Differential Privacy Dynamics of Langevin Diffusion and Noisy Gradient Descent

Anonymous Author(s)
Affiliation
Address
email

Abstract

What is the information leakage of an iterative learning algorithm about its training data, when the internal state of the algorithm is *not* observable? How much is the contribution of each specific training epoch to the final leakage? We study this problem for noisy gradient descent algorithms, and model the *dynamics* of Rényi differential privacy loss throughout the training process. Our analysis traces a provably tight bound on the Rényi divergence between the pair of probability distributions over parameters of models with neighboring datasets. We prove that the privacy loss converges exponentially fast, for smooth and strongly convex loss functions, which is a significant improvement over composition theorems. For Lipschitz, smooth, and strongly convex loss functions, we prove optimal utility for differential privacy algorithms with a small gradient complexity.

1 Introduction

Machine learning models leak a significant amount of information about their training data, through their parameters and predictions [20, 17]. Randomized iterative algorithms can limit this information leakage and bound the differential privacy loss of the learning process [3, 1, 8, 9]. The strength of this certified defense is determined by an *upper bound* on the (Rényi) divergence between the pair of distributions over the model parameters, which can be learned on any neighboring datasets.

The general method to compute the differential privacy bound for gradient-based learning algorithms is to view the process as a number of (identical) differential privacy mechanisms, and to compute the *composition* of their bounds. However, this over-estimates the privacy loss, thus resulting in a lose differential privacy bound. This is because it also accounts for the leakage of intermediate gradient updates, whereas only the final model parameters are observable (referred to as the black-box analysis [12, 18]). Feldman et al. [8, 9] address this issue for the privacy analysis of *one* single epoch, for smooth and convex loss functions. However, in learning a model over multiple training epochs, such a guarantee is similar to the composition-based bounds which are amplified with sub-sampling [8]. Thus, the open challenge, that we tackle in this paper, is to provide an analysis that can tightly reflect the privacy loss of the released model after K training epochs, for any K.

We present a novel analysis for privacy dynamics of noisy gradient descent with smooth and strongly convex loss functions. We model a pair of continuous-time processes as Langevin diffusion [19] that trace the probability distributions over the model parameters of noisy GD. Subsequently, we derive differential equations bounding the *rate* of privacy loss (worst case Rényi divergence between the **coupled stochastic processes** associated with neighboring datasets) throughout the training process. We then prove an exponentially-fast converging privacy bound for noisy GD: (simplified theorem) Under 1-strongly convex and β-smooth loss function $\ell(\theta; \mathbf{x})$ with total gradient sensitivity 1, the noisy GD Algorithm 1, with initial parameter vector $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, 2\sigma^2 \mathbb{I}_d))$ and step size $\eta \leq 1/(2\beta)$, satisfies (α, ε)-Rényi DP with $\varepsilon = O\left(\frac{\alpha}{\sigma^2 n^2}(1 - e^{-\eta K/2})\right)$, where n is the size of the training set.

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

This guarantee shows that the privacy loss **converges exponentially** in the number of iterations K, instead of growing proportionally with $\sqrt{K}$ as in composition-based analysis of the same algorithms. It captures a strong privacy amplification due to the dynamics (and convergence) of differential privacy over an algorithm with a private internal state. Our converging bound is consistent with the prior results that the distribution over model parameters converges after many iterations [6, 7, 21], and also the results on converging differential privacy analysis [24, 15, 10].

In addition to computing a privacy bound, it is also of great importance to have an analysis of the *tightness* of the bound, the *utility* of the randomized algorithm under the computed differential privacy bound, as well as its *gradient complexity* (number of required gradient computations). We prove the **tightness guarantee** for this privacy bound by showing that there exist a loss function and neighboring datasets such that the divergence between corresponding model parameter distributions matches our privacy bound. For Lipschitz, smooth, and strongly convex loss functions, we prove that noisy GD achieves **optimal utility** of order $O(\frac{d}{n^2\varepsilon^2})$, with a *small gradient complexity* of order $O(n\log(n))$. This improves over previous utility results for noisy SGD algorithm [3], with significantly smaller gradient complexity by a factor $\log(n)/n$, and slightly better utility by $\mathrm{polylog}(n)$. Previously, noisy SGD has worse utility than known optimal ERM algorithm for strongly convex and smooth loss functions, such as objective perturbation [5]. To the best of our knowledge, we are the first to show that noisy GD algorithm matches the optimal utility of DP-ERM algorithms for strongly convex smooth loss functions.

Our work provides a tight privacy analysis *throughout* the training process, for a converging noisy GD algorithm. A different approach is to completely ignore the dynamics of differential privacy, and instead analyze privacy *only at the convergence time* (or when the algorithm is near convergence). Wang et al. [24], Minami et al. [15] show that sampling from the Gibbs posterior distribution $\nu(\theta) \propto e^{-\mathcal{L}_{\mathcal{D}}(\theta)/\sigma^2}$ for bounded $\mathcal{L}_{\mathcal{D}}$ satisfies differential privacy. However, sampling exactly from the Gibbs distribution is difficult [4]. Thus, Minami et al. [15], Ganesh and Talwar [10] extend the DP guarantees of Gibbs posterior distribution to gradient-descent based samplers such as Unadjusted Langevin Algorithm (ULA) that can sample from distributions arbitrarily close to Gibbs distribution after a sufficient number of iterations K with extremely small step-size η. Minami et al. [15] compute the distance to convergence in total variation, and Ganesh and Talwar [10] improve the prior bound by measuring the distance in Rényi divergence (building on the rapid convergence results of Vempala and Wibisono [21]). The latter results in a better gradient complexity $\Omega(nd)$, which however still grows with model dimension d. In comparison, our DP guarantees are unaffected by parameter dimension d, which in practice can be much larger than the dataset size n.

We do not foresee any potential negative societal impact of this work. Conversely, this work significantly contributes to the research field on improving the privacy of individuals in ML systems.

2 Preliminaries on differential privacy

Let $\mathcal{X}$ be the data universe, and a dataset $\mathcal{D}$ contain n records from it: $\mathcal{D} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) \in \mathcal{X}^n$. We refer to a dataset pair $\mathcal{D}, \mathcal{D}'$ as *neighboring* if they differ in one data record.

Definition 2.1 ([16] Rényi differential privacy)**.** *Let $\alpha > 1$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ satisfies* (α, ε)-Rényi Differential Privacy (RDP)*, if for any two neighboring datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$, the α Rényi divergence $R_\alpha\left(\mathcal{A}(\mathcal{D}) \| \mathcal{A}(\mathcal{D}')\right) \leq \varepsilon$. For a pair of distributions p, p' over $\mathbb{R}^d$ with* $\mathrm{supp}(p) \subset \mathrm{supp}(p')$*, $R_\alpha\left(p \| p'\right)$ is defined as*

$$R_\alpha\left(p \| p'\right) = \frac{1}{\alpha - 1} \log E_\alpha\left(p \| p'\right), \quad \textit{where} \quad E_\alpha\left(p \| p'\right) = \mathop{\mathbb{E}}_{\theta \sim p'}\left[\frac{p(\theta)^\alpha}{p'(\theta)^\alpha}\right]. \tag{1}$$

We refer to $R_\alpha\left(\mathcal{A}(\mathcal{D}) \| \mathcal{A}(\mathcal{D}')\right)$ also as the *Rényi privacy loss* of algorithm $\mathcal{A}$ on datasets $\mathcal{D}, \mathcal{D}'$. An RDP guarantee can be converted to (ε, δ)-DP guarantee [16, Proposition 5].

Definition 2.2 ([21] Rényi information)**.** *Let $\alpha > 1$. For any two distributions p, p' over $\mathbb{R}^d$ with* $\mathrm{supp}(p) \subset \mathrm{supp}(p')$ *and differentiable $\frac{p(\theta)}{p'(\theta)}$, the* α-Rényi Information *of p with respect to p' is*

$$I_\alpha\left(p \| p'\right) = \frac{4}{\alpha^2} \mathop{\mathbb{E}}_{\theta \sim p'}\left[\left\|\nabla \frac{p(\theta)^{\frac{\alpha}{2}}}{p'(\theta)^{\frac{\alpha}{2}}}\right\|_2^2\right] = \mathop{\mathbb{E}}_{\theta \sim p'}\left[\frac{p(\theta)^{\alpha-2}}{p'(\theta)^{\alpha-2}}\left\|\nabla \frac{p(\theta)}{p'(\theta)}\right\|_2^2\right]. \tag{2}$$

See the Appendix for a comprehensive presentation of preliminaries.

3 Privacy analysis of noisy gradient descent

Let $\mathcal{D} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ be a dataset of size n with records taken from a universe $\mathcal{X}$. For a given machine learning algorithm, let $\ell(\theta; \mathbf{x}) : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$ be a loss function of a parameter vector $\theta \in \mathcal{C}$ on the data point $\mathbf{x}$, where $\mathcal{C}$ is a closed convex set (can be $\mathbb{R}^d$).

A generic formulation of the optimization problem to learn the model parameters, is in the form of empirical risk minimization (ERM) with the following objective, where $\mathcal{L}_{\mathcal{D}}(\theta)$ is the empirical loss of the model, with parameter vector θ, on a dataset $\mathcal{D}$.

$$\theta^* = \arg\min_{\theta \in \mathcal{C}} \mathcal{L}_{\mathcal{D}}(\theta), \quad \text{where} \quad \mathcal{L}_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{D}} \ell(\theta; \mathbf{x}). \tag{3}$$

Releasing this optimization output (i.e., θ^*) can leak information about the dataset $\mathcal{D}$, hence violating data privacy. To mitigate this risk, there exist randomized algorithms to ensure that the (α-Rényi) privacy loss of the ERM algorithm is upper-bounded by ε, i.e., the algorithm satisfies (α, ε)-RDP.

Algorithm 1 $\mathcal{A}_{\text{Noisy-GD}}$: Noisy Gradient Descent

Input: Dataset $\mathcal{D} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$, loss function $\ell(\theta; \mathbf{x})$, closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$, learning rate η, noise variance σ^2, initial parameter vector θ_0.

1: **for** $k = 0, 1, \cdots, K-1$ **do**
2: $\quad g(\theta_k; \mathcal{D}) = \sum_{i=1}^n \nabla \ell(\theta_k; \mathbf{x}_i)$
3: $\quad \theta_{k+1} = \Pi_{\mathcal{C}}(\theta_k - \frac{\eta}{n} g(\theta_k; \mathcal{D}) + \sqrt{2\eta}\mathcal{N}(0, \sigma^2 \mathbb{I}_d))$
4: Output θ_K

In this paper, our objective is to analyze privacy loss of *Noisy Gradient Descent* (Algorithm 1), which is a randomized ERM algorithm. Let θ_k, θ'_k be the parameter vectors at the k'th iteration of $\mathcal{A}_{\text{Noisy-GD}}$ on neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, respectively. We denote by $\Theta_{\eta k}$ and $\Theta'_{\eta k}$ the corresponding random variables that model θ_k and θ'_k. We abuse notation to also denote their probability density functions by $\Theta_{\eta k}$ and $\Theta'_{\eta k}$. In this paper, our objective is to model and analyze the **dynamics of differential privacy** of this algorithm. More precisely, we focus on the following.

1. Compute an RDP bound (i.e., the worst case Rényi divergence $R_\alpha\left(\Theta_K \| \Theta'_K\right)$ between the output distributions of two neighboring datasets) for Algorithm 1, and analyze its tightness.
2. Compute the contribution of each iteration to the privacy loss. As we go from step $k = 1$ to K in Algorithm 1, we investigate how the algorithm's privacy loss changes as it runs the k'th iteration (computed as $R_\alpha\left(\Theta_{\eta k} \middle\| \Theta'_{\eta k}\right) - R_\alpha\left(\Theta_{\eta(k-1)} \middle\| \Theta'_{\eta(k-1)}\right)$).

In the end, we aim to provide a RDP bound that is tight, thus facilitating optimal utility [3]. We emphasize that our goal is to construct a theoretical framework for analyzing privacy loss of releasing the output θ_K of the algorithm, assuming *private* internal states (i.e., $\theta_1, \cdots, \theta_{K-1}$).

3.1 Tracing diffusion for Noisy GD

To analyze the privacy loss of Noisy GD, which is a *discrete-time stochastic process*, we first interpolate each discrete update from θ_k to θ_{k+1} with a piece-wise continuously differentiable diffusion process. Given step-size η and initial parameter vector θ_0, the k'th discrete update in Algorithm 1 is

$$\theta_{k+1} = \Pi_{\mathcal{C}}(\theta_k - \eta \nabla \mathcal{L}_{\mathcal{D}}(\theta_k) + \sqrt{2\eta\sigma^2}\mathbf{Z}_k), \quad \text{with} \quad \mathbf{Z}_k \sim \mathcal{N}(0, \mathbb{I}_d). \tag{4}$$

This discrete jump can be interpolated with the following SDE over time $\eta k \le t < \eta(k+1)$,

$$d\Theta_t = -\nabla \mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) dt + \sqrt{2\sigma^2} d\mathbf{W}_t, \tag{5}$$

where $d\mathbf{W}_t \sim \sqrt{dt}\mathcal{N}(0, \mathbb{I}_d)$ describes the Wiener process on $\mathbb{R}^d$. Integrate this SDE over time gives

$$\Theta_t - \Theta_{\eta k} = (t - \eta k)\nabla \mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) + \sqrt{2(t - \eta k)\sigma^2}\mathbf{Z}_k, \quad \text{with} \quad \mathbf{Z}_k \sim \mathcal{N}(0, \mathbb{I}_d), \tag{6}$$

for any $k\eta \le t < (k+1)\eta$. At the end of step, i.e. at $t \to \eta(k+1)$, we project onto convex set $\mathcal{C}$,

$$\Theta_{\eta(k+1)} = \Pi_{\mathcal{C}}(\lim_{t\to\eta(k+1)} \Theta_t), \quad \text{where} \quad t < \eta(k+1). \tag{7}$$

Therefore θ_{k+1} at $k+1^{\text{th}}$ step of noisy GD, is distributed as this projected random variable $\Theta_{\eta(k+1)}$. Repeating the construction for $k = 0, \cdots, K-1$, we define a piece-wise continuous diffusion process $\{\Theta_t\}_{t\ge 0}$ whose distribution at time $t = \eta k$ is consistent with θ_k in the noisy GD process for any $k \in \{0, \cdots, K-1\}$. We refer to this process as the *tracing diffusion* for noisy GD.

Definition 3.1 (Coupled tracing diffusions)**.** *Let $\Theta_0 = \Theta_0'$ be two identically distributed random variables. We refer to the stochastic processes $\{\Theta_t\}_{t\ge 0}$ and $\{\Theta_t'\}_{t\ge 0}$ that solve SDE* (5) *and projection* (7) *as coupled tracing diffusions for noisy GD on neighboring datasets $\mathcal{D}, \mathcal{D}'$.*

The Rényi divergence $R_\alpha(\Theta_{\eta K} \| \Theta'_{\eta K})$ reflects the Rényi privacy loss of Algorithm 1 with K steps. Conditioned on observing θ_k, the process $\{\Theta_t\}_{\eta k \le t < \eta(k+1)}$ is Langevin diffusion along a constant vector field $\nabla \mathcal{L}_{\mathcal{D}}(\theta_k)$ for duration η. Therefore, the conditional probability distribution $p_{t|\eta k}(\theta|\theta_k)$ follows the following Fokker-Planck equation. For brevity, we use $p_{t|t'}(\theta|\theta')$ to represent the conditional probability density function $p(\Theta_t = \theta | \Theta_{t'} = \theta')$.

$$\frac{\partial p_{t|\eta k}(\theta|\theta_k)}{\partial t} = \nabla \cdot \big(p_{t|\eta k}(\theta|\theta_k) \nabla \mathcal{L}_{\mathcal{D}}(\theta_k)\big) + \sigma^2 \Delta p_{t|\eta k}(\theta|\theta_k)$$

By taking expectations over distribution $p_{\eta k}(\theta_k)$ on both sides, we get the partial differential equation that models the evolution of (unconditioned) probability distribution $p_t(\theta)$ in the tracing diffusion.

Lemma 1. *For SDE* (5)*, the equivalent Fokker-Planck equation at time $\eta k \le t < \eta(k+1)$ is*

$$\frac{\partial p_t(\theta)}{\partial t} = \nabla \cdot (p_t(\theta) \mathop{\mathbb{E}}_{\theta_k \sim p_{\eta k|t}} [\nabla \mathcal{L}_{\mathcal{D}}(\theta_k)|\theta]) + \sigma^2 \Delta p_t(\theta). \tag{8}$$

Using this distribution evolution equation, we model the privacy dynamics in the tracing diffusion process. This process is similar to Langevin diffusion under conditionally expected loss function $\mathop{\mathbb{E}}_{\theta_k \sim p_{\eta k|t}} [\mathcal{L}_{\mathcal{D}}(\theta_k)|\theta]$. Therefore, we first study the privacy dynamics in Langevin diffusion.

3.2 Privacy erosion in tracing (Langevin) diffusion

In the Langevin diffusion process, the privacy loss increases over time, as each gradient computation requires accessing data. We refer to this phenomenon as **privacy erosion**. This increase is determined by the amount of distribution change in Langevin diffusion, characterized by the loss gradient and Gaussian noise in the following Fokker-Planck equation for Langevin diffusion.

$$\frac{\partial p_t(\theta)}{\partial t} = -\nabla \cdot (p_t(\theta) \nabla \mathcal{L}_{\mathcal{D}}(\theta)) + \sigma^2 \Delta p_t(\theta). \tag{9}$$

We compute a bound on the rate (partial derivative) of $R_\alpha\left(\Theta_t \| \Theta_t'\right)$ over time based on this equation in the following lemma, to model privacy erosion in Langevin diffusion. We refer to *coupled Langevin diffusions* as respective Langevin diffusion processes under any neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$.

Lemma 2 (Rate of Rényi privacy loss)**.** *Let $\ell(\theta; \mathbf{x})$ be a loss function on $\mathbb{R}^d$, with a finite total gradient sensitivity S_g. Then, for coupled Langevin diffusions $\{\Theta_t\}_{t\ge 0}$ and $\{\Theta_t'\}_{t\ge 0}$ on neighboring datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$, under loss $\ell(\theta; \mathbf{x})$ and noise variance σ^2, the Rényi privacy loss rate at any $t \ge 0$ is upper bounded by*

$$\frac{\partial R_\alpha\left(\Theta_t \| \Theta_t'\right)}{\partial t} \le \frac{1}{\gamma} \frac{\alpha S_g^2}{4\sigma^2 n^2} - (1-\gamma)\sigma^2 \alpha \frac{I_\alpha\left(\Theta_t \| \Theta_t'\right)}{E_\alpha\left(\Theta_t \| \Theta_t'\right)}. \tag{10}$$

where $\gamma > 0$ is a tuning parameter that we can fix arbitrarily according to our need.

Although this lemma bounds the Rényi privacy loss rate, the term $I_\alpha\left(\Theta_t \| \Theta_t'\right)$ depends on unknown distributions Θ_t, Θ_t', and is intractable to compute. Even with explicit expressions for distributions of Θ_t, Θ_t', the calculation would involve integration in $\mathbb{R}^d$ which is computationally prohibitive for large d. Note that, however, the ratio I_α / E_α is always positive by definition. Therefore, the Rényi privacy loss rate in (10) is bounded by its first component (a constant) given any fixed α.

Theorem 1 (Linear RDP bound)**.** *Let $\ell(\theta; \mathbf{x})$ be a loss function on $\mathbb{R}^d$, with a finite total gradient sensitivity S_g. Then, the Langevin diffusion under loss $\ell(\theta; \mathbf{x})$ and noise variance σ^2 for time T satisfies (α, ε)-RDP for $\varepsilon = \frac{\alpha S_g^2 T}{4\sigma^2 n^2}$.*

This naive privacy analysis gives linear RDP guarantee, which resembles the moment accountant analysis [1]. However, a tighter bound of Rényi privacy loss is possible with finer control of the ratio $I_\alpha(\Theta_t \| \Theta_t')/E_\alpha(\Theta_t \| \Theta_t')$, which by definition depends on the likelihood ratio between Θ_t and Θ_t', thus is connected with Rényi privacy loss itself. When this ratio grows, the Rényi privacy loss rate decreases, thus slowing down privacy loss accumulation, and leading to tighter privacy bound.

Controlling Rényi privacy loss rate under isoperimetry We control the I_α/E_α term in lemma 2 by making an *isoperimetric* assumption known as *log-Sobolev inequality* [2], described as follows.

Definition 3.2 ([11] Log-Sobolev Inequality (c-LSI))**.** *Distribution of a random variable Θ on $\mathbb{R}^d$ satisfies* logarithmic Sobolev inequality *with parameter $c > 0$, i.e. it is c-LSI, if for all functions f in the function set $\mathcal{F}_\Theta = \{f : \mathbb{R}^d \to \mathbb{R} | \nabla f$ is continuous, and $\mathbb{E}(f(\Theta)^2) < \infty\}$, we have*

$$\mathbb{E}[f(\Theta)^2 \log f(\Theta)^2] - \mathbb{E}[f(\Theta)^2] \log \mathbb{E}[f(\Theta)^2] \leq \frac{2}{c}\mathbb{E}[\|\nabla f(\Theta)\|_2^2]. \tag{11}$$

LSI was introduced by Gross [11] as a necessary and sufficient condition for rapid convergence of a diffusion processes. Recently, Vempala and Wibisono [21] showed that this isoperimetry condition is sufficient for rapid convergence of Langevin diffusion in Rényi divergence. Under LSI, they provide the following useful lower bound on I_α/E_α for an arbitrary pair of distributions.

Lemma 3 ([21] c-LSI in terms of Rényi Divergence)**.** *Suppose $\Theta_t, \Theta_t' \in \mathbb{R}^d$ are random variables such that likelihood ratio between Θ_t and Θ_t' lies in $\mathcal{F}_{\Theta_t'}$. Then for any $\alpha \geq 1$,*

$$R_\alpha(\Theta_t \| \Theta_t') + \alpha(\alpha - 1)\frac{\partial R_\alpha(\Theta_t \| \Theta_t')}{\partial \alpha} \leq \frac{\alpha^2}{2c}\frac{I_\alpha(\Theta_t \| \Theta_t')}{E_\alpha(\Theta_t \| \Theta_t')},$$

if and only if Θ_t' satisfies c-LSI.

This lemma shows that the $I_\alpha(\Theta_t \| \Theta_t')/E_\alpha(\Theta_t \| \Theta_t')$ term grows monotonically with the Rényi privacy loss $R_\alpha(\Theta_t \| \Theta_t')$, which implies a throttling privacy loss rate as privacy loss accumulates, by Lemma 2. Using this lemma, we model the **dynamics for Rényi privacy loss** with the following PDE, which describes the relation between privacy loss, its changes over time, and its change over Rényi parameter α under c-LSI. For brevity, let $R(\alpha, t)$ represent $R_\alpha(\Theta_t \| \Theta_t')$.

$$\frac{\partial R(\alpha, t)}{\partial t} \leq \frac{1}{\gamma}\frac{\alpha S_g^2}{4n^2\sigma^2} - 2(1-\gamma)\sigma^2 c\left[\frac{R(\alpha, t)}{\alpha} + (\alpha - 1)\frac{\partial R(\alpha, t)}{\partial \alpha}\right] \tag{12}$$

The initial privacy loss $R(\alpha, 0) = 0$, as $\Theta_0 = \Theta_0'$. The solution for this PDE increases with time $t \geq 0$, and models the erosion of Rényi privacy loss in Langevin diffusion.

3.3 Privacy guarantee for Noisy GD

We now extend the privacy dynamics (12) of Langevin diffusion to the tracing diffusion for noisy GD. We bound gradient sensitivity of conditionally expected loss $\mathop{\mathbb{E}}_{\theta_k \sim p_{\eta k|t}}[\mathcal{L}_\mathcal{D}(\theta_k)|\theta]$ in tracing diffusion.

Lemma 4. *Let $\ell(\theta; \mathbf{x})$ be a β-smooth loss function on closed convex set $\mathcal{C}$, with a finite total gradient sensitivity S_g, then for coupled tracing diffusions $\{\Theta_t\}_{t \geq 0}$ and $\{\Theta_t'\}_{t \geq 0}$ on neighboring datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$, under loss $\ell(\theta; \mathbf{x})$ and noise variance σ^2, the gradient sensitivity of conditionally expected loss is bounded by*

$$\left\| \mathop{\mathbb{E}}_{\theta_k \sim p_{\eta k|t}}[\nabla \mathcal{L}_\mathcal{D}(\theta_k)|\theta] - \mathop{\mathbb{E}}_{\theta_k' \sim p'_{\eta k|t}}[\nabla \mathcal{L}_{\mathcal{D}'}(\theta_k')|\theta] \right\|_2^2 \leq \frac{S_g^2}{n^2(1-\eta\beta)^2}. \tag{13}$$

This is because under smoothness and small step-size, the gradient $\nabla \mathcal{L}_\mathcal{D}(\Theta_{\eta k})$ in each piece of the tracing diffusion $\{\Theta_t\}_{\eta k \leq t < \eta(k+1)}$ can be uniquely expressed by Θ_t and Gaussian noise Z_t, by (6). This facilitates bounding the gradient sensitivity of conditionally expected loss.

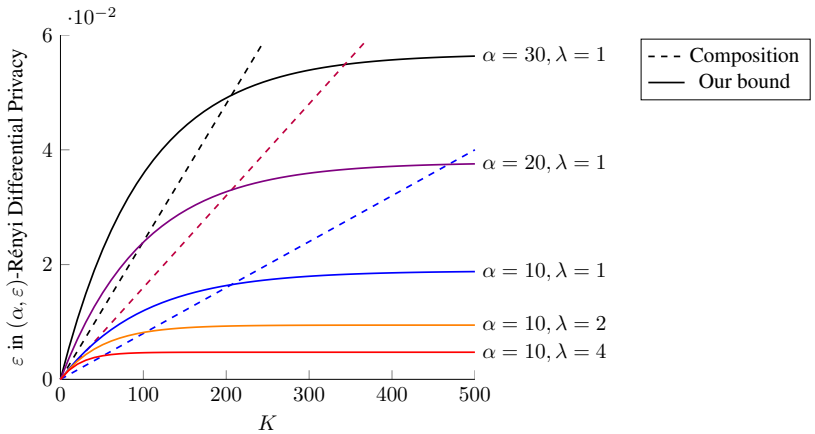

Figure 1: Rényi privacy loss of noisy GD over K iterations, quantified using our DP Dynamics Analysis. We show ε in the (α, ε)-RDP guarantee derived by Corollary 1 (bold lines), and the Baseline composition analysis (dashed lines). We evaluate under the following setting: RDP order $\alpha \in \{10, 20, 30\}$; λ-strongly convex loss function with $\lambda \in \{1, 2, 4\}$; β-smooth loss function with $\beta = 4$; finite ℓ_2-sensitivity S_g for total gradient with $S_g = 4$; size of the data set $n = 50$; step-size $\eta = 0.02$; noise variance $\sigma^2 = 4$. The expressions for computing the privacy loss are: our analysis: $\varepsilon = \frac{\alpha S_g^2}{\lambda \sigma^2 n^2 (1-\eta\beta)^2} \cdot \left(1 - e^{-\lambda \eta K/2}\right)$; and Baseline composition-based analysis: $\varepsilon = \frac{\alpha S_g^2}{4 n^2 \sigma^2} \cdot \eta K$

We then substitute S_g in PDE (12) with $S_g/(1-\eta\beta)$, and compute the following PDE modelling Rényi privacy loss dynamics of tracing diffusion at $\eta k \le t < \eta(k+1)$, under c-LSI condition.

$$\frac{\partial R(\alpha, t)}{\partial t} \le \frac{1}{\gamma} \frac{\alpha S_g^2}{4\sigma^2 n^2 (1-\eta\beta)^2} - 2(1-\gamma)\sigma^2 c \left[\frac{R(\alpha, t)}{\alpha} + (\alpha - 1) \frac{\partial R(\alpha, t)}{\partial \alpha} \right] \tag{14}$$

We solve this PDE under $\gamma = \frac{1}{2}$ for each time piece, and combine multiple pieces by seeing projection as privacy-preserving post-processing step. We derive the RDP guarantee for the Noisy GD algorithm.

Theorem 2 (RDP for noisy GD under c-LSI)**.** *Let $\{\Theta_t\}_{t \ge 0}$ be the tracing diffusion for $\mathcal{A}_{\text{Noisy-GD}}$ on dataset $\mathcal{D}$. Let $\ell(\theta; \mathbf{x})$ be a β-smooth loss function on closed convex set $\mathcal{C}$, with a finite total gradient sensitivity S_g. If Θ_t satisfies c-LSI throughout $0 \le t \le \eta K$, and the step-size $\eta < \frac{1}{\beta}$, then $\mathcal{A}_{\text{Noisy-GD}}$ satisfies (α, ε) Rényi Differential Privacy for*

$$\varepsilon = \frac{\alpha S_g^2}{2c\sigma^4 n^2 (1-\eta\beta)^2} (1 - e^{-\sigma^2 c \eta K}). \tag{15}$$

This theorem offers a strong converging privacy guarantee, on the condition that c-LSI is satisfied throughout the Noisy GD process. We then analyze the LSI constant c for given Noisy GD process.

Isoperimetry constants for noisy GD When the loss function is strongly convex and smooth, we prove that tracing diffusion of noisy GD satisfies LSI. This is because the gradient descent update is Lipschitz under smooth loss, and the Gaussian noise preserves LSI, as discussed in the Appendix.

Lemma 5 (LSI for noisy GD)**.** *If loss function $\ell(\theta; \mathbf{x})$ is λ-strongly convex and β-smooth over a closed convex set $\mathcal{C}$, the step-size is $\eta < \frac{1}{\beta}$, and initial distribution is $\Theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbb{I}_d))$, then the distribution of tracing diffusion Θ_t for noisy GD satisfies c-LSI for any $t \ge 0$ with $c = \frac{\lambda}{2\sigma^2}$.*

Using the LSI constant proved by this lemma, we immediately prove the following RDP bound for noisy GD on Lipschitz smooth strongly convex loss, as a corollary of Theorem 2.

Corollary 1 (Privacy Guarantee for noisy GD)**.** *Let $\ell(\theta; \mathbf{x})$ be a λ-strongly convex, and β-smooth loss function on closed convex set $\mathcal{C}$, with a finite total gradient sensitivity S_g, then the noisy gradient descent algorithm (Algorithm 1) with start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbb{I}_d))$, and step-size $\eta \le \frac{1}{\beta}$, satisfies (α, ε) Rényi Differential Privacy with*

$$\varepsilon = \frac{\alpha S_g^2}{\lambda \sigma^2 n^2 (1-\eta\beta)} (1 - e^{-\lambda \eta K/2}).$$

Figure 1 demonstrates how this RDP guarantee for noisy GD converges with the number of iterations K. Through y-axis, we show the ε guaranteed for noisy GD under various Rényi divergence orders c and strong convexity constant λ. The RDP order α linearly scales the asymptotic guarantee, but does not affect the convergence rate of RDP guarantee. However, the strong convexity parameter λ positively affects the asymptotic guarantee as well as the convergence rate; the larger the strong convexity parameter λ is, the stronger the asymptotic RDP guarantee and the faster the convergence.

4 Tightness analysis: a lower bound on privacy loss of noisy GD

Differential privacy guarantees reflect a *bound* on privacy loss on an algorithm; thus, it is very crucial to also have an analysis of their tightness (i.e., how close they are to the exact privacy loss). We prove that our RDP guarantee in Theorem 2 is tight. To this end, we construct an instance of the ERM optimization problem, for which we show that the Rényi privacy loss of the noisy GD algorithm grows at an order matching our guarantee in Theorem 2.

It is very challenging to lower bound the the exact Rényi privacy loss $R_\alpha\left(\Theta_{K\eta}\middle\|\Theta'_{K\eta}\right)$ in general. This might require having an explicit expression for the probability distribution over the last-iterate parameters θ_k. Computing a close-form expression is, however, feasible when the loss gradients are linear. This is due to the fact that, after a sequence of linear transformations and Gaussian noise additions, the parameters follow a Gaussian distribution. Therefore, we construct such an ERM objective, compute the exact privacy loss, and prove the following lower bound.

Theorem 3 (Lower bound on RDP of $\mathcal{A}_{\text{Noisy-GD}}$)**.** *There exist two neighboring datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$, a start parameter θ_0, and a smooth loss function $\ell(\theta; \mathbf{x})$ on unconstrained convex set $\mathcal{C} = \mathbb{R}^d$, with a finite total gradient sensitivity S_g, such that for any step-size $\eta < 1$, noise variance $\sigma^2 > 0$, and iteration $K \in \mathbb{N}$, the privacy loss of $\mathcal{A}_{\text{Noisy-GD}}$ on $\mathcal{D}, \mathcal{D}'$ is lower-bounded by*

$$R_\alpha\left(\Theta_{\eta K}\middle\|\Theta'_{\eta K}\right) \geq \frac{\alpha S_g^2}{4\sigma^2 n^2}\left(1 - e^{-\eta K}\right). \tag{16}$$

We prove this lower bound using the ℓ_2-squared norm loss as ERM objective: $\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^n \frac{\|\theta - \mathbf{x}_i\|_2^2}{n}$. We assume bounded data domain s.t. the gradient has finite sensitivity. With start parameter $\theta_0 = 0^d$, the kth step parameter θ_k is distributed as Gaussian with mean $\mu_k = \eta\bar{\mathbf{x}}\sum_{i=0}^{k-1}(1-\eta)^i$ and variance $\sigma_k^2 = \frac{2\eta\sigma^2}{n^2}\sum_{i=0}^{k-1}(1-\eta)^{2i}$ in each dimension, where $\bar{x} = \sum_{i=1}^n \mathbf{x}_i/n$ is the empirical dataset mean. We explicitly compute the privacy loss at any step K, which is lower bounded by $\frac{\alpha S_g^2}{4\sigma^2 n^2}(1 - e^{-\eta K})$.

Meanwhile, Theorem 2 gives our RDP upper-bound for Noisy GD on this same ERM objective, because all intermediate Gaussian distributions satisfy c-LSI with $c = \frac{2-\eta}{2\sigma^2}$, as proved in Appendix.

Corollary 2 (RDP guarantee of $\mathcal{A}_{\text{Noisy-GD}}$ on ℓ_2-norm squared loss)**.** *For any two neighboring datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$, start parameter θ_0, step-size η, noise variance σ^2, and $K \in \mathbb{N}$, if the loss function is ℓ_2-norm squared loss $\ell(\theta; \mathbf{x}) = \frac{1}{2}\|\theta - \mathbf{x}\|_2^2$ on unconstrained convex set $\mathcal{C} = \mathbb{R}^d$, with a finite total gradient sensitivity S_g, the privacy loss of $\mathcal{A}_{\text{Noisy-GD}}$ on $\mathcal{D}, \mathcal{D}'$ is upper-bounded by*

$$R_\alpha\left(\Theta_{\eta K}\middle\|\Theta'_{\eta K}\right) \leq \frac{\alpha S_g^2}{(2-\eta)\sigma^2 n^2 (1-\eta)^2}(1 - e^{-\frac{2-\eta}{2}\eta K}). \tag{17}$$

This RDP guarantee converges fast to $O(\frac{\alpha S_g^2}{\sigma^2 n^2})$, which matches the lower bound at every step K, up to a constant of $\frac{2}{(1-\eta)^2} \approx 2$. This immediately shows tightness of our converging RDP guarantee. Meanwhile, composition-based privacy bound grows linearly as training proceeds, as shown in Figure 1. When the number of iterations K is small, however, composition-based bound grows at the same rate with the lower bound, as discussed in Appendix. Therefore, to conclude whether our RDP guarantee is superior to composition-based bound, we need to understand the number of iterations noisy GD needs, to achieve optimal utility. We discuss this in the following section.

5 Utility analysis for noisy gradient descent

The randomness, required for satisfying differential privacy, can adversely affect the utility of the trained model. The standard way to measure the utility of a randomized ERM algorithm (for example,

Table 1: Utility comparison with the prior (ε, δ)-DP ERM algorithms. We assume 1-Lipschitz, β-smooth and λ-strongly convex loss. Size of input dataset is n, and dimension of parameter vector θ is d. For objective perturbation, we assume $\varepsilon \geq \frac{\beta}{2\lambda}$, and loss is twice differentiable. For our result, we assume $\varepsilon \leq 2\log(1/\delta)$. The lower bound is $\Omega\left(\min\left\{1, \frac{d}{\varepsilon^2 n^2}\right\}\right)$ [3]. We ignore numerical constants and multiplicative dependence on $\log(1/\delta)$.

	Method	Utility Upper Bound	Gradient complexity
Bassily et al. [3]	Noisy SGD	$O(\frac{d\log^2(n)}{\lambda n^2 \epsilon^2})$	n^2
Wang et al. [22]	DP-SVRG	$O(\frac{d\log(n)}{\lambda n^2 \epsilon^2})$	$O\left((n + \frac{\beta}{\lambda})\log(\frac{\lambda n^2 \varepsilon^2}{d})\right)$
Zhang et al. [25]	Output Perturbation	$O(\frac{\beta d}{\lambda^2 n^2 \epsilon^2})$	$O(\frac{\beta}{\lambda}\log(\frac{n^2\varepsilon^2}{d}))$
Kifer et al. [13]	Objective Perturbation	$O(\frac{d}{\lambda n^2 \epsilon^2})$	NA
This Paper	Noisy GD	$O(\frac{\beta d}{\lambda^2 n^2 \epsilon^2})$	$O(\frac{\beta^2}{\lambda^2} n \log\left(\frac{n^2\varepsilon^2}{d}\right))$

$\mathcal{A}_{\text{Noisy-GD}}$) is to quantify its worst case *excess empirical risk*, which is defined as

$$\max_{\mathcal{D}\in\mathcal{X}^n} \mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta) - \mathcal{L}_{\mathcal{D}}(\theta^*)], \tag{18}$$

where θ is the output of the randomized algorithm $\mathcal{A}_{\text{Noisy-GD}}$ on $\mathcal{D}$, θ^* is the solution to the standard (no privacy) ERM (3), and the expectation is computed over the randomness of the algorithm.

We provide the *optimal* excess empirical risk (utility) of noisy GD algorithm under (α, ε')-RDP constraint. The notion of *optimality* for utility is defined as the smallest upper-bound for excess empirical risk that can be guaranteed under (α, ε')-RDP constraint by tuning the algorithm's hyper-parameters (such as the noise variance σ^2 and the number of iterations K). We focus here on smooth and strongly convex loss functions with a finite total gradient sensitivity.

Lemma 6 (Excess empirical risk for smooth and strongly convex loss)**.** *For L-Lipschitz, λ-strongly convex and β-smooth loss function $\ell(\theta; \mathbf{x})$ over a closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$, step-size $\eta \leq \frac{\lambda}{2\beta^2}$, and start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\sigma^2))$, the excess empirical risk of Algorithm 1 is bounded by*

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \leq \frac{2\beta L^2}{\lambda^2} e^{-\lambda\eta K} + \frac{2\beta d\sigma^2}{\lambda}, \tag{19}$$

where θ^* is the minimizer of $\mathcal{L}_{\mathcal{D}}(\theta)$ in $\mathcal{C}$, and d is the dimension of parameter.

This lemma shows decreasing excess empirical risk for noisy GD algorithm under Lipschitz smooth strongly convex loss function as the number of iterations K increases. The utility is determined by K and the noise variance σ^2, which are constrained under (α, ε')-RDP. Using our tight RDP guarantee in Corollary 1, we prove optimal utility for noisy GD.

Theorem 4 (Upper bound for (α, ε')-RDP and (ε, δ)-DP Noisy GD)**.** *For Lipschitz smooth strongly convex loss function $\ell(\theta; \mathbf{x})$ on a bounded closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$, and dataset $\mathcal{D} \in \mathcal{X}^n$ of size n, if the step-size $\eta = \frac{\lambda}{2\beta^2}$ and the initial parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbb{I}_d))$, then the noisy GD Algorithm 1 is (α, ε')-Rényi differentially private, where $\alpha > 1$ and $\varepsilon' > 0$, and satisfies*

$$\mathbb{E}[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)] = O(\frac{\alpha\beta d L^2}{\varepsilon'\lambda^2 n^2}), \tag{20}$$

by setting noise variance $\sigma^2 = \frac{\alpha S_g^2}{\lambda\varepsilon' n^2(1-\eta\beta)^2}$, and number of updates $K^* = \frac{2\beta^2}{\lambda^2}\log(\frac{n^2\varepsilon'}{\alpha d})$.

Equivalently, for $\varepsilon \leq 2\log(1/\delta)$ and $\delta > 0$, Algorithm 1 is (ε, δ)-differentially private, and satisfies

$$\mathbb{E}[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)] = O(\frac{\beta d L^2 \log(1/\delta)}{\epsilon^2\lambda^2 n^2}), \tag{21}$$

by setting noise variance $\sigma^2 = \frac{8\log(1/\delta)S_g^2}{\lambda\varepsilon^2 n^2(1-\eta\beta)^2}$, and number of updates $K^* = \frac{2\beta^2}{\lambda^2}\log(\frac{n^2\varepsilon^2}{8\log(1/\delta)d})$.

Our algorithm achieves this utility guarantee with $O(\frac{\beta^2}{\lambda^2} n \log\left(\frac{\varepsilon^2 n^2}{d}\right))$ gradient computations of $\nabla \ell(\theta; \mathbf{x})$, which is faster than noisy SGD algorithm [3] with a factor of n. However, we additionally assume smoothness for the loss function. Our gradient complexity also matches that of other efficient gradient perturbation and output perturbation methods [22, 25], as shown in Table 1.

This utility matches the following theoretical lower bound in Bassily et al. [3] for the best attainable utility of (ϵ, δ)-differentially private algorithms on Lipschitz smooth strongly convex loss functions.

Theorem 5 ([3] Lower bound for (ε, δ)-DP algorithms)**.** *Let $n, d \in \mathbb{N}$, $\varepsilon > 0$, and $\delta = o(\frac{1}{n})$. For every (ε, δ)-differentially private algorithm $\mathcal{A}$ (whose output is denoted by θ^{priv}), there is a dataset $\mathcal{D} \in \mathcal{X}^n$ such that, with probability at least $1/3$ (over the algorithm random coins), we must have*

$$\mathcal{L}_{\mathcal{D}}(\theta^{priv}) - \mathcal{L}_{\mathcal{D}}(\theta^*) = \Omega\left(\min\left\{1, \frac{d}{\varepsilon^2 n^2}\right\}\right), \tag{22}$$

where θ^* minimizes a constructed 1-Lipschitz, 1-strongly convex objective $\mathcal{L}_D(\theta)$ over convex set $\mathcal{C}$.

Our utility matches this lower bound upto the constant factor $\log(1/\delta)$, when assuming $\frac{\beta}{\lambda^2} = O(1)$. This improves upon the previous gradient perturbation methods [3, 23] by a factor of $\log(n)$, and matches the utility of previously know optimal ERM algorithm for Lipschitz smooth strongly convex loss functions, such as objective perturbation [5, 13] and output perturbation [25].

Utility gain from tight privacy guarantee As shown in Table 1, our utility guarantee for noisy GD is logarithmically better than that for noisy SGD in Bassily et al. [3], although the two algorithms are extremely similar. This is because we use our tight RDP guarantee, while Bassily et al. [3] use a composition-based privacy bound. More specifically, noisy SGD needs n^2 iterations to achieve the optimal utility, as shown in Table 1. This number of iterations is large enough for the composition-based privacy bound to grow above our RDP guarantee, thus leaving room for improving privacy utility trade-off, as we further discuss in Appendix. This concludes that our tight privacy guarantee enables providing a superior privacy-utility trade-off, for Lipschitz, strongly convex, and smooth loss functions.

Our algorithm also has significantly smaller gradient complexity than noisy SGD [3], for strongly convex loss functions, by a factor of $n/\log n$. We use a (moderately large) constant step-size, thus achieving fast convergence to optimal utility. However, noisy SGD [3] uses a decreasing step-size, thus requiring more iterations to reach optimal utility.

6 Conclusions

We have developed a novel theoretical framework for analyzing the dynamics of privacy loss for noisy gradient descent algorithms. Our theoretical results show that by hiding the internal state of the training algorithm (over many iterations over the whole data), we can tightly analyze the rate of information leakage throughout training, and derive a bound that converges at a exponential rate to a bound, which is significantly better than that of composition-based approaches.

Limitations. Our main result is a tight privacy guarantee for Noisy GD, but on smooth and strongly convex loss functions. The assumptions have obvious advantages in enabling the tightness and utility analysis, and are very similar to that of the prior work on privacy amplification by iteration [8]. However, they hinder the application of this analysis for non-smooth and non-convex loss functions, which are commonly used for example in deep learning. Relaxing these assumptions in our analysis would require a novel approach for satisfying LSI condition. Specifically, in Lemma 5, we prove LSI for Noisy GD under smoothness and strong convexity assumptions. However, this proof might be overly conservative, as the LSI condition does hold for continuous-time Langevin diffusion under non-smooth non-convex loss [14]. That said, for discrete processes, such as the unadjusted Langevin algorithms, we are only aware of the theoretical frameworks for LSI which require these assumptions [10]. This leaves an open challenge to extend the analysis of private state noisy gradient descent algorithm to non-convex non-smooth loss functions.

References

[1] 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.

[2] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

[3] 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.

[4] William M Bolstad. *Understanding computational Bayesian statistics*, volume 644. John Wiley & Sons, 2009.

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

[6] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society*, 2017.

[7] Alain Durmus, Eric Moulines, et al. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[8] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

[9] 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.

[10] Arun Ganesh and Kunal Talwar. Faster differentially private samplers via rényi divergence analysis of discretized langevin mcmc. *Advances in Neural Information Processing Systems*, 33, 2020.

[11] Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4): 1061–1083, 1975.

[12] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *arXiv preprint arXiv:2006.07709*, 2020.

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

[14] Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. *arXiv preprint arXiv:1902.00621*, 2019.

[15] Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, pages 956–964, 2016.

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

[17] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[18] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. *arXiv preprint arXiv:2101.04535*, 2021.

[19] Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, pages 982–990, 2014.

[20] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[21] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pages 8094–8106, 2019.

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

[23] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *arXiv preprint arXiv:1802.05251*, 2018.

[24] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502, 2015.

[25] Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. *arXiv preprint arXiv:1703.09947*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 1
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Complete proofs are in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]