Dimension Independent Generalization of DP-SGD for Overparameterized Smooth Convex Optimization

Yi-An Ma (yianma@ucsd.edu)
University of California, San Diego

Teodor Marinov (tvmarinov@google.com)
Google Research

Tong Zhang
Google Research and The Hong Kong University of Science and Technology

Abstract

This paper considers the generalization performance of differentially private convex learning. We demonstrate that the convergence analysis of Langevin algorithms can be used to obtain new generalization bounds with differential privacy guarantees for DP-SGD. More specifically, by using some recently obtained dimension-independent convergence results for stochastic Langevin algorithms with convex objective functions, we obtain \( O(n^{-1/4}) \) privacy guarantees for DP-SGD with the optimal excess generalization error of \( \tilde{O}(n^{-1/2}) \) for certain classes of overparameterized smooth convex optimization problems. This improves previous DP-SGD results for such problems that contain explicit dimension dependencies, so that the resulting generalization bounds become unsuitable for overparameterized models used in practical applications.

1 Introduction

We consider a set of i.i.d. training data \( S_n = \{z_1, \ldots, z_n\} \), each drawn from a distribution \( D \) on \( \mathcal{Z} \). Consider a loss function \( \ell(w, z) \) that is convex in \( w \), where \( w \in \mathbb{R}^d \) and \( z \in \mathcal{Z} \). In supervised learning, the goal is find an estimator \( \hat{w} \) from \( S_n \) to minimize the generalization error

\[
\mathbb{E}_{z \sim D} \ell(\hat{w}, z).
\]

A frequently used method is to use stochastic gradient descent (SGD) to approximately minimize the empirical risk on the training data:

\[
\hat{w} \approx \arg \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i).
\]

It can be shown using tools from statistical learning theory that the SGD method that aims to minimize the empirical risk in (2) also achieves a small generalization error in (1). In particular, it
can be shown that under suitable conditions on the loss function $\ell(w, z)$, the excess generalization error achieved by SGD, compared to an arbitrary $w \in \mathbb{R}^d$, is bounded as

$$E_{z \sim \mathcal{D}} \ell(\hat{w}, z) - E_{z \sim \mathcal{D}} \ell(w, z) = \tilde{O}(1/\sqrt{n}),$$

where the constant in $\tilde{O}(\cdot)$ depends on the norm $\|w\|_2$, properties of $\ell(w, z)$, and log factors.

However, the standard SGD method for solving (2) does not have good privacy guarantees. To overcome this problem, there have been significant interests in designing statistical estimators that are differentially private. That is, such estimators are insensitive to a modification of a single data point in $S_n$. In practice, the most widely used algorithm is DP-SGD, which adds a Gaussian noise in each SGD step. Mathematically this method is equivalent to the stochastic gradient Langevin algorithm (SGLD) for Monte Carlo Sampling [Welling and Teh, 2011]. The purpose of this paper is to demonstrate that the existing convergence analysis of Langevin algorithm can be leveraged to derive generalization bound of the DP-SGD method with good differential privacy guarantees. Using some recently obtained dimension-independent convergence results for Lagenvin algorithms, we establish differential privacy analysis of DP-SGD that improves existing bounds in terms of dimension dependency for certain classes of smooth convex optimization problems.

2 Related Work

There have been significant interests in developing machine learning algorithms with differential privacy guarantees. In practice, the most popular algorithm is DP-SGD, and its differential privacy property has been investigated in [Abadi et al., 2016]. However, it is known that the requirement of differential privacy guarantee may negatively affect the generalization performance of a machine learning algorithm. Therefore in practice, the trade-off between differential privacy and generalization of SGD becomes an important consideration. This issue was not considered in [Abadi et al., 2016].

To better understand the relationship between differential privacy and generalization, more recent works focused on the theoretical analysis of differentially private algorithms for solving stochastic convex optimization problems. In particular, it was shown that for generic smooth stochastic convex optimization without additional structures, one can achieve a $O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d} \log(1/\delta)}{\epsilon n}\right)$ excess generalization error bound for $(\epsilon, \delta)$-differential privacy (see Definition 1 below for the definition of DP) [Bassily et al., 2019]. The work of Bassily et al. [2014] show that the above rate is not improvable in general, when considering constrained optimization problems. Additional results showed that similar results can be extended to nonsmooth problems and for less computational costs [Feldman et al., 2020, Bassily et al., 2020]. Unfortunately, such a result means that in the overparameterized regime where $d \geq n$, any privacy guarantees at a level smaller than constant will lead to worse generalization results than that of SGD. Such results imply that DP-SGD significantly affects the performance of machine learning algorithms in the overparameterized regime.

The purpose of this paper is to show with additional structures that are naturally satisfied for classes of machine learning problems such as logistic regression, the pessimistic dimension dependent excess generalization lower bound in [Bassily et al., 2019] for smooth convex optimization problems can be avoided. In particular, we are able to obtain a dimension independent excess generalization bound at the optimal rate of $\tilde{O}(1/\sqrt{n})$ with a differential privacy guarantee at the level of $\epsilon = \tilde{O}(n^{-1/4})$. This implies that even for overparameterized models commonly used in practice, DP-
SGD can produce differentially private machine learning models without negatively affecting the generalization performance.

3 Main Results

**Algorithm 1:** DP-SGD (Stochastic Gradient Langevin Algorithm)

**Input:** decreasing sequence of stepsize $\eta_t$ and $\beta_0$

1. Draw $w_1$ from $N(0, \beta_0 I)$

2. for $t = 2, \ldots, T$ do

3. Sample mini-batch $\mathcal{M}_{t-1} \subseteq [n]$ (without replacement).

4. Construct mini-batched gradient $G_{t-1}(w_{t-1}) = \frac{1}{|\mathcal{M}_{t-1}|} \sum_{i \in \mathcal{M}_{t-1}} \nabla \ell(w_{t-1}, z_i)$. Let $\tilde{w}_{t-1} = w_{t-1} - \eta_t G_{t-1}(w_{t-1}).$ (3)

5. Sample $w_t \sim N((1 - \lambda_t \eta_t)\tilde{w}_{t-1}, \lambda_t \eta_t (2 - \lambda_t \eta_t) \beta_0 I).$ (4)

end

Mathematically, the DP-SGD algorithm corresponds to Algorithm 1, which is also referred to as Stochastic Gradient Langevin Dynamics (SGLD) in the MCMC sampling literature.

**Definition 1.** Let $S_n \in \mathcal{Z}^n$ be a dataset. A randomized learning algorithm $A$ maps $S_n$ to an output $w \in \Omega$. We say $A$ is $(\epsilon, \delta)$-DP (differentially private) if for any two datasets $S_n$ and $S'_n$ differing in at most one item, the following statement holds for all $E \subset \Omega$:

$$P(A(S_n) \in E) \leq e^\epsilon P(A(S'_n) \in E) + \delta.$$ 

We are interested in the trade-off of generalization, differential privacy, and computation in DP-SGD. We consider two settings, one is the single-pass setting, and the other is multi-pass setting. Before presenting the main results, we introduce the following notations and assumptions. Define

$$\ell(w, \mathcal{D}) = \mathbb{E}_{z \sim \mathcal{D}} \ell(w, z).$$

Moreover for any distribution $p$ on $\mathbb{R}^d$, define

$$\ell(p, \mathcal{D}) = \mathbb{E}_{w \sim p} \ell(w, \mathcal{D}).$$

We make the following assumptions.

**Assumption 1.** Each loss function $\ell(w, z)$ is convex in $w$.

**Assumption 2.** We assume that

$$\sup_{w,z} \| \nabla_w \ell(w, z) \|_2 \leq G, \quad \text{and} \quad \sup_w \nabla^2_w \ell(w, \mathcal{D}) \preceq \tilde{H}_\mathcal{D}.$$
Algorithm 1 is where $O$ samples to output $w_T$.

Further, Algorithm 1 does not use more than $1$ absolute constant.

Assumption 3 implies that we can carry out the convergence analysis in the simpler online setting as in the single-pass setting each $z_i$ is going to be observed only once and there will be a martingale structure which allows us to decouple $w_t$ from $G_t(w_t)$. In fact the size of the mini-batches only matters in the privacy analysis and for showing the generalization bound we could take $|M_t| = 1$.

**Single Pass Setting**

As we already mentioned, in the single pass setting we to through the data only once, that is we use $z_i \in S$ in a single gradient computation. We assume that the data size $n$ is known a priori. We have the following result.

**Theorem 1.** Assume that we run Algorithm 1 for $T = n$ iterations in the single pass setting with $\eta_t = \sqrt{\frac{1}{4T^2}}, |M_t| = \sqrt{\frac{T}{2(T-1)}}, \lambda_t = \frac{1}{\eta_t^n}, \beta_0 = \frac{(\epsilon+\log(1/\delta))\eta_0^2}{2\epsilon T}$, the last iterate, $w_T \sim P_T$, of Algorithm 1 is $(2\epsilon, \delta)$-DP and satisfies for all $w \in \mathbb{R}^d$:

$$\mathbb{E}[\ell(P_T, D)] \leq \ell(w, D) + \frac{\ln(n)(G\|w\|^2+1.5\eta_0^2G)}{\eta_0^2n^\alpha} + \frac{\eta_0^2\log(1/\delta)\text{trace}(H_D)}{\epsilon^2n}.$$ Further, Algorithm 1 does not use more than $n$ samples to output $w_T \sim P_T$.

**Multi Pass Setting**

Although single pass setting is sufficient for theoretical purposes, in practice, one often runs DP-SGD multiple times over the data. Another advantage of the multi-pass analysis is that the DP guarantee is with respect to the all intermediate iterates, as opposed to the single pass setting, and thus suitable for distributed computing when intermediate iterates need to be shared across multiple machines. Therefore we also include a multi-pass analysis for consistency with practice.

**Theorem 2.** Let $\eta_t = G^{-1}\sqrt{n}/\sqrt{8\ln(2.5t^2/n\delta)}$ and $|M_{t-1}| = 1$ for $t \geq 1$, $\lambda_1 = 1/\eta_1$ and $\lambda_t = \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}$ for $t \geq 2$. Assume $T = n^\alpha \epsilon^2 \geq n$ for $1 \leq \alpha \leq 2$ and $\beta_0 = \eta_0^2(n/T)$ for some $\eta_0 > 0$. Then Algorithm 1 is $(3\epsilon\sqrt{\ln(2/\delta)}+3\epsilon^2, \delta)$-DP with respect to all intermediate iterates $\{w_t\}$. Moreover, the following bound on the average generalization error holds for all $w \in \mathbb{R}^d$:

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \ell(p_t, D)\right] \leq \ell(w, D) + O\left(\frac{G\|w\|^2\sqrt{T/\delta}}{\eta_0^2n^\alpha} + \frac{\eta_0G}{\sqrt{n}\ln(1/\delta)} + \frac{\eta_0^2\text{trace}(H_D)}{\epsilon^2n^{\alpha-1}}\right),$$

where $O(\cdot)$ hides an absolute constant.

Making different choices of $T = n^\alpha \epsilon^2$ will lead to different generalization error with the same $(3\epsilon\sqrt{\ln(2/\delta)}+3\epsilon^2, \delta)$ differential privacy. Assume that trace$(H_D)$ is bounded, then we take $\epsilon = n^{-1/4}$ and $\alpha = 2$. This implies $(\epsilon, \delta)$-DP with $T = n^{1.5}$ computation, and excess generalization error of $O(1/\sqrt{n})$.

We also note that the results of Theorem 1 and Theorem 2 are comparable up to logarithmic factors. Although we are not concerned with logarithmic factors in this paper, we note that the logarithmic factor in Theorem 1 can be improved by using the step-size schedule in [Jain et al., 2019], and the logarithmic factor in Theorem 2 can be improved by using the strong composition analysis with RDP instead of with DP as in this paper [Abadi et al., 2016, Mironov, 2017].
3.1 Example: Ridge separable problems

We note that in the general setting where \( \text{trace}(\tilde{H}_D) \) depends on the dimensionality \( d \), we can obtain the following generalization performance for DP-SGD

\[
\ell(w, D) + O \left( \frac{G\|w\|_2}{\sqrt{n}} + \frac{G\|w\|_2\sqrt{d \log(1/\delta)}}{en} \right),
\]

which was shown by [Feldman et al., 2020]. However, in many applications, the trace of the Hessian \( \tilde{H}_D \) could be constant or much smaller compared to \( d \). We now give one such example.

We consider the binary logistic regression (or smoothed SVM losses) as a concrete example, where

\[
\ell(w, z) = \phi(w^\top x, y),
\]

and \( z = (x, y) \). Assume \( \|x\|_2 \leq 1 \) for all \( x \). If \( \phi'(a, y) \leq \gamma_1 \) and \( \phi''(a, y) \leq \gamma_2 \), then

\[
G \leq \gamma_1, \quad \text{trace}(\tilde{H}_D) \leq \gamma_2.
\]

Theorem \( \[ \) and Theorem \( \[ \) now imply an excess generalization error of at most \( \tilde{O}(n^{-1/2}) \) for any setting of \( \epsilon = \Omega(n^{-1/4}) \).

In comparison, for overparameterized problems with \( d \geq n \), we have excess generalization error of

\[
\tilde{O} \left( \frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{en} \right) = \tilde{O} \left( \frac{1}{\epsilon \sqrt{n}} \right)
\]

at the level of \( (\epsilon, \delta) \) differential privacy from earlier results. Such results are inferior to our result for overparameterized models used in practice. Our result is also closely related to those of [Song et al., 2021], which also considered the problem of deriving dimensional independent results. Their results employ the rank of the Hessian instead of the trace of the Hessian as in this paper. In applications, rank may be as large as \( O(n) \), while trace is generally small. We note that the results in [Feldman et al., 2020, Song et al., 2021] with which we compare hold for the more general setting of constrained convex optimization, where \( w \in \mathcal{P} \) for some convex set \( \mathcal{P} \) describing the constraints. The results presented in this work hold only in the unconstrained setting. Obtaining bounds which depend on the trace of the Hessian in the constrained setting is an interesting open problem.

4 Analysis of DP-SGD

Before presenting our analysis, we introduce the following additional notations. Define

\[
\ell(w, S_n) = \frac{1}{n} \sum_{i=1}^n \ell(w, z_i).
\]

Moreover for any distribution \( p \) on \( \mathbb{R}^d \), define

\[
\ell(p, S_n) = \mathbb{E}_{w \sim p} \ell(w, S_n).
\]

Define for any distribution \( q \) on \( \mathbb{R}^d \): \( H(p\|q) = \mathbb{E}_{w \sim p} \ln \frac{p(w)}{q(w)} \).

We divide the main steps in our analysis into the convergence of the DP-SGD procedure, generalization analysis, and DP-analysis. We then put these results together to prove the main results of the paper. The following summarizes the key intermediate steps. Detailed proofs of the required technical results can be found in Section 5.
4.1 Proof of Theorem 2

Convergence Analysis

The main convergence result for Algorithm 1, stated in the following proposition, is adapted from [Freund et al., 2021].

Proposition 1. Let \( p_t(w) \) denote the distribution of \( w_t \), and let \( q(w) = N(0, \beta_0 I) \), then

\[
\sum_{t=1}^{T} \left[ \ell(p_t, z_t) + \beta_t H(p_t \| q) \right] \leq \inf_{p} \sum_{t=1}^{T} \left[ \ell(p, z_t) + \beta_t H(p \| q) \right] + \sum_{t=1}^{T} \frac{\eta t}{2} G^2,
\]

where we set \( \lambda_1 = 1/\eta_1; \forall t \geq 1, \lambda_t \geq \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \); and \( \beta_t = \lambda_t \beta_0 \).

Bounding the term \( \inf_{p} \sum_{t=1}^{T} [\ell(p, z_t) + \beta_t H(p \| q)] \) results in the following corollary.

Corollary 1. For Algorithm 1, we have the following regret bound for any fixed \( w \in \mathbb{R}^d \):

\[
\mathbb{E} \sum_{t=1}^{T} \left[ \ell(p_t, S_n) - \ell(w, S_n) \right] \leq \frac{\lambda_T \| w \|^2}{2} + \tilde{\lambda}_T \beta_0 \ln \left( \frac{\tilde{\lambda}_T H_D + I}{\eta} \right) + \sum_{t=1}^{T} \frac{\eta t}{2} G^2,
\]

where the expectation is with respect to the randomization in the algorithm, and \( \tilde{\lambda}_T = \sum_{t=1}^{T} \lambda_t \).

Generalization Analysis

In order to obtain optimal trade-off of generalization and differential privacy in the setting when \( |\mathcal{M}| = 1, \forall t \in [T] \), we need to run DP-SGD more than once over the training dataset. In this situation, the standard online learning analysis doesn’t apply. It is also difficult to employ standard empirical process theory to analyze such systems in the general setting due to the lack of general covering number results. In this section, we derive generalization bounds using the stability analysis for SGD of [Hardt et al., 2016], which was also employed by [Bassily et al., 2020] to analyze DP-SGD. For simplicity, we only consider the expected version, with proofs included for completeness.

Proposition 2. Given any fixed \( w \) on \( \mathbb{R}^d \), we have for all \( t \):

\[
\mathbb{E} \ell(p_t, D) \leq \ell(w, D) + \mathbb{E} \left[ \ell(p_t, S_n) - \ell(w, S_n) \right] + \frac{2G^2}{n} \sqrt{(n + t) \sum_{s \leq t} \eta_s^2},
\]

We can now combine Proposition 2 and Corollary 1 to obtain the following generalization result.

Corollary 2. For Algorithm 1, we have the following generalization bound:

\[
\mathbb{E} \frac{1}{T} \sum_{t=1}^{T} \ell(p_t, D) \leq \inf_{w} \left[ \ell(w, D) + \frac{\| w \|^2}{2T \eta_T} \right] + \frac{\beta_0}{T \eta_T} \ln \left( \frac{\eta T H_D + I}{\eta} \right) + \sum_{t=1}^{T} \frac{\eta t}{2} G^2 + \frac{2G^2}{n} \sqrt{(n + T) \sum_{t \leq T} \eta_t^2}.
\]
Our analysis follows that of [Bassily et al., 2014]. Although similar to [Abadi et al., 2016, Bassily et al., 2020], we can remove a \( \ln(1/\delta) \) factor by using the Rényi entropy approach of [Mironov, 2017]. We consider a simpler derivation below using the strong composition theorem from [Dwork et al., 2010], together with DP-guarantee of Gaussian mechanism, and amplification theorem of [Balle et al., 2018] to handle SGD.

**Proposition 3.** If we take \( \eta_t = G^{-1} \sqrt{\beta_0/\sqrt{8 \ln(2.5t^2/n\delta)}}, \) then Algorithm 1, using sampling with replacement and setting \( |M_t| = 1 \), is
\[
\left( \sqrt{2T \ln(2/\delta)} e/n + Te^2/n^2, \delta \right)
\]
differentially private.

**Proof of Theorem 2**

Taking \( \eta_t \) as in Proposition 3 implies the DP part of the result. Further, combining this choice with Corollary 2 implies a generalization bound of (assuming \( T \geq n \)):
\[
\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} \ell(p_t, D) \right] \leq \ell(w, D) + \frac{||w||_2^2}{2T\eta T} + \frac{\beta_0}{2T\eta T} \ln \left| \eta T \tilde{H}_D + I \right| + \sum_{t=1}^{T} \frac{\eta G^2}{2n} + \frac{2G^2}{n} \sqrt{(n + T) \sum_{t \leq T} \eta_t^2}
\]
\[
= \ell(w, D) + O \left( \frac{G||w||_2^2 \sqrt{\ln(T/\delta)}}{2\sqrt{T\beta_0}} + \beta_0 \text{trace}(\tilde{H}_D) + \frac{G\sqrt{\beta_0}T}{n\sqrt{\ln(1/\delta)}} \right),
\]
where we have used the inequality \( \ln |I + A| \leq \text{trace}(A) \), and \( O(\cdot) \) hides an absolute constant. This bound implies Theorem 2.

### 4.2 Proof of Theorem 1

**Convergence and Generalization Analysis**

In the single-pass mini-batched setting we can obtain a simple generalization guarantee by taking expectation for the averaged iterate and using Corollary 1. The average iterate, however, may not enjoy good privacy properties. Following [Shamir and Zhang, 2013] it is possible to show a last iterate convergence guarantee stated below.

**Corollary 3.** For the single-pass setting in Algorithm 1 with fixed step-size \( \eta_t = \eta, \lambda_t = \frac{1}{T\eta}, \forall t \in [T] \) it holds that
\[
\mathbb{E} \left[ \ell(p_T, D) \right] - \ell(w, D) \leq \frac{\ln(T)||w||_2^2}{T\eta} + \beta_0 \text{trace}(\tilde{H}_D) + \frac{3\ln(T)}{2} - \eta G^2,
\]
for any fixed \( w \in \mathbb{R}^d \). Further, if the mini-batch size is set as \( |M_t| = \left\lceil \sqrt{\frac{T}{2(T-t+1)}} \right\rceil \), then the number of samples used by Algorithm 1 is at most \( T \).
DP Analysis

In the single-pass setting our analysis follows that of [Feldman et al., 2020]. To state our result formally we first need to introduce the notion of Rényi differential privacy.

**Definition 2.** Let \( \mu \) and \( \nu \) be two measures on \( \mathcal{X} \) which are absolutely continuous with respect to the Lebesgue measure with densities \( f_\mu \) and \( f_\nu \). The \( \alpha \)-Rényi divergence, for \( 1 < \alpha \leq \infty \) between \( \mu \) and \( \nu \) is defined as \( D_\alpha(\mu \| \nu) = \frac{1}{\alpha-1} \log \int_\mathcal{X} \left( \frac{f_\mu(x)}{f_\nu(x)} \right)^\alpha f_\mu dx \), if \( \mu \ll \nu \) and \( D_\alpha(\mu \| \nu) = \infty \) otherwise.

**Definition 3 ([Mironov, 2017]).** Let \( \mathcal{A} : \mathcal{D} \to \mathbb{R} \) be a randomized algorithm and let \( \mathcal{P}_A : \mathcal{D} \to \mathbb{P}(\mathbb{R}) \) be the induced probability measure by \( \mathcal{A} \). \( \mathcal{A} \) is (\( \alpha, \epsilon \))-Rényi differentially private (RDP) if for any two neighboring datasets \( \mathcal{D}, \mathcal{D}' \in \mathcal{D} \) it holds that \( D_\alpha(\mathcal{P}_A(\mathcal{D}) \| \mathcal{P}_A(\mathcal{D}')) \leq \epsilon \).

**Proposition 4.** Algorithm \( \Pi \) in the single-pass setting with \( \eta_t = \eta \), \( \lambda_t = \frac{1}{t\eta} \), and \( |\mathcal{M}_t| = \left\lfloor \sqrt{\frac{T}{2(1-\epsilon t+1)}} \right\rfloor \) is (\( \alpha, \epsilon \))-RDP with \( \epsilon = 2\frac{\alpha^2G^2}{\beta_0} \), for the last iterate.

**Proof of Theorem \( \Pi \)**

Fix any \( w \in \mathbb{R}^d \). Using the results in Corollary \( \beta \) and Proposition \( \Pi \) with \( |\mathcal{M}_t| = \sqrt{\frac{T}{2(1-\epsilon t+1)}} \), \( \beta_0 = \frac{\alpha^2}{2\epsilon^2} \), \( \eta = \eta_0 \sqrt{\frac{1}{4\epsilon^2}} \), and \( \lambda_t = \frac{1}{t\eta} \) implies Algorithm \( \Pi \) is (\( \alpha, \epsilon \))-RDP for the last iterate and satisfies

\[
\mathbb{E}[\ell(P_T, \mathcal{D})] - \ell(w, \mathcal{D}) \leq \frac{\ln(T)\left(\frac{|w|\|G\|}{\eta_0} + \frac{3}{2}\eta_0 G \right)}{\sqrt{T}} + \frac{\alpha \eta_0^2 \text{trace}(H_D)}{2\epsilon T}.
\]

The following lemma allows us to convert the RDP result into a differential privacy guarantee.

**Lemma 1 ([Mironov, 2017]).** If an algorithm \( \mathcal{A} \) is (\( \alpha, \epsilon \))-RDP then for any \( \delta > 0 \) it satisfies \( (\epsilon + \frac{\log(1/\delta)}{\alpha-1}, \delta) \) differential privacy (DP).

Setting \( \alpha = 1 + \frac{\log(1/\delta)}{\epsilon} \), we obtain the desired result.

5 Proofs of Technical Lemmas and Propositions

5.1 Convergence

**Proof of Proposition \( \Pi \)** Define: \( H(p) = \mathbb{E}_{w \sim p} \ln p(w) \). We also let \( g_t(w) = \frac{M}{2} \|w\|_2^2 \). We additionally use the following definition.

**Definition 4.** Given two probability distributions \( p(u) \) and \( p'(v) \) on \( \mathbb{R}^d \), and let \( \Pi(p, p') \) be the class of distributions \( q(u, v) \) on \( \mathbb{R}^d \times \mathbb{R}^d \) so that the marginals \( q(u) = p(u) \) and \( q(v) = p'(v) \). The \( W_2 \) Wasserstein distance of \( p \) and \( p' \) is defined as

\[
W_2(p, p')^2 = \min_{q \in \Pi(p, p')} \mathbb{E}_{(u, v) \sim q} \|u - v\|_2^2.
\]

Then we have the following two lemmas from [Freund et al., 2021].
Lemma 2 (For Regularized Entropy). We have for $t \geq 1$:

$$2\eta_t \left[ \mathbb{E}_{w \sim p_t} \ell(w) + \beta_t H(p_t) \right] - \mathbb{E}_{w \sim p_t} g_t(w) + \beta_t H(p_t) \right]$$

$$\leq \left[ \left( 1 - \lambda t \eta_t \right) W_2(\tilde{p}_{t-1}, p)^2 - W_2(p_t, p)^2 \right],$$

where $\tilde{p}_{t-1}$ is the distribution of $\tilde{w}_{t-1}$. Note that $1 - \lambda t \eta_t = 0$, so we do not have to define $\tilde{p}_0$.

Lemma 3. We have for all $t \geq 1$:

$$2\eta_t \left[ \mathbb{E}_{w \sim p_t} \ell(w, z_t) - \mathbb{E}_{w \sim p_t} \ell(w, z_t) \right]$$

$$\leq \left[ W_2(p_t, p)^2 - W_2(\tilde{p}_t, p)^2 \right] + \eta_t^2 \mathbb{E}_{w \sim p_t} \left\| \nabla \ell_t(w) \right\|_2^2$$

$$\leq \left[ W_2(p_t, p)^2 - W_2(\tilde{p}_t, p)^2 \right] + \eta_t^2 G^2.$$

By summing the inequalities in Lemma 2 and Lemma 3 and divide the resulting inequality by $\eta_t$, we obtain

$$2\mathbb{E}_{w \sim p_t} \left[ \ell(w, z_t) + g_t(w) + \beta_t \ln p_t(w) \right] - 2\mathbb{E}_{w \sim p_t} \left[ \ell(w, z_t) + g_t(w) + \beta_t \ln p(w) \right]$$

$$\leq \frac{1}{\eta_t - 1} W_2(\tilde{p}_{t-1}, p)^2 - \frac{1}{\eta_t} W_2(\tilde{p}_t, p)^2 + \eta_t G^2.$$

We can now sum over $t = 1$ to $t = T$:

$$\sum_{t=1}^{T} \mathbb{E}_{w \sim p_t} \left[ \ell(w, z_t) + \beta_t \ln \frac{p_t(w)}{q(w)} \right] \leq \inf_p \mathbb{E}_{w \sim p} \sum_{t=2}^{T} \left[ \ell(w, z_t) + \beta_t \ln \frac{p(w)}{q(w)} \right] + \sum_{t=1}^{T} \eta_t G^2.$$

Proof of Corollary 1. We employ the following estimate.

Lemma 4. Let $\sum_{t=1}^{T} \beta_t = \tilde{\lambda} T \beta_0$, and let $Q_T(w) = \sum_{t=1}^{T} \ell(w, S_n) + \frac{\tilde{\lambda}}{2} \left\| w \right\|_2^2$. We have

$$\mathbb{E} \inf_p \mathbb{E}_{w \sim p} \sum_{t=1}^{T} \left[ \ell(w, z_t) + \beta_t \ln \frac{p(w)}{q(w)} \right] \leq \inf_w \mathbb{E} Q_T(w) + \tilde{\lambda} T \beta_0 \mathbb{E} \left[ \frac{T}{\lambda T} \tilde{H} + I \right],$$

where the expectation is with respect to the randomization in the SGLD steps.

Proof. Let $w^* = \arg \min_w Q_T(w)$. Combining the smoothness assumption and the fact that we can upper bound the Hessian of $Q_T$ as $\tilde{H} + \tilde{\lambda} T I$ we have

$$Q_T(w) \leq Q_T(w^*) + \frac{1}{2} (w - w^*)^\top \tilde{H} (w - w^*) + \frac{\tilde{\lambda}}{2} \left\| w - w^* \right\|_2^2.$$

(5)
Therefore
\[
\inf_p \mathbb{E}_{w \sim p} \sum_{t=1}^{T} \left[ \ell(w, z_t) + \beta_t \ln \frac{p(w)}{q(w)} \right]
= -\tilde{\lambda}_T \beta_0 \ln \mathbb{E}_{w \sim q} \exp \left( -\frac{1}{\tilde{\lambda}_T \beta_0} \sum_{t=1}^{T} \ell(w, z_t) \right)
= -\tilde{\lambda}_T \beta_0 \ln \int \frac{1}{(2\pi \beta_0)^{d/2}} \exp \left( -\frac{1}{2\beta_0} Q_T(w) \right) dw
\leq Q_T(w^*) - \tilde{\lambda}_T \beta_0 \ln \int \frac{1}{(2\pi \beta_0)^{d/2}} \exp \left( -\frac{1}{2\beta_0} \sum_{t=1}^{T} (w - w^*)^\top \tilde{H}_D (w - w^*) - \frac{1}{2\beta_0} \|w - w^*\|_2^2 \right) dw
\leq Q_T(w^*) + \tilde{\lambda}_T \beta_0 \ln \left\| \frac{T}{\tilde{\lambda}_T} \tilde{H}_D + I \right\|.
\]
The inequality is a consequence of (5).

We can plug Lemma 4 into the right hand side of Proposition 1 and note that
\[
\beta_t \mathbb{E}_{w \sim p_t} \ln \frac{p_t(w)}{q(w)} \geq 0.
\]
This implies Corollary 1.

5.1.1 Last iterate convergence

Proof of Corollary 3. We proceed as in [Shamir and Zhang, 2013]. Let \( f(p) = \mathbb{E}_{w \sim p, z_t \sim D} [\ell(w, z_t) + g_t(w)] + \beta_0 H(p) \). Lemma 2 and Lemma 3 imply
\[
\sum_{t=T-k}^{T} f(p_t) - (k + 1) f(p) \leq \mathbb{E} \left[ \frac{W^2(\tilde{p}_{T-k-1}, p)}{2\eta} \right] + (k + 1) G^2 \frac{\eta}{2},
\]
for any \( p \) measurable with respect to the sigma-algebra induced by the randomness of the updates in the first \( T - k \) rounds. We set \( p = p_{T-k-1} \) which implies
\[
-f(p_{T-k-1}) \leq \frac{1}{k+1} \left( \mathbb{E} \left[ \frac{W^2(\tilde{p}_{T-k-1}, p_{T-k-1})}{2\eta} \right] + (k + 1) G^2 \frac{\eta}{2} - \sum_{t=T-k}^{T} f(p_t) \right).
\]
Let \( \mathcal{S}_k = \frac{1}{k+1} \sum_{t=T-k}^{T} f(p_t) \). Then we have
\[
(k + 1) \mathcal{S}_k = \sum_{t=T-k}^{T} f(p_t) = \sum_{t=T-k}^{T} f(p_t) - f(p_{T-k-1}) = (k + 2) \mathcal{S}_{k+1} - f(p_{T-k-1})
\leq (k + 2) \mathcal{S}_{k+1} - \mathcal{S}_k + \mathbb{E} \left[ \frac{W^2(\tilde{p}_{T-k-1}, p_{T-k-1})}{2\eta(k+1)} \right] + G^2 \frac{\eta}{2}
\leq \mathcal{S}_k \leq \mathcal{S}_{k+1} + \mathbb{E} \left[ \frac{W^2(\tilde{p}_{T-k-1}, p_{T-k-1})}{2\eta(k+1)(k+2)} \right] + G^2 \frac{\eta}{2(k+2)}
\leq f(p_T) = \mathcal{S}_0 \leq \mathcal{S}_T + \sum_{k=0}^{T-1} \mathbb{E} \left[ \frac{W^2(\tilde{p}_{T-k-1}, p_{T-k-1})}{2\eta(k+1)^2} \right] + \eta G^2 \ln(T) + \frac{G^2}{2}.
Finally we bound $W_2^2(p_t, p_t') \leq \mathbb{E}_{\tilde{w}_t \sim \tilde{p}_t} [||\tilde{w}_t - \eta \nabla \ell(\tilde{w}_t, z_t) - \tilde{w}_t||^2] \leq \eta^2 G^2$. Combining with our bound on $\mathcal{G}_T$ from Corollary [1] we have that for any $w$ it holds that

$$\mathbb{E}[\mathbb{E}_{w_{PT}} [\ell(w_{PT}, D) + g_T(w_{PT})]] - \ell(w, D) - g_T(w) + \beta_0 H(p_T \| q) \leq \frac{\tilde{\lambda}_T \| w \|^2}{2T} + \frac{\tilde{\lambda}_T \beta_0}{T} \ln \left| \frac{T}{\tilde{\lambda}_T} \tilde{H}_D + 1 \right| + \frac{3}{2} \eta G^2 \ln(T).$$

Using the setting $\lambda_t = \frac{1}{\eta}$ we can bound $\tilde{\lambda}_T \leq \frac{\ln(T)}{\eta}$. Together with the fact $\beta_0 H(p_T \| q) \geq 0$ we have

$$\mathbb{E} [\ell(p_T, D)] - \ell(w, D) \leq \frac{\ln(T)\| w \|^2}{T \eta} + \beta_0 \text{trace}(\tilde{H}_D) + \frac{3 \ln(T)}{2} \eta G^2.\qed$$

5.2 Generalization

Proof of Proposition [2] We need the following lemma for the stability of SGLD.

Lemma 5. Consider $S_n = \{z_1, \ldots, z_n\}$, and $S'_n = \{z_1, \ldots, z_{n-1}, z'_n\}$. Let $p_t$ and $p'_t$ be the distributions of model parameter after $t$ iterations of Algorithm [7] with $|M_{s-1}| = 1$ for all $s$. Assume that $\ell(w, z)$ is $L$-smooth for all $z$ for some $L > 0$, and $\eta_t \leq 1/L$. Then for all $t$:

$$W_2^2(p_t, p'_t) \leq 4G^2(t/n^2 + 1/n) \sum_{s \leq t} \eta_s^2.$$

Proof of Lemma [2] Consider $w_s (\tilde{w}_s)$, and $w'_s (\tilde{w}'_s)$ obtained from the algorithm with $S_n$ and $S'_n$, under the same randomization (except when we draw $z_n$ from $S_n$, we will draw $z'_n$ from $S'_n$). This gives a specialized couple of $p_s$ and $p'_s$. We can track $\mathbb{E}\| w_s - w'_s \|^2_2$, which gives an upper bound of $W_2^2(p_s, p'_s)^2$.

Consider any stage $t = s$. If we draw $M_{s-1}$ with $|cM_{s-1}| = 1$ that does not contain $n$. Then we have the updates

$$\tilde{w}_{s-1} = w_{s-1} - \eta_t \nabla \ell(w_{s-1}, z_i), \quad \tilde{w}'_{s-1} = w'_{s-1} - \eta_t \nabla \ell(w'_{s-1}, z_i).$$

Since $\eta_t \leq 1/L$, we have

$$\| \tilde{w}_{s-1} - \tilde{w}'_{s-1} \|^2 = \| [w_{s-1} - w'_{s-1}] - \eta_t [\nabla \ell(w_{s-1}, z_i) - \nabla \ell(w'_{s-1}, z_i)] \|^2 \leq \| w_{s-1} - w'_{s-1} \|^2.$$

If we draw $M_{s-1}$ (with $|M_{s-1}| = 1$) that contains $n$ (and thus $z_n$ and $z'_n$ for the two datasets) instead, then we have the update

$$\| \tilde{w}_{s-1} - \tilde{w}'_{s-1} \|^2 = \| [w_{s-1} - w'_{s-1}] - \eta_s [\nabla \ell(w_{s-1}, z_n) - \nabla \ell(w'_{s-1}, z'_n)] \|^2 \leq \| w_{s-1} - w'_{s-1} - \eta_s \delta_s \|^2,$$
where \( \| \delta_s \|_2 = \| \nabla \ell(w_{s-1}, z_n) - \nabla \ell(w'_s, z'_n) \|_2 \leq 2G \). Since the probability that \( M_{s-1} \) contains \( n \) is \( 1/n \), we thus obtain
\[
E\| \bar{w}_{s-1} - \bar{w}'_{s-1} \|_2^2 \\
\leq E\| w_{s-1} - w'_{s-1} \|_2^2 + \frac{4 \eta_n G}{n} E\| w_{s-1} - w'_{s-1} \| + \frac{4 \eta_n^2 G^2}{n} \\
\leq \frac{\eta_n^2}{\eta_s^2} E\| w_{s-1} - w'_{s-1} \|_2^2 + \frac{\eta_n^2}{\eta_s^2} \cdot 4 \eta_n^2 G^2 + \frac{4 \eta_n^2 G^2}{n}.
\]
The last inequality used \( 2ab \leq \frac{a^2}{\eta_s^2} - \frac{\eta_n^2}{\eta_s^2} a^2 + \frac{\eta_n^2}{\eta_s^2} \eta_s^2 b^2 \). Since with the same randomization,
\[
E\| w_s - w'_s \|_2^2 = (1 - \lambda_s \eta_s)^2 E\| \bar{w}_{s-1} - \bar{w}'_{s-1} \|_2^2 = \frac{\eta_s^2}{\eta_s^2 - 1} E\| \bar{w}_{s-1} - \bar{w}'_{s-1} \|_2^2
\]
and \( \frac{\eta_s^2}{\eta_s^2 - 1} \leq s \), we obtain
\[
E\| w_s - w'_s \|_2^2 \leq E\| w_{s-1} - w'_{s-1} \|_2^2 + \frac{4 \eta_n^2 G^2}{n^2} + \frac{4 \eta_n^2 G^2}{n}.
\]
We may take \( w_1 = w'_1 \) because they both come from distribution \( N(0, \beta_0 I) \). By summing over \( s = 2 \) to \( s = t \), we obtain the desired result. \( \square \)

We are now ready to prove Proposition 2. Consider two independent samples \( S_n = \{z_1, \ldots, z_n\} \), and \( S'_n = \{z'_1, \ldots, z'_n\} \). Let \( S^{(i)}_n = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\} \). Let \( p_t^{(i)} \) be the distribution obtained from Algorithm I using \( S^{(i)}_n \), \( p_t \) be the distribution using \( S_n \). We have
\[
E[\ell(p_t, D) - \ell(w, D)] = E\frac{1}{n} \sum_{i=1}^{n} [\ell(p_t^{(i)}, z_i) - \ell(p_t, z_i)] + E[\ell(p_t, S_n) - \ell(w, S_n)]
\]
\[
= E\frac{1}{n} \sum_{i=1}^{n} E_{(w, w'') \sim q_s} [\ell(w', z_i) - \ell(w'', z_i)] + E[\ell(p_t, S_n) - \ell(w, S_n)]
\]
\[
\leq G E\frac{1}{n} \sum_{i=1}^{n} E_{(w, w'') \sim q_s} \| w' - w'' \|_2 + E[\ell(p_t, S_n) - \ell(w, S_n)]
\]
\[
\leq E\frac{1}{n} \sum_{i=1}^{n} GW_2(p_t^{(i)}, p_t) + E[\ell(p_t, S_n) - \ell(w, S_n)].
\]
In the above derivation, the first equality used the fact that \( \ell(p_t, z) \) with \( z \sim D \) has the same distribution as that of \( \ell(p_t^{(i)}, z_i) \). In the second equality, \( q_s(w', w'') \) is the optimal \( W_2 \) coupling in \( \Pi(p_t^{(i)}, p_t) \) to achieve the optimality in \( W_2 \) distance. The first inequality used the fact that \( \ell(w, z) \) is Lipschitz in \( w \). The second inequality used the definition of \( W_2 \)-distance and the Jensen’s inequality. Since \( S_n^{(i)} \) and \( S_n \) differ by one element, we obtain the desired result by applying Lemma 5. \( \square \)

### 5.3 Differential Privacy

We first state the following result.
Lemma 6. Consider \( S_m = \{z_1, z_2, \ldots, z_m\} \), and a randomized algorithm \( w \sim A(\tilde{w}, S_m) \) with the one-step update rule as follows

\[
    w = \alpha \tilde{w} - \eta \nabla \ell(\tilde{w}, z) + \sigma \epsilon, \quad z \sim U(S_m), \quad \epsilon \sim N(0, I),
\]

where \( U(S_m) \) denotes the uniform sampling over elements of \( S_m \). Then \( A \) is deferentially private.

Proof. Consider \( z, z' \in \mathcal{Z} \), and the update rule

\[
    w = \alpha \tilde{w} - \eta \nabla \ell(\tilde{w}, z) + \sigma \epsilon, \quad \epsilon \sim N(0, I)
\]

with probability distribution \( p(w) \), and

\[
    w' = \alpha \tilde{w} - \eta \nabla \ell(\tilde{w}, z') + \sigma \epsilon, \quad \epsilon \sim N(0, I)
\]

with probability distribution \( p'(w') \).

We know from Theorem A.1 of [Dwork et al., 2014] that the Gaussian mechanism (4) applied to any dataset (without considering the randomization effect) is \((\epsilon(\delta), \delta)\) deferentially private, where \( \epsilon(\delta) = \sqrt{\ln(1.25/\delta)\eta G/\sigma} \). We can now apply the privacy amplification result, Theorem 9 of [Balle et al., 2018], to deal sampling without replacement, and obtain the desired deferential privacy guarantee for \( A \).

We can now use the strong composition theorem to obtain the following result. We state the following slightly more general result of Theorem 3.3 of [Dwork et al., 2010], which allows step dependent \((\epsilon, \delta)\).

Proposition 5. Consider an online learning algorithm \( A: \mathcal{Z}^n \to \Omega^T \). If at each step \( t \), \( A([w_1, \ldots, w_{t-1}], S_n) \) is \((\epsilon_t, 0.5\delta/t(t-1))\) deferentially private for \( t \geq 2 \), then \( A \) is deferentially private.

Proof. We simply note that: \( \delta/2 + \sum_{t \geq 2} \frac{0.5\delta}{t(t-1)} = \delta \). The proof is essentially the same as that of [Dwork et al., 2010].

We then have the following proof of Proposition 3.

Proof of Proposition 3. We note that Algorithm 1 can be regarded as the composition of \( T \) steps of mechanisms in Lemma 4 where at each step \( t \), we can take \( \eta = \eta_t^2/\eta_{t-1} \), and \( \sigma^2 = (1 - (\eta_t/\eta_{t-1})^2)\beta_0 \) to obtain \((\epsilon_t, 0.5\delta/t(t-1))\) differential privacy for each step \( t \geq 2 \) of Algorithm 1, where

\[
    \epsilon_t = \ln \left( 1 + n^{-1} \exp \left( \sqrt{\frac{8\ln(2.5t(t-1)/tm\delta)}{(\eta_t^2 - \eta_{t-1}^2)\beta_0}} \right) \right) \leq \ln(1 + \epsilon_0),
\]

where \( \epsilon_0 = \epsilon/n \). In the derivation, we used \( t \ln t^2 - (t-1) \ln(t-1)^2 \geq \ln t^2 \). By applying Proposition 5 we obtain

\[
    \left( \sqrt{2T \ln(2/\delta)} \ln(1 + \epsilon_0) + T(\ln(1 + \epsilon_0)(e^{\ln(1+\epsilon_0)} - 1), \delta) \right)
\]

differential privacy. Since \( \ln(1 + \epsilon) \leq \epsilon \), we obtain the desired bound.
5.4 Rényi Differential Privacy

We are going to use Theorem 3.1 from [Feldman et al., 2020]. Before we state the theorem, however, we have to recall the update for the projected noisy stochastic gradient descent with mini-batches algorithm from [Feldman et al., 2020] given below:

\[ w_t = w_{t-1} - \eta_t \left( \mathcal{G}(w_{t-1}) + \xi_{t-1} \right), \]

where \( \xi_{t-1} \sim \mathcal{N}(0, \sigma^2_{t-1}I) \).

**Theorem 3 (Theorem 3.1 [Feldman et al., 2020]).** Let \( \mathcal{K} \subseteq \mathbb{R}^d \) be a convex set and \( \{f(\cdot, z)\}_{z \in \mathcal{Z}} \) be convex, \( G \)-Lipschitz and \( L \)-smooth functions over \( \mathcal{K} \). Then for every mini-batch sequence \( \{M_t\}_{t \in [T]} \), step-size sequence \( \{\eta_t\}_{t \in [T]} \) with \( \eta_t \leq \frac{2}{T} \), \( \forall t \in [T] \), noise parameter \( \{\sigma_t\}_{t \in T}, \alpha \geq 1 \), starting point \( w_0 \) and \( S_n \in \mathbb{Z}^n \) it holds that projected noisy stochastic gradient descent with mini-batches (PNSGD) with the above parameters satisfies

\[ D_\alpha(w_T\|w'_T) \leq \max_{t \in [T]} \frac{2\alpha G^2 \eta^2 T}{|M_t|^2 \sum_{s=t}^{T} \eta^2_s \sigma^2_s}. \]

*Proof of Proposition 4.* While Algorithm 1 is not exactly an instance of PNSGD we argue that we can apply Theorem 3 with step size \( \eta_t = \eta \) and \( \sigma^2_t = \frac{\eta \lambda_t (2 - \eta \lambda_t)}{\tau \eta \lambda_t (2 - \eta \lambda_t)} \). To apply Theorem 3 it is sufficient to argue that the update in Algorithm 1 is contractive for any step-size sequence with \( \eta_t \leq \frac{2}{T} \). We can write the update as

\[ w_{t+1} = (1 - \lambda_t \eta_t) (w_t - \eta_t \mathcal{G}(w_t)) + \xi_t, \]

where \( \xi_t \sim \mathcal{N}(0, \lambda_t \eta_t (2 - \lambda_t \eta_t) \beta_0 I) \). For two neighboring sequences \( (w_t)_{t \in [T]} \) and \( (w'_t)_{t \in [T]} \) we have

\[ \|w_{t+1} - w'_{t+1}\|_2 \leq (1 - \lambda_t \eta_t) \|w_t - \eta_t \mathcal{G}(w_t) - w'_t - \eta_t \mathcal{G}(w'_t)\|_2 < \|w_t - w'_t\|_2, \]

where in the last inequality we use that \( (1 - \lambda_t \eta_t) < 1 \) together with the fact that for \( \eta_t \leq \frac{2}{T} \) the gradient mapping, \( w_t - \eta_t \mathcal{G}(w_t) \), is contractive because \( \mathcal{G}(\cdot) \) is \( L \)-Lipschitz. We can now apply Theorem 3 to Algorithm 1. Further, because \( 2 \geq (2 - \lambda_t \eta_t) \geq 1 \), we can ignore this factor and so we have

\[ D_\alpha(w_T\|w'_T) \leq \max_{t \in [T]} \frac{2\alpha G^2 \eta^2 T}{|M_t|^2 \sum_{s=t}^{T} \lambda_t}. \]

For the choice of \( |M_t| = \sqrt{T-t+1} \) and \( \lambda_t = \frac{1}{\eta t} \) we have

\[ \max_{t \in [T]} \frac{2\alpha G^2 \eta^2 T}{|M_t|^2 \sum_{s=t}^{T} \lambda_t} = \max_{t \in [T]} \frac{2\alpha G^2 \eta^2 (T-t+1)}{\beta_0 T \sum_{s=t}^{T} \frac{1}{s}} \leq \frac{2\alpha G^2 \eta^2 T}{\beta_0}. \]

\[ \Box \]

6 Conclusion

We obtain dimension independent generalization bounds for DP-SGD using convergence analysis of SGLD of [Freund et al., 2021]. We demonstrate that the derived bounds improve previous results for certain classes of overparameterized smooth convex learning problems. Our results imply that for certain practical applications with overparameterized machine learning models, it is possible to obtain strong differential-privacy guarantees without much degradation in the generalization performance.
References

Martin Abadi, Andy Chu, Ian 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, pages 308–318, 2016.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31, 2018.

R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In Proceedings of 34th Conference on Neural Information Processing Systems (Neurips), 2020.

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

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. arXiv preprint arXiv:1908.09970, 2019.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, pages 51–60. IEEE, 2010.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci., 9(3-4):211–407, 2014.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, pages 439–449, 2020.

Yoav Freund, Yi-An Ma, and Tong Zhang. When is the convergence time of langevin algorithms dimension independent? a composite optimization viewpoint. arXiv preprint arXiv:2110.01827, 2021.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In International Conference on Machine Learning, pages 1225–1234. PMLR, 2016.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. In Conference on Learning Theory, pages 1752–1755. PMLR, 2019.

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

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In International conference on machine learning, pages 71–79. PMLR, 2013.
Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading curse of dimensionality in unconstrained private glm's via private gradient descent. In *AISTAT 2021*, 2021.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688. Omnipress, 2011.