Learning with User-Level Privacy

Anonymous Author(s)
Affiliation
Address
email

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, 31]. 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 [23] 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 change if *any single* element of a user's data changes. 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] and bounding user contributions in ML models [4, 27]. Differentially private SQL with bounded user

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

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 (central) user-level privacy. 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 [25]. 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 [26, 9, 28, 29] 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, 30] 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[1] and the gradients are sub-Gaussian random vectors, we prove matching upper

[1] 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.

(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 $(X_1, \ldots, X_n)$. 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.

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

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

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[2]

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

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.*

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

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

[2] 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 $\mathsf{median}(m_1, \ldots, m_n)$ using techniques from [46].

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 [22, 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[3], 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

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

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 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:

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

[3] 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.

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 $1 - \gamma$,*

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

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*

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

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 $\|P_1 - P_2\|_{\mathsf{TV}} \leq \beta$. 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 $\mathbb{E}\left[\|\mathsf{A}(X^n) - \mu\|_2^2\right]$. In the user-level setting (in the homogeneous case), one observes a dataset $\mathcal{S}$ sampled as in (2) and wishes to minimizes $\mathbb{E}[\|\mathsf{A}(\mathcal{S}) - \mathbb{E}P_0\|_2^2]$ under user-level privacy constraints. We first focus on the scalar case.

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 $\xi \sim \text{Lap}\left(0, \frac{8\tau}{\varepsilon n}\right)$ and return

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

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

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 [28], detailed in Algorithm 6 in Appendix D.1, to privately estimate a crude interval 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 [28], 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: $X^n := (X_1, X_2, ..., X_n), X_i \in [-B, B]^n$, 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 $U = d^{-1/2}HD$, where 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 $\mathsf{A}(X^n) = \textbf{WinsorizedMeanHighD}(X^n, \varepsilon, \delta, \tau, B, \alpha)$ 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 $\beta = \min\left\{1, \gamma + \alpha + \frac{d^2B\sqrt{\log(dn/\alpha)}}{\tau}\exp\left(-\frac{n\varepsilon}{24\sqrt{d\log(1/\delta)}}\right)\right\}$.

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 $n \ge (c_1\sqrt{d\log(1/\delta)}/\varepsilon)\log(m(dn + n^2\varepsilon^2))$ for a numerical constant c_1, we have*[4]

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

This bound is optimal up to logarithmic factors. When only A1 holds with $\Delta \le \mathsf{poly}(d, \frac{1}{n}, \frac{1}{m}, \frac{1}{\varepsilon})$, the same error bounds holds (up to constant) for estimating $\mathbb{E}_{P_u}[Z]$ for any $u \in [n]$.

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

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.

Algorithm We recall the ERM setting with user-level DP. We observe $\mathcal{S} = (S_1, \ldots, S_n)$ 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.

Theorem 3 (Privacy and utility guarantees for ERM)**.** *Assume A2 holds and recall that $\widetilde{G} = \sigma\sqrt{d}$, assume[5] $n = \tilde{\Omega}(\sqrt{dT}/\varepsilon)$ 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 $1-\alpha$:*

(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

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

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

Algorithm 3 Winsorized First-Order Optimization

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

(iii) Otherwise, defining the gradient mapping[6] $\mathsf{G}_{F,\gamma}(\theta) := \frac{1}{\gamma}[\theta - \Pi_\Theta(\theta - \gamma\nabla F(\theta))]$, *we have*

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

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 $\nu^2 = \tilde{O}(T\widetilde{G}^2\frac{d}{n^2m\varepsilon^2})$. 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 [24].

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 $T = \frac{H}{\mu}\log\left(n^2m(\underline{G}/\widetilde{G}^2)\frac{\mu R\varepsilon^2}{d}\right)$ and $\alpha = \frac{\sigma^2d^2}{\mu^2n^2m\varepsilon^2R^2}$. For $\theta^*_{\mathcal{S}} \in \operatorname{argmin}_{\theta'\in\Theta}\mathcal{L}(\theta'; \mathcal{S})$, it holds*[7]

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

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 $n = \tilde{\Omega}(\min\{\sqrt[3]{d^2mH^2R^2/(G\underline{G}\varepsilon^4)}, HR\sqrt{m}/(\sigma\varepsilon)\})$, 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*

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

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

[7] 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.

Furthermore, our results still holds 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 [30] 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 [30]. 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 $\{(\lambda_t, n_t)\}_{t \leq T}$ 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 lossesof 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. □

References

[1] J. Acharya, Z. Sun, and H. Zhang. Differentially private Assouad, Fano, and Le Cam. *To appear at Conference on Algorithmic Learning Theory 2021*, 2021.

[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. *Pre-Proceedings of NeurIPS 2020*, 2020.

[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] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *arXiv:1405.0782 [cs.IT]*, 2014.

[23] 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.

[24] 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.

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

[26] 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.

[27] A. Epasto, M. Mahdian, J. Mao, V. Mirrokni, and L. Ren. Smoothly bounding user contributions in differential privacy. *Pre-Proceedings of NeurIPS 2020*, 2020.

[28] 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.

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

[30] 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.

[31] 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.

[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. *arXiv preprint arXiv:1912.04977*, 2019.

[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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] in the first page of the supplementary.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] in the first page of the supplementary.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]