# Privacy Amplification by Iteration

Vitaly Feldman[1], Ilya Mironov[1], Kunal Talwar[1], and Abhradeep Thakurta[2]

[1]Google Brain, {vitalyfm,mironov,kunal}@google.com.
[2]UC Santa Cruz, aguhatha@ucsc.edu.

August 20, 2018

**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.

# 1 Introduction

Differential privacy [DMNS06] 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* [KLN$^+$08, BBKN14, WFS15, BDRS18, WBK18, ACG$^+$16]. In its basic form, for $\varepsilon = O(1)$, an $\varepsilon$-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 $\alpha$ (denoted $\mathrm{D}_\alpha$) 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 $(\varepsilon, \delta)$-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 3.3.

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

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 [DKM$^+$06]. Using amplification by sampling as in DP-SGD by Abadi et al. [ACG$^+$16] 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 $\varepsilon$ becomes $\approx q\varepsilon(1+\exp(\varepsilon))$ which is close to $2q\varepsilon$ for small $\varepsilon$ but grows quickly, and for instance, precludes setting $\varepsilon \geq 1/q$ for small $q$. Our main result applies for arbitrary $\sigma$ so that even if each $\sigma$ 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 [PAE+17, AKZ+17]. The public corpus can be based on opt-in population, such as a product's developers or early testers, data shared by volunteers [Chu05], or be released through a legal process [KY04].

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 $\sqrt{d}/\sqrt{n}$ [BST14a]. 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 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 dominated by the 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 2. We present our main technique in Section 3. Section 4 shows how this technique can be applied to versions of the noisy stochastic gradient descent algorithm. Finally, in Section 5, we apply this framework to derive the applications mentioned above.

## 1.1 Related Work

The field of differentially private convex optimization spans almost a decade [CM08, CMS11, JKT12, KST12, ST13, DJW13, Ull15, JT14, BST14b, TTZ15, STU17, WLK+17, INS+19]. Many of these results are optimal under different regimes such as empirical loss, population loss, the low-dimensional setting ($d \ll n$) or the high-dimensional setting $d \gg n$. Some of the algorithms (e.g., output perturbation [CMS11] and objective perturbation [CMS11, KST12]) 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 loss, 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 loss (see Table 1.1 for details).

In this table we also compare the local differential privacy of the algorithms [KLN+08]. 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 $\varepsilon_{\mathsf{local}}$ 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

| Algorithm | Excess loss for one task | Excess loss for $k \le n$ tasks | LDP ($\varepsilon_{\mathsf{local}}$) for one task |
|---|---|---|---|
| Noisy SGD + sampling$^{\dagger}$ [BST14b] | $\tilde{O}\left(\sqrt{\frac{d}{\varepsilon n}}\right)$ | $\tilde{O}\left(\sqrt{\frac{d}{\varepsilon n}}\right)$ | $\tilde{O}(n)$ |
| Noisy SGD [DJW13, STU17] | $\tilde{O}\left(\sqrt{\frac{d}{\varepsilon^2 n}}\right)$ | $\tilde{O}\left(\sqrt{\frac{dk}{\varepsilon^2 n}}\right)$ | $\varepsilon$ |
| Output perturbation$^{*}$ [CMS11, WLK$^{+}$17] | $\tilde{O}\left(\sqrt{\frac{d}{\varepsilon^2 n}}\right)$ | $\tilde{O}\left(\sqrt{\frac{dk}{\varepsilon^2 n}}\right)$ | $\infty$ |
| This work | $\tilde{O}\left(\sqrt{\frac{d}{\varepsilon^2 n}}\right)$ | $\tilde{O}\left(\sqrt{\frac{d}{\varepsilon^2 n}}\right)$ | $\varepsilon$ |

Table 1: The excess loss corresponds to the excess population loss. Comparison for a single pass over the dataset (i.e., at most $n$ gradient evaluations). For brevity, the table hides dependence on $\mathsf{poly}\ln(1/\delta)$. ($^{\dagger}$) This bound is not stated explicitly but can be derived by setting the parameters in [BST14b] appropriately. ($^{*}$) For output perturbation, we used the variant that can be implemented via SGD in a single pass [WLK$^{+}$17]. LDP stands for local differential privacy.

updates (which is a standard step in federated learning [MMR$^{+}$17]). In contrast, analyses based on secrecy of the sample (e.g., [KLN$^{+}$08, ACG$^{+}$16]) require that either data be stored centrally (thus eliminating local differential privacy guarantees) or all-to-all communications.

## 2 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.

### 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** ($L$-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 $w, w' \in \mathcal{K}$: $|f(w) - f(w)| \le L \left\|w - w'\right\|_2$.*

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

## 2.2 Probability Measures

In this work, we will primarily be interested in the $d$-dimensional Euclidean space $\mathbb{R}^d$ endowed with the $\ell_2$ metric and the Lebesgue measure. Our main result holds in a more general setting of Banach spaces.

We say a distribution $\mu$ is *absolutely continuous* with respect to $\nu$ if $\mu(A) = 0$ whenever $\nu(A) = 0$ for all measurable sets $A$. We will denote this by $\mu \ll \nu$.

Given two distributions $\mu$ and $\nu$ 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 [Rén61])**.** *Let $1 < \alpha < \infty$ and $\mu, \nu$ be measures with $\mu \ll \nu$. The Rényi divergence of order $\alpha$ between $\mu$ and $\nu$ 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 $\infty$. Rényi divergence of orders $\alpha = 1, \infty$ is defined by continuity.*

**Proposition 5** ([vEH14])**.** *The following hold for any $\alpha \in (1, \infty)$, and distributions $\mu, \mu', \nu, \nu'$:*

**Additivity:** $\mathrm{D}_\alpha(\mu \times \mu' \parallel \nu \times \nu') = \mathrm{D}_\alpha(\mu \parallel \nu) + \mathrm{D}_\alpha(\mu' \parallel \nu')$.

**Post-Processing:** *For any (deterministic) function $f$, $\mathrm{D}_\alpha(f(\mu) \parallel f(\nu)) \le \mathrm{D}_\alpha(\mu \parallel \nu)$, where we $f(\mu)$ denotes the distribution of $f(X)$ where $X \sim \mu$.*

As usual, we denote by $\mu * \nu$ the convolution of $\mu$ and $\nu$, 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** ($\infty$-Wasserstein Distance)**.** *The* $\infty$-Wasserstein distance *between distributions $\mu$ and $\nu$ on a Banach space $(\mathcal{Z}, \|\cdot\|)$ is defined as*

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

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

The following is immediate from the definition.

**Lemma 7.** *The following are equivalent for any distributions $\mu$, $\nu$ over $\mathcal{Z}$:*

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

Next we define a hybrid[1] between these two families of distances that plays a central role in our work.

[1] Here we use a *budgeted* version of the definition, putting a hard constraint on the $W_\infty$ portion of the distance, as it is most convenient for reasoning about differential privacy. A *Lagrangian* version of the definition may be more natural in other applications.

**Definition 8** (Shifted Rényi Divergence)**.** *Let $\mu$ and $\nu$ 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 $\mu$ and $\nu$ is defined as*

$$\mathrm{D}_\alpha^{(z)}(\mu \parallel \nu) \doteq \inf_{\mu' \colon W_\infty(\mu,\mu') \leq z} \mathrm{D}_\alpha\big(\mu' \,\big\|\, \nu\big).$$

The following follows from the definition:

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

**Monotonicity:** *For $0 \leq z \leq z'$, $\mathrm{D}_\alpha^{(z')}(\mu \parallel \nu) \leq \mathrm{D}_\alpha^{(z)}(\mu \parallel \nu)$.*

**Shifting:** *For any $x \in \mathcal{Z}$, $\mathrm{D}_\alpha^{(\|x\|)}(\mu \parallel \nu) \leq \mathrm{D}_\alpha(\mu * \mathbf{x} \parallel \nu)$, 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 $\zeta$ 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 $\alpha$ between $\zeta$ and the same distribution $\zeta$ shifted by a vector of length at most $a$:*

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

We denote the standard Gaussian distribution over $\mathbb{R}^d$ with variance $\sigma^2$ by $\mathcal{N}(0, \sigma^2 \mathbb{I}_d)$. By the well-known properties of Gaussians, for any $x \in \mathbb{R}^d$, and $\sigma$, $\mathrm{D}_\alpha\big(\mathcal{N}(0, \sigma^2 \mathbb{I}_d) \,\big\|\, \mathcal{N}(x, \sigma^2 \mathbb{I}_d)\big) = \alpha \|x\|_2^2 / 2\sigma^2$. 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 $\mu$ and $\nu$ respectively, we will often abuse notation and write $\mathrm{D}_\alpha(U \parallel V)$, $W_\infty(U, V)$ and $\mathrm{D}_\alpha^{(z)}(U \parallel V)$ to mean $\mathrm{D}_\alpha(\mu \parallel \nu)$, $W_\infty(\mu, \nu)$ and $\mathrm{D}_\alpha^{(z)}(\mu \parallel \nu)$, respectively.

## 2.3 (Rényi ) Differential Privacy

The notion of differential privacy (Definition 11) is by now a de facto standard for statistical data privacy [DMNS06, Dwo06, DR14]. 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 $(\varepsilon, \delta)$ quantify the amount of information leakage. A common choice of these parameters is $\varepsilon \approx 0.1$ and $\delta = 1/n^{\omega(1)}$, where $n$ refers to size of the dataset.

**Definition 11** ([DMNS06, DKM$^+$06])**.** *A randomized algorithm $\mathcal{A}$ is $(\varepsilon, \delta)$-differentially private ($(\varepsilon, \delta)$-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 $x_1, \ldots, x_n$ is said to satisfy $(\varepsilon, \delta)$-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*

$$\Pr[\mathcal{A}(x_1, \ldots, x_i, \ldots, x_n) \in \mathcal{O}] \leq e^\varepsilon \Pr[\mathcal{A}(x_1, \ldots, x_i', \ldots, x_n) \in \mathcal{O}] + \delta.$$

Another related model of privacy is local differential privacy [KLN+08]. 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 [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 $(\varepsilon, \delta)$-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]. In the current work, we will use the nomenclature based on the notion of the Rényi divergence (Definition 4).

**Definition 13** ([Mir17])**.** *For $1 \leq \alpha \leq \infty$ and $\varepsilon \geq 0$, a randomized algorithm $\mathcal{A}$ is* $(\alpha, \varepsilon)$-Rényi differentially private*, or $(\alpha, \varepsilon)$-RDP if for all neighboring data sets $S$ and $S'$ we have*

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

Per-person RDP can be defined in an analogous way. The following two lemmas [Mir17] allow translating Rényi differential privacy to $(\varepsilon, \delta)$-differential privacy, and give a composition rule for RDP.

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

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

**Lemma 15.** *If $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are randomized algorithms satisfying, respectively, $(\alpha, \varepsilon_1)$-RDP,…,$(\alpha, \varepsilon_k)$-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 16** (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 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 [Nes04]; for completeness, we give a proof in Appendix A.

**Proposition 18.** *Suppose that a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex and $\beta$-smooth. Then the function $\psi$ 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 $\psi_t\colon \mathcal{Z} \to \mathcal{Z}$, 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 $\zeta_{t+1}$. For brevity, we will denote the random variable output by this process after $T$ steps as $\mathrm{CNI}_T(X_0, \{\psi_t\}, \{\zeta_t\})$.*

# 3 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 $\zeta$ ensures that $X_1$ and $X_1'$ are $R_\alpha(\zeta, 1)$-close in $\alpha$-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 3.3.

## 3.1 The Shift-Reduction Lemma

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

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

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

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

*Proof.* Let $U$ be distributed as $\mu$ and $V$ as $\nu$. We first show the result for the case when $z = 0$. Let $\mu'$ be the distribution certifying $\mathrm{D}_\alpha^{(a)}(\mu \parallel \nu)$, that is $\mathrm{D}_\alpha(\mu' \parallel \nu) = \mathrm{D}_\alpha^{(a)}(\mu \parallel \nu)$ and $W_\infty(\mu, \mu') \leq a$. Let $(U, W)$

be the random variable whose existence is given by Lemma 7. That is, $\|W\| \le a$ with probability 1, $U \sim \mu$ and $U + W \sim \mu'$. Let $Y$ be an independent random variable distributed as $\zeta$. We can write

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

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

$$\begin{aligned}&\exp((\alpha-1)\mathrm{D}_\alpha((U+W,-W+Y) \parallel (V,Y)) \\ &\quad= \int\int \left(\frac{p_{(U+W,-W+Y)}(v,y)}{p_{(V,Y)}(v,y)}\right)^\alpha \zeta(y)\nu(v)\mathrm{d}y\mathrm{d}v \\ &\quad= \int\int \left(\frac{p_{U+W}(v)\cdot p_{-W+Y|U+W=v}(y)}{\nu(v)\cdot\zeta(y)}\right)^\alpha \zeta(y)\nu(v)\mathrm{d}y\mathrm{d}v \\ &\quad= \int \left(\frac{p_{U+W}(v)}{\nu(v)}\right)^\alpha \cdot \left(\int \left(\frac{p_{-W+Y|U+W=v}(y)}{\zeta(y)}\right)^\alpha \zeta(y)\mathrm{d}y\right)\nu(v)\mathrm{d}v \\ &\quad\le \int \left(\frac{p_{U+W}(v)}{\nu(v)}\right)^\alpha \nu(v)\mathrm{d}v \cdot \operatorname*{ess\,sup}_{(v',w)\sim p_{(U+W,W)}} \int \left(\frac{p_{-W+Y|W=w,U+W=v'}(y)}{\zeta(y)}\right)^\alpha \zeta(y)\mathrm{d}y \\ &\quad\le \int \left(\frac{\mu'(v)}{\nu(v)}\right)^\alpha \nu(v)\mathrm{d}v \cdot \operatorname*{ess\,sup}_{w\sim p_W} \int \left(\frac{p_{-W+Y|W=w}(y)}{\zeta(y)}\right)^\alpha \zeta(y)\mathrm{d}y \\ &\quad\le \exp((\alpha-1)\mathrm{D}_\alpha\big(\mu' \,\big\|\, \nu\big)) \cdot \exp\left((\alpha-1)R_\alpha(\zeta,a)\right). \qquad \text{(Proposition 5)}\end{aligned}$$

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

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

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

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 $\|W_1\| \le z$ and $\|W_2\| \le a$ with probability 1. We write

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

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

$$\begin{aligned}\mathrm{D}_\alpha^{(a)}(U + W_1 \parallel V) &\le \mathrm{D}_\alpha(U + W_1 + W_2 \parallel V) \\ &= \mathrm{D}_\alpha(U + W \parallel V) \\ &= \mathrm{D}_\alpha^{(z+a)}(U \parallel V).\end{aligned}$$

This completes the proof. □

### 3.2 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 $\psi$ and $\psi'$ are contractive maps on $(\mathcal{Z}, \|\cdot\|)$ and $\sup_x \|\psi(x) - \psi'(x)\| \le s$. Then for r.v.'s $X$ and $X'$ over $\mathcal{Z}$,*

$$\mathrm{D}_\alpha^{(z+s)}\big(\psi(X) \,\big\|\, \psi'(X')\big) \le \mathrm{D}_\alpha^{(z)}\big(X \,\big\|\, X'\big).$$

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

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

Thus $(\psi(X), \psi'(Y))$ is a coupling establishing the claimed upper bound on $\mathrm{D}_\alpha^{(z+s)}(\psi(X) \,\|\, \psi'(Y))$. □

### 3.3 Privacy Amplification by Iteration

We are now ready to prove our main result. We prove a general statement that can handle changes in several $\psi$'s; this enables us to easily analyze algorithms that access data points more than once[2]. Recall that $R_\alpha$ is introduced in Definition 10 and measures the maximal Rényi divergence of order $\alpha$ between a noise distribution and its shifted copy.

**Theorem 22.** *Let $X_T$ and $X'_T$ denote the output of CNI$_T(X_0, \{\psi_t\}, \{\zeta_t\})$ and CNI$_T(X_0, \{\psi'_t\}, \{\zeta_t\})$. 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 $z_t \doteq \sum_{i \le t} s_i - \sum_{i \le t} a_i$. If $z_t \ge 0$ for all $t$, then*

$$\mathrm{D}_\alpha^{(z_T)}\big(X_T \,\big\|\, X'_T\big) \le \sum_{t=1}^{T} R_\alpha(\zeta_t, a_t).$$

*In particular, if $z_T = 0$, then*

$$\mathrm{D}_\alpha\big(X_T \,\big\|\, X'_T\big) \le \sum_{t=1}^{T} R_\alpha(\zeta_t, a_t).$$

*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 CNI$(X_0, \{\psi_t\}, \{\zeta_t\})$ (resp., CNI$(X_0, \{\psi'_t\}, \{\zeta_t\})$. We argue that for all $t \le T$,

$$\mathrm{D}_\alpha^{(z_t)}\big(X_t \,\big\|\, X'_t\big) \le \sum_{i=1}^{t} R_\alpha(\zeta_i, a_i).$$

[2]Since 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 $\zeta_{t+1}$.

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

This completes the induction step and the proof. □

# 4 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 $S = (x_1, \ldots, x_n)$, starting point $w_0$, rate parameter $\eta$, and noise scale $\sigma$ 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 $\mathrm{PNSGD}(S, w_0, \eta, \sigma)$ and describe it formally in Algorithm 1.

**Algorithm 1** Projected noisy stochastic gradient descent (PNSGD)

**Input:** Data set $S = \{x_1, \ldots, x_n\}$, $f \colon \mathcal{K} \times \mathcal{X} \to \mathbb{R}$ a convex function in the first parameter, learning rate $\eta$, starting point $w_0 \in \mathcal{K}$, noise parameter $\sigma$.

1: **for** $t \in \{0, \ldots, n-1\}$ **do**
2: $\quad v_{t+1} \leftarrow w_t - \eta(\nabla_w f(w, x_{t+1}) + Z)$, where $Z \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$.
3: $\quad w_{t+1} \leftarrow \Pi_{\mathcal{K}}(v_{t+1})$, where $\Pi_{\mathcal{K}}(w) = \arg\min_{\theta \in \mathcal{K}} \|\theta - w\|_2$ is the $\ell_2$-projection on $\mathcal{K}$.
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 $\mathrm{PNSGD}(S, w_0, \eta, \sigma)$ 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 $\eta \cdot \sigma$. 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 $\beta$-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 $S \in \mathcal{X}^n$, PNSGD$(S, w_0, \eta, \sigma)$ satisfies $\left(\alpha, \frac{\alpha \cdot \varepsilon}{n+1-t}\right)$-RDP for its $t$'th input, where $\varepsilon = \frac{2L^2}{\sigma^2}$.*

*Proof.* Let $S \doteq (x_1, \ldots, x_n)$ and $S' \doteq (x_1, \ldots, x_{t-1}, x'_t, x_{t+1}, \ldots, x_n)$ be two arbitrary datasets that at index $t$. As discussed above, under the smoothness condition $\eta \leq 2/\beta$ the steps of 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

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

We can now apply Theorem 22 with $a_1, \ldots, a_{t-1} = 0$ and $a_t, \ldots, a_n = \frac{2\eta L}{n-t+1}$. 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

$$\mathrm{D}_\alpha\big(X_n \,\big\|\, X'_n\big) \leq \frac{\alpha}{2\eta^2\sigma^2} \sum_{i=1}^{n} a_t^2 \leq \frac{2\alpha L^2}{\sigma^2 \cdot (n-t+1)}$$

as claimed. □

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 Skip-PNSGD$(S, w_0, \eta, \sigma)$ as the algorithm that picks randomly and uniformly $t_0 \in \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ 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 Skip-PNSGD$(S, w_0, \eta, \sigma)$ are at least as good as those we gave for 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 $\beta$-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 $S \in \mathcal{X}^n$, Skip-PNSGD$(S, w_0, \eta, \sigma)$ satisfies $\left(\alpha, \frac{\alpha \cdot \varepsilon}{n+1-t}\right)$-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 Stop-PNSGD$(S, w_0, \eta, \sigma)$. To analyze this algorithm we will need to prove a weak[3] form of convexity for the Rényi divergence that might have other applications.

**Lemma 25.** *Let $\mu_1, \ldots, \mu_n$ and $\nu_1, \ldots, \nu_n$ be probability distributions over some domain $\mathcal{Z}$ such that for all $i \in [n]$, $\mathrm{D}_\alpha(\mu_i \,\|\, \nu_i) \leq c/(\alpha - 1)$ for some $c \in (0, 1]$. Let $\rho$ be a probability distribution over $[n]$ and denote by $\mu_\rho$ (or $\nu_\rho$) the probability distribution over $\mathcal{Z}$ obtained by sampling $i$ from $\rho$ and then outputting a random sample from $\mu_i$ (respectively, $\nu_i$). Then*

$$\mathrm{D}_\alpha(\mu_\rho \,\|\, \nu_\rho) \leq (1 + c) \cdot \mathop{\mathbb{E}}_{i \sim \rho} [\mathrm{D}_\alpha(\mu_i \,\|\, \nu_i)].$$

[3]The weakness here is the strong (if necessary) assumption that $\mathrm{D}_\alpha(p_i \,\|\, q_i) \leq c/(\alpha - 1)$ for some $c \leq 1$.

*Proof.* Let $\mu'_\rho$ (or $\nu'_\rho$) be the probability distribution over $[n] \times \mathcal{Z}$ obtained by sampling $i$ from $\rho$ and then sampling a random $x$ from $\mu_i$ (respectively, $\nu_i$) and outputting $(i, x)$. We can obtain $\mu_\rho$ from $\mu'_\rho$ by applying the function that removes the first coordinate and the same function applied to $\nu'_\rho$ gives $\nu_\rho$. Therefore, by the post-processing properties of the Rényi divergence, we obtain that $\mathrm{D}_\alpha(\mu_\rho \parallel \nu_\rho) \leq \mathrm{D}_\alpha\big(\mu'_\rho \,\big\|\, \nu'_\rho\big)$. Now observe that for every $i \in [n]$ and $x \in \mathcal{Z}$, $\mu'_\rho(i, x) = \rho(i) \cdot \mu_i(x)$. Therefore,

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

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

We can now state and prove the privacy guarantees for Stop-PNSGD$(\eta, \sigma)$.

**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 $\beta$-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$(S, w_0, \eta, \sigma)$ satisfies $\left(\alpha, \frac{4\alpha L^2 \cdot \ln n}{n\sigma^2}\right)$-RDP.*

*Proof.* Let $S \doteq (x_1, \ldots, x_n)$ and $S' \doteq (x_1, \ldots, x_{t-1}, x'_t, x_{t+1}, \ldots, x_n)$ 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$(S, w_0, \eta, \sigma)$ on $S$ after $T$ steps and analogously define $X'_T$ for Stop-PNSGD$(S', w_0, \eta, \sigma)$. If $t > T$ then the algorithm does not reach $x_t$ (or $x'_t$) and hence $\mathrm{D}_\alpha(X_T \parallel X'_T) = 0$. Otherwise, we can use Theorem 23 with $n = T$ to obtain that

$$\mathrm{D}_\alpha\big(X_T \,\big\|\, X'_T\big) \leq \frac{2\alpha L^2}{\sigma^2 \cdot (T - t + 1)}.$$

By definition, the output of Stop-PNSGD$(S, w_0, \eta, \sigma)$ 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

$$\mathrm{D}_\alpha\big(Y_T \,\big\|\, Y'_T\big) \le 2 \cdot \frac{1}{n} \sum_{T \in [n]} \mathrm{D}_\alpha\big(X_T \,\big\|\, X'_T\big) \le 2 \cdot \frac{1}{n} \sum_{T=t}^{n} \frac{2\alpha L^2}{\sigma^2 \cdot (T-t+1)}$$
$$\le \frac{4\alpha L^2 \cdot \ln(n-t+1)}{n\sigma^2} \le \frac{4\alpha L^2 \cdot \ln n}{n\sigma^2}.$$

□

In Appendix B, 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 [ACG+16], the approach here works in the distributed setting and leads to a significantly simpler proof.

Finally, we remark that PNSGD$(S, w_0, \eta, \sigma)$ and its variants described above satisfy local differential privacy (even without the smoothness assumption). Specifically,

**Lemma 27.** *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 functions over $\mathcal{K}$. Then, for every $\eta > 0, \alpha > 1$, starting point $w_0 \in \mathcal{K}$, and dataset $S \in \mathcal{X}^n$, (Stop/Skip)-PNSGD$(S, w_0, \eta, \sigma)$ satisfies local $\left(\alpha, \frac{2\alpha L^2}{\sigma^2}\right)$-RDP. In particular, for every $\varepsilon, \delta > 0$ and $\sigma = 2L\sqrt{2\ln(1.25/\delta)}/\varepsilon$ it satisfies local $(\varepsilon, \delta)$-DP.*

# 5 Applications

We now show how to use the algorithms we have analyzed to derive new results for privacy-preserving convex optimization. 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 [DKM+06], 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 [WLK+17] 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.

## 5.1 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., [Bub15]). 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 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 $t = 0, \ldots, T-1$.

**Theorem 28.** *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 $G$ 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 $w_1, \ldots, w_T$ denote the iterates produced by 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 PSGD$(G, w_0, \eta, T)$. As a result we get the following baseline guarantees for Stop-PNSGD$(S, w_0, \eta, \sigma)$ we defined in Section 4 (namely, these guarantees do not use our amplification analysis and do not require smoothness).

**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 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 $S \in \mathcal{X}^n$, Stop-PNSGD$(S, w_0, \eta, \sigma)$ satisfies local $(\varepsilon, \delta)$-DP. In addition, if $S$ consists of samples drawn i.i.d. from an arbitrary distribution $\mathcal{P}$ over $\mathcal{X}$, then*

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

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

*Proof.* By Lemma 27, setting $\sigma \doteq 2L\sqrt{2\ln(1.25/\delta)}/\varepsilon$ ensures local $(\varepsilon, \delta)$-DP. Now we observe that $G(w) = \nabla f(w, x) + Z$ where $x$ is drawn from $\mathcal{P}$ and $Z$ is drawn from $\mathcal{N}(0, \sigma^2\mathbb{I}_d)$ is an unbiased gradient oracle for $F(w)$. Further,

$$\mathbb{E}[\|G(w)\|^2] = \mathop{\mathbb{E}}_{x\sim\mathcal{P}}[\|\nabla f(w, x)\|^2] + d\sigma^2 \leq L^2 + \frac{8dL^2\ln(1.25/\delta)}{\varepsilon^2}.$$

Hence, we can apply Theorem 28 for $L_G = L\sqrt{1 + \frac{8d\ln(1.25/\delta)}{\varepsilon^2}}$ and $\eta = 2R/(L_G\sqrt{n})$ to obtain that

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

□

## 5.2 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 28.

**Theorem 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, $\beta$-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 $\eta = \sqrt{8}R/\sqrt{n(L^2 + d\sigma^2)} \leq 2/\beta$, then Skip-PNSGD$(S, w_0, \eta, \sigma)$ satisfies local $(\varepsilon, \delta)$-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*

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

*where $W$ denotes the output of 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 and Lemma 27. Let us denote by Stop$(n/2)$-PNSGD$(S, w_0, \eta, \sigma)$ the algorithm that runs PNSGD$(S, w_0, \eta, \sigma)$ with a randomly and uniformly chosen stopping time $T \in \{\lceil n/2 \rceil, \ldots, n\}$. Observe that the distribution of the output of Skip-PNSGD$(S, w_0, \eta, \sigma)$ on $S \sim \mathcal{P}^n$ is identical to the output distribution of Stop$(n/2)$-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). Stop$(n/2)$-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 Stop$(n/2)$-PNSGD are the same as those for Stop-PNSGD (Theorem 29) for a dataset consisting of $n/2$ points. □

### 5.3 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 31.** *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, $\beta$-smooth functions over $\mathcal{K}$. Let $S_{priv} \in \mathcal{X}^{n-m}$ and $S_{pub} \in \mathcal{X}^m$ be two datasets and $S \doteq (S_{priv}, S_{pub})$. For every $\varepsilon > 0$, $\delta > 0$, starting point $w_0 \in \mathcal{K}$, $\sigma = 2L\sqrt{\ln(1.25/\delta)/m}/\varepsilon$, if $\eta = \sqrt{8}R/\sqrt{n(L^2 + d\sigma^2)} \leq 2/\beta$, then Skip-PNSGD$(S, w_0, \eta, \sigma)$ satisfies $(\varepsilon, \delta)$-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*

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

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

### 5.4 Multiple Convex Optimizations

The privacy guarantees in Theorem 26 do not improve on the $(\varepsilon, \delta)$-DP guarantees for an individual task since in order to convert RDP guarantees to $(\varepsilon, \delta)$-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 loss as the one obtained in Theorem 29. When solving $k$ convex optimization tasks on the same

dataset, standard analysis requires increasing the noise scale $\sigma$ (and hence the bound on excess loss) by a factor of $\sqrt{k}$ to keep the same $(\varepsilon, \delta)$-DP level. In contrast, our analysis allows to bound $(\varepsilon, \delta)$-DP directly and only requires increasing $\sigma$ 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 [BNS16b]. The question of optimizing multiple loss functions was also studied in [Ull15, FGV15]. 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 32.** *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, $\beta$-smooth functions over $\mathcal{K}$ and $w_0 \in \mathcal{K}$ be a starting point. For $\varepsilon, \delta \in (0,1)$ let $q \doteq \max\left\{\frac{2k\ln n}{n}, 2\ln(1/\delta)\right\}$, $\sigma \doteq \frac{4L\sqrt{q\ln(1/\delta)}}{\varepsilon}$, $\eta \doteq 4R/\sqrt{n(L^2 + d\sigma^2)}$. For a dataset $S \in \mathcal{X}^n$ and $i \in [k]$, let $W_i$ denote the output of Stop-PNSGD$(S, w_0, \eta, \sigma)$ on the $i$'th family of functions (with independent randomness). Then the entire output $(W_1, \ldots, W_k)$ satisfies $(\varepsilon, \delta)$-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 $i$,*

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

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

*Proof.* By, the composition properties of RDP and Theorem 26, we have that the output of $k$ executions of Stop-PNSGD$(S, w_0, \eta, \sigma)$ satisfies $\left(\alpha, \frac{4k\alpha L^2 \cdot \ln n}{n\sigma^2}\right)$-RDP, whenever $\sigma \geq L\sqrt{2(\alpha-1)\alpha}$. We let $\alpha \doteq \frac{\sigma\sqrt{\ln(1/\delta)}}{L\sqrt{q}}$. Note that this ensures that

$$\sigma = \frac{\alpha L\sqrt{q}}{\sqrt{\ln(1/\delta)}} \geq \frac{\alpha L\sqrt{2\ln(1/\delta)}}{\sqrt{\ln(1/\delta)}} = \sqrt{2}\alpha L > L\sqrt{2(\alpha-1)\alpha}.$$

Note that for our choice of $\sigma = \frac{4L\sqrt{q\ln(1/\delta)}}{\varepsilon}$ we get that $\alpha = \frac{4\ln(1/\delta)}{\varepsilon} > 2$.
By Lemma 14, our bound on RDP implies $(\varepsilon', \delta)$-DP for

$$\varepsilon' = \frac{4k\alpha L^2 \cdot \ln n}{n\sigma^2} + \frac{\ln(1/\delta)}{\alpha - 1} \leq \frac{4k \cdot \ln(1/\delta)\ln n}{\alpha q n} + \frac{2\ln(1/\delta)}{\alpha} \leq \frac{2\ln(1/\delta)}{\alpha} + \frac{2\ln(1/\delta)}{\alpha} \leq \varepsilon.$$

Given the value of $\sigma$, we obtain the bound on the excess population loss from Theorem 28 in the same way as in the proof of Theorem 29. □

### 5.5 Removing the Smoothness Assumption

In this section we show that our assumption on smoothness of the loss function $f(w, x)$ can effectively be removed in several of our applications. We do this by convolving $f$ with the Gaussian distribution of an appropriate variance. While smoothing a non-smooth objective is a standard technique in optimization (for example [Nes05, DBW12]) we are not aware of bounds that are stated in the form we need. Specifically, we prove the following theorem in Appendix C.

**Theorem 33.** *Consider an L-Lipschitz convex function $f\colon \mathcal{K} \to \mathbb{R}$ defined over the convex set $\mathcal{K} \subseteq \mathbb{R}^d$. For every $\lambda > 0$, there exists a convex function $\hat{f}\colon \mathbb{R}^d \to \mathbb{R}$ with the following properties: i) $\hat{f}$ is convex, L-Lipschitz, and $L/\lambda$-smooth over $\mathcal{K}$, and ii) for all $w \in \mathcal{K}$, $|\hat{f}(w) - f(w)| \leq L\lambda\sqrt{d}$.*

In our Theorems 30 and 32 we use $\eta \leq \frac{R\varepsilon}{L\sqrt{n\ln(1/\delta)}}$. Therefore we need our smoothness parameter $\beta \leq 2\frac{L\sqrt{n\ln(1/\delta)}}{R\varepsilon}$. This means that it suffices to set $\lambda \doteq \frac{R\varepsilon}{2\sqrt{n\ln(1/\delta)}}$, which by Theorem 33, leads to approximation error of $\frac{LR\varepsilon\sqrt{d}}{2\sqrt{n\ln(1/\delta)}}$. Note that this additional error is dominated by the excess population loss whenever $\ln(1/\delta) \geq \varepsilon$ (which is typically the case). For completeness, we state the immediate corollary of Theorem 33 for per-person privacy formally.

**Corollary 34.** *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 $\mathcal{K}$. For every $\varepsilon > 0$, $\delta > 0$, starting point $w_0 \in \mathcal{K}$, $\sigma = 2L\sqrt{2\ln(1.25/\delta)}/\varepsilon$, dataset $S \in \mathcal{X}^n$, $\eta = \sqrt{8}R/\sqrt{n(L^2 + d\sigma^2)}$, and index $t \in [n]$, then Skip-PNSGD$(S, w_0, \eta, \sigma)$ executed on $\{f(\cdot, x)\}_{x\in\mathcal{X}}$ smoothed with $\lambda = \frac{R\varepsilon}{2\sqrt{n\ln(1/\delta)}}$ satisfies $(\varepsilon/\sqrt{n-t+1}, \delta)$-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*

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

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

We remark that in our application that uses public data (Corollary 31) the additional error introduced by general smoothing might no longer be dominated by bound on the excess population loss we prove. However, better smoothing techniques can be used for many important classes of functions. For example, generalized linear models can be smoothed with the smoothing error being on the same order as the statistical error (or $LR/\sqrt{n}$). Specifically, these are functions of the form $f(w) = \ell(\langle w, \theta\rangle, y)$ for some parameter $\theta \in \mathbb{R}^d$ and convex loss function $\ell\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. To smoothen such a function it suffices to convolve $\ell$ with a one-dimensional Gaussian kernel.

## Acknowledgements

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

## 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.

[AKZ+17] Brendan Avent, Aleksandra Korolova, David Zeber, Torgeir Hovden, and Benjamin Livshits. BLENDER: Enabling local search with a hybrid differential privacy model. In *Proceedings of the 26th USENIX Security Symposium*, pages 747–764, 2017.

[BBKN14] Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine learning*, 94(3):401–437, March 2014.

[BDRS18] Mark Bun, Cynthia Dwork, Guy N. Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated CDP. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 74–86, 2018.

[BNS+16a] Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059, 2016.

[BNS16b] Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 369–380, 2016.

[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.

[BST14a] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv preprint arXiv:1405.7085*, 2014.

[BST14b] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization, revisited. In *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 464–473, 2014.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[Chu05] George M. Church. The personal genome project. *Molecular systems biology*, 1(1), 2005.

[CM08] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS)*, pages 289–296, 2008.

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

[DBW12] John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM J. on Optimization*, 22(2):674–701, 2012.

[DFH+14] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.

[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*, pages 486–503, 2006.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pages 265–284, 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 *International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 1–12, 2006.

[FGV15] Vitaly Feldman, Cristobal Guzman, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. *CoRR*, abs/1512.09170, 2015. Extended abstract in SODA 2017.

[INS+19] Roger Iyengar, Joseph P. Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *IEEE Security & Privacy*, 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.

[JT14] Prateek Jain and Abhradeep Guha Thakurta. (Near) dimension independent risk bounds for differentially private learning. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32(1), pages 476–484, 2014.

[KLN+08] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 531–540, 2008.

[KST12] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *25th Annual Conference on Learning Theory (COLT)*, pages 25.1–25.40, 2012.

[KY04] Bryan Klimt and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer, 2004.

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

[MMR+17] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1273–1282, 2017.

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

[Nes05] Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, May 2005.

[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.

[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.

[TTZ15] Kunal Talwar, Abhradeep Guha 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.

[vEH14] Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014.

[WBK18] Yu-Xiang Wang, Borja Balle, and Shiva Kasiviswanathan. Subsampled Rényi differential privacy and analytical moments accountant. *arXiv preprint arXiv:1808.00087*, 2018.

[WFS15] Yu-Xiang Wang, Stephen E. Fienberg, and Alexander 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)*, pages 2493–2502, 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 *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD)*, pages 1307–1322, 2017.

# A Contractivity of GD for smooth functions

Contractivity of a Gradient Descent step for a smooth convex function is a well-known result in convex optimization (see, e.g., Nesterov [Nes04]). We reproduce a proof below.

**Proposition 18.** *Suppose that a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex, twice differentiable*[4]*, and $\beta$-smooth. Then the function $\psi$ defined as:*

$$\psi(x) \doteq w - \eta \nabla f(w)$$

*is contractive as long as $\eta < 2/\beta$.*

*Proof.* Let $w, w' \in \mathbb{R}^d$. We wish to show that

$$\|\psi(w) - \psi(w')\| \leq \|w - w'\|.$$

We write

$$\begin{aligned}\psi(w) - \psi(w') &= w - w' - \eta(\nabla f(w) - \nabla f(w') \\ &= w - w' + \eta(w - w')^\top \nabla^2 f(z) \\ &= (w - w')(\mathbb{I} - \eta \nabla^2 f(z)),\end{aligned}$$

for some $z$ on the line joining $w$ and $w'$. By smoothness and convexity, the Hessian has eigenvalues in $[0, \beta]$. Thus,

$$\|\psi(w) - \psi(w')\| \leq \|w - w'\| \|\mathbb{I} - \eta \nabla^2 f(z)\|.$$

Since $0 \preceq \nabla^2 f(z) \preceq \beta \mathbb{I}$, the claim follows. □

# B Analyzing Multiple-Epoch SGD

In this section, we show how our techniques can be used to prove privacy for a fixed-ordering version of a multiple-epoch SGD algorithm for minimizing a convex $\beta$-smooth loss function where $\eta \leq 2/\beta$. Formally, we consider the following algorithm:

**Algorithm 2** Projected noisy multiple-epoch stochastic gradient descent (PNMSGD)

**Input:** Data set $S = \{x_1, \ldots, x_n\}$, $f\colon \mathcal{K} \times \mathcal{X} \to \mathbb{R}$ a function convex in the first parameter, learning rate $\eta$, starting point $w_0 \in \mathcal{K}$, noise parameter $\sigma$.

1: **for** $j \in \{0, \ldots, n-1\}$ **do**
2:  **for** $i \in \{0, \ldots, n-1\}$ **do**
3:  $t \leftarrow nj + i$
4:  $v_{t+1} \leftarrow w_t - \eta(\nabla_w f(w, x_{i+1}) + Z)$, where $Z \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$.
5:  $w_{t+1} \leftarrow \Pi_{\mathcal{K}}(v_{t+1})$, where $\Pi_{\mathcal{K}}(w) = \arg\min_{\theta \in \mathcal{K}} \|\theta - w\|_2$ is the $\ell_2$-projection.
6: **return** the final iterate $w_{n^2}$.

An algorithm similar to this was analyzed by [BST14a] who used privacy amplification by sampling to prove that it satisfies $(\varepsilon, \delta)$-DP when $\sigma^2 = \frac{32L^2 \ln(n/\delta) \ln(1/\delta)}{\varepsilon^2}$. This analysis can be improved using the techniques of [ACG+16] to ensure that $\sigma^2 = \Theta\left(\frac{L^2 \ln(1/\delta)}{\varepsilon^2}\right)$ suffices for a suitable range of $\varepsilon$.

[4] This constraint makes the proof simpler but is technically unnecessary.

The privacy bound for Algorithm 2 follows in a rather straightforward way from Theorem 22. Let $S$ and $S'$ be two datasets that differ in the $i$th example. Algorithm 2 run on $S$ (resp. $S'$) defines a contractive noise iteration CNI$(X_0, \{\psi_t\}, \{\zeta_t\})$ (resp. CNI$(X_0, \{\psi'_t\}, \{\zeta_t\})$). Letting $s_t \doteq \sup_w \|\psi_t(w) - \psi'_t(w)\|$, we have

- $s_t = 0$ whenever $t \not\equiv i \pmod n$, since $\psi_t = \psi'_t$.
- $s_t \leq 2\eta L$ for $t \equiv i \pmod n$.

We set

$$a_t = \begin{cases} 0 & \text{if } t < i, \\ \frac{2\eta L}{n} & \text{if } i \leq t \leq n(n-1)+i, \\ \frac{2\eta L}{n-i} & \text{if } n(n-1)+i \leq t \leq n^2. \end{cases}$$

It is easy to check that for this definition, $z_t \geq 0$ and that $z_{n^2} = 0$. Applying Theorem 22 and noting that $\zeta_t = \mathcal{N}(0, (\eta\sigma)^2 \mathbb{I}_d)$ for all $t$, we get that

$$\begin{aligned} \mathrm{D}_\alpha\big(X_{n^2} \,\big\|\, X'_{n^2}\big) &\leq \frac{\alpha}{2\eta^2\sigma^2} \cdot \sum_t a_t^2 \\ &= \frac{2\alpha L^2}{\sigma^2} \cdot \left( \frac{n(n-1)}{n^2} + \frac{n-i}{(n-i)^2} \right) \\ &\leq \frac{8\alpha L^2}{\sigma^2}. \end{aligned}$$

It follows that Algorithm 2 satisfies $(\alpha, \frac{8\alpha L^2}{\sigma^2})$-RDP. Applying Lemma 14, with $\alpha = \frac{2\ln(1/\delta)}{\varepsilon}$, we conclude that the algorithm satisfies $(\varepsilon, \delta)$-DP for $\sigma = \frac{32L^2\ln(1/\delta)}{\varepsilon^2}$. Compared to the approach from [ACG+16], we have a significantly cleaner proof without needing any assumptions on $\sigma$. We remark that the two algorithms differ slightly. Here we fix an ordering and make $n$ passes over the data points in the same order, whereas the algorithm in [BST14a] takes $n^2$ steps, each on a uniformly random data point. To obtain utility guarantees for this algorithm one can appeal to standard regret bounds for online algorithms (e.g., [Bub15]). These bounds imply an upper bound on the empirical loss of the randomly chosen iterate. To obtain bounds on the population loss one can appeal to the generalization properties of differential privacy [DFH+14, BNS+16a].

# C Smoothing via Convolution with the Gaussian Kernel

*Proof of Theorem 33.* Consider the Gaussian kernel $\zeta = \mathcal{N}(0, \lambda^2 \mathbb{I}_d)$. Before convolving $f$ with this kernel we need to extend $f$ beyond $\mathcal{K}$. Let $h(w) \doteq \min_{v \in \mathcal{K}} f(v) + L\|w - v\|_2$, where $h(w)$ is defined over the complete $\mathbb{R}^d$. (The function $h$ is also called the convex Lipschitz extension of $f$.) We define the approximation to the function $f(w)$ as

$$\hat{f}(w) \doteq \mathop{\mathbb{E}}_{Z \sim \zeta} \left[ h(w + Z) \right].$$

The function $\hat{f}$ satisfies the following properties:

1. **Total on $\mathbb{R}^d$:** By definition, the function $\hat{f}$ is well-defined on $\mathbb{R}^d$.
2. **Lipschitzness and convexity:** Since the function $f(w)$ is convex and $L$-Lipschitz, the Lipschitz extension function $h(w)$ is also convex and $L$-Lipschitz. Hence, $\hat{f}(w)$ is both convex and $L$-Lipschitz as it is defined as a convolution of $h(w)$ with the Gaussian probability kernel.

3. **Smoothness:** $\hat{f}(w)$ is $L/\lambda$-smooth. For all $w, w' \in \mathbb{R}^d$,

$$\left\|\nabla \hat{f}(w) - \nabla \hat{f}(w')\right\|_2 \leq \frac{L}{\lambda} \left\|w - w'\right\|_2.$$

Let $p_Z$ denote the probability density function of the random variable $Z$. By definition,

$$\begin{aligned}
\left\|\nabla \hat{f}(w) - \nabla \hat{f}(w')\right\|_2 &= \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,\lambda^2\mathbb{I}_d)} [\partial h(w+Z)] - \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,\lambda^2\mathbb{I}_d)} \left[\partial h(w'+Z)\right] \right\|_2 \\
&= \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,\lambda^2\mathbb{I}_d)} [\partial h(w+Z)] - \mathop{\mathbb{E}}_{Z'\sim\mathcal{N}(w'-w,\lambda^2\mathbb{I}_d)} \left[\partial h(w+Z')\right] \right\|_2 \\
&= \left\| \int_z \partial h(w+z) \left(p_Z(z) - p_{Z'}(z)\right) \mathrm{d}z \right\|_2 \\
&\leq \sup_{w\in\mathbb{R}^d} \|\partial h(w)\|_2 \cdot \int_z |p_Z(z) - p_{Z'}(z)| \, \mathrm{d}z \\
&\leq L \cdot 2 \, \mathsf{TV}(p_Z, p_{Z'}),
\end{aligned}$$

where $\mathsf{TV}$ refers to the total variation distance. To complete the proof note that by Pinsker's inequality,

$$\mathsf{TV}\left(\mathcal{N}(0,\lambda^2\mathbb{I}_d), \mathcal{N}(w'-w,\lambda^2\mathbb{I}_d)\right) \leq \sqrt{\frac{1}{2}\mathrm{D}_1(\mathcal{N}(0,\lambda^2\mathbb{I}_d) \parallel \mathcal{N}(w'-w,\lambda^2\mathbb{I}_d))} \leq \frac{\|w-w'\|_2}{2\lambda}.$$

4. **Approximation error:** For all $w \in \mathcal{K}$, $\left|\hat{f}(w) - f(w)\right| \leq L\lambda\sqrt{d}$. By definition,

$$\begin{aligned}
\left|\hat{f}(w) - f(w)\right| &= \left| \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,\lambda^2\mathbb{I}_d)} [h(w+Z) - h(w)] \right| \\
&\leq \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,\lambda^2\mathbb{I}_d)} [|h(w+Z) - h(w)|] \\
&\leq L \cdot \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,\lambda^2\mathbb{I}_d)} [\|Z\|_2] \\
&= L\lambda\sqrt{d}.
\end{aligned}$$

□