Private Stochastic Convex Optimization: Optimal Rates in Linear Time

Vitaly Feldman[†], Tomer Koren[†*], and Kunal Talwar[†]

[†]Google Research, Mountain View
[*]School of Computer Science, Tel Aviv University

December 5, 2019

Abstract

We study differentially private (DP) algorithms for *stochastic convex optimization*: the problem of minimizing the population loss given i.i.d. samples from a distribution over convex loss functions. A recent work of Bassily et al. (2019) has established the optimal bound on the excess population loss achievable given n samples. Unfortunately, their algorithm achieving this bound is relatively inefficient: it requires $O(\min\{n^{3/2}, n^{5/2}/d\})$ gradient computations, where d is the dimension of the optimization problem.

We describe two new techniques for deriving DP convex optimization algorithms both achieving the optimal bound on excess loss and using $O(\min\{n, n^2/d\})$ gradient computations. In particular, the algorithms match the running time of the optimal non-private algorithms. The first approach relies on the use of variable batch sizes and is analyzed using the privacy amplification by iteration technique of Feldman et al. (2018). The second approach is based on a general reduction to the problem of localizing an approximately optimal solution with differential privacy. Such localization, in turn, can be achieved using existing (non-private) uniformly stable optimization algorithms. As in the earlier work, our algorithms require a mild smoothness assumption. We also give a linear-time algorithm achieving the optimal bound on the excess loss for the strongly convex case, as well as a faster algorithm for the non-smooth case.

1 Introduction

Stochastic convex optimization (SCO) is the problem of minimizing the expected loss (also referred to as *population loss*) $F(w) = \mathbb{E}_{x\sim\mathcal{P}}[f(w,x)]$ for convex loss functions of w over some d-dimensional convex body $\mathcal{K}$ given access to i.i.d. samples $x_1, \ldots, x_n$ from the data distribution $\mathcal{P}$. The performance of an algorithm for the problem is measured by bounding the *excess (population) loss* of a solution w, that is the value $F(w) - \min_{v\in\mathcal{K}} F(v)$. This problem is central to numerous applications in machine learning and arises for example in least squares/logistic regression, or minimizing a convex surrogate loss for a classification problem. It also serves as the basis for the development of continuous optimization algorithms in the non-convex setting. In this work we study this problem with the constraint of differential privacy with respect to the set of samples [DMNS06].

Placing a differential privacy constraint usually comes at a cost in terms of utility. In this case, it is measured by the excess population loss of the solution, for a given number of samples n. Additionally, runtime efficiency of an optimization method is crucial for modern applications on large high-dimensional datasets, and this is the primary reason for the popularity of stochastic gradient descent-based methods. This motivates the problem of understanding the trade-offs between computational efficiency, and excess population loss in the presence of privacy constraints.

Differentially private convex optimization is one of most well-studied problems in private data analysis [CM08, CMS11, JKT12, KST12, ST13, SCS13, DJW13, Ull15, JT14, BST14, TTZ15, STU17, WLK$^+$17, WYX17, INS$^+$19]. However, most of the prior work focuses on the easier problem of minimizing the empirical loss $\hat{F}(w) = \frac{1}{n}\sum_i f(w, x_i)$ (referred to as empirical risk minimization (ERM)) for which tight upper and lower bounds on the excess loss are known in a variety of settings. Upper bounds for the differentially private ERM can be translated to upper bounds on the population loss by appealing to *uniform convergence* of empirical loss to population loss, namely an upper bound on $\sup_{w\in\mathcal{K}}(F(w)-\hat{F}(w))$. However, in general,[1] this approach leads to suboptimal bounds: it is known that there exist distributions over loss functions over $\mathbb{R}^d$ for which the best bound on uniform convergence is $\Omega(\sqrt{d/n})$ [Fel16]. As a result, in the high-dimensional settings often considered in modern ML (when $n = \Theta(d)$), bounds based on uniform convergence are $\Omega(1)$ and do not lead to meaningful bounds on population loss.

The first work to address the population loss for SCO with differential privacy (DP-SCO) is [BST14] who give a bound of order $\max\{d^{1/4}\big/\sqrt{n}, \varepsilon^{-1}\sqrt{d}\big/n\}$ [BST14, Sec. F].[2] For the most relevant case where $d = \Theta(n)$ and $\varepsilon = \Theta(1)$, this results in a bound of $\Omega(n^{-1/4})$ on excess population loss. More recent work of Bassily et al. [BFTT19] demonstrates the existence of an efficient algorithm that achieves the tight bound of $O(1\big/\sqrt{n} + \varepsilon^{-1}\sqrt{d}\big/n)$. Notably, this bound is comparable to the non-private SCO bound of $O(1/\sqrt{n})$ as long as $d/\varepsilon^2 = O(n)$. Their algorithm is based on solving the ERM via noisy stochastic gradient descent (SGD) [BST14] but requires relatively large batch sizes for the privacy analysis. As a result, their algorithm uses $O(\min\{n^{3/2}, n^{5/2}\big/d\})$ gradient computations. This is substantially less efficient than the optimal non-private algorithms for the problem which require only n gradient evaluations. They also give a near-linear-time algorithm under an additional strong assumption that the Hessian of each loss function is rank-1 over the entire domain.

Along the other axis, several of the aforementioned works on differentially private ERM [WLK$^+$17, WYX17, INS$^+$19] are geared towards finding computationally efficient algorithms for the problem, often at the cost of worse utility bounds.

[1] At the same time, uniform convergence suffices to derive optimal bounds on the excess population loss in a number of special cases, such as regression for generalized linear models.

[2] For clarity, in the introduction we focus on the dependence on d, n and ε for (ε, δ)-DP, suppressing the dependence on δ and on parameters of the loss function such as Lipschitz constant and the diameter of $\mathcal{K}$.

1.1 Linear-time Algorithms

We describe two new techniques for deriving linear-time algorithms that achieve the (asymptotically) optimal bounds on the excess population loss. Thus our results show that for the problem of Stochastic Convex Optimization, under mild assumptions, a privacy constraint come for free. For $d \leq n$, there is no overhead in terms of either excess loss or the computational efficiency. When $d \geq n$, the excess loss provable increases, but the optimal bounds can still be achieved without any computational overhead. Unlike the earlier algorithm [BFTT19] that solves the ERM and relies on uniform stability of the algorithm to ensure generalization, our algorithms directly optimize the population loss.

Formally, our algorithms satisfy the following bounds:

Theorem 1.1. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set of diameter D and $\{f(\cdot, x)\}_{x\in\mathcal{X}}$ be a family of convex L-Lipschitz and β-smooth functions over $\mathcal{K}$. For every $\rho > 0$, there exists an algorithm $\mathcal{A}$ that given a starting point $w_0 \in \mathcal{K}$, and $S \in \mathcal{X}^n$ returns a point $\hat{w}$. For all $\alpha \geq 1$, $\mathcal{A}$ uses n evaluations of the gradient of $f(w,x)$ and satisfies $\left(\alpha, \alpha\rho^2/2\right)$-RDP as long as $\beta \leq c\frac{L}{D}\min(\sqrt{n}, \rho n/\sqrt{d})$, where c is a universal constant. Further, if S consists of samples drawn i.i.d. from a distribution $\mathcal{P}$ over $\mathcal{X}$, then*

$$\mathbb{E}[F(\hat{w})] \leq F^* + O\left(DL\cdot\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\rho n}\right)\right),$$

where, for all $w \in \mathcal{K}$, $F(w) \doteq \mathbb{E}_{x\sim\mathcal{P}}[f(w,x)]$, $F^* \doteq \min_{w\in\mathcal{K}} F(w)$ and the expectation is taken over the random choice of S and randomness of $\mathcal{A}$.

Our guarantees are stated in terms of Rényi differential privacy (RDP) [Mir17] for all orders α and can also be equivalently stated as 0-mean $(\rho^2/2)$-concentrated differential privacy (or $(\rho^2/2)$-zCDP) [BS16]. Standard properties of RDP/zCDP imply that our algorithms satisfy $(2\rho\sqrt{\ln(1/\delta)}, \delta)$-DP for all $\delta > 0$ as long as $\rho \leq \sqrt{\ln(1/\delta)}$. Thus for (ε, δ)-DP our bound is

$$\mathbb{E}[F(\hat{w})] \leq F^* + O\left(DL\cdot\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)\right),$$

matching the tight bound in [BFTT19]. We now overview the key ideas and tools used in these techniques.

Variable batch sizes: Our first algorithm relies on a one-pass noisy SGD with gradually growing batch sizes. Namely, at step t out of T the batch size is proportional to $1/\sqrt{T-t+1}$. The analysis of this algorithm relies on two tools. The first one is privacy amplification by iteration [FMTT18]. This privacy amplification technique ensures that for the purposes of analyzing the privacy guarantees of a point x_i used at step t one can effectively treat all the noise added at subsequent steps as also added to the gradient of the loss at x_i. A direct application of this technique to noisy SGD results in different privacy guarantees for different points [FMTT18] and, as a result, the points used in the last $o(n)$ steps will not have sufficient privacy guarantees. However, we show that by increasing the batch size in those steps we can achieve the optimal privacy guarantees for all the points.

A limitation of relying on this analysis technique is that the privacy guarantees apply only to the algorithm that outputs the last iterate of SGD. In contrast, the optimization guarantees usually apply to the average of all the iterates (see Section 5 for an example in which the privacy guarantees for the average iterate are much worse than those for the last iterate). Thus the second tool we rely on is the recent work of Jain et al. [JNN19] showing that, for an appropriate choice of step sizes in SGD, the last iterate has the (asymptotically) optimal excess population loss. (Without the special step sizes the last iterate has excess loss larger by a $\log n$ factor [SZ13, HLPR19].) See Section 3 for additional details of this approach.

No Privacy Amplification for the Average Iterate: It is natural to ask if the last iterate analysis is really needed or if the average iterate itself be proven to have good privacy properties. In Section 5, we address this question and show that in general the average iterate can be very non-private even when the noise is sufficient to give strong privacy guarantees for the last iterate.

Localization: Our second approach is based on an (implicit) reduction to an easier problem of localizing an approximate minimizer of the population loss. Specifically, the reduction is to a differentially private algorithm that given a point w_0 that is within distance R from the minimizer of the loss, finds a point $\hat{w}$ that is within distance $R/2$ from a point that approximately minimizes the loss. By iteratively using such a localizing algorithm with appropriately chosen parameters, a sufficiently good solution will be found after a logarithmic number of applications of the algorithm. Each application operates on its own subset of the dataset and thus this reduction preserves the privacy guarantees of the localizing algorithm.

A simple way to implement a localization algorithm is to start with non-private SCO algorithm whose output has optimal L_2 sensitivity. Namely, solutions produced by the algorithm on any two datasets that differ in one point are at distance on the order of $R/\sqrt{n}$ (this property is also referred to as *uniform stability* in the parameter space). Given such an algorithm one can simply add Gaussian noise to the output. This is a standard approach to differentially private optimization referred to as output perturbation [CMS11, WLK+17]. However, for the purposes of localization, we only need to be within $R/2$ of the solution output by the algorithm and so we can add much more noise than in the standard applications, thereby getting substantially better privacy guarantees.

We note that in order to ensure that the addition of Gaussian noise localizes the solution with probability at least $1-\alpha$ we would need to increase the noise variance by an additional $\ln(1/\alpha)$ factor making the resulting rate suboptimal by a logarithmic factor. Thus, instead we rely on the fact that for algorithms based on SGD the bound on excess loss can be stated in terms of the second moment of the distance to the optimum.

We can now plug in existing uniformly stable algorithms for SCO. Specifically, it is known that under mild smoothness assumptions, one-pass SGD finds a solution that both achieves optimal bounds on the excess population loss and stability [HRS15, FV19]. This leads to the second algorithm satisfying the guarantees in Theorem 1.1. See Section 4 for additional details of this approach.

Non-smooth case: Both of our algorithms require essentially the same and relatively mild smoothness assumption: namely that the smoothness parameter is at most $\sqrt{n}$ (ignoring the scaling with D, L and for simplicity focusing on the case when $d = O(n)$ and $\varepsilon = 1$). Bassily et al. [BFTT19] show that optimal rates are still achievable even without this smoothness assumption. Their algorithm for the problem relies on using the prox operator instead of gradient steps which is known to be equivalent to gradient steps on the loss function smoothed via the Moreau-Yosida envelope. Unfortunately, computing the prox step with sufficient accuracy requires many gradient computations and very high accuracy is needed due to potential error accumulation. As a result, implementing the algorithm in [BFTT19] requires $O(n^5)$ gradient computations.

Our reduction based technique gives an alternative and simpler way to deal with the non-smooth case. One can simply plug in a uniformly stable algorithms for SCO in the non-smooth case from [SSSSS10]. This algorithm relies on solving ERM with an added strongly convex $\lambda\|w\|_2^2$ term. In this case the analysis of the accuracy to which the ERM needs to be solved is straightforward. However achieving such accuracy with high probability requires $O(n^2)$ gradient computations thus giving an $O(n^2)$ algorithm for the non-smooth version of our problem. Improving this running time is a natural avenue for future work. We remark that finding a faster uniformly stable (non-private) SCO for the non-smooth case is an interesting problem in itself.

Strongly convex case: When the loss functions are strongly convex, the optimal (non-private) excess population loss is of the order of $O(1/n)$ rather than $O(1/\sqrt{n})$. The excess loss due to privacy is known to be $\Omega(d/\varepsilon^2 n^2)$. The best known upper bounds for this problem due to [BST14] are $O(\sqrt{d}/\varepsilon n)$. We show a nearly linear time algorithm that has excess risk matching the known lower bounds. As in the convex case, when $d \leq n$, privacy has virtually no additional cost in terms of utility or efficiency.

2 Preliminaries

2.1 Convex Loss Minimization

Let $\mathcal{X}$ be the domain of data sets, and $\mathcal{P}$ be a distribution over $\mathcal{X}$. Let $S = \{x_1, \ldots, x_n\}$ be a data set drawn i.i.d. from $\mathcal{P}$. Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set denoting the space of all models. Let $f \colon \mathcal{K} \times \mathcal{X} \to \mathbb{R}$ be a loss function, which is convex in its first parameter (the second parameter is a data point and dependence on this parameter can be arbitrary). The excess population loss of solution w is defined as

$$\mathbb{E}_{x \sim \mathcal{P}}\left[f(w,x)\right] - \min_{v \in \mathcal{K}} \mathbb{E}_{x \sim \mathcal{P}}\left[f(v,x)\right].$$

In order to argue differential privacy we place certain assumptions on the loss function. To that end, we need the following two definitions of Lipschitz continuity and smoothness.

Definition 2.1 (L-Lipschitz continuity)**.** *A function $f \colon \mathcal{K} \to \mathbb{R}$ is L-*Lipschitz continuous *over the domain $\mathcal{K} \subseteq \mathbb{R}^d$ if the following holds for all $w, w' \in \mathcal{K}$: $|f(w) - f(w)| \leq L \left\|w - w'\right\|_2$.*

Definition 2.2 (β-smoothness)**.** *A function $f \colon \mathcal{K} \to \mathbb{R}$ is β-*smooth *over the domain $\mathcal{K} \subseteq \mathbb{R}^d$ if for all $w, w' \in \mathcal{K}$, $\|\nabla f(w) - \nabla f(w')\|_2 \leq \beta \left\|w - w'\right\|_2$.*

2.2 Probability Measures

In this work, we will primarily be interested in the d-dimensional Euclidean space $\mathbb{R}^d$ endowed with the ℓ_2 metric and the Lebesgue measure. We say a distribution μ is *absolutely continuous* with respect to ν if $\mu(A) = 0$ whenever $\nu(A) = 0$ for all measurable sets A. We will denote this by $\mu \ll \nu$.

Given two distributions μ and ν on a Banach space $(\mathcal{Z}, \|\cdot\|)$, one can define several notions of distance between them. The primary notion of distance we consider is Rényi divergence:

Definition 2.3 (Rényi Divergence [Rén61])**.** *Let $1 < \alpha < \infty$ and μ, ν be measures with $\mu \ll \nu$. The Rényi divergence of order α between μ and ν is defined as*

$$\mathrm{D}_\alpha(\mu \parallel \nu) \doteq \frac{1}{\alpha - 1} \ln \int \left(\frac{\mu(z)}{\nu(z)}\right)^{\alpha} \nu(z)\,\mathrm{d}z.$$

Here we follow the convention that $\frac{0}{0} = 0$. If $\mu \not\ll \nu$, we define the Rényi divergence to be ∞. Rényi divergence of orders $\alpha = 1, \infty$ is defined by continuity.

2.3 (Rényi) Differential Privacy

The notion of differential privacy is by now a de facto standard for statistical data privacy [DMNS06, Dwo06, DR14].

Definition 2.4 ([DMNS06, DKM+06]). *A randomized algorithm $\mathcal{A}$ is (ε, δ)-differentially private ((ε, δ)-DP) if, for all datasets S and S' that differ in a single data element and for all events $\mathcal{O}$ in the output space of $\mathcal{A}$, we have*

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

Starting with Concentrated Differential Privacy [DR16], definitions that allow more fine-grained control of the privacy loss random variable have proven useful. The notions of zCDP [BS16], Moments Accountant [ACG+16], and Rényi differential privacy (RDP) [Mir17] capture versions of this definition. This approach improves on traditional (ε, δ)-DP accounting in numerous settings, often leading to significantly tighter privacy bounds as well as being applicable when the traditional approach fails [PAE+17, PSM+18].

Definition 2.5 ([Mir17]). *For $1 \leq \alpha \leq \infty$ and $\varepsilon \geq 0$, a randomized algorithm $\mathcal{A}$ is* (α, ε)-Rényi differentially private, *or (α, ε)-RDP if for all neighboring data sets S and S' we have*

$$\mathrm{D}_{\alpha}\big(\mathcal{A}(S) \,\big\|\, \mathcal{A}(S')\big) \leq \varepsilon.$$

The following two lemmas allow translating Rényi differential privacy to (ε, δ)-differential privacy, and give a composition rule for RDP.

Lemma 2.6 ([Mir17, BS16]). *If $\mathcal{A}$ satisfies (α, ε)-Rényi differential privacy, then for all $\delta \in (0, 1)$ it also satisfies $\left(\varepsilon + \frac{\ln(1/\delta)}{\alpha-1}, \delta\right)$-DP. In particular, if $\mathcal{A}$ satisfies $(\alpha, \alpha\rho^2/2)$-RDP for every $\alpha \geq 1$ then for all $\delta \in (0, 1)$ it also satisfies $\left(\rho^2/2 + \rho\sqrt{2\ln(1/\delta)}, \delta\right)$-DP.*

The standard composition rule for Rényi differential privacy, when the outputs of all algorithms are revealed, takes the following form.

Lemma 2.7 ([Mir17]). *If $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are randomized algorithms satisfying, respectively, (α, ε_1)-RDP,...,(α, ε_k)-RDP, then their composition defined as $(\mathcal{A}_1(S), \ldots, \mathcal{A}_k(S))$ is $(\alpha, \varepsilon_1 + \cdots + \varepsilon_k)$-RDP. Moreover, the i'th algorithm can be chosen on the basis of the outputs of algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_{i-1}$.*

2.4 Contractive Noisy Iteration

We start by recalling the definition of a contraction.

Definition 2.8 (Contraction). *For a Banach space $(\mathcal{Z}, \|\cdot\|)$, a function $\psi\colon \mathcal{Z} \to \mathcal{Z}$ is said to be* contractive *if it is 1-Lipschitz. Namely, for all $x, y \in \mathcal{Z}$,*

$$\|\psi(x) - \psi(y)\| \leq \|x - y\|.$$

A canonical example of a contraction is projection onto a convex set in the Euclidean space.

Proposition 2.9. *Let $\mathcal{K}$ be a convex set in $\mathbb{R}^d$. Consider the* projection operator*:*

$$\Pi_{\mathcal{K}}(x) \doteq \arg\min_{y\in\mathcal{K}} \|x - y\|.$$

The map $\Pi_{\mathcal{K}}$ is a contraction.

Another example of a contraction, which will be important in our work, is a gradient descent step for a smooth convex function. The following is a standard result in convex optimization [Nes04].

Proposition 2.10. *Suppose that a function $f \colon \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth. Then the function ψ defined as:*

$$\psi(w) \doteq w - \eta \nabla_w f(w)$$

is contractive as long as $\eta \leq 2/\beta$.

We will be interested in a class of iterative stochastic processes where we alternate between adding noise and applying some contractive map.

Definition 2.11 (Contractive Noisy Iteration (CNI)). *Given an initial random state $X_0 \in \mathcal{Z}$, a sequence of contractive functions $\psi_t \colon \mathcal{Z} \to \mathcal{Z}$, and a sequence of noise distributions $\{\mathcal{D}_t\}$, we define the Contractive Noisy Iteration (CNI) by the following update rule:*

$$X_{t+1} \doteq \psi_{t+1}(X_t) + Z_{t+1},$$

where Z_{t+1} is drawn independently from $\mathcal{D}_{t+1}$. For brevity, we will denote the random variable output by this process after T steps as $CNI_T(X_0, \{\psi_t\}, \{\mathcal{D}_t\})$.

As usual, we denote by $\mu * \nu$ the convolution of μ and ν, that is the distribution of the sum $X + Y$ where we draw $X \sim \mu$ and $Y \sim \nu$ independently.

Definition 2.12. *For a noise distribution $\mathcal{D}$ over a Banach space $(\mathcal{Z}, \|\cdot\|)$ we measure the magnitude of noise by considering the function that for $a > 0$, measures the largest Rényi divergence of order α between $\mathcal{D}$ and the same distribution $\mathcal{D}$ shifted by a vector of length at most a:*

$$R_\alpha(\mathcal{D}, a) \doteq \sup_{x \colon \|x\| \leq a} \mathrm{D}_\alpha(\mathcal{D} * \mathbf{x} \parallel \mathcal{D}).$$

We denote the standard Gaussian distribution over $\mathbb{R}^d$ with variance σ^2 by $\mathcal{N}(0, \sigma^2 \mathbb{I}_d)$. By the well-known properties of Gaussians, for any $x \in \mathbb{R}^d$, and σ, $\mathrm{D}_\alpha\big(\mathcal{N}(0, \sigma^2 \mathbb{I}_d) \parallel \mathcal{N}(x, \sigma^2 \mathbb{I}_d)\big) = \alpha \|x\|_2^2 / 2\sigma^2$. This implies that in the Euclidean space, $R_\alpha(\mathcal{N}(0, \sigma^2 \mathbb{I}_d), a) = \frac{\alpha a^2}{2\sigma^2}$.

When U and V are sampled from μ and ν respectively, we will often abuse notation and write $\mathrm{D}_\alpha(U \parallel V)$.

2.5 Privacy Amplification by Iteration

The main result in [FMTT18] states that

Theorem 2.13. *Let X_T and X'_T denote the output of $CNI_T(X_0, \{\psi_t\}, \{\mathcal{D}_t\})$ and $CNI_T(X_0, \{\psi'_t\}, \{\mathcal{D}_t\})$. Let $s_t \doteq \sup_x \|\psi_t(x) - \psi'_t(x)\|$. Let $a_1, \ldots, a_T$ be a sequence of reals and let $z_t \doteq \sum_{i \leq t} s_i - \sum_{i \leq t} a_i$. If $z_t \geq 0$ for all t, then*

$$\mathrm{D}_\alpha\big(X_T \parallel X'_T\big) \leq \sum_{t=1}^{T} R_\alpha(\mathcal{D}_t, a_t).$$

We now give a simple corollary of this general theorem for the case when the iterative processes differ in a single index and, in addition, the noise distribution with parameter σ ensures that Rényi divergence for a shift of a scales as a^2/σ^2. As discussed above, this is exactly the case for Gaussian distribution.

Corollary 2.14. *Let X_T and X'_T denote the output of $CNI_T(X_0, \{\psi_t\}, \{\mathcal{D}_t\})$ and $CNI_T(X_0, \{\psi'_t\}, \{\mathcal{D}_t\})$. Let $s_i \doteq \sup_x \|\psi_i(x) - \psi'_i(x)\|$. Assume that there exists $t \in [T]$ such that for all $i \neq t$, $s_i = 0$. For $\alpha \geq 1$ assume that there exists γ such that for every $\zeta > 0$ and $a \geq 0$, and $i \in [T]$, $R_\alpha(\mathcal{D}_i, a) \leq \gamma \frac{a^2}{\sigma_i^2}$ for some σ_i. Then*

$$\mathrm{D}_\alpha\big(X_T \,\|\, X'_T\big) \leq \gamma \frac{s_t^2}{\sum_{i=t}^T \sigma_i^2}.$$

Proof. We use Theorem 2.13 with $a_i = 0$ for $i < t$ and $a_i = \frac{s_t \sigma_i^2}{\sum_{v=t}^T \sigma_v^2}$. The resulting bound we get

$$\mathrm{D}_\alpha\big(X_T \,\|\, X'_T\big) \leq \sum_{i=t}^T \gamma \frac{a_i^2}{\sigma_i^2} = \sum_{i=t}^T \gamma \left(\frac{s_t \sigma_i^2}{\sum_{i=t}^T \sigma_i^2}\right)^2 \cdot \frac{1}{\sigma_i^2} = \gamma s_t^2 \sum_{i=t}^T \frac{\sigma_i^2}{\left(\sum_{i=t}^T \sigma_i^2\right)^2} = \gamma \frac{s_t^2}{\sum_{i=t}^T \sigma_i^2}.$$ □

3 DP SCO via Privacy Amplification by Iteration

We start by describing a general version of noisy SGD and analyze its privacy using the privacy amplification by iteration technique from [FMTT18]. Recall that in our problem we are given a family of convex loss functions over some convex set $\mathcal{K} \subseteq \mathbb{R}^d$ parameterized by $x \in \mathcal{X}$, that is $f(w, x)$ is convex and differentiable in the first parameter for every $x \in \mathcal{X}$. Given a dataset $S = (x_1, \ldots, x_n)$, starting point w_0, a number of steps T, batch size parameters $B_1, \ldots, B_T$ such that B_t are positive integers and $\sum_{t \in [T]} B_t = n$, rate parameters $\eta_1, \ldots, \eta_T$, and noise scales $\sigma_1, \ldots, \sigma_T$ the algorithm works as follows. Starting from $w_0 \in \mathcal{K}$ perform the following update $v_{t+1} \doteq w_t - \eta_{t+1}(\nabla_w F_{t+1}(w_t) + \xi_{t+1})$ and $w_{t+1} \doteq \Pi_{\mathcal{K}}(v_{t+1})$, where (1) F_{t+1} is the average of loss functions for samples in batch $t + 1$, that is

$$F_{t+1}(w) \doteq \frac{1}{B_{t+1}} \sum_{i=1+\sum_{s \leq t} B_s}^{i=\sum_{s \leq t+1} B_s} f(w, x_i);$$

(2) ξ_{t+1} is a freshly drawn sample from $\mathcal{N}(0, \sigma_{t+1}^2 \mathbb{I}_d)$; and (3), $\Pi_{\mathcal{K}}$ denotes the Euclidean projection to set $\mathcal{K}$. We refer to this algorithm as PNSGD($S, w_0, \{B_t\}, \{\eta_t\}, \{\sigma_t\}$) and describe it formally in Algorithm 1. For a value a we denote the fixed sequence of parameters $(a, \ldots, a)$ of length T by $\{a\}$.

Algorithm 1 Projected noisy stochastic gradient descent (PNSGD)

```
Input: Data set $S = \{x_1, \ldots, x_n\}$, $f\colon \mathcal{K} \times \mathcal{X} \to \mathbb{R}$ a convex function in the first parameter, starting point
    $w_0 \in \mathcal{K}$, batch sizes $\{B_t\}$, step sizes $\{\eta_t\}$, noise parameters $\{\sigma_t\}$.
for $t \in \{0, \ldots, n-1\}$ do
    $v_{t+1} \leftarrow w_t - \eta_{t+1}(\nabla_w F_{t+1}(w_t)) + \xi_{t+1})$, where $\xi_{t+1} \sim \mathcal{N}(0, \sigma_{t+1}^2 \mathbb{I}_d)$.
    $w_{t+1} \leftarrow \Pi_{\mathcal{K}}(v_{t+1})$, where $\Pi_{\mathcal{K}}(w) = \arg\min_{\theta \in \mathcal{K}} \|\theta - w\|_2$ is the $\ell_2$-projection on $\mathcal{K}$.
return the final iterate $w_n$.
```

3.1 Privacy Guarantees for Noisy SGD

As in [FMTT18], the key property that allows us to treat noisy gradient descent as a contractive noisy iteration is the fact that for any convex function, a gradient step is contractive as long as the function satisfies a relatively mild smoothness condition (see Proposition 2.10). In addition, as is well known, for any convex set

$\mathcal{K} \in \mathbb{R}^d$, the (Euclidean) projection to $\mathcal{K}$ is contractive (see Proposition 2.9). Naturally, a composition of two contractive maps is a contractive map and therefore we can conclude that PNSGD$(S, w_0, \{B_t\}, \{\eta_t\}, \{\sigma_t\})$ is an instance of contractive noisy iteration. More formally, consider the sequence $v_0 = w_0, v_1, \ldots, v_n$. In this sequence, v_{t+1} is obtained from v_t by first applying a contractive map that consists of projection to $\mathcal{K}$ followed by the gradient step at w_t and then addition of Gaussian noise of scale $\eta_{t+1} \cdot \sigma_{t+1}$. Note that the final output of the algorithm is $w_n = \Pi_{\mathcal{K}}(v_n)$ but it does not affect our analysis of privacy guarantees as it can be seen as an additional post-processing step.

More formally, for this algorithm we prove the following privacy guarantees.

Theorem 3.1. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set and $\{f(\cdot, x)\}_{x\in\mathcal{X}}$ be a family of convex L-Lipschitz and β-smooth functions over $\mathcal{K}$. Then, for every batch-size sequence $\{B_t\}_{t\in[T]}$, step-size sequence $\{\eta_t\}_{t\in[T]}$ such that $\eta_t \leq 2/\beta$ for all $t \in [T]$, noise parameters $\{\sigma_t\}_{t\in[T]}$, $\alpha \geq 1$, starting point $w_0 \in \mathcal{K}$, and $S \in \mathcal{X}^n$, PNSGD$(S, w_0, \{B_t\}, \{\eta_t\}, \{\sigma_t\})$ satisfies $\left(\alpha, \alpha \cdot \rho^2/2\right)$-RDP, where*

$$\rho = 2L \cdot \max_{t\in[T]} \left\{ \frac{\eta_t}{B_t\sqrt{\sum_{s=t}^T \eta_s^2\sigma_s^2}} \right\}.$$

Proof. For $k \in [n]$, let $S \doteq (x_1, \ldots, x_n)$ and $S' \doteq (x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_n)$ be two arbitrary datasets that differ at index k and let t be the index of the batch in which k-th example is used by PNSGD with batch-size sequence $\{B_t\}_{t\in[T]}$. Note that each F_t is an average of β-smooth, L-Lipschitz convex functions and thus is itself β-smooth, L-Lipschitz and convex over $\mathcal{K}$. Thus, as discussed above, under the condition $\eta_t \leq 2/\beta$, the steps of PNSGD$(S, w_0, \{B_t\}, \{\eta_t\}, \{\sigma_t\})$ are a contractive noisy iteration. Specifically, on the dataset S, the CNI is defined by the initial point w_0, sequence of functions $g_t(w) \doteq \Pi_{\mathcal{K}}(w) - \eta_t\nabla F_t(\Pi_{\mathcal{K}}(w))$ and sequence of noise distributions $\mathcal{D}_t = \mathcal{N}(0, (\eta_t\sigma_t)^2\mathbb{I}_d)$. Similarly, on the dataset S', the CNI is defined in the same way with the exception of $g'_t(w) \doteq \Pi_{\mathcal{K}}(w) - \eta_t\nabla F'_t(\Pi_{\mathcal{K}}(w))$, where F'_t includes loss function for x'_k instead of x_k. Namely, $F'_t(w) = F'_t(w) + (f(w, x'_k) - f(w, x_k))/B_t$.

By our assumption, $f(w, x)$ is L-Lipschitz for every $x \in \mathcal{X}$ and $w \in \mathcal{K}$ and therefore

$$\sup_w \|g_t(w) - g'_t(w)\|_2 = \frac{\eta_t}{B_t} \sup_w \|\nabla f(\Pi_{\mathcal{K}}(w), x_k) - \nabla f(\Pi_{\mathcal{K}}(w), x'_k)\|_2 \leq \frac{2\eta_t L}{B_t}.$$

We can now apply Corollary 2.14 with $\gamma = \alpha/2$. Note that $s_t \leq \frac{2\eta_t L}{B_t}$ and thus we obtain that

$$\mathrm{D}_\alpha\left(X_n \,\|\, X'_n\right) \leq \frac{\alpha}{2} \cdot \frac{4L^2\eta_t^2}{B_t^2} \cdot \frac{1}{\sum_{s=t}^T \eta_s^2\sigma_s^2} = \frac{2\alpha L^2\eta_t^2}{B_t^2 \cdot \sum_{s=t}^T \eta_s^2\sigma_s^2}.$$

Maximizing this expression over all indices $i \in [n]$ gives the claim. □

The important property of this analysis is that it allows for batch size to be used to improve the privacy guarantees. The specific batch size choice depends on the step sizes and noise rates. Next we describe the setting of these parameters that ensures convergence at the optimal rate.

3.2 Utility Guarantees for the Last Iterate of SGD

In order to analyze the performance of the noisy projected gradient descent algorithm we will use the convergence guarantees for the last iterate of SGD given in [SZ13, JNN19]. For the purpose of these results we let $F(w)$ be an arbitrary convex function over $\mathcal{K}$ for which we are given an unbiased stochastic (sub-)gradient oracle G. That is for every $w \in \mathcal{K}$, $\mathbb{E}[G(w)] \in \partial F(w)$. Let $\mathrm{PSGD}(G, w_0, \{\eta_t\}_{t\in[T]})$ denote the execution of the following process: starting from point w_0, use the update $w_{t+1} \doteq \Pi_{\mathcal{K}}(w_t + \eta_{t+1} G(w_t))$ for $t = 0, \ldots, T-1$. Shamir and Zhang [SZ13] prove that the suboptimality of the last iterate of SGD with the learning rate η_t being proportional to $1/\sqrt{t}$ scales as $(\log T)/\sqrt{T}$. This variant of SGD relies on relatively large step sizes in the early iterates which would translate into a relatively strong assumption on smoothness in Theorem 3.1. However, it is known [Har19] that the analysis in [SZ13] also applies to the fixed learning rate η_t scaling as $1/\sqrt{T}$ (in fact, it is simpler and gives a slightly better constants in this case).

Theorem 3.2 ([SZ13])**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex body of diameter D, let $F(w)$ be an arbitrary convex function over $\mathcal{K}$ and let G be an unbiased stochastic (sub-)gradient oracle G for F. Assume that for every $w \in \mathcal{K}$, $\mathbb{E}[\|G(w)\|_2^2] \leq L_G^2$. For $T \in \mathbb{N}$ and $w_0 \in \mathcal{K}$, let $w_1, \ldots, w_T$ denote the iterates produced by PSGD$(G, w_0, \{D/(L_G\sqrt{T})\})$. Then*

$$\mathbb{E}[F(w_T)] \leq F^* + \frac{DL_G(2 + \ln T)}{\sqrt{T}},$$

where $F^* \doteq \min_{w\in\mathcal{K}} F(w)$ and the expectation is taken over the randomness of G.

Further, Jain et al. [JNN19] show that the $\ln T$ factor can be eliminated by using faster decaying rates. Their step-size schedule is defined as follows.

Definition 3.3. *For an integer T, let $\ell = \lceil \log_2 T \rceil$. For $0 \leq i \leq \ell$, let $T_i = T - \lceil T \cdot 2^{-i} \rceil$ and let $T_{\ell+1} = T$. For a constant c, every $0 \leq i \leq \ell$ and $T_i < t \leq T_{i+1}$, we define $\eta_t = \frac{c2^{-i}}{\sqrt{T}}$. We denote the resulting sequence of step sizes by $\bar{\eta}_{\mathrm{JNN}}(c)$.*

Jain et al. [JNN19] prove that the following guarantees hold for SGD with step sizes given by $\bar{\eta}_{\mathrm{JNN}}(c)$.

Theorem 3.4 ([JNN19])**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex body of diameter D, let $F(w)$ be an arbitrary convex function over $\mathcal{K}$ and let G be an unbiased stochastic (sub-)gradient oracle G for F. Assume that for every $w \in \mathcal{K}$, $\mathbb{E}[\|G(w)\|_2^2] \leq L_G^2$. For $T \in \mathbb{N}$ and $w_0 \in \mathcal{K}$, let $w_1, \ldots, w_T$ denote the iterates produced by PSGD$(G, w_0, \bar{\eta}_{\mathrm{JNN}}(D/L_G))$. Then*

$$\mathbb{E}[F(w_T)] \leq F^* + \frac{15DL_G}{\sqrt{T}},$$

where $F^* \doteq \min_{w\in\mathcal{K}} F(w)$ and the expectation is taken over the randomness of G.

We remark that the results in [JNN19] are stated for an oracle G that gives (sub)-gradients bounded by G_L almost surely. This condition is necessary for the high-probability version of their result but a bound on the variance of G suffices to upper bound $\mathbb{E}[F(w_T)]$. In addition, while the results are stated for a fixed gradient oracle, the same results hold when a different stochastic gradient oracle G_t is used in step t as long as all the oracles satisfy the assumptions (namely, $\mathbb{E}[G_t(w)] \in \partial F(w)$ and $\mathbb{E}[\|G_t(w)\|_2^2] \leq L_G^2$ for all t).

3.3 Batch Size Analysis

Finally we derive the privacy and utility guarantees for noisy SGD by calculating the batch sizes needed to ensure the privacy guarantees for the settings in Theorems 3.2 and 3.4. The sum of batch sizes in turn gives us the number of samples n necessary to implement T steps of these algorithms.

Theorem 3.5. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set of diameter D and $\{f(\cdot, x)\}_{x\in\mathcal{X}}$ be a family of convex L-Lipschitz and β-smooth functions over $\mathcal{K}$. For $T \in \mathbb{N}$, $\rho > 0$, and all $t \in [T]$ let $B_t = \lceil 2\sqrt{d/(T-t+1)}/\rho \rceil$, $n = \sum_{t\in[T]} B_t$, $\eta = D/(L\sqrt{2T})$, $\sigma = L/\sqrt{d}$, If $\eta \leq 2/\beta$ then for all $\alpha \geq 1$, starting point $w_0 \in \mathcal{K}$, and $S \in \mathcal{X}^n$, PNSGD$(S, w_0, \{B_t\}, \{\eta\}, \{\sigma\})$ satisfies $\left(\alpha, \alpha \cdot \rho^2/2\right)$-RDP. Further, if S consists of samples drawn i.i.d. from a distribution $\mathcal{P}$, then $n \leq T + 4\sqrt{dT}/\rho$ and*

$$\mathbb{E}[F(w_T)] \leq F^* + \frac{\sqrt{8}DL(2 + \ln T)}{\sqrt{T}} \leq F^* + \sqrt{32}DL \cdot \ln(10n) \cdot \left(\frac{1}{\sqrt{n}} + \frac{2\sqrt{d}}{\rho n}\right),$$

where, for all $w \in \mathcal{K}$, $F(w) \doteq \mathbb{E}_{x\sim\mathcal{P}}[f(w, x)]$, $F^* \doteq \min_{w\in\mathcal{K}} F(w)$ and the expectation is taken over the random choice of S and noise added by PNSGD.

Proof. We first establish the privacy guarantees. By Theorem 3.1, all we need is to verify that for our choice of $\{B_t\}$, σ and η we have for every $t \in [T]$,

$$4L^2 \cdot \max_{t\in[T]} \left\{ \frac{\eta^2}{B_t^2 \cdot \sum_{s=t}^{T} \eta^2\sigma^2} \right\} = 2L^2 \max_{t\in[T]} \left\{ \frac{1}{B_t^2 \cdot (T-t+1)L^2/d} \right\} \leq \rho^2.$$

This implies that

$$n = \sum_{t\in[T]} \left\lceil \sqrt{\frac{4d}{\rho^2(T-t+1)}} \right\rceil \leq \sum_{t\in[T]} \sqrt{\frac{4d}{\rho^2(T-t+1)}} + 1 = T + \frac{2\sqrt{d}}{\rho} \sum_{t\in[T]} \frac{1}{\sqrt{t}} \leq T + \frac{4\sqrt{dT}}{\rho},$$

where we used the fact that $\sum_{t\in[T]} \frac{1}{\sqrt{t}} \leq 2(\sqrt{T+1} - 1) + 1 \leq 2\sqrt{T}$.

To establish the utility guarantees, we first note that for all $t \in [T]$,

$$\nabla F_t(w) = \frac{1}{B_t} \sum_{i=1+\sum_{s\leq t-1} B_s}^{i=\sum_{s\leq t} B_s} \nabla_w f(w, x_i).$$

Thus for S sampled i.i.d. from $\mathcal{P}$ and index i in batch t, $\mathbb{E}[\nabla_w f(w, x_i)] = \nabla F(w)$. In particular, for $\xi_t \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}[\nabla F_t(w) + \xi_t] = \nabla F(w)$ and therefore each $\nabla F_t(w) + \xi_t$ gives an independent sample from a stochastic gradient oracle for F. Our setting of the noise scale $\sigma = L/\sqrt{d}$ ensures that for every $t \in [T]$

$$\mathbb{E}_{S\sim\mathcal{P}^n, \xi_t\sim\mathcal{N}(0,\sigma^2)} \left[\|\nabla F_t(w) + \xi_t\|_2^2 \right] = \frac{\mathbb{E}_{x\sim\mathcal{P}} \left[\|\nabla_w f(w, x)\|_2^2 \right]}{B_t} + d\sigma^2 \leq \frac{L^2}{B_t} + L^2 \leq 2L^2.$$

This implies that for our choice of parameters PNSGD$(S, w_0, \{B_t\}, \{\eta\}, \{\sigma\})$ can be seen as an execution PSGD$(G, w_0, \{D/(L_G\sqrt{T})\})$ with stochastic gradient oracles with variance upper-bounded by $L_G^2 = 2L^2$. Plugging this value in Theorem 3.2 gives our bound on the utility of the algorithm. To obtain the bound in terms of n we note that $n \leq T + \frac{4\sqrt{dT}}{\rho}$, implies that $T \geq \frac{n^2}{16d/\rho^2 + 4n}$ and thus

$$\frac{1}{\sqrt{T}} \leq \frac{2}{\sqrt{n}} + \frac{4\sqrt{d}}{\rho n}. \qquad \square$$

Next, we give a differentially private version of the step-size schedule from [JNN19].

Theorem 3.6. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set of diameter D and $\{f(\cdot, x)\}_{x\in\mathcal{X}}$ be a family of convex L-Lipschitz and β-smooth functions over $\mathcal{K}$. For $T \in \mathbb{N}$, $\rho > 0$, and all $t \in [T]$ let $B_t = \lceil 4\sqrt{3d/(T-t+1)}/\rho \rceil$, $n = \sum_{t\in[T]} B_t$, $\{\eta_t\} = \bar{\eta}_{\text{JNN}}(D/(\sqrt{2}L))$, $\sigma = L/\sqrt{d}$, If $\eta_1 \leq 2/\beta$ then for all $\alpha \geq 1$, starting point $w_0 \in \mathcal{K}$, and $S \in \mathcal{X}^n$, PNSGD$(S, w_0, \{B_t\}, \{\eta_t\}, \{\sigma\})$ satisfies $\left(\alpha, \alpha \cdot \rho^2/2\right)$-RDP. Further, if S consists of samples drawn i.i.d. from a distribution $\mathcal{P}$, then $n \leq T + 4\sqrt{dT}/\rho$ and*

$$\mathbb{E}[F(w_T)] \leq F^* + \frac{15\sqrt{2}DL}{\sqrt{T}} \leq 30\sqrt{2}DL \cdot \left(\frac{1}{\sqrt{n}} + \frac{4\sqrt{3d}}{\rho n}\right),$$

where, for all $w \in \mathcal{K}$, $F(w) \doteq \mathbb{E}_{x\sim\mathcal{P}}[f(w, x)]$, $F^* \doteq \min_{w\in\mathcal{K}} F(w)$ and the expectation is taken over the random choice of S and noise added by PNSGD.

Proof. The utility guarantees for this algorithm follow from the same argument as in the proof of Theorem 3.5 together with Theorem 3.4. As before, by Theorem 3.1, all we need to establish the privacy guarantees is to verify that for our choice of $\{B_t\}, \sigma$ and $\{\eta_t\}$ we have for every $t \in [T]$,

$$4L^2 \cdot \max_{t\in[T]} \left\{ \frac{\eta_t^2}{B_t^2 \cdot \sum_{s=t}^T \eta_s^2 \sigma^2} \right\} \leq \rho^2. \tag{1}$$

We first observe that for $t = T$ we have that

$$\frac{\eta_t^2}{B_t^2 \cdot \sum_{s=t}^T \eta_s^2 \sigma^2} = \frac{d}{B_T^2 L^2} \leq \frac{\rho^2}{48L^2}. \tag{2}$$

For $t \in [T-1]$, let i be such that $T_i < t \leq T_{i+1}$. Then we note that for $c = D/(\sqrt{2}L)$

$$\eta_t = c\frac{2^{-i}}{\sqrt{T}} \geq c\frac{\lceil T2^{-i} \rceil - 1}{T\sqrt{T}} \geq c\frac{T - T_i - 1}{T^{3/2}} \geq c\frac{T-t}{T^{3/2}}.$$

and therefore

$$\sum_{s=t}^T \eta_s^2 \geq c^2 \sum_{s=t}^T \frac{(T-s)^2}{T^3} \geq \frac{c^2}{T^3} \cdot \frac{(T-t)^2(T-t+1)}{3}.$$

In addition, using the fact that for $t \leq T-1$, $i \leq \ell - 1$ we have that

$$\eta_t = 2c\frac{2^{-i-1}}{\sqrt{T}} \leq 2c\frac{\lceil T2^{-i-1} \rceil}{T\sqrt{T}} = 2c\frac{T - T_{i+1}}{T^{3/2}} \leq 2c\frac{T-t}{T^{3/2}}.$$

Thus

$$\frac{\eta_t^2}{B_t^2 \cdot \sum_{s=t}^T \eta_s^2 \sigma^2} = \frac{1}{B_t^2 \sigma^2} \cdot \frac{4c^2(T-t)^2}{T^3} \cdot \frac{3T^3}{c^2(T-t)^2(T-t+1)} = \frac{12d}{B_t^2 L^2 (T-t+1)} \leq \frac{\rho^2}{4L^2}.$$

Plugging this and eq. (2) into eq.(1) we obtain that the privacy condition holds.

As in the proof of Theorem 3.6, we obtain that

$$n = \sum_{t\in[T]} B_t \leq T + \frac{8\sqrt{3}\sqrt{dT}}{\rho}$$

and thus $T \geq \frac{n^2}{192d/\rho^2+4n}$. This means that

$$\frac{1}{\sqrt{T}} \leq \frac{2}{\sqrt{n}} + \frac{8\sqrt{3d}}{\rho n},$$

implying the claimed bound on utility in terms of n. □

As a corollary we get the proof of our main claim.

Corollary 3.7 (Thm. 1.1 restated)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set of diameter D and $\{f(\cdot, x)\}_{x\in\mathcal{X}}$ be a family of convex L-Lipschitz and $(2D\sqrt{2T}/L)$-smooth functions over $\mathcal{K}$. For every $\rho > 0$, there exists an algorithm $\mathcal{A}$ that given a starting point $w_0 \in \mathcal{K}$, and $S \in \mathcal{X}^n$ returns a point $\hat{w}$. For all $\alpha \geq 1$, $\mathcal{A}$ satisfies $(\alpha, \alpha \cdot \rho^2/2)$-RDP and uses n evaluations of the gradient of $f(w, x)$. Further, if S consists of samples drawn i.i.d. from a distribution $\mathcal{P}$ over $\mathcal{X}$, then*

$$\mathbb{E}[F(\hat{w})] \leq F^* + O\left(DL \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\rho n}\right)\right),$$

where, for all $w \in \mathcal{K}$, $F(w) \doteq \mathbb{E}_{x\sim\mathcal{P}}[f(w, x)]$, $F^* \doteq \min_{w\in\mathcal{K}} F(w)$ and the expectation is taken over the random choice of S and randomness of $\mathcal{A}$.

4 Localization-based algorithms

In this section, we describe the Iterative Localization framework, and give two instantiations of it. Our localization algorithm will be based on adding Gaussian noise to an algorithm whose output has low L_2-sensitivity (also referred to as uniform stability of the parameter). We first briefly recall the relevant definitions and the resulting privacy guarantees.

Definition 4.1. *A deterministic algorithm (or function) $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ has L_2-sensitivity of γ if for all pairs of datasets $S, S' \in \mathcal{X}^n$ that differ in a single element we have that $\|\mathcal{A}(S) - \mathcal{A}(S')\|_2 \leq \gamma$.*

The well-known property of the Gaussian mechanism is that it can convert any algorithm with bounded L_2-sensitivity to a differentially private one.

Lemma 4.2. *Let $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ be a deterministic function with L_2-sensitivity γ. Then for any $\rho > 0$, an algorithm that outputs $\mathcal{A}(S) + \xi$ where $\xi \sim \mathcal{N}(0, \frac{\gamma^2}{\rho^2}\mathbb{I}_d)$ satisfies $(\alpha, \frac{1}{2}\alpha\rho^2)$-RDP for all $\alpha \geq 1$.*

Suppose we have an algorithm A that given a point $w \in B(w^*, D)$ and a sequence of samples from F, outputs a $w' \in B(w^*, D/4)$. We will want this A to have small sensitivity, and small suboptimality, both of which scale, say, linearly with D. Given such an A, we can iteratively invoke it with geometrically decreasing D, adding noise at the end of each phase to ensure privacy. Crucially, once the diameter bound D becomes small enough, the noise added is small enough that the suboptimality due the added noise is negligible. This would allow us to incur only a logarithmic overhead in terms of sample complexity, while ensuring privacy and good utility bounds.

We next describe two instantiations of this Iterative Localization framework. To get better bounds, we only bound the second moment of the distance $\|w' - w^*\|$, instead of requiring that $\|w' - w^*\|$ is uniformly bounded. The first instantiation uses SGD as algorithm A, and applies to convex functions. The sensitivity bound here comes from bounding the learning rate and holds under mild smoothness assumptions. The second instantiation will apply to arbitrary convex functions, and optimizes a regularized objective to ensure a sensitivity bound.

4.1 SGD-based Iterative Localization

Our algorithm is based on a sequence of phases such that each phase (implicitly) localizes an approximate minimizer of the population loss. Specifically, given a point w_i such that for some w_i^*, $\mathbb{E}[\|w_i - w_i^*\|_2^2] \le D$, the algorithm outputs a point w_{i+1} such that for some point w_{i+1}^*, $\mathbb{E}[\|w_{i+1} - w_{i+1}^*\|_2^2] \le D/4$ and, in addition, $\mathbb{E}[F(w_{i+1}^*)] - \mathbb{E}[F(w_i^*)] \le \tau/2^{-i}$ where τ is the desired excess loss.

Our algorithm relies on the fact that SGD on sufficiently smooth loss functions has low L_2-sensitivity [HRS15, FV19].

Lemma 4.3. *Each iterate of one-pass online projected gradient descent with fixed step size η over a sequence of β-smooth L-Lipschitz convex functions has L_2-sensitivity of at most $2L\eta$ as long as $\eta \le 2/\beta$. In particular, the same applies to the average of all the iterates.*

Algorithm 2 Phased-SGD algorithm

```
Input: Data set $S = \{x_1, \ldots, x_n\}$, convex $f\colon \mathcal{K} \times \mathcal{X} \to \mathbb{R}$, initial point $w_0 \in \mathcal{K}$, step size $\eta$, privacy
    parameter $\rho$.
set $k = \lceil \log_2 n \rceil$
for $i = 1, \ldots, k$ do
    set $n_i = 2^{-i} n$ and $\eta_i = 4^{-i}\eta$.
    initialize an PSGD algorithm (over domain $\mathcal{K}$) at $w_{i-1}$ and run with step size $\eta_i$ for $n_i$ steps; let $\overline{w}_i$ be
    the average iterate.
    set $w_i = \overline{w}_i + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma_i^2 \mathbb{I}_d)$ with $\sigma_i = 4L\eta_i/\rho$.
return the final iterate $w_k$.
```

Theorem 4.4. *Assume that $\|w_0 - w^*\|_2 \le D$ (this is the case, for example, when $\mathcal{K}$ has diameter at most D), and set*

$$\eta = \frac{D}{L} \min\left\{ \frac{4}{\sqrt{n}}, \frac{\rho}{\sqrt{d}} \right\}.$$

Then for the output of Algorithm 2, we have

$$\mathbb{E}[F(w_k)] - F(w^*) \le 10LD\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\rho n} \right),$$

provided that $\beta \le 1/\eta$.

To prove the theorem, we first provide utility and privacy guarantees for each individual phase of the algorithm.

Lemma 4.5. *Assume that $\eta_i \le 2/\beta$. Then for any $\alpha \ge 1$, the output w_i of phase i in Algorithm 2 satisfies $(\alpha, \alpha\rho^2/2)$-RDP, and for any $w \in \mathcal{K}$,*

$$\mathbb{E}[F(\overline{w}_i)] - F(w) \le \frac{\mathbb{E}[\|w_{i-1} - w\|_2^2]}{2\eta_i n_i} + \frac{\eta_i L^2}{2}. \tag{3}$$

Proof. The privacy guarantee follows from Lemma 4.2 together with the fact that PSGD (when viewed as a deterministic mapping from a data set to a final iterate) with step size $\eta_i \le 2/\beta$ has L_2-sensitivity bounded by $2L\eta_i$ (this is a consequence of Lemma 4.3). The utility guarantee follows from standard convergence bounds for PSGD. □

We can now prove Theorem 4.4.

Proof of Theorem 4.4. Denote $\overline{w}_0 = w^*$ and $\xi_0 = w_0 - w^*$; by assumption, $\|\xi_0\|_2 \leq D$. Using Lemma 4.5, the total error of the algorithm can be bounded by

$$\begin{aligned}\mathbb{E}[F(w_k)] - F(w^*) &= \sum_{i=1}^{k} \mathbb{E}[F(\overline{w}_i) - F(\overline{w}_{i-1})] + \mathbb{E}[F(w_k) - F(\overline{w}_k)] \\ &\leq \sum_{i=1}^{k} \left(\frac{\mathbb{E}[\|\xi_{i-1}\|_2^2]}{2\eta_i n_i} + \frac{\eta_i L^2}{2} \right) + L \cdot \mathbb{E}[\|\xi_k\|_2].\end{aligned}$$

Recall that by definition $\eta \leq (D/L) \cdot (\rho/\sqrt{d})$, so that for all $i \geq 0$,

$$\mathbb{E}[\|\xi_i\|_2^2] = d\sigma_i^2 = d(4^{-i} L\eta/\rho)^2 \leq (4^{-i} D)^2.$$

In particular, we have $\mathbb{E}[\|\xi_k\|_2] \leq \sqrt{\mathbb{E}[\|\xi_k\|_2^2]} = 4^{-k} D$. Hence,

$$\begin{aligned}\mathbb{E}[F(w_k)] - F(w^*) &\leq \sum_{i=1}^{k} 2^{-i} \left(\frac{8D^2}{\eta n} + \frac{\eta L^2}{2} \right) + 4^{-k} LD \\ &\leq \sum_{i=1}^{\infty} 2^{-i} LD \left(\frac{8}{n} \max\left\{ \sqrt{n}, \frac{\sqrt{d}}{\rho} \right\} + \frac{1}{2\sqrt{n}} \right) + \frac{LD}{n^2} \\ &\leq 9LD \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\rho n} \right) + \frac{LD}{n^2}.\end{aligned}$$

□

4.2 Non-smooth DP-SCO: Phased ERM

In this section, we demonstrates that the general approach based on localization can also be applied to the non-smooth case. We only require $f(\cdot, x)$ to be convex and L-Lipschitz for any $x \in \mathcal{X}$. Our algorithm is similar to the one in the previous section, except that we replace the PSGD subroutine in step 3 of the algorithm with a regularized ERM computation. The L_2 regularization in this case is the standard technique for ensuring low sensitivity that we require. In addition, low-sensitivity ensures uniform stability and thus generalization of the solution to the population. To get a more efficient algorithm, we use an approximate optimizer instead of an exact one. The suboptimality of this optimization should be small enough that the sensitivity of the resulting algorithm can still be controlled. To solve the regularized problem, we employ SGD that ensures the suboptimality bound, and hence the sensitivity bound, with high probability. To allow for a small failure probability of this approach, we will only give (ε, δ)-DP guarantees for the algorithm. We will use the following standard variant ofLemma 4.2:

Lemma 4.6. *Let $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ be a randomized function such that for all pairs of datasets $S, S' \in \mathcal{X}^n$ that differ in a single element there is a coupling such that $\|\mathcal{A}(S) - \mathcal{A}(S')\|_2 \leq \gamma$ except with probability δ. Then for any $\rho > 0$, an algorithm that outputs $\mathcal{A}(S) + \xi$ where $\xi \sim \mathcal{N}(0, \frac{\gamma^2 \ln \frac{1}{\delta}}{\varepsilon^2} \mathbb{I}_d)$ satisfies $(\varepsilon, 2\delta)$-DP.*

We first prove the relevant properties of the regularized ERM algorithm.

Algorithm 3 Phased-ERM algorithm

```
Input: Data set $S = \{x_1, \dots, x_n\}$, convex $f\colon \mathcal{K} \times \mathcal{X} \to \mathbb{R}$, initial point $w_0 \in \mathcal{K}$, step size $\eta$, privacy
    parameters $\varepsilon, \delta$.
set $k = \lceil \log_2 n \rceil$
for $i = 1, \dots, k$ do
    set $n_i = 2^{-i} n$ and $\eta_i = 4^{-i} \eta$.
    compute $\tilde{w}_i \in \mathcal{K}$ such that $F_i(\tilde{w}_i) - \arg\min_{w \in \mathcal{K}} F_i(w) \le L^2 \eta_i / n_i$ with prob. $(1 - \delta)$ for
        $F_i(w) = \frac{1}{n_i} \sum_{t=1}^{n_i} f(w, x_t) + \frac{1}{\eta_i n_i} \|w - w_{i-1}\|_2^2.$
    set $w_i = \tilde{w}_i + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma_i^2 \mathbb{I}_d)$ with $\sigma_i = 4L\eta_i \sqrt{\ln \frac{1}{\delta}} / \varepsilon$.
return the final iterate $w_k$.
```

Lemma 4.7. *The output w_i of phase i of Algorithm 3 satisfies $(\varepsilon, 2\delta)$-DP, and for any $w \in \mathcal{K}$,*

$$\mathbb{E}[F(\tilde{w}_i)] - F(w) \le \frac{\mathbb{E}[\|w_{i-1} - w\|_2^2]}{\eta_i n_i} + 3\eta_i L^2. \tag{4}$$

Further, $\tilde{w}_i$ can be found with $O(n_i^2 \sqrt{\ln \frac{1}{\delta}})$ gradient calls (to gradients of f).

Proof. The objective F_i minimized on phase i is L-Lipschitz and λ_i-strongly convex for $\lambda_i = 2/(\eta_i n_i)$; denote by $\overline{w}_i \in \mathcal{K}$ its minimizer. From the results in [BE02, SSSSS09] we know that the minimizer $\overline{w}_i$ has L_2 sensitivity bounded by $4L/(\lambda_i n_i) = 2L\eta_i$, and furthermore, that for any $w \in \mathcal{K}$,

$$\mathbb{E}[F(\overline{w}_i)] - F(w) \le \frac{\lambda_i}{2n_i} \mathbb{E}[\|w_{i-1} - w\|_2^2] + \frac{4L^2}{\lambda_i n_i} = \frac{\mathbb{E}[\|w_{i-1} - w\|_2^2]}{\eta_i n_i} + 2L^2 \eta_i.$$

For the approximate minimizer $\tilde{w}_i$, we have by strong convexity that except with probability δ

$$\frac{\lambda_i}{2} \|\tilde{w}_i - \overline{w}_i\|^2 \le F_i(\tilde{w}_i) - F_i(\overline{w}_i) \le \frac{L^2 \eta_i}{n_i},$$

which implies that $\|\tilde{w}_i - \overline{w}_i\| \le L\eta_i$. In particular, $\tilde{w}_i$ has sensitivity of at most $4L\eta_i$, which gives the privacy guarantee via Lemma 4.6. Finally, for any $w \in \mathcal{K}$ we have

$$\mathbb{E}[F(\tilde{w}_i)] - F(w) = \mathbb{E}[F(\overline{w}_i) - F(w)] + \mathbb{E}[F(\tilde{w}_i) - F(\overline{w}_i)] \le \frac{\mathbb{E}[\|w_{i-1} - w\|_2^2]}{\eta_i n_i} + 3L^2 \eta_i,$$

which implies the claim on utility. Finally, to see the running time statement, recall that for optimizing an L-Lipschitz and λ-strongly convex function to within Δ accuracy with probability $\ge 1 - \delta$ using SGD, one needs $\widetilde{O}((L^2/\lambda\Delta)\sqrt{\log(1/\delta)})$ stochastic gradient computations (e.g., [HK14]). Hence, the number of gradient calls needed for computing $\tilde{w}_i$, being an Δ_i-approximate minimizer of a λ_i-strongly convex function for $\Delta_i = L^2 \eta_i / n_i$ and $\lambda_i = 1/(\eta_i n_i)$, is $O((L^2/\lambda_i \Delta_i)\sqrt{\log(1/\delta)}) = O(n_i^2 \sqrt{\log(1/\delta)})$. □

The proof of the following result is identical to that Theorem 4.4, with Lemma 4.7 replacing Lemma 4.5.

Theorem 4.8. *Assume that $\|w_0 - w^*\|_2 \le D$, and set*

$$\eta = \frac{D}{L} \min\left\{ \frac{4}{\sqrt{n}}, \frac{\varepsilon}{\sqrt{d \ln(1/\delta)}} \right\}.$$

Then for the output of Algorithm 2, we have

$$\mathbb{E}[F(w_k)] - F(w^*) = O\left(LD\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n} \right) \right).$$

Further, a version of this algorithm can be implemented with $\widetilde{O}(n^2)$ stochastic gradient computations.

4.3 The strongly convex case

Suppose that the population loss of interest F is λ-strongly convex and L-Lipschitz over the domain $\mathcal{K}$. In this case, the optimal statistical rate is $O(L^2/\lambda n)$ [HK14], and the private ERM can be optimized with an error of $\tilde{O}(dL^2/\varepsilon^2 \lambda n^2)$. The best known bound for Private Stochastic Convex Optimization for this case is due to [BST14] who give an upper bound of $\tilde{O}(L^2\sqrt{d}/\lambda\varepsilon n)$. As in the convex case, we show that the optimal rate is in fact the larger of the two lower bounds, and is attained by a linear-time algorithm.

We will show this by a folklore reduction to the convex case (see, e.g., [HK14] for a similar instantiation of this reduction). Assume a private stochastic (non-strongly) convex optimization algorithm $\mathcal{A}$ with the following utility guarantee when initialized at $w_0 \in \mathcal{K}$:

$$\mathbb{E}[F(w_{\mathcal{A}})] - F(w^*) \le cLD\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\rho n} \right)$$

for some universal constant $c > 0$, where $D > 0$ is such that $\|w_0 - w^*\|_2 \le D$. (E.g., this can be one of Algorithms 1 to 3 under their respective assumptions and settings of ρ.) Consider the following algorithm: starting from a given $w_0 \in \mathcal{K}$, repeat the private optimization algorithm for $k = \lceil 2 \log\log(n/c) \rceil$ times, where each run is initialized at the output of the previous phase and is run for $n_i = \lceil n/k \rceil$ iterations. We prove the following:

Theorem 4.9. *The algorithm described above is private (with the same privacy parameters as those of $\mathcal{A}$) and outputs a solution whose expected population risk is at most*

$$O\left(\frac{L^2}{\lambda} \left(\frac{1}{n} + \frac{d}{\rho^2 n^2} \right) \log^2 \log \frac{n}{c} \right).$$

Up to the $O(\log\log(n))$ overhead, we obtain the optimal rate. This overhead can be easily removed by using a more careful increasing sequence of n_i's.

Under β-smoothness assumptions and when the condition number is $\beta/\lambda = O(\max\{\sqrt{n}, \sqrt{d}/\rho\})$, the inner stochastic convex optimization problem is sufficiently smooth so that we can use Algorithm 2 as the basic private optimization algorithm and get a linear time algorithm for stochastic strongly convex optimization. Without any smoothness assumptions, we can invoke the reduction with Algorithm 3 and get a quadratic-time algorithm with the optimal rate.

Proof of Theorem 4.9. Denote the output of phase i by w_i. Let $\Delta_i = \mathbb{E}[F(w_i)] - F(w^*)$ be the expected suboptimality after phase i, and let $D_i^2 = \mathbb{E}\|w_i - w^*\|^2$ for all i. The λ-strong convexity of F implies

$\frac{1}{2}\lambda D_i^2 \leq \Delta_i$ for all $i \geq 0$, and in particular, $\Delta_0 \leq LD_0 \leq \sqrt{2\Delta_1 L^2/\lambda}$ or equivalently, $\Delta_0 \leq 2L^2/\lambda$. Thus, by the guarantee of the private convex optimization algorithm, we have for all i that

$$\Delta_{i+1} \leq cLD_i \left(\frac{1}{\sqrt{n_i}} + \frac{\sqrt{d}}{\rho n_i} \right) \leq cL\sqrt{\frac{2\Delta_i}{\lambda}} \left(\frac{1}{\sqrt{n_i}} + \frac{\sqrt{d}}{\rho n_i} \right).$$

Let us denote by E_i the expression $c^2(2L^2/\lambda)\big(\frac{1}{\sqrt{n_i}} + \frac{\sqrt{d}}{\rho n_i}\big)^2$. Since $E_i/E_{i+1} \leq 4$, the above inequality can be rearranged as

$$\forall\, i \geq 0, \qquad \frac{\Delta_{i+1}}{16E_{i+1}} \leq \frac{\sqrt{\Delta_i E_i}}{16E_{i+1}} = \frac{E_i}{16E_{i+1}}\sqrt{\frac{\Delta_i}{E_i}} \leq \sqrt{\frac{\Delta_i}{16E_i}}.$$

This implies that for $k \geq k_0 = \lceil \log\log(\Delta_0/E_0) \rceil$, it holds that $\Delta_k \leq 2E_k$. Thus, after $k \leq 2\log\log(n/c)$ phases, we hold a solution with error

$$\mathbb{E}[F(w_k)] - F(w^*) \leq \frac{8c^2L^2}{\lambda}\left(\frac{1}{n} + \frac{d}{\rho^2 n^2}\right). \qquad \square$$

5 No Privacy Amplification by Averaged Iteration

A common technique in convex optimization is to use iterate averaging. A plausible conjecture is that the average of the iterates enjoys privacy properties similar to the last iterate. Indeed in a Contractive Noisy Iteration with uniform noise, the privacy for the last iterate and that for the average iterate are within constant factors of each other when the contractive map is the identity.

We show that this does not hold true in general. Consider the contractive noise process defined by contractive maps:

$$\phi_i(\mathbf{x}) = \begin{cases} \mathbf{x} & \text{if } i \leq k \\ \mathbf{0} & \text{otherwise} \end{cases}$$

Here k is a parameter we will set appropriately. Thus the contractive noise process is

$$X_{t+1} = \begin{cases} X_t + \mathcal{N}(\mathbf{0}, \sigma^2) & \text{if } t \leq k \\ \mathcal{N}(\mathbf{0}, \sigma^2) & \text{otherwise} \end{cases}$$

The sum of X_t's thus is easily seen to be distributed as:

$$kX_0 + \sum_{i \leq k}(k - i + 1)\mathcal{N}(\mathbf{0}, \sigma^2) + \sum_{i=k+1}^{T} \mathcal{N}(\mathbf{0}, \sigma^2).$$

Simplifying, the average iterate is distributed as:

$$\frac{k}{T} \cdot X_0 + \mathcal{N}(\mathbf{0}, \frac{O(k^3) + (T-k)}{T^2}\sigma^2).$$

For $\sigma = \frac{1}{\sqrt{T}}$, where the final iterate has $(\alpha, O(\alpha))$-RDP, this simplifies to

$$\frac{k}{T}X_0 + \mathcal{N}(\mathbf{0}, \frac{O(k^3) + (T-k)}{T^2}\sigma^2) = \frac{k}{T}\left(X_0 + \mathcal{N}(\mathbf{0}, O(\frac{k}{T}) + \frac{(T-k)}{Tk^2})\right)$$

While for $k = 1$ and for $k = T$, this amount of noise gives $(\alpha, O(\alpha))$-RDP, for intermediate values of k, e.g. $k \in [T^{\frac{1}{3}}, T^{\frac{2}{3}}]$, the effective amount of noise is not sufficient to mask X_0.

A similar lower bound can be realized for online convex optimization. Consider the sequence of loss functions over $\mathbb{R}$ defined as:

$$\ell_t(w) = \begin{cases} (w-b)^2 & t = 1 \\ 0 & 2 \leq t \leq k \\ w^2 & k+1 \leq t \leq T \end{cases}$$

Here k is a parameter to be set appropriately, and $b \in \{-1, 1\}$. Suppose that that learning rate $\eta = \frac{1}{\sqrt{T}}$ and the noise scale at each step is $\frac{1}{\sqrt{T}}$. If the noise added to the gradient at step t is ξ_t, then one can verify that the average iterate is

$$\frac{k}{T}(b\eta) + \sum_{t \leq k}(O(\eta) + (k-t))\eta\xi_t + \sum_{t=k+1}^{T} O(\eta)\eta\xi_t.$$

In other words, the average iterate is distributed as

$$\frac{k\eta}{T}\left(b + \mathcal{N}(0, O(\frac{k}{T}) + \frac{(T-k)}{Tk^2})\right)$$

This is the same behaviour as in the counterexample above. Thus the average is not $(\alpha, O(\alpha)) - RDP$. This example can be easily modified to handle suffix averaging over a $\Omega(T)$-sized suffix.

References

[ACG+16] Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 308–318, 2016.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2002.

[BFTT19] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. *CoRR*, abs/1908.09970, 2019. Extended abstract in Proceedings of NeurIPS 2019.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography—14th International Conference, TCC 2016-B, Part I*, pages 635–658, 2016.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 464–473. IEEE, 2014.

[CM08] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *NIPS*. MIT Press, 2008.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

[DJW13] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 429–438, 2013.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.

[DR16] Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016.

[Dwo06] Cynthia Dwork. Differential privacy. In *ICALP*, 2006.

[Fel16] Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems*, pages 3576–3584, 2016.

[FMTT18] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. *CoRR*, abs/1808, 2018. Extended abstract in Proceedings of FOCS 2018.

[FV19] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *arXiv preprint arXiv:1902.10710*, 2019.

[Har19] Nicholas Harvey. Personal communication, 2019.

[HK14] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[HLPR19] Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *COLT*, pages 1579–1613, 2019.

[HRS15] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

[INS+19] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *IEEE S and P (Oakland)*, 2019.

[JKT12] Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *25th Annual Conference on Learning Theory (COLT)*, pages 24.1–24.34, 2012.

[JNN19] Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. In *COLT*, pages 1752–1755, 2019.

[JT14] Prateek Jain and Abhradeep Thakurta. (near) dimension independent risk bounds for differentially private learning. In *ICML*, 2014.

[KST12] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1, 2012.

[Mir17] Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Optimization. A Basic Course.* Springer US, 2004.

[PAE$^+$17] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[PSM$^+$18] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with PATE. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[Rén61] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.

[SCS13] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, 2013.

[SSSSS09] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic Convex Optimization. In *COLT*, 2009.

[SSSSS10] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *JMLR*, 2010.

[ST13] Adam Smith and Abhradeep Thakurta. Differentially private feature selection via stability arguments, and the robustness of the LASSO. In *Conference on Learning Theory (COLT)*, pages 819–850, 2013.

[STU17] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *IEEE Security & Privacy*, pages 58–77, 2017.

[SZ13] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *ICML*, pages 71–79, 2013.

[TTZ15] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Nearly optimal private LASSO. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 2, pages 3025–3033, 2015.

[Ull15] Jonathan Ullman. Private multiplicative weights beyond linear queries. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 303–312. ACM, 2015.

[WLK$^+$17] Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *SIGMOD*. ACM, 2017.

[WYX17] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.