

Privacy Amplification by Iteration

Vitaly Feldman*, Ilya Mironov*, Kunal Talwar* and Abhradeep Thakurta*[†]

*Google Brain

Email: {vitalyfm, mironov, kunal}@google.com.

[†]UC Santa Cruz

Email: aguhatha@ucsc.edu.

Abstract—Many commonly used learning algorithms work by iteratively updating an intermediate solution using one or a few data points in each iteration. Analysis of differential privacy for such algorithms often involves ensuring privacy of each step and then reasoning about the cumulative privacy cost of the algorithm. This is enabled by composition theorems for differential privacy that allow releasing of all the intermediate results. In this work, we demonstrate that for contractive iterations, not releasing the intermediate results strongly amplifies the privacy guarantees.

We describe several applications of this new analysis technique to solving convex optimization problems via noisy stochastic gradient descent. For example, we demonstrate that a relatively small number of non-private data points from the same distribution can be used to close the gap between private and non-private convex optimization. In addition, we demonstrate that we can achieve guarantees similar to those obtainable using the privacy-amplification-by-sampling technique in several natural settings where that technique cannot be applied.

I. INTRODUCTION

Differential privacy [2] is a standard concept for capturing privacy of statistical algorithms. In its original formulation, (pure) differential privacy is parameterized by a single real number—the so-called privacy budget—which characterizes the privacy loss of an individual contributor to the input dataset.

As applications of differential privacy start to proliferate, they bring to the fore the problem of administering the privacy budget, with specific emphasis on *privacy composition* and *privacy amplification*.

Privacy composition enables modular design and analysis of complex and heterogeneous algorithms from simpler building blocks by controlling the total privacy budget of their combination. Improving on “naïve” composition, which simply (but very consequentially!) states that the privacy budgets of composition blocks sum up, “advanced” composition theorems allow subadditive accumulation of the privacy budgets. All existing proofs of advanced composition theorems assume that all intermediate outputs are revealed, whether the composite mechanism requires it or not.

Privacy amplification goes even further by bounding the privacy budget—for select mechanisms—of a combination

to be *less* than the privacy budget of its parts. The only systematically studied instance of this phenomenon is *privacy amplification by sampling* [3]–[8]. In its basic form, for $\varepsilon = O(1)$, an ε -differentially private mechanism applied to a secretly sampled p fraction of the input satisfies $O(p\varepsilon)$ -differential privacy. More recent results demonstrate that privacy can be amplified in proportion to p^2 (for a Gaussian additive noise mechanism and appropriate relaxations of differential privacy).

This work introduces a new amplification argument—*amplification by iteration*—that in certain contexts can be seen as an alternative to privacy amplification by sampling. As an exemplar of the kind of algorithms we wish to analyze, we consider noisy stochastic gradient descent for a smooth and convex objective.

Our preferred privacy notion for formally stating our contributions is Rényi differential privacy (RDP). For the purpose of this introduction, it suffices to keep in mind that RDP is parameterized with $1 < \alpha \leq \infty$ and measures the Rényi divergence of order α (denoted D_α) between the output distributions of a randomized algorithm on two neighboring datasets. It is a relaxation of (pure) differential privacy which has been instrumental for achieving tighter bounds on privacy cost in a number of recent papers on privacy-preserving machine learning. In addition, to being a privacy definition in its own right, one can easily translate RDP bounds to usual (ε, δ) -DP bounds.

Our first contribution is a general theorem that states that, under certain conditions on an iterative process, the process shrinks the Rényi divergence between distributions. We will focus on the simplest form of these conditions in which the mechanism is a composition of a sequence of *contractive* (or 1-Lipschitz) maps and an additive Gaussian noise mechanism. This is a natural setting for several differentially private optimization algorithms. A more general treatment that allows other Banach spaces and noise distributions appears in Section III-C.

Theorem 1 (Informal). *Let $x_0 \in \mathbb{R}^d$ and let X_T be obtained from x_0 by iterating*

$$x_{t+1} \doteq \psi_{t+1}(x_t) + Z_{t+1}$$

for some sequence of contractive maps $\{\psi_t\}_{t=1}^T$ and $Z_{t+1} \sim$

⁰Extended abstract. Full version appears as [1].

$\mathcal{N}(0, \sigma^2 \mathbb{I}_d)$. Let X'_T denote the output of the same process started at some x'_0 . Then for every $\alpha \geq 1$, $D_\alpha(X_T \| X'_T) \leq \frac{\alpha \|x_0 - x'_0\|}{2T\sigma^2}$.

We note that in this result we measure the divergence only between the final steps, in other words, the intermediate steps of the iteration are not revealed. This theorem is a special case of our more general result Theorem 22. This result translates a *metric* assumption of bounded distance between x_0 and x'_0 to an information-theoretic conclusion of bounded Rényi divergence between X_T and X'_T . While standard facts about the Gaussian distribution allow one to make such a statement for a one-step process, the intermediate arbitrary contractive steps essentially rule out a first principles approach to proving such a theorem. We use a careful induction argument that rests on controlling the “distance” between X_t and X'_t . We start by measuring the metric distance when $t = 0$ and gradually transform this to an information theoretic divergence at $t = T$. We interpolate between these two using a new *hybrid* distance measure that we refer to as *shifted divergence*. We believe that this notion should find additional applications in the analyses of stochastic processes. Our bounds are tight (with no loss in constants) and show that the worst-case for such a result is when all the contractive maps are the identity map.

This result has some surprising implications. Consider an iterative mechanism that processes one input record at a time, n iterations in total. The immediate application of this result to this mechanism leads to the following observation about individuals’ privacy loss. The person whose record was processed last experiences privacy loss afforded by the Gaussian noise added at the last iteration. At the same time, the person whose record was processed *first* suffers the least amount of privacy loss, equal to $1/n$ of the last one’s. Importantly, the order in which the inputs were considered need not be random or secret for this analysis to be applicable. In contrast, privacy amplification by sampling depends crucially on the sample’s randomness and secrecy.

We outline some applications of this analysis in privacy-preserving machine learning via convex optimization.

Distributed stochastic gradient descent.: In this setting records are stored locally, and the parties engage in a distributed computation to train a model [9]. Using amplification by sampling as in DP-SGD by Abadi et al. [8] would require keeping secret the set of parties taking part in each step of the algorithm. When the communication channel is not trusted, hiding whether or not a party takes part in a certain step would essentially require all parties to communicate in all steps, leading to an unreasonable amount of communication. In addition, the assumptions that the sample of parties participating in each step is a random subset may itself be difficult to enforce in many settings.

Our approach does not need the order of participating parties to be random or hidden. It is sufficient to hide the model

itself until a certain number of update steps are applied. This approach then allows significantly reducing communication costs to be proportional to the size of the mini-batch (the number of records consumed by each update). Additionally, our approach can amplify privacy even when the noise added in each step is too small to guarantee much privacy. This is in contrast to amplification by sampling, which requires the unamplified privacy cost to be small to start with: a starting ε becomes $\approx q\varepsilon(1 + \exp(\varepsilon))$ which is close to $2q\varepsilon$ for small ε but grows quickly, and for instance, precludes setting $\varepsilon \geq 1/q$ for small q . Our main result applies for arbitrary σ so that even if each σ is very small (say, $1/\sqrt{n}$) the final privacy is non-vacuous. A smaller noise scale then permits a smaller size of each mini-batch, further reducing the communication cost. On the negative side, the privacy guarantee we get varies between examples: examples used early in the SGD get stronger privacy than those occurring late.

Multi-query setting.: Our approach above gives better privacy than competing approaches to the parties taking part early in the computation, while giving similar guarantees to the last user. This better per-user privacy guarantee can allow one to solve several such convex optimization problems on the same set of users, at no increase in the worst-case privacy cost. Specifically, if we have n parties, then we can solve $\tilde{\Omega}(n)$ such convex optimization problems at the same privacy cost as answering one of them. More generally, the privacy cost grows linearly in $\tilde{O}(\sqrt{\max\{k/n, 1\}})$. To our knowledge, except for privacy-amplification-by-sampling, existing techniques such as output perturbation have utility bounds that grow linearly in \sqrt{k} .

Public/private data.: The setting in which some public data from the same distribution as private data is available has been recently identified as promising and practically important [10], [11]. The public corpus can be based on opt-in population, such as a product’s developers or early testers, data shared by volunteers [12], or be released through a legal process [13].

In this model, the last iterations of the iterative algorithm can be done over the public samples whose privacy need not be preserved. Since data points used early lose less privacy, we can add much less noise at each step. In effect, having m public samples decreases the error due to the addition of noise by a factor of \sqrt{m} . In the absence of public data, privacy comes at a provable cost: while the statistical error due to sampling scales as $1/\sqrt{n}$ independently of the dimension, the error of the differentially private version scales as \sqrt{d}/\sqrt{n} [14]. Our results imply that for convex optimization problems satisfying very mild smoothness assumptions, given $\tilde{O}(d)$ public data points, we can ensure that the additional error due to privacy is comparable to the statistical error.

We remark that our technique requires that the optimized functions satisfy a mild smoothness assumption. However,

as we show in the full version, in our applications we can always achieve the desired level of smoothness by convolving the optimized functions with the Gaussian kernel. Such convolution introduces an additional error but this error is usually dominated by the additional error necessary to ensure privacy.

Organization.: The rest of the paper is organized as follows. After discussing some additional related work, we start with some preliminaries in Section II. We present our main technique in Section III. Section IV shows how this technique can be applied to versions of the noisy stochastic gradient descent algorithm. Finally, in Section V, we apply this framework to derive the applications mentioned above.

A. Related Work

The field of differentially private convex optimization spans almost a decade [15]–[27]. Many of these results are optimal under different regimes such as empirical risk, population risk, the low-dimensional setting ($d \ll n$) or the high-dimensional setting $d \gg n$. Some of the algorithms (e.g., output perturbation [16] and objective perturbation [16], [18]) require finding a global optimum of an optimization problem to argue privacy and utility, while the others are based on the variants of noisy stochastic gradient descent. In this section we restrict ourselves to only the population risk, and allow comparisons to algorithms that can be implemented with one pass of stochastic gradient descent over the data set S for a direct comparison (which is close to the typical application of optimization algorithms in machine learning). We note that our analysis technique also applies to multi-pass and batch versions of gradient descent. In this setting our algorithm achieves close to optimal bounds on population risk (see Table I-A for details).

In this table we also compare the local differential privacy of the algorithms [3]. In several settings (such as distributed learning) we want the published outcome of the optimization algorithm to satisfy a strong level of (central) differential privacy while still guaranteeing ϵ_{local} differential privacy. Local differential privacy protects the user’s data even from the aggregating server or an adversary who can obtain the complete transcript of communication between the server and the user.

We note that some architectures may not be compatible with all privacy-preserving techniques or guarantees. For instance, we assume secrecy of intermediate computations, which rules out sharing intermediate updates (which is a standard step in federated learning [28]). In contrast, analyses based on secrecy of the sample (e.g., [3], [8]) require that either data be stored centrally (thus eliminating local differential privacy guarantees) or all-to-all communications.

II. PRELIMINARIES

We recall definitions and tools from the learning theory, probability theory, and differential privacy and define the

notion of shifted divergence. In the process we set up the notation that we will use throughout the paper.

A. Convex Risk Minimization

Let \mathcal{X} be the domain of data sets, and \mathcal{P} be a distribution over \mathcal{X} . Let $S = \{x_1, \dots, x_n\}$ 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 $f: \mathcal{K} \times \mathcal{X} \rightarrow \mathbb{R}$ 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 risk of solution w is defined as

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

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

Definition 3 (*β -smoothness*). A function $f: \mathcal{K} \rightarrow \mathbb{R}$ is β -smooth over the domain $\mathcal{K} \subseteq \mathbb{R}^d$ if for all $w, w' \in \mathcal{K}$, $\|\nabla f(w) - \nabla f(w')\|_2 \leq \beta \|w - w'\|_2$.

B. 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. Our main result holds in a more general setting of Banach spaces.

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 first family of distances we consider is independent of the norm:

Definition 4 (*Rényi Divergence [29]*). 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

$$D_\alpha(\mu \parallel \nu) \doteq \frac{1}{\alpha - 1} \ln \int \left(\frac{\mu(z)}{\nu(z)} \right)^\alpha \nu(z) dz.$$

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.

Proposition 5 ([30]). The following hold for any $\alpha \in (1, \infty)$, and distributions μ, μ', ν, ν' :

Additivity: $D_\alpha(\mu \times \mu' \parallel \nu \times \nu') = D_\alpha(\mu \parallel \nu) + D_\alpha(\mu' \parallel \nu')$.

Algorithm	Excess risk		LDP (ϵ_{local})
	for one task	for $k \leq n$ tasks	for one task
Noisy SGD + sampling [†] [23]	$\tilde{O}\left(\sqrt{\frac{d}{\epsilon n}}\right)$	$\tilde{O}\left(\sqrt{\frac{d}{\epsilon n}}\right)$	$\tilde{O}(n)$
Noisy SGD [20], [25]	$\tilde{O}\left(\sqrt{\frac{d}{\epsilon^2 n}}\right)$	$\tilde{O}\left(\sqrt{\frac{dk}{\epsilon^2 n}}\right)$	ϵ
Output perturbation* [16], [26]	$\tilde{O}\left(\sqrt{\frac{d}{\epsilon^2 n}}\right)$	$\tilde{O}\left(\sqrt{\frac{dk}{\epsilon^2 n}}\right)$	∞
This work	$\tilde{O}\left(\sqrt{\frac{d}{\epsilon^2 n}}\right)$	$\tilde{O}\left(\sqrt{\frac{d}{\epsilon^2 n}}\right)$	ϵ

Table I

THE EXCESS RISK CORRESPONDS TO THE EXCESS POPULATION RISK. COMPARISON FOR A SINGLE PASS OVER THE DATASET (I.E., AT MOST n GRADIENT EVALUATIONS). FOR BREVITY, THE TABLE HIDES DEPENDENCE ON $\text{poly} \ln(1/\delta)$. ([†]) THIS BOUND IS NOT STATED EXPLICITLY BUT CAN BE DERIVED BY SETTING THE PARAMETERS IN [23] APPROPRIATELY. (*) FOR OUTPUT PERTURBATION, WE USED THE VARIANT THAT CAN BE IMPLEMENTED VIA SGD IN A SINGLE PASS [26]. LDP STANDS FOR LOCAL DIFFERENTIAL PRIVACY.

Post-Processing: For any (deterministic) function f , $D_\alpha(f(\mu) \| f(\nu)) \leq D_\alpha(\mu \| \nu)$, where we $f(\mu)$ denotes the distribution of $f(X)$ where $X \sim \mu$.

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.

We will also need the following “norm-aware” statistical distance:

Definition 6 (∞ -Wasserstein Distance). The ∞ -Wasserstein distance between distributions μ and ν on a Banach space $(\mathcal{Z}, \|\cdot\|)$ is defined as

$$W_\infty(\mu, \nu) \doteq \inf_{\gamma \in \Gamma(\mu, \nu)} \text{ess sup}_{(x, y) \sim \gamma} \|x - y\|,$$

where $(x, y) \sim \gamma$ means that the essential supremum is taken relative to measure γ over $\mathcal{Z} \times \mathcal{Z}$ parameterized by (x, y) . Here $\Gamma(\mu, \nu)$ is the collection of couplings of μ and ν , i.e., the collection of measures on $\mathcal{Z} \times \mathcal{Z}$ with marginals μ and ν on the first and second factors respectively.

The following is immediate from the definition.

Lemma 7. The following are equivalent for any distributions μ, ν over \mathcal{Z} :

- 1) $W_\infty(\mu, \nu) \leq s$.
- 2) There exists jointly distributed r.v.'s (U, V) such that $U \sim \mu, V \sim \nu$ and $\Pr[\|U - V\| \leq s] = 1$.
- 3) There exists jointly distributed r.v.'s (U, W) such that $U \sim \mu, U + W \sim \nu$ and $\Pr[\|W\| \leq s] = 1$.

Next we define a hybrid between these two families of distances that plays a central role in our work.

Definition 8 (Shifted Rényi Divergence). Let μ and ν be distributions defined on a Banach space $(\mathcal{Z}, \|\cdot\|)$. For parameters $z \geq 0$ and $\alpha \geq 1$, the z -shifted Rényi divergence between μ and ν is defined as

$$D_\alpha^{(z)}(\mu \| \nu) \doteq \inf_{\mu': W_\infty(\mu, \mu') \leq z} D_\alpha(\mu' \| \nu).$$

The following follows from the definition:

Proposition 9. The shifted Rényi divergences satisfy the following for any μ, ν , and any $\alpha \in (1, \infty)$:

Monotonicity: For $0 \leq z \leq z'$, $D_\alpha^{(z')}(\mu \| \nu) \leq D_\alpha^{(z)}(\mu \| \nu)$.

Shifting: For any $x \in \mathcal{Z}$, $D_\alpha^{(\|x\|)}(\mu \| \nu) \leq D_\alpha(\mu * \mathbf{x} \| \nu)$, where we let \mathbf{x} denote the distribution of the random variable that is always equal to x (note that $\mu * \mathbf{x}$ is the distribution of $U + x$ for $U \sim \mu$).

Definition 10. For a noise distribution ζ 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 ζ and the same distribution ζ shifted by a vector of length at most a :

$$R_\alpha(\zeta, a) \doteq \sup_{x: \|x\| \leq a} D_\alpha(\zeta * \mathbf{x} \| \zeta).$$

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 , $D_\alpha(\mathcal{N}(0, \sigma^2 \mathbb{I}_d) \| \mathcal{N}(x, \sigma^2 \mathbb{I}_d)) = \alpha \|x\|_2^2 / 2\sigma^2$. 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 $D_\alpha(U \| V)$, $W_\infty(U, V)$ and $D_\alpha^{(z)}(U \| V)$ to mean $D_\alpha(\mu \| \nu)$, $W_\infty(\mu, \nu)$ and $D_\alpha^{(z)}(\mu \| \nu)$, respectively.

C. (Rényi) Differential Privacy

The notion of differential privacy (Definition 11) is by now a de facto standard for statistical data privacy [2], [31], [32]. At a semantic level, the privacy guarantee ensures that an adversary learns almost the same thing about an individual independent of the individual’s presence or absence in the data set. The parameters (ϵ, δ) quantify the amount of information leakage. A common choice of these parameters is $\epsilon \approx 0.1$ and $\delta = 1/n^{\omega(1)}$, where n refers to size of the dataset.

Definition 11 ([2], [9]). A randomized algorithm \mathcal{A} is (ϵ, δ) -differentially private ((ϵ, δ) -DP) if, for all neighboring data

sets S and S' and for all events \mathcal{O} in the output space of \mathcal{A} , we have

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

The notion of neighboring data sets is domain-dependent, and it is commonly taken to capture the contribution of a single individual. In the simplest case S and S' differ in one record, or equivalently, $d_H(S, S') = 1$, where $d_H(S, S')$ is the Hamming distance. We also define

Definition 12 (Per-person Privacy). *An algorithm \mathcal{A} operating on a sequence of data points x_1, \dots, x_n is said to satisfy (ε, δ) -differentially privacy at index i if for any pair of sequences that differ in the i th position, and for any event \mathcal{O} in the output space of \mathcal{A} , we have*

$$\begin{aligned} \Pr[\mathcal{A}(x_1, \dots, x_i, \dots, x_n) \in \mathcal{O}] \\ \leq e^\varepsilon \Pr[\mathcal{A}(x_1, \dots, x'_i, \dots, x_n) \in \mathcal{O}] + \delta. \end{aligned}$$

Another related model of privacy is local differential privacy [3]. In this model each user executes a differentially private algorithm on their individual input which is then used for arbitrary subsequent computation (we omit the formal definition as it is not used in our work).

Starting with Concentrated Differential Privacy [33], definitions that allow more fine-grained control of the privacy loss random variable have proven useful. The notions of zCDP [34], Moments Accountant [8], and Rényi differential privacy (RDP) [35] 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 [10], [36]. In the current work, we will use the nomenclature based on the notion of the Rényi divergence (Definition 4).

Definition 13 ([35]). *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*

$$D_\alpha(\mathcal{A}(S) \parallel \mathcal{A}(S')) \leq \varepsilon.$$

Per-person RDP can be defined in an analogous way. The following two lemmas [35] allow translating Rényi differential privacy to (ε, δ) -differential privacy, and give a composition rule for RDP.

Lemma 14. *If \mathcal{A} satisfies (α, ε) -Rényi differential privacy, then for all $\delta \in (0, 1)$ it also satisfies $(\varepsilon + \frac{\ln(1/\delta)}{\alpha-1}, \delta)$ -differential privacy. Moreover, pure $(\varepsilon, 0)$ -differential privacy coincides with (∞, ε) -RDP.*

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

Lemma 15. *If $\mathcal{A}_1, \dots, \mathcal{A}_k$ are randomized algorithms satisfying, respectively, (α, ε_1) -RDP, \dots , (α, ε_k) -RDP, then their composition defined as $(\mathcal{A}_1(S), \dots, \mathcal{A}_k(S))$ is $(\alpha, \varepsilon_1 + \dots + \varepsilon_k)$ -RDP. Moreover, the i 'th algorithm can be chosen on the basis of the outputs of algorithms $\mathcal{A}_1, \dots, \mathcal{A}_{i-1}$.*

D. Contractive Noisy Iteration

We start by recalling the definition of a contraction.

Definition 16 (Contraction). *For a Banach space $(\mathcal{Z}, \|\cdot\|)$, a function $\psi: \mathcal{Z} \rightarrow \mathcal{Z}$ 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 17. *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 [37].

Proposition 18. *Suppose that a function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 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 19 (Contractive Noisy Iteration (CNI)). *Given an initial random state $X_0 \in \mathcal{Z}$, a sequence of contractive functions $\psi_t: \mathcal{Z} \rightarrow \mathcal{Z}$, and a sequence of noise distributions $\{\zeta_t\}$, 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 ζ_{t+1} . For brevity, we will denote the random variable output by this process after T steps as $CNI_T(X_0, \{\psi_t\}, \{\zeta_t\})$.

III. COUPLED DESCENT

In this section, we prove a bound on the Rényi divergence between the outputs of two contractive noisy iterations. Suppose that X_0 and X'_0 are two random states such that $W_\infty(X_0, X'_0) \leq 1$. The map's contractivity and the fact that we are adding noise ζ ensures that X_1 and X'_1 are $R_\alpha(\zeta, 1)$ -close in α -Rényi divergence. By the post-processing property

of Rényi divergence, X_T and X'_T are similarly close. Our main theorem says that this can be substantially improved if we do not release the intermediate steps. The noise added in subsequent steps further decreases the Rényi divergence even when contractive steps are taken in between the noise addition.

While the final result is a statement about Rényi divergences, the shifted Rényi divergences play a crucial role in the proof. We start with an important technical lemma that for the noise addition step, allows one to reduce the shift parameter z . We will then show how contractive maps affect the shifted divergence. Armed with these results, we prove the main theorem in Section III-C.

A. The Shift-Reduction Lemma

In this section we prove the key lemma that relates $D_\alpha^{(z)}(\mu * \zeta \parallel \nu * \zeta)$ to $D_\alpha^{(z+a)}(\mu \parallel \nu)$. Recall that we use $R_\alpha(\zeta, a)$ to measure how well noise distribution ζ hides changes in our norm $\|\cdot\|$ (see Definition 10):

$$R_\alpha(\zeta, a) \doteq \sup_{x: \|x\| \leq a} D_\alpha(\zeta * \mathbf{x} \parallel \zeta).$$

Lemma 20 (Shift-Reduction Lemma). *Let μ, ν and ζ be distributions over a Banach space $(\mathcal{Z}, \|\cdot\|)$. Then for any $a \geq 0$,*

$$D_\alpha^{(z)}(\mu * \zeta \parallel \nu * \zeta) \leq D_\alpha^{(z+a)}(\mu \parallel \nu) + R_\alpha(\zeta, a).$$

Proof: Let U be distributed as μ and V as ν . We first show the result for the case when $z = 0$. Let μ' be the distribution certifying $D_\alpha^{(a)}(\mu \parallel \nu)$, that is $D_\alpha(\mu' \parallel \nu) = D_\alpha^{(a)}(\mu \parallel \nu)$ and $W_\infty(\mu, \mu') \leq a$. Let (U, W) be the random variable whose existence is given by Lemma 7. That is, $\|W\| \leq a$ with probability 1, $U \sim \mu$ and $U + W \sim \mu'$. Let Y be an independent random variable distributed as ζ . We can write

$$\begin{aligned} D_\alpha(\mu * \zeta \parallel \nu * \zeta) &= D_\alpha(U + Y \parallel V + Y) \\ &= D_\alpha(U + W - W + Y \parallel V + Y) \\ &\leq D_\alpha((U + W, -W + Y) \parallel (V, Y)), \end{aligned}$$

where we have used the post-processing property of Rényi divergence. Note that the distribution (V, Y) is a product distribution, whereas the factors of $(U + W, -W + Y)$ are dependent. Denoting the p_X the density function of a

random variable X , we expand

$$\begin{aligned} &\exp((\alpha - 1)D_\alpha((U + W, -W + Y) \parallel (V, Y))) \\ &= \int \int \left(\frac{p_{(U+W, -W+Y)}(v, y)}{p_{(V, Y)}(v, y)} \right)^\alpha \zeta(y) \nu(v) dy dv \\ &= \int \int \left(\frac{p_{U+W}(v) \cdot p_{-W+Y|U+W=v}(y)}{\nu(v) \cdot \zeta(y)} \right)^\alpha \zeta(y) \nu(v) dy dv \\ &= \int \left(\frac{p_{U+W}(v)}{\nu(v)} \right)^\alpha \times \\ &\quad \left(\int \left(\frac{p_{-W+Y|U+W=v}(y)}{\zeta(y)} \right)^\alpha \zeta(y) dy \right) \nu(v) dv \\ &\leq \int \left(\frac{p_{U+W}(v)}{\nu(v)} \right)^\alpha \nu(v) dv \times \\ &\quad \text{ess sup}_{v' \sim \nu} \int \left(\frac{p_{-W+Y|U+W=v'}(y)}{\zeta(y)} \right)^\alpha \zeta(y) dy \\ &\leq \int \left(\frac{p_{U+W}(v)}{\nu(v)} \right)^\alpha \nu(v) dv \times \\ &\quad \text{ess sup}_{v' \sim \nu} \text{ess sup}_{w \sim p_W} \int \left(\frac{p_{-W+Y|W=w, U+W=v'}(y)}{\zeta(y)} \right)^\alpha \zeta(y) dy \\ &\leq \int \left(\frac{\mu'(v)}{\nu(v)} \right)^\alpha \nu(v) dv \times \\ &\quad \text{ess sup}_{v' \sim \nu} \text{ess sup}_{w \sim p_W} \int \left(\frac{p_{-W+Y|W=w}(y)}{\zeta(y)} \right)^\alpha \zeta(y) dy \\ &\leq \exp((\alpha - 1)D_\alpha(\mu' \parallel \nu)) \cdot \exp((\alpha - 1)R_\alpha(\zeta, a)). \end{aligned}$$

(Proposition 5)

Taking logs and dividing by $(\alpha - 1)$, we get the claim for $z = 0$.

The general z case reduces readily to the $z = 0$ case. Define

$$h_z(x) = \begin{cases} x & \text{if } \|x\| \leq z, \\ \frac{x}{\|x\|} z & \text{otherwise.} \end{cases}$$

It is easy to see that $\|h_z(x)\| \leq z$ for all x , and that $\|x - h_z(x)\| \leq a$ whenever $\|x\| \leq z + a$.

As before, let (U, W) be r.v.'s from the joint distribution guaranteed by Lemma 7. Let $W_1 \doteq h_z(W)$ and $W_2 \doteq W - W_1$. It follows that $\|W_1\| \leq z$ and $\|W_2\| \leq a$ with probability 1. We write

$$\begin{aligned} D_\alpha^{(z)}(U + Y \parallel V + Y) &= D_\alpha^{(z)}(U + W_1 + Y - W_1 \parallel V + Y) \\ &\leq D_\alpha(U + W_1 + Y \parallel V + Y) \\ &\leq D_\alpha^{(a)}(U + W_1 \parallel V) + R_\alpha(\zeta, a), \end{aligned}$$

where we have used the $z = 0$ case in the last step. On the other hand,

$$\begin{aligned} D_\alpha^{(a)}(U + W_1 \parallel V) &\leq D_\alpha(U + W_1 + W_2 \parallel V) \\ &= D_\alpha(U + W \parallel V) \\ &= D_\alpha^{(z+a)}(U \parallel V). \end{aligned}$$

This completes the proof. ■

B. Contractive Maps

We next show that contractive maps cannot increase a shifted divergence. In the lemma below we give a more general version that allows using different contractive maps.

Lemma 21 (Contraction reduces $D_\alpha^{(z)}$). *Suppose that ψ and ψ' are contractive maps on $(\mathcal{Z}, \|\cdot\|)$ and $\sup_x \|\psi(x) - \psi'(x)\| \leq s$. Then for r.v.'s X and X' over \mathcal{Z} ,*

$$D_\alpha^{(z+s)}(\psi(X) \parallel \psi'(X')) \leq D_\alpha^{(z)}(X \parallel X').$$

Proof: By definition of $D_\alpha^{(z)}(\cdot \parallel \cdot)$ and Lemma 7, there is a joint distribution (X, Y) such that $D_\alpha(Y \parallel X') = D_\alpha^{(z)}(X \parallel X')$ and $\Pr[\|X - Y\| \leq z] = 1$. By the post-processing property of Rényi divergence, we have that $D_\alpha(\psi'(Y) \parallel \psi'(X')) \leq D_\alpha(Y \parallel X') = D_\alpha^{(z)}(X \parallel X')$. Moreover,

$$\begin{aligned} \|\psi(X) - \psi'(Y)\| &\leq \|\psi(X) - \psi(Y)\| + \|\psi(Y) - \psi'(Y)\| \\ &\leq \|X - Y\| + s \\ &\leq z + s. \end{aligned}$$

Thus $(\psi(X), \psi'(Y))$ is a coupling establishing the claimed upper bound on $D_\alpha^{(z+s)}(\psi(X) \parallel \psi'(Y))$. ■

C. Privacy Amplification by Iteration

We are now ready to prove our main result. We prove a general statement that can handle changes in several ψ 's; this enables us to easily analyze algorithms that access data points more than once¹. Recall that R_α is introduced in Definition 10 and measures the maximal Rényi divergence of order α between a noise distribution and its shifted copy.

Theorem 22. *Let X_T and X'_T denote the output of $\text{CNI}_T(X_0, \{\psi_t\}, \{\zeta_t\})$ and $\text{CNI}_T(X_0, \{\psi'_t\}, \{\zeta_t\})$. Let $s_t \doteq \sup_x \|\psi_t(x) - \psi'_t(x)\|$. Let a_1, \dots, 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*

$$D_\alpha^{(z_T)}(X_T \parallel X'_T) \leq \sum_{t=1}^T R_\alpha(\zeta_t, a_t).$$

In particular, if $z_T = 0$, then

$$D_\alpha(X_T \parallel X'_T) \leq \sum_{t=1}^T R_\alpha(\zeta_t, a_t).$$

Proof: The proof is by induction where we use the contraction-reduces- $D_\alpha^{(z)}$ lemma and then reduce the shift amount by a_t using the shift-reduction lemma.

Let X_t (resp., X'_t) denote the t 'th iterate of the $\text{CNI}(X_0, \{\psi_t\}, \{\zeta_t\})$ (resp., $\text{CNI}(X_0, \{\psi'_t\}, \{\zeta_t\})$). We argue that for all $t \leq T$,

$$D_\alpha^{(z_t)}(X_t \parallel X'_t) \leq \sum_{i=1}^t R_\alpha(\zeta_i, a_i).$$

¹Since Rényi divergence does not satisfy the triangle inequality, blackbox analyses of such algorithms use the group privacy properties of RDP that can be loose.

The base case is $t = 0$. By definition, $X_0 = X'_0$ and $z_0 = 0$. For the inductive step, let Z_{t+1} denote the random variable drawn from ζ_{t+1} .

$$\begin{aligned} D_\alpha^{(z_{t+1})}(X_{t+1} \parallel X'_{t+1}) &= \\ D_\alpha^{(z_{t+1})}(\psi_{t+1}(X_t) + Z_{t+1} \parallel \psi'_{t+1}(X'_t) + Z_{t+1}) & \\ \leq D_\alpha^{(z_{t+1}+a_{t+1})}(\psi_{t+1}(X_t) \parallel \psi'_{t+1}(X'_t)) + R_\alpha(\zeta_{t+1}, a_{t+1}) & \quad (\text{Lemma 20}) \\ = D_\alpha^{(z_t+s_{t+1})}(\psi_{t+1}(X_t) \parallel \psi'_{t+1}(X'_t)) + R_\alpha(\zeta_{t+1}, a_{t+1}) & \quad (\text{Definition of } z_{t+1}) \\ \leq D_\alpha^{(z_t)}(X_t \parallel X'_t) + R_\alpha(\zeta_{t+1}, a_{t+1}) & \quad (\text{Lemma 21}) \\ \leq \sum_{i=1}^t R_\alpha(\zeta_i, a_i) + R_\alpha(\zeta_{t+1}, a_{t+1}). & \\ & \quad (\text{induction hypothesis}) \end{aligned}$$

This completes the induction step and the proof. ■

IV. PRIVACY GUARANTEES FOR NOISY STOCHASTIC GRADIENT DESCENT

We will now apply our analysis technique to derive the privacy parameters of several versions of the noisy stochastic gradient descent algorithm (also referred to as Stochastic Gradient Langevin Dynamics) defined as follows. 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 $S = (x_1, \dots, x_n)$, starting point w_0 , rate parameter η , and noise scale σ the algorithm works as follows. Starting from $w_0 \in \mathcal{K}$ perform the following update $v_{t+1} \doteq w_t - \eta(\nabla_w f(w_t, x_{t+1}) + Z)$ and $w_{t+1} \doteq \Pi_{\mathcal{K}}(v_{t+1})$, where Z is a freshly drawn sample from $\mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ and $\Pi_{\mathcal{K}}$ denotes the Euclidean projection to set \mathcal{K} . We refer to this algorithm as $\text{PNSGD}(S, w_0, \eta, \sigma)$ and describe it formally in Algorithm 1.

Algorithm 1 Projected noisy stochastic gradient descent (PNSGD)

Input: Data set $S = \{x_1, \dots, x_n\}$, $f: \mathcal{K} \times \mathcal{X} \rightarrow \mathbb{R}$ a convex function in the first parameter, learning rate η , starting point $w_0 \in \mathcal{K}$, noise parameter σ .

- 1: **for** $t \in \{0, \dots, n-1\}$ **do**
 - 2: $v_{t+1} \leftarrow w_t - \eta(\nabla_w f(w, x_{t+1}) + Z)$, where $Z \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$.
 - 3: $w_{t+1} \leftarrow \Pi_{\mathcal{K}}(v_{t+1})$, where $\Pi_{\mathcal{K}}(w) = \arg \min_{\theta \in \mathcal{K}} \|\theta - w\|_2$ is the ℓ_2 -projection on \mathcal{K} .
 - 4: **return** the final iterate w_n .
-

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 18). 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 17). Naturally, a composition of two contractive maps is a contractive map and therefore we can conclude that $\text{PNSGD}(S, w_0, \eta, \sigma)$ is an instance of contractive noisy iteration. More formally, consider the sequence $v_0 = w_0, v_1, \dots, v_n$. 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 $\eta \cdot \sigma$. Note that the final output of the algorithm is $w_n = \Pi_{\mathcal{K}}(v_n)$ but it does not affect our analysis of divergence as it can be seen as an additional post-processing step.

For this baseline algorithm we prove that points that are used earlier have stronger privacy guarantees due to noise injected in subsequent steps.

Theorem 23. *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 $\eta \leq 2/\beta, \sigma > 0, \alpha > 1, t \in [n]$, starting point $w_0 \in \mathcal{K}$, and $S \in \mathcal{X}^n$, $\text{PNSGD}(S, w_0, \eta, \sigma)$ satisfies $(\alpha, \frac{\alpha \cdot \varepsilon}{n+1-t})$ -RDP for its t 'th input, where $\varepsilon = \frac{2L^2}{\sigma^2}$.*

Proof: Let $S \doteq (x_1, \dots, x_n)$ and $S' \doteq (x_1, \dots, x_{t-1}, x'_t, x_{t+1}, \dots, x_n)$ be two arbitrary datasets that at index t . As discussed above, under the smoothness condition $\eta \leq 2/\beta$ the steps of $\text{PNSGD}(S, w_0, \eta, \sigma)$ are a contractive noisy iteration. Specifically, on dataset S , the CNI is defined by the initial point w_0 , sequence of functions $g_i(w) \doteq \Pi_{\mathcal{K}}(w) - \eta \nabla f(\Pi_{\mathcal{K}}(w), x_i)$ and sequence of noise distributions $\zeta_i \sim \mathcal{N}(0, (\eta\sigma)^2 \mathbb{I}_d)$. Similarly, on the dataset S' , the CNI is defined in the same way with the exception of $g'_t(w) \doteq \Pi_{\mathcal{K}}(w) - \eta \nabla f(\Pi_{\mathcal{K}}(w), x'_t)$. By our assumption, $f(w, x)$ is L -Lipschitz for every $x \in \mathcal{X}$ and $w \in \mathcal{K}$ and therefore

$$\begin{aligned} & \sup_w \|g_t(w) - g'_t(w)\|_2 \\ &= \sup_w \|\eta \nabla f(\Pi_{\mathcal{K}}(w), x_t) - \eta \nabla f(\Pi_{\mathcal{K}}(w), x'_t)\|_2 \leq 2\eta L. \end{aligned}$$

We can now apply Theorem 22 with $a_1, \dots, a_{t-1} = 0$ and $a_t, \dots, a_n = \frac{2\eta L}{n-t+1}$. Note that $s_t = 2\eta L$ and $s_i = 0$ for $i \neq t$. In addition, $z_i \geq 0$ for all $i \leq n$ and $z_n = 0$. Hence we obtain that

$$D_\alpha(X_n \parallel X'_n) \leq \frac{\alpha}{2\eta^2 \sigma^2} \sum_{i=1}^n a_i^2 \leq \frac{2\alpha L^2}{\sigma^2 \cdot (n-t+1)}$$

as claimed. \blacksquare

We now consider privacy guarantees for several variants of this baseline approach. These variants are needed to ensure utility guarantees, that require that the algorithm output one of the iterates randomly. Specifically, we define the algorithm $\text{Skip-PNSGD}(S, w_0, \eta, \sigma)$ as the algorithm that picks randomly and uniformly $t_0 \in \{0, 1, \dots, \lfloor n/2 \rfloor\}$ and then skips the first t_0 points. That is, it makes only

$n - t_0$ steps and at step t the update is $w'_{t+1} = w_t - \eta(\nabla_w f(w_t, x_{t+1+t_0}) + Z)$. It is easy to see that the privacy guarantees $\text{Skip-PNSGD}(S, w_0, \eta, \sigma)$ are at least as good as those we gave for $\text{PNSGD}(S, w_0, \eta, \sigma)$ in Theorem 23.

Theorem 24. *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 $\eta \leq 2/\beta, \sigma > 0, \alpha > 1, t \in [n]$, starting point $w_0 \in \mathcal{K}$, and $S \in \mathcal{X}^n$, $\text{Skip-PNSGD}(S, w_0, \eta, \sigma)$ satisfies $(\alpha, \frac{\alpha \cdot \varepsilon}{n+1-t})$ -RDP for point with index t , where $\varepsilon = \frac{2L^2}{\sigma^2}$.*

Finally, we consider a version of PNSGD with random stopping. Namely, instead of running for n steps the algorithm picks $T \in [n]$ randomly and uniformly, makes T steps and outputs w_T . We refer to this version as $\text{Stop-PNSGD}(S, w_0, \eta, \sigma)$. To analyze this algorithm we will need to prove a weak² form of convexity for the Rényi divergence that might have other applications.

Lemma 25. *Let μ_1, \dots, μ_n and ν_1, \dots, ν_n be probability distributions over some domain \mathcal{Z} such that for all $i \in [n]$, $D_\alpha(\mu_i \parallel \nu_i) \leq c/(\alpha - 1)$ for some $c \in (0, 1]$. Let ρ be a probability distribution over $[n]$ and denote by μ_ρ (or ν_ρ) the probability distribution over \mathcal{Z} obtained by sampling i from ρ and then outputting a random sample from μ_i (respectively, ν_i). Then*

$$D_\alpha(\mu_\rho \parallel \nu_\rho) \leq (1 + c) \cdot \mathbb{E}_{i \sim \rho} [D_\alpha(\mu_i \parallel \nu_i)].$$

Proof: Let μ'_ρ (or ν'_ρ) be the probability distribution over $[n] \times \mathcal{Z}$ obtained by sampling i from ρ and then sampling a random x from μ_i (respectively, ν_i) and outputting (i, x) . We can obtain μ_ρ from μ'_ρ by applying the function that removes the first coordinate and the same function applied to ν'_ρ gives ν_ρ . Therefore, by the post-processing properties of the Rényi divergence, we obtain that $D_\alpha(\mu_\rho \parallel \nu_\rho) \leq D_\alpha(\mu'_\rho \parallel \nu'_\rho)$. Now observe that for every $i \in [n]$ and $x \in \mathcal{Z}$,

²The weakness here is the strong (if necessary) assumption that $D_\alpha(p_i \parallel q_i) \leq c/(\alpha - 1)$ for some $c \leq 1$.

$\mu'_\rho(i, x) = \rho(i) \cdot \mu_i(x)$. Therefore,

$$\begin{aligned}
& D_\alpha(\mu'_\rho \parallel \nu'_\rho) \\
&= \frac{1}{\alpha-1} \ln \mathbb{E}_{(i,x) \sim \nu'_\rho} \left[\left(\frac{\mu'_\rho(i, x)}{\nu'_\rho(i, x)} \right)^\alpha \right] \\
&= \frac{1}{\alpha-1} \ln \mathbb{E}_{i \sim \rho} \left[\mathbb{E}_{x \sim \nu_i} \left[\left(\frac{\mu_i(x)}{\nu_i(x)} \right)^\alpha \right] \right] \\
&= \frac{1}{\alpha-1} \ln \mathbb{E}_{i \sim \rho} \left[e^{(\alpha-1) \cdot D_\alpha(\mu_i \parallel \nu_i)} \right] \\
&\leq \frac{1}{\alpha-1} \ln \mathbb{E}_{i \sim \rho} \left[1 + (1+c)(\alpha-1) \cdot D_\alpha(\mu_i \parallel \nu_i) \right] \\
&= \frac{1}{\alpha-1} \ln \left(1 + (1+c)(\alpha-1) \cdot \mathbb{E}_{i \sim \rho} [D_\alpha(\mu_i \parallel \nu_i)] \right) \\
&\leq \frac{1}{\alpha-1} \left((1+c)(\alpha-1) \cdot \mathbb{E}_{i \sim \rho} [D_\alpha(\mu_i \parallel \nu_i)] \right) \\
&= (1+c) \cdot \mathbb{E}_{i \sim \rho} [D_\alpha(\mu_i \parallel \nu_i)],
\end{aligned}$$

where to obtain the inequality in the fourth line we used the fact that for every $a \leq c \leq 1$, $e^a \leq 1+a+a^2 \leq 1+(1+c)a$. \blacksquare

We can now state and prove the privacy guarantees for Stop-PNSGD(η, σ).

Theorem 26. *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 $\eta \leq 2/\beta$, $\alpha > 1$, starting point $w_0 \in \mathcal{K}$, $\sigma \geq L\sqrt{2(\alpha-1)\alpha}$, and dataset $S \in \mathcal{X}^n$, Stop-PNSGD(S, w_0, η, σ) satisfies $\left(\alpha, \frac{4\alpha L^2 \cdot \ln n}{n\sigma^2}\right)$ -RDP.*

Proof: Let $S \doteq (x_1, \dots, x_n)$ and $S' \doteq (x_1, \dots, x_{t-1}, x'_t, x_{t+1}, \dots, x_n)$ be two arbitrary datasets that differ in the element at index t . For every value of $T \in [n]$, let X_T denote the output of Stop-PNSGD(S, w_0, η, σ) on S after T steps and analogously define X'_T for Stop-PNSGD(S', w_0, η, σ). If $t > T$ then the algorithm does not reach x_t (or x'_t) and hence $D_\alpha(X_T \parallel X'_T) = 0$. Otherwise, we can use Theorem 23 with $n = T$ to obtain that

$$D_\alpha(X_T \parallel X'_T) \leq \frac{2\alpha L^2}{\sigma^2 \cdot (T-t+1)}.$$

By definition, the output of Stop-PNSGD(S, w_0, η, σ) corresponds to picking T randomly and uniformly from $[n]$ and then outputting X_T . We denote the resulting random variable by Y_n and denote Y'_n the corresponding random variable for S' . By our assumption, $\sigma \geq L\sqrt{2(\alpha-1)\alpha}$ and therefore for every $t \geq T$,

$$\frac{2\alpha L^2}{\sigma^2 \cdot (T-t+1)} \leq \frac{2\alpha L^2}{\sigma^2} \leq \frac{1}{\alpha-1}.$$

Hence the conditions of Lemma 25 are satisfied with $c = 1$. This implies that

$$\begin{aligned}
D_\alpha(Y_T \parallel Y'_T) &\leq 2 \cdot \frac{1}{n} \sum_{T \in [n]} D_\alpha(X_T \parallel X'_T) \\
&\leq 2 \cdot \frac{1}{n} \sum_{T=t}^n \frac{2\alpha L^2}{\sigma^2 \cdot (T-t+1)} \\
&\leq \frac{4\alpha L^2 \cdot \ln(n-t+1)}{n\sigma^2} \leq \frac{4\alpha L^2 \cdot \ln n}{n\sigma^2}.
\end{aligned}$$

\blacksquare

In the full version, we present a simple analysis of a multiple-pass version of the SGD algorithm. While it gives results that are quantitatively similar to what can be achieved using privacy amplification by sampling results from [8], the approach here works in the distributed setting and leads to a significantly simpler proof.

Finally, we remark that PNSGD(S, w_0, η, σ) and its variants described above satisfy local differential privacy (even without the smoothness assumption).

V. APPLICATIONS

We now show how to use the algorithms we have analyzed to derive new results for privacy-preserving convex optimization. Omitted proofs and the use of the smoothing to remove the smoothness assumptions appear in the full version [1].

One of the applications we discussed is concerned with a distributed model, where the input records are spread across users' devices. In the "Our data, ourselves" model proposed by [9], each user's device holds their data, and there is no central trusted party. Under reasonable assumptions on the devices, one can simulate a trusted party by means of a Secure Multi-party Computation protocol. While one can assume that all peer-to-peer channels are encrypted, it is reasonable to assume that an attacker can detect the presence or absence of communication. Additionally, in many settings of interest, bandwidth is at a premium and the number of users is large enough that all-to-all communication becomes an implementation bottleneck.

These constraints rule out algorithms that require all parties to be active in every iteration. Consequently, since the presence or absence of communication may be observed by an adversary, we cannot apply privacy amplification by sampling. While algorithms such as bolt-on differential privacy [26] may be usable in the trusted central party setting, their privacy guarantee is uniform and weaker than ours. Our approach gives some baseline local differential privacy and a stronger global privacy guarantee for most users.

A. Private Stochastic Optimization of Smooth Functions

We will present our results for stochastic convex optimization. Specifically, let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex body contained in a ball of radius R around the origin. Let \mathcal{P}

be a distribution over convex L -Lipschitz functions over \mathcal{K} and let $F(w) \doteq \mathbb{E}_{f \sim \mathcal{P}}[f(w)]$. We will assume that each data point corresponds to an independent sample from \mathcal{P} and the goal is to optimize $F(x)$. In order to analyze the performance of the noisy projected gradient descent algorithm for this problem we will need the following classical result about stochastic convex optimization (e.g., [38]). For the purposes of this result $F(w)$ can 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 $\text{PSGD}(G, w_0, \eta, T)$ denote the execution of the following process: starting from point w_0 , use the update $w_{t+1} \doteq \Pi_{\mathcal{K}}(w_t + \eta G(w_t))$ for $t = 0, \dots, T - 1$.

Theorem 27. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex body contained in a ball of radius R , let $F(w)$ be an arbitrary convex function over \mathcal{K} and let G be an unbiased stochastic (sub-)gradient oracle for F . Assume that for every $w \in \mathcal{K}$, $\mathbb{E}[\|G(w)\|_2^2] \leq L_G^2$. For $\eta = 2R/(L_G\sqrt{T})$ and $w_0 \in \mathcal{K}$, let w_1, \dots, w_T denote the iterates produced by $\text{PSGD}(G, w_0, \eta, T)$. Then*

$$\frac{1}{T} \sum_{t \in [T]} \mathbb{E}[F(w_t)] \leq F^* + \frac{4RL_G}{\sqrt{T}},$$

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

Note that this result gives a bound on the expected value of F averaged over all the iterates. Equivalently, it can be seen as the expected value of $F(w_t)$ with the expectation also taken over t being chosen randomly and uniformly from $[T]$. This corresponds to the random stopping of $\text{PSGD}(G, w_0, \eta, T)$. As a result we get the following baseline guarantees for $\text{Stop-PNSGD}(S, w_0, \eta, \sigma)$ we defined in Section IV (namely, these guarantees do not use our amplification analysis and do not require smoothness).

Theorem 28. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex body contained in a ball of radius R and $\{f(\cdot, x)\}_{x \in \mathcal{X}}$ be a family of convex L -Lipschitz functions over \mathcal{K} . Then for every $\varepsilon > 0$, $\delta > 0$, starting point $w_0 \in \mathcal{K}$, $\sigma = 2L\sqrt{2\ln(1.25/\delta)}/\varepsilon$, $\eta = 2R/\sqrt{n(L^2 + d\sigma^2)}$ and dataset $S \in \mathcal{X}^n$, $\text{Stop-PNSGD}(S, w_0, \eta, \sigma)$ satisfies local (ε, δ) -DP. In addition, if S consists of samples drawn i.i.d. from an arbitrary distribution \mathcal{P} over \mathcal{X} , then*

$$\mathbb{E}_{S \sim \mathcal{P}^n} [F(W)] \leq F^* + \frac{4RL}{\sqrt{n}} \cdot \sqrt{1 + \frac{8d\ln(1.25/\delta)}{\varepsilon^2}},$$

where W denotes the output of $\text{Stop-PNSGD}(S, w_0, \eta, \sigma)$ and $F(w) \doteq \mathbb{E}_{x \sim \mathcal{P}}[f(w, x)]$.

B. Per-person Privacy

We will now show how to combine our stronger privacy guarantees for some of the individuals in the dataset with the utility guarantees in Theorem 27.

Theorem 29. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex body contained in a ball of radius R and $\{f(\cdot, x)\}_{x \in \mathcal{X}}$ be a family of convex L -Lipschitz, β -smooth functions over \mathcal{K} . For every $\varepsilon > 0$, $\delta > 0$, starting point $w_0 \in \mathcal{K}$, $\sigma = 2L\sqrt{2\ln(1.25/\delta)}/\varepsilon$, dataset $S \in \mathcal{X}^n$ and index $t \in [n]$, if $\eta = \sqrt{8R}/\sqrt{n(L^2 + d\sigma^2)} \leq 2/\beta$, then $\text{Skip-PNSGD}(S, w_0, \eta, \sigma)$ satisfies local (ε, δ) -DP and $(\varepsilon/\sqrt{n-t+1}, \delta)$ -DP at index t . In addition, if S consists of samples drawn i.i.d. from an arbitrary distribution \mathcal{P} over \mathcal{X} , then*

$$\mathbb{E}_{S \sim \mathcal{P}^n} [F(W)] \leq F^* + \frac{4\sqrt{2}RL}{\sqrt{n}} \cdot \sqrt{1 + \frac{8d\ln(1.25/\delta)}{\varepsilon^2}},$$

where W denotes the output of $\text{Skip-PNSGD}(S, w_0, \eta, \sigma)$ and $F(w) \doteq \mathbb{E}_{x \sim \mathcal{P}}[f(w, x)]$.

Proof: Our privacy guarantees follow directly from Theorem 23. Let us denote by $\text{Stop}(n/2)$ - $\text{PNSGD}(S, w_0, \eta, \sigma)$ the algorithm that runs $\text{PNSGD}(S, w_0, \eta, \sigma)$ with a randomly and uniformly chosen stopping time $T \in \{[n/2], \dots, n\}$. Observe that the distribution of the output of $\text{Skip-PNSGD}(S, w_0, \eta, \sigma)$ on $S \sim \mathcal{P}^n$ is identical to the output distribution of $\text{Stop}(n/2)$ - $\text{PNSGD}(S, w_0, \eta, \sigma)$ on $S \sim \mathcal{P}^n$. (This is true since in both cases the starting point, the distribution on the number of steps and the stochastic gradient oracle are identical). $\text{Stop}(n/2)$ - $\text{PNSGD}(S, w_0, \eta, \sigma)$ can be seen as running Stop-PNSGD on $n/2$ points starting from some random point W_0 (where W_0 is the output of PNSGD on the first $n/2$ points). The utility guarantees for Stop-PNSGD hold for an arbitrary starting point and therefore the utility guarantees for $\text{Stop}(n/2)$ - PNSGD are the same as those for Stop-PNSGD (Theorem 28) for a dataset consisting of $n/2$ points. ■

C. Utility of Public Data

In a variety of settings the algorithm may also have access to a relatively small amount of data from the same distribution that do not require privacy protection. We demonstrate that by using the non-private data points at the end of the training process our per-index privacy guarantees directly lead to substantially improved utility guarantees. In particular, given $\Theta(d\ln(1/\delta)/\varepsilon^2)$ non-private points the utility guarantees of our algorithm match (up to a constant factor) those of non-private learning on the entire dataset.

Corollary 30. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex body contained in a ball of radius R and $\{f(\cdot, x)\}_{x \in \mathcal{X}}$ be a family of convex L -Lipschitz, β -smooth functions over \mathcal{K} . Let $S_{\text{priv}} \in \mathcal{X}^{n-m}$ and $S_{\text{pub}} \in \mathcal{X}^m$ be two datasets and $S \doteq (S_{\text{priv}}, S_{\text{pub}})$. For every $\varepsilon > 0$, $\delta > 0$, starting point $w_0 \in \mathcal{K}$, $\sigma = 2L\sqrt{\ln(1.25/\delta)}/m/\varepsilon$, if $\eta = \sqrt{8R}/\sqrt{n(L^2 + d\sigma^2)} \leq 2/\beta$, then $\text{Skip-PNSGD}(S, w_0, \eta, \sigma)$ satisfies (ε, δ) -DP relative to S_{priv} . In addition, if S consists of samples drawn i.i.d. from*

an arbitrary distribution \mathcal{P} over \mathcal{X} , then

$$\mathbb{E}_{S \sim \mathcal{P}^n} [F(W)] \leq F^* + \frac{4\sqrt{2}RL}{\sqrt{n}} \cdot \sqrt{1 + \frac{8d \ln(1.25/\delta)}{m\varepsilon^2}},$$

where W denotes the output of *Skip-PNSGD*(S, w_0, η, σ) and $F(w) \doteq \mathbb{E}_{x \sim \mathcal{P}} [f(w, x)]$.

D. Multiple Convex Optimizations

The privacy guarantees in Theorem 26 do not improve on the (ε, δ) -DP guarantees for an individual task since in order to convert RDP guarantees to (ε, δ) -DP we need to set $\alpha > 1/\varepsilon$ (see Lemma 14). At the same time, Theorem 26 requires setting $\sigma = \Omega(L\alpha)$ which would give (roughly) the same bound on excess population risk as the one obtained in Theorem 28. When solving k convex optimization tasks on the same dataset, standard analysis requires increasing the noise scale σ (and hence the bound on excess risk) by a factor of \sqrt{k} to keep the same (ε, δ) -DP level. In contrast, our analysis allows to bound (ε, δ) -DP directly and only requires increasing σ by a factor of $\max\{\tilde{O}(k/n), 1\}$. We note that in the context of PAC learning sample complexity of solving multiple learning problems with differential privacy was studied in [39]. The question of optimizing multiple loss functions was also studied in [21], [40]. The bounds given there are incomparable to ours: the multiplicative-weights-update-based approaches there give better bounds when $k \gg n$ and d is small but for $k \leq n$ the bounds given there are worse.

For simplicity of presentation we will state this result for solving a fixed set of k tasks with identical parameters. Composition properties of RDP imply that the bounds can be extended to using problems with different parameters and also allow choosing the tasks in an adaptive way (i.e., after observing the outcome of the previous tasks).

Theorem 31. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex body contained in a ball of radius R and $\{f_i(\cdot, x)\}_{i \in [k], x \in \mathcal{X}}$ be k families of convex L -Lipschitz, β -smooth functions over \mathcal{K} and $w_0 \in \mathcal{K}$ be a starting point. For $\varepsilon, \delta \in (0, 1)$ let $q \doteq \max\{\frac{2k \ln n}{n}, 2 \ln(1/\delta)\}$, $\sigma \doteq \frac{4L\sqrt{q \ln(1/\delta)}}{\varepsilon}$, $\eta \doteq 4R/\sqrt{n(L^2 + d\sigma^2)}$. For a dataset $S \in \mathcal{X}^n$ and $i \in [k]$, let W_i denote the output of *Stop-PNSGD*(S, w_0, η, σ) on the i 'th family of functions (with independent randomness). Then the entire output (W_1, \dots, W_k) satisfies (ε, δ) -DP whenever $\eta \leq 2/\beta$. In addition, if S consists of samples drawn i.i.d. from an arbitrary distribution \mathcal{P} over \mathcal{X} , then for every i ,*

$$\mathbb{E}_{S \sim \mathcal{P}^n} [F_i(W_i)] \leq F_i^* + \frac{4RL}{\sqrt{n}} \cdot \sqrt{1 + \frac{16dq \ln(1/\delta)}{\varepsilon^2}},$$

where $F_i(w) \doteq \mathbb{E}_{x \sim \mathcal{P}} [f_i(w, x)]$.

ACKNOWLEDGMENTS

We thank Úlfar Erlingsson and Tomer Koren for useful suggestions and insightful discussions of this work.

REFERENCES

- [1] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," *CoRR*, vol. abs/1808, 2018.
- [2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference (TCC)*, 2006, pp. 265–284.
- [3] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" in *49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2008, pp. 531–540.
- [4] A. Beimel, H. Brenner, S. P. Kasiviswanathan, and K. Nissim, "Bounds on the sample complexity for private learning and private data release," *Machine learning*, vol. 94, no. 3, pp. 401–437, Mar. 2014.
- [5] Y.-X. Wang, S. E. Fienberg, and A. J. Smola, "Privacy for free: Posterior sampling and stochastic gradient Monte Carlo," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML)*, 2015, pp. 2493–2502.
- [6] M. Bun, C. Dwork, G. N. Rothblum, and T. Steinke, "Composable and versatile privacy via truncated CDP," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2018, pp. 74–86.
- [7] Y.-X. Wang, B. Balle, and S. Kasiviswanathan, "Subsampled Rényi differential privacy and analytical moments accountant," *arXiv preprint arXiv:1808.00087*, 2018.
- [8] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 308–318.
- [9] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation." in *EUROCRYPT*, 2006, pp. 486–503.
- [10] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.
- [11] B. Avent, A. Korolova, D. Zeber, T. Hovden, and B. Livshits, "BLENDER: Enabling local search with a hybrid differential privacy model," in *Proceedings of the 26th USENIX Security Symposium*, 2017, pp. 747–764.
- [12] G. M. Church, "The personal genome project," *Molecular systems biology*, vol. 1, no. 1, 2005.
- [13] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *European Conference on Machine Learning*. Springer, 2004, pp. 217–226.
- [14] R. Bassily, A. Smith, and A. Thakurta, "Differentially private empirical risk minimization: Efficient algorithms and tight error bounds," *arXiv preprint arXiv:1405.7085*, 2014.

- [15] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS)*, 2008, pp. 289–296.
- [16] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Machine Learning Research*, vol. 12, pp. 1069–1109, Jul. 2011.
- [17] P. Jain, P. Kothari, and A. Thakurta, "Differentially private online learning," in *25th Annual Conference on Learning Theory (COLT)*, 2012, pp. 24.1–24.34.
- [18] D. Kifer, A. Smith, and A. Thakurta, "Private convex empirical risk minimization and high-dimensional regression," in *25th Annual Conference on Learning Theory (COLT)*, 2012, pp. 25.1–25.40.
- [19] A. Smith and A. Thakurta, "Differentially private feature selection via stability arguments, and the robustness of the LASSO," in *Conference on Learning Theory (COLT)*, 2013, pp. 819–850.
- [20] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, 2013, pp. 429–438.
- [21] J. Ullman, "Private multiplicative weights beyond linear queries," in *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. ACM, 2015, pp. 303–312.
- [22] P. Jain and A. G. Thakurta, "(Near) dimension independent risk bounds for differentially private learning," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, vol. 32(1), 2014, pp. 476–484.
- [23] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization, revisited," in *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, 2014, pp. 464–473.
- [24] K. Talwar, A. G. Thakurta, and L. Zhang, "Nearly optimal private LASSO," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, 2015, pp. 3025–3033.
- [25] A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning?" in *IEEE Security & Privacy*, 2017, pp. 58–77.
- [26] X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton, "Bolt-on differential privacy for scalable stochastic gradient descent-based analytics," in *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD)*, 2017, pp. 1307–1322.
- [27] R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang, "Towards practical differentially private convex optimization," in *IEEE Security & Privacy*, 2019.
- [28] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54, 2017, pp. 1273–1282.
- [29] A. Rényi, "On measures of entropy and information," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1961, pp. 547–561.
- [30] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.
- [31] C. Dwork, "Differential privacy," in *International Colloquium on Automata, Languages, and Programming (ICALP)*, 2006, pp. 1–12.
- [32] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, Aug. 2014.
- [33] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," *CoRR*, vol. abs/1603.01887, 2016. [Online]. Available: <http://arxiv.org/abs/1603.01887>
- [34] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography—14th International Conference, TCC 2016-B, Part I*, 2016, pp. 635–658.
- [35] I. Mironov, "Rényi differential privacy," in *30th IEEE Computer Security Foundations Symposium (CSF)*, 2017, pp. 263–275.
- [36] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable private learning with PATE," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.
- [37] Y. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course*. Springer US, 2004.
- [38] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015. [Online]. Available: <http://dx.doi.org/10.1561/22000000050>
- [39] M. Bun, K. Nissim, and U. Stemmer, "Simultaneous private learning of multiple concepts," in *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, 2016, pp. 369–380.
- [40] V. Feldman, C. Guzman, and S. Vempala, "Statistical query algorithms for mean vector estimation and stochastic convex optimization," *CoRR*, vol. abs/1512.09170, 2015, extended abstract in SODA 2017. [Online]. Available: <http://arxiv.org/abs/1512.09170>