An Improved Privacy and Utility Analysis of Differentially Private SGD with Bounded Domain and Smooth Losses

Hao Liang[1], **Wanrong Zhang**[2], **Xinlei He**[1], **Kaishun Wu**[1], **Hong Xing**[1,3*]
[1]Information Hub, The Hong Kong University of Science and Technology (Guangzhou)
[2]Harvard University
[3]Department of ECE, The Hong Kong University of Science and Technology
hliang346@connect.hkust-gz.edu.cn, wanrongzhang@fas.harvard.edu, {xinleihe, wuks}@hkust-gz.edu.cn, hongxing@ust.hk

Abstract

Differentially Private Stochastic Gradient Descent (DPSGD) is widely used to protect sensitive data during the training of machine learning models, but its privacy guarantee often comes at a large cost of model performance due to the lack of tight theoretical bounds quantifying privacy loss. While recent efforts have achieved more accurate privacy guarantees, they still impose some assumptions prohibited from practical applications, such as convexity and complex parameter requirements, and rarely investigate in-depth the impact of privacy mechanisms on the model's utility. In this paper, we provide a rigorous privacy characterization for DPSGD with general L-smooth and non-convex loss functions, revealing converged privacy loss with iteration in bounded-domain cases. Specifically, we track the privacy loss over multiple iterations, leveraging the noisy smooth-reduction property, and further establish comprehensive convergence analysis in different scenarios. In particular, we show that for DPSGD with a bounded domain, (i) the privacy loss can still converge without the convexity assumption, (ii) a smaller bounded diameter can improve both privacy and utility simultaneously under certain conditions, and (iii) the attainable big-O order of the privacy utility trade-off for DPSGD with gradient clipping (DPSGD-GC) and for DPSGD-GC with bounded domain (DPSGD-DC) and strongly convex population risk function, respectively. Experiments via membership inference attack (MIA) in a practical setting validate insights gained from the theoretical results.

Code — https://github.com/HauLiang/DPSGD-DC
Extended version — https://arxiv.org/abs/2502.17772

1 Introduction

Differentially Private Stochastic Gradient Descent (DPSGD) (Abadi et al. 2016) has emerged as the leading defense mechanism to protect personal sensitive data in training of machine learning models. However, achieving good performance with DPSGD often comes with a significant privacy cost. A fundamental question, therefore, is how to precisely quantify the privacy loss associated with DPSGD.

Previous methods for quantifying privacy loss include strong composition (Dwork, Rothblum, and Vadhan 2010; Bassily, Smith, and Thakurta 2014; Kairouz, Oh, and Viswanath 2015), moments accountant (Abadi et al. 2016), Rényi Differential Privacy (RDP) (Mironov 2017; Mironov, Talwar, and Zhang 2019), and Gaussian Differential Privacy (GDP) (Dong, Roth, and Su 2022), along with several numerical composition methods (Koskela, Jälkö, and Honkela 2020; Gopi, Lee, and Wutschitz 2021). These methods primarily rely on composition theorems, assuming that all intermediate models are revealed during the training procedure, which leads to an overestimation of privacy loss. While numerical composition methods aim to tightly characterize the privacy loss, they still operate under this same assumption.

To address this overestimation, recent works have focused solely on the privacy guarantees of the final output. For instance, the privacy amplification by iteration (Feldman et al. 2018) demonstrated that withholding intermediate results significantly enhances privacy guarantees for smooth and convex objectives. Building upon this, Chourasia, Ye, and Shokri (2021) suggest that the privacy loss of DPGD, the full batch version of DPSGD, may converge exponentially fast for smooth and strongly convex objectives. Furthermore, results by Ye and Shokri (2022) as well as Ryffel, Bach, and Pointcheval (2022) extended this analysis to assess the privacy loss of DPSGD, although both studies rely on the assumption of strong convexity.

More recently, the work by Altschuler and Talwar (2022) and its extension (Altschuler, Bok, and Talwar 2024) established a constant upper bound on privacy loss after a burn-in period for Lipschitz continuous and smooth convex losses over a bounded domain. However, this analytical result is limited by its reliance on the convexity assumption and strict restrictions on the Rényi parameter α, which hinders its broader applicability. In order to relax several strong assumptions, Kong and Ribero (2024) provided an analysis of weakly-convex smooth losses in the case where data is traversed cyclically. Later, Chien and Li (2024) suggest precisely tracking the privacy leakage incurred before reaching the constant upper bound by solving an optimization problem. However, this result is formulated as a complex optimization problem rather than a closed-form expression, making it hard to operationalize. Notably, most recent methods necessitate double clipping of both gradients and parameters due to the additional bounded domain assumption.

*Corresponding author: Hong Xing (hongxing@ust.hk).

These methods, however, do not provide a thorough utility analysis or experimental results, leaving their practical performance and trade-offs underexplored.

We outline the main contributions of this paper below and provide a comparison of the key assumptions and theoretical results with the most relevant works in Table 1.

1.1 Contributions

In this paper, we present a precise analytical characterization of privacy bounds for DPSGD that focus on smooth loss functions without relying on convexity assumptions or restrictive Rényi parameter conditions. Our general results encompass DPSGD applied to both unbounded and bounded domains. Additionally, we establish utility guarantees based on the derived RDP bounds, offering an intuitive perspective on privacy-utility trade-offs. To demonstrate the practicality and validity of our theoretical findings, we conduct extensive numerical simulations, which confirm the effectiveness and rationality of the proposed bounds. Our contributions are as follows:

- We analyze the noisy smooth-reduction behavior of the shifted Rényi divergence for smooth objectives. This analysis enables the derivation of closed-form RDP guarantees for DPSGD applied to both unbounded and bounded domains.
- We establish the convergence behavior for DPSGD with smooth loss functions in unbounded domains and strongly convex smooth loss functions in bounded domains. Our results provide the privacy-utility trade-offs under the computed RDP bounds.
- To validate these theoretical findings, we examine the privacy parameters estimated by the membership inference attack (MIA). Extensive experiments illustrate the effectiveness and rationality of the proposed bounds.

To clearly illustrate the effectiveness of our proposed privacy bound, we here provide detailed comparisons with several prominent existing approaches, including: a) Feldman et al. (2018), (b) the combined analysis by Mironov (2017) and Mironov, Talwar, and Zhang (2019), c) Altschuler and Talwar (2022), and d) Kong and Ribero (2024). The detailed setting can be found in Appendix D.1. As shown in Figure 1, our analysis demonstrates strict improvement on all existing privacy bounds, except for Altschuler and Talwar (2022). The reason for which their bound appeared tighter is their additional assumptions—including convexity and more restrictive conditions (as summarized in Table 1)—while our analysis relies only on the smoothness of the loss function and far weaker assumptions.

1.2 Other Related Works

In addition to privacy analysis, the utility (convergence) of private optimization algorithms has been extensively studied. This line of works typically focus on understanding how the number of iterations affects the convergence behavior of the algorithm. Below, we provide a brief review of utility analysis for DPSGD.

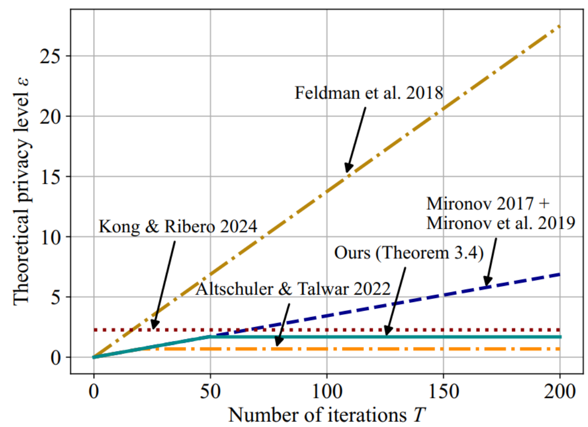


Figure 1: Comparison of our theoretical (α, ε)-RDP bound for DPSGD-DC with existing approaches. Detailed assumptions required by each method have been summarized in Table 1.

In 2014, Bassily, Smith, and Thakurta (2014) analyzed the optimal utility guarantees of DPSGD under the assumption of Lipschitz continuity, considering both convex and strongly convex cases. Then, based on the additional assumption of the gradient distribution, Chen, Wu, and Hong (2020) studied the convergence of DPSGD with gradient clipping (DPSGD-GC) and derived a utility bound for the non-convex setting. The work by Song et al. (2021) explored the convergence of DPSGD-GC for generalized linear models, noting that, in the worst case, the utility can remain constant relative to the original objective. Later, Fang et al. (2023) further refined this analysis for smooth and unconstrained problems, providing more precise convergence results. However, many of these studies fix a specific value for the clipping threshold C, which may be adjusted due to privacy requirements.

More recently, Koloskova, Hendrikx, and Stich (2023) characterized the convergence guarantees for DPSGD-GC across various clipping thresholds C in the non-convex setting. While this work provides valuable convergence insights for DPSGD-GC, recent privacy characterizations have introduced the need for double clipping—clipping both gradients and parameters—due to the additional assumption of bounded domains. The convergence analysis involving double clipping has not been thoroughly explored in the existing literature.

1.3 Organization

The rest of this paper is organized as follows: In the next section, we recall the relevant preliminaries. Our main results are presented in Section 3. Numerical results are provided in Section 4. Finally, Section 5 concludes with a discussion of future research directions motivated by our findings. Proof details are deferred to Appendices.

Reference	Assumptions	Domain	Privacy Guarantee	Utility Analysis?
Feldman et al. (2018)	convex, L-smooth, M-Lipschitz, $\eta \leq 2/L$	unbounded	$\mathcal{O}\left(\frac{\alpha M^2}{b^2 \sigma_{\text{DP}}^2} T\right)$	$\surd$
Altschuler and Talwar (2022)	convex, L-smooth, M-Lipschitz, $\eta \leq 2/L$ $b \leq n/5$, $\sigma_{\text{DP}} > 8\sqrt{2}M/b$, $\alpha \leq \alpha^*(b/n, \frac{b\sigma_{\text{DP}}}{2\sqrt{2}M})$	bounded	$\mathcal{O}\left(\frac{\alpha M^2}{n^2 \sigma_{\text{DP}}^2} \min\left\{T, \frac{Dn}{\eta M}\right\}\right)$	$\times$
Kong and Ribero (2024)[‡]	m-weakly convex, L-smooth, $\eta \leq \frac{1}{2(m+L)}$	bounded	$\mathcal{O}\left(\frac{\alpha}{\eta^2 \sigma_{\text{DP}}^2} (D\sqrt{1 + 2\eta m[1 + \frac{m}{2(L+m)}]} + \frac{\eta C}{b})^2\right)$	$\times$
Chien and Li (2024)	M-Lipschitz L-smooth or (L, λ)-Hölder continuous gradient	unbounded /bounded	w/o analytical form	$\times$
Ours[†]	L-smooth	unbounded	$\mathcal{O}\left(\frac{\alpha C^2}{nb\sigma_{\text{DP}}^2} T\right)$	$\surd$
Ours[†]	L-smooth	bounded	$\mathcal{O}\left(\frac{\alpha C^2}{nb\sigma_{\text{DP}}^2} \min\left\{T, \frac{(1+\eta L)^2 nbD^2}{\eta^2 C^2}\right\}\right)$	$\surd$

Table 1: Comparison of the (α, ε)-RDP guarantee and assumptions needed by different works for DPSGD, where b is the batch size, n is the dataset size, η is the step size, C is the gradient clipping norm bound, D is the diameter of the parameter domain, σ_{DP} is the noise scale, and T is the number of iterations. "‡" is exclusively suitable for cyclic data traversal cases. "†" indicates that a tighter bound can be obtained under additional assumptions on the Rényi parameters. $\alpha^*(q, \sigma)$ is defined as the largest α that satisfies both $\alpha \leq K\sigma^2/2 - 2\log\sigma$ and $\alpha \leq \left(K^2\sigma^2/2 - \log 5 - 2\log\sigma\right) / \left(K + \log(q\alpha) + 1/(2\sigma^2)\right)$, where $K = \log(1 + 1/(q(\alpha - 1)))$.

2 Preliminaries

In this section, we introduce the foundational concepts and definitions relevant to our analysis. We start with our notations, which will be used throughout this paper.

Notations. Let $\Pr[\cdot]$ denote the probability of a random event, and $\mathbb{P}_{\boldsymbol{\mu}}$ be the law of a random variable $\boldsymbol{\mu}$. We refer to two datasets $\mathcal{D}$ and $\mathcal{D}'$ as *adjacent* if they differ from each other by adding or removing only one data point.

2.1 Rényi Differential Privacy (RDP)

We first recall the formal definition of differential privacy (DP) and RDP.

Definition 2.1. (Differential privacy (Dwork et al. 2006)). For $\epsilon \geq 0, \delta \in [0, 1]$, a randomized mechanism $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$ is (ϵ, δ)-*DP* if, for every pair of adjacent datasets, $\mathcal{D}, \mathcal{D}' \subseteq \mathcal{X}$, and for any subset of outputs $\mathcal{S} \subseteq \mathcal{Y}$, we have

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta. \quad (1)$$

Throughout this paper, we use RDP, a more efficient approach for tracking privacy loss than DP, as our primary framework for privacy analysis. RDP provides a relaxation of DP based on *Rényi divergence*, which is defined as follows.

Definition 2.2. (Rényi divergence (Rényi 1961)). For adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$, a randomized mechanism $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$, and an outcome $s \in \mathcal{Y}$, the Rényi divergence of a finite order $\alpha \neq 1$ between $\mathcal{M}(\mathcal{D})$ and $\mathcal{M}(\mathcal{D}')$ is defined as

$$\begin{aligned} &D_\alpha(\mathbb{P}_{\mathcal{M}(\mathcal{D})} \| \mathbb{P}_{\mathcal{M}(\mathcal{D}')}) \\ &= \frac{1}{\alpha - 1} \log \mathbb{E}_{s \sim \mathbb{P}_{\mathcal{M}(\mathcal{D}')}} \left\{ \left(\frac{\Pr[\mathcal{M}(\mathcal{D}) = s]}{\Pr[\mathcal{M}(\mathcal{D}') = s]} \right)^\alpha \right\}. \end{aligned} \quad (2)$$

On the grounds of Rényi divergence, RDP is defined by the following definition.

Definition 2.3. (Rényi differential privacy (Mironov 2017)). For $\alpha > 1, \varepsilon \geq 0$, a randomized mechanism $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$ satisfies (α, ε)-*RDP* if, for every pair of adjacent datasets, $\mathcal{D}, \mathcal{D}' \subseteq \mathcal{X}$, it holds that

$$D_\alpha(\mathbb{P}_{\mathcal{M}(\mathcal{D})} \| \mathbb{P}_{\mathcal{M}(\mathcal{D}')}) \leq \varepsilon. \quad (3)$$

Note that RDP can be easily transformed into an equivalent characterization in terms of DP, as demonstrated by the following lemma.

Lemma 2.4. (From (α, ε)-RDP to (ϵ, δ)-DP (Mironov 2017)). *If $\mathcal{M}$ is an (α, ε)-RDP mechanism, it is also $(\varepsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-DP for any $0 < \delta < 1$.*

Based on the assumption that intermediate training models are not revealed, the technique *privacy amplification by iteration* substantially improves the privacy guarantee analysis (Feldman et al. 2018), which is grounded on the concept of *shifted Rényi divergence*, as follows.

Definition 2.5. (Shifted Rényi divergence (Feldman et al. 2018)). Let $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ be two random variables. Then, for any $z \geq 0$ and $\alpha > 1$, the z-shifted Rényi divergence is defined as

$$\mathcal{D}_\alpha^{(z)}(\mathbb{P}_{\boldsymbol{\mu}} \| \mathbb{P}_{\boldsymbol{\nu}}) = \inf_{\mathbb{P}_{\boldsymbol{\mu}'} : W_\infty(\mathbb{P}_{\boldsymbol{\mu}}, \mathbb{P}_{\boldsymbol{\mu}'}) \leq z} \mathcal{D}_\alpha(\mathbb{P}_{\boldsymbol{\mu}'} \| \mathbb{P}_{\boldsymbol{\nu}}), \quad (4)$$

where $W_\infty(\cdot, \cdot)$ denotes the ∞-Wasserstein distance.[1]

The privacy amplification by iteration utilizes the following lemma, which we frequently employ in the sequel.

Lemma 2.6. (Shift-reduction (Feldman et al. 2018)). *Let $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ be two random variables. Then, for any $a \geq 0$ and $z \geq 0$, we have*

$$\mathcal{D}_\alpha^{(z)}(\mathbb{P}_{\boldsymbol{\mu}} * \mathbb{P}_{\boldsymbol{\zeta}} \| \mathbb{P}_{\boldsymbol{\nu}} * \mathbb{P}_{\boldsymbol{\zeta}}) \leq \mathcal{D}_\alpha^{(z+a)}(\mathbb{P}_{\boldsymbol{\mu}} \| \mathbb{P}_{\boldsymbol{\nu}}) + \frac{\alpha a^2}{2\sigma^2}, \quad (5)$$

where $\boldsymbol{\zeta}$ is a multi-variate Gaussian random variable with zero mean and covariance given by $\sigma^2 \boldsymbol{I}_d$, denoted by $\boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d)$; and $\mathbb{P}_{\boldsymbol{\mu}} * \mathbb{P}_{\boldsymbol{\zeta}}$ denotes the distribution of the sum $\boldsymbol{\mu} + \boldsymbol{\zeta}$ with $\boldsymbol{\mu}$ and $\boldsymbol{\zeta}$ being drawn independently.

Algorithm 1: Differentially Private Stochastic Gradient Descent with Double Clipping (DPSGD-DC)

Input: Dataset $\mathcal{D}$, stochastic loss function $l_\xi(\boldsymbol{\theta}) : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}$, learning rate η, noise scale σ_{DP}, dataset size n, batch size b, gradient norm bound C, parameter domain $\mathcal{K}$ with diameter D, number of iterations T;
Initialize $\boldsymbol{\theta}_0 \leftarrow \mathbf{0}$ and $t \leftarrow 0$;
repeat
1) batch sampling:
take a random mini-batch $\mathcal{B}_t$ with sampling probability $q = b/n$;
2) compute and clip the gradients:
$\nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{\theta}_t; \mathcal{D}) \leftarrow \frac{1}{b} \sum_{\xi \in \mathcal{B}_t} \text{clip}_C(\nabla l_\xi(\boldsymbol{\theta}_t))$,
where $\text{clip}_C(\boldsymbol{x}) = \boldsymbol{x} \cdot \min(1, \frac{C}{\|\boldsymbol{x}\|})$;
3) update and project the parameters:
$\boldsymbol{\theta}_{t+1} \leftarrow \Pi_{\mathcal{K}}(\boldsymbol{\theta}_t - \eta(\nabla \mathcal{L}_{B_t}(\boldsymbol{\theta}_t; \mathcal{D}) + \boldsymbol{\zeta}_t))$,
where $\Pi_{\mathcal{K}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{x} \in \mathcal{K}} \|\boldsymbol{\theta} - \boldsymbol{x}\|$ and $\boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{DP}}^2 \boldsymbol{I}_d)$;
4) update the iteration counter:
$t \leftarrow t + 1$;
until $t > T$
Output: Final-round model parameters $\boldsymbol{\theta}_T$.

2.2 DPSGD with Double Clipping (DPSGD-DC)

In this paper, we also consider the DPSGD with both gradient clipping and parameter projection (Algorithm 1), termed as *DPSGD-DC*. This method begins with applying the vanilla SGD procedure using gradient clipping and Gaussian perturbation for model updates. Then, the updated parameters are projected into a bounded domain $\mathcal{K} = \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\| \le D\}$ as follows[2]

$$\boldsymbol{\theta}_{t+1} = \Pi_{\mathcal{K}}\left(\boldsymbol{\theta}_t - \eta(\nabla \mathcal{L}_{B_t}(\boldsymbol{\theta}_t; \mathcal{D}) + \boldsymbol{\zeta}_t)\right), \qquad (6)$$

with

$$\Pi_{\mathcal{K}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{x} \in \mathcal{K}} \|\boldsymbol{\theta} - \boldsymbol{x}\|, \qquad (7)$$

where η denotes the learning rate; $\boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{DP}}^2 \boldsymbol{I}_d)$ denotes a multi-variate Gaussian random variable at iteration t; the loss function accrued on a model using data sample $\xi \in \mathcal{D}$ is defined as $l_\xi(\cdot) : \mathcal{D} \times \mathbb{R}^d \mapsto \mathbb{R}$; and $\nabla \mathcal{L}_{B_t}(\boldsymbol{\theta}_t; \mathcal{D})$ is the clipped gradient of the loss function averaging over a mini-batch $\mathcal{B}_t$, i.e., $\nabla \mathcal{L}_{B_t}(\boldsymbol{\theta}_t; \mathcal{D}) = \frac{1}{b} \sum_{\xi \in \mathcal{B}_t} \text{clip}_C(\nabla l_\xi(\boldsymbol{\theta}_t))$, with $b = |\mathcal{B}_t|$ denoting the size of the mini-batch and $\text{clip}_C(\boldsymbol{x}) = \boldsymbol{x} \cdot \min(1, \frac{C}{\|\boldsymbol{x}\|})$. Note that if $\mathcal{K} = \mathbb{R}^d$, it reduces to DPSGD with only gradient clipping, i.e., *DPSGD-GC*.

3 Main Theoretical Results

In this section, we construct RDP bounds to analyze privacy loss associated with releasing only the final-round model of DPSGD-GC and DPSGD-DC, respectively. To gain insights into how such privacy loss affects their respective training performance, we further provide utility analysis and derive the corresponding privacy-utility trade-offs.

[1]See Definition A.7 in Appendix A.

[2]A detailed discussion of the bounded domain assumption is provided in Appendix E.1.

Assumption 3.1. (L-smoothness of the loss function). The loss function $l_\xi(\cdot) : \mathcal{D} \times \mathbb{R}^d \mapsto \mathbb{R}$ is smooth with constant $L > 0$, if for any $\xi \in \mathcal{D}$ and all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, $l_\xi(\boldsymbol{\theta})$ is continuously differentiable, and its gradient $\nabla l_\xi(\cdot)$ in terms of $\boldsymbol{\theta}$ is L-Lipschitz, i.e.,

$$\left\|\nabla l_\xi(\boldsymbol{\theta}) - \nabla l_\xi(\boldsymbol{\theta}')\right\| \le L \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|. \qquad (8)$$

3.1 Privacy Analysis of DPSGD

In this subsection, first, we divide the original additive Gaussian noise $\boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{DP}}^2 \boldsymbol{I}_d)$ (c.f. (6)) into two parts: $\boldsymbol{\varrho}_t \sim \mathcal{N}(\mathbf{0}, \beta\sigma_{\text{DP}}^2 \boldsymbol{I}_d)$ and $\boldsymbol{\varsigma}_t \sim \mathcal{N}(\mathbf{0}, (1-\beta)\sigma_{\text{DP}}^2 \boldsymbol{I}_d)$. The first part, together with the clipped SGD update, constitutes the noisy update function, defined as

$$\psi(\boldsymbol{\theta}_t) \triangleq \boldsymbol{\theta}_t - \eta \left(\frac{1}{b} \sum_{\xi \in \mathcal{B}_t} \text{clip}_C(\nabla l_\xi(\boldsymbol{\theta}_t)) + \boldsymbol{\varrho}_t \right), \qquad (9)$$

and the privacy loss of which is characterized by Lemma 3.2 presented shortly; the other part, $-\eta\boldsymbol{\varsigma}_t$, aims for reducing the shift amount of the shifted Rényi divergence leveraging Lemma 2.6.

Next, we provide the following key lemma that provides the upper bound on the shifted Rényi divergence between the distributions of the noisy update function applied on two adjacent datasets with a smooth loss function.

Lemma 3.2. (Noisy smooth-reduction). *Let $\psi(\cdot)$ and $\psi'(\cdot)$ be noisy update functions (c.f. (9)) of DPSGD based on adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$, respectively, and n be the size of $\mathcal{D}$. If the loss function is L-smooth (Assumption 3.1), for any random variables $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, we have*

$$\mathcal{D}_\alpha^{((1+\eta L)z)}(\mathbb{P}_{\psi(\boldsymbol{\mu})} || \mathbb{P}_{\psi'(\boldsymbol{\nu})}) \le \mathcal{D}_\alpha^{(z)}(\mathbb{P}_{\boldsymbol{\mu}} || \mathbb{P}_{\boldsymbol{\nu}}) + \frac{2\alpha C^2}{\beta n b \sigma_{\text{DP}}^2}. \qquad (10)$$

Proof sketch. We summarize the proof steps as follows. First, we transform the shifted Rényi divergence to the standard Rényi divergence utilizing the smoothness of losses and equivalent definitions of ∞-Wasserstein distance. Next, the post-processing (Lemma A.9) and partial convexity inequality (Lemma A.10) allow us to derive the privacy loss associated with SGD sampling. Finally, we apply the strong composition lemma (Lemma A.11) of RDP, and obtain the privacy loss associated with these two Gaussian distributions. The complete proof can be found in Appendix B.1. □

If we further assume the mini-batch size $b \le \frac{n}{5}$, the RDP parameter $\alpha \le \alpha^*(\frac{b}{n}, \frac{b\sqrt{\beta}\sigma_{\text{DP}}}{2C})$, and the per-dimension Gaussian noise scale $\sigma_{\text{DP}} > \frac{8C}{b\sqrt{\beta}}$, a strengthened result can be obtained. Similar findings hold for all of our following theoretical results, and the complete version of Lemma 3.2 can be found in Appendix B.1.

Note that this result generalizes the contraction-reduction lemma (see Lemma A.1) (Feldman et al. 2018) and its variants (Altschuler and Talwar 2022; Altschuler, Bok, and Talwar 2024) in existing literature, which all rely on the convexity of the loss function to ensure that the update function is contractive.[3] It characterizes the privacy dynamics of shifted

[3]A function is said to be contractive if it is 1-Lipschitz.

Rényi divergence for noisy stochastic updates with general smooth loss functions. Based on this building block, we are ready to present the privacy guarantees for DPSGD-GC and DPSGD-DC, respectively.

Theorem 3.3. (Privacy guarantee for DPSGD-GC). *Given the number of total iterations T, dataset size n, batch size b, stepsize η, $\alpha > 1$, gradient-clipping threshold C, and noise scale σ_{DP}, if the loss function is L-smooth (Assumption 3.1), then the DPSGD-GC algorithm satisfies (α, ε)-RDP for*

$$\varepsilon = \mathcal{O}\left(\frac{\alpha C^2}{nb\sigma_{\text{DP}}^2} T\right). \tag{11}$$

Proof sketch. We establish our result utilizing the recursive hypothesis from T to 0 and the flexibility of the shifted Rényi divergence. For the base case at $t = 0$, we have $\mathcal{D}_\alpha^{(z_0)}(\mathbb{P}_{\boldsymbol{\theta}_0} || \mathbb{P}_{\boldsymbol{\theta}_0'}) = 0$, since the initialization satisfy $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0'$. For the recursive step, we apply Lemma 3.2 and subsequently reduce Lemma 2.6 with auxiliary variables to derive the recurrence relationship. Finally, by tracking the privacy loss across all iterations, we derive an upper bound on the RDP loss. The complete proof can be found in Appendix B.2. □

Theorem 3.4. (Privacy guarantee for DPSGD-DC). *Given the number of total iterations T, dataset size n, batch size b, stepsize η, $\alpha > 1$, gradient-clipping threshold C, diameter of the model-parameter domain D, and noise scale σ_{DP}, if the loss function is L-smooth (Assumption 3.1), then the DPSGD-DC algorithm satisfies (α, ε)-RDP for*

$$\varepsilon = \mathcal{O}\left(\frac{\alpha C^2}{nb\sigma_{\text{DP}}^2} \min\left\{T, \frac{(1+\eta L)^2 nbD^2}{\eta^2 C^2}\right\}\right). \tag{12}$$

Proof sketch. This proof follows a similar approach to Theorem 3.3 but terminates the recursion early at iteration τ. By judiciously setting the shift amount $z_\tau = D$ at $t = \tau$, we obtain $\mathcal{D}_\alpha^{(z_\tau)}(\mathbb{P}_{\boldsymbol{\theta}_\tau} || \mathbb{P}_{\boldsymbol{\theta}_\tau'}) = 0$ due to the bounded domain assumption. Finally, by appropriately setting the values of the auxiliary shift variables, we derive the converged privacy loss, as detailed in Appendix B.3. □

Remark 3.5. Comparing Theorem 3.4 to 3.3, we observe that the privacy loss for DGSGD-DC can still converge to a constant for non-convex smooth loss function when the model-parameter domain is bounded by D. However, unlike the convex cases (Altschuler and Talwar 2022; Altschuler, Bok, and Talwar 2024), where the upper bound scales linearly with D, the non-convex setting yields a bound that scales quadratically with D. This suggests that upper bounds on RDP for bounded-domain scenarios are inherently looser without the assumption on convexity, which is consistent with intuition and recent findings (Kong and Ribero 2024; Chien and Li 2024).

3.2 Utility Analysis of DPSGD

In this subsection, we provide upper bounds on utility functions that reflect convergence behavior of DPSGD-GC and DPSGD-DC under (α, ε)-RDP constraint, respectively. All bounds are expressed as expectations over the randomness of SGD sampling and Gaussian noise.

Assumption 3.6. (Bounded SGD variance). The gradient $\nabla l_\xi(\cdot)$ has bounded variance. That is, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, we have

$$\mathbb{E}_{\xi \sim \mathbb{P}_{\mathcal{D}}}\left[\|\nabla l_\xi(\boldsymbol{\theta}) - \nabla l(\boldsymbol{\theta})\|^2\right] \leq \sigma_{\text{SGD}}^2. \tag{13}$$

Note that Assumption 3.1 implies that the population risk function $l(\cdot) = \mathbb{E}_\xi[l_\xi(\cdot)]$ is also smooth with constant $L > 0$. Based on Assumption 3.1 and 3.6, the following lemma provides an upper bound on the minimum expected norm of the gradient for DPSGD-GC with smooth population risk functions.

Lemma 3.7. (Utility bound for DPSGD-GC (Koloskova, Hendrikx, and Stich 2023)). *Assume that the population risk function $l(\cdot)$ has smoothness parameter L and SGD variance bounded from above by σ_{SGD}^2 (Assumption 3.6). The DPSGD-GC running for T iterations with step-size $\eta \leq \frac{1}{9L}$ has the minimum expected norm of the population risk's gradient upper bounded by*

$$\begin{aligned} &\min_{t\in[0,T]} \mathbb{E}\left[\|\nabla l(\boldsymbol{\theta}_t)\|\right] \\ &\leq \mathcal{O}\Big(\frac{1}{\eta CT} + \frac{1}{\sqrt{\eta T}} + \min\big(\sigma_{\text{SGD}}, \frac{\sigma_{\text{SGD}}^2}{C}\big) \\ &+ \sqrt{\eta L}\frac{\sigma_{\text{SGD}}}{\sqrt{b}} + \frac{dL\eta}{C}\sigma_{\text{DP}}^2 + \sqrt{dL\eta}\sigma_{\text{DP}}\Big), \end{aligned} \tag{14}$$

Proposition 3.8. (Privacy-utility trade-off for DPSGD-GC). *Assuming that the conditions in Lemma 3.7 are satisfied, for DPSGD-GC with L-smooth population risk and $\sigma_{\text{DP}}^2 = \mathcal{O}\big(\frac{\alpha C^2 T}{\varepsilon nb}\big)$, we have the following results:*

$$\begin{aligned} &\min_{t\in[0,T]} \mathbb{E}\left[\|\nabla l(\boldsymbol{\theta}_t)\|\right] \\ &\leq \mathcal{O}\Big(\frac{1}{\eta CT} + \frac{1}{\sqrt{\eta T}} + \min\left\{\sigma_{\text{SGD}}, \frac{\sigma_{\text{SGD}}^2}{C}\right\} \\ &+ \sqrt{\eta L}\frac{\sigma_{\text{SGD}}}{\sqrt{b}} + \frac{\alpha dL\eta CT}{\varepsilon nb} + \frac{\sqrt{\alpha dL\eta T}C}{\sqrt{\varepsilon nb}}\Big). \end{aligned} \tag{15}$$

Following this proposition, we can derive the achievable privacy-utility trade-off for DPSGD-GC as $\mathcal{O}(\max\{\frac{dL\log(1/\delta)}{\epsilon^2 n^2}, \sigma_{\text{SGD}}^{4/3}[\frac{dL\log(1/\delta)}{\epsilon^2 n^2}]^{1/3}\})$, and we defer the details to Appendix C.2.

Assumption 3.9. (μ-strongly convex). The population risk function $l(\cdot)$ is strongly convex with constant $\mu > 0$ if and only if the following inequality holds for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, i.e.,

$$\left[\nabla l(\boldsymbol{\theta}) - \nabla l\left(\boldsymbol{\theta}'\right)\right]^T \left(\boldsymbol{\theta} - \boldsymbol{\theta}'\right) \geq \mu \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|^2. \tag{16}$$

For DPSGD-DC, in the special case of smooth and strongly convex population risk, we obtain the following upper bound that describes the minimum expected square root of the optimality gap.

Theorem 3.10. (Utility bound for DPSGD-DC). *Assume that the population risk function $l(\cdot)$ has smoothness parameter L, strongly convex parameter μ (Assumption 3.9), SGD variance bounded from above by σ_{SGD}^2 (Assumption 3.6), and $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \in \operatorname{int}\mathcal{K}$. The DPSGD-DC running for T iterations with step-size $\eta \leq \frac{9}{20L}$ has a minimum* *expected square root of the optimality gap upper bounded by*

$$\min_{t\in[0,T]} \mathbb{E}\left[\sqrt{l(\boldsymbol{\theta}_t) - l(\boldsymbol{\theta}^*)}\right] \le \mathcal{O}\bigg(\frac{\sqrt{L}D^2}{\eta CT} + \frac{D}{\sqrt{\eta T}} + \min\left\{\frac{L^{3/4}}{\mu^{5/4}}\sigma_{\text{SGD}}, \sqrt{\frac{\sigma_{\text{SGD}}^3}{\mu C}}\right\} + \frac{\sqrt{\eta}\sigma_{\text{SGD}}}{\sqrt{b}} + \frac{d\eta\sigma_{\text{DP}}^2\sqrt{L}}{C} + \sqrt{d\eta}\sigma_{\text{DP}}\bigg). \tag{17}$$

Proof sketch. We present an intuitive sketch below, with the complete proof provided in Appendix C.1. The primary challenges stem from the gradient clipping operation, the SGD procedure, and the parameter projection step. To address these, we divide the proof into several cases. For instance, in the case where the clipping threshold $C \le 10\sigma_{\text{SGD}}(\frac{L}{\mu})^{\frac{1}{2}}$ and $\|\nabla l(\boldsymbol{\theta}_t)\| \ge 35\sigma_{\text{SGD}}(\frac{L}{\mu})^{\frac{3}{4}}$, first, we leverage the non-expansiveness of the projection operator to analyze the impact of the parameter projection. Next, we introduce an auxiliary clipping factor $\gamma_\xi = \min(1, \frac{C}{\|\nabla l_\xi(\boldsymbol{\theta}_t)\|})$ and apply the Markov inequality to quantify the effects of clipped SGD. By careful step-size design and manipulations, we obtain a recurrence relation that characterizes the evolution over two successive time steps. Finally, by averaging over t, we derive an upper bound for this case. The proof for other cases follows a similar procedure but incorporates different auxiliary variables tailored to the true gradient. By summing up all cases, we derive an upper bound on the error for DPSGD-DC. □

Remark 3.11. Note that our results adopt a different utility metric other than the more commonly used $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[l(\boldsymbol{\theta}_t) - l(\boldsymbol{\theta}^*)]$, as this metric better facilitates our analysis and gaining insights into the results.

Using the RDP guarantees in Theorem 3.4, we immediately obtain the following results.

Proposition 3.12. (Privacy-utility trade-off for DPSGD-DC). *Assuming that the conditions in Theorem 3.10 are satisfied, for DPSGD-DC with L-smooth and μ-strongly convex population risk, $\sigma_{\text{DP}}^2 = \mathcal{O}\big(\frac{\alpha C^2}{\varepsilon nb}\min\{T, \overline{T}\}\big)$ and $\overline{T} = \frac{(1+\eta L)^2 nbD^2}{\eta^2 C^2}$, we have the following results:*

$$\min_{t\in[0,T]} \mathbb{E}\left[\sqrt{l(\boldsymbol{\theta}_t) - l(\boldsymbol{\theta}^*)}\right] \le \mathcal{O}\bigg(\frac{\sqrt{L}D^2}{\eta CT} + \frac{D}{\sqrt{\eta T}} + \min\left\{\frac{L^{3/4}}{\mu^{5/4}}\sigma_{\text{SGD}}, \sqrt{\frac{\sigma_{\text{SGD}}^3}{\mu C}}\right\} + \frac{\sqrt{\eta}\sigma_{\text{SGD}}}{\sqrt{b}} + \frac{\alpha d\eta C\sqrt{L}}{\varepsilon nb}\min\{T, \overline{T}\} + \sqrt{\frac{\alpha d\eta}{\varepsilon nb}\min\{T, \overline{T}\}}C\bigg). \tag{18}$$

Proposition 3.12 suggests that the utility bound for DPSGD-DC comprises six terms. The first two terms capture optimization-related factors, reflecting the influence of clipping and projection in convergence behavior. The third term accounts for the inherent bias introduced by gradient clipping, while the fourth term reflects the effect due to SGD sampling. Finally, the last two terms quantify the impact of the injected DP noise.

Following Proposition 3.12, we can derive the achievable privacy-utility trade-off for DPSGD-DC as

$$\mathcal{O}\bigg(\max\bigg\{\frac{D^2 dL\log(1/\delta)}{\epsilon^2 n^2}, \frac{\sigma_{\text{SGD}}^{3/2} D^{1/2}}{\mu^{1/2}}[\frac{dL\log(1/\delta)}{\epsilon^2 n^2}]^{1/4}, \frac{\sigma_{\text{SGD}} D\sqrt{d\log(1/\delta)}}{\sqrt{b}\epsilon}\bigg\}\bigg), \tag{19}$$

and we provide the details in Appendix C.3.

Remark 3.13. Unlike previous works on DPSGD-DC (Altschuler and Talwar 2022; Altschuler, Bok, and Talwar 2024; Kong and Ribero 2024; Chien and Li 2024), which mainly focused on RDP analysis, our results explicitly demonstrate how gradient clipping and projection affect the utility of DPSGD-DC. As seen from Theorem 3.4, we establish an eventually constant upper bound on privacy loss dependent on the bounded domain diameter D, which yields tighter privacy guarantees with smaller D. Meanwhile, Proposition 3.12 demonstrates that when the conditions in Theorem 3.10 are satisfied, a smaller D also leads to a lower upper bound on the utility, thus enhancing the privacy-utility trade-off.[4]

4 Experiment Evaluation

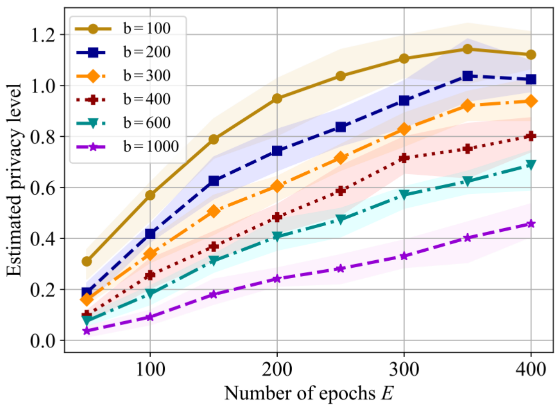
In this section, we present empirical results estimating the privacy level via MIA. Following Kairouz, Oh, and Viswanath (2015), we estimate (ϵ, δ)-DP using the false positive rate (FPR) and false negative rate (FNR) of its attack model on the test data, applying the following formula:

$$\hat{\epsilon} = \max\left\{\log\frac{1-\delta-\text{FPR}}{\text{FNR}}, \log\frac{1-\delta-\text{FNR}}{\text{FPR}}\right\}. \tag{20}$$

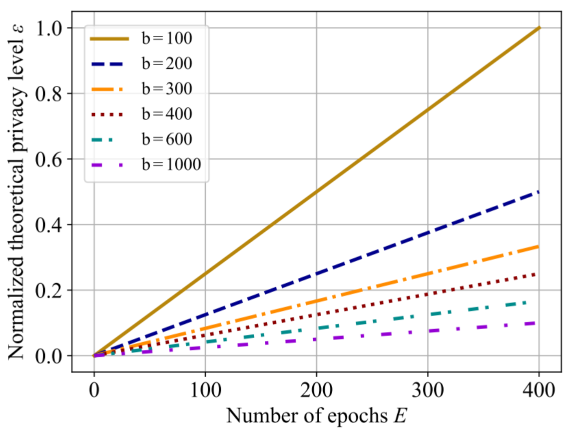
We emphasize that the MIA serves primarily as a tool to provide a lower bound on privacy, capturing the trend with privacy level changes and validating the consistency of the theoretical bounds. By comparing experimental results with the theoretical trends, we aim for demonstrating the reasonableness of the derived privacy bounds, rather than offering an exact measure of privacy leakage. This approach allows us to examine how privacy evolves with varying experimental conditions under the same privacy attack. Additional information on the MIA setting and implementation details can be found in Appendix D.

Effect of the Batch Size. We evaluate DPSGD-GC with various batch sizes $b \in \{100, 200, 300, 400, 600, 1000\}$. Figure 2 illustrates the evolution of the estimated and theoretical privacy level. We also report the evolution of training loss in Figure 4 in Appendix D.3. As expected, the estimated privacy level $\hat{\epsilon}$ increases with the number of epochs, and larger batch sizes provide stronger privacy protection. These observations align with our theoretical results in Theorem 3.3. Additionally, DPSGD-GC converges more slowly

[4]A brief discussion on selecting D is provided in Appendix E.3.

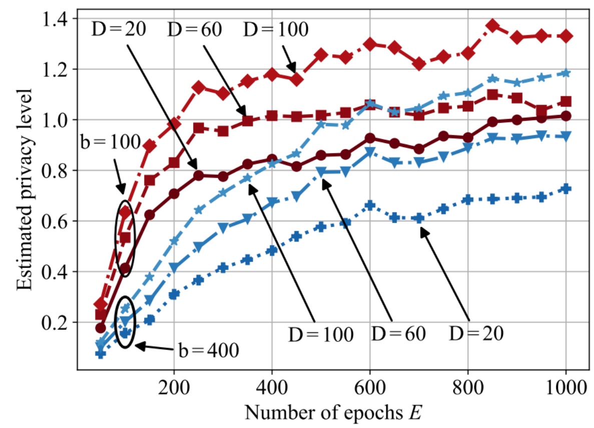


(a)

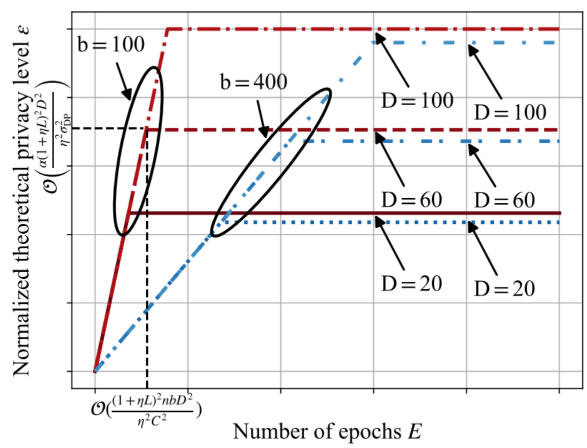


(b)

Figure 2: The evolution of the: (a) estimated and (b) normalized theoretical privacy level during DPSGD-GC with different batch sizes. The shaded error bars correspond to intervals covering 95% of the realized values, obtained from the 10 Monte Carlo trials. Note that the privacy bounds in terms of the number of epochs, E, can be derived by substituting $T = \lceil \frac{n}{b} \rceil E$ into our main results.



(a)



(b)

Figure 3: The evolution of the privacy level during DPSGD-DC with different diameters of the bounded domain: (a) estimated by MIA and (b) normalized theoretical privacy level ε with $\alpha = 1.1$. The red and blue lines correspond to the cases with batch sizes of 100 and 400, respectively.

with larger batch sizes, consistent with Lemma 3.7, further highlighting the trade-off between privacy and utility.

Effect of the Bounded Domain Diameter. We then conduct experiments on DPSGD-DC using various diameters for the bounded domain $D \in \{20, 60, 100\}$ with batch sizes of 100 and 400, respectively. The estimated and theoretical privacy parameter of DPSGD-DC for different bounded domain diameters D is shown in Figure 3. As can be observed, DPSGD-DC provides stronger privacy guarantees with a smaller D, as limiting the domain diameter restricts the range of parameter variations to a narrower interval. Additionally, privacy leakage tends to stabilize as the number of epochs increases. This observation is not surprising, as the bounded domain assumption provides a constant upper privacy bound for DPSGD-DC, as shown in Theorem 3.4.

5 Conclusions

In this paper, we have rigorously analyzed the privacy and utility guarantee of DPSGD, considering both gradient clipping (DPSGD-GC) and double clipping (DPSGD-DC). Our analysis extends the existing privacy bounds of DPSGD to general smooth and non-convex problems without relying on other assumptions. While previous works have focused solely on privacy characterization, we have also derived utility bounds corresponding to our RDP guarantees. This dual characterization admits a more comprehensive understanding of the privacy-utility trade-offs in DPSGD, providing valuable insights for developing more effective differentially private optimization algorithms.

Acknowledgments

The authors would like to thank the anonymous reviewers, SPC, and AC for their insightful comments. The work of H. Xing was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010123, and in part by the Guangzhou Municipal Science and Technology Project under Grants 2024A04J4527 and 2023A03J0663.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. Vienna Austria: ACM.

Altschuler, J.; and Talwar, K. 2022. Privacy of Noisy Stochastic Gradient Descent: More Iterations without More Privacy Loss. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, 3788–3800. Curran Associates, Inc.

Altschuler, J. M.; Bok, J.; and Talwar, K. 2024. On the privacy of noisy stochastic gradient descent for convex optimization. *SIAM Journal on Computing*, 53(4): 969–1001.

Bassily, R.; Smith, A.; and Thakurta, A. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE 55th Annual Symposium on Foundations of Computer Science*, 464–473. Philadelphia, PA: IEEE.

Chen, X.; Wu, S. Z.; and Hong, M. 2020. Understanding gradient clipping in private sgd: A geometric perspective. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, 13773–13782. Curran Associates, Inc.

Chien, E.; and Li, P. 2024. Convergent privacy loss of noisy-sgd without convexity and smoothness. *arXiv preprint arXiv:2410.01068*.

Chourasia, R.; Ye, J.; and Shokri, R. 2021. Differential privacy dynamics of Langevin diffusion and noisy gradient descent. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, 14771–14781. Curran Associates, Inc.

Dong, J.; Roth, A.; and Su, W. J. 2022. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1): 3–37.

Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; and Naor, M. 2006. Our data, ourselves: Privacy via distributed noise generation. In Vaudenay, S., ed., *Advances in Cryptology-EUROCRYPT*, 486–503. St. Petersburg, Russia: Springer Berlin Heidelberg.

Dwork, C.; Rothblum, G. N.; and Vadhan, S. 2010. Boosting and differential privacy. In *IEEE 51st Annual Symposium on Foundations of Computer Science*, 51–60. Las Vegas, NV: IEEE.

Fang, H.; Li, X.; Fan, C.; and Li, P. 2023. Improved convergence of differential private sgd with gradient clipping. In *The Eleventh International Conference on Learning Representations*.

Feldman, V.; Mironov, I.; Talwar, K.; and Thakurta, A. 2018. Privacy amplification by iteration. In *IEEE 59th Annual Symposium on Foundations of Computer Science*, 521–532. Paris, France: IEEE.

Gopi, S.; Lee, Y. T.; and Wutschitz, L. 2021. Numerical Composition of Differential Privacy. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, 11631–11642. Curran Associates, Inc.

Kairouz, P.; Oh, S.; and Viswanath, P. 2015. The composition theorem for differential privacy. In Bach, F.; and Blei, D., eds., *International Conference on Machine Learning*, 1376–1385. Lille, France: PMLR.

Koloskova, A.; Hendrikx, H.; and Stich, S. U. 2023. Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning*, 17343–17363. PMLR.

Kong, W.; and Ribero, M. 2024. Privacy of the last iterate in cyclically-sampled DP-SGD on nonconvex composite losses. *arXiv preprint arXiv:2407.05237*.

Koskela, A.; Jälkö, J.; and Honkela, A. 2020. Computing tight differential privacy guarantees using FFT. In Chiappa, S.; and Calandra, R., eds., *International Conference on Artificial Intelligence and Statistics*, 2560–2569. PMLR.

Mironov, I. 2017. Rényi differential privacy. In *IEEE 30th Computer Security Foundations Symposium*, 263–275. Santa Barbara, CA, USA: IEEE.

Mironov, I.; Talwar, K.; and Zhang, L. 2019. Rényi differential privacy of the sampled Gaussian mechanism. *arXiv preprint arXiv:1908.10530*.

Rényi, A. 1961. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, 547–562. University of California Press.

Ryffel, T.; Bach, F.; and Pointcheval, D. 2022. Differential privacy guarantees for stochastic gradient Langevin dynamics. *arXiv preprint arXiv:2201.11980*.

Song, S.; Steinke, T.; Thakkar, O.; and Thakurta, A. 2021. Evading the curse of dimensionality in unconstrained private glms. In Banerjee, A.; and Fukumizu, K., eds., *International Conference on Artificial Intelligence and Statistics*, 2638–2646. PMLR.

Ye, J.; and Shokri, R. 2022. Differentially private learning needs hidden state (or much faster convergence). In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, 703–715. Curran Associates, Inc.