NON-ASYMPTOTIC ANALYSIS OF STOCHASTIC GRADIENT DESCENT UNDER LOCAL DIFFERENTIAL PRIVACY GUARANTEE

Anonymous authors
Paper under double-blind review

ABSTRACT

In private machine learning algorithms, Differentially Private Stochastic Gradient Descent (DP-SGD) plays an important role. Despite this, there have been few studies that have explored the theoretical analysis that can be derived from DP-SGD, particularly in a more challenging scenario where individual users retain the autonomy to specify their differential privacy budgets. In this work, we conduct a comprehensive non-asymptotic analysis of the convergence of the DP-SGD algorithm as well as its variants. This will allow individual users to assign different privacy guarantees when releasing models trained by DP-SGD. Most importantly, we provide readers with practical guidelines regarding the effect of various hyperparameters, such as step size, parameter dimensions, and privacy budgets, on convergence rates. The problem we consider includes the most commonly used loss functions in standard machine learning algorithms. For strongly convex loss functions, we establish an upper bound on the expected distance between the estimators and the global optimum. In the case of non-strongly convex functions, we analyze the upper bound difference between the loss incurred by the estimators and the optimal loss. Our proposed estimators are validated in the theoretical and practical realms by rigorous mathematical derivation and numerous numerical tests.

1 INTRODUCTION

Machine learning (ML) models have demonstrated their effectiveness in analyzing sensitive user data across a variety of domains, including medical imaging (Tang, 2019; Lundervold & Lundervold, 2019), healthcare (Esteva et al., 2017; Wiens et al., 2019), finance (Caruana et al., 2015), and social media content (Kosinski et al., 2013; Wu et al., 2016). However, ML algorithms frequently require extensive training data, which may include sensitive personal information, then this poses an unacceptable risk to individual privacy (Shokri et al., 2017). As a result, optimizers that protect the privacy of the model while training it are becoming increasingly important. Differential Privacy (DP) (Dwork et al., 2006), as a gold-standard concept, provides rigorous guarantees for privacy protection that focus on limiting the influence of any individual data point when handling sensitive information. Training these ML models under the framework of differential privacy has been widely accepted for protecting algorithms against unintentional leakage of private training data (Shokri & Shmatikov, 2015; Papernot et al., 2017; Carlini et al., 2021; Yu et al., 2022; Mehta et al., 2022; Golatkar et al., 2022).

To ensure differential privacy during the training process, a widely used approach is the noisy stochastic gradient descent, commonly referred to as DP-SGD (Song et al., 2013; Abadi et al., 2016; Papernot & Steinke, 2022; Du et al., 2022; Sander et al., 2023). DP-SGD diverges from conventional stochastic gradient descent by initially clipping individual gradients and then adding noise to the average of clipped gradients during the model parameter updates. Such modifications successfully limit the sensitivity of each update, thereby enabling the introduced noise to guarantee differential privacy. Traditional privacy accounting employs a worst-case methodology, assigning uniform privacy parameters to all data samples. However, from the perspective of ML, different examples can exert varying influences on a learning algorithm, as highlighted by (Koh & Liang, 2017; Feldman & Zhang, 2020).

Recently, Local Differential Privacy (LDP) has emerged as a more rigorous differential privacy technique for preserving personal information privacy (Kasiviswanathan et al., 2011; Duchi et al., 2013). Unlike traditional settings, LDP operates under the assumption that the data collector cannot be trusted. In this framework, each individual applies a differential privacy mechanism to their raw data locally, before transmitting the perturbed version to the untrusted server. Until the server receives the perturbed data from all individuals, it proceeds to calculate the relevant statistics and subsequently publishes the statistical findings. Due to its better privacy protection, LDP rapidly attracted substantial interest and found widespread application in the industrial sector. Prominent technology companies like Google (Erlingsson et al., 2014), Apple (Tang et al., 2017; Apple, 2017), and Microsoft (Ding et al., 2017) have already integrated LDP into their respective product portfolios.

To date, a wealth of studies on LDP have surfaced (Duchi et al., 2013; 2018; Duchi & Rogers, 2019; Gopi et al., 2020; Butucea et al., 2023). (Duchi et al., 2013; 2018; Duchi & Rogers, 2019; Gopi et al., 2020; Butucea et al., 2023), but there is a notable gap in the research focusing on non-asymptotic analysis, especially in the context of convergence guarantees of noisy stochastic approximation schemes. In the LDP framework, the convergence rate is notably regarded as crucial in characterizing the statistical efficiency loss incurred by differentially private optimizers. In this paper, we consider the DP-SGD procedure for the minimization of a convex objective function. The core principle behind this method involves adding noise to every iterate of SGD in a way that causes each iterate to satisfy a targeted differential privacy guarantee. While the underlying idea has gained considerable traction in existing literature, our contribution is prominent in the following aspects:

- Our primary contribution is to provide a comprehensive, non-asymptotic analysis of the DP-SGD algorithm as well as the Polyak-Ruppert averaging algorithm (Polyak & Juditsky, 1992) that includes least-squares and logistic regressions with and without strong convexity assumptions.
- A further key contribution of our work is conducting a comparative analysis between the DP-SGD estimator and the associated Polyak-Ruppert averaging estimator. Specifically, we illustrate that both estimators exhibit identical sensitivity to parameter dimension and differential privacy budget. However, they exhibit divergent convergence rates under varying step-size configurations.
- Lastly, through extensive numerical experiments, we validate that the proposed estimators attain the targeted differential privacy and convergence guarantees, underscoring their efficacy in both theoretical and practical settings.

The structure of this paper is organized as follows: We begin with an overview of both DP and GDP concepts. Subsequently, we introduce our proposed methodology, delving into the DP-SGD algorithms and their non-asymptotic theoretical analysis. We conclude by presenting experimental results that underscore the efficacy of our approach.

2 PRELIMINARIES

Definition 1 ((ε, δ)-DP (Dwork et al., 2006)). *A mechanism $M\colon \mathcal{X}^n \to \mathcal{R}$, taking a dataset consisting of individuals as its input, is (ε, δ)-differentially private if, for every pair of adjacent datasets S, $S' \subset \mathcal{X}^n$ that differ in the record of a single individual and for every measurable event $E \subseteq \mathcal{R}$,*

$$\Pr(M(S) \in E) \leq e^{\varepsilon} \cdot \Pr(M(S') \in E) + \delta,$$

where the probability measure $\Pr$ is induced by the randomness of M only. When $\delta = 0$, then M is called ε-differentially private (ε-DP).

While the concept of (ε, δ)-DP has broad applicability in various domains such as healthcare, finance, and social networks, it comes with several notable limitations. In this paper, we specifically employ Gaussian Differential Privacy (GDP), a specific instantiation of f-DP. To formally introduce this privacy definition, let P and P' denote the distributions of $M(S)$ and $M(S')$, respectively. Given any rejection rule $0 \leq \phi \leq 1$, the type-I and type-II errors are defined as: $\alpha_\phi = E_P(\phi)$ and $\beta_\phi = 1 - E_{P'}(\phi)$. Let $T(P, P')$ denote the trade-off function of P and P', that is $T(P, P') = \inf_\phi\{\beta_\phi : \alpha_\phi \leq \alpha\}$. The formal definition of μ-GDP is then given as follows.

Definition 2 (μ-GDP (Dong et al., 2022)). *A mechanism $M\colon \mathcal{X}^n \to \mathcal{R}$, is μ-GDP, if*

$$T(M(S), M(S')) \geq T(N(0,1), N(\mu, 1))$$

for all neighboring datasets S, $S' \subset \mathcal{X}^n$ and $\mu \geq 0$.

For statisticians, GDP offers an intuitively appealing interpretation: determining whether an individual is part of a given dataset is at least as challenging as distinguishing between two normal distributions—specifically $N(0,1)$ and $N(\mu,1)$ based on one draw, where $\mu \geq 0$. We then list several useful facts about designing μ-GDP algorithms.

Proposition 1 (Dong et al. (2022)). *Let $f : \mathcal{X}^n \to \mathcal{R}^d$ be a deterministic function with finite sensitivity $\Delta(f) = \sup_{S,S' \subset \mathcal{X}^n, \text{adjacent}} \|f(S) - f(S')\|_2 < \infty$. For all $\mu > 0$ and $\omega \in \mathcal{R}^d$ with coordinates i.i.d samples drawn from $N(0,1)$, $f(S) + \Delta(f)\omega/\mu$ is μ-GDP.*

This mechanism boasts computational efficiency and serves as a foundational building block for more complex algorithms. Subsequently, we introduce the parallel composition property of GDP, which enables us to design a noisy SGD algorithm with a tight privacy guarantee.

Proposition 2 (Parallel composition (Smith et al., 2021)). *Let a sequence of K mechanisms $M_k : \mathcal{X}^n \to \mathcal{R}$, each be μ_k-GDP. Let S_k be disjoint subsets of $S \subset \mathcal{X}^n$. The joint mechanism is defined as the sequence of $M_k(S_k \cup S)$ (given also the output of the previous $k-1$ mechanisms) is $\max\{\mu_1, \mu_2, \ldots, \mu_K\}$-GDP.*

3 METHODOLOGY

Let $x_1, \ldots, x_n$ be independent and identically distributed samples from probability distribution Π representing the private information of each user. For statistical estimation and machine learning problems, we consider the general situation where the true d-dimensional model parameter $\theta^* \in \mathcal{R}^d$ is the minimizer of a convex objection function $F(\theta)$: $\mathcal{R}^d \to \mathcal{R}$, that is,

$$\theta^* = \operatorname{argmin} \{F(\theta) = E_{x \in \Pi} f(\theta, x)\},$$

where x is a random variable drawn from a probability distribution Π and $f(\theta, x)$ is the loss function. A widely used optimization method for minimizing $F(\theta)$ is the stochastic gradient descent (Robbins & Monro, 1951; Polyak & Juditsky, 1992). In particular, let θ_0 be any given initial point. Recall that the SGD updates the iterate as follows:

$$\theta_i = \theta_{i-1} - \eta_i \nabla f(\theta_{i-1}, x_i), \quad i \geq 1, \tag{1}$$

where x_i is the ith sample randomly drawn from the distribution Π, $\nabla f(\theta, x)$ denotes the gradient of $f(\theta, x)$ with respect to the first argument, and η_i is the ith step size, which we refer to as the learning rate.

Within the framework of the GDP mechanism, we introduce a differentially private estimator utilizing noisy SGD and provide a non-asymptotic convergence analysis. Intuitively, the heightened risk of privacy leakage stems from the numerous data and gradient queries inherent in the SGD algorithm. To mitigate this, we utilize the GDP to gain privacy by restricting the class of estimators to satisfy a uniform boundedness condition.

Condition 1. *The gradient of the loss function $f(\theta, x)$ is such that $\sup_{\theta \in \mathcal{R}^d, x \in \Pi} \|\nabla f(\theta, x)\|_2 \leq C_0 < \infty$, where C_0 is some positive constant.*

Actually, Condition 1 implies that the objective loss function $f(\theta, x)$ possesses a predefined gradient bound, denoted as C_0. Consequently, this ensures that the sensitivity of $\nabla f(\theta, x)$ does not exceed $2C_0$. Therefore, to achieve a certain level of privacy, our private version of SGD considers the following noisy version of the iterates (1):

$$\theta_i = \theta_{i-1} - \eta_i \nabla f(\theta_{i-1}, x_i) + \frac{2\eta_i C_0}{\mu_i} \omega_i, \quad i \geq 1, \tag{2}$$

where $\{\omega_i\}_{i \geq 1}$ is a sequence of i.i.d. standard d-dimensional Gaussian random vectors and μ_i is the ith individual's certain level of privacy. Taking $C_0 = 0$ in the iterates (2) recovers the standard SGD algorithm in (1).

From (2), each individual has different privacy budgets μ_i, $i = 1, \ldots, n$. This is the notion of the local version of differential privacy (Kasiviswanathan et al., 2011; Yang et al., 2020), which is a valuable tool to protect the privacy of individual data owners without the need for a trusted third

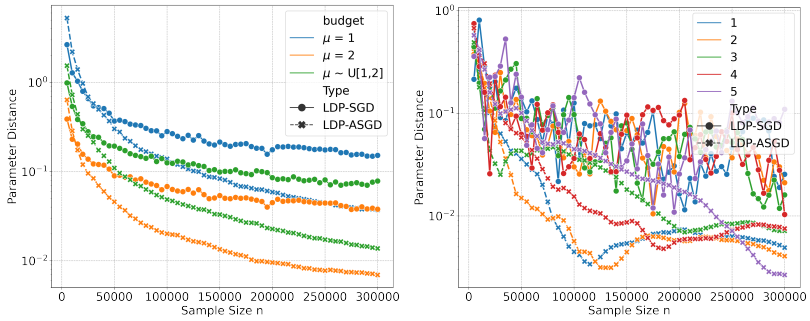

Figure 1: Trajectories of the distance between DP-SGD estimators and the optimal for linear regression. **Left:** Trajectories for two types of estimators under varying privacy budgets with results averaged over 200 replicates. The parameter distance for privacy budgets sampled from $\mathrm{Unif}(1, 2)$ lies between the cases of $\mu = 1$ and $\mu = 2$. **Right:** Trajectories for the first five replications when $\mu = 2$. The LDP-ASGD estimator is more stable than the LDP-SGD estimator.

party. Unlike traditional differential privacy (called centralized differential privacy), local differential privacy (LDP) empowers each participant to apply a differential privacy mechanism to their own raw data locally. The perturbed data is then securely transmitted to an untrusted server. Indeed, stochastic approximation schemes are sequentially interactive cases, which is a typical example of LDP (Duchi et al., 2018). The estimator θ_i defined in (2) is referred to LDP-SGD estimator.

Theorem 1. *The LDP-SGD estimator θ_i obtained from (2) is $\max\{\mu_1, \ldots, \mu_i\}$-GDP, where $i \geq 1$.*

This theorem implies that the proposed noisy SGD estimator θ_n is $\max\{\mu_1, \ldots, \mu_n\}$-GDP. A special case is that the proposed estimator θ_n is μ-GDP when $\mu_1 = \cdots = \mu_n$ without considering different privacy budget for each individual.

It is well-established that classical SGD with a learning rate that is inversely proportional to the number of iterations can perform suboptimally if the proportionality constant is improperly set or when strong convexity assumptions are not met (Nemirovski et al., 2009). To address these challenges, the celebrated Polyak-Ruppert averaging technique significantly enhances the stability and efficiency of stochastic approximation algorithms by averaging over the sequence of iterates (Ruppert, 1988; Polyak & Juditsky, 1992). This technique has been shown to yield information-theoretically optimal asymptotic variance when utilizing learning rates with slower decay rates, in conjunction with uniform averaging.

In this section, we also consider the Polyak-Ruppert averaging private estimator $\bar{\theta}_n = \sum_{i=1}^{n} \theta_i / n$ after n iterations, which is referred to LDP-ASGD estimator. Note that this estimator can also be recursively updated: $\bar{\theta}_n = (n-1)\bar{\theta}_{n-1}/n + \theta_n/n$. The parallel composition proposition 2 along with Theorem 1 easily imply the following results:

Theorem 2. *The LDP-ASGD estimator $\bar{\theta}_n$ is $\max\{\mu_1, \ldots, \mu_n\}$-GDP.*

Visualization of motivation of this paper is displayed in Figure 1. Referring to the left panel of Figure 1, we permit each user to assign individual privacy budgets that align more closely with real-world scenarios. Meanwhile, as illustrated in the right panel of the same figure, the LDP-ASGD estimator exhibits greater stability compared to the LDP-SGD estimator.

4 Theoretical Properties Analysis

4.1 Strongly Convex Objectives

In this subsection, we focus on studying a direct non-asymptotic analysis of the LDP-SGD and LDP-ASGD algorithms. To better illustrate the theoretical properties of the proposed estimators, we impose the following regularity conditions on the objective functions.

Condition 2. *Assume that the objective function $F(\theta)$ is differentiable, M-smooth, and m-strongle convex, meaning that*

(i). $F(\theta_1) - F(\theta_2) \leq \langle \nabla F(\theta_2), \theta_1 - \theta_2 \rangle + \frac{M}{2}\|\theta_1 - \theta_2\|^2, \quad \forall \theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^d,$

(ii). $F(\theta_1) - F(\theta_2) \geq \langle \nabla F(\theta_2), \theta_1 - \theta_2 \rangle + \frac{m}{2}\|\theta_1 - \theta_2\|^2, \quad \forall \theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^d.$

Strong convexity and smoothness are both standard conditions for the convergence analysis of stochastic gradient optimization methods; similar conditions can be found in Gower et al. (2019); Vaswani et al. (2022). In order to establish the results, we also require the Hessian to be Lipschitz continuous, which is commonly assumed to establish the convergence of SGD's method under strong convexity; see Moulines & Bach (2011); Jin et al. (2021); Godichon-Baggioni et al. (2023).

Condition 3. *For each $n \geq 1$, the Hessian $\nabla^2 f(\theta, x_n)$ is C_1-Lipschitz continuous, i.e.,*

$$\|\nabla^2 f(\theta_1, x_n) - \nabla^2 f(\theta^*, x_n)\| \leq C_1 \|\theta_1 - \theta^*\|, \quad \forall \theta_1 \in \Theta \subseteq \mathbb{R}^d.$$

Before presenting the non-asymptotic bounds, we first introduce the following family of functions: $\psi_\beta : \mathbb{R}_+ \backslash \{0\} \to \mathbb{R}$ given by:

$$\psi_\beta(t) = \frac{t^\beta - 1}{\beta}, \quad \text{if } \beta \neq 0$$

Notice that the function $\beta \mapsto \psi_\beta(t)$ is continuous for all $t > 0$. In addition, for $\beta > 0, \psi_\beta(t) < t^\beta/\beta$, while for $\beta < 0$, we have $\psi_\beta(t) < -1/\beta$ (both with asymptotic equality when t is large).

Theorem 3. *Assume Conditions 1, 2 hold, if $\eta_n = \eta n^{-\alpha}$, we then have, for $\alpha \in [0, 1)$*

$$E(\|\theta_n - \theta^*\|^2) \leq C_0^2 \left\{ \frac{\eta}{mn^\alpha} + \eta^2 \psi_{1-2\alpha}(n) \exp(-m\eta n^{1-\alpha}/2) \right\} (1 + 64d/\min_k\{\mu_k^2\}) + \exp\{-2m\eta\psi_{1-\alpha}(n)\} E(\|\theta_0 - \theta^*\|^2).$$

Remark 1. *From Theorem 3, we know that the non-asymptotic upper bound of the LDP-SGD estimator θ_n is influenced by several factors: the initial condition $E(\|\theta_0 - \theta^*\|^2)$, the initial step size η, the decay rate α, the dimensionality d of the parameters, and the smallest private budget $\min_k\{\mu_k^2\}$ across all samples. Specifically,*

- ***Initial condition:*** *The initial condition diminishes sub-exponentially for $\alpha \in [0, 1)$.*
- ***Convergence rate:*** *The leading asymptotic term is $C_0^2\eta(1 + 64d/\min_k\{\mu_k^2\})/mn^\alpha$. As such, the predominant convergence rate is $O(n^{-\alpha})$ for $\alpha \in [0, 1)$. For a scenario where $\alpha = 1$, convergence of the LDP-SGD estimator θ_n is not assured.*
- ***Parameter dimension:*** *The bound exhibits a linear relationship with the dimension d, a consequence of Gaussian noise being introduced to every parameter coordinate.*
- ***Private budget:*** *As μ_k increases, the overall bound of $E(\|\theta_n - \theta^*\|^2)$ reduces in a quadratic speed.*

The following theorem presents the non-asymptotic upper bounds for the LDP-ASGD estimator.

Theorem 4. *Assume Conditions 1, 2, 3 hold, if $\eta_n = \eta n^{-\alpha}$, we then have, for $\alpha \in [0, 1)$*

$$\begin{aligned} E(\|\bar{\theta}_n - \theta^*\|^2) \leq & \frac{1}{n} \operatorname{tr}\left[E\{\nabla^2 f(\theta^*, x_k)\}^{-1} S E\{\nabla^2 f(\theta^*, x_k)\}^{-1}\right] + \frac{4\alpha^2\tilde{\sigma}_\mu^2}{m^2n^2\eta}\psi_{\alpha-1}(n) + \frac{\tilde{\sigma}_\mu^2}{m^2n^{2-\alpha}\eta} \\ & + \left(\frac{1}{mn^2\eta_1^2} + \frac{1}{mn^{2(1-\alpha)}\eta^2} + \frac{4M^2}{mn^2}\right) E(\|\theta_0 - \theta^*\|^2) + \frac{\tilde{\sigma}_\mu^2}{mn^{2(1-\alpha)}}\psi_{1-2\alpha}(n) \\ & + \frac{C_1C_{3,0}\eta^2}{n^2m^2}\psi_{1-2\alpha}(n) + \frac{2C_1C_{4,0}\eta^3}{n^2m^2}\psi_{1-3\alpha}(n) + \frac{4M^2\eta\tilde{\sigma}_\mu^2}{m^2n^2}\psi_{1-\alpha}(n) + \frac{\tilde{\sigma}_\mu^2}{mn} \\ & + \frac{C_1}{4mn^2}E(\|\theta_0 - \theta^*\|^4) + \left(\frac{4\alpha^2}{mn^2\eta^2} + \frac{4M^2}{mn^2}\right)\{E(\|\theta_0 - \theta^*\|^2) + \eta^2\tilde{\sigma}_\mu^2\}B \\ & + \frac{4M^2\eta\tilde{\sigma}_\mu^2}{m^2n^2}\left\{E(\|\theta_0 - \theta^*\|^4) + C_{3,0}\eta^3\psi_{1-3\alpha}(n) + C_{4,0}\eta^4\psi_{1-4\alpha}(n)\right\}B, \end{aligned}$$

where $\tilde{\sigma}_\mu^2 = C_0^2(1+64d/\min_k\{\mu_k^2\})$, $B = \sum_{k=1}^n \exp\left\{-m\eta k^{1-\alpha}/4 + \psi_{1-2\alpha}(k) + 4\eta^3\psi_{1-3\alpha}(k)\right\}$, and $S = E\{\nabla f(\theta^*, x)\nabla f(\theta^*, x)^\top\}$.

Remark 2. *From Theorem 4, notice that the non-asymptotic upper bound of the LDP-ASGD estimator $\bar{\theta}_n$ is influenced by the same factors in a similar way to as Theorem 3.*

- ***Summation*** B***:*** *For all $\alpha \in [0, 1)$, B is finite, while for $\alpha = 1$, $B = O(n)$.*
- ***Initial Condition:*** *For all $\alpha \in [0, 1)$, the initial condition diminishes at rate $O(n^{-2})$.*
- ***Convergence rate:*** *For $\alpha \in [0, 1/2]$, the predominant convergence rate is $O(n^{-1})$, while for $\alpha \in (1/2, 1)$, the predominant convergence rate becomes $O(n^{-2(1-\alpha)})$.*
- ***Other factors:*** *The constant $\tilde{\sigma}_\mu^2$ suggests that the bound of $E(\|\bar{\theta}_n - \theta^*\|^2)$ is affected by the factors d and μ in the same manner as $E(\|\theta_n - \theta^*\|^2)$.*

4.2 Non-strongly convex obejectives

In this subsection, we do not assume that the function $F(\theta)$ is strongly convex, but impose the following condition.

Condition 4. *Assume the objective function $F(\theta)$ attains its global minimum at a certain $\theta^* \in \Theta$.*

Notice that θ^* is not unique when the objective function $F(\theta)$ is not strongly convex, we only get a bound on loss function values as follows.

Theorem 5. *Assume Conditions 1, 2(i), 3, 4 hold, then if $\eta_n = \eta n^{-\alpha}$, for $\alpha \in (1/3, 1)$,*

$$E\{F(\theta_n) - F(\theta^*)\} \leq \begin{cases} \dfrac{2(\delta_0 + \tilde{\sigma}_\mu^2 \eta^2 \psi_{1-2\alpha}(n))^{1/2}(1 + 4M^{1/2}\tilde{\sigma}_\mu \eta^{3/2})/\eta}{\psi_{1-\alpha}(n)}, & \alpha \in (\frac{2}{3}, 1), \\ \dfrac{2(\delta_0 + \tilde{\sigma}_\mu^2 \eta^2 \psi_{1-2\alpha}(n))^{1/2}(1 + 4M^{1/2}\tilde{\sigma}_\mu \eta^{3/2})/\eta}{\psi_{\alpha/2}(n)}, & \alpha \in (\frac{1}{2}, \frac{2}{3}], \\ \dfrac{2(\delta_0 + \tilde{\sigma}_\mu^2 \eta^2)(1 + 4M^{1/2}\tilde{\sigma}_\mu \eta^{3/2})/\eta}{(1-2\alpha)^{1/2}\psi_{3\alpha/2-1/2}(n)}, & \alpha \in [\frac{1}{3}, \frac{1}{2}], \end{cases}$$

where $\delta_0 = E(\|\theta_0 - \theta^*\|^2)$ and $\tilde{\sigma}_\mu^2 = C_0^2(1 + 64d/\min_k\{\mu_k^2\})$.

Remark 3. *From Theorem 5, we note that the upper bound of $E\{F(\theta_n) - F(\theta^*)\}$ shares the same constant $\tilde{\sigma}_\mu^2$ as in strongly-convex scenario, suggesting a similar relationship with d and μ as discussed before. More specifically,*

- ***Initial Condition:*** *For all $\alpha \in (1/3, 1)$, the initial condition diminishes at the same rate as $E\{F(\theta_n) - F(\theta^*)\}$.*
- ***Convergence rate:*** *For $\alpha \in (1/3, 1/2]$, the convergence rate is $O(n^{-(3\alpha-1)/2})$, for $\alpha \in (1/2, 2/3]$, the convergence rate is $O(n^{-\alpha/2})$, while for $\alpha \in (2/3, 1)$ the convergence rate becomes $O(n^{-(1-\alpha)})$.*

The following theorem is shown in a similar way to Theorem 4.

Theorem 6. *Assume Conditions 1, 2(i), 3, 4 hold, then if $\eta_n = \eta n^{-\alpha}$, for $\alpha \in [0, 1)$, we have*

$$E\{F(\bar{\theta}_n) - F(\theta^*)\} \leq \frac{1}{2\eta n^{1-\alpha}}\{E(\|\theta_0 - \theta^*\|^2) + \tilde{\sigma}_\mu^2 \eta^2 \psi_{1-2\alpha}(n)\} + \frac{\tilde{\sigma}_\mu^2 \eta \psi_{1-\alpha}(n)}{2n},$$

where $\tilde{\sigma}_\mu^2 = C_0^2(1 + 64d/\min_k\{\mu_k^2\})$.

Remark 4. *The convergence rate of $E\{F(\bar{\theta}_n) - F(\theta^*)\}$ follows $O(n^{-\alpha})$ when $\alpha \in [0, 1/2)$. However, in the range $1/2 < \alpha < 1$, it adheres to $O(n^{\alpha-1})$. Consequently, the optimal asymptotic rate stands at $O(n^{-1/2})$. It's noteworthy that through averaging, the algorithm extends the α range ensuring convergence, shifting it from $(1/3, 1)$ to $(0, 1)$, and concurrently accelerates the rate of convergence. Nonetheless, in real-world scenarios when $\alpha > 1/2$, the constant term $\{E(\|\theta_0 - \theta^*\|^2) + \tilde{\sigma}_\mu^2 \eta^2 \psi_{1-2\alpha}(n)\}$ could vastly exceed its square root, leading to a bigger loss in the averaged estimators.*

α	Strongly		Non-strongly		
	$(0, 1/2)$	$(1/2, 1)$	$(1/3, 1/2)$	$(1/2, 2/3)$	$(2/3, 1)$
LDP-SGD	$n^{-\alpha}$	$n^{-\alpha}$	$n^{-(3\alpha-1)/2}$	$n^{-\alpha/2}$	$n^{-(1-\alpha)}$
LDP-ASDG	n^{-1}	$n^{-2(1-\alpha)}$	$n^{-\alpha}$	$n^{-(1-\alpha)}$	$n^{-(1-\alpha)}$

Table 1: Summary of convergence rate in different situations.

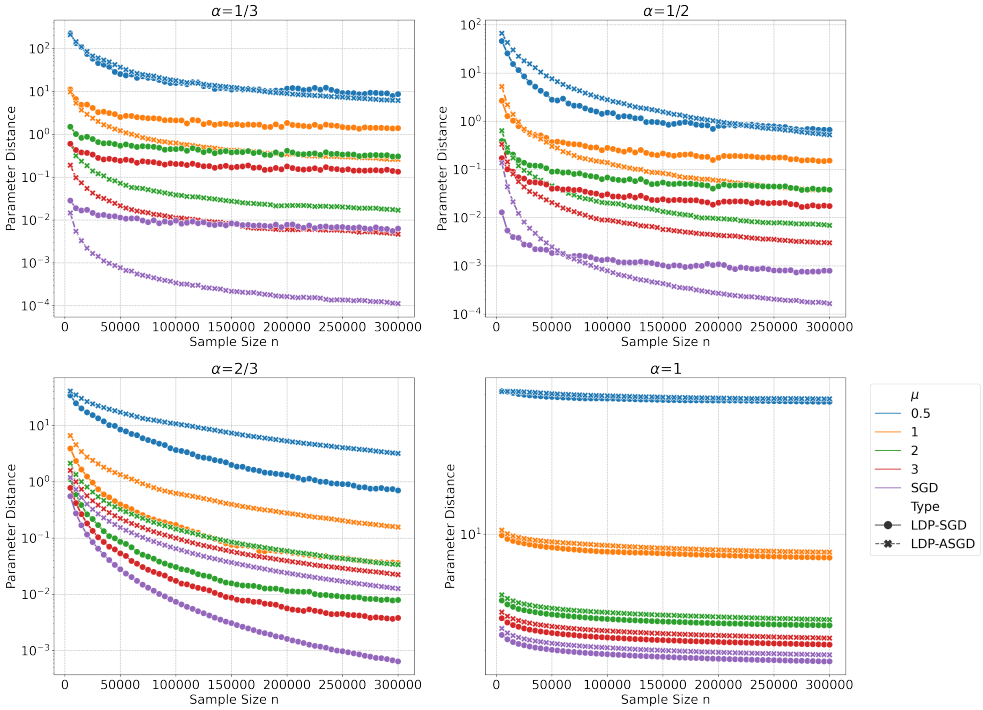

Figure 2: Trajectories of the distance between DP-SGD estimators and the optimal for linear regression with $d = 5$.

5 Experiments

In this section, we illustrate our theoretical results through two numerical simulations. In particular, we show the behavior of expected distance for LDP-SGD and LDP-ASGD estimators in both strongly and non-strongly convex cases.

5.1 Linear regression

We consider a linear regression model given by $y_i = x_i^\top \beta + \epsilon_i$, where the disturbances $\epsilon_1, \ldots, \epsilon_n$ are independently and identically drawn from $N\left(0, \sigma^2\right)$. Here, the covariates are denoted as $x_i = (1, z_i)^\top$, with z_i following an i.i.d. sampled from $N\left(0, \sigma_z^2 \mathbb{I}_d\right)$. The loss function we have chosen is expressed as:

$$\mathcal{L}_n(\beta, \sigma) = \frac{1}{n} \sum_{i=1}^{n} \left(\sigma \rho_c \left(\frac{y_i - x_i^\top \beta}{\sigma} \right) + \frac{1}{2} \kappa_c \sigma \right) w\left(x_i\right),$$

where ρ_c denotes the Huber loss function with a tuning parameter c, κ_c is a constant chosen to ensure the consistency of $\hat{\sigma}$, and $w(x_i) = \min(1, 2/\|x_i\|_2^2)$ serves to de-weight outlying covariates. By this construction, the global sensitivities of $\nabla_\theta \mathcal{L}_n(\theta)$ is $\sqrt{8c^2 + c^4/4}$.

We generate data utilizing a sample size of $n = 300000$ based on the linear model, with the true parameters assigned as $\beta = \mathbf{1}_{d+1}$, $\sigma = 2$, $\sigma_z = 1$. Additionally, we adopt $c = 1.345$ for the loss function. Our objective is to evaluate the performance of LDP-SGD and LDP-ASGD estimators;

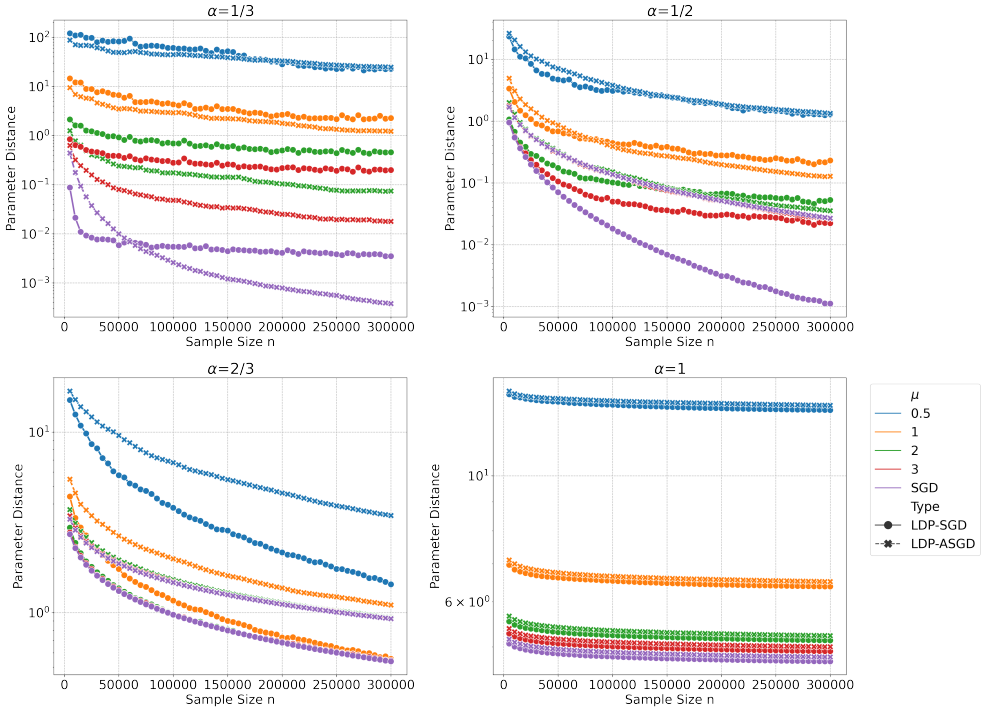

Figure 3: Trajectories of the distance between DP-SGD estimators and the optimal for logistic regression with $d = 5$.

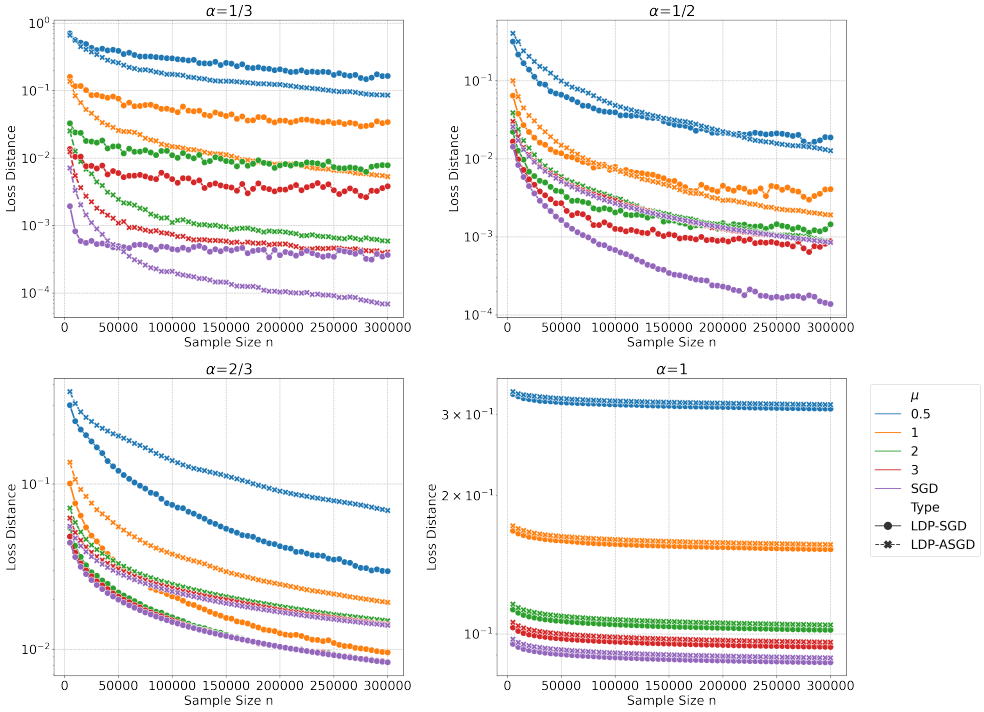

Figure 4: Trajectories of the distance between the loss incurred by the estimators and the optimal loss for logistic regression with $d = 5$.

hence, we conduct experiments under various settings: parameter dimensions $d = 5, 10, 20$; initial step size $\eta = 0.2$ with decay rates $\alpha = 1/3, 1/2, 2/3, 1$; privacy budgets $\mu = 0.5, 1, 2, 3$; and a non-private SGD approach (without the addition of noise). For each configuration, we perform 200 replications and plot the trajectories of the distance between the estimators and the optimum, as well as the loss distance, throughout the training process. The results showcasing parameter distance with $d = 5$ are illustrated in Figure 2, and detailed outcomes for other settings are provided in the Appendix.

From Figure 2, it is evident that the empirical findings cohere seamlessly with our theoretical predictions. For $\alpha \in (0, 1/2)$, the LDP-ASGD estimator exhibits faster convergence compared to the LDP-SGD estimator, resulting in a diminished distance across all privacy configurations. However, when $\alpha \in (2/3, 1)$, the convergence of the LDP-ASGD estimator is slower than that of the LDP-SGD estimator, leading to an increased distance. Notably, when $\alpha = 1$, neither estimator achieves convergence. Furthermore, a reduction in the privacy budget μ is associated with an increment in distance, attributable to the incorporation of more noise at each iteration.

5.2 Logistic regression

In the context of the logistic regression simulation, data were generated based on the model $y_i \sim \text{Bernoulli}(\{1 + \exp(-x_i^\top \beta)\}^{-1})$. This procedure employs the same value of β and adopts a similar approach for generating the covariates x_i as was applied in the linear regression simulation. When incorporating Mallows weights, we utilized a weighted variant of the standard cross-entropy loss:

$$\mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i \log\left(\frac{1}{1+\exp\left(-x_i^\top\beta\right)}\right) - (1-y_i)\log\left(\frac{\exp\left(-x_i^\top\beta\right)}{1+\exp\left(-x_i^\top\beta\right)}\right)\right) w\left(x_i\right),$$

where the weight function is denoted as $w(x_i) = \min(1, 2/\|x_i\|_2^2)$.

We generate data employing a sample size of $n = 300000$ drawn from the logistic model, with the true parameters set as $\beta = \mathbf{1}_{d+1}$ and $\sigma_z = 1$. The parameter configurations adopted align with those utilized in the linear regression scenario. It is important to note that, in the context of logistic regression, the assumption of strong convexity is no longer applicable. Consequently, our primary focus shifts toward the disparity between the loss incurred by the estimators and the optimal loss. To further assess the robustness of Theorems 3 and 4 in scenarios void of strong convexity, we also present results pertaining to parameter distance. The results related to parameter distance for $d = 5$ are depicted in Figure 3, while those concerning loss distance are showcased in Figure 4. Detailed results for alternative configurations are provided in the Appendix.

As shown in Figure 3, for $\alpha \in (0, 1)$, both LDP-SGD and LDP-ASGD estimators exhibit a convergence tendency towards the optimal, evident from the diminishing distance, even in the absence of strong convexity in the loss. The nearest achievable distance is observed to be approximately 10^{-2}, indicating potential deviations from the exact global optimal parameter. Turning our attention to the loss distance, the insights derived from Figure 4 are consistent with the theoretical predictions outlined in Theorems 5 and 6. For any $\alpha \in (0, 1)$, the LDP-ASGD estimator exhibits faster convergence compared to the LDP-SGD estimator. However, when $\alpha > 1/2$, the constant term bounding the LDP-ASGD estimator may become more significant, potentially resulting in a larger loss distance for the averaged estimators.

6 Conclusion and Future Works

In this paper, we have provided the non-asymptotic analysis of the convergence of the DP-SGD algorithm as well as its averaged version under the LDP framework, which allows different individual users to have different privacy budgets. Importantly, we have analyzed the theoretical properties of the proposed estimators, with and without strong convexity assumptions. Our theory shows that the convergence rates of the considered estimators are affected by various hyperparameters, including step size, parameter dimensions, and privacy budgets. This furnishes readers with practical guidelines on how to select hyperparameters. Comprehensive simulation studies yield positive affirmation of the asymptotic theory.

While the contributions mentioned above are significant, this work leaves several interesting avenues for future work. First, we mainly focus on differentiable loss functions, as these are commonly used

loss functions. Further investigation is warranted to consider non-differentiable objectives such as pinball or hinge losses. Second, we have focused on results with fixed parameter dimensions and we expect to study the modification DP-SGD algorithm under high parameter dimensions; see Agarwal et al. (2012); Chen et al. (2020). We will leave it for future research.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. *Advances in Neural Information Processing Systems*, 25, 2012.

Differential Privacy Team Apple. Learning with privacy at scale. *Apple Machine Learning, Journal*, 1:1–25, 2017.

Cristina Butucea, Angelika Rohde, and Lukas Steinberger. Interactive versus noninteractive locally differentially private estimation: Two elbows for the quadratic functional. *The Annals of Statistics*, 51(2):464–486, 2023.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Jian Du, Song Li, Xiangyi Chen, Siheng Chen, and Mingyi Hong. Dynamic differential-privacy preserving sgd. *Proceedings of the 39 th International Conference on Machine Learning*, 2022.

John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pp. 1161–1191. PMLR, 2019.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Antoine Godichon-Baggioni, Nicklas Werge, and Olivier Wintenberger. Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *ESAIM: Probability and Statistics*, 27: 482–514, 2023.

Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8376–8386, 2022.

Sivakanth Gopi, Gautam Kamath, Janardhan Kulkarni, Aleksandar Nikolov, Zhiwei Steven Wu, and Huanyu Zhang. Locally private hypothesis selection. In *Conference on Learning Theory*, pp. 1785–1816. PMLR, 2020.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pp. 5200–5209. PMLR, 2019.

Yanhao Jin, Tesi Xiao, and Krishnakumar Balasubramanian. Statistical inference for polyak-ruppert averaged zeroth-order stochastic gradient algorithm. *arXiv preprint arXiv:2102.05198*, 2021.

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.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15): 5802–5805, 2013.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

Harsh Mehta, Abhradeep Thakurta, Alexey Kurakin, and Ashok Cutkosky. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:2205.02973*, 2022.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. *In International Conference on Learning Representations*, 2022.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *In International Conference on Learning Representations*, 2017.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Tom Sander, Pierre Stock, and Alexandre Sablayrolles. Tan without a burn: Scaling laws of dp-sgd. In *International Conference on Machine Learning*, pp. 29937–29949. PMLR, 2023.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Josh Smith, Hassan Jameel Asghar, Gianpaolo Gioiosa, Sirine Mrabet, Serge Gaspers, and Paul Tyler. Making the most of parallel composition in differential privacy. *arXiv preprint arXiv:2109.09078*, 2021.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013.

Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

Xiaoli Tang. The role of artificial intelligence in medical imaging research. *BJR| Open*, 2(1): 20190031, 2019.

Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad. Towards noise-adaptive, problem-adaptive (accelerated) stochastic gradient descent. In *International Conference on Machine Learning*, pp. 22015–22059. PMLR, 2022.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

Shaomei Wu, Hermes Pique, and Jeffrey Wieland. Using artificial intelligence to help blind people 'see'facebook, 2016.

Mengmeng Yang, Lingjuan Lyu, Jun Zhao, Tianqing Zhu, and Kwok-Yan Lam. Local differential privacy and its applications: A comprehensive survey. *arXiv preprint arXiv:2008.03686*, 2020.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *In Proceedings of the 10th International Conference on Learning Representations*, 2022.