Privacy of Noisy Stochastic Gradient Descent: More Iterations without More Privacy Loss

Anonymous Author(s)
Affiliation
Address
email

Abstract

A central issue in machine learning is how to train models on sensitive user data. Industry has widely adopted a simple algorithm: Stochastic Gradient Descent with noise (a.k.a. Stochastic Gradient Langevin Dynamics). However, foundational theoretical questions about this algorithm's privacy loss remain open—even in the seemingly simple setting of smooth convex losses over a bounded domain. Our main result resolves these questions: for a large range of parameters, we characterize the differential privacy up to a constant. This result reveals that all previous analyses for this setting have the wrong qualitative behavior. Specifically, while previous privacy analyses increase ad infinitum in the number of iterations, we show that after a small burn-in period, running SGD longer leaks no further privacy. Our analysis departs completely from previous approaches based on fast mixing, instead using techniques based on optimal transport (namely, Privacy Amplification by Iteration) and the sampled Gaussian mechanism (namely, Privacy Amplification by Sampling). Our techniques readily extend to other settings, e.g., strongly convex losses, non-uniform stepsizes, arbitrary batch sizes, and random or cyclic batches.

1 Introduction

Convex optimization is a fundamental task in machine learning. When models are learnt on sensitive data, privacy becomes a major concern—motivating a large body of work on differentially private convex optimization [8, 3, 4, 13, 20, 18, 27, 5]. In practice, the most common approach for training private models is Noisy-SGD, i.e., Stochastic Gradient Descent with noise added each iteration. This algorithm is simple and natural, has optimal utility bounds [3, 4], and is implemented in mainstream machine learning platforms such as Tensorflow (TF Privacy) [1], PyTorch (Opacus) [32], and JAX [6].

Yet, despite the simplicity and ubiquity of this Noisy-SGD algorithm, we do not understand basic questions about its *privacy loss*[1]—i.e., how sensitive the output of Noisy-SGD is with respect to the training data. Specifically:

Question 1.1. *What is the privacy loss of Noisy-SGD as a function of the number of iterations?*

Even in the seemingly simple setting of smooth convex losses over a bounded domain, this fundamental question has remained wide open. In fact, even more basic questions are open:

Question 1.2. *Does the privacy loss of Noisy-SGD increase ad infinitum in the number of iterations?*

The Supplementary Material contains a full version of this paper with proofs, more discussion, and extensions to strongly convex losses, cyclic batch updates, and non-uniform stepsizes.

[1] Throughout, we follow the literature by writing "privacy loss" to refer to the differential privacy parameters.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

The purpose of this paper is to understand these fundamental theoretical questions. Specifically, we resolve Questions 1.1 and 1.2 by characterizing the privacy loss of Noisy-SGD up to a constant factor in this (and other) settings for a large range of parameters. Below, we first provide context by describing previous analyses in §1.1, and then describe our result in §1.2 and techniques in §1.3.

1.1 Previous approaches and limitations

Although there is a large body of research devoted to understanding the privacy loss of Noisy-SGD, all existing analyses have (at least) one of the following two drawbacks.

Privacy bounds that increase ad infinitum. One type of analysis approach yields upper bounds on the DP (resp., Rényi DP) that scale as $\sqrt{T}$ (resp., T). This includes the original analyses of Noisy-SGD, which were based on the techniques of Privacy Amplification by Sampling and Advanced Composition [3, 2], as well as alternative analyses based on the technique of Privacy Amplification by Iteration [12]. A key issue with these analyses is that they increase unboundedly in the number of iterations T. This limits the number of iterations that Noisy-SGD can be run given a reasonable privacy budget, typically leading to suboptimal optimization error in practice. Is this a failure of existing analysis techniques or an inherent fact about the privacy loss of Noisy-SGD?

Privacy bounds that apply for large T and require strong additional assumptions. The second type of analysis approach yields convergent upper bounds on the privacy loss, but requires strong additional assumptions. This approach is based on connections to sampling. The high-level intuition is that Noisy-SGD is a discretization of a continuous-time algorithm with bounded privacy loss. Specifically, Noisy-SGD can be interpreted as the Stochastic Gradient Langevin Dynamics (SGLD) algorithm [30], which is a discretization of a continuous-time Markov process that has a simple stationary distribution—and this stationary distribution is equivalent to the exponential mechanism [21], and thus is itself differentially private under certain assumptions.

However, making this argument rigorous requires strong additional assumptions and/or the resolution of long-standing open questions about both the discretization error and mixing time of SGLD (in metrics relevant to privacy analysis like Rényi divergence). A large effort in this direction recently culminated in the breakthrough work by Chourasia et al. [9] which proves that *full batch*[2] Langevin dynamics (a.k.a., Noisy-GD rather than Noisy-SGD) has a privacy loss that converges as $T \to \infty$ in this setting where the smooth losses are additionally assumed to be *strongly convex*.

Unfortunately, the assumption of strong convexity seems unavoidable with these techniques. Indeed, in the absence of strong convexity, it is not even known if Noisy-SGD converges to a private stationary distribution, let alone if this convergence occurs in a reasonable amount of time. (The tour-de-force work [7] shows mixing in (large) polynomial time, but only in total variation distance which does not have implications for privacy.) There are fundamental challenges for proving such a result. From a sampling perspective, fast mixing is typically proven via expansion properties such as log Sobolev inequalities, and these break down without strong convexity. From an optimization perspective, SGD is only a weakly contractive process without strong convexity, which means that its instability increases with the number of iterations [17]—or in other words, it is plausible that Noisy-SGD could run for a long time while memorizing training data, which would of course mean it is not a privacy-preserving algorithm. As such, given state-of-the-art analyses in both sampling and optimization, it is unclear if the privacy loss of Noisy-SGD should even remain bounded; i.e., it is unclear what answer one should even expect for Question 1.2, let alone Question 1.1.

1.2 Contributions

The purpose of this paper is to resolve Questions 1.1 and 1.2. To state our result requires first recalling the parameters of the problem. Throughout, we prefer to state our results in terms of Rényi Differential Privacy (RDP); these RDP bounds are easily translated to DP bounds, as mentioned below in Remark 1.4. See the preliminaries section §2 for definitions and background on privacy.

[2] Recently, Ye and Shokri [31] and Ryffel et al. [25], in works concurrent to the present paper, extended the result of [9] to SGLD by removing the full batch assumption; we also obtain the same result by a direct extension of our (completely different) techniques, see the Supplementary Materials. Note that both papers [31, 25] still require strongly convex losses, and in fact state in their conclusions that removing this assumption is an open problem that "would pave the way for wide adoption by data scientists." Our main result resolves this question.

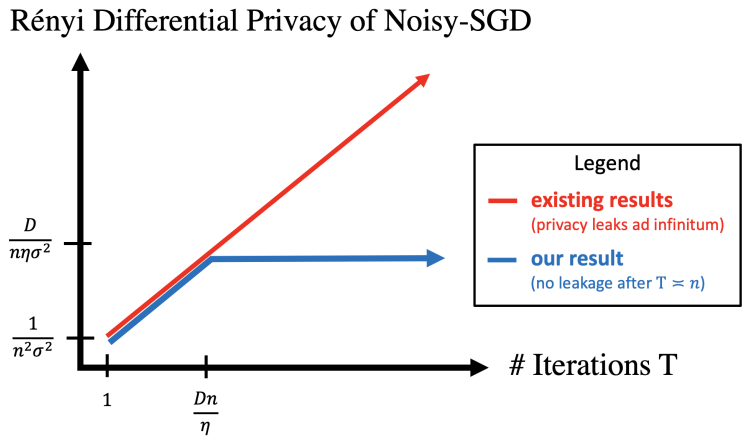

Figure 1: Even in the basic setting of smooth convex optimization over a bounded domain, existing implementations and theoretical analyses of Noisy-SGD—the standard algorithm for private optimization—leak privacy ad infinitum as the number of iterations T increases. Our main result establishes that after a small burn-in period of $\bar{T} = \Theta(n)$ iterations, there is no further privacy loss. For simplicity, this plot sets $L = 1$ (wlog by rescaling), $\alpha = O(1)$ (the regime in practice), and omits logarithmic factors; see the main text for our precise (and optimal) dependence on all parameters.

We consider the basic Noisy-SGD algorithm run on a dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$, where each x_i defines a convex L-Lipschitz and M-smooth loss function $f(\cdot; x_k)$ on a convex set $\mathcal{K}$ of diameter D. For any step size $\eta \leqslant 2/M$, batch size b, and initialization $\omega_0 \in \mathcal{K}$, we iterate T times the update $\omega_{t+1} \leftarrow \Pi_{\mathcal{K}}[\omega_t - \eta(G_t + Z)]$, where G_t denotes the average gradient vector on a random batch of size b, $Z \sim \mathcal{N}(0, \sigma^2 I_d)$ is an isotropic Gaussian, and $\Pi_{\mathcal{K}}$ denotes the Euclidean projection onto $\mathcal{K}$.

Known privacy analyses of Noisy-SGD give an (α, ε)-RDP upper bound of

$$\varepsilon \lesssim \frac{\alpha L^2}{n^2\sigma^2} T \tag{1.1}$$

which increases ad infinitum as the number of iterations $T \to \infty$ [3, 2, 12]. Our main result is a tight characterization of the privacy loss, answering Question 1.1. This result also answers Question 1.2 and shows that previous analyses have the wrong qualitative behavior: after a small burn-in period of $\bar{T} = \Theta(\frac{nD}{L\eta})$ iterations, Noisy-SGD leaks no further privacy. See Figure 1.

Theorem 1.3 (Informal statement of main result: tight characterization of the privacy of Noisy-SGD)**.** *For all small enough α, Noisy-SGD satisfies (α, ε)-RDP for*

$$\varepsilon \lesssim \frac{\alpha L^2}{n^2\sigma^2} \min\left\{T, \frac{Dn}{L\eta}\right\}. \tag{1.2}$$

Moreover, for a large range of parameters, this bound is tight.

Observe that the privacy bound in Theorem 1.3 is identical to the previous bound (1.1) when the number of iterations T is small, but stops increasing when $T \geqslant \bar{T}$. Intuitively, $\bar{T}$ can be interpreted as the smallest number of iterations required for two SGD processes run on adjacent datasets to, with reasonable probability, reach opposite ends of the constraint set.[3] In other words, $\bar{T}$ is effectively the smallest number of iterations before the final iterates could be maximally distinguishable.

To prove Theorem 1.3, we show a matching upper bound (in §3) and lower bound (in §4). These two bounds are formally stated as Theorems 3.1 and 4.1, respectively. See §1.3 for an overview of our techniques and how they depart from previous approaches.

We conclude this section with several remarks about Theorem 1.3.

[3]Details: Noisy-SGD updates on adjacent datasets differ by at most $\eta L/b$ if the different datapoint is in the current batch (i.e., with probability b/n), and otherwise are identical. Thus, in expectation and with high probability, it takes $\bar{T} = (Dn)/(\eta L)$ iterations for the two Noisy-SGD processes to differ by distance D.

Remark 1.4 (Tight DP characterization)**.** *While Theorem 1.3 characterizes the privacy loss of* Noisy-SGD *in terms of RDP, our results can be restated in terms of standard DP bounds. Specifically, by a standard conversion from RDP to DP [23, Proposition 1], if δ is not smaller than exponentially small in $-b\sigma/n$ and if the resulting bound is $\varepsilon \lesssim 1$, it follows that* Noisy-SGD *is (ε, δ)-DP for*

$$\varepsilon \lesssim \frac{L}{n\sigma}\sqrt{\min\left\{T, \frac{Dn}{L\eta}\right\}\log 1/\delta}. \tag{1.3}$$

A matching DP lower bound, for the same wide range of parameters as in Theorem 1.3, is proved along the way when we establish our RDP lower bound in §4.

Discussion of extensions. Our analysis techniques immediately extend to other settings; we highlight several below and provide full details in the Supplementary Materials.

Strongly convex losses. If the convexity assumption on the losses is replaced by strong convexity, then Noisy-SGD enjoys better privacy. Specifically, the bound in Theorem 1.3 extends identically except that the threshold $\bar{T}$ for no further privacy loss improves from linear to *logarithmic* in n, namely $\tilde{O}(\kappa)$ where κ denotes the condition number of the losses. This matches the independent results of [25, 31] (see Footnote 2), and uses completely different techniques from them.

Cyclic batch updates. Noisy-SGD is sometimes run with batches that are chosen cyclically rather than randomly. The bound in Theorem 1.3 extends identically except that there is an initial transient phase in which there is unavoidably a large privacy loss since the algorithm may choose to update using a sensitive datapoint in an early batch.

Non-uniform stepsizes. Noisy-SGD is sometimes run with decaying stepsizes $\eta_t = t^{-c}$ for $c \in (0, 1)$. The bound in Theorem 1.3 extends identically, with η replaced by η_t. This means that the privacy increases after the threshold, but at the slower rate of $T^c \ll T$ for RDP (resp., $T^{c/2} \ll T^{1/2}$ for DP).

Discussion of assumptions in Theorem 1.3.

Boundedness. While all utility analyses for Noisy-SGD necessarily exploit some form of diameter bound on the iterates, all previous privacy analyses are oblivious to any type of diameter bound. This is why all previous privacy analyses unavoidably increase ad infinitum (c.f., our lower bound in Theorem 1.3). Our privacy analysis is the first to exploit boundedness of the constraint set, which is crucial for obtaining privacy bounds that converge as $T \to \infty$. (We remark that in the strongly convex setting, boundedness is unnecessary; see the Supplementary Materials.)

Smoothness. The smoothness assumption on the losses can be relaxed by running Noisy-SGD on a smoothed version of the objective. This can be done using standard techniques, e.g., Gaussian convolution smoothing [12, §5.5], or Moreau-Yousida smoothing by replacing gradient steps with proximal steps [4, §4].

Convexity. Convexity appears to be essential for privacy bounds that do not increase ad infinitum in T. Our analysis uses convexity in an essential way to ensure that Noisy-SGD is a contractive process. However, convexity is too restrictive when training deep networks, and it is an interesting open question if the assumption of convexity be relaxed. Any such result appears to require entirely new techniques—if even true. The key technical challenge is that for any iterative process whose set of reachable fixed points is non-convex, there will be non-contractivity at the boundary between basins of attraction—and this precludes arguments based on Privacy Amplification by Iteration.

Mild assumptions on parameters. The lower bound makes the mild assumption that the diameter D of the decision set is neither asymptotically smaller than the movement size from one gradient step, nor asymptotically smaller than the standard deviation from $\bar{T}$ random increments of Gaussian noise $\mathcal{N}(0, \eta^2\sigma^2)$ so that the noise does not overwhelm the learning. The upper bound uses the technique of Privacy Amplification by Sampling, and thus inherits an upper bound assumption on α from the analysis of the Sampled Gaussian mechanism. This neither affects numerical bounds which can be computed for any α (see §2.3), nor does it affect the asymptotic (ε, δ)-DP bounds in most parameter regimes of interest. We also remark that our results require no restriction on α if the batches are chosen cyclically, and thus also if full-batch gradients are used (see the Supplementary Materials).

1.3 Techniques

Upper bound on privacy. Our analysis technique departs completely from recent approaches based on fast mixing. This allows us to bypass the many long-standing technical challenges for sampling from Gibbs distributions in the absence of strong convexity (details in §1.1).

Instead, our analysis combines the techniques of Privacy Amplification by Iteration and Privacy Amplification by Sampling. As discussed in §1.1, previous analyses based on these techniques yield loose privacy bounds that increase ad infinitum in T. Indeed, bounds based on Privacy Amplification by Sampling inevitably diverge since they "pay" for releasing the entire sequence of T iterates, each of which leaks more information about the private data [3, 2]. Privacy Amplification by Iteration avoids releasing the entire sequence by directly arguing about the final iterate; however, previous arguments based on this technique are unable to exploit a diameter bound on the constraint set, and therefore inevitably lead to privacy bounds which grow unboundedly in T [12].

Our analysis is based on the observation that when the iterates are constrained to a bounded domain, we can combine the techniques of Privacy Amplification by Iteration and Privacy Amplification by Sampling in order to only pay for the privacy loss from the final $\bar{T} = (Dn)/(L\eta)$ iterations. Specifically, we partition the execution of Noisy-SGD into the initial $T - \bar{T}$ iterations and the final $\bar{T}$ iterations. Recall that to show privacy, we must show that Noisy-SGD produces a similar output if run on a given dataset or another datast that is changed slightly. On one hand, if the partially learned model $\omega_{T-\bar{T}}$ were fixed, then the privacy loss from the last $\bar{T}$ iterations could be bounded by the standard analysis of Noisy-SGD based on Privacy Amplification by Sampling and Advanced Composition. On the other hand, irrespective of the input dataset, $\omega_{T-\bar{T}}$ lies in a set of diameter D. If the remaining iterations were now run on the same dataset, the Privacy Amplification by Iteration Argument can be used to show that the algorithm "forgets" this initial displacement by up to D. (This is formalized in Proposition 3.2, which is the first Privacy Amplification by Iteration argument that is not vacuous as $T \to \infty$). In other words, we can control the privacy loss if we change the dataset either in the first $T - \bar{T}$ iterations or in the final $\bar{T}$ iterations. At a high-level, our analysis proceeds by carefully defining sequences of iterates that allow us to run these arguments in parallel, letting us change the dataset in all iterations.

Lower bound on privacy. We construct two adjacent datasets on which the Noisy-SGD iterates are generated by random walks—one symmetric, one biased—constrained within an interval of length D. In this way, we reduce the question of how large is the privacy loss of Noisy-SGD, to the question of how distinguishable is a constrained symmetric random walk from a constrained biased random walk. The core technical challenge is that the distributions of the iterates of these random walks are intractable to reason about explicitly—due to the highly non-linear interactions between the projections and random increments. Briefly, our key technique here is a way to modify these processes so that on one hand, their distinguishability is essentially the same, and the other hand, no projections occur with high probability—allowing us to explicitly compute their distributions and thus also their distinguishability. Details in §4.

1.4 Other related work

Private sampling. The mixing time of (stochastic) Langevin dynamics has been extensively studied in recent years starting with [10, 11], and other more recent works have analyzed mixing in other notions of distance. From the point of view of privacy, more stringent notions of mixing are needed, and mixing in Rènyi divergence was studied by Vempala and Wibisono [29] and Ganesh and Talwar [14]. In addition to the aforementioned works, several recent works have studied algorithms for sampling from the distribution $\exp(\varepsilon \sum_i f(\omega, x_i))$ or its regularized versions from a privacy viewpoint. [22] proposed a mechanism of this kind that works when f may be unbounded and showed that it satisfies (ε, δ)-DP. Recently, [16] gave better privacy bounds for such a regularized exponential mechanism, and designed an efficient sampler based only on function evaluation. [15] showed a number of results about the continuous Langevin diffusion, showing that it gives optimal utility bounds for different variants of private optimization problems with different parameters.

Utility bounds. In the field of private optimization, one separately analyzes two properties of algorithms: their privacy and their utility (a.k.a., optimization error). These two properties can then be combined to obtain privacy-utility tradeoffs. Utility bounds for SGD are very well-understood [26] and these analyses have allowed understanding the utility of Noisy-SGD for empirical [3] and

population [4] losses. While in theory a single pass of SGD achieves the optimal population risk, in practice one typically benefits from multi-pass algorithms. The goal of the present paper is to develop tools to analyze the privacy loss of multi-epoch SGD.

2 Preliminaries

In this section, we recall relevant preliminaries about convex optimization (§2.1), differential privacy (§2.2), and two by-now-standard techniques for analyzing the privacy of optimization algorithms—namely, Privacy Amplification by Sampling (§2.3) and Privacy Amplification by Iteration (§2.4).

Notation. We write $\mathbb{P}_X$ to denote the law of a random variable X, and $\mathbb{P}_{X|Y=y}$ to denote the law of X given the event $Y = y$. We write $Z_{S:T}$ as shorthand for the vector concatenating $Z_S, \ldots, Z_T$. We write $f_{\#}\mu$ to denote the push-forward of a distribution μ under a (possibly random) function f, i.e., the law of $f(X)$ where $X \sim \mu$. We write $\lambda\mu + (1-\lambda)\nu$ to denote the mixture distribution that is μ with probability λ and ν with probability $1-\lambda$. We write $\mathcal{A}(\mathcal{X})$ to denote the output of an algorithm $\mathcal{A}$ run on input $\mathcal{X}$; this is a probability distribution if $\mathcal{A}$ is a randomized algorithm.

2.1 Convex optimization

Throughout, we minimize loss functions $f : \mathcal{K} \times \mathcal{X} \to \mathbb{R}$, which are convex in the first parameter (a.k.a., the machine learning model to train) and arbitrary in the second parameter (a.k.a., the data point). Throughout, the set $\mathcal{K}$ of possible models is a convex subset of $\mathbb{R}^d$. In order to establish both optimization and privacy guarantees, two additional assumptions are required on the loss functions. Recall that a differentiable function $g : \mathcal{K} \to \mathbb{R}^d$ is: L-Lipschitz if $\|g(\omega) - g(\omega')\| \leqslant L\|\omega - \omega'\|$ for all $\omega, \omega' \in \mathcal{K}$; and M-smooth if $\|\nabla g(\omega) - \nabla g(\omega')\| \leqslant M\|\omega - \omega'\|$ for all $\omega, \omega' \in \mathcal{K}$.

2.2 Differential privacy and Rényi differential privacy

In order to prove DP guarantees, we work with the related notion of Rényi Differential Privacy (RDP) introduced by [23], since RDP is more amenable to our analysis techniques and is readily converted to DP guarantees [23, Proposition 1]. RDP measures how distinguishable the output of an algorithm is when run on two "adjacent" datasets—i.e., two datasets which differ on at most one datapoint—in Rényi divergence. We therefore first recall the definition of Rényi divergence.

Definition 2.1 (Rényi divergence)**.** *The Rényi divergence between probability measures μ and ν of order $\alpha \in (1, \infty)$ is $\mathcal{D}_\alpha(\mu \parallel \nu) = \frac{1}{\alpha-1}\log\int(\mu(x)/\nu(x))^\alpha \nu(x)dx$, if $\mu \ll \nu$, and ∞ otherwise. Here $0/0 = 0$ and $x/0 = \infty$ for $x > 0$. The Rényi divergences of order $\alpha \in \{1, \infty\}$ are defined by continuity.*

Definition 2.2 (Rényi Differential Privacy)**.** *A randomized algorithm $\mathcal{A}$ satisfies (α, ε)-RDP if $\mathcal{D}_\alpha(\mathcal{A}(\mathcal{X}) \parallel \mathcal{A}(\mathcal{X}')) \leqslant \varepsilon$ for any two adjacent datasets $\mathcal{X}, \mathcal{X}'$.*

Following, we recall two basic properties of RDP that we use repeatedly in our analysis. Proofs can be found, e.g., in [28, Theorem 9] and [23, Proposition 3], respectively.

Lemma 2.3 (Post-processing property of Rényi divergence)**.** *For any Rényi parameter $\alpha \geqslant 1$, any (possibly random) function h, and any probability distributions μ, ν,*

$$\mathcal{D}_\alpha(h_{\#}\mu \parallel h_{\#}\nu) \leqslant \mathcal{D}_\alpha(\mu \parallel \nu).$$

Lemma 2.4 (Strong composition for RDP)**.** *For any Rényi parameter $\alpha \geqslant 1$, and any two sequences of random variables $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_k$,*

$$\mathcal{D}_\alpha\left(\mathbb{P}_{X_{1:k}} \middle\| \mathbb{P}_{Y_{1:k}}\right) \leqslant \sum_{i=1}^{k} \sup_{x_{1:i-1}} \mathcal{D}_\alpha\left(\mathbb{P}_{X_i|X_{1:i-1}=x_{1:i-1}} \parallel \mathbb{P}_{Y_i|Y_{1:i-1}=x_{1:i-1}}\right).$$

2.3 Privacy Amplification by Sampling

A core technique in the DP literature is Privacy Amplification by Sampling [19], which quantifies the idea that a private algorithm run on a small random sample of the input become more private. We recall a convenient statement of this from [24, Theorem 11] in terms of the RDP of the "Sampled Gaussian Mechanism" (which is a composition of subsampling and additive Gaussian noise).

Definition 2.5 (Rényi Divergence of the Sampled Gaussian Mechanism)**.** *For Rényi parameter $\alpha \geqslant 1$, mixing probability $q \in (0,1)$, and noise parameter $\sigma > 0$, define*

$$S_\alpha(q,\sigma) := \mathcal{D}_\alpha\left(\mathcal{N}(0,\sigma^2) \,\|\, (1-q)\mathcal{N}(0,\sigma^2) + q\mathcal{N}(1,\sigma^2)\right).$$

Lemma 2.6 (Bound on Rényi Divergence of the Sampled Gaussian Mechanism)**.** *Consider Rényi parameter $\alpha > 1$, mixing probability $q \in (0,1/5)$, and noise level $\sigma \geqslant 4$. If $\alpha \leqslant \alpha^*(q,\sigma)$, then*

$$S_\alpha(q,\sigma) \leqslant 2\alpha q^2/\sigma^2.$$

This bound restricts to Rényi parameter at most $\alpha^*(q,\sigma)$, which is defined to be the largest α satisfying $\alpha \leqslant M\sigma^2/2 - \log(\sigma^2)$ and $\alpha \leqslant (M^2\sigma^2/2 - \log(5\sigma^2))/(M + \log(q\alpha) + 1/(2\sigma^2))$, where $M := \log(1+1/(q(\alpha-1))$. While we use Lemma 2.6 to prove the asymptotic bounds in Theorem 1.3, we emphasize that i) our bounds can be computed numerically for *any* $\alpha \geqslant 1$ (see the Supplementary Materials), and ii) this upper bound does not preclude α from the typical parameter regime of interest, see the discussion in §1.2.

2.4 Privacy Amplification by Iteration

The Privacy Amplification by Iteration technique of [12] bounds the privacy loss of an iterative algorithm without "releasing" the entire sequence of iterates—unlike arguments based on Privacy Amplification by Sampling, c.f., §1.3. This technique applies to processes generated by a Contractive Noisy Iteration (CNI). We begin by recalling this definition, albeit in a slightly more general form that allows for *random* contractions, which is critical for analyzing the random updates of Noisy-SGD.

Definition 2.7 (Contractive Noisy Iteration)**.** *Consider a (random) initial state $X_0 \in \mathbb{R}^d$, a sequence of (random) contractions $\phi_t : \mathbb{R}^d \to \mathbb{R}^d$, and a sequence of noise distributions ξ_t. The Contractive Noisy Iteration $\mathrm{CNI}(X_0, \{\phi_t\}, \{\xi_t\})$ is the final iterate X_T produced by the process $X_{t+1} := \phi_{t+1}(X_t) + Z_{t+1}$, where Z_{t+1} is drawn independently from ξ_{t+1}.*

Proposition 2.8 (Original[4] PABI bound)**.** *Let X_T, X'_T denote the outputs of $\mathrm{CNI}(X_0, \{\phi_t\}, \{\xi_t\})$, $\mathrm{CNI}(X_0, \{\phi'_t\}, \{\xi_t\})$ where $\xi_t = \mathcal{N}(0, \sigma_t^2 I_d)$. Denote $s_t := \sup_x \|\phi_t(x) - \phi'_t(x)\|$, and consider any $a_1, \ldots, a_T$ such that $z_t := \sum_{i=1}^t (s_i - a_i)$ is non-negative for all t and satisfies $z_T = 0$. Then*

$$\mathcal{D}_\alpha\left(\mathbb{P}_{X_T} \,\|\, \mathbb{P}_{X'_T}\right) \leqslant \frac{\alpha}{2}\sum_{t=1}^T \frac{a_t^2}{\sigma_t^2}.$$

3 Upper bound on privacy

In this section, we prove the upper bound in Theorem 1.3. The formal statement of this result is as follows; see §1.2 and §2.3 for a discussion of the bound on α.

Theorem 3.1 (Privacy upper bound for Noisy-SGD)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set of diameter D, and let $\{f(\cdot, x)\}_{x\in\mathcal{X}}$ be a family of convex functions over $\mathcal{K}$ that are L-Lipschitz and M-smooth. For any number of iterations T, dataset size $n \in \mathbb{N}$, batch size $b \leqslant n$, stepsize $\eta \leqslant 2/M$, noise parameter $\sigma > 0$, and initialization $\omega_0 \in \mathcal{K}$, Noisy-SGD satisfies (α,ε)-RDP if $1 < \alpha \leqslant \alpha^*(\frac{b}{n}, \frac{b\sigma}{2\sqrt{2}L})$ and*

$$\varepsilon \lesssim \frac{\alpha L^2}{n^2\sigma^2}\min\left\{T, \frac{Dn}{L\eta}\right\}.$$

3.1 Privacy Amplification by Iteration bounds that are not vacuous as $T \to \infty$

Recall from the preliminaries section §2.4 that PABI arguments, while tight for a small number of iterations T, provide vacuous bounds as $T \to \infty$ (c.f., Proposition 2.8). The following proposition overcomes this by establishing privacy bounds which are *independent* of the number of iterations T. This result only requires additionally making the mild assumption that the CNI iterates are bounded.

[4] Strictly speaking, Proposition 2.8 is a generalization of [12, Theorem 22] since it allows for randomized contractions in the CNI. See the Supplementary Materials for a proof of this more general statement.

Proposition 3.2 (New PABI bound that is not vacuous as $T \to \infty$)**.** *Let X_T, X'_T, and s_t be as in Proposition 2.8. Suppose that all contractions ϕ_t, ϕ'_t have ranges lying in some set $\mathcal{K}$ of diameter D. Consider any $\tau \in \{0, \dots, T-1\}$ and $a_{\tau+1}, \dots, a_T$ such that $z_t := D + \sum_{i=\tau+1}^{t}(s_i - a_i)$ is non-negative for all t and satisfies $z_T = 0$. Then:*

$$\mathcal{D}_\alpha\left(\mathbb{P}_{X_T} \parallel \mathbb{P}_{X'_T}\right) \leqslant \frac{\alpha}{2} \sum_{t=\tau+1}^{T} \frac{a_t^2}{\sigma_t^2}.$$

Proof idea. We first recall the basic idea of the proof of the original PABI bound (Proposition 2.8). Briefly, that proof uses as a potential a certain *shifted Rényi divergence*, which allows for uncertainty in one of the two distributions as measured in optimal transport distance (making it geometrically aware). The original argument bounds the shifted divergence at iteration T (which is precisely the RDP in question), by the shifted divergence at iteration $T-1$, and so on to the shifted divergence at iteration 0 (which vanishes since Noisy-SGD always has the same initialization). In this way, by keeping track of the privacy loss at each iteration, PABI bounds the RDP. The main idea behind Proposition 3.2 is to change this argument by simply stopping the induction earlier. Specifically, unroll to iteration τ, and then use boundedness of the iterates to control the shifted divergence at that intermediate time τ. For brevity, the proof is deferred to the Supplementary Materials.

3.2 Proof sketch of Theorem 3.1 (full details in the Supplementary Materials)

Step 1: Coupling the iterates. Suppose $\mathcal{X} = \{x_1, \dots, x_n\}$ and $\mathcal{X}' = \{x'_1, \dots, x'_n\}$ are adjacent datasets; that is, they agree $x_i = x'_i$ on all indices $i \in [n] \setminus \{i^*\}$ except for at most one index $i^* \in [n]$. Consider running Noisy-SGD on either $\mathcal{X}$ or $\mathcal{X}'$ for T iterations—call the resulting iterates $\{W_t\}_{t=0}^T$ and $\{W'_t\}_{t=0}^T$, respectively—where we start from the same point $w_0 \in \mathcal{K}$ and couple the sequence of random batches $\{B_t\}_{t=0}^{T-1}$ and the random noise in each iteration. That is, $W_{t+1} = \Pi_{\mathcal{K}}[W_t - \frac{\eta}{b}\sum_{i\in B_t} \nabla_\omega f(W_t; x_i) + Y_t + Z_t]$ and $W'_{t+1} = \Pi_{\mathcal{K}}[W'_t - \frac{\eta}{b}\sum_{i\in B_t} \nabla_\omega f(W'_t; x_i) + Y_t + Z'_t]$ for all $t \leqslant T-1$, where we have split the noise into terms $Y_t \sim \mathcal{N}(0, \eta^2\sigma_1^2)$, $Z_t \sim \mathcal{N}(0, \eta^2\sigma_2^2)$, $Z'_t \sim \mathcal{N}(0, \eta^2\sigma_2^2) + \frac{\eta}{b}\left[\nabla_\omega f(W'_t; x_{i^*}) - \nabla_\omega f(W'_t; x'_{i^*})\right] \cdot \mathbb{1}_{i^* \in S_t}$, for any $\sigma_1, \sigma_2 > 0$ satisfying $\sigma_1^2 + \sigma_2^2 = \sigma^2$ (we set $\sigma_1 = \sigma_2 = \sigma/\sqrt{2}$ later for asymptotics).

Importantly, notice that in the definition of W'_{t+1}, the gradient is taken w.r.t. points in $\mathcal{X}$ rather than $\mathcal{X}'$; this is corrected via the bias in the noise term Z'_t. Notice also that this bias term is only realized with probability $1 - b/n$ because the probability that i^* is in a random size-b subset of $[n]$ is b/n.

Step 2: Construction of auxiliary CNI sequences. The processes $\{W_t\}_{t=0}^T$ and $\{W'_t\}_{t=0}^T$ are not Contractive Noisy Iterations (c.f., Definition 2.7) and thus we cannot directly apply the new PABI argument (Proposition 3.2). In this step, we modify these processes to construct CNI that are amenable to this analysis. For a parameter $\tau \in \{0, \dots, T-1\}$ to be chosen shortly, define $\{V_t\}_{t=0}^T$ and $\{V'_t\}_{t=0}^T$ by $V_t := W_t - \sum_{s=\tau}^{t-1} Z_s + Y_t$ and $V'_t := W'_t - \sum_{s=\tau}^{t-1} Z'_s + Y_t$. Define also the random functions $\phi_t, \phi'_t : \mathbb{R}^d \to \mathbb{R}^d$ by $\phi_t(v) := \Pi_{\mathcal{K}}[v - \frac{\eta}{b}\sum_{i\in B_t} \nabla_\omega f(v + \sum_{s=\tau}^{t-1} Z_s - Y_t; x_i) + (Z_t + \sum_{s=\tau}^{t-1} Z_s)] - \sum_{s=\tau}^{t} Z_s$ and $\phi'_t(v') := \Pi_{\mathcal{K}}[v' - \frac{\eta}{b}\sum_{i\in B_t} \nabla_\omega f(v' + \sum_{s=\tau}^{t-1} Z'_s - Y_t; x_i) + \left(Z'_t + \sum_{s=\tau}^{t-1} Z'_s\right)] - \sum_{s=\tau}^{t} Z'_s$ for all $t \leqslant T-1$. Note that $Z_t + \sum_{s=\tau}^{t-1} Z_s$ is different from $\sum_{s=\tau}^{t} Z_s$ since it contains Z_t even if $t < \tau$; this nuance ensures that Observation 3.3 holds for all $t \geqslant 0$, not just $t \geqslant \tau$.

We now make two observations that, together, establish that $\{V_t\}_{t=0}^T$ and $\{V'_t\}_{t=0}^T$ are CNI. The former is a straightforward calculation; the latter uses the fact that a stochastic gradient update on a smooth convex function is contractive. For brevity, proofs are deferred to the Supplementary Materials.

Observation 3.3. *For all t, it holds that $V_{t+1} = \phi_t(V_t) + Y_{t+1}$ and $V'_{t+1} = \phi'_t(V'_t) + Y_{t+1}$.*

Observation 3.4. *For all t, the maps ϕ_t, ϕ'_t are contractive.*

Step 3: Bounding the privacy loss of the auxiliary CNI sequences. Recall that we seek to upper bound $\mathcal{D}_\alpha(\mathbb{P}_{W_T} \parallel \mathbb{P}_{W'_T})$. We argue that:

$$\begin{aligned}
\mathcal{D}_\alpha\left(\mathbb{P}_{W_T} \parallel \mathbb{P}_{W'_T}\right) &\leqslant \mathcal{D}_\alpha\left(\mathbb{P}_{W_T, Z_{\tau:T-1}} \parallel \mathbb{P}_{W'_T, Z'_{\tau:T-1}}\right) \\
&\leqslant \mathcal{D}_\alpha\left(\mathbb{P}_{Z_{\tau:T-1}} \parallel \mathbb{P}_{Z'_{\tau:T-1}}\right) + \sup_z \mathcal{D}_\alpha\left(\mathbb{P}_{W_T|Z_{\tau:T-1}=z} \parallel \mathbb{P}_{W'_T|Z'_{\tau:T-1}=z}\right) \\
&\leqslant \mathcal{D}_\alpha\left(\mathbb{P}_{Z_{\tau:T-1}} \parallel \mathbb{P}_{Z'_{\tau:T-1}}\right) + \sup_z \mathcal{D}_\alpha\left(\mathbb{P}_{V_{T-1}|Z_{\tau:T-1}=z} \parallel \mathbb{P}_{V'_{T-1}|Z'_{\tau:T-1}=z}\right) \qquad (3.1)
\end{aligned}$$

Above, the first step is by the post-processing inequality for the Rényi divergence (Lemma 2.3). The second step is by the strong composition rule for the Rényi divergence (Lemma 2.4). The third step is also by the post-processing inequality (Lemma 2.3) because, conditional on $Z_{\tau:T-1} = Z'_{\tau:T-1} = z$, it holds that $W_T = h(V_{T-1})$ and $W'_T = h(V'_{T-1})$ for the same random function h.

Step 3a: Bounding the first term in (3.1), using Privacy Amplification by Sampling. In words, this is the privacy loss from releasing $T - \tau$ noisy gradients. This can be bounded by Privacy Amplification by Sampling (Lemma 2.6) and Advanced Composition (Lemma 2.4) via a standard argument [2]. This yields the bound $(T-\tau)S_\alpha\left(\frac{b}{n}, \frac{b\sigma_2}{2L}\right)$; full details in the Supplementary Materials.

Step 3b: Bounding the second term in (3.1), using Privacy Amplification by Iteration. Observe that conditional on $Z_{\tau:T-1} = Z'_{\tau:T-1}$, the maps $\phi_t \equiv \phi'_t$ are identical for all $t \geqslant \tau$. Thus, by Observations 3.3 and 3.4, the iterates V_{T-1}, V'_{T-1} are produced by CNI with the same contraction mappings $\phi_t \equiv \phi'_t$ for all $t \geqslant \tau$. Thus we may apply the new Privacy Amplification by Iteration bound (Proposition 3.2) with $s_t \equiv 0$ and $a_t \equiv D/(T-\tau-1)$ to obtain the bound $\alpha D^2/(2\eta^2\sigma_1^2(T-\tau-1))$.

Step 4: Putting the bounds together. By plugging into (3.1) the bounds in Steps 3a-3b, we conclude that Noisy-SGD is (α, ε)-RDP for

$$\varepsilon \leqslant \min_{\tau \in \{0,\ldots,T-1\}} \left\{ (T-\tau)\, S_\alpha\left(\frac{b}{n}, \frac{b\sigma_2}{2L}\right) + \frac{\alpha D^2}{2\eta^2\sigma_1^2(T-\tau-1)} \right\} \tag{3.2}$$

By bounding $S_\alpha(\frac{b}{n}, \frac{b\sigma_2}{2L})$ using Lemma 2.6, setting $\sigma_1 = \sigma_2 = \frac{\sigma}{\sqrt{2}}$, setting $\tau = T - \Theta(\frac{Dn}{L\eta})$ if $\frac{D}{\eta L} \lesssim \frac{T}{n}$, and otherwise using the simple bound (1.1) which scales linearly in T, and simplifying, we obtain the desired privacy bound on ε, proving Theorem 3.1; full details in the Supplementary Materials.

4 Lower bound on privacy

In this section, we prove the lower bound in Theorem 1.3. This is stated formally below and holds even for linear loss functions in one dimension, in fact even when all but one of the loss functions are zero. See §1.2 for a discussion of the mild assumptions on the diameter. Below let $\bar{T} := 0.75\frac{Dn}{L\eta}$.

Theorem 4.1 (Privacy lower bound for Noisy-SGD)**.** *There exist universal constants c_D, c_σ, c_α, $\bar{\alpha}$ and a family $\{f(\cdot, x)\}_{x \in \mathcal{X}}$ of L-Lipschitz linear functions over the interval $\mathcal{K} = [-D/2, D/2]$ such that the following holds. Consider running Noisy-SGD from arbitrary initialization ω_0 with any parameters satisfying $D \geqslant c_D \eta L$ and $\sigma^2 \leqslant c_\sigma D^2/(\eta^2 \bar{T})$. Then Noisy-SGD is not $(\bar{\alpha}, \varepsilon)$-RDP for*

$$\varepsilon \leqslant c_\alpha \frac{\bar{\alpha} L^2}{n^2 \sigma^2} \min\left\{T, \bar{T}\right\}. \tag{4.1}$$

Construction. Consider datasets $\mathcal{X} = \{x_1, \ldots, x_{n-1}, x_n\}$ and $\mathcal{X}' = \{x_1, \ldots, x_{n-1}, x'_n\}$ which differ only on x'_n, and corresponding functions which are all zero $f(\cdot; x_1) = \cdots = f(\cdot; x_n) = 0$, except for $f(\omega; x'_n) = L(D - \omega)$. Clearly these functions are linear and L-Lipschitz. The intuition behind this construction is that running Noisy-SGD on $\mathcal{X}$ or $\mathcal{X}'$ generates a random walk that is clamped to stay within the interval $\mathcal{K}$—but with the key difference that running Noisy-SGD on dataset $\mathcal{X}$ generates a *symmetric* random walk $\{\omega_t\}$, whereas running Noisy-SGD on dataset $\mathcal{X}'$ generates a *biased* random walk $\{\omega'_t\}$ that biases right with probability b/n each step. That is,

$$\omega_{t+1} = \Pi_{\mathcal{K}}\Big[\omega_t + Z_t\Big] \qquad \text{and} \qquad \omega'_{t+1} = \Pi_{\mathcal{K}}\Big[\omega'_t + Y_t + Z_t\Big]$$

where the processes are initialized at $\omega_0 = \omega'_0 = 0$, each random increment $Z_t \sim \mathcal{N}(0, \eta^2\sigma^2)$ is an independent Gaussian, and each bias Y_t is $\eta L/b$ with probability b/n and otherwise is 0.

Key obstacle for analysis: intractability of reasoning about the iterates' distributions. The high-level intuition behind this construction is simple to state: the bias of the random walk $\{\omega'_t\}$ makes it distinguishable (to the minimax optimal extent, as show) from the symmetric random walk $\{\omega_t\}$. However, making this intution precise is challenging because the distributions of the iterates ω_t, ω'_t are intractable to reason about explicitly—due to the highly non-linear interactions between the projections and the random increments. Thus we must establish the distinguishability of ω_T, ω'_T in a way that avoids reasoning explicitly about their distributions. We show how to do this in the Supplementary Materials and give a full proof of Theorem 4.1 there.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. TensorFlow: A system for large-scale machine learning. In *Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Conference on Computer and Communications Security*, pages 308–318, 2016.

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

[4] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, pages 11282–11291, 2019.

[5] Raef Bassily, Cristobal Guzman, and Anupama Nandi. Non-Euclidean differentially private stochastic convex optimization. *arXiv:2103.01278*, 2021.

[6] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

[7] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with Projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

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

[9] Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential privacy dynamics of Langevin diffusion and noisy gradient descent. In *Advances in Neural Information Processing Systems*, volume 34, pages 14771–14781, 2021.

[10] Arnak Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 6 2017.

[11] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25:2854–2882, 11 2019.

[12] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *Symposium on Foundations of Computer Science*, pages 521–532. IEEE, 2018.

[13] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Symposium on the Theory of Computing*, pages 439–449, 2020.

[14] Arun Ganesh and Kunal Talwar. Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC. In *Advances in Neural Information Processing Systems*, volume 33, pages 7222–7233, 2020.

[15] Arun Ganesh, Abhradeep Thakurta, and Jalaj Upadhyay. Langevin diffusion: An almost universal algorithm for private Euclidean (convex) optimization. arXiv:2204.01585v2, 2022.

[16] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. arXiv:2203.00263v1, 2022.

[17] 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, 2016.

[18] Prateek Jain and Abhradeep Guha Thakurta. (Near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484. PMLR, 2014.

[19] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

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

[21] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Symposium on Foundations of Computer Science*, pages 94–103, 2007.

[22] Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[23] Ilya Mironov. Rényi differential privacy. In *Computer Security Foundations Symposium*, pages 263–275. IEEE, 2017.

[24] Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the Sampled Gaussian Mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[25] Theo Ryffel, Francis Bach, and David Pointcheval. Differential privacy guarantees for stochastic gradient Langevin dynamics. arXiv:2201.11980, 2022.

[26] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Conference on Computational Learning Theory*, 2009.

[27] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Nearly optimal private Lasso. In *Advances in Neural Information Processing Systems*, volume 28, pages 3025–3033, 2015.

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

[29] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32*, pages 8092–8104. 2019.

[30] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, page 681–688, Madison, WI, USA, 2011.

[31] Jiayuan Ye and Reza Shokri. Differentially private learning needs hidden state (or much faster convergence. arXiv:2203.05363, 2022.

[32] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

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]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] On the contrary, the main results of this paper have positive societal impacts: we show that machine learning models can be trained for longer while still protecting the privacy of sensitive user data.
   (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]
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]