Almost Sure Convergence Analysis of Differentially Private Stochastic Gradient Methods

Amartya Mukherjee and Jun Liu

***Abstract*— Differentially private stochastic gradient descent (DP-SGD) has become the standard algorithm for training machine learning models with rigorous privacy guarantees. Despite its widespread use, the theoretical understanding of its long-run behavior remains limited: existing analyses typically establish convergence in expectation or with high probability, but do not address the almost sure convergence of single trajectories. In this work, we prove that DP-SGD converges almost surely under standard smoothness assumptions, both in nonconvex and strongly convex settings, provided the step sizes satisfy some standard decaying conditions. Our analysis extends to momentum variants such as the stochastic heavy ball (DP-SHB) and Nesterov's accelerated gradient (DP-NAG), where we show that careful energy constructions yield similar guarantees. These results provide stronger theoretical foundations for differentially private optimization and suggest that, despite privacy-induced distortions, the algorithm remains pathwise stable in both convex and nonconvex regimes.**

I. Introduction

In the training of machine learning models, maintaining the privacy of training data is of paramount importance. Particularly in domains such as health and finance, it is important that an adversary cannot reconstruct training data from the trained models. Unfortunately, generative models risk overfitting to their training data, thus generating data indistinguishable from the training set and compromising the privacy of users. For example, the work of [3] reviews risks of leaking training data in image and text generation models.

To protect the privacy of training data, a mathematically rigorous framework for privacy titled differential privacy (DP [5]) has gained interest in recent years. Differentially private stochastic gradient descent (DP-SGD [1]) is a modification of stochastic gradient descent (SGD) that offers privacy guarantees by introducing gradient clipping and noise injection. While DP guarantees strong privacy protection, it often comes at the cost of slower convergence and degraded model utility due to the bias induced by clipping and the noise injection. This paper will analyze the convergence rates of various DP-SGD methods under different noise injection schemes and dataset conditions. We explore how gradient clipping and noise scaling affect model performance.

Since the advent of DP-SGD, its convergence analysis has been of interest to both the machine learning and control communities. Recent literature has focused on the optimization and generalization trade-offs introduced by differential privacy, but not on its almost-sure stability. For example, the work of [6] provides convergence analysis on a modified DP-SGD that replaces gradient clipping with an affine function of the gradient of the objective function. The authors of [15] extend the analysis to momentum-based variants and show that the additive Gaussian noise used for privacy can dominate the second-moment estimates in adaptive methods like Adam, effectively neutralizing their curvature adaptation and creating severe ill-conditioning under heavy-tailed data distributions. They demonstrate that bias-corrected DP-Adam (DP-AdamBC) mitigates this issue by subtracting the variance of the DP noise, improving convergence on imbalanced datasets. The comprehensive work of [9] derives convergence rates of SGD with clipping in deterministic and stochastic settings. Lastly, DP has also been studied in distributed optimization settings [7], symbolic systems [2], and multi-agent systems [8].

Overall, most existing analyses provide convergence guarantees only in expectation or with high probability, leaving open the question of whether individual trajectories stabilize. This gap is critical, since practical deployments of DP-SGD often train for many epochs under noisy, biased gradient updates introduced by clipping and Gaussian perturbation. Our paper builds upon these works by providing almost sure convergence guarantees for DP-SGD in convex and non-convex settings. We approach this by proving that a weighted average of the norm of the gradient of the objective function converges almost surely, and therefore that the best iterate converges. We extend our analysis further to variants of SGD that include momentum, where we also provide almost sure convergence guarantees of the last iterates.

II. Preliminaries and Assumptions

We provide a formal definitions of DP and SGD, and introduce some assumptions commonly used in the convergence analysis of SGD [10], [4].

Definition 1 (Differential Privacy (DP) [5]): A randomized mechanism $\mathscr{M} : \mathscr{D} \to \mathscr{R}$ with domain $\mathscr{D}$ and range $\mathscr{R}$ satisfies (ε, δ) differential privacy, where $\varepsilon > 0, \delta > 0$, if for any datasets $d, d' \subset \mathscr{D}$ differing by at most one entry, and for any subset of outputs $S \subset \mathscr{R}$, it holds that

$$P(\mathscr{M}(d) \in S) \leq e^{\varepsilon} P(\mathscr{M}(d') \in S) + \delta. \tag{1}$$

Problem statement: We are interested in solving the fol-

Amartya Mukherjee and Jun Liu are with the Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (email: (a29mukhe,j.liu)@uwaterloo.ca).

lowing unconstrained minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}), \tag{2}$$

where $f:\mathbb{R}^d\to\mathbb{R}$, using stochastic gradient methods that satisfy (ε,δ)-DP. Let f^* be the true minimum. In convex settings, we want to prove that $f(\mathbf{x}_t)-f^*\to 0$ as $t\to\infty$. In non-convex settings, we want to prove that $\nabla f(\mathbf{x}_t)\to 0$ as $t\to\infty$.

Definition 2 (Stochastic Gradient Descent (SGD)): The iteration of SGD is given by

$$\mathbf{x}_{t+1}=\mathbf{x}_t-\alpha_t\mathbf{g}_t, \tag{3}$$

where $\mathbf{g}_t=\nabla f(\mathbf{x};\xi_t)$ is the stochastic gradient at $\mathbf{x}_t$ with a random process ξ_t and α_t is a step size. Throughout this paper, we denote $\nabla f(\mathbf{x}_t):=\mathbb{E}[\mathbf{g}_t]$ as the expectation of the stochastic gradient over all ξ_t.

Definition 3 (Differentially Private SGD (DP-SGD) [1]): DP-SGD is a modification of SGD, where gradients are clipped and noise is added to the clipped gradients.

$$\mathbf{x}_{t+1}=\mathbf{x}_t-\alpha_t\mathbf{g}_t^{DP}, \tag{4}$$

where the differentially private stochastic gradient $\mathbf{g}_t^{DP}$ is given by

$$\mathbf{g}_t^{DP}=\text{clip}_q(\nabla f(\mathbf{x}_t;\xi_t))+q\zeta_t, \tag{5}$$

where the clip_q function is defined for $q>0$ by

$$\text{clip}_q(\nabla f(\mathbf{x}_t;\xi_t))=\min\left(1,\frac{q}{\|\nabla f(\mathbf{x}_t;\xi_t)\|}\right)\nabla f(\mathbf{x}_t;\xi_t), \tag{6}$$

where $\zeta_t\sim\mathcal{N}(0,\sigma_{DP}^2 I)$, and $\|\cdot\|$ denotes the 2-norm.

Remark 1: DP-SGD can be adapted to other stochastic gradient methods such as stochastic heavy ball and stochastic Nesterov accelerated gradient. A DP-SGD update satisfies (ε,δ)-DP if σ_{DP}^2 and q satisfy $\sigma_{DP}^2>2\log(1.25/\delta)q^2/\varepsilon^2$.

Based on the formulation of DP-SGD, we also introduce the notion of clipping probability.

Definition 4 (Clipping Probability): Define the process η_t as the clipping probability

$$\eta_t=P(\|\nabla f(\mathbf{x}_t,\xi_t)\|>q|\mathbf{x}_t). \tag{7}$$

We make the following assumptions that are commonly used in the SGD literature [12].

Assumption 1 (L-smoothness): f is bounded from below by $f^*:=\inf_{\mathbf{x}\in\mathbb{R}^d}f(\mathbf{x})$ and its gradient ∇f is L-Lipschitz i.e. $\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|\le L\|\mathbf{x}-\mathbf{y}\|$, for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$.

Assumption 2 (μ-strongly convex): There exists a positive constant $\mu>0$ such that

$$f(\mathbf{y})\ge f(\mathbf{x})+\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle+\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2, \tag{8}$$

for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ A consequence of f being μ-strongly convex is that

$$\frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2\ge f(\mathbf{x})-f^*. \tag{9}$$

Assumption 3 (Directional Invariance): There exists a constant $D>0$ such that

$$\mathbb{E}\left[\langle\nabla f(\mathbf{x}_t),\frac{\nabla f(\mathbf{x}_t;\xi)}{\|\nabla f(\mathbf{x}_t;\xi)\|}\rangle\mid\mathbf{x}_t\right]\ge D\|\nabla f(\mathbf{x}_t)\| \tag{10}$$

holds whenever $\nabla f(\mathbf{x}_t)\neq 0$.

Remark 2: This assumption essentially states that the direction of $\nabla f(\mathbf{x};\xi)$ is preserved if we normalize it, and that the distribution of $\nabla f(\mathbf{x};\xi)$ is not extremely skewed.

III. Background and Lemmas on Supermartingales

The analysis in this paper follows from the following result derived in [14]. From this section onward, we use the shorthand notation $\mathbb{E}_t[\cdot]:=\mathbb{E}[\cdot|\mathbf{x}_t]$.

Proposition 1: Let $\{X_t\}$, $\{Y_t\}$, and $\{Z_t\}$ be three sequences of random variables that are adapted to a filtration $\{\mathscr{F}_t\}$. Let $\{\gamma_t\}$ be a sequence of nonnegative real numbers such that $\prod_{t=1}^{\infty}(1+\gamma_t)<\infty$. Suppose that the following conditions hold:

1) X_t,Y_t,Z_t are nonnegative for all $t\ge 1$.
2) $\mathbb{E}[Y_{t+1}|\mathscr{F}_t]\le(1+\gamma_t)Y_t-X_t+Z_t$ for all $t\ge 1$.
3) $\sum_{t=1}^{\infty}Z_t<\infty$ holds almost surely.

Then, we have

$$\sum_{t=1}^{\infty}X_t<\infty\quad\text{almost surely,} \tag{11}$$

and Y_t converges almost surely.

The following result from [10] is used for convergence results in non-convex settings.

Lemma 1 (Lemma 2 of [10]): Let $\{X_t\}$ be a sequence of nonnegative real numbers and $\{\alpha_t\}$ be a decreasing sequence of positive real numbers such that the following conditions hold:

$$\sum_{t=1}^{\infty}\alpha_t X_t<\infty,\quad\sum_{t=1}^{\infty}\frac{\alpha_t}{\sum_{i=1}^{t-1}\alpha_i}=\infty. \tag{12}$$

Then,

$$\min_{1\le i\le t}X_i=o\left(\frac{1}{\sum_{i=1}^{t-1}\alpha_i}\right). \tag{13}$$

We derive some properties of the differentially private stochastic gradient that aids in our core theory.

Proposition 2: For all $\mathbf{x}_t$ and ξ_t,

$$\mathbb{E}_t\|\mathbf{g}_t^{DP}\|^2\le q^2+q^2 d\sigma_{DP}^2. \tag{14}$$

Furthermore, if Assumption 3 holds, then

$$-\mathbb{E}_t[\langle\nabla f(\mathbf{x}_t),\mathbf{g}_t^{DP}\rangle]\le-(1-\eta_t)\|\nabla f(\mathbf{x}_t)\|^2-D\eta_t q\|\nabla f(\mathbf{x}_t)\|. \tag{15}$$

Proof: We first expand $\mathbb{E}_t\|\mathbf{g}_t^{DP}\|^2$:

$$\mathbb{E}_t\|\mathbf{g}_t^{DP}\|^2 = \mathbb{E}_t\langle \text{clip}_q(\nabla f(\mathbf{x}_t;\xi_t)) + q\zeta_t, \text{clip}_q(\nabla f(\mathbf{x}_t;\xi_t)) + q\zeta_t\rangle.$$

Since ξ_t and ζ_t are independent, we can separate the terms:

$$\begin{aligned}\mathbb{E}_t\|\mathbf{g}_t^{DP}\|^2 &= \mathbb{E}_t\|\text{clip}_q(\nabla f(\mathbf{x}_t;\xi_t))\|^2 + \mathbb{E}_t\|q\zeta_t\|^2 \\ &\quad + 2\mathbb{E}_t\langle \text{clip}_q(\nabla f(\mathbf{x}_t;\xi_t)), q\zeta_t\rangle \\ &= \mathbb{E}_t\|\text{clip}_q(\nabla f(\mathbf{x}_t;\xi_t))\|^2 + q^2 d\sigma_{DP}^2 + 0 \\ &\le q^2 + q^2 d\sigma_{DP}^2,\end{aligned}$$

where we exploit that $\|\text{clip}_q(\cdot)\| \le q$ for any vector. If Assumption 3 holds, then

$$\begin{aligned}-\mathbb{E}_t[\langle \nabla f(\mathbf{x}_t), \mathbf{g}_t^{DP}\rangle] &= -(1-\eta_t)\|\nabla f(\mathbf{x}_t)\|^2 \\ &\quad - \eta_t q\mathbb{E}_t[\langle \nabla f(\mathbf{x}_t), \tfrac{\nabla f(\mathbf{x}_t;\xi_t)}{\|\nabla f(\mathbf{x};\xi_t)\|}\rangle] \\ &\le -(1-\eta_t)\|\nabla f(\mathbf{x}_t)\|^2 - D\eta_t q\|\nabla f(\mathbf{x}_t)\|.\end{aligned}$$

■

IV. Almost Sure Convergence Rate Analysis with Privacy Guarantees

In this section, we derive convergence rates for three differentially private stochastic gradient methods. we use the following notation:

$$\Phi_t(\mathbf{x}) := (1-\eta_t)\|\nabla f(\mathbf{x})\|^2 + D\eta_t q\|\nabla f(\mathbf{x})\|, \tag{16}$$

$$\Phi_t^{\mu}(\mathbf{x}) := (1-\eta_t)(2\mu(f(\mathbf{x}) - f^*)) + D\eta_t q\sqrt{2\mu(f(\mathbf{x}) - f^*)}. \tag{17}$$

If Assumption 2 holds, then $\Phi_t^{\mu}(\mathbf{x}) \le \Phi_t(\mathbf{x})$ as $2\mu(f(\mathbf{x}) - f^*) \le \|\nabla f(\mathbf{x})\|^2$.

A. Stochastic Gradient Descent

The iteration of DP-SGD is given in Definition 3.

Theorem 1 (Convergence of DP-SGD): Consider the iterates of DP-SGD. Let η_t be the clipping probability defined in Definition 4. If Assumptions 1, 3 hold and $\alpha_t = \Theta(\frac{1}{t^{1-\theta}})$ for some $\theta \in (0, \frac{1}{2})$, then

$$\min_{1\le i\le t}\Phi_i(\mathbf{x}_i) = o\Big((\sum_{i=1}^{t-1}\alpha_i)^{-1}\Big) \tag{18}$$

almost surely. Furthermore, if Assumption 2 holds, then

$$\min_{1\le i\le t}\Phi_i^{\mu}(\mathbf{x}_i) = o\Big((\sum_{i=1}^{t-1}\alpha_i)^{-1}\Big) \tag{19}$$

Proof: By Assumption 1, we have

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \alpha_t\langle \nabla f(\mathbf{x}_t), \mathbf{g}_t^{DP}\rangle + \frac{L\alpha_t^2}{2}\|\mathbf{g}_t^{DP}\|^2.$$

Taking the expectation $\mathbb{E}_t[\cdot]$ of both sides gives

$$\begin{aligned}\mathbb{E}_t[f(\mathbf{x}_{t+1}) - f^*] \le\ & f(\mathbf{x}_t) - f^* - \alpha_t\mathbb{E}_t[\langle \nabla f(\mathbf{x}_t), \mathbf{g}_t^{DP}\rangle] \\ &+ \frac{L\alpha_t^2}{2}\mathbb{E}_t\|\mathbf{g}_t^{DP}\|^2.\end{aligned}$$

By Proposition 2,

$$\mathbb{E}_t\|\mathbf{g}_t^{DP}\|^2 \le q^2 + q^2 d\sigma_{DP}^2, \tag{20}$$

$$-\mathbb{E}_t[\langle \nabla f(\mathbf{x}_t), \mathbf{g}_t^{DP}\rangle] \le -\Phi_t(\mathbf{x}_t). \tag{21}$$

Plugging these terms simplifies the expression to

$$\begin{aligned}\mathbb{E}_t[f(\mathbf{x}_{t+1}) - f^*] \le & f(\mathbf{x}_t) - f^* - \alpha_t\Phi_t(\mathbf{x}_t) \\ &+ \frac{L\alpha_t^2}{2}q^2(1 + d\sigma_{DP}^2).\end{aligned} \tag{22}$$

Non-Convex Case. By Proposition 1, equation (22) gives

$$\sum_{t=1}^{\infty}\alpha_t\Phi_t(\mathbf{x}_t) < \infty.$$

Thus from Lemma 1,

$$\min_{1\le i\le t}\Phi_i(\mathbf{x}_i) = o\Big((\sum_{i=1}^{t-1}\alpha_i)^{-1}\Big).$$

Strongly Convex Case. From strong convexity:

$$\|\nabla f(\mathbf{x}_t)\|^2 \ge 2\mu(f(\mathbf{x}_t) - f^*).$$

equation (22) simplifies to

$$\begin{aligned}\mathbb{E}_t[f(\mathbf{x}_{t+1}) - f^*] \le & (1 - 2\alpha_t(1-\eta_t)\mu)(f(\mathbf{x}_t) - f^*) \\ &- \alpha_t D\eta_t q\|\nabla f(\mathbf{x}_t)\| + \frac{L\alpha_t^2}{2}q^2(1 + d\sigma_{DP}^2) \\ \le & (1 - 2\alpha_t(1-\eta_t)\mu)(f(\mathbf{x}_t) - f^*) \\ &- \alpha_t D\eta_t q\sqrt{2\mu(f(\mathbf{x}_t) - f^*)} \\ &+ \frac{L\alpha_t^2}{2}q^2(1 + d\sigma_{DP}^2).\end{aligned}$$

By Proposition 1, we conclude that

$$\sum_{t=1}^{\infty}\alpha_t[(1-\eta_t)(2\mu(f(\mathbf{x}_t) - f^*)) + D\eta_t q\sqrt{2\mu(f(\mathbf{x}_t) - f^*)}] < \infty,$$

and therefore from Lemma 1,

$$\min_{1\le i\le t}\Phi_i^{\mu}(\mathbf{x}_i) = o\Big((\sum_{i=1}^{t-1}\alpha_i)^{-1}\Big),$$

which concludes the proof. ■

B. Stochastic Heavy-Ball Method

The iteration of the differentially private stochastic heavy-ball (DP-SHB) method is given by

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t\mathbf{g}_t^{DP} + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}), \tag{23}$$

where $\beta \in [0, 1)$ is the weight given to the momentum component. To simplify our analysis, we express the DP-SHB iteration as a system of two variables. Define

$$\mathbf{z}_t = \mathbf{x}_t + \frac{\beta}{1-\beta}\mathbf{v}_t, \quad \mathbf{v}_t = \mathbf{x}_t - \mathbf{x}_{t-1}. \tag{24}$$

The iteration of SHB can be rewritten as

$$\mathbf{v}_{t+1} = \beta\mathbf{v}_t - \alpha\mathbf{g}_t^{DP}, \quad \mathbf{z}_{t+1} = \mathbf{z}_t - \frac{\alpha_t}{1-\beta}\mathbf{g}_t^{DP}. \tag{25}$$

This update rule is derived in [11].

Theorem 2 (Convergence of DP-SHB): Let $\{\mathbf{x}_t\}$ be the iterates of DP-SHB. Let η_t be the clipping probability defined in Definition 4. Define the differentially private stochastic gradient as in equation (4). If Assumptions 1, 3 hold and $\alpha_t = \Theta(\frac{1}{t^{1-\theta}})$ for some $\theta \in (0, \frac{1}{2})$, then

$$\min_{1\le i\le t} \Phi_i(\mathbf{x}_i) = o\Big(\big(\sum_{i=1}^{t-1}\alpha_i\big)^{-1}\Big) \tag{26}$$

almost surely. Furthermore, if Assumption 2 holds, then

$$\min_{1\le i\le t} \Phi_i^{\mu}(\mathbf{x}_i) = o\Big(\big(\sum_{i=1}^{t-1}\alpha_i\big)^{-1}\Big) \tag{27}$$

Proof: Define the energy function

$$Y_t := f(\mathbf{z}_t) - f^* + c\|\mathbf{v}_t\|^2.$$

For a constant $c > 0$. Moreover,

$$\|\mathbf{v}_{t+1}\|^2 = \beta^2\|\mathbf{v}_t\|^2 + \alpha_t^2\|\mathbf{g}_t^{\mathrm{DP}}\|^2 - 2\alpha_t\beta\langle\mathbf{v}_t, \mathbf{g}_t^{\mathrm{DP}}\rangle.$$

Taking $\mathbb{E}_t[\cdot]$ of both sides gives

$$\mathbb{E}_t\|\mathbf{v}_{t+1}\|^2 \le \beta^2\|\mathbf{v}_t\|^2 + \alpha_t^2\mathbb{E}_t\|\mathbf{g}_t^{\mathrm{DP}}\|^2 - 2\alpha_t\beta\mathbb{E}_t\langle\mathbf{v}_t, \mathbf{g}_t^{\mathrm{DP}}\rangle.$$

Using Young's inequality and Proposition 2, we introduce a constant $c_1 > 0$ such that

$$\begin{aligned}\mathbb{E}_t\|\mathbf{v}_{t+1}\|^2 \le\ & \beta^2\|\mathbf{v}_t\|^2 + \alpha_t^2 q^2(1 + d\sigma_{DP}^2) \qquad (28)\\ & + c_1\|\mathbf{v}_t\|^2 + \tfrac{\alpha_t^2\beta^2}{c_1}\mathbb{E}_t\|\mathbf{g}_t^{DP}\|^2\\ \le\ & (\beta^2 + c_1)\|\mathbf{v}_t\|^2 + \alpha_t^2 q^2(1 + \tfrac{\beta^2}{c_1})(1 + d\sigma_{DP}^2). \qquad (29)\end{aligned}$$

By Assumption 1,

$$f(\mathbf{z}_{t+1}) \le f(\mathbf{z}_t) - \tfrac{\alpha_t}{1-\beta}\langle\nabla f(\mathbf{z}_t), \mathbf{g}_t^{DP}\rangle + \tfrac{L\alpha_t^2}{2(1-\beta)^2}\|\mathbf{g}_t^{DP}\|^2.$$

Taking $\mathbb{E}_t[\cdot]$ of both sides gives

$$\mathbb{E}_t f(\mathbf{z}_{t+1}) \le f(\mathbf{z}_t) - \tfrac{\alpha_t}{1-\beta}\mathbb{E}_t\langle\nabla f(\mathbf{z}_t), \mathbf{g}_t^{DP}\rangle + \tfrac{L\alpha_t^2}{2(1-\beta)^2}\mathbb{E}_t\|\mathbf{g}_t^{DP}\|^2. \tag{30}$$

To bound $\mathbb{E}_t\langle\nabla f(\mathbf{z}_t), \mathbf{g}_t^{DP}\rangle$, we can expand it:

$$\begin{aligned}& -\mathbb{E}_t\langle\nabla f(\mathbf{z}_t), \mathbf{g}_t^{DP}\rangle\\ &= -\mathbb{E}_t\langle\nabla f(\mathbf{x}_t), \mathbf{g}_t^{DP}\rangle - \mathbb{E}_t\langle\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t), \mathbf{g}_t^{DP}\rangle\\ &\le -\Phi_t(\mathbf{x}_t) + \|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\|\|\mathbb{E}_t\mathbf{g}_t^{DP}\|\\ &\le -\Phi_t(\mathbf{x}_t) + Lq\sqrt{1 + d\sigma_{DP}^2}\|\mathbf{z}_t - \mathbf{x}_t\|\\ &\le -\Phi_t(\mathbf{x}_t) + \tfrac{Lq\beta}{1-\beta}\sqrt{1 + d\sigma_{DP}^2}\|\mathbf{v}_t\|.\end{aligned}$$

Letting $K := \frac{Lq\beta}{1-\beta}\sqrt{1 + d\sigma_{DP}^2}$, equation (30) simplifies to

$$\mathbb{E}_t f(\mathbf{z}_{t+1}) \le f(\mathbf{z}_t) - \tfrac{\alpha_t}{1-\beta}\Phi_t(\mathbf{x}_t) + \tfrac{\alpha_t}{1-\beta}K\|\mathbf{v}_t\| + \tfrac{K^2}{2L}\alpha_t^2.$$

Finally, we control the extra $\|\mathbf{v}_t\|$ term using Young's inequality by introducing a new constant $c_2 > 0$ such that

$$\tfrac{\alpha_t}{1-\beta}K\|\mathbf{v}_t\| \le \tfrac{c_2(1-\beta)}{2}\|\mathbf{v}_t\|^2 + \tfrac{K^2}{2c_2(1-\beta)^3}\alpha_t^2.$$

Combining with equation (28), we can now express the energy function Y_t as a supermartingale

$$\begin{aligned}&\mathbb{E}_t[f(\mathbf{z}_{t+1}) - f^* + c\|\mathbf{v}_{t+1}\|^2]\\ &\le f(\mathbf{z}_t) - f^* - \tfrac{\alpha_t}{1-\beta}\Phi_t(\mathbf{x}_t) + (\tfrac{c_2(1-\beta)}{2} + c\beta^2 + cc_1)\|\mathbf{v}_t\|^2\\ &\quad + \alpha_t^2[\tfrac{K^2}{2L} + q^2c(1 + \tfrac{\beta^2}{c_1})(1 + d\sigma_{DP}^2)]. \qquad (31)\end{aligned}$$

We can choose positive values for c, c_1, and c_2 carefully such that $c = \frac{c_2(1-\beta)}{2} + c\beta^2 + cc_1$, for example, $c_1 = \frac{1-\beta^2}{2}$, $c = \frac{c_2}{1+\beta}$. And let c_3 be the coefficient of α_t^2 for simplification. This yields the clean recursion

$$\mathbb{E}[Y_{t+1}] \le Y_t - \tfrac{\alpha_t}{1-\beta}\Phi_t(\mathbf{x}_t) + c_3\alpha_t^2. \tag{32}$$

Non-Convex Case. Summing the recursion and applying Proposition 1 give

$$\sum_{t=1}^{\infty}\frac{\alpha_t}{1-\beta}\Phi_t(\mathbf{x}_t) < \infty,$$

which by Lemma 1 implies

$$\min_{1\le i\le t} \Phi_i(\mathbf{x}_i) = o\Big(\big(\sum_{i=1}^{t-1}\alpha_i\big)^{-1}\Big), \tag{33}$$

almost surely.

Strongly convex case. If f is μ-strongly convex, then $\|\nabla f(\mathbf{x}_t)\|^2 \ge 2\mu(f(\mathbf{x}_t) - f^*)$. Plugging this yields

$$\min_{1\le i\le t} \Phi_i^{\mu}(\mathbf{x}_i) = o\Big(\big(\sum_{i=1}^{t-1}\alpha_i\big)^{-1}\Big) \tag{34}$$

This completes the proof. ■

C. Stochastic Nesterov's Accelerated Gradient

The iteration of the differentially private stochastic Nesterov's accelerated gradient (DP-NAG) is given by

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \alpha_t\mathbf{g}_t^{DP}, \tag{35}$$

$$\mathbf{x}_{t+1} = \mathbf{y}_t + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}), \tag{36}$$

where $\beta \in [0, 1)$ is the weight given to the momentum component.

Theorem 3 (Convergence of DP-NAG): Let $\{\mathbf{x}_t\}$ be the iterates of DP-NAG. Let η_t be the clipping probability defined in Definition 4. Define the differentially private stochastic gradient as in equation (4). If Assumptions 1, 3 hold and $\alpha_t = \Theta(\frac{1}{t^{1-\theta}})$ for some $\theta \in (0, \frac{1}{2})$, then

$$\min_{1\le i\le t} \Phi_i(\mathbf{x}_i) = o\Big(\big(\sum_{i=1}^{t-1}\alpha_i\big)^{-1}\Big) \tag{37}$$

almost surely. Furthermore, if Assumption 2 holds, then

$$\min_{1\le i\le t} \Phi_i^{\mu}(\mathbf{x}_i) = o\Big(\big(\sum_{i=1}^{t-1}\alpha_i\big)^{-1}\Big). \tag{38}$$

Proof: Define $\mathbf{v}_t$ and $\mathbf{z}_t$ as in equation (24). The iteration of DP-NAG can be rewritten as

$$\mathbf{v}_{t+1} = \beta \mathbf{v}_t - \beta \alpha_t \mathbf{g}_t^{DP}, \quad \mathbf{z}_{t+1} = \mathbf{z}_t - \tfrac{\alpha_t}{1-\beta} \mathbf{g}_t^{DP}. \tag{39}$$

The proof is identical to that of Theorem 2 with

$$\|\mathbf{v}_{t+1}\|^2 = \beta^2[\|\mathbf{v}_t\|^2 + \alpha_t^2 \|\mathbf{g}_t^{\mathrm{DP}}\|^2 - 2\alpha_t \langle \mathbf{v}_t, \mathbf{g}_t^{\mathrm{DP}} \rangle].$$

■

V. Last-iterate Convergence Analysis

The convergence analysis results above show that the "best" iterate converges to zero almost surely in the strongly convex and non-convex case. To extend the almost sure convergence guarantee from best-iterate results to the last iterate, we need a device that controls oscillations of the gradient norm across iterations. Even if $\sum_t \alpha_t \|\nabla f(\mathbf{x}_t)\|^2 \to 0$, this condition alone does not imply that $\nabla f(\mathbf{x}_t) \to 0$ because the sequence could fluctuate indefinitely. The key tool to overcome this difficulty is a lemma of [13], which ensures that if the weighted sum of squared gradients is finite and the gradient sequence does not vary too quickly, then the gradients themselves converge to zero. We restate a suitable version below.

Lemma 2 (Lemma 1 of [13]): Let $\{b_t\}$ and $\{\alpha_t\}$ be two nonnegative sequences and $\{w_t\}$ be a sequence of vectors. Assume $\sum_{t=1}^{\infty} \alpha_t b_t^p < \infty$ and $\sum_{t=1}^{\infty} \alpha_t = \infty$, where $p \geq 1$. Furthermore, assume that there exists some $L > 0$ such that

$$|b_{t+\tau} - b_t| \leq L\Big(\sum_{i=t}^{t+\tau-1} \alpha_i b_i + \Big\| \sum_{i=t}^{t+\tau-1} \alpha_i w_i \Big\| \Big),$$

where w_t is such that $\sum_{t=1}^{\infty} \alpha_t w_t$ converges. Then b_t converges to 0. See also Lemma 10 of [11] for the case $p > 0$.

Lemma 2 provides a general criterion for establishing last-iterate convergence: it reduces the problem to showing that the cumulative bias and noise terms introduced by DP-SGD induced by clipping and Gaussian perturbations form a convergent series. We verify this condition under our assumptions by combining the supermartingale recursion established in Theorem 2 with additional bias control for the clipped gradient. This allows us to use Lemma 2 and conclude that the last-iterate gradients vanish almost surely. The formal statement is given in Theorem 4.

Theorem 4: Consider the iterates of DP-SHB and DP-NAG. Let Assumptions 1 and 3 hold, and assume $q \geq 1$. Let the step size $\{\alpha_t\}$ satisfy $\sum_{t=1}^{\infty} \alpha_t = \infty, \sum_{t=1}^{\infty} \alpha_t^2 < \infty$. Then we have $\nabla f(\mathbf{x}_t) \to 0$ almost surely as $t \to \infty$.

Proof: We revisit the convergence proof for DP-SHB. By L-smoothness of f,

$$\begin{aligned}
\mathbb{E}_t\big[f(\mathbf{z}_{t+1})\big] &\leq f(\mathbf{z}_t) - \tfrac{\alpha_t}{1-\beta} \mathbb{E}_t\big[\langle \nabla f(\mathbf{z}_t), \mathbf{g}_t^{\mathrm{DP}} \rangle\big] \\
&\quad + \tfrac{L}{2}\Big(\tfrac{\alpha_t}{1-\beta}\Big)^2 \mathbb{E}_t\big[\|\mathbf{g}_t^{\mathrm{DP}}\|^2\big] \\
&\leq f(\mathbf{z}_t) - \tfrac{\alpha_t}{1-\beta} \|\nabla f(\mathbf{z}_t)\|^2 \\
&\quad - \tfrac{\alpha_t}{1-\beta} \mathbb{E}_t\big[\langle \nabla f(\mathbf{z}_t), \mathbf{g}_t^{\mathrm{DP}} - \nabla f(\mathbf{z}_t) \rangle\big] \\
&\quad + \tfrac{L}{2}\Big(\tfrac{\alpha_t}{1-\beta}\Big)^2 q^2(1 + d\sigma_{DP}^2).
\end{aligned}$$

Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
&-\mathbb{E}_t\big[\langle \nabla f(\mathbf{z}_t), \mathbf{g}_t^{\mathrm{DP}} - \nabla f(\mathbf{z}_t) \rangle\big] \\
&\leq \|\nabla f(\mathbf{z}_t)\| \|\mathbb{E}_t \mathbf{g}_t^{\mathrm{DP}} - \nabla f(\mathbf{z}_t)\| \\
&\leq \|\nabla f(\mathbf{z}_t)\| [\|\mathbb{E}_t \mathbf{g}_t^{\mathrm{DP}} - \mathrm{clip}_q(\nabla f(\mathbf{z}_t))\| \\
&\quad + \|\mathrm{clip}_q(\nabla f(\mathbf{z}_t)) - \nabla f(\mathbf{z}_t)\|] \\
&\leq \|\nabla f(\mathbf{z}_t)\| [q\|\mathbf{x}_t - \mathbf{z}_t\| + \max(\|\nabla f(\mathbf{z}_t)\| - q, 0)] \\
&\leq \|\nabla f(\mathbf{z}_t)\| [\tfrac{qL\beta}{1-\beta} \|\mathbf{v}_t\| + \max(\|\nabla f(\mathbf{z}_t)\| - q, 0)],
\end{aligned} \tag{40}$$

where $\|\mathbb{E}_t \mathbf{g}_t^{\mathrm{DP}} - \mathrm{clip}_q(\nabla f(\mathbf{z}_t))\| \leq q\|\mathbf{x}_t - \mathbf{z}_t\|$ comes from exploiting the q-Lipschitz property of $\mathrm{clip}_q(\cdot)$. Combining with equation (28), we have

$$\begin{aligned}
&\mathbb{E}_t\big[f(\mathbf{z}_{t+1}) - f^* + \|\mathbf{v}_{t+1}\|^2\big] \\
&\leq f(\mathbf{z}_t) - f^* - \tfrac{\alpha_t}{1-\beta} \min(\|\nabla f(\mathbf{z}_t)\|^2, q\|\nabla f(\mathbf{z}_t)\|) \\
&\quad + \tfrac{\alpha_t}{1-\beta} \|\nabla f(\mathbf{z}_t)\| \tfrac{L\beta}{1-\beta} \|\mathbf{v}_t\| + \tfrac{L}{2}\Big(\tfrac{\alpha_t}{1-\beta}\Big)^2 q^2(1 + d\sigma_{DP}^2) \\
&\quad + (\beta^2 + c_1)\|\mathbf{v}_t\|^2 + \alpha_t^2 q^2 (1 + \tfrac{\beta^2}{c_1})(1 + d\sigma_{DP}^2) \\
&\leq f(\mathbf{z}_t) - f^* - \tfrac{\alpha_t}{1-\beta} \min(\|\nabla f(\mathbf{z}_t)\|^2, q\|\nabla f(\mathbf{z}_t)\|) \\
&\quad + \tfrac{\alpha_t^2 L^2 \beta^2}{c_4(1-\beta)^4} \|\nabla f(\mathbf{z}_t)\|^2 + (\beta^2 + c_1 + c_4)\|\mathbf{v}_t\|^2 + \alpha_t^2 C_2,
\end{aligned}$$

where $c_1 > 0$ comes from using Young's inequality, and C_2 is the coefficient of α_t^2. Finally, for sufficiently large t, there exists a positive constant c_5 such that

$$-\tfrac{\alpha_t}{1-\beta} + \tfrac{\alpha_t^2 L^2 \beta^2}{c_4(1-\beta)^4} \leq -\tfrac{c_5}{1-\beta} \alpha_t.$$

This simplifies our bound to

$$\begin{aligned}
&\mathbb{E}_t\big[f(\mathbf{z}_{t+1}) - f^* + \|\mathbf{v}_{t+1}\|^2\big] \\
&\leq f(\mathbf{z}_t) - f^* - \tfrac{\alpha_t}{1-\beta} \min((1+c_5)\|\nabla f(\mathbf{z}_t)\|^2, q\|\nabla f(\mathbf{z}_t)\|) \\
&\quad + (\beta^2 + c_1 + c_4)\|\mathbf{v}_t\|^2 + \alpha_t^2 C_2.
\end{aligned}$$

By Proposition 1, we conclude that

$$\sum_{t=1}^{\infty} \alpha_t \min((1+c_5)\|\nabla f(\mathbf{z}_t)\|^2, q\|\nabla f(\mathbf{z}_t)\|) < \infty, \tag{41}$$

almost surely. Furthermore, with a careful choice of c_1 and c_4 such that $\beta^2 + c_1 + c_4 < 1$, by Proposition 1, we conclude that

$$\sum_{t=1}^{\infty} \alpha_t \|\mathbf{v}_t\|^2 < \infty.$$

For the next part of the proof, we want to show that the inequality in Lemma 2 holds. Define the "error" sequence

$$\mathbf{w}_t := \mathbf{g}_t^{\mathrm{DP}} - \nabla f(\mathbf{z}_t) \quad \text{and} \quad \alpha_t' := \tfrac{\alpha_t}{1-\beta}.$$

By $\mathbf{z}_{t+1} = \mathbf{z}_t - \alpha'_t(\nabla f(\mathbf{z}_t) + \mathbf{w}_t)$. Since ∇f is L-Lipschitz, for any $\tau \geq 1$,

$$\begin{aligned}
&\big|\|\nabla f(\mathbf{z}_{t+\tau})\| - \|\nabla f(\mathbf{z}_t)\|\big| \\
&\leq \big\|\nabla f(\mathbf{z}_{t+\tau}) - \nabla f(\mathbf{z}_t)\big\| \\
&\leq L\Big\|\mathbf{z}_{t+\tau} - \mathbf{z}_t\Big\| \\
&\leq L\Big\|\sum_{i=t}^{t+\tau-1} \alpha'_i\big(\nabla f(\mathbf{z}_i) + \mathbf{w}_i\big)\Big\| \\
&\leq L\sum_{i=t}^{t+\tau-1} \alpha'_i\|\nabla f(\mathbf{z}_i)\| + L\Big\|\sum_{i=t}^{t+\tau-1} \alpha'_i\mathbf{w}_i\Big\|.
\end{aligned}$$

Therefore, setting $b_t := \|\nabla f(\mathbf{z}_t)\|$,

$$|b_{t+\tau} - b_t| \leq L\sum_{i=t}^{t+\tau-1} \alpha'_i b_i + L\Big\|\sum_{i=t}^{t+\tau-1} \alpha'_i\mathbf{w}_i\Big\|. \tag{42}$$

We first show $\sum_t \alpha'_t b_t^2 < \infty$. From equation (32), the sequence $\sum_t \alpha'_t\Phi_t(\mathbf{x}_t)$ is finite. Using $\|\mathbf{z}_t - \mathbf{x}_t\| \to 0$, L-smoothness implies $\|\nabla f(\mathbf{z}_t)\|^2 \leq 2\|\nabla f(\mathbf{x}_t)\|^2 + 2L^2\|\mathbf{z}_t - \mathbf{x}_t\|^2$. Thus

$$\sum_t \alpha'_t\|\nabla f(\mathbf{z}_t)\|^2 \leq 2\sum_t \alpha'_t\|\nabla f(\mathbf{x}_t)\|^2 + \frac{2L^2\beta^2}{(1-\beta)^3}\sum_t \alpha_t\|v_t\|^2 < \infty.$$

Next, we show $\sum \alpha'_t\mathbf{w}_t$ converges almost surely. Decompose

$$\begin{aligned}
\mathbf{w}_t = &\underbrace{\big(\mathrm{clip}_q(\nabla f(\mathbf{x}_t;\xi_t)) - \mathbb{E}_t[\mathrm{clip}_q(\nabla f(\mathbf{x}_t;\xi_t))]\big)}_{=:U_t} \\
&+ \underbrace{\mathbb{E}_t[\mathrm{clip}_q(\nabla f(\mathbf{x}_t;\xi_t))] - \nabla f(\mathbf{z}_t)}_{=:T_t} + \underbrace{q\zeta_t}_{=:G_t}.
\end{aligned}$$

Here U_t is a martingale difference with $\mathbb{E}_t\|U_t\|^2 \leq q^2$ and G_t is zero-mean Gaussian with variance $q^2\sigma_{\mathrm{DP}}^2 I$. Hence $\sum_t \alpha'_t U_t$ and $\sum_t \alpha'_t G_t$ converge almost surely due to being martingales bounded in $\mathscr{L}^2$ [16, Theorem 12.1].

Finally, the transfer term $\sum_t \alpha'_t T_t$ expands to $\sum_t \alpha'_t \frac{qL\beta}{1-\beta}\|\mathbf{v}_t\| + \max(\|\nabla f(\mathbf{z}_t)\| - q, 0)$ as shown in equation (40), and $\sum_t \alpha_t\|\mathbf{v}_t\| < \infty$. If $q \geq 1$, then $\max(\|\nabla f(\mathbf{z}_t)\| - q, 0) \leq \min(\|\nabla f(\mathbf{z}_t)\|^2, q\|\nabla f(\mathbf{z}_t)\|)$. And the convergence of $\sum_t \alpha'_t \min(\|\nabla f(\mathbf{z}_t)\|^2, q\|\nabla f(\mathbf{z}_t)\|)$ follows by equation (41).

With (i) and (ii), Lemma 2 with $p = 2$, $b_t = \|\nabla f(\mathbf{z}_t)\|$, $\alpha'_t = \alpha_t/(1-\beta)$, and (42) implies $\|\nabla f(\mathbf{z}_t)\| \to 0$, almost surely. For DP-NAG, define $\mathbf{y}_t = \mathbf{x}_t + \beta(\mathbf{x}_t - \mathbf{x}_{t-1})$ (the look-ahead point) and use the update $\mathbf{x}_{t+1} = \mathbf{y}_t - \alpha_t\mathbf{g}_t^{\mathrm{DP}}(\mathbf{y}_t)$. The same proof applies with $\mathbf{z}_t$ replaced by $\mathbf{y}_t$ and $\alpha'_t = \alpha_t/(1-\beta)$. ■

Remark 3: Almost sure convergence of $f(\mathbf{x}_t)$ trivially follows if f is μ-strongly convex as $\mu(f(\mathbf{x}_t) - f^*) \leq \frac{1}{2}\|\nabla f(\mathbf{x}_t)\|^2$.

VI. Conclusion

In this paper, we established the first almost sure convergence guarantees for differentially private stochastic gradient descent (DP-SGD) and its momentum variants, including DP-SHB and DP-NAG. Our analysis adapts supermartingale techniques to handle the combined challenges of gradient clipping and Gaussian noise injection, which break the unbiasedness and smooth descent properties that underlie classical SGD proofs. We showed that, under standard assumptions, the iterates converge almost surely to stationary points in the non-convex setting and to the global minimizer in the strongly convex setting. Our results provide pathwise convergence, ensuring that individual runs of DP-SGD stabilize rather than merely converging in expectation. This strengthens the theoretical foundation for deploying DP-SGD in practice, where guarantees for single trajectories are often more relevant than averaged behaviors.

Several directions remain open. Our analysis provides almost sure convergence guarantees regardless of the choices of the clipping parameter q or the variance of the injected noise σ_{DP}^2. However, increasing either of these parameters will naturally slow down the convergence rate in practice. Deriving convergence rates that depend on these parameters will be an interesting area for future work.

References

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

[2] B. Chen, K. Leahy, A. Jones, and M. Hale. Differential privacy for symbolic systems with application to markov chains. *Automatica*, 152:110908, 2023.

[3] C. Chen, J. Fu, and L. Lyu. A pathway towards responsible ai generated content. *arXiv preprint arXiv:2303.01325*, 2023.

[4] T. T. Doan. Finite-time analysis of markov gradient descent. *IEEE Transactions on Automatic Control*, 68(4):2140–2153, 2022.

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

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

[7] L. Huang, J. Wu, D. Shi, S. Dey, and L. Shi. Differential privacy in distributed optimization with gradient tracking. *IEEE Transactions on Automatic Control*, 69(9):5727–5742, 2024.

[8] V. Katewa, A. Chakrabortty, and V. Gupta. Differential privacy for network identification. *IEEE Transactions on Control of Network Systems*, 7(1):266–277, 2019.

[9] A. Koloskova, H. Hendrikx, and S. U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, 2023.

[10] J. Liu and Y. Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory*, 2022.

[11] J. Liu and Y. Yuan. Almost sure convergence rates analysis and saddle avoidance of stochastic gradient methods. *Journal of Machine Learning Research*, 25(271):1–40, 2024.

[12] Y. Nesterov. Introductory Lectures on Convex Optimization. *Applied Optimization*, 87, 2004.

[13] F. Orabona. Almost sure convergence of SGD on smooth nonconvex functions. *Blogpost at https://parameterfree.com/2020/10/05/almost-sure-convergence-of-sgd-on-smooth-non-convex-functions*, 2020.

[14] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[15] Q. Tang, F. Shpilevskiy, and M. Lécuyer. DP-AdamBC: Your DP-Adam is actually DP-SGD (unless you apply bias correction). In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[16] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.