Correct Before You Clip: A Bias Correction Method for Differentially Private Stochastic Gradient Descent

Anonymous authors
Paper under double-blind review

Abstract

Differentially Private Stochastic Gradient Descent (DPSGD) is a foundational method for privacy-preserving machine learning, but its performance is compromised by the bias introduced by gradient clipping. While recent methods show progress in mitigating the utility-damaging bias of gradient clipping in DPSGD, they often compromise the computational efficiency and ease of tuning essential for deploying modern large models and datasets. We propose a novel algorithm called DP-MBC, which introduces a pre-clipping mechanism that constructs a corrected gradient estimate by adjusting the current gradient with a momentum-driven correction. We provide a rigorous convergence analysis for DP-MBC under non-convex settings and show that it satisfies differential privacy guarantees. Extensive experiments using large-scale image and sentence classification benchmarks show that DP-MBC achieves better privacy-utility trade-off than existing state-of-the-art methods, with simplified hyperparameter tuning and reduced memory and computational demands.

1 Introduction

While large-scale datasets have driven the remarkable success of deep learning, their inclusion of sensitive personal information makes trained models vulnerable to attacks such as membership inference (Shokri et al., 2017; Liu et al., 2022) and model inversion (Zhao et al., 2021; Nguyen et al., 2023). Differential Privacy (DP) (Dwork et al., 2014) provides a rigorous mathematical framework to address these risks. The core idea of DP is to ensure that the presence of any single individual in the dataset has a minimal effect on the algorithm's output. This is achieved by injecting calibrated noise.

The seminal approach for applying DP in deep learning is Differentially Private Stochastic Gradient Descent (DPSGD) (Abadi et al., 2016). DPSGD modifies the standard gradient update in two key ways to ensure privacy during the training of deep learning models. First, DPSGD computes per-sample gradients and clips their ℓ_2 norm to a fixed clipping threshold C. Second, after aggregating the clipped gradients, calibrated Gaussian noise is added to those gradients with noise magnitude scaled according to the bounded threshold C and the desired privacy budget ϵ. While DPSGD provides provable privacy guarantee, it also degrades the model utility, making the privacy-utility trade-off a key challenge (Jayaraman & Evans, 2019; Xu et al., 2021).

A key factor contributing to this model utility degradation is the bias induced by gradient clipping. (Koloskova et al., 2023). The bias appears because the updated direction of the average clipped per-sample gradients can be dramatically misaligned with the true batch gradients, profoundly affecting model performance (Zhang et al., 2024). To address this, existing research follows three main directions. First, the pre-clipping correction approach aims to refine the per-sample gradients before they are clipped (Xiao et al., 2023; Bao et al., 2023). The second direction focuses on modifying the clipping function to assign smaller weights to the per-sample gradients that exhibit greater bias during aggregation (Xia et al., 2023; Huang et al., 2024). The third direction is based on post-clipping compensation where the bias from clipping is estimated privately and then used for correction in the next iteration (Zhang et al., 2024).

Among these directions, Xiao *et al.* (Xiao et al., 2023) proposed the Inner-Outer Momentum (IOM) approach, which mitigates clipping bias by reducing sampling variance via two momentum buffers, Inner and

Outer. IOM smooths per-sample gradients with inner momentum before clipping and reduces batch gradient variance using outer momentum. However, IOM relies on a long-term history of gradients to reduce sampling variance of each per-sample gradient. Furthermore, it is constrained by high computational complexity and a substantial amount of hyperparameter tuning, which makes it less feasible in real applications.

To address those limitations in existing techniques, we propose a novel bias reduction algorithm: Differentially Private Momentum-based Bias Correction (DP-MBC) with lower computational overhead. The core of DP-MBC is inspired by a non-private optimizer STORM (Cutkosky & Orabona, 2019). DP-MBC uses a pre-clipping mechanism to build a low sampling variance gradient estimate, reducing the bias introduced by the subsequent clipping operation.

Specifically, DP-MBC adjusts the current gradient by incorporating both the historical momentum and a correction factor derived from the gradients at the previous model training iterations. We provide a theoretical analysis to demonstrate DP-MBC's convergence in non-convex settings with DP guarantees. We further evaluate DP-MBC using different image and sentence classification datasets. The empirical results show that our proposed approach outperforms existing state-of-the-art DPSGD variants. We summarize our main contributions below:

- We introduce DP-MBC, an efficient and robust pre-clipping correction method for DPSGD that reduces clipping bias with minimal computational and memory overhead.
- We provide a rigorous convergence analysis for DP-MBC in non-convex setting with provable differential privacy guarantees.
- We conduct extensive experiments on several benchmark datasets, demonstrating that DP-MBC achieves a better privacy-utility trade-off compared to various existing state-of-the-art DPSGD variants.

2 Related Work

Clipping Bias Reduction in DPSGD: DPSGD (Abadi et al., 2016) is the well-known standard for private deep learning. A key limitation of DPSGD is that the bias from gradient clipping prevents the model from converging to its optimum, ultimately degrading the model performance. An early line of research aims to eliminate this clipping bias while controlling DP noise by carefully tuning the clipping threshold C (Iyengar et al., 2019; Feldman et al., 2020; Chen et al., 2020b; Qian et al., 2021; Zhang et al., 2022). However, the theoretical guarantees of these methods rely on strong assumptions, such as the bound of the gradient magnitude or the coefficient of the gradient symmetry, which are difficult to verify in practice.

Given the difficulty of tuning the clipping threshold in practice, existing research has explored three main approaches: (i). Modifying the clipping function itself to reduce aggregation bias (Xia et al., 2023; Huang et al., 2024). Xia *et al.* proposed a non-monotonic normalization-based function that suppresses the influence of small gradients to reduce aggregation bias. However, this approach relies solely on gradient magnitudes and often yields limited performance improvements. (ii.) Post-clipping compensation, which attempts to eliminate clipping bias by introducing a corrective term (Zhang et al., 2024). However, this approach requires the compensation term to be privatized, introducing additional DP noise. (iii). Pre-clipping correction, which seeks to reduce per-sample sampling variance before clipping to mitigate bias (Xiao et al., 2023; Bao et al., 2023). These three approaches, however, are constrained by high computational complexity and require extensive hyperparameter tuning.

KFSGD (Zhang et al., 2025) is another recent state-of-the-art method that aims to improve DP optimizer performance. Though it does not aim to reduce clipping bias, KFSGD treats the privatized gradient as a noisy observation and employs a simplified Kalman filter to construct a refined gradient estimate. However, our approach differs fundamentally in how the correction is applied compared to KFSGD.

Sampling Variance Reduction in Non-private Optimization: The motivation behind pre-clipping correction for reducing sampling variance is a well-established direction in non-private stochastic optimization. SVRG (Johnson & Zhang, 2013) is among the first methods to address this using periodic full gradient

calculations in convex settings. Subsequently, Defazio *et al.* proposed SAGA (Defazio et al., 2014), which maintains all previous gradients adopted across iterations. Nguyen *et al.* (Nguyen et al., 2017) proposed a recursive approach to update gradient estimators (SARAH), ensuring better convergence for smooth convex functions. SPIDER (Fang et al., 2018) and SpiderBoost (Wang et al., 2019) are proposed to reduce the sampling variance in non-convex settings by using integer path technique to reduce sampling variance within a single loop framework. However, these methods depend on access to all historical gradients, which poses a significant challenge in differential privacy settings. This is because privatizing such sensitive information requires additional privacy budget, complicating their use under strict DP constraints.

More recently, Cutkosky *et al.* (Cutkosky & Orabona, 2019) proposed STORM, which employs a recursive momentum update to achieve fast convergence in non-convex optimization without relying on large batch sizes or full-gradient computations. Inspired by STORM, sampling variance reduction techniques have since been applied in wide range of settings, including composition optimization (Jiang et al., 2022; 2023), adaptive algorithms (Kavis et al., 2022; Jiang et al., 2024a), and sign-based optimization (Jiang et al., 2024b).

3 Preliminaries and Problem Formulation

In this work, we focus on the supervised learning setting. The goal is to train a model with parameters $x \in \mathbb{R}^d$ by minimizing an empirical loss function over a dataset D containing N data points. This is formally expressed as the Empirical Risk Minimization (ERM) problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{\xi \in D} f(x; \xi), \tag{1}$$

where $f(x;\xi)$ represents the loss function for a single data point ξ. This optimization problem is solved using iterative methods and we use $(\cdot)_t$ to denote the variables in iteration t. We denote the lower bound as $f^* = \inf f(x)$.

Differential Privacy: Differential Privacy (DP) (Dwork et al., 2014) is used as a standard privacy technique to protect sensitive information within a dataset D. DP is formally defined as follows:

Definition 1 ((ϵ, δ)-Differential Privacy)**.** *A randomized algorithm $\mathcal{M}$ satisfies (ϵ, δ)-DP if for any two adjacent datasets D and D' that differ by a single element, and for any subset of outputs $S \subseteq Range(\mathcal{M})$, it holds that:*

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in S] + \delta,$$

where ϵ and δ denote the privacy budget and the failure factor, respectively. To achieve DP, Gaussian noise (Balle et al., 2018) can be added based on the ℓ_2 sensitivity of the function.

Definition 2 (ℓ_2-Sensitivity)**.** *For a function $\mathcal{A} : \mathcal{D} \to \mathbb{R}^d$, ℓ_2-sensitivity $\Delta_{\mathcal{A}}$ is the maximum possible change in output when applied to any pair of adjacent datasets D and D':*

$$\Delta_{\mathcal{A}} = \sup_{D, D'} \|\mathcal{A}(D) - \mathcal{A}(D')\|.$$

Definition 3 (Gaussian Mechanism)**.** *Suppose an algorithm $\mathcal{A} : \mathcal{D} \to \mathbb{R}^d$ has ℓ_2 sensitivity $\Delta_{\mathcal{A}}$. The Gaussian mechanism $\mathcal{M}(x) = \mathcal{A}(x) + w$, with $w \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, is (ϵ, δ)-DP if the noise σ and privacy parameters satisfy:*

$$\Phi\left(\frac{\Delta_{\mathcal{A}}}{2\sigma} - \frac{\epsilon\sigma}{\Delta_{\mathcal{A}}}\right) - e^{\epsilon}\Phi\left(-\frac{\Delta_{\mathcal{A}}}{2\sigma} - \frac{\epsilon\sigma}{\Delta_{\mathcal{A}}}\right) \leq \delta,$$

where $\Phi(t) = Pr[\mathcal{N}(0,1) \leq t]$ is the cumulative density function of the standard normal distribution.

Differentially Private Stochastic Gradient Descent (DPSGD): DPSGD is a DP variant of stochastic gradient descent. In each training iteration, the algorithm randomly samples a mini-batch of training data and computes per-sample gradients $g_{t,\xi} = \nabla f(x_t; \xi)$. It then clips $g_{t,\xi}$ to bound their ℓ_2-norm (which controls

the global sensitivity), aggregates and adds calibrated Gaussian noise to the aggregated gradients. Finally, DPSGD updates the model parameters using the privatized gradient:

$$\bar{g}_t = \frac{1}{B}\sum_{\xi\in\mathcal{B}_t}\left(g_{t,\xi}\cdot\min\{1,\frac{C}{\|g_{t,\xi}\|}\}\right) + w_t \tag{2}$$

$$x_{t+1} = x_t - \eta\bar{g}_t \tag{3}$$

The privacy guarantees of DPSGD are primarily established through two key techniques. (1) Privacy Amplification by Subsampling (Balle et al., 2018; 2020): By including each data point in a minibatch with a certain probability, the influence of any single individual's data on the model updates is reduced, thereby decreasing the overall privacy loss. (2) Moments Accountant (Abadi et al., 2016): By maintaining tight bounds on the moments of the privacy loss, this technique provides a more accurate analysis of the total privacy cost over multiple iterations.

Motivation: Among the techniques proposed to reduce clipping bias in DPSGD, the Inner-Outer Momentum (IOM) method Xiao et al. (2023) is considered state-of-the-art. IOM smooths per-sample gradients prior to clipping via inner momentum mechanism, and applies outer momentum to reduce variance in the batch-averaged gradients. The key improvement of IOM is its inner momentum mechanism, which maintains an exponential moving average of historical gradients for each sample ξ before clipping:

$$g_{t,\xi} = \sum_{l=t-k}^{t}\gamma_0^{t-l}\nabla f(x_l;\xi).$$

This smoother $g_{t,\xi}$ is then clipped, aggregated across the batch, and perturbed with calibrated noise to ensure differential privacy before being used for the model update. Although IOM is empirically effective, its design leads to a difficult trade-off between utility and efficiency. Its main limitations are two fold.

(1) **High Complexity:** To compute the inner momentum, the algorithm must store the model parameters from the past k iterations and recompute the gradient for each sample in the batch using these historical model parameters, resulting in significant computational and storage overhead.

(2) **Hyperparameter Tuning:** The method introduces several highly sensitive and independent hyperparameters, including the inner and outer momentum coefficients γ_0, γ_1, and the momentum length k. Under differential privacy, this challenge is amplified as tuning these hyperparameters itself can leak some sensitive information. Consequently, optimal tuning of these hyperparameters requires either the use of a public dataset, which is often unavailable, or spending a part of the privacy budget on pre-training, which leaves less privacy budget for the model training.

4 DP Momentum-based Bias Correction

To address these limitations, we propose a novel algorithm: Differentially Private Momentum-based Bias Correction (DP-MBC). Algorithm 1 outlines the main steps in our proposed method.

Pre-clipping Gradient Correction (Lines 1-4). The core innovation of DP-MBC lies in constructing a corrected per-sample gradient vector $g_{t,\xi}$ for each sample $\xi \in \mathcal{B}_t$ prior to clipping. Assuming x_t and x_{t-1} represent the model parameters in current (t) and previous ($t-1$) iterations, respectively, the corrected gradient is calculated as:

$$g_{t,\xi} = \nabla f(x_t;\xi) - \alpha\nabla f(x_{t-1};\xi) + \alpha m_{t-1}. \tag{4}$$

Here, m_{t-1} is the global momentum from the previous iteration (line 8), serving as a stable and privatized historical estimation of the true gradient. The term $(\nabla f(x_t;\xi) - \alpha\nabla f(x_{t-1};\xi))$ acts as a correction factor, capturing the change in the gradient between consecutive iterations. By combining these two components, $g_{t,\xi}$ becomes a low-variance estimator that more closely aligns with the true gradient direction.

Algorithm 1 Differentially Private Momentum-based Bias Correction (DP-MBC)

Input: Dataset D, initial model parameters x_0, batch size B, iteration number T, learning rate η, momentum factor α,
clipping threshold C
Output: Final model parameters x_T

1: **for** $t = 0$ to $T-1$ **do**
2: Sample minibatch $\mathcal{B}_t$ of size B from D
3: **for** $\xi \in \mathcal{B}_t$ **do**
4: $g_{t,\xi} = \nabla f(x_t;\xi) - \alpha \nabla f(x_{t-1};\xi) + \alpha m_{t-1}$
5: $\tilde{g}_{t,\xi} = g_{t,\xi} \cdot \min\left(1, \frac{C}{\|g_{t,\xi}\|}\right)$
6: **end for**
7: $\bar{g}_t = \frac{1}{B}\sum_{\xi \in \mathcal{B}_t} \tilde{g}_{t,\xi} + w_t$, where $w_t \sim \mathcal{N}(0, \sigma^2 I_d)$
8: $m_t = \alpha m_{t-1} + (1-\alpha)\bar{g}_t$
9: $x_{t+1} = OptimizerUpdate(x_t, \eta, \bar{g}_t)$
10: **end for**
11: **return** x_T

Per-sample Clipping (Line 5). To enforce the sensitivity bound required by the Gaussian mechanism (Definition 3), the corrected gradient $g_{t,\xi}$ is clipped to a predefined threshold C:

$$\tilde{g}_{t,\xi} = g_{t,\xi} \cdot \min\left(1, \frac{C}{\|g_{t,\xi}\|}\right). \tag{5}$$

This operation ensures that, for any sample ξ, its contribution to the update is bounded such that $\|\tilde{g}_{t,\xi}\| \leq C$.

Noisy Aggregation (Line 7). The clipped gradients are then aggregated by averaging over the minibatch $\mathcal{B}_t$, and calibrated Gaussian noise w_t is added to provide the formal (ϵ, δ)-DP guarantee. This generates the final noisy batch gradient $\bar{g}_t$:

$$\bar{g}_t = \frac{1}{B}\sum_{\xi \in \mathcal{B}_t} \tilde{g}_{t,\xi} + w_t, \quad \text{where } w_t \sim \mathcal{N}(0, \sigma^2 I_d). \tag{6}$$

The variance of DP noise σ^2 is scaled according to Definition 3.

Momentum State and Model Update (Lines 8-9). The final stage of each training iteration separates the momentum state update from the model parameter update, which is a key feature of our design. In this stage, first, the global momentum m_t is updated using an exponential moving average of the noisy gradient $\bar{g}_t$:

$$m_t = \alpha m_{t-1} + (1-\alpha)\bar{g}_t. \tag{7}$$

Since $\bar{g}_t$ is protected by DP, the momentum term m_t also inherits this privacy guarantee due to the post-processing property. The momentum m_t is then passed to next iteration, $t+1$ to update the gradient correction step (line 4). This mechanism allows DP-MBC to be composed with any standard optimizer such as SGD with momentum (Sutskever et al., 2013) or Adam (Kingma & Ba, 2015). The corrected and privatized gradient $\bar{g}_t$ is then passed to a standard optimizer for the model update:

$$x_{t+1} = OptimizerUpdate(x_t, \eta, \bar{g}_t). \tag{8}$$

Comparison with State-Of-The-Art Methods

Comparison to IOM. As detailed above, DP-MBC shares IOM's high-level goal of mitigating bias through pre-clipping correction, but does so with a more efficient and robust mechanism avoiding IOM's practical drawbacks in complexity and hyperparameter tuning. Specifically,

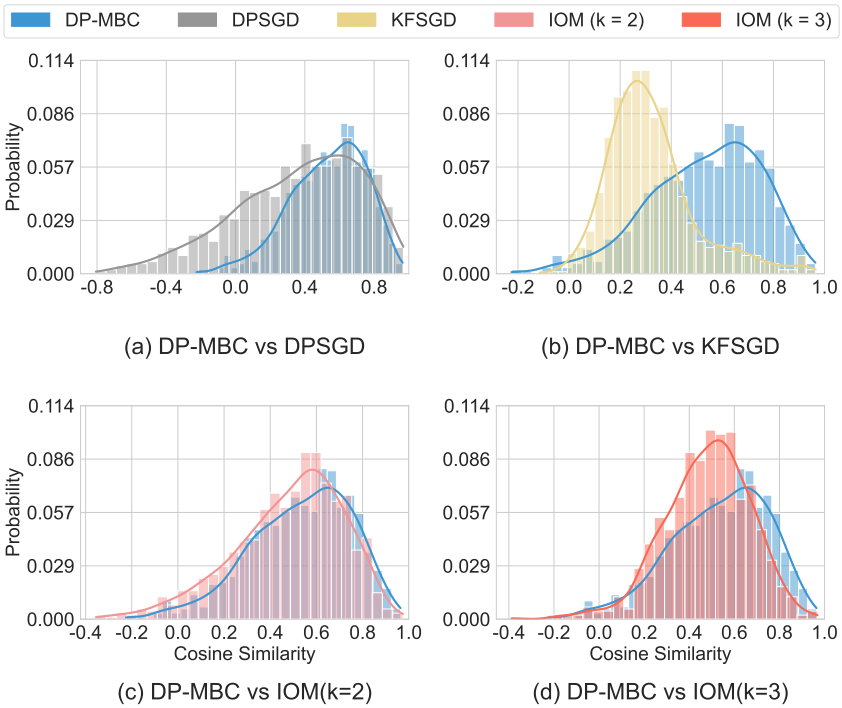

Figure 1: Comparison of cosine similarity probability density functions between DP-MBC and the methods DPSGD, KFSGD, IOM ($k = 2$), and IOM ($k = 3$) on Fashion-MNIST. All experiments run 5 times recording the cosine similarity every 10 iterations.

- **Solving High Complexity:** Unlike IOM, which requires storing up to k past model states, our method relies only on the immediately preceding iterate x_{t-1} and its corresponding gradient. This significantly reduces both memory and computational overhead.
- **Reducing Hyperparameter Tuning Space:** Our method eliminates IOM's complex hyperparameters (γ_0, γ_1, k), relying only on a single momentum factor α, which simplifies the tuning process.

Comparison to KFSGD. As we mentioned before, KFSGD (Zhang et al., 2025) is another recent method that aims to refine gradient estimate. While both KFSGD and DP-MBC methods construct a corrected gradient, our approach differs fundamentally in how the correction is applied.

KFSGD aims to reduce the effect of DP noise while our method aims to reduce the clipping bias. KFSGD is inspired by signal processing, where the gradient information is treated as a discrete signal, and a better filter is found to reduce the effect of DP noise. However, its theoretical analysis assumes that the clipping threshold is never activated, effectively ignoring the impact of clipping bias. In contrast, DP-MBC is explicitly designed to operate effectively without rely on such a strong assumption.

KFSGD computes a privatized gradient as a linear combination of gradients at the current point x_t and an explored point $x_t + \gamma d_{t-1}$, with the weighting controlled by κ. However, it introduces two additional hyperparameters κ and γ, whose optimal values can be difficult to determine across different datasets. In contrast, DP-MBC uses a global DP-protected momentum m_{t-1} as a stable historical estimation to correct each per-sample gradient.

Cosine Similarity Analysis of Gradient Correction. To illustrate the effectiveness of our proposed momentum-based bias correction, we measured the cosine similarity between the clipped batch-averaged gradient and the true batch-averaged gradient. A higher cosine similarity indicates closer alignment between the two gradients. As shown in Figure 1, we compared the cosine similarity of our method against DPSGD,

KFSGD, and IOM on the Fashion-MNIST dataset. Our DP-MBC yields a higher proportion of gradients with larger cosine similarity compared to the other techniques, indicating that it remains statistically closer to the true batch-averaged gradient. The detailed experimental setting is described in Section 6.1.

5 Theoretical Analysis

We now present the privacy and convergence analysis of Algorithm 1. We first introduce the assumptions we used in our proofs below.

Assumption 1 (L-Smoothness)**.** *The loss function $f(x;\xi)$ is a L-smooth function, which means that for any $x, y \in \mathbb{R}^d$:*

$$\|\nabla f(x;\xi) - \nabla f(y;\xi)\| \leq L\|x - y\|.$$

Assumption 2 (Bounded Variance)**.** *The variance of the stochastic gradients is bounded. For any $x \in \mathbb{R}^d$, we have:*

$$\mathbb{E}_\xi[\|\nabla f(x;\xi) - \nabla f(x)\|^2] \leq r^2.$$

Assumption 1 and Assumption 2 are commonly used in non-convex optimization (Reddi et al., 2018; Cutkosky & Orabona, 2019; Levy et al., 2021). Assumption 1 also implies that the empirical risk function f is L-smooth, that is,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

5.1 Privacy Guarantee

The following theorem formalizes the (ϵ, δ)-DP guarantee of Algorithm 1:

Theorem 1 (Differential Privacy Guarantee of DP-MBC)**.** *There exist constants c_1 and c_2 such that, given the sampling probability $q = B/N$ and the number of iterations T, for any $\epsilon \leq c_1 q^2 T$ and $\delta > 0$, Algorithm 1 is (ϵ, δ)-DP if the noise scale σ is chosen such that:*

$$\sigma \geq c_2 \frac{q\sqrt{T\log(1/\delta)}}{\epsilon}$$

Proof. The privacy guarantee follows from the standard analysis of the subsampled Gaussian mechanism. At each iteration t, the contribution of any single sample ξ is the vector $g_{t,\xi}$. The clipping operation in line 5 of Algorithm 1 ensures that the ℓ_2-norm of each per-sample contribution is bounded by C, i.e. $\|\tilde{g}_{t,\xi}\| = g_{t,\xi} \cdot \min\left(1, \frac{C}{\|g_{t,\xi}\|}\right) \leq C$. The subsequent updates to m_t and x_{t+1} in lines 8 and 9 are postprocessing steps based on the private output $\bar{g}_t$, and therefore do not consume additional privacy budget.

The overall (ϵ, δ)-DP guarantee is obtained by composing the privacy loss over T iterations using the Moments Accountant (Abadi et al., 2016), which accounts for privacy amplification due to the subsampling at rate $q = B/N$. This analysis is equivalent to that of vanilla DPSGD, and leads to the same requirement for the noise scale σ. □

5.2 Convergence Guarantee

We now formalize Theorem 2 to bound the expected gradient norm with the number of iterations T as follows:

Theorem 2. *Assume Assumptions 1 and 2 hold. Suppose $\eta \leq min\left(\sqrt{\frac{d\sigma^2}{3\alpha^2 L^2(C^2+d\sigma^2)}}, \frac{(1-\alpha)r}{L\sqrt{2(C^2+d\sigma^2)}}\right)$ and $\alpha \geq max\left(1 - \sqrt[4]{\frac{C^2}{54r^2}}, \frac{48\beta+1-\sqrt{96\beta+1}}{48\beta}\right)$, and run Algorithm 1 for T iterations, then $\mathbb{E}[\|\nabla f(x_t)\|]$ is bounded by:*

$$\mathcal{O}\left(\sqrt{\frac{f(x_0) + \beta\|\nabla f(x_0)\| - f^*}{(1-\alpha)\eta T}} + \frac{f(x_0) - f^*}{\eta C T} + \sqrt{\frac{\beta + L\eta^2}{(1-\alpha)\eta}(C^2 + d\sigma^2)} + \frac{L\eta(C^2 + d\sigma^2)}{C} + \sqrt{(1-\alpha)}r\right),$$

where $\beta \geq \frac{110\eta\alpha^2 + \eta\alpha}{1 - 3\alpha^2 - 18\alpha^2(1-\alpha)^2}$.

Table 1: Test accuracy (%) comparison on various image classification datasets. We report the mean with standard deviation.

Dataset	Model	(ϵ, δ)	DPSGD	IOM	KFSGD	DP-MBC
MNIST	4-layer CNN	(3, 1e-5)	98.07 ± 0.10	**98.58 ± 0.08**	98.44 ± 0.13	98.57 ± 0.04
Fashion-MNIST	4-layer CNN	(3, 1e-5)	86.45 ± 0.42	87.49 ± 0.29	87.36 ± 0.32	**87.78 ± 0.30**
CIFAR-10	SimCLRv2	(2, 1e-5)	92.62 ± 0.03	93.40 ± 0.07	93.23 ± 0.14	**93.45 ± 0.19**
ImageNette	ResNet9	(8, 1e-4)	60.91 ± 0.71	61.91 ± 0.71	66.99 ± 0.42	**67.69 ± 1.11**
CelebA Multi-label	ResNet9	(8, 5e-6)	88.49 ± 0.13	88.80 ± 0.05	88.96 ± 0.03	**89.38 ± 0.09**

Our bound identifies three types of error:

- **Optimization Error**: The first two terms, $\sqrt{\frac{f(x_0)+\beta\|\nabla f(x_0)\|-f^*}{(1-\alpha)\eta T}}$ and $\frac{f(x_0)-f^*}{\eta C T}$, constitute the optimization error, which vanishes as the number of iterations T increases.
- **Clipping Threshold and DP Noise**: The third and fourth terms, $\sqrt{\frac{\beta+L\eta^2}{(1-\alpha)\eta}(C^2+d\sigma^2)}$ and $\frac{L\eta(C^2+d\sigma^2)}{C}$, are influenced by the clipping threshold C and DP noise scale σ, which could be eliminated by choosing a small stepsize η.
- **Constant Bias**: The final term, $\sqrt{(1-\alpha)}r$, is a constant bias that is not affected by the choice of T or η.

A key advantage of DP-MBC is the reduction of the constant bias term. While the bias in vanilla DP-SGD is $\mathcal{O}(r)$ (Xiao et al., 2023; Koloskova et al., 2023), our approach reduces it to $\mathcal{O}(\sqrt{(1-\alpha)}r)$. By choosing α close to 1, this bias term can be significantly reduced, enabling Algorithm 1 to converge closer to the true minimum. Importantly, this improvement does not rely on strong assumptions made in previous work, such as independent sampling variance between iterations (Xiao et al., 2023) or inactive gradient clipping (Zhang et al., 2025). The core of our proof is outlined below.

Proof Sketch. We first introduce two key lemmas in our approach. For simplicity, we use $\Delta_{g,t}$ and $\Delta_{m,t}$ to represent $(g_{t,\xi} - \nabla f(x_t))$ and $(m_t - \nabla f(x_t))$, respectively.

Lemma 1 provides a bound on the sampling error of our corrected gradient estimator. It demonstrates that the sampling error is recursively dependent on the momentum sampling error of the previous iteration step and the inherent stochastic variance.

Lemma 1. *Suppose* $\eta \leq \frac{(1-\alpha)r}{L\sqrt{2(C^2+d\sigma^2)}}$, *the sampling variance of the corrected gradient estimator* $g_{t,\xi}$ *is bounded as:*

$$\mathbb{E}[\|\Delta_{g,t}\|^2] \leq \alpha^2\|\Delta_{m,t-1}\|^2 + 3(1-\alpha)^2 r^2. \tag{9}$$

Next, Lemma 2 establishes a recursive bound for the momentum sampling error. It shows that the error of the momentum term of the current iteration t is depends on the momentum term of the previous iteration $(t-1)$, the quality of the current gradient estimate (as bounded by Lemma 1), the magnitude of the true gradient, and the DP noise.

Lemma 2. *Suppose* $\eta \leq \sqrt{\frac{d\sigma^2}{3\alpha^2 L^2(C^2+d\sigma^2)}}$, *the momentum error term* $\Delta_{m,t}$ *is bounded by:*

$$\mathcal{O}(\alpha^2\|\Delta_{m,t-1}\|^2 + (1-\alpha)^2\mathbb{E}[\|\Delta_{g,t}\|^2] + (1-\alpha)^2\|\nabla f(x_t)\|^2 + d\sigma^2). \tag{10}$$

Table 2: Test accuracy (%) comparison on the NLP datasets. We report the mean with standard deviation for $\epsilon = 3$ and $\epsilon = 8$.

Method	$\epsilon = 3$				$\epsilon = 8$			
	MNLI	QQP	QNLI	SST-2	MNLI	QQP	QNLI	SST-2
DPSGD	79.25 ± 0.49	84.46 ± 0.15	83.40 ± 2.52	88.19 ± 0.69	80.43 ± 0.03	85.58 ± 0.27	85.79 ± 0.42	89.38 ± 0.33
IOM	80.41 ± 0.97	85.74 ± 0.55	84.57 ± 1.71	88.02 ± 1.42	81.49 ± 0.63	85.68 ± 0.45	85.86 ± 0.94	89.22 ± 0.76
KFSGD	80.99 ± 0.27	86.17 ± 0.18	86.58 ± 0.58	**90.06 ± 0.76**	81.73 ± 0.36	86.88 ± 0.34	87.24 ± 0.48	90.56 ± 0.48
DP-MBC	**81.28 ± 0.30**	**86.21 ± 0.15**	**86.95 ± 0.91**	90.02 ± 0.78	**82.10 ± 0.41**	**87.00 ± 0.16**	**87.77 ± 0.21**	**90.71 ± 0.51**

Our main convergence proof begins with the property of L-smoothness, which is:

$$\mathbb{E}[f(x_{t+1})] - f(x_t) \leq -\eta\mathbb{E}[\langle \nabla f(x_t), \bar{g}_t \rangle] + \frac{L\eta^2}{2}\mathbb{E}[\|\bar{g}_t\|^2].$$

The core challenge lies in lower-bounding the inner product term $\mathbb{E}[\langle \nabla f(x_t), \bar{g}_t \rangle]$ which is complicated by the non-linear bias introduced by clipping and the added DP noise.

Our analysis is motivated by a key observation from the work by Koloskova et. al (Koloskova et al., 2023): the impact of the bias depends on the magnitude of the true gradient $\|\nabla f(x_t)\|$. Specifically, when $\|\nabla f(x_t)\|$ is sufficiently large, it can offset the sampling variance, allowing the directions of the clipped and unclipped batch averaged gradient to be the same. However, when $\|\nabla f(x_t)\|$ is small, a more detailed analysis of the bias term becomes necessary. This insight motivates our two-case proof strategy, based on whether $\|\nabla f(x_t)\|$ is above or below a threshold $6\mathbb{E}[\|\Delta_{g,t}\|]$.

In the first case, when the gradient norm is relatively large (i.e. $|\nabla f(x_t)| \geq 6\mathbb{E}[|\Delta_{g,t}|]$), our analysis aims to show that the expected update direction remains unchanged despite noise and clipping. We introduce two events A_t and its complement event A_t^c, where

$$A_t = \{\xi \mid \|g_{t,\xi} - \nabla f(x_t)\| \leq 3\sqrt{\mathbb{E}[\|\Delta_{g,t}\|]}\}.$$

Applying the Markov's inequality, we can establish that A_t occurs with a probability of at least (8/9). Conditioned on A_t and A_t^c, we further show that the clipping operation has a limited impact on the gradient's direction. As a result, the expected inner product between the true gradient $\|\nabla f(x_t)\|$ and the clipped batch-averaged gradient $\bar{g}_t$ admits a positive bound:

$$\mathbb{E}[\langle \nabla f(x_t), \bar{g}_t \rangle] \geq \mathcal{O}(C\|\nabla f(x_t)\|).$$

This ensures a per-step reduction in the objective function, leading to convergence rates that decrease with the number of iterations T without any bias.

For the second case $\|\nabla f(x_t)\| < 6\mathbb{E}[\|\Delta_{g,t}\|]$, we directly decompose $\mathbb{E}[\langle \nabla f(x_t), \bar{g}_t \rangle]$ and derive the upper bound as follow:

$$\mathbb{E}[\langle \nabla f(x_t), \bar{g}_t \rangle] \leq \|\nabla f(x_t)\|^2 + \alpha\langle \nabla f(x_t), \Delta_{m,t-1} \rangle.$$

This decomposition isolates the desired descent component ($\|\nabla f(x_t)\|^2$). However, this reveals a new challenge, a cross-term that couples the momentum error from the previous step, $\Delta_{m,t-1}$. To handle this recursive dependency, we employ a Lyapunov function, defined as:

$$\mathcal{L}_t = f(x_t) + \beta\|\Delta_{m,t}\|^2.$$

We then focus on the one-step change of $\mathbb{E}[\mathcal{L}_{t+1} - \mathcal{L}_t]$. By applying Lemmas 1 and 2, and appropriately choosing β, α, and η, we derive the key descent inequality for the Lyapunov function as follows:

$$\mathbb{E}[\mathcal{L}_{t+1}] - \mathbb{E}[\mathcal{L}_t] \leq -(1-\alpha)\eta\mathbb{E}[\|\nabla f(x_t)\|^2] + \mathcal{O}(\eta(1-\alpha)^2 r^2 + \eta^2(C^2 + d\sigma^2)).$$

Finally, iterating this descent inequality over T iterations and combining the results from both cases yields the convergence guarantee stated in Theorem 2. Detailed derivations are provided in the appendix. □

6 Experimental Evaluation

We empirically evaluate the performance of DP-MBC, against several state-of-the-art clipping bias reduction baselines. To demonstrate its effectiveness, our evaluation spans different tasks across two application domains, image classification and natural language processing (NLP), under various datasets and different privacy budgets.

6.1 Experimental Setting

Dataset. We evaluate our method on five image classification datasets, including MNIST (Deng, 2012), Fashion-MNIST (Xiao et al., 2017), the more complex CIFAR-10 color image dataset, ImageNette (a 10-class subset of ImageNet (Deng et al., 2009) with ten labels), and a large-scale CelebA dataset (Liu et al., 2015) for a multi-label face attribute classification task.

For NLP tasks, we use four representative tasks from the GLUE benchmark (Wang et al., 2018). These datasets are: MNLI for natural language inference, QQP for paraphrase detection, QNLI for question answering, and SST-2 for sentiment analysis.

Baselines. We compare our proposed method against three baseline methods. We use vanilla DPSGD (Abadi et al., 2016), the standard algorithm for private deep learning via gradient clipping and noise addition. The second one is IOM (Xiao et al., 2023), a state-of-the-art clipping bias reduction approach that uses two momentums to improve the utility of DPSGD. The last one is KFSGD (Zhang et al., 2025), a method that uses the Kalman filter to obtain a better gradient estimation for DP training.

Algorithms. As optimization algorithms, we utilize SGD with momentum (Sutskever et al., 2013) for MNIST, Fashion-MNIST, and CIFAR-10, and Adam (Kingma & Ba, 2015) for all other datasets. It is worth noting that DP-MBC is orthogonal to the choice of optimizer, and can in principle be combined with a wide range of optimization algorithms.

Models and Hyperparameters. We use the model architectures and hyperparameter setting as in prior work (Bu et al., 2024) to ensure a fair comparison. Specifically for MNIST and Fashion-MNIST, we use a simple 4-layer CNN. For CIFAR-10, we use a pretrained SimCLRv2 model (Chen et al., 2020a) as a strong feature extractor. For the ImageNette and CelebA dataset, we train a ResNet9 model. For all tasks in the GLUE benchmark, we fine-tune a pre-trained RoBERTa-base model (Liu et al., 2019). We set the hyperparameter α to 0.9, rather than tuning it separately for each dataset. We also analyze how α interacts with other hyperparameters to affect model performance. For the additional hyperparameters introduced by IOM and KFSGD, we follow the values recommended by the original authors.

Hardware and software information All methods are implemented using Python (version 3) and all experiments are run on an i7-7800X CPU with a NVIDIA RTX A6000 GPU (48 GB). All experiments are repeated five times, with the mean and standard deviation reported.

6.2 Performance Comparison with Baselines

6.2.1 Image Classification Results.

As shown in Table 1, DP-MBC consistently outperforms all baselines across most datasets under different privacy settings. In particular, on the ImageNette dataset, DP-MBC achieves an accuracy of 67.69%, outperforming the vanilla DPSGD by around 7% and surpassing the KFSGD method by around 0.7%. Meanwhile, DP-MBC consistently outperforms all baseline methods on other datasets, achieving 87.78% on Fashion-MNIST, 93.45% on CIFAR-10, and 89.38% on CelebA. For the MNIST dataset, our method achieves an accuracy of 98.57%, which is comparable to the best performing baseline, IOM.

6.2.2 NLP Results.

Similar to the image classification tasks, we also achieve the better performance on the GLUE benchmark, as shown in Table 2. We evaluated all methods in two privacy budgets: $\epsilon = 3$ and $\epsilon = 8$. Under $\epsilon = 3$, DP-MBC

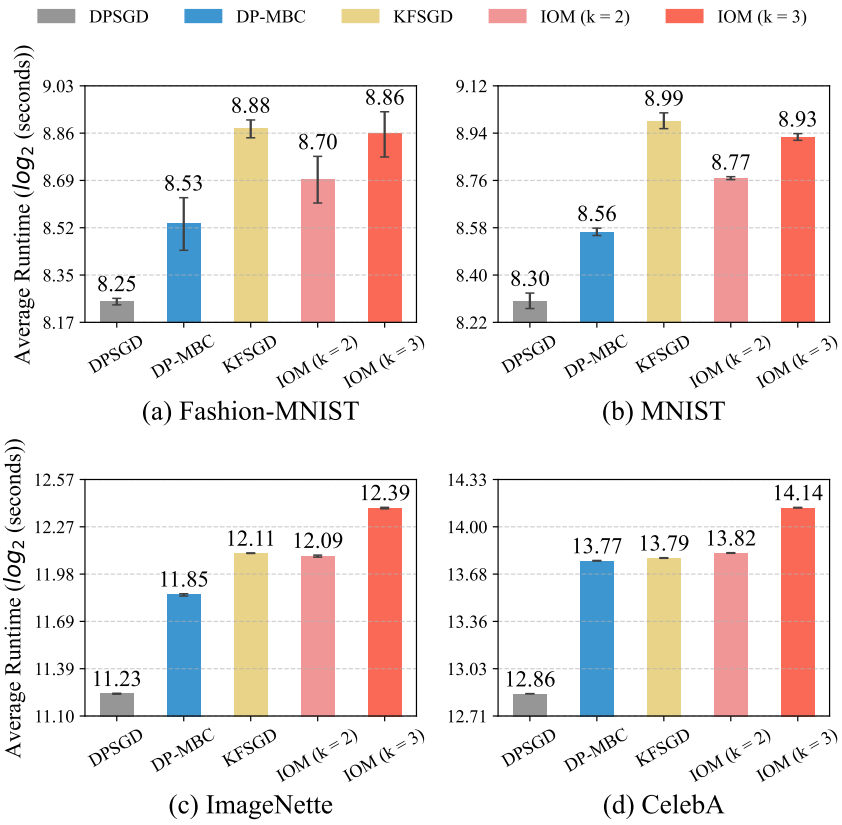

Figure 2: Average runtime (seconds, log_2 scale) comparison on various image classification tasks.

outperforms all baselines in three of four tasks (MNLI, QQP, and QNLI), highlighting its robustness in high-privacy regimes. The only exception is the SST-2 task, where KFSGD exhibits a marginal advantage over DP-MBC. When the privacy budget is relaxed to $\epsilon = 8$, performance improves for all methods. However, DP-MBC still extends its superiority, achieving the best results on all four GLUE tasks. For example, on QNLI, the test accuracy of DP-MBC reaches 87.77%, which is 0.5% higher than the strongest baseline, KFSGD.

6.2.3 Statistical Significance Analysis.

To assess the statistical significance of our improvements, we apply the Wilcoxon signed-rank test (Wilcoxon, 1992) across 13 dataset–budget combinations each with 5 independent runs, resulting in 65 paired comparisons per method. The p-values obtained when comparing DP-MBC against DPSGD (1.86×10^{-9}), IOM (9.19×10^{-6}), and KFSGD (3.96×10^{-5}) are all far below the threshold of 0.05. These results provide strong evidence that DP-MBC achieves statistically significant performance improvements over the baselines.

6.3 Runtime Analysis

Figure 2 shows the average runtime of 5 independent runs with standard deviations for DP-MBC and baseline methods across four image classification datasets. While vanilla DPSGD remains the fastest due to its simplicity, it suffers from significant clipping bias and poor utility. KFSGD and IOM, on the other hand, have substantially increased runtime because of their complex filtering and excess correction mechanisms. In contrast, DP-MBC achieves competitive computational efficiency, requiring less runtime than KFSGD and IOM. This scalability is critical for large-scale learning, where models must handle massive datasets

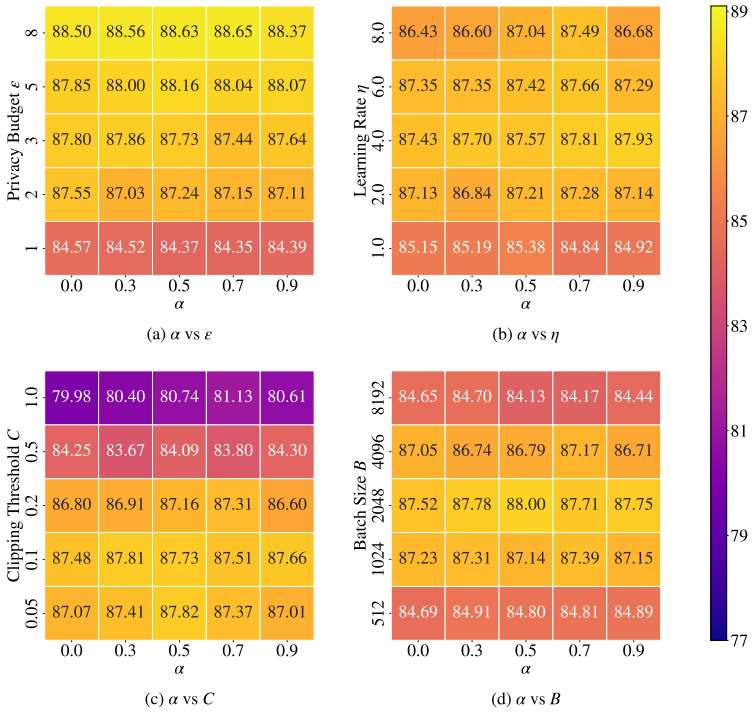

Figure 3: Hyperparameter analysis of DP-MBC on Fashion-MNIST. The heatmaps show test accuracy (%) of different momentum factors α versus the privacy budget ϵ, the learning rate η, the clipping threshold C, and the batch size B.

under resource constraints, highlighting the applicability of DP-MBC for practical deployment where strong privacy and efficiency are required.

6.4 Hyperparameter Analysis

We conduct experiments to investigate the impact of the momentum factor α, and its interaction with other hyperparameters. Figure 3 presents four heatmaps showing the test accuracy on the Fashion-MNIST dataset for different α values against the privacy budget ϵ, the learning rate η, the clipping threshold C, and the batch size B. We also validate the stability of α across different pairs of learning rate η and batch size B. Figure 4 shows the stability analysis of α under different pairs of η and B.

6.4.1 Interaction with Privacy Budget ϵ

The optimal momentum factor α is positively correlated with the privacy budget ϵ. As illustrated in Figure 3(a), in the strong privacy setting ($\epsilon = 1$ *or* 2), the high levels of injected noise render historical gradients unreliable. Consequently, accumulating this noisy history with a high α is detrimental. Conversely, as we relax the privacy constraint ($\epsilon \geq 3$), we could always find an $\alpha > 0$ that outperforms the test accuracy when $\alpha = 0$. Here, the benefit of using momentum factor becomes clear as a larger α effectively smooths out the gradient and accelerates convergence.

6.4.2 Interaction with Learning Rate η

Figure 3(b) shows that the momentum factor α is benefit for choosing the learning rate due to its stabilizing effect. By using the momentum-based bias correction, DP-MBC provides a more consistent optimization direction, allowing it to converge more effectively, especially at higher learning rates. The middle heatmap in Figure 3 illustrates that for large learning rates ($\eta \geq 4$), $\alpha > 0$ consistently improves accuracy over $\alpha = 0$, reaching a peak test accuracy of 87.93% at $\eta = 4$ and $\alpha = 0.9$.

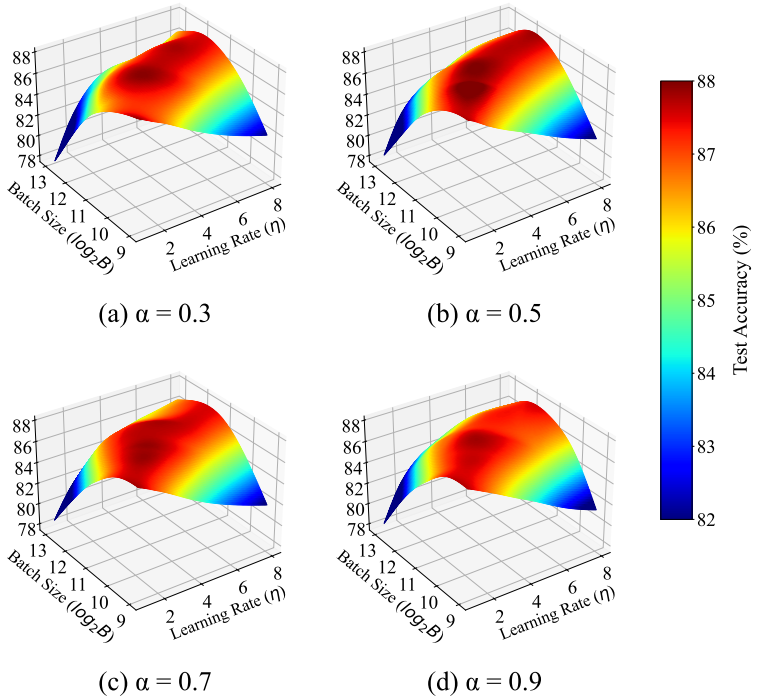

Figure 4: Stability analysis of DP-MBC with respect to the momentum factor α with learning rates η and batch sizes B.

6.4.3 Interaction with Clipping Threshold C

Prior work has shown that selecting a small clipping threshold is crucial for achieving high utility in DP training (Li et al., 2022; Bu et al., 2024). Thus, DP-MBC demonstrates particularly high test accuracy with small clipping thresholds. This is because a small clipping threshold C can introduce significant bias by distorting most gradient directions, where DP-MBC is designed to mitigate this bias. As shown in Figure 3(c), even with a relatively small clipping threshold of $C = 0.05$ and $\alpha = 0.5$ enables the model to reach a peak performance of 87.82%.

6.4.4 Interaction with Batch Size B

As illustrated in Figure 3(d), when the batch size is small, non-zero α (> 0) values consistently yield higher accuracy (e.g., $\alpha = 0.5$ improves accuracy to around 88% when $B = 2048$). This is because the introduction of α mitigates clipping bias, which grows with sampling variance and is more significant under small batch sizes. As batch size B increases, sampling variance diminishes and the relative benefit of α decreases. Moreover, in real-world scenarios, very large batch sizes are often infeasible or prohibitively memory-intensive, further underscoring the practical effectiveness of our proposed approach.

6.4.5 Stability Analysis of α

Figure 4 demonstrates the stability of the momentum factor α. For any fixed configuration of learning rate η and batch size B, the test accuracy shows a minor variation between different non-zero values of α. This indicates that the performance of DP-MBC is robust to the choice of α, thereby simplifying hyperparameter selection in practice.

Table 3: Empirical privacy evaluation via membership inference attacks. We report the True Positive Rate (TPR) at a fixed 1% False Positive Rate (FPR), denoted as $TPR@1\%FPR$ (Carlini et al., 2022).

Dataset	Method	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 8$	Non-private
MNIST	DPSGD	2.64%	3.01%	3.42%	4.23%	7.38%	14.24%
	DP-MBC	2.62%	2.95%	3.31%	4.26%	7.38%	
Fashion-MNIST	DPSGD	2.68%	2.97%	3.22%	3.82%	6.62%	7.69%
	DP-MBC	2.71%	2.95%	3.25%	3.83%	6.59%	

6.5 Empirical Privacy Evaluation

To empirically validate the privacy guarantee of DP-MBC, we conduct a membership inference attack (MIA) using the Likelihood Ratio Attack (LiRA) (Carlini et al., 2022). We adopt the True Positive Rate (TPR) at a 1% False Positive Rate (FPR), as defined in (Carlini et al., 2022), as a metric and evaluate on MNIST and Fashion-MNIST using a ResNet-9 model. A lower value of this metric indicates that the model is more resilient against MIA attacks, effectively protecting the presence of training samples from an adversary.

As illustrated in Table 3, the privacy resilience of DP-MBC is comparable to that of the vanilla DPSGD across various privacy budgets ϵ. Both DPSGD and DP-MBC significantly suppress the privacy risk compared to the non-private optimizer, which exhibits a leakage of 14.24% on MNIST and 7.69% on Fashion-MNIST. These results show that DP-MBC effectively reduces membership inference risks, achieving a better privacy-utility trade-off.

7 Conclusion

In this paper, we introduce DP-MBC, a novel method to improve the privacy-utility trade-off in privacy-preserving deep learning. DP-MBC is built on a momentum-based corrected gradient estimator that provably reduces the constant bias term of vanilla DP-SGD. We provide a rigorous convergence guarantee with a formal differential privacy proof. Extensive experiments on image and sentence classification tasks show that our method consistently outperforms the state-of-the-art baselines across various datasets.

In future work, we aim to design an adaptive mechanism to set α and integrate our method with advanced optimizers such as AdamW (Loshchilov & Hutter, 2019). In addition, we will also apply our DP-MBC method to new domains, such as the private training of large language models (LLMs) (Makni et al., 2025) or graph neural networks (GNNs) (Guo et al., 2025).

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *SIGSAC Conference on Computer and Communications Security*, CCS '16, pp. 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31, 2018.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1):1–32, 2020.

Wenxuan Bao, Francesco Pittaluga, Vijay Kumar BG, and Vincent Bindschaedler. Dp-mix: mixup-based data augmentation for differentially private learning. *Advances in Neural Information Processing Systems*, 36:12154–12170, 2023.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *Advances in Neural Information Processing Systems*, 36, 2024.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020a.

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33:13773–13782, 2020b.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Zhiyu Guo, Yang Liu, Xiang Ao, and Qing He. Grasp: Differentially private graph reconstruction defense with structured perturbation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 767–777, 2025.

Tao Huang, Qingyu Huang, Xin Shi, Jiayang Meng, Guolong Zheng, Xu Yang, and Xun Yi. Enhancing DP-SGD through non-monotonous adaptive scaling gradient weight. *CoRR*, abs/2411.03059, 2024. doi: 10.48550/ARXIV.2411.03059. URL `https://doi.org/10.48550/arXiv.2411.03059`.

Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE symposium on security and privacy (SP)*, pp. 299–316. IEEE, 2019.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX security symposium (USENIX security 19)*, pp. 1895–1912, 2019.

Wei Jiang, Gang Li, Yibo Wang, Lijun Zhang, and Tianbao Yang. Multi-block-single-probe variance reduced estimator for coupled compositional optimization. *Advances in Neural Information Processing Systems*, 35:32499–32511, 2022.

Wei Jiang, Jiayu Qin, Lingyu Wu, Changyou Chen, Tianbao Yang, and Lijun Zhang. Learning unnormalized statistical models via compositional optimization. In *International Conference on Machine Learning*, pp. 15105–15124. PMLR, 2023.

Wei Jiang, Sifan Yang, Yibo Wang, and Lijun Zhang. Adaptive variance reduction for stochastic optimization under weaker assumptions. *Advances in Neural Information Processing Systems*, 37:22047–22080, 2024a.

Wei Jiang, Sifan Yang, Wenhao Yang, and Lijun Zhang. Efficient sign-based optimization: Accelerating convergence via variance reduction. *Advances in Neural Information Processing Systems*, 37:33891–33932, 2024b.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Ali Kavis, Stratis Skoulakis, Kimon Antonakopoulos, Leello Tadesse Dadi, and Volkan Cevher. Adaptive stochastic variance reduction for non-convex finite-sum minimization. *Advances in Neural Information Processing Systems*, 35:23524–23538, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pp. 17343–17363, Hawai, 2023. PMLR, PMLR.

Kfir Y. Levy, Ali Kavis, and Volkan Cevher. STORM+: fully adaptive SGD with recursive momentum for nonconvex optimization. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 20571–20582, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/ac10ff1941c540cd87c107330996f4f6-Abstract.html`.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=bVuP3ltATMz`.

Lan Liu, Yi Wang, Gaoyang Liu, Kai Peng, and Chen Wang. Membership inference attacks against machine learning models via prediction sensitivity. *IEEE Transactions on Dependable and Secure Computing*, 20 (3):2341–2347, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Mehdi Makni, Kayhan Behdin, Gabriel Afriat, Zheng Xu, Sergei Vassilvitskii, Natalia Ponomareva, Rahul Mazumder, and Hussein Hazimeh. Sparta: An optimization framework for differentially private sparse fine-tuning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2090–2101, 2025.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pp. 2613–2621. PMLR, 2017.

Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Re-thinking model inversion attacks against deep neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16384–16393, Vancouver, Canada, 2023. IEEE.

Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding gradient clipping in incremental gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 1504–1512. PMLR, 2021.

S Reddi, Manzil Zaheer, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Proceeding of 32nd conference on neural information processing systems (NIPS 2018)*, volume 2, 2018.

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.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. pmlr, 2013.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32, 2019.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pp. 196–202. Springer, 1992.

Tianyu Xia, Shuheng Shen, Su Yao, Xinyi Fu, Ke Xu, Xiaolong Xu, and Xing Fu. Differentially private learning with per-sample adaptive clipping. In *the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10444–10452, Washington, DC, USA, 2023. AAAI.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, pp. 1–6, 2017.

Hanshen Xiao, Zihang Xiang, Di Wang, and Srinivas Devadas. A theory to instruct differentially-private learning via clipping bias reduction. In *IEEE Symposium on Security and Privacy (SP)*, pp. 2170–2189, San Francisco, CA, USA, 2023. IEEE, IEEE.

Depeng Xu, Wei Du, and Xintao Wu. Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1924–1932, 2021.

Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning, ICML 2022*, 2022.

Xinwei Zhang, Zhiqi Bu, Steven Wu, and Mingyi Hong. Differentially private SGD without clipping bias: An error-feedback approach. In *International Conference on Learning Representations (ICLR)*, pp. 1–27, Vienna, Austria, 2024. ICLR. URL `https://openreview.net/forum?id=uFbWHyTlPn`.

Xinwei Zhang, Zhiqi Bu, Borja Balle, Mingyi Hong, Meisam Razaviyayn, and Vahab Mirrokni. Disk: Differentially private optimizer with simplified kalman filter for noise reduction. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=Lfy9q7Icp9`.

Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. Exploiting explanations for model inversion attacks. In *IEEE/CVF international conference on computer vision*, pp. 682–692, Montreal, BC, Canada, 2021. IEEE.

A Proof of Lemma 1

Proof. We start by expanding the term $\mathbb{E}[\|\Delta_{g,t}\|^2]$ according to the definition of $\Delta_{g,t}$:

$$\mathbb{E}[\|\Delta_{g,t}\|^2] = \mathbb{E}[\|g_{t,\xi} - \nabla f(x_t)\|^2]. \tag{11}$$

By utilizing $\mathbb{E}[\|X - c\|^2] = \mathbb{E}[\|X - \mathbb{E}[X]\|^2] + \mathbb{E}[\|\mathbb{E}[X] - c\|^2]$, we have:

$$\mathbb{E}[\|g_{t,\xi} - \nabla f(x_t)\|^2] = \mathbb{E}[\|g_{t,\xi} - \mathbb{E}[g_{t,\xi}]\|^2] + \mathbb{E}[\|\mathbb{E}[g_{t,\xi}] - \nabla f(x_t)\|^2]. \tag{12}$$

We first derive the bound of the first term $\mathbb{E}[\|g_{t,\xi} - \mathbb{E}[g_{t,\xi}]\|^2]$. By using the definition of $g_{t,\xi}$, and $\mathbb{E}[\nabla f(x_t;\xi)] = \nabla f(x_t)$, we have:

$$\mathbb{E}[\|g_{t,\xi} - \mathbb{E}[g_{t,\xi}]\|^2] = \mathbb{E}[\|\nabla f(x_t;\xi) - \nabla f(x_t) - \alpha(\nabla f(x_{t-1};\xi) - \nabla f(x_{t-1}))\|^2], \tag{13}$$

by adding and subtracting the term $\nabla f(x_{t-1};\xi) - \nabla f(x_{t-1})$, and $\mathbb{E}[\|a+b\|^2] \leq 2\mathbb{E}[\|a\|^2] + 2\mathbb{E}[\|b\|^2]$, Equation 13 is bounded by:

$$\begin{aligned} &2\mathbb{E}[\|\nabla f(x_t;\xi) - \nabla f(x_t) - (\nabla f(x_{t-1};\xi) - \nabla f(x_{t-1}))\|^2] \\ &+2\mathbb{E}[\|(1-\alpha)(\nabla f(x_{t-1};\xi) - \nabla f(x_{t-1}))\|^2]. \end{aligned} \tag{14}$$

By applying $\mathbb{E}[\|X - c\|^2] = \mathbb{E}[\|X - \mathbb{E}[X]\|^2] + \mathbb{E}[\|\mathbb{E}[X] - c\|^2]$ and Assumption 2 to the first term and second term, respectively, we derive the upper bound of (14), which is:

$$2\mathbb{E}[\|\nabla f(x_t;\xi) - (\nabla f(x_{t-1};\xi)\|^2] - 2\|\nabla f(x_t) - \nabla f(x_{t-1})\|^2 + 2(1-\alpha)^2 r^2. \tag{15}$$

By dropping the term $-2\|\nabla f(x_t) - \nabla f(x_{t-1})\|^2$ and applying Assumption 1, the upper bound is given by:

$$2L^2\eta^2\mathbb{E}[\|\bar{g}_t\|_2^2] + 2(1-\alpha)^2 r^2, \tag{16}$$

then, by utilizing the independence of DP noise and gradient, we further derive the bound of $\mathbb{E}[\|\bar{g}_t\|_2^2]$ and have:

$$2L^2\eta^2\mathbb{E}[\|\bar{g}_t\|_2^2] + 2(1-\alpha)^2 r^2 \leq 2L^2\eta^2(C^2 + d\sigma^2) + 2(1-\alpha)^2 r^2, \tag{17}$$

setting $\eta \leq \frac{(1-\alpha)r}{L\sqrt{2(C^2+d\sigma^2)}}$, we obtain the final upper bound of the first term:

$$\mathbb{E}[\|g_{t,\xi} - \mathbb{E}[g_{t,\xi}]\|^2] \leq 3(1-\alpha)^2 r^2. \tag{18}$$

For the second term $\mathbb{E}[\|\mathbb{E}[g_{t,\xi}] - \nabla f(x_t)\|^2]$, by expanding $\mathbb{E}[g_{t,\xi}]$, we could obtain:

$$\begin{aligned} &\mathbb{E}[\|\mathbb{E}[g_{t,\xi}] - \nabla f(x_t)\|^2] \\ =&\mathbb{E}[\|\mathbb{E}[\nabla f(x_t;\xi) - \alpha\nabla f(x_{t-1};\xi) + \alpha m_{t-1}] - \nabla f(x_t)\|^2] \\ =&\alpha^2\|\Delta_{m,t-1}\|^2. \end{aligned} \tag{19}$$

Substituting (18) and (19) to (12), we have:

$$\mathbb{E}[\|\Delta_{g,t}\|^2] \leq \alpha^2\|\Delta_{m,t-1}\|^2 + 3(1-\alpha)^2 r^2, \tag{20}$$

which completes the proof. □

B Proof of Lemma 2

Proof. We start from the definition of $\Delta_{m,t}$:

$$\Delta_{m,t} = \alpha m_{t-1} + (1-\alpha)\bar{g}_t - \nabla f(x_t). \tag{21}$$

By adding and subtracting terms $\nabla f(x_{t-1})$, we have:

$$\Delta_{m,t} = \alpha(m_{t-1} - \nabla f(x_{t-1})) + \alpha(\nabla f(x_{t-1}) - \nabla f(x_t)) + (1-\alpha)(\bar{g}_t - \nabla f(x_t)) \tag{22}$$

For the first term $\alpha(m_{t-1} - \nabla f(x_{t-1}))$, by utilizing the definition of $\Delta_{m,t}$, we obtain:

$$\mathbb{E}[\|\alpha(m_{t-1} - \nabla f(x_{t-1}))\|^2] = \alpha^2 \|\Delta_{m,t-1}\|^2. \tag{23}$$

For the second term $\alpha(\nabla f(x_{t-1}) - \nabla f(x_t))$, by applying L-Smoothness assumption (Assumption 1), we could derive the upper bound of $\mathbb{E}[\|\alpha(\nabla f(x_{t-1}) - \nabla f(x_t))\|^2]$ as:

$$\mathbb{E}[\|\alpha(\nabla f(x_{t-1}) - \nabla f(x_t))\|^2] \leq \alpha^2 \mathbb{E}[\|\eta L \bar{g}_{t-1}\|^2] = \alpha^2 \eta^2 L^2 \mathbb{E}[\|\bar{g}_{t-1}\|^2]. \tag{24}$$

Then, given $\mathbb{E}[\|\bar{g}_{t-1}\|^2]$ is bounded by $C^2 + d\sigma^2$, we have:

$$\mathbb{E}[\|\alpha(\nabla f(x_{t-1}) - \nabla f(x_t))\|^2] \leq \alpha^2 \eta^2 L^2 (C^2 + d\sigma^2). \tag{25}$$

Now, we bound the third term $\mathbb{E}[\|(1-\alpha)(\bar{g}_t - \nabla f(x_t))\|^2] = (1-\alpha)^2 \mathbb{E}[\|\bar{g}_t - \nabla f(x_t)\|^2]$. By using the definition of $\bar{g}_t$, we have:

$$\begin{aligned}
&\mathbb{E}[\|\bar{g}_t - \nabla f(x_t)\|^2] \\
=&\mathbb{E}[\|\frac{1}{B}\sum_{\xi \in \mathcal{B}_t} \tilde{g}_{t,\xi} + w_t - \nabla f(x_t)\|^2] \\
=&\mathbb{E}[\|\frac{1}{B}\sum(\tilde{g}_{t,\xi} - g_{t,\xi} + g_{t,\xi}) + w_t - \nabla f(x_t)\|^2] \\
\leq&2\mathbb{E}[\|\frac{1}{B}\sum g_{t,\xi} - \nabla f(x_t)\|^2] + 2\mathbb{E}[\|\frac{1}{B}\sum(\tilde{g}_{t,\xi} - g_{t,\xi})\|^2] + d\sigma^2,
\end{aligned} \tag{26}$$

where the last inequality re-arranging the terms, applying $\mathbb{E}[\|w_t\|^2] = d\sigma^2$, and utilizing $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. For the first term $\mathbb{E}[\|\frac{1}{B}\sum g_{t,\xi} - \nabla f(x_t)\|^2]$, by applying Jensen's Inequality, we obtain:

$$\mathbb{E}[\|\frac{1}{B}\sum g_{t,\xi} - \nabla f(x_t)\|^2] \leq \frac{1}{B}\sum_{\xi \in \mathcal{B}_t} \mathbb{E}[\|g_{t,\xi} - \nabla f(x_t)\|^2] \leq \mathbb{E}[\|\Delta_{g,t}\|^2]. \tag{27}$$

For the second term $\mathbb{E}[\|\frac{1}{B}\sum(\tilde{g}_{t,\xi} - g_{t,\xi})\|^2]$, given the inherent property of clipping, we always have $\|\tilde{g}_{t,\xi} - g_{t,\xi}\|^2 \leq \|g_{t,\xi}\|_2^2$. Therefore, similar to (27), we have:

$$\mathbb{E}[\|\frac{1}{B}\sum(\tilde{g}_{t,\xi} - g_{t,\xi})\|^2] \leq \frac{1}{B}\sum_{\xi \in \mathcal{B}_t} \mathbb{E}[\|g_{t,\xi}\|^2], \tag{28}$$

by further applying $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we derive the upper bound, which is denoted as:

$$\mathbb{E}[\|\frac{1}{B}\sum(\tilde{g}_{t,\xi} - g_{t,\xi})\|^2] \leq 2\|\nabla f(x_t)\|^2 + 2\mathbb{E}[\|\Delta_{g,t}\|^2]. \tag{29}$$

Therefore, substituting (27) and (29) into (26), we obtain the upper bound of $\mathbb{E}[\|(1-\alpha)(\bar{g}_t - \nabla f(x_t))\|^2]$, denoted as:

$$\mathbb{E}[\|(1-\alpha)(\bar{g}_t - \nabla f(x_t))\|^2] \leq 2(1-\alpha)^2(3\mathbb{E}[\|\Delta_{g,t}\|^2] + 2\|\nabla f(x_t)\|^2) + d\sigma^2 \tag{30}$$

Finally, using $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$ and substituting (23), (25), and (30) into (22), we obtain:

$$\mathbb{E}[\|\Delta_{m,t}\|^2] \leq 3\alpha^2\|\Delta_{m,t-1}\|^2 + 4d\sigma^2 + 6(1-\alpha)^2(3\mathbb{E}[\|\Delta_{g,t}\|^2] + 2\|\nabla f(x_t)\|^2), \tag{31}$$

where $\eta \leq \sqrt{\frac{d\sigma^2}{3\alpha^2 L^2 (C^2 + d\sigma^2)}}$.
This completes the proof. □

C Proof of Theorem 2

Proof. From the L-smoothness of $f(x)$ (Assumption 1) and the upper bound of $\mathbb{E}[\|\bar{g}_t\|^2]$, we have:

$$\begin{aligned}
&\mathbb{E}[f(x_{t+1}) - f(x_t)] \\
\leq &- \eta\mathbb{E}[\langle \nabla f(x_t), \bar{g}_t\rangle] + \frac{L\eta^2}{2}\mathbb{E}[\|\bar{g}_t\|^2] \\
\leq &- \eta\mathbb{E}[\langle \nabla f(x_t), \bar{g}_t\rangle] + \frac{L\eta^2}{2}(C^2 + d\sigma^2) \\
\leq &- \eta\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}] + \frac{L\eta^2}{2}(C^2 + d\sigma^2),
\end{aligned} \tag{32}$$

where the last inequality is derived by given the DP noise is independent and the linearity of $\mathbb{E}[\cdot]$. We then consider two cases based on the difference between $\nabla f(x_t)$ and $\mathbb{E}[\|\Delta_{g,t}\|]$.

Case 1: $\|\nabla f(x_t)\| \geq 6\mathbb{E}[\|\Delta_{g,t}\|]$.
We first define two complementary events based on the random sample ξ at step t:

$$\begin{aligned}
A_t &= \{\xi \mid \|g_{t,\xi} - \nabla f(x_t)\| \leq 3\sqrt{\mathbb{E}[\|\Delta_{g,t}\|]}\} \\
A_t^c &= \{\xi \mid \|g_{t,\xi} - \nabla f(x_t)\| > 3\sqrt{\mathbb{E}[\|\Delta_{g,t}\|]}\}
\end{aligned}$$

By using Markov's inequality, we obtain:

$$P(A_t^c) \leq \frac{\mathbb{E}[\|\Delta_{g,t}\|]^2}{(3\sqrt{\mathbb{E}[\|\Delta_{g,t}\|]})^2} \leq \frac{1}{9}, \tag{33}$$

which implies $P(A_t) \geq 8/9$. By using the law of total expectation, we get:

$$\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}\rangle] = P(A_t)\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}\rangle \mid A_t] + P(A_t^c)\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}\rangle \mid A_t^c] \tag{34}$$

Under event A_t, by the definition of clipping operation, we have:

$$\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}\rangle \mid A_t] = \mathbb{E}[min(1, \frac{C}{\|g_{t,\xi}\|}) \times \langle \nabla f(x_t), g_{t,\xi}\rangle] \geq \frac{C}{2\|\nabla f(x_t)\|} \times \mathbb{E}[\langle \nabla f(x_t), g_{t,\xi}\rangle], \tag{35}$$

where the last inequality is given by $\|g_{t,\xi}\| \leq \|g_{t,\xi} - \nabla f(x_t)\| + \|\nabla f(x_t)\| \leq 2\|\nabla f(x_t)\|$. Then, we have:

$$\begin{aligned}
\langle \nabla f(x_t), g_{t,\xi}\rangle &= \langle \nabla f(x_t), f(x_t) + (g_{t,\xi} - f(x_t))\rangle \\
&= \|\nabla f(x_t)\|^2 + \langle \nabla f(x_t), (g_{t,\xi} - f(x_t))\rangle \\
&\geq \|\nabla f(x_t)\|^2 - \|\nabla f(x_t)\|\|(g_{t,\xi} - f(x_t))\| \\
&\geq \|\nabla f(x_t)\|^2 - \frac{1}{2}\|\nabla f(x_t)\|^2 \\
&= \frac{1}{2}\|\nabla f(x_t)\|^2,
\end{aligned} \tag{36}$$

where the first inequality is derived by $\langle a, b\rangle \geq -\|a\|\|b\|$, and the second inequality is derived by $\|g_{t,\xi} - \nabla f(x_t)\| \leq 2\|\nabla f(x_t)\|$.

Under event A_t^c, we obtain:

$$\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}\rangle \mid A_t^c] = \langle \nabla f(x_t), \mathbb{E}[\tilde{g}_{t,\xi}]\rangle \geq -\|\nabla f(x_t)\|\|\mathbb{E}[\tilde{g}_{t,\xi}]\| \geq -C\|\nabla f(x_t)\|, \tag{37}$$

where the first inequality applies $\langle a, b\rangle \geq -\|a\|\|b\|$, and the second inequality utilizes $\|\mathbb{E}[\tilde{g}_{t,\xi}]\| \leq C$. Substituting these bounds into (34), we obtain:

$$\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}\rangle] \geq \frac{8}{9} \times \frac{C}{4}\|\nabla f(x_t)\| - \frac{1}{9} \times C\|\nabla f(x_t)\| \geq \frac{C}{9}\|\nabla f(x_t)\|.$$

Substituting this into (32), we have:

$$\mathbb{E}[f(x_{t+1}) - f(x_t)] \leq -\eta \frac{C}{9} \|\nabla f(x_t)\| + \frac{L\eta^2}{2}(C^2 + d\sigma^2) \tag{38}$$

Re-arranging terms and averaging over $t \in [0, T-1]$, we have

$$\mathbb{E}[\|\nabla f(x_t)\|] \leq \frac{9(f(x_0) - f^*)}{\eta C T} + \frac{9L\eta(C^2 + d\sigma^2)}{2C}. \tag{39}$$

Case 2: $\|\nabla f(x_t)\| < 6\mathbb{E}[\|\Delta_{g,t}\|]$.
We first divide $-\eta\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}\rangle]$ into two parts, denoted as:

$$-\eta\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}\rangle] = -\eta\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi}\rangle] = -\eta\langle \nabla f(x_t), \mathbb{E}[g_{t,\xi}]\rangle - \eta\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi} - g_{t,\xi}\rangle]. \tag{40}$$

For the first term, we have:

$$\begin{aligned}
&-\eta\langle \nabla f(x_t), \mathbb{E}[g_{t,\xi}]\rangle \\
=&-\eta\langle \nabla f(x_t), \nabla f(x_t) - \alpha\nabla f(x_{t-1}) + \alpha m_{t-1}\rangle \\
=&-\eta\|\nabla f(x_t)\|^2 - \eta\alpha\langle \nabla f(x_t), m_{t-1} - \nabla f(x_{t-1})\rangle \\
\leq&-\eta\|\nabla f(x_t)\|^2 + \frac{\eta\alpha}{2}\|\nabla f(x_t)\|^2 + \frac{\eta\alpha}{2}\|\Delta_{m,t-1}\|^2,
\end{aligned} \tag{41}$$

where the last inequality applies $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$. For the second term, by utilizing Cauchy-Schwarz inequality, we obtain:

$$-\eta\mathbb{E}[\langle \nabla f(x_t), \tilde{g}_{t,\xi} - g_{t,\xi}\rangle] \leq \eta\mathbb{E}[\|\nabla f(x_t)\| \cdot \|\tilde{g}_{t,\xi} - g_{t,\xi}\|], \tag{42}$$

Then, by utilizing Young's Inequality, we have:

$$\begin{aligned}
&\eta\mathbb{E}[\|\nabla f(x_t)\| \cdot \|\tilde{g}_{t,\xi} - g_{t,\xi}\|] \\
\leq&\frac{\eta}{2}\|\nabla f(x_t)\|^2 + \frac{\eta}{2}\mathbb{E}[\|\tilde{g}_{t,\xi} - g_{t,\xi}\|^2] \\
\leq&\frac{\eta}{2}\|\nabla f(x_t)\|^2 + \eta\|\nabla f(x_t)\|^2 + \eta\mathbb{E}[\|\Delta_{g,t}\|^2] \\
\leq&10\eta\mathbb{E}[\|\Delta_{g,t}\|^2],
\end{aligned} \tag{43}$$

where the second inequality applies (29), and the third inequality uses $\|\nabla f(x_t)\| < 6\mathbb{E}[\|\Delta_{g,t}\|]$. Combining those together, we obtain an upper bound of $\mathbb{E}[f(x_{t+1}) - f(x_t)]$, which is:

$$\begin{aligned}
&\mathbb{E}[f(x_{t+1}) - f(x_t)] \\
\leq&-(\eta - \frac{\eta\alpha}{2})\|\nabla f(x_t)\|^2 + \frac{\eta\alpha}{2}\|\Delta_{m,t-1}\|^2 + 55\eta\mathbb{E}[\|\Delta_{g,t}\|^2] + \frac{L\eta^2}{2}(C^2 + d\sigma^2) \\
\leq&-(\eta - \frac{\eta\alpha}{2})\|\nabla f(x_t)\|^2 + (55\eta\alpha^2 + \frac{\eta\alpha}{2})\|\Delta_{m,t-1}\|^2 + 55\eta(1-\alpha)^2 r^2 + \frac{L\eta^2}{2}(C^2 + d\sigma^2),
\end{aligned} \tag{44}$$

where the last inequality applies Lemma 1. Define $\mathcal{L}_t = f(x_t) + \beta\|\Delta_{m,t}\|^2$, we have:

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(x_{t+1})] - \mathcal{L}(x_t) \\
\leq&-(\eta - \frac{\eta\alpha}{2})\|\nabla f(x_t)\|^2 - (\beta - 55\eta\alpha^2 - \frac{\eta\alpha}{2})\|\Delta_{m,t-1}\|^2 + 55\eta(1-\alpha)^2 r^2 + \frac{L\eta^2}{2}(C^2 + d\sigma^2) + \beta\mathbb{E}[\|\Delta_{m,t}\|^2] \\
\leq&-(\eta - \frac{\eta\alpha}{2})\|\nabla f(x_t)\|^2 - (\beta - 55\eta\alpha^2 - \frac{\eta\alpha}{2})\|\Delta_{m,t-1}\|^2 + 55\eta(1-\alpha)^2 r^2 + \frac{L\eta^2}{2}(C^2 + d\sigma^2) + 3\alpha^2\beta\|\Delta_{m,t-1}\|^2 \\
&+ 6(1-\alpha)^2\beta(3\alpha^2\|\Delta_{m,t-1}\|^2 + 9(1-\alpha)^2 r^2) + 12(1-\alpha)^2\beta\|\nabla f(x_t)\|^2 + 4d\sigma^2\beta \\
=&-(\eta - \frac{\eta\alpha}{2} - 12(1-\alpha)^2\beta)\|\nabla f(x_t)\|^2 - (\beta - 55\eta\alpha^2 - \frac{\eta\alpha}{2} - 3\alpha^2\beta - 18\alpha^2(1-\alpha)^2\beta)\|\Delta_{m,t-1}\|^2 \\
&+ \frac{L\eta^2}{2}(C^2 + d\sigma^2) + 54(1-\alpha)^4\beta r^2 + 4\beta d\sigma^2 + 55\eta(1-\alpha)^2 r^2.
\end{aligned} \tag{45}$$

By choosing:

$$\beta \geq \frac{110\eta\alpha^2 + \eta\alpha}{1 - 3\alpha^2 - 18\alpha^2(1-\alpha)^2} \tag{46}$$

$$\alpha \geq min\left(1 - \sqrt[4]{\frac{C^2}{54r^2}}, \frac{48\beta + 1 - \sqrt{96\beta + 1}}{48\beta}\right), \tag{47}$$

we have:

$$\eta - \frac{\eta\alpha}{2} - 12(1-\alpha)^2\beta \geq (1-\alpha)\eta \tag{48}$$

$$\beta - 55\eta\alpha^2 - \frac{\eta\alpha}{2} - 3\alpha^2\beta - 18\alpha^2(1-\alpha)^2\beta \geq 0 \tag{49}$$

$$54(1-\alpha)^4\beta r^2 \leq 4\beta C^2. \tag{50}$$

Therefore, we have the following:

$$\mathbb{E}[\mathcal{L}(x_{t+1})] - \mathcal{L}(x_t) \leq -(1-\alpha)\eta\|\nabla f(x_t)\|^2 + (4\beta + \frac{L\eta^2}{2})(C^2 + d\sigma^2) + 55\eta(1-\alpha)^2 r^2. \tag{51}$$

Averaging t from 0 to $T-1$, we obtain:

$$\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{f(x_0) + \beta\|\nabla f(x_0)\| - f^*}{(1-\alpha)\eta T} + \frac{8\beta + L\eta^2}{2(1-\alpha)\eta}(C^2 + d\sigma^2) + 55(1-\alpha)r^2. \tag{52}$$

Using Jensen's inequality, we derive the upper bound of $\mathbb{E}[\|\nabla f(x_t)\|]$:

$$\begin{aligned} &\mathbb{E}[\|\nabla f(x_t)\|] \\ \leq &\sqrt{\mathbb{E}[\|\nabla f(x_t)\|^2]} \\ \leq &\sqrt{\frac{f(x_0) + \beta\|\nabla f(x_0)\| - f^*}{(1-\alpha)\eta T}} + \sqrt{\frac{8\beta + L\eta^2}{2(1-\alpha)\eta}(C^2 + d\sigma^2)} + \sqrt{55(1-\alpha)}r, \end{aligned} \tag{53}$$

where the last inequality is given by $\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$.
Finally, combining the two cases, we get:

$$\begin{aligned} &\mathbb{E}[\|\nabla f(x_t)\|] \\ \leq &\mathcal{O}\left(\sqrt{\frac{f(x_0) + \beta\|\nabla f(x_0)\| - f^*}{(1-\alpha)\eta T}} + \frac{(f(x_0) - f^*)}{\eta C T} + \sqrt{\frac{\beta + L\eta^2}{(1-\alpha)\eta}(C^2 + d\sigma^2)} + \frac{L\eta(C^2 + d\sigma^2)}{C} + \sqrt{(1-\alpha)}r\right), \end{aligned} \tag{54}$$

which completes the proof. □