcal{L}(\theta'; \mathcal{S})$, it holds*[8]

[6] For precise log factors, see Appendix E.3.

[7] In the unconstrained case—$\Theta = \mathbb{R}^d$—this corresponds to an ϵ-stationary point as $\mathsf{G}_{F,\gamma}(x) = \nabla F(x)$.

[8] A logarithmic dependence on T is hiding in the result. Since $T = \tilde{O}(H/\mu)$, we implicitly assume H/μ is polynomial in the stated parameters, which is satisfied when we later apply these results to regularized objectives.

$$\mathbb{E}[\|\widehat{\theta} - \theta^*_{\mathcal{S}}\|_2^2] = \tilde{O}\left(\frac{\sigma^2 d^2}{\mu^2 n^2 m \varepsilon^2}\right).$$

5 Stochastic Convex Optimization with User-level Privacy

In this section we address the SCO task of (5) under user-level DP constraints. Our approach (which we show in Algorithm 4) solves a sequence of carefully regularized ERM problems, drawing on the guarantees of the previous section. Recall that $\widetilde{G} = \sigma\sqrt{d}$ and $\underline{G} = \min\{G, \widetilde{G}\}$, and that ℓ is H-smooth under assumption A3. In this section, we assume that ℓ is convex. We first present our results and state an upper and lower bound for SCO with user-level privacy constraints.

Theorem 4 (Phased ERM for SCO)**.** *Algorithm 4 is user-level (ε, δ)-DP. When A2 holds and $n = \tilde{\Omega}(\min\{\sqrt[3]{d^2 m H^2 R^2/(G\underline{G}\varepsilon^4)}, HR\sqrt{m}/(\sigma\varepsilon)\})$, or, equivalently, $H = \tilde{O}(\sqrt{\frac{n^2\varepsilon^2\sigma^2}{R^2 m} + \frac{G\underline{G}n^3\varepsilon^4}{d^2 R^2 m}})$ for all P and ℓ satisfying Assumptions A3 and A4, we have*

$$\mathbb{E}\left[\mathcal{L}(\mathsf{A}_{\mathsf{PhasedERM}}(\mathcal{S}); P_0)\right] - \min_{\theta' \in \Theta} \mathcal{L}(\theta'; P_0) = \tilde{O}\left(\frac{R\sqrt{G\underline{G}}}{\sqrt{mn}} + R\widetilde{G}\frac{\sqrt{d}}{n\sqrt{m\varepsilon}}\right).$$

Furthermore, our results still hold in the heterogeneous setting (Assumption A1) whenever $\Delta \leq \mathsf{poly}(d, \frac{1}{n}, \frac{1}{m}, \frac{1}{\varepsilon})$; the risk guarantee being with respect to any user distribution P_u.

Theorem 5 (Lower bound for SCO)**.** *There exists a distribution P and a loss ℓ satisfying Assumptions A3 and A4 such that for any algorithm A satisfying (ε, δ)-DP at user-level, we have*

$$\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}); P)\right] - \min_{\theta' \in \Theta} \mathcal{L}(\theta'; P) = \Omega\left(\frac{R\underline{G}}{\sqrt{mn}} + R\underline{G}\frac{\sqrt{d}}{n\sqrt{m\varepsilon}}\right).$$

When $G = \Theta(\sigma\sqrt{d})$, the upper bound matches the lower bound up to logarithmic factors. We present the algorithm and proof for Theorem 4 in Section 5.1. Theorem 5 is proved in Section 5.2.

5.1 Upper bound: minimizing a sequence of regularized ERM problems

We now present Algorithm 4, which achieves the upper bound of Theorem 4. It is similar in spirit to Phased ERM [29] and EpochGD [34], in that at each round we minimize a regularized ERM problem with fresh samples and increased regularization, initializing each round from the final iterate of the previous round. This allows us to localize the optimum with exponentially increasing accuracy without blowing up our privacy budget. We solve each round using Algorithm 3 to guarantee privacy and obtain an *approximate* minimizer. We show the guarantee in Corollary 2 is enough to achieve optimal rates. We provide the proof of Theorem 4 in Appendix F and present a sketch here.

Algorithm 4 $\mathsf{A}_{\mathsf{PhasedERM}}$: Phased ERM

Require: Private dataset: $\mathcal{S} = (S_1, \ldots, S_n) \in (\mathcal{Z}^m)^n$: $n \times m$ i.i.d samples from P, H-smooth, convex loss function ℓ, convex set $\Theta \subset \mathbb{R}^d$, privacy parameters $\varepsilon \leq 1, \delta \leq 1/n^2$, sub-Gaussian parameter σ.

1: Set $T = \lceil \log_2(\frac{Gn\sqrt{m}\varepsilon}{\sigma d}) \rceil$, $\lambda = \sqrt{\frac{G\underline{G}}{nm} + \frac{\sigma^2 d^2}{n^2 m \varepsilon^2}}/R$
2: **for** $t = 1$ to T **do**
3: Set $n_t = \frac{n}{2^t}$, $\lambda_t = 4^t\lambda$.
4: Sample $\mathcal{S}_t$, n_t users that have not participated in previous rounds. Using Algorithm 3, compute an approximate minimizer $\widehat{\theta}_t$, to the accuracy of Corollary 2, for the objective

$$\mathcal{L}_{\lambda_t, \widehat{\theta}_{t-1}}(\theta; \mathcal{S}_t) = \frac{1}{mn_t}\sum_{u \in \mathcal{S}_t}\sum_{j=1}^{m} \ell(\theta, z_j^{(u)}) + \frac{\lambda_t}{2}\|\theta - \widehat{\theta}_{t-1}\|_2^2. \tag{7}$$

5: **end for**
6: **return** $\widehat{\theta}_T$.

Proof sketch of Theorem 4. The privacy guarantee comes directly from the privacy guarantee of Algorithm 3 and the fact that $\mathcal{S}_t$ are non-overlapping. The proof for utility is similar to the proof

of Theorem 4.8 in [29]. In round t of Algorithm 4, we consider the true minimizer θ_t^* and the approximate minimizer $\widehat{\theta}_t$. By stability [14], we can bound the generalization error of θ_t^* (see Proposition 5 in Appendix F) and, by Corollary 2, we can bound $\mathbb{E}\|\widehat{\theta}_t - \theta_t^*\|_2^2$. We finally choose $\{(\lambda_t, n_t)\}_{t \le T}$ such that the assumptions of Corollary 2 hold and to minimize the final error. □

5.2 Lower bound: SCO is harder than Gaussian mean estimation

First of all, note that it suffices to prove the lower bounds in the homogeneous setting as any level of heterogeneity only makes the problem harder. Theorem 5 holds for (ε, δ)-user-level DP—importantly, this is a setting for which lower bounds are generally more challenging (we provide a related lower bound for ε-user-level DP in Appendix A.2). We present the proof in Appendix F.2 and a sketch here.

Proof sketch of Theorem 5. The (constrained) minimax lower bound decomposes into a statistical rate and a privacy rate. The statistical rate is optimal (see, e.g., [44, 2]), thus we focus on the privacy rate. We consider linear losses of the form $\ell(\theta; z) = -\langle \theta, z \rangle$. We show that optimizing $\mathcal{L}(\theta; P) = \mathbb{E}_P[\ell(\theta; Z)]$ over $\theta \in \Theta$ is harder than the mean estimation task for P. Intuitively, $\mathcal{L}(\theta; P) = -\langle \theta, \mathbb{E}Z \rangle$ attains its minimum at $\theta^* = R\mathbb{E}[Z]/\|\mathbb{E}[Z]\|_2$ and finding θ^* provides a good estimate of (the direction of) $\mathbb{E}[Z]$. We make this formal in Proposition 6. Next, for Gaussian mean estimation, we reduce, in Proposition 3, user-level DP to item-level DP with lower variance by having each user contribute their sample average (which is a sufficient statistic). We conclude with the results of [38] (see Proposition 7) by proving in Corollary 6 that estimating the direction of the mean with item-level privacy is hard. □

Acknowledgments

The authors would like to thank Hilal Asi and Karan Chadha for comments on an earlier draft as well as Yair Carmon, Peter Kairouz, Gautam Kamath, Sai Praneeth Karimireddy, Thomas Steinke and Sebastian Stich, for useful discussions and pointers to very relevant references.

References

[1] J. Acharya, Z. Sun, and H. Zhang. Differentially private Assouad, Fano, and Le Cam. In V. Feldman, K. Ligett, and S. Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pages 48–78. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/acharya21a.html.

[2] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[3] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563, 2006.

[4] K. Amin, A. Kulesza, A. Munoz, and S. Vassilvtiskii. Bounding user contributions: A bias-variance trade-off in differential privacy. In *International Conference on Machine Learning*, pages 263–271, 2019.

[5] Apple Privacy Team. Learning with privacy at scale, 2017. Available at https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html.

[6] S. Augenstein, H. B. McMahan, D. Ramage, S. Ramaswamy, P. Kairouz, M. Chen, R. Mathews, and B. A. y Arcas. Generative models for effective ml on private, decentralized datasets. In *International Conference on Learning Representations*, 2019.

[7] R. F. Barber and J. C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv:1412.4451 [math.ST]*, 2014.

[8] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[9] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016.

[10] R. Bassily, V. Feldman, K. Talwar, and A. G. Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pages 11279–11288, 2019.

[11] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4381–4391. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/2e2c4bf7ceaa4712a72dd5ee136dc9a8-Paper.pdf.

[12] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems 20*, 2007.

[13] S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

[14] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[15] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the Forty-Eigth Annual ACM Symposium on the Theory of Computing*, 2016. URL https://arxiv.org/abs/1506.07216.

[16] S. Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

[17] M. Bun and T. Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. In *Advances in Neural Information Processing Systems*, pages 181–191, 2019.

[18] T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

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

[20] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[21] B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3571–3580, 2017.

[22] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pages 265–284, 2006.

[23] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

[24] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. 2014.

[25] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015.

[26] A. Epasto, M. Mahdian, J. Mao, V. Mirrokni, and L. Ren. Smoothly bounding user contributions in differential privacy. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13999–14010. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a0dc078ca0d99b5ebb465a9f1cad54ba-Paper.pdf.

[27] V. Feldman and T. Steinke. Generalization for adaptively-chosen estimators via stable median. In S. Kale and O. Shamir, editors, *ICML*, volume 65 of *Proceedings of Machine Learning Research*, pages 728–757, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

[28] V. Feldman and T. Steinke. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory*, pages 535–544. PMLR, 2018.

[29] V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[30] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, New York, NY, USA, 2015. ACM. doi: 10.1145/2810103.2813677. URL http://doi.acm.org/10.1145/2810103.2813677.

[31] B. Ghazi, R. Kumar, and P. Manurangsi. User-level private learning via correlated sampling. *arXiv preprint arXiv:2110.11208*, 2021.

[32] Google. Enabling developers and organizations to use differential privacy, 2019. Available at https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html.

[33] C. Guzmán and A. Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015.

[34] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436, 2011.

[35] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):e1000167, 2008.

[36] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv:1902.03736 [math.PR]*, 2019.

[37] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. ISSN 1935-8237. doi: 10.1561/2200000083. URL http://dx.doi.org/10.1561/2200000083.

[38] G. Kamath, J. Li, V. Singhal, and J. Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

[39] G. Kamath, V. Singhal, and J. Ullman. Private mean estimation of heavy-tailed distributions. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2204–2235. PMLR, 09–12 Jul 2020.

[40] V. Karwa and S. Vadhan. Finite sample differentially private confidence intervals. *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, 2018.

[41] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the Association for Computing Machinery*, 45(6):983–1006, 1998.

[42] D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

[43] A. Kulunchakov and J. Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research*, 21(155):1–52, 2020.

[44] D. Levy and J. C. Duchi. Necessary and sufficient geometries for gradient methods. In *Advances in Neural Information Processing Systems 32*, 2019. URL https://arxiv.org/abs/1909.10455.

[45] J. Liu and K. Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 298–309, 2019.

[46] Y. Liu, A. Theertha Suresh, F. Yu, S. Kumar, and M. Riley. Learning discrete distributions: user vs item-level privacy. In *Advances in Neural Information Processing Systems*, 2020.

[47] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of the Twenty Second Annual Conference on Computational Learning Theory*, 2009.

[48] H. B. McMahan, G. Andrew, U. Erlingsson, S. Chien, I. Mironov, N. Papernot, and P. Kairouz. A general approach to adding differential privacy to iterative training procedures. *arXiv preprint arXiv:1812.06210*, 2018.

[49] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

[50] S. Mei, Y. Bai, A. Montanari, et al. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[51] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Conference on Learning Theory*, 2009.

[52] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1310–1321, 2015.

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

[54] A. T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan. Distributed mean estimation with limited communication. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3329–3337, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[55] United States Census Bureau. Statistical safeguards, 2018. Available at https://www.census.gov/about/policies/privacy/statistical_safeguards.html.

[56] R. Vershynin. *High Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2019.

[57] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[58] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2512–2520. IEEE, 2019.

[59] R. J. Wilson, C. Y. Zhang, W. Lam, D. Desfontaines, D. Simmons-Marengo, and B. Gipson. Differentially private SQL with bounded user contribution. *Proceedings on Privacy Enhancing Technologies*, 2:230–250, 2020.

[60] B. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[61] B. E. Woodworth, J. Wang, A. Smith, B. McMahan, and N. Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. In *Advances in Neural Information Processing Systems 31*, 2018.

[62] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

[63] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/d6ef5f7fa914c19931a55bb262ec879c-Paper.pdf.

Discussion

In this work, we explore the fundamental limits of learning under user-level privacy constraints. Importantly, we provide practical algorithms with significantly improved privacy cost in the regime where the number of samples per user $m \gg 1$. However, our work provides generalization guarantees under a limited heterogeneity assumption. Extending our work to more heterogeneous settings is an interesting research direction. Secondly, our work focuses on establishing information-theoretic limits and we do not optimize the runtime of our algorithms. For example, in the case of SCO, our algorithm runs in $\min\{(nm)^{3/2}, n^2m^{3/2}/\sqrt{d}\}$ time, while achieving the optimal item-level private rate requires at most $\min\{nm, (nm)^2/d\}$ time [10]. Developing faster algorithms in these settings is a possible future direction.

Potential negative societal impact

Our work is theoretical in nature and we do not foresee major direct negative societal consequences. Because of the growing prevalence of data collection from all sources (mobile, browser, medical records etc.), providing meaningful guarantees—such as user-level DP—while preserving adequate accuracy is an important direction of research. Our work suffers from the same potential negative impact as any work in the broad differential privacy area in two ways: first, a simple way to guarantee privacy is to limit data collection or delete data the users provided in the past. Second, the guarantees we provide are contingent on careful choices of ε and δ as well as rigorous and independent methodologies for evaluating the privacy of deployed models.

A Function Classes with Bounded Metric Entropy under Pure DP

We consider the general task of learning hypothesis class with finite metric entropy (i.e., such that there exists a finite Δ-cover under a certain norm) and bounded loss under *pure* user-level DP constraints.

For this setting, we present Algorithm 5, which we complement with an information-theoretic lower bound. As in the previous sections, we consider a sample set $\mathcal{S} = (S_1, \dots, S_n)$, with $S_u = \{z_j^{(u)}\}_{j\in[m]} \subset \mathcal{Z}$. We begin by considering the case of a finite parameter space: for $K \in \mathbb{N}, K < +\infty$, we have

$$\Theta = \left\{\theta^{(1)}, \dots, \theta^{(K)}\right\}. \tag{8}$$

For $0 \le B < \infty$, we denote $\mathcal{F}_B \coloneqq \{\ell \colon \Theta \times \mathcal{Z} \to \mathbb{R} : \|\ell\|_\infty \le B\}$ the set of B-bounded functions and $\mathcal{A}_\varepsilon^{\mathsf{user}}$ the set ε-user-level DP estimators from $\mathcal{Z}^n$ to Θ, the goal of this section is to elicit the constrained minimax rate [62, 7, 1]

$$\mathfrak{M}_{m,n}^{\mathsf{user}}(\Theta, \mathcal{F}_B, \varepsilon) \coloneqq \sup_{\mathcal{Z}, \mathcal{P}\subset\mathcal{P}(\mathcal{Z})} \inf_{\mathsf{A}\in\mathcal{A}_\varepsilon^{\mathsf{user}}} \sup_{\ell\in\mathcal{F}_B, P\in\mathcal{P}} \mathbb{E}_{\mathcal{S}\overset{\text{iid}}{\sim}(P^m)^n}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}); P) - \inf_{\theta\in\Theta}\mathcal{L}(\theta; P)\right].$$

We start with providing the estimator, which combines the private mean estimator of Section 3 with the private selection techniques of [45]. Given a collection of ε-DP mechanisms, the latter provides an ε-DP way to find an (approximate) minimum by sampling from each mechanism at random *with the same data* and returning the maximum of the values observed. In our setup, each mechanism A_k will be a private release of $\mathcal{L}(\theta^{(k)}; \mathcal{S})$.

A.1 Combining mean estimation and private selection

Our first step is to show that the conditions of Section 3 are met, that is, the data are concentrated with high probability.

Lemma 1. *Let $\mathcal{S} = (S_1, \dots, S_n) \overset{\text{iid}}{\sim} (P^m)^n$ and $\alpha \in (0, 1)$. With probability greater than $1 - \alpha$, it holds that*

$$\max_{k\in K}\max_{u\in[n]}\left|\mathcal{L}(\theta^{(k)}; S_u) - \mathcal{L}(\theta^{(k)}; P)\right| \le \frac{B}{2}\sqrt{\frac{\log(|\Theta|\cdot n) + \log(2/\alpha)}{m}}. \tag{9}$$

In other words, $(\mathcal{S}, \mathcal{Q}_\Theta)$ is $(B/(2\sqrt{m})\sqrt{\log(2Kn/\alpha)}, \alpha)$ uniformly concentrated where $\mathcal{Q}_\Theta = \{\mathcal{L}(\theta; \cdot) : \theta \in \Theta\}$.

Proof. The proof is straightforward: for a fixed $\theta^{(k)} \in \Theta$ and $u \in [n]$, the random variable $\mathcal{L}(\theta^{(k)}; S_u)$ is $\frac{B^2}{4m}$-sub-Gaussian around its mean $\mathcal{L}(\theta^{(k)}; P)$. A union bound over the samples and parameters concludes the proof. □

Conditioned on that event, the data are well concentrated and the results of Theorem 1 apply. We now describe the algorithm and then go on to prove privacy and utility guarantees. We call it "idealized" because it is not computationally efficient. Roughly, the running time scales as $|\Theta|/\alpha$ to obtain good accuracy with probability greater than $1 - \alpha$. In certain problems, $|\Theta|$ can be exponential in the dimension (e.g., the Lipschitz stochastic optimization problem considered in Remark 1), which makes it computationally intractable.

Algorithm 5 Idealized estimator for learning with bounded losses

Input: Privacy parameter ε, probability of stopping $\gamma \in (0, 1]$, concentration parameter $\tau > 0$, finite parameter set Θ, dataset $\mathcal{S} = \{S_1, \ldots, S_n\}$
Denote
$$\mathsf{A}_k(S) := \textbf{WinsorizedMean1D}\Big(\{\mathcal{L}(\theta^{(k)}; S_u)\}_{u\in[n]}, \varepsilon/3, \tau\Big)$$
Initialize $\mathcal{T} = \emptyset$.
for $t = 0, \ldots, \infty$ **do**
  Sample $J_t \sim \mathsf{Uniform}(\{1, \ldots, |\Theta|\})$.
  Sample $V_t \sim \mathsf{A}_{J_t}(\mathcal{S})$.
  Update $\mathcal{T} \to \mathcal{T} \cup \{(J_t, V_t)\}$.
  Sample $w_t \sim \mathsf{Bernoulli}(\gamma)$, if $w_t = 1$, break;
end for
$t^* \to \operatorname{argmin}_t V_t$.
return (J_{t^*}, V_{t^*}).

We state the privacy and utility of our algorithm. The result follows from the utility guarantees of the mean estimator (Algorithm 1) and the guarantees of private selection in [45].

Theorem 6. *Let $\alpha \in (0, 1]$ and let us consider Algorithm 5 with $q = 1/K = 1/|\Theta|$ and $\tau = \frac{B}{2}\sqrt{(\log(Kn) + \log(10/\alpha)/m}$. Assuming that $n \geq \frac{8}{\varepsilon}\log\Big(\frac{25\log(5/\alpha)}{\alpha^2} \cdot \frac{KB}{\tau}\Big)$, the following holds:*

(i) The mechanism of Algorithm 5 is ε-user-level DP.
(ii) Let J_{t^*} be the output of Algorithm 5, with probability greater than $1 - \alpha$ it achieves the following utility

$$\mathcal{L}(\theta^{(J_{t^*})}; \mathcal{S}) - \inf_{\theta' \in \Theta} \mathcal{L}(\theta'; \mathcal{S}) \leq 8\frac{B}{n\sqrt{m\varepsilon}}\log\left(25K \cdot \frac{\log(5/\alpha)}{\alpha^2}\right)\sqrt{\log(Kn) + \log(10/\alpha)}. \tag{10}$$

Proof. We first state the privacy guarantee followed by the utility guarantee.

Proof of (i) Since each A_k is $\varepsilon/3$-user-level DP, Theorem 3.2 in [45] guarantees that the output of Algorithm 5 is ε-user-level DP.

Proof of (ii) The proof is adapted from Theorem 5.2 in [45]. First of all, with probability greater than $1 - \alpha_1$, as we prove in Lemma 1, the data are uniformly concentrated for all $\theta^{(k)}$, meaning

$$\max_{k\in K}\max_{u\in[n]}\Big|\mathcal{L}(\theta^{(k)}; S_u) - \mathcal{L}(\theta^{(k)}; P)\Big| \leq \left\{\frac{B}{2}\sqrt{\frac{\log(|\Theta| \cdot n) + \log(2/\alpha_1)}{m}} =: \tau\right\}.$$

We condition on this event (Event 1) for the rest of the proof. Let $\alpha_1 \in (0, 1]$ and $\gamma \in (0, 1]$. Let T_s denotes the time that the algortihm exists the loop, which is number of queries the algorithm makes.

Let us denote k^*, the best hypothesis in Θ i.e.

$$k^* = \operatorname*{argmin}_{k\leq K} \mathcal{L}(\theta^{(k)}; \mathcal{S}).$$

We choose γ such that k^* is queried with probability greater than $1-\alpha_1$, i.e., if $E_{\neg k^*}$ is the event (denote $\neg E_{\neg k^*}$ as Event 2) that the algorithm finishes without querying k^*, we choose γ such that $\mathbb{P}(E_{\neg k^*}) \le \alpha_1$. More precisely,

$$\begin{aligned}
\mathbb{P}(E_{\neg k^*}) &= \sum_{l=1}^{\infty} \mathbb{P}(E_{\neg k^*} | T_s = l)\mathbb{P}(T_s = l) \\
&= \sum_{l=1}^{\infty} \left(1 - \frac{1}{K}\right)^l \cdot (1-\gamma)^{l-1} \cdot \gamma \\
&= \left(1 - \frac{1}{K}\right)\gamma \sum_{l=0}^{\infty} \left[\left(1 - \frac{1}{K}\right)(1-\gamma)\right]^l \\
&= \frac{\left(1 - \frac{1}{K}\right)\gamma}{1 - \left(1 - \frac{1}{K}\right)(1-\gamma)}.
\end{aligned}$$

Choosing $\gamma = \alpha_1/K$ guarantees that $\mathbb{P}(E_{\neg k^*}) \le \alpha_1$. Let $L := \frac{\log(1/\alpha_1)}{\gamma} = \log(1/\alpha_1)\frac{K}{\alpha_1}$, we have

$$\mathbb{P}(T_s > L) = \mathbb{P}(\omega_1 = \ldots = \omega_L = 0) = (1-\gamma)^L \le \exp(-L\gamma) = \alpha_1.$$

Hence with probability at least $1-\alpha_1$, the algorithm ends in less than L throws (Event 3). Conditioned on this event, by Theorem 1 and union bound, with probability greater than $1 - L \cdot \frac{B}{\tau}\exp(-n\varepsilon/8)$, the output of A_{J_t} for all $t \le T_s$ is

$$\mathsf{A}_{J_t}(S) = \mathcal{L}(\theta^{(J_t)}; \mathcal{S}) + \mathsf{Lap}\left(\frac{8\tau}{n\varepsilon}\right) = \frac{1}{m \cdot n} \sum_{j \in [m], u \in [n]} \ell\left(\theta^{(J_t)}; z_j^{(u)}\right) + \mathsf{Lap}\left(\frac{8\tau}{n\varepsilon}\right),$$

which we denote as Event 4. For a Laplace distribution, computing the tail gives that $\mathbb{P}(|\mathsf{Lap}(\lambda)| \ge u) \le \exp(-u/\lambda)$ and with a union bound and change of variables it holds that if $Y_1, Y_2, \ldots, Y_L \overset{\text{iid}}{\sim} \mathsf{Lap}(\frac{8\tau}{n\varepsilon})$, then with probability greater than $1-\alpha_1$

$$\max_{i=1,\ldots L} |Y_i| \le \frac{8\tau}{n\varepsilon} \log\left(\frac{L}{\alpha_1}\right).$$

In other words, except with probability α_1, the noise is bounded by $\frac{8\tau}{n\varepsilon}\log(L/\alpha_1)$ (Event 5). Conditioned on all these events, the parameter $\theta^{(J_{t^*})}$ that the algorithm outputs is sub-optimal by at most $\frac{16\tau}{n\varepsilon}\log(L/\alpha_1)$ as in the worst-case the noise is $+\frac{8\tau}{n\varepsilon}\log(L/\alpha_1)$ for J_{t^*} and $-\frac{8\tau}{n\varepsilon}\log(L/\alpha_1)$ for k^*. Setting $\alpha_1 = \alpha/5$ and as we assume that $n \ge \frac{8}{\varepsilon}\log\left(\frac{25\log(5/\alpha)}{\alpha^2} \cdot \frac{KB}{\tau}\right)$, we conclude the proof by taking a union bound over all 5 events.

□

Corollary 3. *Assume* $n \ge \tilde{\Omega}(1)\,\frac{1}{\varepsilon}\max\{\frac{1}{Km}, \log(Km)\}$. *It holds that*

$$\mathfrak{M}^{\mathsf{user}}_{m,n}(\Theta, \mathcal{F}_B, \epsilon) = \tilde{O}\left(B\left\{\sqrt{\frac{\log K}{m \cdot n}} + \frac{\log^{3/2}(Knm\varepsilon)}{n\sqrt{m}\varepsilon}\right\}\right), \tag{11}$$

where $\tilde{O}, \tilde{\Omega}$ *ignores only numerical constants and log-log factors in this case.*

Proof. We get the result directly from Theorem 6, by setting $\alpha = \log K/(n\sqrt{m}\varepsilon)$, applying standard uniform convergence results for bounded losses with finite parameter set (Hoeffding bound) and ignoring log-log factors. □

Corollary 4 (Parameter sets with finite metric entropy)**.** *Let us further assume that our loss functions are* G*-Lipschitz with respect to some norm* $\|\cdot\|$ *with (finite) covering number* $\mathsf{N}_{\|\cdot\|}(\Theta, \Delta)$*—i.e. there exists a set* $\Gamma_{\|\cdot\|,\Delta} \subset \Theta$ *such that* $|\Gamma_{\|\cdot\|,\Delta}| = \mathsf{N}_{\|\cdot\|}(\Theta, \Delta)$ *and for all* $\theta \in \Theta$*, there exists* $\tau \in \Gamma_{\|\cdot\|,\Delta}$

such that $\|\theta - \tau\| \leq \Delta$. *In this case, for any* $\Delta > 0$ *and applying Algorithm 5 with parameter set* Γ *guarantees that*

$$\mathfrak{M}^{\mathsf{user}}_{m,n}(\Theta, \mathcal{F}_{B,(G,\|\cdot\|)}, \varepsilon) = \tilde{O}(1) \inf_{\Delta>0}\left\{B\left[\sqrt{\frac{\log \mathsf{N}_{\|\cdot\|}(\Theta,\Delta)}{m\cdot n}} + \frac{\log^{3/2}\big(\mathsf{N}_{\|\cdot\|}(\Theta,\Delta)nm\varepsilon\big)}{n\sqrt{m}\varepsilon}\right] + G\Delta\right\}.$$

Remark 1. For $\|\cdot\| = \ell_2, \Theta = \mathbb{B}^d_\infty(0,1)$ and setting $\Delta = \frac{B}{G}\left\{\sqrt{d/(mn)} + d^{3/2}/(n\varepsilon\sqrt{m})\right\}$, we directly get

$$\mathfrak{M}^{\mathsf{user}}_{m,n}(\mathbb{B}^d_\infty(0,1), \mathcal{F}_{B,(G,\ell_2)}, \varepsilon) = \tilde{O}\left\{B\sqrt{\frac{d}{m\cdot n}} + B\frac{d^{3/2}}{n\sqrt{m}\varepsilon}\right\}.$$

The first term, which corresponds to the statistical rate, is optimal (see e.g. Proposition 2 in [44]). Whether the privacy rate is optimal remains open.

A.2 Information-theoretic lower bound

We now prove a lower bound on $\mathfrak{M}^{\mathsf{user}}_{m,n}(\Theta, \mathcal{F}_B, \varepsilon)$ when $|\Theta| = K < \infty$. We follow the standard machinery of reducing estimation to testing [62, 57] but under privacy constraints [7, 1].

Theorem 7 (Lower bound for finite-hypothesis class)**.** *Let* $K, m, n \in \mathbb{N}, K < \infty, \varepsilon \in \mathbb{R}_+$, *and* $0 \leq B < \infty$. *Assume* $\log_2 K \geq 32 \log 2$ *and* $n \geq \log_2 K \max\{\frac{1}{192\sqrt{m}\varepsilon}, \frac{1}{96m}\}$, *there exists a sample space* $\mathcal{Z}$ *and parameter set* Θ *with* $|\Theta| = K$ *and* $|\mathcal{Z}| = \lceil \log_2 K \rceil$ *such that the following holds*

$$\mathfrak{M}^{\mathsf{user}}_{m,n}(\Theta, \mathcal{F}_B, \varepsilon) = \Omega\left(B\sqrt{\frac{\log|\Theta|}{m\cdot n}} + B\frac{\log|\Theta|}{n\sqrt{m\epsilon}}\right). \tag{12}$$

We detail the proof of the theorem below. The proof relies on a (standard) generalization of Fano's method, whcih reduces optimization to multiple hypothesis tests. We refer to the results of [1] to obtain the lower bounds in the case of a constrained—in this case, ϵ-DP—estimators. For the user-level case, we simply consider that samples from an m-fold product of measures—the separation does not change but the KL-divergence increase by at most a m factor and TV-distance increase by at most a $\sqrt{m}$ factor thus yielding the final answer.

Proposition 2 (Acharya et al. [1, Corollary 4])**.** *Let* $\mathcal{P}$ *be a collection of distributions over a common sample space* $\mathcal{Z}$ *and a loss function* $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$. *For* $P, Q \in \mathcal{P}$, *define*

$$\mathsf{sep}_{\mathcal{L}}(P, Q; \Theta) := \sup\left\{\delta \geq 0 \;\middle|\; \text{for all } \theta \in \Theta, \begin{array}{l}\mathcal{L}(\theta, P) \leq \delta \text{ implies } \mathcal{L}(\theta, Q) \geq \delta \\ \mathcal{L}(\theta, Q) \leq \delta \text{ implies } \mathcal{L}(\theta, P) \geq \delta\end{array}\right\}.$$

Let $\mathcal{V}$ *be a finite index set and* $\mathcal{P}_\mathcal{V} := \{P_v\}_{v\in\mathcal{V}}$ *be a collection of distributions contained in* $\mathcal{P}$ *such that for* $\Delta \geq 0$, $\min_{v\neq v'} \mathsf{sep}(P_v, P_{v'}, \Theta) \leq \Delta$. *Then*

$$\mathfrak{M}^{\mathsf{item}}_n(\Theta, \mathcal{F}, \epsilon) \geq \frac{\Delta}{4}\max\left\{1 - \frac{I(X_1^n; V) + \log 2}{\log|\mathcal{V}|}, \min\left\{1, \frac{|\mathcal{V}|}{\exp(c_0 n\varepsilon \mathrm{d}_{\mathsf{TV}}(\mathcal{P}_\mathcal{V}))}\right\}\right\},$$

where $V \sim \mathsf{Uniform}(\mathcal{V}), c_0 = 10, \mathrm{d}_{\mathsf{TV}}(\mathcal{P}_\mathcal{V}) := \max_{v\neq v'}\|P_v - P_{v'}\|_{\mathsf{TV}}$ *and* $I(X;Y)$ *is the (Shannon) mutual information.*

Proof of Theorem 7. We follow the standard steps: we first compute the separation, we bound the testing error for any (constrained) estimator in the item-level DP case (with Proposition 2) and finally, we show how to adapt the proof to obtain the user-level DP lower bound.

Separation For simplicity, assume $K = 2^d$, if not, the problem is harder than for $\underline{K} = 2^{\lfloor \log_2 K \rfloor} \leq K$ which is of the same order. Let us define the sample space $\mathcal{Z}$, the parameter set Θ and the loss function ℓ we consider.

We define

$$\mathcal{Z} = \Theta := \{-1, +1\}^d \;\text{ and }\; \ell(\theta; z) := B\sum_{j\leq d} \mathbf{1}_{\theta_j = z_j}.$$

We consider $\mathcal{V}$ an $d/2$-ℓ_1 packing of $\{\pm 1\}^d$ of size at least $\exp(d/8)$—which the Gilbert-Varshimov bound (see e.g., [56, Ex. 4.2.16]) guarantees the existence of—and consider the following family of distribution $\mathcal{P} = \{P_v : v \in \mathcal{V}\}$ such that if $X \sim P_v$ then

$$X = \begin{cases} v_j e_j & \text{with probability } \frac{1+\Delta}{2d} \\ -v_j e_j & \text{with probability } \frac{1-\Delta}{2d}. \end{cases} \tag{13}$$

For $\theta \in \Theta$, we have that

$$\mathcal{L}(\theta; P_v) = \mathbb{E}_{P_v}\left[B \sum_{j \le d} \mathbf{1}_{\theta_j = Z_j}\right] = B \sum_{j \le d} \frac{1 + \theta_j v_j \Delta}{2d}.$$

Naturally, $\mathcal{L}(\theta; P_v)$ achieves its minimum at $\theta_v^* = -v$ such that $\inf_{\theta' \in \theta} \mathcal{L}(\theta; P_v) = B\frac{1-\Delta}{2}$. We now compute the separation by noting that

$$\mathsf{sep}_{\mathcal{L}}(P_v, P_{v'}, \Theta) \ge \frac{1}{2} \min_{\theta' \in \Theta} \{\mathcal{L}(\theta'; P_v) + \mathcal{L}(\theta'; P_{v'}) - \mathcal{L}(\theta_v^*; P_v) - \mathcal{L}(\theta_{v'}^*; P_{v'})\}. \tag{14}$$

A quick computation shows that $\mathsf{sep}_{\mathcal{L}}(P_v, P_{v'}, \Theta) \ge \frac{B\Delta}{8}$ by noting that $\mathrm{d}_{\mathsf{Ham}}(v, v') \ge d/4$.

Obtaining the item-level lower bound We can now use the results of Proposition 2. We have that $\min_{v \neq v'} \mathsf{sep}_{\mathcal{L}}(P_v, P_{v'}, \Theta) \ge \frac{B\Delta}{8}$. The identity $\mathrm{D}_{\mathrm{KL}}(P_v, P_{v'}) = \Delta \log \frac{1+\Delta}{1-\Delta} \le 3\Delta^2$ implies that $I(Z^n; V) \le 3n\Delta^2$. Similarly, Pinsker's inequality guarantees that

$$\mathrm{d}_{\mathsf{TV}} \le \sqrt{\frac{1}{2} \max_{v \neq v'} \mathrm{D}_{\mathrm{KL}}(P_v, P_{v'})} \le \sqrt{3/2}\Delta.$$

We put everything together and it holds that for $\Delta \in [0, 1]$,

$$\mathfrak{M}_n^{\mathsf{item}}(\Theta, \mathcal{F}, \epsilon) \ge \frac{B\Delta}{32} \max\left\{1 - \frac{3n\Delta^2 + \log 2}{d/8}, \min\left\{1, \frac{\exp(d/8)}{\exp(30n\varepsilon\Delta)}\right\}\right\}. \tag{15}$$

Since $d \ge 32 \log 2$, $\Delta = \sqrt{d/(96n)}$ guarantees that $1 - \frac{3n\Delta^2 + \log 2}{d/8} \ge 1/2$. On the other hand, setting $\Delta = \frac{5}{960}\frac{d}{n\varepsilon}$, guarantees that $\min\left\{1, \frac{\exp(d/8)}{\exp(30n\varepsilon\Delta)}\right\} \ge 1/2$. The assumption on n guarantees that these two values are in $[0, 1]$ and thus setting $\Delta^* = \max\left\{\sqrt{d/(96n)}, \frac{1}{192}\frac{d}{n\varepsilon}\right\}$ which implies that

$$\mathfrak{M}_n^{\mathsf{item}}(\Theta, \mathcal{F}, \epsilon) \ge \frac{B}{32}\left\{\sqrt{\frac{d}{96n}} + \frac{1}{192}\frac{d}{n\varepsilon}\right\}.$$

Concluding for user-level DP Let $m \in \mathbb{N}, m \ge 1$. For the user-level DP lower bound, the proof remains the same except that the collection $\mathcal{P}_{\mathcal{V}}$ becomes $\{P_v^m\}_{v \in \mathcal{V}}$ i.e. the m-fold product distribution of P_v. The separation remains exactly the same but we now have

$$\mathrm{D}_{\mathrm{KL}}(P_v^m, P_{v'}^m) \le 3m\Delta^2 \text{ and } \mathrm{d}_{\mathsf{TV}}(\mathcal{P}_{\mathcal{V}}) \le \sqrt{\frac{3m}{2}}\Delta.$$

Under the assumption $\Delta^* = \max\left\{\sqrt{d/(96mn)}, \frac{1}{192}\frac{d}{n\sqrt{m}\varepsilon}\right\}$ is less than 1 and thus concludes the proof. □

Note, the upper bound of Theorem 6 and the lower bound above match only up to $\sqrt{\log K}$. Given that K can be exponential in the dimension—e.g. in the case of Θ being a cover of an ℓ_p ball—the bound is only tight for "small" hypothesis class. However, it seems this extra-factor cannot be removed using the techniques we present in this paper, as we need to both obtain uniform concentration and bound the maximum of i.i.d. noise over K samples—both of which are tight. We leave the problem of finding an optimal estimator for this problem to future work.

B Limit of Learning with a Fixed Number of Users

In this section, we consider the following binary testing problem between P_1 and P_2 supported on $\{+B, -B\}$ where

$$P_0(+B) = 1, \quad P_0(-B) = 0,$$
$$P_1(+B) = 0, \quad P_1(-B) = 1.$$

We prove the following result.

Theorem 8. *For all user-level (ε, δ)-DP algorithm $\mathsf{A} : \{+B, -B\}^{m\times n} \to [0, 1]$, let $\mathcal{S}$ be $n \times m$ i.i.d samples from $P_\vartheta, \vartheta \in \{0, 1\}$, we have when $\delta < 1/2ne^{n\varepsilon}$,*

$$\max_{\theta\in\{0,1\}} \mathbb{E}\left[(\mathsf{A}(\mathcal{S}) - \vartheta)^2\right] = \Omega(e^{-n\varepsilon}).$$

Before proving the theorem, we describe the implications of the theorem to applications considered in this work. Let $\mathcal{A}^{\mathsf{user}}_{\varepsilon,\delta}$ denote the set of all user-level (ε, δ)-DP algorithms.

Reduction from mean estimation P_0 and P_1 are both bounded distributions. Moreover, we have $\mu_\vartheta = B(2\vartheta - 1)$. For any user-level (ε, δ)-DP mean estimator $\widehat{\mu} : \{+B, -B\}^{m\times n} \to [-B, +B]$, set $\mathsf{A}_{\widehat{\mu}}(\mathcal{S}) = (\widehat{\mu} + B)/(2B) \in [0, 1]$, we have $\forall\vartheta \in \{0, 1\}$,

$$\mathbb{E}\left[(\widehat{\mu}(\mathcal{S}) - \mu_\vartheta)^2\right] = 4B^2\mathbb{E}\left[(\mathsf{A}_{\widehat{\mu}}(\mathcal{S}) - \vartheta)^2\right].$$

We have

$$\begin{aligned}\inf_{\widehat{\mu}\in\mathcal{A}^{\mathsf{user}}_{\varepsilon,\delta}} \max_{\vartheta\in\{0,1\}} \mathbb{E}\left[(\widehat{\mu}(\mathcal{S}) - \mu_\vartheta)^2\right] &= 4B^2 \inf_{\widehat{\mu}\in\mathcal{A}^{\mathsf{user}}_{\varepsilon,\delta}} \max_{\vartheta\in\{0,1\}} \mathbb{E}\left[(\mathsf{A}_{\widehat{\mu}}(\mathcal{S}) - \vartheta)^2\right] \\ &\geq 4B^2 \inf_{\mathsf{A}\in\mathcal{A}^{\mathsf{user}}_{\varepsilon,\delta}} \max_{\vartheta\in\{0,1\}} \mathbb{E}\left[(\mathsf{A}(\mathcal{S}) - \vartheta)^2\right] = \Omega(B^2e^{-n\varepsilon}).\end{aligned}$$

Reduction from SCO Let $\Theta = [-1, 1]$ and $\ell(\theta, Z) = \theta \cdot Z$. Setting $B = G$. The loss is linear (and thus convex), G-Lipschitz and satisfies Assumptions A3 and A4. For P_ϑ,

$$\mathcal{L}(\theta, P_\vartheta) = \theta G(2\vartheta - 1).$$

Hence the minimizer is $\theta^*_\vartheta = 1 - 2\vartheta$ and

$$\mathcal{L}(\theta, P_\vartheta) - \mathcal{L}(\theta^*_\vartheta, P_\vartheta) = (2\vartheta - 1)G(\theta - 1 + 2\vartheta) = G(1 - \theta(2\vartheta - 1)) \geq \frac{G}{2}(\theta - 2\vartheta + 1)^2 = \frac{G}{2}(\theta - \mu_\vartheta)^2.$$

With similar arguments as in the mean estimation reduction, we get

$$\inf_{\mathsf{A}\in\mathcal{A}^{\mathsf{user}}_{\varepsilon,\delta}} \max_{\vartheta\in\{0,1\}} \mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}); P_\vartheta) - \min_{\theta\in[-1,1]} \mathcal{L}(\theta; P_\vartheta)\right] = \Omega(Ge^{-n\varepsilon}).$$

Reduction from Bounded Losses In the reduction from SCO, the loss is uniformly bounded and thus this is a sub-problem of the boundeed loss class and the same bound holds.

Finally, let us prove the theorem.

Proof of Theorem 8. Note that there is only two possible sets that each user can observe. Let S_+ be the multiset consisting of m copies of $+B$ and Let S_- be the multiset consisting of m copies of $-B$. Let $\beta_1 = \mathbb{P}(\mathsf{A}((S_+)^n) < 1/2)$ and $\beta_0 = \mathbb{P}(\mathsf{A}((S_-)^n) \geq 1/2)$. We first show that these two probabilities cannot be simultaneously small.

Since $(S_+)^n$ can be changed into $(S_-)^n$ by changing n users' samples, by group property of differential privacy,

$$1 - \beta_1 = \mathbb{P}(\mathsf{A}((S_+)^n) \geq 1/2) \leq e^{n\varepsilon}\mathbb{P}(\mathsf{A}((S_-)^n) \geq 1/2) + ne^{n\varepsilon}\delta = e^{n\varepsilon}\beta_0 + ne^{n\varepsilon}\delta.$$

Similarly, we get

$$1 - \beta_0 \leq e^{n\varepsilon}\beta_1 + ne^{n\varepsilon}\delta.$$

Combining the two, we get:

$$\beta_0 + \beta_1 \geq \frac{2(1 - n\delta e^{n\varepsilon})}{1 + e^{n\varepsilon}} \geq \frac{1}{1 + e^{n\varepsilon}}.$$

Note that when $\vartheta = 1$, we have $\mathbb{P}(\mathcal{S} = (S_+)^n) = 1$. Hence

$$\mathbb{E}_{P_1}\left[(\mathsf{A}(\mathcal{S}) - 1)^2\right] \geq \frac{1}{4}\mathbb{P}(\mathsf{A}((S_+)^n) < 1/2).$$

Similarly,

$$\mathbb{E}_{P_0}\left[(\mathsf{A}(\mathcal{S}) - 0)^2\right] \geq \frac{1}{4}\mathbb{P}(\mathsf{A}((S_-)^n) \geq 1/2).$$

We conclude the proof by noting that

$$\max_{\vartheta \in \{0,1\}} \mathbb{E}\left[(\mathsf{A}(\mathcal{S}) - \vartheta)^2\right] \geq \frac{1}{2}\Big(\mathbb{E}_{P_0}\left[(\mathsf{A}(\mathcal{S}) - 0)^2\right] + \mathbb{E}_{P_1}\left[(\mathsf{A}(\mathcal{S}) - 1)^2\right]\Big).$$

□

C Extension to Limited Heterogeneity Setting

In this section, we show that our results and techniques developed under the homogeneous setting (Assumption A2) can be extended to the setting with limited heterogeneity (Assumption A1).

In particular, we show that applying the algorithms under the i.i.d setting in a black-box fashion will work with an additional bounded error under the limited heterogeneity setting, stated in the theorem below.

Theorem 9. *Let $\mathsf{A} : \mathcal{Z}^{m\times n} \to \Theta$ be a learning algorithm and $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}_+$ be a loss function with $\max_{z\in\mathcal{Z}} \max_{\theta\in\Theta} \mathcal{L}(\theta; z) \leq B$. Given samples $\mathcal{S} = (S_1, \ldots, S_n) \sim \otimes_{u\in[n]}(P_u)^m$, if under Assumption A2, we have*

$$\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}); P_0)\right] - \min_{\theta'\in\Theta} \mathcal{L}(\theta'; P_0) \leq L(m, n),$$

then under Assumption A1, we have

$$\max_u \left\{\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}); P_u)\right] - \min_{\theta'\in\Theta} \mathcal{L}(\theta'; P_u)\right\} \leq L(m, n) + B(mn + 2)\Delta.$$

Before proving the theorem, we can see that for any learning task, when $\Delta < L(m, n)/(B(mn+2))$, we can get the same performance as in the homogeneous case up to constant factors. This is only inverse polynomial in the problem parameters for all considered tasks.

Proof. We first show that $\mathcal{S}$ have a similar distribution under Assumption A2 and A1 when Δ is small. By sub-additivity of total variantion distance. Under Assumption A1, we have

$$\|\otimes_{u\in[n]}(P_u)^m - (P_0)^{n\times m}\|_{\mathsf{TV}} \leq mn\Delta. \tag{16}$$

By definition of TV distance, there exists a coupling $(\mathcal{S}, \mathcal{S}')$ where $\mathcal{S} \sim \otimes_{u\in[n]}(P_u)^m$, $\mathcal{S}' \sim (P_0)^{n\times m}$ and

$$\mathbb{P}(\mathcal{S} \neq \mathcal{S}') \leq mn\Delta.$$

Since $\max_{z\in\mathcal{Z}} \max_{\theta\in\Theta} \mathcal{L}(\theta; z) \leq B$, we have

$$\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}); P_0)\right] - \mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}'); P_0)\right] \leq B \times \mathbb{P}(\mathcal{S} \neq \mathcal{S}') \leq Bmn\Delta. \tag{17}$$

Under Assumption A1, for all $u \in [n]$, $\|P_u - P_0\|_{\mathsf{TV}} \leq \Delta$. For all $\theta \in \Theta$,

$$\mathcal{L}(\theta; P_0) - \mathcal{L}(\theta; P_u) \leq B\Delta,$$

Hence we have

$$\min_{\theta'\in\Theta} \mathcal{L}(\theta'; P_0) - \min_{\theta'\in\Theta} \mathcal{L}(\theta'; P_u) \leq \max_{\theta\in\Theta} |\mathcal{L}(\theta; P_0) - \mathcal{L}(\theta; P_u) \leq B\Delta, \tag{18}$$

and

$$\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}'); P_u)\right] - \mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}'); P_0)\right] \leq B\Delta. \tag{19}$$

Therefore, for all $u \in [n]$,

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}); P_u)\right] - \min_{\theta' \in \Theta} \mathcal{L}(\theta'; P_u) \\
=\ &\left(\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}); P_u)\right] - \mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}'); P_u)\right]\right) + \left(\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}'); P_u)\right] - \mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}'); P_0)\right]\right) \\
&+ \left(\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}'); P_0)\right] - \min_{\theta' \in \Theta} \mathcal{L}(\theta'; P_0)\right) + \left(\min_{\theta' \in \Theta} \mathcal{L}(\theta'; P_0) - \min_{\theta' \in \Theta} \mathcal{L}(\theta'; P_u)\right) \\
\leq\ &L(m, n) + B(mn + 2)\Delta,
\end{aligned}$$

where we bound each term using (16), (17), (18) and (19) respectively. □

D Proofs for Section 3

D.1 Private range estimation

Algorithm 6 PrivateRange$(X^n, \varepsilon, \tau, B)$: Private Range Estimation [27]

Require: $X^n := (X_1, X_2, ..., X_n) \in [-B, B]^n$, τ : concentration radius, privacy parameter $\varepsilon > 0$.

1: Divide the interval $[-B, B]$ into $l = B/\tau$ disjoint bins[9], each with width 2τ. Let T be the set of middle points of intervals.

2: $\forall i \in [n]$, let $X_i' = \min_{x \in T} |X_i - x|$ be the point in T closest to X_i.

3: $\forall x \in T$, define cost function

$$c(x) = \max\{|\{i \in [n] \mid X_i' < x\}|, |\{i \in [n] \mid X_i' > x\}|\}.$$

4: Sample $x \in T$ based on the following distribution:

$$\mathbb{P}(\hat{\mu} = x) = \frac{e^{-\varepsilon c(x)/2}}{\sum_{x' \in T} e^{-\varepsilon c(x')/2}}.$$

5: Return $R = [\hat{\mu} - 2\tau, \hat{\mu} + 2\tau]$.

D.2 Proof of Theorem 1

Theorem 1. *Let X^n be a dataset supported on $[-B, B]$. The output of Algorithm 1, denoted by $\mathsf{A}(X^n)$, is ε-DP. Furthermore, if X^n is (τ, γ)-concentrated, it holds that*

$$\mathsf{A}(X^n) \sim_\beta \frac{1}{n} \sum_{i=1}^n X_i + \mathit{Lap}\left(\frac{8\tau}{n\varepsilon}\right),$$

where $\beta = \min\left\{1, \gamma + \frac{B}{\tau} \exp\left(-\frac{n\varepsilon}{8}\right)\right\}$. Moreover, Algorithm 1 runs in time $\tilde{O}(n + \log(B/\tau))$.

Proof. The privacy guarantee of the algorithm follows from the composition theorem of DP and the privacy guarantees of the exponential and Laplace mechanisms. For utility, it is enough to show that with probability at least $1 - (\gamma + \frac{B}{\tau} \exp\left(-\frac{n\varepsilon}{8}\right)), \forall i \in [n], X_i$ is not truncated, i.e. $X_i \in [\hat{\mu} - 2\tau, \hat{\mu} + 2\tau]$.

Recall that X_i' is the middle of the interval in which X_i falls. By the definition of (τ, γ)-concentration, with probability at least $1 - \gamma, \forall i \in [n]$,

$$|X_i - x_0| \leq \tau.$$

This implies that $\forall i \in [n]$,

$$|X_i' - x_0| \leq 2\tau,$$

hence so is the $\left(\frac{1}{4}, \frac{3}{4}\right)$-quantile of $\{X_i'\}_{i=1}^n$. According to [27] (Theorem 3.1), Algorithm 6 outputs $\left(\frac{1}{4}, \frac{3}{4}\right)$-quantile of $\{X_i'\}_{i=1}^n$ with probability at least $1 - \frac{B}{\tau} e^{-\frac{n\varepsilon}{8}}$. The proof follows by a union bound of both events. □

[9] The last interval is of length $2B - (t-1)\tau$ if τ doesn't divide B.

D.3 Proof of Theorem 2

Theorem 2. *Let* $\mathsf{A}(X^n) = \textbf{\textit{WinsorizedMeanHighD}}(X^n, \varepsilon, \delta, \tau, B, \gamma)$ *be the output of Algorithm 2.* $\mathsf{A}(X^n)$ *is* (ε, δ)*-DP. Furthermore, if* X^n *is* (τ, γ)*-concentrated in* ℓ_2*-norm, there exists an estimator* $\mathsf{A}'(X^n)$ *such that* $\mathsf{A}(X^n) \sim_\beta \mathsf{A}'(X^n)$ *and*

$$\mathbb{E}[\mathsf{A}'(X^n)|X^n] = \frac{1}{n}\sum_{i=1}^n X_i \text{ and } \mathrm{Var}(\mathsf{A}'(X^n)|X^n) \le c_0 \frac{d\tau^2 \log(dn/\alpha)\log(1/\delta)}{n^2\varepsilon^2}, \tag{6}$$

where $c_0 = 102,400$ *and* $\beta = \min\left\{1, 2\gamma + \frac{d^2 B\sqrt{\log(dn/\gamma)}}{\tau}\exp\left(-\frac{n\varepsilon}{24\sqrt{d\log(1/\delta)}}\right)\right\}$.

We start by proving the following Lemma, which states that if the data is concentrated in ℓ_2-norm with radius τ, then after a random rotation, the points are concentrated in ℓ_∞-norm with radius $\tau/\sqrt{d}$ up to logarithmic factors.

Lemma 2. *Let* $U = \frac{1}{\sqrt{d}}\mathbf{H}D$*, where* $\mathbf{H}$ *is the Walsh Hadamard matrix and* D *is a diagonal matrix with i.i.d. uniformly random* $\{+1, -1\}$ *entries. Let* $x_1, x_2, \ldots, x_n$ *and* x_0 *be vectors in* $\mathbb{R}^d$*. With probability at least* $1-\alpha$*, then the following holds.*

$$\max_i \|Ux_i - Ux_0\|_\infty \le \frac{10\max_i \|x_i - x_0\|_2\sqrt{\log\frac{nd}{\alpha}}}{\sqrt{d}}.$$

Proof. Let $z_i = x_i - x_0$. It suffices to show that

$$\max_i \|Uz_i\|_\infty \le \frac{10\max_i \|z_i\|_2\sqrt{\log\frac{nd}{\alpha}}}{\sqrt{d}}.$$

holds with probability at least $1-\alpha$. Let $y_i = Uz_i$ and let $y_{i,j}$ denote the j^{th} coordinate of y_j. Let D_j denote that j^{th} diagonal of D. Then

$$y_{i,j} = \frac{1}{\sqrt{d}}\sum_k \mathbf{H}_{j,k}D_k z_{i,k}$$

Hence,

$$\mathbb{E}[y_{i,j}] = \frac{1}{\sqrt{d}}\sum_k \mathbf{H}_{j,k}\mathbb{E}[D_k]z_{i,k} = 0.$$

However, observe that changing one coordinate of D, say D_k changes the value of $y_{i,j}$ by at most

$$y_{i,j} - y'_{i,j} \le \frac{2}{\sqrt{d}}z_{i,k} \le \frac{2\|z_i\|_2}{\sqrt{d}}.$$

Hence, by the McDiarmid's inequality with probability at least $1-\alpha'$

$$|y_{i,j}| \le \frac{10\|z_i\|_2\sqrt{\log\frac{1}{\alpha'}}}{\sqrt{d}}.$$

Choosing $\alpha' = \alpha/nd$ and applying union bound over all coordinates of all vectors yields the desired bound. □

Thus, after applying the random rotation, we have with probability $1-2\gamma$ that for all $j \in [d]$, $\{Y_i(j)\}_{u\in[n]}$ is $(\tau', 0)$-concentrated with $\tau' = 10\tau\sqrt{\log(nd/\alpha)/d}$. Hence conditioned on this event, by Theorem 1 and a union bound over d coordinates, after applying **WinsorizedMean1D** to each dimension, we have that for all $j \in [d]$, $\bar{Y}(j) \sim_\beta \bar{Y}'(j)$ where $\beta = 1 - \frac{\sqrt{d}B}{\tau'}\exp\left(-\frac{n\varepsilon'}{8}\right)$ and

$$\bar{Y}'(j) = \frac{1}{n}\sum_{i=1}^n Y_i(j) + \mathrm{Lap}\left(\frac{8\tau'}{n\varepsilon'}\right),$$

Plugging in values of τ' and ε', it can be seen that $\bar{Y}'$ satisfies the conditions in the theorem. By subadditivity of TV distance, we have

$$\bar{Y} \sim_{d\beta} \bar{Y}'.$$

The theorem follows by noting the random rotation is an orthogonal transform and preserves variance.

D.4 Proof of Corollary 1

For all $i \in [n]$, let $X_i = \frac{1}{m}\sum_{j=1}^{m} Z_j^{(i)}$, i.e., the average of user i's samples. Since $\|Z_j^{(i)}\| \le B$, we know that X^n is $(B\sqrt{\log(2n/\gamma)/(2m)}, \gamma)$-concentrated (e.g., see [36]). Hence by Theorem 2, if we apply Algorithm 2 to X^n, we have $\mathsf{A}(X^n) \sim_\beta \mathsf{A}'(X^n)$ with $\beta = \min\{1, \gamma + \alpha + \frac{d^2 B\sqrt{\log(dn/\alpha)}}{\tau}\exp(-\frac{n\varepsilon}{24\sqrt{d\log(1/\delta)}})\}$ with $\tau = B\sqrt{\log(2n/\gamma)/(2m)}$ and

$$\mathbb{E}[\mathsf{A}'(X^n)|X^n] = \frac{1}{n}\sum_{i=1}^{n} X_i \;\text{ and }\; \mathrm{Var}(\mathsf{A}'(X^n)|X^n) \le c_0\frac{d\tau^2\log(dn/\alpha)\log(1/\delta)}{n^2\varepsilon^2}.$$

Hence

$$\mathbb{E}[\mathsf{A}'(X^n)] = \mathbb{E}[\mathbb{E}[\mathsf{A}'(X^n)|X^n]] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu.$$

$$\begin{aligned}\mathrm{Var}(\mathsf{A}'(X^n)) &= \mathbb{E}[\mathrm{Var}(\mathsf{A}'(X^n)|X^n)] + \mathrm{Var}(\mathbb{E}[\mathsf{A}'(X^n)|X^n]) \\ &\le \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + c_0\frac{d\tau^2\log(dn/\alpha)\log(1/\delta)}{n^2\varepsilon^2} \\ &= \frac{\mathrm{Var}(P_0)}{mn} + c_0\frac{dB^2\log(2n/\gamma)\log(dn/\alpha)\log(1/\delta)}{mn^2\varepsilon^2}.\end{aligned}$$

Combining the two, we have

$$\mathbb{E}\left[\|\mathsf{A}'(X^n) - \mu\|_2^2\right] \le \frac{\mathrm{Var}(P_0)}{mn} + c_0\frac{dB^2\log(2n/\gamma)\log(dn/\alpha)\log(1/\delta)}{mn^2\varepsilon^2}.$$

Since $\mathsf{A}(X^n) \sim_\beta \mathsf{A}'(X^n)$, we have

$$\mathbb{E}\left[\|\mathsf{A}(X^n) - \mu\|_2^2\right] \le \frac{\mathrm{Var}(P_0)}{mn} + c_0\frac{dB^2\log(2n/\gamma)\log(dn/\alpha)\log(1/\delta)}{mn^2\varepsilon^2} + \beta B^2.$$

Taking $\alpha = \gamma = \frac{c_0 d}{3mn^2\varepsilon^2}$, we have when $n \ge c_1\frac{\sqrt{d\log(1/\delta)}}{\varepsilon}\log(dm^{3/2}\varepsilon^2)$ for a constant c_1, we have

$$\mathbb{E}\left[\|\mathsf{A}(X^n) - \mu\|_2^2\right] \le \frac{\mathrm{Var}(P_0)}{mn} + c_0\frac{2dB^2\log(mn^2\varepsilon^2/d)\log(mn^3\varepsilon^2))\log(1/\delta)}{mn^2\varepsilon^2}.$$

Tightness of Corollary 1. The first term is the classic statistical rate even with unconstrained access to the samples. We prove the tightness of the second term using the following family of truncated Gaussian distributions. The proof follows a similar line of argument of the proof for Theorem 5 in Section F.2. For a mean $\mu \in \mathbb{R}^d$, a covariance $\Sigma \in \mathbb{R}^{d\times d}$ and $B > 0$, we consider the family of ℓ_∞-truncated Gaussians, meaning

$$Z \sim \mathsf{N}^{\mathrm{tr}}(\mu, \Sigma, B) \;\text{ if }\; Z_0 \sim \mathsf{N}(\mu, \Sigma) \;\text{ and set for all } j \in [d]\;\; Z(j) = \frac{Z_0(j)}{\max\{1, |Z_0(j)|/B\}}. \tag{20}$$

In other words, the standard high-dimensional Gaussian distribution where the mass outside of $\mathbb{B}_\infty^d(0, B)$ has been projected back onto the hyperrectangle coordinate-wise.

In this proof, we will take $\Sigma = \sigma^2 I_d$. We first state the following Lemma, proved in Section F.2, which shows that when B is large enough compared to $\|\mu\|_2$ and σ, then the expectation of $\mathsf{N}^{\mathrm{tr}}(\mu, \sigma^2 I_d, B/\sqrt{d})$ and μ are exponentially close in ℓ_2-norm.

Lemma 3. *Suppose* $\|\mu\|_2 + 10\sqrt{d}\sigma < G$,

$$\|\mathbb{E}_{Z\sim\mathsf{N}^{\mathrm{tr}}(\mu,\sigma^2 I_d, G/\sqrt{d})}[Z] - \mu\|_2 = O\big(\sigma e^{-10d}\big).$$

Reducing to standard Gaussian mean estimation We will take $\sigma = B/20\sqrt{d}$ and $\|\mu\|_2 \le B/2$, Hence assuming m, n is polynomial in d, $O\big(\sigma e^{-10d}\big)$ is small compared to the bound in Corollary 1.

Note that we can always simulate a sample from $\mathsf{N}^{\mathrm{tr}}(\mu, \sigma^2 I_d, B/\sqrt{d})$ using a sample from $\mathsf{N}(\mu, \sigma^2 I_d)$ by performing truncation. Taking $\sigma = B/20\sqrt{d}$, it would be enough to prove the following:

$$\inf_{\widehat{\mu} \in \mathcal{A}^{\mathsf{item}}_{\varepsilon,\delta}} \sup_{\mu : \|\mu\|_2 \le B/2} \mathbb{E}_{\mathcal{S} \overset{\mathrm{iid}}{\sim} \mathsf{N}(\mu, \sigma^2 I_d)} \left[\|\widehat{\mu}(\mathcal{S}) - \mu\|_2^2 \right] = \tilde{\Omega}\left(\frac{d^2\sigma^2}{mn^2\varepsilon^2} \right),$$

where $\mathcal{A}^{\mathsf{user}}_{\varepsilon,\delta}$ denotes set of all user-level (ε, δ)-DP algorithms. The next proposition, based on the fact that sample mean is a sufficient statistic for i.i.d Gaussian samples, shows that we can reduce the problem to Gaussian mean estimation under item-level DP, with a smaller variance. The proposition is proved in Section F.2.

Proposition 3 (From multiple samples to one good sample)**.** *Suppose each user $u \in [n]$ observe $(Z_1^{(u)}, \dots, Z_m^{(u)}) \overset{\mathrm{iid}}{\sim} \mathsf{N}(\mu, \sigma^2 I_d)$. For any (ε, δ) user-level DP algorithm $\mathsf{A}^{\mathsf{user}}$, there exists an (ε, δ)-item-level DP algorithm $\mathsf{A}^{\mathsf{item}}$ that takes as input $(\bar{Z}^{(1)}, \dots, \bar{Z}^{(n)})$ with $\bar{Z}^{(u)} := \frac{1}{m}\sum_{j \le m} Z_j^{(u)}$ and has the same performance as $\mathsf{A}^{\mathsf{user}}$.*

Since $\bar{Z}^{(u)}$ is a sample from $\mathsf{N}(\mu, \frac{\sigma^2}{m} I_d)$, it remains to prove

$$\inf_{\widehat{\mu} \in \mathcal{A}^{\mathsf{item}}_{\varepsilon,\delta}} \sup_{\mu : \|\mu\|_2 \le B/2} \mathbb{E}_{Z^n \overset{\mathrm{iid}}{\sim} \mathsf{N}(\mu, \frac{\sigma^2}{m} I_d)} \left[\|\widehat{\mu}(Z^n) - \mu\|_2^2 \right] = \tilde{\Omega}\left(\frac{d^2\sigma^2}{mn^2\varepsilon^2} \right),$$

where $\mathcal{A}^{\mathsf{item}}$ denotes set of all item-level (ε, δ)-DP algorithms. This directly follows from Kamath et al. [38, Lemma 6.7], concluding the proof.

D.5 Mean Estimation of Sub-Gaussian Distribution

In this section, we prove error guarantees for mean estimation of sub-Gaussian distributions. We note that known results in mean estimation of Gaussian distributions and moment bounded distributions [38, 39] imply this bound. We include it here for the sake of completeness to demonstrate the strength of our techniques.

Corollary 5. *Suppose P is a σ-sub-Gaussian distribution supported on $[-B, B]^d$ with mean μ. Assume $n \ge (c_1\sqrt{d\log(1/\delta)}/\varepsilon)\log(B(dn+n^2\varepsilon^2)/\sigma)$ for a numerical constant $c_1 < \infty$, if $X^n \overset{\mathrm{iid}}{\sim} P$, the output $\mathsf{A}(X^n)$ of Algorithm 2 satisfies*[10]

$$\mathbb{E}\left[\|\mathsf{A}(X^n) - \mu\|_2^2 \right] = \tilde{O}\left(\frac{d\sigma^2}{n} + \frac{d^2\sigma^2}{n^2\varepsilon^2} \right).$$

Furthermore, the bound is tight up to logarithmic factors.

The proof is almost parallel to the proof of Corollary 1 by noting that X^n is $(\sigma\sqrt{d\log(2n/\gamma)}, \gamma)$-concentrated and

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i \right) = \frac{d\sigma^2}{n}.$$

The tightness of the result follows from Theorem 3.1 and Lemma 3.1 in [18], which proves lower bounds for mean estimation of k-dimensional random variables supported on $[-\sigma, \sigma]^k$ under (ε, δ)-DP constraints.

D.6 Uniform concentration: answering many queries privately

The statistical query framework subsumes many learning algorithms. For example, we easily express stochastic gradient methods for solving ERM in the language of SQ algorithms (see beginning of Section 4). In the next theorem, we show that with a uniform concentration assumption we can answer a sequence of adaptively chosen queries with variance—or, equivalently, privacy cost—proportional to the concentration radius of the queries instead of the full range.

Theorem 10. *If $(Z^n, \mathcal{Q}_B^d)$ is (τ, γ)-uniformly concentrated, then for any sequence of (possibly adaptively chosen) queries $\phi_1, \phi_2, ..., \phi_K \in \mathcal{Q}_B^d$, there exists an (ε, δ)-DP algorithm A, such that A outputs $v_1, v_2, ..., v_K$ satisfying $(v_1, v_2, ..., v_K) \sim_\beta (v_1', v_2', ..., v_K')$, where*

$$\mathbb{E}\left[v_k' | Z^n \right] = \frac{1}{n}\sum_{i=1}^n \phi_k(Z_i) \quad \textit{and} \quad \mathrm{Var}(v_k' | Z^n) \le \frac{8c_0 dK\tau^2 \log(Kdn/\gamma)\log^2(2K/\delta)}{n^2\varepsilon^2} = \tilde{O}\left(\frac{dK\tau^2}{n^2\varepsilon^2} \right),$$

[10] For precise log factors, see Appendix D.5.

where $c_0 = 102400$ *and* $\beta = \min\left\{1, 2\gamma + \frac{d^2 KB\sqrt{\log(dKn/\gamma)}}{\tau}\exp\left(-\frac{n\varepsilon}{48\sqrt{2dK\log(2/\delta)\log(2K/\delta)}}\right)\right\}$.

The algorithm for Theorem 10 is simply applying Algorithm 2 to $\{\phi_k(Z_i)\}_{i\in[n]}$ with $\varepsilon_0 = \frac{\varepsilon}{2\sqrt{2K\log(2/\delta)}}$ and $\delta_0 = \frac{\delta}{2K}$ for each query. Algorithm 3 is an illustration of an application of this result.

Proof. For each query $\phi_k, k \in [K]$, the algorithm computes $\phi_k(Z_i), i \in [n]$ and returns

$$v_k = \textbf{WinsorizedMeanHighD}\big(\{\phi_k(Z_i)\}_{i\in[n]}, \varepsilon_0, \delta_0, \tau, B, \gamma/K\big)$$

where

$$\varepsilon_0 = \frac{\varepsilon}{2\sqrt{2K\log(2/\delta)}}, \quad \delta_0 = \frac{\delta}{2K}.$$

Privacy guarantee. The proof is immediate and hinges on the strong-composition theorem. Under the standard strong composition results of [23, Theorem III.3], for any $\delta' \in (0, 1]$, the output of Algorithm 3 is $(\bar{\varepsilon}, \bar{\delta})$-user-level DP with

$$\bar{\varepsilon} = K\varepsilon_0(\exp(\varepsilon_0) - 1) + \sqrt{2K\ln(1/\delta')}\varepsilon_0, \qquad \delta = K\delta_0 + \delta'.$$

Plugging in values of ε_0, δ_0 concludes the proof.

Utility guarantee. The proof follows is very similar to the proof of Theorem 2 with $\alpha = \gamma/K$. We conclude by using the subadditivity of the TV distances (or equivalently, a union bound) over all K queries. □

E Proofs from Section 4

E.1 Uniform Concentration

Proposition 1 (Concentration of random gradients)**.** *Let* $S_u \overset{\text{iid}}{\sim} P_u$, $|S_u| = m$ *for* $u \in [n]$ *and* $\alpha \geq 0$. *Under Assumptions A3 and A4, with probability greater than* $1 - \alpha$ *it holds that*

$$\max_{u\in[n]}\sup_{\theta\in\Theta}\|\nabla\mathcal{L}(\theta; S_u) - \nabla\mathcal{L}(\theta; P_u)\|_2 = O\left(\sigma\sqrt{\frac{d\log\left(\frac{RHm}{d\sigma}\right)}{m} + \frac{\log\left(\frac{n}{\alpha}\right)}{m}}\right).$$

Proof. The proof relies on a standard covering number argument. We know that $\sup_{\theta_1,\theta_2\in\Theta}\|\theta_1 - \theta_2\| \leq R$. This implies that $\Theta \subset \mathbb{B}_2^d(\theta_0, R)$, where $\mathbb{B}_2^d(v, r)$ is the d-dimensional ℓ_2-ball centered at $v \in \mathbb{R}^d$ of radius r. Without loss of generality, we assume $\theta_0 = 0$, i.e. the constraint set Θ is centered at 0.

Let us consider $\Gamma_{\|\cdot\|_2}(\Theta, \Delta) =: \Gamma$, a Δ-net of Θ for the ℓ_2 norm, i.e. such that $|\Gamma| < \infty$ and that for all $\theta, \vartheta \in \Theta$, $\|\theta - \vartheta\|_2 \leq \Delta$. Standard results (e.g. Vershynin [56, Corollary 4.2.13]) guarantee that there exists such a set and that its cardinality is smaller than $(1 + 2R/\Delta)^d$.

Since ℓ is uniformly H-smooth, for any sample S we immediately have that

$$\sup_{\theta\in\Theta}\|\nabla\mathcal{L}(\theta; S) - \nabla\mathcal{L}(\theta; P)\|_2 \leq \max_{\vartheta\in\Gamma}\|\nabla\mathcal{L}(\vartheta; S) - \nabla\mathcal{L}(\vartheta; P)\|_2 + 2H\Delta.$$

Consequently, letting $t > 0$, we have that

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\|\nabla\mathcal{L}(\theta; S) - \nabla\mathcal{L}(\theta; P)\|_2 \geq t\right) \leq \mathbb{P}\left(\max_{\vartheta\in\Gamma}\|\nabla\mathcal{L}(\vartheta; S) - \nabla\mathcal{L}(\vartheta; P)\| \geq t/2\right) + \mathbb{P}(H\Delta \geq t/4).$$

For the second term, we simply need to ensures that when choosing t and Δ, it holds that $H\Delta < t/4$. Let us now bound the first term. Once again, let us consider Ξ a $1/2$-net of $\mathbb{B}_2^d(0, 1)$. For any $v \in \mathbb{R}^d$, it holds that

$$\|v\|_2 = \sup_{\|u\|_2\leq 1}\langle u, v\rangle \leq \max_{\tilde{u}\in\Xi}\langle\tilde{u}, v\rangle + \sup_{w\in\mathbb{B}_2^d(0,1/2)}\langle w, v\rangle = \max_{\tilde{u}\in\Xi}\langle\tilde{u}, v\rangle + \frac{1}{2}\|v\|_2,$$

which implies that $\|v\|_2 \leq 2\max_{\tilde{u}\in\Xi}\langle\tilde{u}, v\rangle$. Thus,

$$\begin{aligned}\mathbb{P}\left(\max_{\vartheta\in\Gamma}\|\nabla\mathcal{L}(\vartheta;S)-\nabla\mathcal{L}(\vartheta;P)\|_2 \geq t/2\right) &\leq \mathbb{P}\left(\max_{\vartheta\in\Gamma, v\in\Xi}\langle v, \nabla\mathcal{L}(\vartheta;S)-\nabla\mathcal{L}(\vartheta;P)\rangle \geq t/4\right)\\ &\leq |\Gamma|\cdot|\Xi| e^{-\frac{mt^2}{2\sigma^2}}\\ &= 5^d\big(1+\tfrac{2R}{\Delta}\big)^d e^{-\frac{mt^2}{2\sigma^2}},\end{aligned}$$

where the penultimate line follows from a union bound and Assumption A4 which guarantees that $\nabla\mathcal{L}(\vartheta;S)$ is a σ^2/m-sub-Gaussian vector. We set $t = \sigma\sqrt{\frac{2}{m}(d\log(5+10R/\Delta)+\log(n/\alpha))}$. Picking $\Delta = \min\{1, \frac{\sqrt{2}\sigma}{4H}\sqrt{\frac{d}{m}}\}$ and applying a union bound over n points conclude the proof.

□

E.2 Stochastic gradient algorithms

Algorithm 7 Generic optimization algorithm

1: **Input:** Number of steps T, stochastic first-order oracle O_{F,ν^2}, optimization algorithm with $\{\mathcal{O}, \mathsf{Query}, \mathsf{Update}, \mathsf{Aggregate}\}$, initial output o_0.
2: **for** $t = 0, \ldots, T-1$ **do**
3: $\quad\theta_t \leftarrow \mathsf{Query}(o_t)$.
4: $\quad g_t \leftarrow \mathsf{O}_{F,\nu^2}(\theta_t)$.
5: $\quad o_{t+1} \leftarrow \mathsf{Update}(o_t, g_t)$.
6: **end for**
7: **return** $\widehat{\theta}_T \leftarrow \mathsf{Aggregate}(o_0, \ldots, o_T)$.

Proposition 4 (Convergence of stochastic gradient methods)**.** *Let $F : \Theta \to \mathbb{R}$ be an H-smooth function. Assume that we have access to a stochastic first-order gradient oracle with variance bounded by ν^2, denoted by O_{F,ν^2}. In each of the following cases, let T be the desired number of calls to O_{F,ν^2}, there exist an optimization algorithm—defined by* Update*,* Query *and* Aggregate *and used as in Algorithm 7—with output $\widehat{\theta}_T \in \Theta$ such that the following convergence guarantees hold.*

(i) [16, Theorem 6.3] Assume F is convex, then it holds that

$$\mathbb{E}[F(\widehat{\theta}_T) - \inf_{\theta'\in\Theta} F(\theta')] \leq O\left(\frac{HR^2}{T} + \frac{\nu R}{\sqrt{T}}\right). \tag{21}$$

(ii) [43, Corollary 32] Assume that F is μ-strongly-convex, and that we have access to $\theta_0 \in \Theta$ such that $F(\theta_0) - \inf_{\theta'\in\Theta} F(\theta') \leq \Delta_0$, then it holds that

$$\mathbb{E}[F(\widehat{\theta}_T) - \inf_{\theta'\in\Theta} F(\theta')] \leq O\left(\Delta_0 \exp\left(-\frac{\mu}{H}T\right) + \frac{\nu^2}{\mu T}\right). \tag{22}$$

(iii) [20, Corollary 3.6] Let us define the gradient mapping $\mathsf{G}_{F,\gamma}$

$$\mathsf{G}_{F,\gamma}(\theta) := \frac{1}{\gamma}[\theta - \Pi_\Theta(\theta - \gamma\nabla F(\theta))].$$

Assume that we have access to θ_0 such that $\|\mathsf{G}_{F,1/H}(\theta_0)\|_2 - \inf_{\theta'}\|\mathsf{G}_{F,1/H}(\theta')\|_2 \leq \Delta_1$, it holds that

$$\mathbb{E}\|\mathsf{G}_{F,1/H}(\widehat{\theta}_T)\|_2^2 \leq O\left(\frac{H\Delta}{T} + \nu\sqrt{\frac{H\Delta_1}{T}}\right). \tag{23}$$

Remark 2. For convex functions, the algorithm is fixed-stepsize, averaged, projected SGD. For strongly-convex functions, the algorithm consists of projected SGD with a fixed stepsize and non-uniform averaging followed by a single restart with decreasing stepsize. Finally, in the non-convex case, the Query and Update sub-routine are also projected SGD with fixed stepsize while the Aggregate selects one of the past iterates uniformly at random.

E.3 Proof of Theorem 3

Theorem 3 (Privacy and utility guarantees for ERM)**.** *Assume A2 holds and recall that* $\widetilde{G} = \sigma\sqrt{d}$*, assume*[11] $n = \tilde{\Omega}(\sqrt{dT}/\varepsilon)$ *and let* $\widehat{\theta}$ *be the output of Algorithm 3. There exists variants of projected SGD (e.g. the ones we present in Proposition 4) such that, with probability greater than* $1-\gamma$*:*

(i) If for all $z \in \mathcal{Z}$, $\ell(\cdot; z)$ *is convex, then*

$$\mathbb{E}\left[\mathcal{L}(\widehat{\theta}; \mathcal{S}) - \inf_{\theta' \in \Theta} \mathcal{L}(\theta'; \mathcal{S}) \,\middle|\, \mathcal{S}\right] = \tilde{O}\left(\frac{R^2 H}{T} + R\widetilde{G}\frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\right).$$

(ii) If for all $z \in \mathcal{Z}$, $\ell(\cdot; z)$ *is* μ*-strongly-convex, then*

$$\mathbb{E}\left[\mathcal{L}(\widehat{\theta}; \mathcal{S}) - \inf_{\theta' \in \Theta} \mathcal{L}(\theta'; \mathcal{S}) \,\middle|\, \mathcal{S}\right] = \tilde{O}\left(GR\exp\left(-\tfrac{\mu}{H}T\right) + \widetilde{G}^2\frac{d}{\mu n^2 m\varepsilon^2}\right).$$

(iii) Otherwise, defining the gradient mapping[12] $\mathsf{G}_{F,\gamma}(\theta) := \frac{1}{\gamma}[\theta - \Pi_\Theta(\theta - \gamma\nabla F(\theta))]$*, we have*

$$\mathbb{E}\left[\|\mathsf{G}_{\mathcal{L}(\cdot;\mathcal{S}),1/H}(\widehat{\theta})\|_2^2 | \mathcal{S}\right] = \tilde{O}\left(\frac{H^2 R}{T} + HR\widetilde{G}\frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\right).$$

For $\varepsilon \le 1, \delta > 0$*, Algorithm 3 instantiated with any first-order gradient algorithm is* (ε, δ)*-user-level DP. In the case that only A1 holds, the same guarantees hold whenever* $\Delta \le \mathsf{poly}(d, \frac{1}{n}, \frac{1}{m}, \frac{1}{\varepsilon})$*.*

Proof. First note that the gradient estimation steps (Step 5 and 6) in Algorithm 3 can be viewed as answering T adaptively chosen queries.

Privacy guarantees. The privacy guarantee follows directly from Theorem 10.

Utility guarantees. By Proposition 1, we have the gradients are $(\tau, \gamma/3)$-concentrated with $\tau = \sigma\sqrt{d\log\left(\frac{RHm}{d\sigma}\right)/m + \log\left(\frac{3n}{\gamma}\right)/m}$. Hence, Theorem 10 guarantees that

$$(\bar{g}_0, \ldots, \bar{g}_{T-1}) \sim_\beta (\bar{g}'_0, \ldots, \bar{g}'_{T-1}),$$

where $\beta = \min\left\{1, \frac{2\gamma}{3} + \frac{d^2 TB\sqrt{\log(3dTn/\gamma)}}{\tau}\exp\left(-\frac{n\varepsilon}{48\sqrt{2dT\log(2/\delta)\log(2T/\delta)}}\right)\right\}$ and $\forall i \in [T]$, $\bar{g}'_0$ is from $\mathsf{O}_{\mathcal{L}(\cdot;\mathcal{S}),\nu^2}(\theta_t)$ with

$$\nu^2 \le \frac{8c_0 dT\tau^2\log(3Tdn/\gamma)\log^2(2T/\delta)}{n^2\varepsilon^2} \le \frac{8c_0 d^2 T\sigma^2\log(3Tdn/\gamma)\log^2(2T/\delta)\log(3RHmn/d\sigma\gamma)}{n^2\varepsilon^2}.$$

Moreover, when $n \ge \tilde{\Omega}(1)\sqrt{dT\log(2/\delta)\log(2T/\delta)\log(dmTB/\sigma\gamma)}/\varepsilon$, where $\tilde{\Omega}(1)$ hides log-log factors, we have $\beta < \gamma$.

Convergence rates Finally, depending on the assumptions on the function $\mathcal{L}(\cdot; \mathcal{S})$, we use the various results of Proposition 4 for the value of ν above. To make the results simpler we note that for (ii) of Proposition 4, we upper bound Δ_0 by GR and for (iii), we upper bound Δ_1 by HR^2. This concludes the proof.

□

F Proofs for Section 5

F.1 Proofs for Theorem 4

We begin with a result that guarantees that the (regularized) empirical risk minimizer has good generalization properties. It relies on a combination of convex analysis and stability arguments. This proof exists in the literature (see, e.g. [51]), we add it here for completeness and with some small variation: (1) that the optimization is constrained (2) that Assumption A4 might improve stability when $\sigma\sqrt{d} \le G$.

[11] For precise log factors, see Appendix E.3.

[12] In the unconstrained case—$\Theta = \mathbb{R}^d$—this corresponds to an ϵ-stationary point as $\mathsf{G}_{F,\gamma}(x) = \nabla F(x)$.

Proposition 5 (Generalization properties of regularized ERM). *Let* $(Z_1, \dots, Z_N) \overset{\text{iid}}{\sim} P$. *Let* $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ *be convex, G-Lipschitz with respect to the* $\|\cdot\|_2$ *and such that Assumption A4 holds. Let us denote* $\underline{G} = \min\{G, \sigma\sqrt{d}\}$. *Let*

$$\theta^*_{S,\lambda,\vartheta} \coloneqq \operatorname*{argmin}_{\theta\in\Theta}\left\{\mathcal{L}(\theta; S) + \frac{\lambda}{2}\|\theta - \vartheta\|_2^2\right\}.$$

The following holds

$$\mathbb{E}\big[\mathcal{L}(\theta^*_{S,\lambda,\vartheta}; P)\big] - \mathcal{L}(\theta; P) \le \frac{\lambda}{2}\mathbb{E}\big[\|\theta - \vartheta\|_2^2\big] + \tilde{O}(1)\,\frac{G\underline{G}}{N\lambda}, \;\; \textit{for all} \;\; \theta \in \Theta. \tag{24}$$

Proof. We first show the stability of the minimizer of the regularized empirical risk. Let us consider $S_0 = \{Z_1, \dots, Z_N\}$ and $S_1 = \{Z'_1, \dots, Z'_N\}$ where $Z_j = Z'_j$ for all $j \neq i$ in $[N]$. We first show that

$$\left\|\theta^*_{S,\lambda,\vartheta} - \theta^*_{S',\lambda,\vartheta}\right\|_2 \le \tilde{O}(1)\,\frac{G}{N\lambda}.$$

For conciseness, we denote $\mathcal{L}_b(\theta) \coloneqq \mathcal{L}(\theta; S_b) + \frac{\lambda}{2}\|\theta - \vartheta\|_2^2$ and $\theta^*_{S_b,\lambda,\vartheta} = \theta_b$ for $b \in \{0, 1\}$. Since $\mathcal{L}_0$ is λ-strongly-convex, its gradients are co-coercive, meaning

$$\frac{\lambda}{2}\|\theta_0 - \theta_1\|_2^2 \le \langle \nabla\mathcal{L}_0(\theta_0) - \nabla\mathcal{L}_0(\theta_1), \theta_0 - \theta_1\rangle.$$

First, let us note that $\nabla\mathcal{L}_0(\theta_1) = \nabla\mathcal{L}_1(\theta_1) + \frac{1}{N}(\nabla\ell(\theta_1; Z_i) - \nabla\ell(\theta_1; Z'_i))$. In other words,

$$\frac{\lambda}{2}\|\theta_0 - \theta_1\|_2^2 \le \langle \nabla\mathcal{L}_0(\theta_0), \theta_0 - \theta_1\rangle + \langle \nabla\mathcal{L}_1(\theta_1), \theta_1 - \theta_0\rangle + \frac{1}{N}\langle \nabla\ell(\theta_1; Z_i) - \nabla\ell(\theta_1; Z'_i), \theta_1 - \theta_0\rangle.$$

Since θ_b is the minimizer of $\mathcal{L}_b(\cdot)$ constrained in Θ for $b \in \{0, 1\}$, by first-order optimiality condition, it holds that

$$\langle \nabla\mathcal{L}_b(\theta_b), \theta_b - \theta_{1-b}\rangle \le 0.$$

Consequently,

$$\frac{\lambda}{2}\|\theta_0 - \theta_1\|_2^2 \le \frac{1}{N}\langle \nabla\ell(\theta_1; Z_i) - \nabla\ell(\theta_1; Z'_i), \theta_1 - \theta_0\rangle \le \frac{1}{N}\|\nabla\ell(\theta_1; Z_i) - \nabla\ell(\theta_1; Z'_i)\|_2\|\theta_1 - \theta_0\|_2.$$

Since $\ell(\cdot; z)$ is G-Lipschitz for all $z \in \mathcal{Z}$, we have that $\|\nabla\ell(\theta_1; Z_i) - \nabla\ell(\theta_1; Z'_i)\|_2 \le 2G$. However, with the addition of Assumption A4, Proposition 1 (applied with $m = 1$) guarantees that with probability greater than $1 - \alpha$,

$$\sup_{\theta\in\Theta}\|\nabla\mathcal{L}(\theta; Z_i) - \nabla\mathcal{L}(\theta; P)\| \le \tilde{O}(1)\,\sigma\sqrt{d},$$

where we note that the dependence is only *logarithmic* in α. This immediately yields that with probability greater than $1 - \alpha$,

$$\frac{\lambda}{2}\|\theta_0 - \theta_1\|_2 \le \tilde{O}(1)\,\frac{G}{\lambda N}.$$

Finally, this implies that

$$\text{for all } z \in \mathcal{Z}, \mathbb{E}[|\ell(\theta_0; z) - \ell(\theta_1; z)|] \le G\mathbb{E}[\|\theta_0 - \theta_1\|_2] \le \tilde{O}(1)\,\frac{G\underline{G}}{\lambda N},$$

by G-Lipschitzness of ℓ and setting $\alpha = \frac{G}{\lambda N R}$, or in the language of stability (see e.g. [14]), $S \to \theta^*_{S,\lambda,\vartheta}$ is $\frac{G\underline{G}}{\lambda N}$-uniformly-stable. Standard stability arguments let us conclude the proof. □

We now state and prove Theorem 4.

Theorem 4 (Phased ERM for SCO). *Algorithm 4 is user-level* (ε, δ)*-DP. When A2 holds and* $n = \tilde{\Omega}(\min\{\sqrt[3]{d^2mH^2R^2/(G\underline{G}\varepsilon^4)}, HR\sqrt{m}/(\sigma\varepsilon)\})$, *or, equivalently,* $H = \tilde{O}(\sqrt{\frac{n^2\varepsilon^2\sigma^2}{R^2m} + \frac{G\underline{G}n^3\varepsilon^4}{d^2R^2m}})$ *for all* P *and* ℓ *satisfying Assumptions A3 and A4, we have*

$$\mathbb{E}\left[\mathcal{L}(\mathsf{A}_{\mathsf{PhasedERM}}(\mathcal{S}); P_0)\right] - \min_{\theta'\in\Theta}\mathcal{L}(\theta'; P_0) = \tilde{O}\left(\frac{R\sqrt{G\underline{G}}}{\sqrt{mn}} + R\widetilde{G}\frac{\sqrt{d}}{n\sqrt{m}\varepsilon}\right).$$

Furthermore, our results still hold in the heterogeneous setting (Assumption A1) whenever $\Delta \le \mathsf{poly}(d, \frac{1}{n}, \frac{1}{m}, \frac{1}{\varepsilon})$*; the risk guarantee being with respect to any user distribution* P_u.

Proof. The proof hinges on repeatedly using of Corollary 2 and Proposition 5 after decomposing the excess risk. Recall that $\widehat{\theta}_t$ is the output of round t i.e.

$$\widehat{\theta}_t \approx \operatorname*{argmin}_{\theta\in\Theta} \mathcal{L}(\theta; S_t) + \frac{\lambda_t}{2}\|\theta - \widehat{\theta}_{t-1}\|_2^2.$$

We denote by θ_t^* the *true* minimizer at round t i.e.

$$\theta_t^* \coloneqq \operatorname*{argmin}_{\theta\in\Theta} \mathcal{L}(\theta; S_t) + \frac{\lambda_t}{2}\|\theta - \widehat{\theta}_{t-1}\|_2^2.$$

Let us denote $\theta^* = \operatorname{argmin}_{\theta\in\Theta} \mathcal{L}(\theta; P)$, we decompose the regret in the following way

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\widehat{\theta}_T; P) - \mathcal{L}(\theta^*; P)] = \underbrace{\mathbb{E}\Big[\mathcal{L}(\widehat{\theta}_T; P) - \mathcal{L}(\theta_T^*; P)\Big]}_{=:\Delta_0} + \underbrace{\sum_{t=2}^{T} \mathbb{E}\big[\mathcal{L}(\theta_t^*; P) - \mathcal{L}(\theta_{t-1}^*; P)\big]}_{=:\Delta_1} \\ + \underbrace{\mathbb{E}[\mathcal{L}(\theta_1^*; P) - \mathcal{L}(\theta^*; P)]}_{=:\Delta_2}.\end{aligned}$$

By Proposition 5 and because Θ is bounded by R, we directly have that

$$\Delta_2 \le \frac{\lambda_1 R^2}{2} + \tilde{O}\bigg(\frac{G\underline{G}}{\lambda_1 n_1 m}\bigg).$$

Turning to Δ_1, for every $t \in \{2, \dots, T\}$, again by Proposition 5, it holds that

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}(\theta_t^*; P) - \mathcal{L}(\theta_{t-1}^*; P)\big] &\le \frac{\lambda_t}{2}\mathbb{E}\Big[\|\theta_{t-1}^* - \widehat{\theta}_{t-1}\|_2^2\Big] + \tilde{O}\bigg(\frac{G\underline{G}}{\lambda_t n_t m}\bigg). \\ &\le \tilde{O}\bigg(\frac{\lambda_t}{2}\frac{\sigma^2 d^2}{\lambda_{t-1}^2 n_{t-1}^2 m\varepsilon^2} + \frac{G\underline{G}}{\lambda_t n_t m}\bigg) \\ &\le \tilde{O}\bigg(\frac{\sigma^2 d^2}{\lambda_{t-1} n_{t-1}^2 m\varepsilon^2} + \frac{G\underline{G}}{\lambda_t n_t m}\bigg),\end{aligned}$$

where the second inequality is an application of Corollary 2[13] and the third is because $\lambda_{t-1} = \lambda_t/4$. Noting that $\lambda_{t-1} n_{t-1}^2 = 2^{t-1}\lambda n$, we have

$$\Delta_1 \le \tilde{O}\bigg((T-1)\frac{\sigma^2 d^2}{\lambda n^2 m\varepsilon^2} + \frac{G\underline{G}}{\lambda n m}\sum_{t=2}^{T} 2^{-t}\bigg) = \tilde{O}\bigg(\frac{\sigma^2 d^2}{\lambda n^2 m\varepsilon^2} + \frac{G\underline{G}}{\lambda n m}\bigg),$$

where we use that T is logarithmic. Finally, using Corollary 2, and that $\mathcal{L}(\cdot; P)$ is G-Lipschitz, we have that

$$\Delta_0 \le \mathbb{E}\Big[G\|\theta_T^* - \widehat{\theta}_T\|_2\Big] \le G\sqrt{\mathbb{E}\Big[\|\theta_T^* - \widehat{\theta}_T\|_2^2\Big]} = \tilde{O}\bigg(\frac{G\sigma d}{2^T \lambda n\sqrt{m}\varepsilon}\bigg).$$

Combining the upper bounds, we have

$$\mathbb{E}\Big[\mathcal{L}(\widehat{\theta}_T; P) - \mathcal{L}(\theta^*; P)\Big] = \tilde{O}\bigg(\frac{G\sigma d}{2^T \lambda n\sqrt{m}\varepsilon} + \frac{\sigma^2 d^2}{\lambda n^2 m\varepsilon^2} + \frac{G\underline{G}}{\lambda n m} + \frac{\lambda_1 R^2}{2} + \frac{G\underline{G}}{\lambda_1 n_1 m}\bigg),$$

and setting $T = \lceil \log_2(\frac{Gn\sqrt{m}\varepsilon}{\sigma d})\rceil$ and $\lambda = \sqrt{\frac{\sigma^2 d^2}{n^2 m\varepsilon^2} + \frac{G\underline{G}}{nm}}/R$ yields the final result.

□

[13]The condition on n for the corollary holds when the condition on n is satisfied in the Theorem statement.

F.2 Proofs of Theorem 5

Theorem 5 (Lower bound for SCO)**.** *There exists a distribution P and a loss ℓ satisfying Assumptions A3 and A4 such that for any algorithm A satisfying (ε, δ)-DP at user-level, we have*

$$\mathbb{E}\left[\mathcal{L}(\mathsf{A}(\mathcal{S}); P)\right] - \min_{\theta' \in \Theta} \mathcal{L}(\theta'; P) = \Omega\left(\frac{RG}{\sqrt{mn}} + RG\frac{\sqrt{d}}{n\sqrt{m\varepsilon}}\right).$$

The first term is a lower bound for SCO without any constraints [44, 2]. We only prove the second term here. Note that without loss of generality, we can assume $G \geq 20\sigma\sqrt{d}$ and prove a lower bound of $\Omega(R\widetilde{G}\sqrt{d}/n\sqrt{m\varepsilon})$. Else, we set $\sigma' = G/(20\sqrt{d})$ and embed the original problem into a lower-dimensional (thus easier) problem where the gradients are σ'^2sub-Gaussian. In the rest of the section, we consider $\Theta = \mathbb{B}_2^d(0, R)$ for $R > 0$. As we explained in Section 5.2, we consider the following loss[14]

$$\ell(\theta; z) \coloneqq -\langle \theta, z \rangle.$$

Finally, we define (a collection) of data distributions. For a mean $\mu \in \mathbb{R}^d$, a covariance $\Sigma \in \mathbb{R}^{d \times d}$ and $B > 0$, we consider the family of ℓ_∞-truncated Gaussians. Recall the definition in (20),

$$Z \sim \mathsf{N}^{\mathrm{tr}}(\mu, \Sigma, B) \text{ if } Z_0 \sim \mathsf{N}(\mu, \Sigma) \text{ and set for all } j \in [d] \;\; Z(j) = \frac{Z_0(i)}{\max\{1, |Z_0(j)|/B\}}.$$

In other words, the standard high-dimension Gaussian distribution where the mass outside of $\mathbb{B}_\infty^d(0, B)$ has been radially projected back on the sphere on each dimension.

Consequently, considering the data distribution $P = \mathsf{N}^{\mathrm{tr}}(\mu, \sigma^2 I_d, G/\sqrt{d})$, ℓ is almost surely G-Lipschitz. Additionally, both assumptions A3 and A4 hold.

We now formally state the reduction from SCO to Gaussian mean-estimation. The main difficulty is that the mean of $\mathsf{N}^{\mathrm{tr}}(\mu, \sigma^2 I_d, G/\sqrt{d})$ and $\mathsf{N}(\mu, \sigma^2 I_d)$ do not coincide. However, we show that when G is sufficiently large compared to $\|\mu\|_2$—which implies that we rarely clip—then the reduction holds.

Proposition 6 (Reduction from SCO to Gaussian mean estimation with item-level DP constraints)**.** *Let $B > 0, \sigma > 0, G > 0$ such that $B + 10\sigma\sqrt{d} < G$, we consider the following collections of distributions*

$$\mathcal{P}_{\sigma,B} \coloneqq \left\{\mathsf{N}(\mu, \sigma^2 I_d) : \|\mu\|_2 \in [B/2, B]\right\} \text{ and } \mathcal{P}^{\mathrm{tr}}_{\sigma,B,G/\sqrt{d}} \coloneqq \left\{\mathsf{N}^{\mathrm{tr}}(\mu, \sigma^2 I_d, G/\sqrt{d}) : \|\mu\|_2 \in [B/2, B]\right\}.$$

The following reduction holds

$$\inf_{\substack{\mathsf{A}:\mathcal{Z}^n \to \Theta \\ \mathsf{A} \in \mathcal{A}^{\mathsf{item}}_{\varepsilon,\delta}}} \sup_{P \in \mathcal{P}^{\mathrm{tr}}_{\sigma,B,G/\sqrt{d}}} \mathbb{E}_P\left[\mathcal{L}(\mathsf{A}(Z^n); P) - \inf_{\theta' \in \Theta} \mathcal{L}(\theta'; P)\right] \geq$$
$$\frac{BR}{4} \inf_{\substack{\widehat{u}:\mathcal{Z}^n \to \mathbb{S}^{d-1} \\ \widehat{u} \in \mathcal{A}^{\mathsf{item}}_{\varepsilon,\delta}}} \sup_{P \in \mathcal{P}_{\sigma,B}} \mathbb{E}_P\left[\|\widehat{u}(Z^n) - \mu/\|\mu\|_2\|_2^2\right] + O\big(R\sigma e^{-10d}\big),$$

where we recall that $\mathcal{A}^{\mathsf{item}}_{\varepsilon,\delta}$ is the set of (ε, δ)-item-level DP algorithm for which the domain and co-domain are clear from context.

Before proving the proposition, we prove a Lemma that says, as previewed, that when G is large enough compared to $\|\mu\|_2$ and σ, then the expectation of $\mathsf{N}^{\mathrm{tr}}(\mu, \sigma^2 I_d, G/\sqrt{d})$ and μ are exponentially close in ℓ_2-norm.

Lemma 3. *Suppose $\|\mu\|_2 + 10\sqrt{d}\sigma < G$,*

$$\|\mathbb{E}_{Z \sim \mathsf{N}^{\mathrm{tr}}(\mu, \sigma^2 I_d, G/\sqrt{d})}[Z] - \mu\|_2 = O\big(\sigma e^{-10d}\big).$$

[14] The negative sign is here for convenience; a positive sign would entail reducing it to finding the *negative* normalized mean.

Proof of Lemma 3. It would be enough to show that $\forall i \in [d]$,

$$|\mathbb{E}_{Z\sim \mathsf{N}^{\mathrm{tr}}(\mu,\sigma^2 I_d, G/\sqrt{d})}[Z](i) - \mu(i)| = O(\sigma e^{-10d}/\sqrt{d}).$$

Let $\alpha = \frac{\mu(i)+G/\sqrt{d}}{\sigma}$, $\beta = \frac{\mu(i)-G/\sqrt{d}}{\sigma}$ and $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ be the density function of $N(0,1)$. We have

$$\mathbb{E}_{Z\sim \mathsf{N}^{\mathrm{tr}}(\mu,\sigma^2 I_d, G/\sqrt{d})}[Z](i) = \mu(i) - \sigma \frac{\phi(\alpha) - \phi(\beta)}{\int_\alpha^\beta \phi(x)dx}.$$

Plugging in $\|\mu\|_2 + 10\sqrt{d}\sigma < G$ we obtain the lemma. □

We can now prove the proposition.

Proof. Let $P = \mathsf{N}^{\mathrm{tr}}(\mu, \sigma^2 I_d, G/\sqrt{d})$ and denote $\mu^{\mathrm{tr}} = \mathbb{E}_P[Z]$ the mean of the truncated Gaussian. We consider $\theta_0 = -R\frac{\mu}{\|\mu\|_2}$, in other words the minimum of $\mathcal{L}(\theta; P)$, if the Gaussian was not truncated. Let $\theta \in \Theta$, we have that

$$\begin{aligned}
\mathcal{L}(\theta; P) - \mathcal{L}(\theta^*; P) &\geq -\mathcal{L}(\theta; P) - \mathcal{L}(\theta_0; P) \\
&= -\langle \theta - \theta_0, \mu^{\mathrm{tr}} \rangle \\
&= -\langle \theta - \theta_0, \mu \rangle - \langle \theta - \theta_0, \mu^{\mathrm{tr}} - \mu \rangle \\
&\geq -\langle \theta - \theta_0, \mu \rangle + 2 \inf_{\theta'} \langle \theta', \mu^{\mathrm{tr}} - \mu \rangle \\
&= -\langle \theta - \theta_0, \mu \rangle + O(R\sigma e^{-10d}),
\end{aligned}$$

where the final line uses the fact that $\inf_{\|v\|_2 \leq R} \langle u, v \rangle = -R\|u\|_2$ and Lemma 3.

Moreover, we have

$$\begin{aligned}
\langle \theta_0 - \theta, \mu \rangle &= R\|\mu\|_2 \left(1 - \left\langle \tfrac{\theta}{R}, \tfrac{\mu}{\|\mu\|_2} \right\rangle\right) \\
&\geq \frac{R\|\mu\|_2}{2} \left(\left\| \tfrac{\theta}{R} \right\|_2^2 + \left\| \tfrac{\mu}{\|\mu\|_2} \right\|_2^2 - 2\left\langle \tfrac{\theta}{R}, \tfrac{\mu}{\|\mu\|_2} \right\rangle \right) \\
&= \frac{R\|\mu\|_2}{2} \left\| \tfrac{\theta}{R} - \tfrac{\mu}{\|\mu\|_2} \right\|_2^2,
\end{aligned}$$

where we used that $\|\theta/R\| \leq 1$ and completed the square.

We now finally prove the main statement of the proposition. The first observation is that, since the loss is linear, we only need to consider estimators $\mathsf{A} : \mathcal{Z}^n \to \Theta$ such that $\|\mathsf{A}(z^n)\|_2 = R$ for all

$z^n \in \mathcal{Z}^n$, as the minimum is always on the boundary[15]. Consequently, we have

$$\begin{aligned}
&\inf_{\mathsf{A}:|\mathsf{A}|_2\le R} \sup_{P\in\mathcal{P}^{\mathrm{tr}}_{\sigma,B,G/\sqrt{d}}} \mathbb{E}[\mathcal{L}(\mathsf{A}(Z^n);P) - \min_{\theta'\in\Theta}\mathcal{L}(\theta';P)] \\
&= \inf_{\mathsf{A}:|\mathsf{A}|_2= R} \sup_{P\in\mathcal{P}^{\mathrm{tr}}_{\sigma,B,G/\sqrt{d}}} \mathbb{E}[\mathcal{L}(\mathsf{A}(Z^n);P) - \min_{\theta'\in\Theta}\mathcal{L}(\theta';P)] \\
&\ge \inf_{\mathsf{A}:|\mathsf{A}|_2= R} \sup_{P\in\mathcal{P}^{\mathrm{tr}}_{\sigma,B,G/\sqrt{d}}} \mathbb{E}\frac{R\|\mu\|_2}{2}\left\|\tfrac{\mathsf{A}(Z^n)}{R} - \tfrac{\mu}{\|\mu\|_2}\right\|_2^2 + O\big(R\rho e^{-10d}\big) \\
&\ge \inf_{\mathsf{A}:|\mathsf{A}|_2= R} \sup_{P\in\mathcal{P}^{\mathrm{tr}}_{\sigma,B,G/\sqrt{d}}} \frac{RB}{4}\mathbb{E}\left\|\tfrac{\mathsf{A}(Z^n)}{R} - \tfrac{\mu}{\|\mu\|_2}\right\|_2^2 + O\big(R\rho e^{-10d}\big) \\
&= \inf_{\widehat{u}:\|\hat{u}\|=1} \sup_{P\in\mathcal{P}^{\mathrm{tr}}_{\sigma,B,G/\sqrt{d}}} \frac{RB}{4}\mathbb{E}\left\|\widehat{u}(Z^n) - \tfrac{\mu}{\|\mu\|_2}\right\|_2^2 + O\big(R\rho e^{-10d}\big) \\
&\ge \inf_{\widehat{u}:\|\hat{u}\|=1} \sup_{P\in\mathcal{P}_{\sigma,B}} \frac{RB}{4}\mathbb{E}\left\|\widehat{u}(Z^n) - \tfrac{\mu}{\|\mu\|_2}\right\|_2^2 + O\big(R\rho e^{-10d}\big),
\end{aligned}$$

where the last line uses that we can always sample from $\mathsf{N}^{\mathrm{tr}}(\mu,\sigma^2 I_d, G/\sqrt{d})$ using samples from $\mathsf{N}(\mu,\sigma^2 I_d)$ and truncating them, thus the problem over $\mathcal{P}^{\mathrm{tr}}_{\sigma,B,G}$ is harder than over $\mathcal{P}_{\sigma,B}$. This concludes the proof. □

Because of this reduction, for the remainder of this proof we consider Gaussian mean estimation with user-level DP constraints. Recall that in this setting, we have n users, each having m i.i.d. samples from $\mathsf{N}(\mu,\sigma^2 I_d)$. However, the lower bound of [38] only holds for *item-level* DP constraints. In the next proposition, we show that mean estimation of $\mathsf{N}(\mu,\sigma^2 I_d)$ with n users and m samples per user under user-level DP constraints is equivalent to mean estimation of $\mathsf{N}(\mu,\frac{\sigma^2}{m}I_d)$ with n samples under *item-level constraints*. In other words, any user-level DP estimator taking as input $n\cdot m$ samples is equivalent to an item-level DP estimator taking as input n samples corresponding the each user's average.

Proposition 3 (From multiple samples to one good sample)**.** *Suppose each user $u\in[n]$ observe $(Z_1^{(u)},\dots,Z_m^{(u)}) \overset{\text{iid}}{\sim} \mathsf{N}(\mu,\sigma^2 I_d)$. For any (ε,δ) user-level DP algorithm $\mathsf{A}^{\mathsf{user}}$, there exists an (ε,δ)-item-level DP algorithm $\mathsf{A}^{\mathsf{item}}$ that takes as input $(\bar{Z}^{(1)},\dots,\bar{Z}^{(n)})$ with $\bar{Z}^{(u)} \coloneqq \frac{1}{m}\sum_{j\le m} Z_j^{(u)}$ and has the same performance as $\mathsf{A}^{\mathsf{user}}$.*

Proof. First of all, note that for Gaussians with unknown mean but known variance, the sample mean is a sufficient statistic. As such, we have that for all $u\in[n]$

$$\text{the distribution of } (Z_1^{(u)},\dots,Z_m^{(u)})|\bar{Z}^{(u)} \text{ does not depend on } \mu.$$

Let us now consider an arbitrary user-level DP estimator $\mathsf{A}^{\mathsf{user}}$ and show how to construct an equivalent item-level DP estimator. When provided with $(\bar{Z}^{(1)},\dots,\bar{Z}^{(n)})$, for each $j\le m$, we can sample

$$\tilde{S}_u = (\tilde{Z}_1^{(u)},\dots\tilde{Z}_m^{(u)}) \overset{\text{iid}}{\sim} (Z_1^{(u)},\dots,Z_m^{(u)})|\bar{Z}^{(u)} \tag{25}$$

and return $\mathsf{A}^{\mathsf{item}}((\bar{Z}^{(u)})_{u\le n}) = \mathsf{A}^{\mathsf{user}}((\tilde{S}_1,\dots,\tilde{S}_n))$. Since the distributions are equal given $\bar{Z}^{(u)}$, in expectation the error is the same. □

This proposition allows us to reduce Gaussian mean estimation with user-level DP, to Gaussian mean estimation with item-level DP albeit with the variance divided by m. We thus conclude with (a

[15] To make this rigorous, we consider Yao's minimax principe. It holds that $\min_{\mathsf{A}:|\mathsf{A}|_2\le R}\max_\mu \mathbb{E}[\bar{\mathcal{L}}(\mathsf{A}(Z^n);P)] = \max_{\mathcal{D}}\min_{\mathsf{A}:|\mathsf{A}|_2\le R}\mathbb{E}_{\mu\sim\mathcal{D}}\mathbb{E}[\bar{\mathcal{L}}(\mathsf{A}(Z^n);P)|\mu]$ where $\bar{\mathcal{L}}(\theta;P) \coloneqq \mathcal{L}(\theta;P) - \inf_{\theta'}\mathcal{L}(\theta';P)$ and $\mathcal{D}$ is a prior over μ. For a given prior $\mathcal{D}$, the Bayes optimal classifier is the minimum of the posterior mean, which means that $\mathsf{A}(Z^n)$ minimizes $\langle\theta,\mathbb{E}[\mu|Z^n]\rangle$ over $\mathbb{B}_2^d(0,R)$ and thus has norm R. We can thus constrain the class of estimators to be of norm exactly R for any prior $\mathcal{D}$. Another application of Yao's minimax principle guarantees that this is also the case for the original (minimax) problem.

slight modification of) the results of [38]. Indeed, we differ only in that their results show that mean estimation is hard, whereas we require that estimating the *direction of the mean* is hard.

First, let us recall the a modified version of the result in [38][16].

Proposition 7 (Kamath et al. [38, Lemma 6.7])**.** *Let* $Z^n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2 I_d)$ *and assume* $\delta \leq \frac{\sqrt{d}}{48\sqrt{2}Bn\sqrt{\log(100Rn/\sqrt{d})}}$*, then it holds that if* $n < d\sigma/(512B\varepsilon)$*,*

$$\inf_{\hat{\mu}, \widehat{\mu} \in \mathcal{A}^{\text{item}}_{\varepsilon,\delta}} \sup_{\mu : B/2 \leq \|\mu\|_2 \leq B} \mathbb{E}\left[\|\widehat{\mu}(Z^n) - \mu\|_2^2\right] \geq \frac{B^2}{6}.$$

Corollary 6 (Estimating the direction of the mean is hard)**.** *Let* $Z^n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2 I_d)$*, set* $B = \frac{d\rho}{512n\varepsilon}$ *and assume that* $\delta \leq \frac{\sqrt{d}}{48\sqrt{2}Bn\sqrt{\log(100Rn/\sqrt{d})}}$*, then it holds that if* $n < d\sigma/(512B\varepsilon)$*,*

$$\inf_{\substack{\widehat{u} : \|\widehat{u}\|_2 = 1 \\ \widehat{u} \in \mathcal{A}^{\text{item}}_{\varepsilon,\delta}}} \sup_{P \in \mathcal{P}_{\sigma,B}} \mathbb{E}\left[\left\|\widehat{u}(Z^n) - \frac{\mu}{\|\mu\|_2}\right\|_2^2\right] \geq \frac{1}{10}.$$

Proof. We prove the corollary by contradiction. Assume there exists an (ε, δ)-DP estimator $\widehat{u}$ such that

$$\sup_{P \in \mathcal{P}_{\sigma,B}} \mathbb{E}\left[\left\|\widehat{u}(Z^n) - \frac{\mu}{\|\mu\|_2}\right\|_2^2\right] < \frac{1}{10}.$$

Then let $\widehat{\mu} = \frac{3}{4B}\widehat{u}$,

$$\begin{aligned}
\mathbb{E}[\|\widehat{\mu}(Z^n) - \mu\|_2^2] &= \mathbb{E}\left[\left\|\frac{3B}{4}\widehat{u}(Z^n) - \mu\right\|_2^2\right] \\
&\leq \mathbb{E}\left[\left\|\frac{3B}{4}\widehat{u}(Z^n) - \|\mu\|_2 \cdot \widehat{u}(Z^n)\right\|\right] + \mathbb{E}\left[\|\|\mu\|_2 \cdot \widehat{u}(Z^n) - \mu\|_2^2\right] \\
&= \left(\frac{3B}{4} - \|\mu\|_2\right)^2 + \|\mu\|_2^2 \,\mathbb{E}\left[\left\|\widehat{u}(Z^n) - \frac{\mu}{\|\mu\|_2}\right\|_2^2\right] \\
&\leq \frac{B^2}{16} + \frac{B^2}{10} \\
&< \frac{B^2}{6},
\end{aligned}$$

which contradicts with Proposition 7. □

Applying Corollary 6 with $B = \sigma/\sqrt{m}$ concludes the proof of the lower bound.

[16] It is not guaranteed in the lower bound construction of [38] that $B/2 \leq \|\mu\|_2 \leq B$. In their construction, the mean is taken uniformly from $[-\sqrt{2}B/\sqrt{d}, \sqrt{2}B/\sqrt{d}]^d$. However, the probability that the mean in the lower bound construction being out of this range is exponentially small in d. Hence the same lower bound can be obtained by straightforward modifications of the construction.