Enhancing DPSGD via Per-Sample Momentum and Low-Pass Filtering

Xincheng Xu[1], **Thilina Ranbaduge**[2], **Qing Wang**[1], **Thierry Rakotoarivelo**[2], **David Smith**[2]

[1]School of Computing, Australian National University, Australia
[2]Data 61, CSIRO, Australia
{xincheng.xu, qing.wang}@anu.edu.au, {thilina.ranbaduge, thierry.rakotoarivelo, david.smith}@data61.csiro.au

Abstract

Differentially Private Stochastic Gradient Descent (DPSGD) is widely used to train deep neural networks with formal privacy guarantees. However, the addition of differential privacy (DP) often degrades model accuracy by introducing both noise and bias. Existing techniques typically address only one of these issues, as reducing DP noise can exacerbate clipping bias and vice-versa. In this paper, we propose a novel method, *DP-PMLF*, which integrates per-sample momentum with a low-pass filtering strategy to simultaneously mitigate DP noise and clipping bias. Our approach uses per-sample momentum to smooth gradient estimates prior to clipping, thereby reducing sampling variance. It further employs a post-processing low-pass filter to attenuate high-frequency DP noise without consuming additional privacy budget. We provide a theoretical analysis demonstrating an improved convergence rate under rigorous DP guarantees, and our empirical evaluations reveal that DP-PMLF significantly enhances the privacy-utility trade-off compared to several state-of-the-art DPSGD variants.

Code — https://github.com/CharlieX001/DPPMLF
Extended version — https://arxiv.org/abs/2511.08841

Introduction

Deep learning has achieved remarkable success in various domains, such as medical diagnosis (Aggarwal et al. 2021; Chen et al. 2022), recommendation systems (Chen et al. 2023b; Fu, Niu, and Maher 2023), and autonomous driving (Bachute and Subhedar 2021). However, training deep models often requires large amounts of sensitive data, raising privacy concerns. Recent research has shown that trained models not only could reveal the presence of individuals in a dataset (Choquette-Choo et al. 2021; Olatunji, Nejdl, and Khosla 2021), but are also vulnerable to model inversion or reconstruction attacks (Zhao et al. 2021; Wang et al. 2021a; Nguyen et al. 2023).

Differential Privacy (DP) (Dwork, Roth et al. 2014) has become the de facto standard for privacy-preserving deep learning (Tanuwidjaja et al. 2020; Boulemtafes, Derhab, and Challal 2020), offering formal privacy guarantees for training data. Among various DP training algorithms, *Differentially Private Stochastic Gradient Descent (DPSGD)* (Abadi et al. 2016) is widely used for training deep neural networks with privacy guarantees. DPSGD enforces (ϵ,δ)-DP through two key mechanisms: (1) *gradient clipping*, which bounds the ℓ_2 norm of individual gradients to limit the influence of any single training sample, and (2) *noise injection*, where DP noise calibrated to the privacy budget ϵ and failure factor δ is added to the aggregated gradients. However, DPSGD faces a challenging privacy-utility trade-off: per-sample gradient clipping can impede convergence, and the added noise can significantly degrade model performance (Fang et al. 2023).

To improve the utility of DPSGD, recent works have proposed various strategies, such as adaptively adjusting the clipping threshold (Andrew et al. 2021; Bu et al. 2024; Xia et al. 2023), dynamically allocating the privacy budget (Lee and Kifer 2018; Yu et al. 2019; Chen et al. 2023a), projecting gradients into low-dimensional spaces (Zhou, Wu, and Banerjee 2021; Yu et al. 2021a; Asi et al. 2021; Yu et al. 2022), designing models less sensitive to DP noise (Papernot et al. 2021; Wang et al. 2021b; Shamsabadi and Papernot 2023), and incorporating public data (Li et al. 2022; Amid et al. 2022; Golatkar et al. 2022). Despite these advances, these methods face practical challenges: some lack rigorous theoretical guarantees, others are limited to specific model architectures, or require access to public data, all of which hinder the feasibility of DPSGD in real-world applications.

Beyond these practical challenges, a key theoretical concern is the convergence behavior of DPSGD, which is influenced by two factors: DP noise and clipping bias. Generally, selecting a smaller clipping threshold reduces injected DP noise, minimizing its scale but increasing the clipping bias. Conversely, a larger clipping threshold lowers clipping bias but requires injecting more DP noise to maintain privacy guarantees, which potentially leads to significant performance degradation.

Existing methods attempt to mitigate one effect at the expense of the other. Zhang *et al.* (Zhang et al. 2024a) applies a low-pass filter to separate DP noise from the true gradient signal, but introduces an additional bias term in the convergence rate. DiceSGD (Zhang et al. 2024b) utilizes an error-feedback mechanism to correct bias but needs additional DP noise to protect the residual gradient information. Recent studies (Koloskova, Hendrikx, and Stich 2023; Xiao et al. 2023) suggest that clipping bias is not strictly related to clipping threshold, but is also influenced by sampling variance

σ_{SGD}. This insight inspired our work where we try to simultaneously reduce both DP noise and clipping bias, thereby enhancing the overall utility of DPSGD.

In this work, we propose a novel method, *DP-PMLF*, which mitigates both clipping bias and DP noise in DPSGD by integrating per-sample momentum and a low-pass filter. First, per-sample momentum is employed to average historical gradients, thereby reducing the sampling variance and bias introduced by gradient clipping. Second, a low-pass filter is applied as a post-processing step to suppress high-frequency DP noise while preserving the essential low-frequency gradient signal. Our theoretical analysis illustrates an improved convergence guarantee compared to DPSGD under some assumptions, while providing strong privacy guarantees. Empirical results demonstrate our method outperforms resent state-of-the-art techniques. Our main contributions are threefold:

- We propose a novel DPSGD method that simultaneously addresses DP noise and clipping bias through the integration of per-sample momentum and low-pass filtering. To the best of our knowledge, our approach is the first to consider reducing the effect of DP noise and clipping bias simultaneously.
- We theoretically prove DP-PMLF achieves faster convergence compared to the vanilla DPSGD, while maintaining a mathematically proven privacy guarantee.
- Empirical results on different benchmarks demonstrate that our approach achieves a better privacy-utility trade-off compared to various existing state-of-art DPSGD variants across different models and privacy levels.

Related Work

Existing variants on DPSGD can be categorized into two directions: DP noise reduction and clipping bias reduction.

DP Noise Reduction: To mitigate the effect of DP noise, existing approaches commonly use the following techniques: adaptive clipping threshold (Andrew et al. 2021; Bu et al. 2024; Xia et al. 2023), privacy budget allocation (Lee and Kifer 2018; Yu et al. 2019; Chen et al. 2023a), low-rank projection (Zhou, Wu, and Banerjee 2021; Yu et al. 2021a,b; Asi et al. 2021; Yu et al. 2022), specific model design (Papernot et al. 2021; Wang et al. 2021b; Shamsabadi and Papernot 2023), public data assistant (Li et al. 2022; Amid et al. 2022; Golatkar et al. 2022).

However, these methods often lack theoretical guarantees, have limited applicability to specific model architectures, or require access to public data for training. To solve these limitations, Zhang *et al.* (Zhang et al. 2024a) proposed the introduction of a low-pass filter as a post-processing step in DP optimizers. They demonstrated that low-pass filtering effectively suppresses high-frequency DP noise while preserving essential gradient information.

Clipping Bias Reduction: Koloskova *et al.* (Koloskova, Hendrikx, and Stich 2023) demonstrated that DPSGD converges with a constant bias term, irrespective of the chosen clipping threshold or the learning rate. Chen *et al.* (Chen, Wu, and Hong 2020) also proposed a geometric analysis framework that quantifies gradient clipping bias by measuring the disparity between gradient distributions and symmetric distributions, and a technique to add Gaussian noise to gradients before the clipping operation when gradients are highly asymmetric.

Xiao *et al.* (Xiao et al. 2023) found that the clipping bias is proportional to the sampling variance σ_{SGD}. The authors proposed to reduce the clipping bias using inner-outer momentum, enhanced network normalization, batch clipping with public data, and data pre-processing. Zhang *et al.* (Zhang et al. 2024b) introduced an error-feedback mechanism, *DiceSGD*, to accumulate the difference between clipped and unclipped gradients but requires more DP noise than vanilla DPSGD. To avoid clipping operation, Bethune *et al.* (Béthune et al. 2024) proposed *Clipless DPSGD*. This method utilizes Lipschitz-constrained neural networks, which analytically compute gradient sensitivity bounds using projection operations and gradient norm-preserving networks with orthogonal weights. However, *Clipless DPSGD* relies on specific model architectures which limit its practical deployment.

Preliminaries

We begin by outlining the Empirical Risk Minimization (ERM) problem along with its standard assumptions, and then provide an overview of DP.

Empirical Risk Minimization (ERM):

In this paper, we focus on differentially private optimization developed within the Empirical Risk Minimization (ERM) framework, which forms the basis for supervised deep learning. Let D be a dataset of n samples, where each sample ξ is drawn from some underlying distribution. In empirical risk minimization (ERM), we seek a parameter vector $x \in \mathbb{R}^d$ that minimizes the average loss:

$$\min_{x\in\mathbb{R}^d} f(x) \quad \text{with} \quad f(x) = \frac{1}{n}\sum_{\xi\in D} f(x,\xi), \tag{1}$$

where $f(x,\xi)$ is the loss incurred on sample ξ.

For clarity, we denote by $\nabla f^{(\xi)}(x) \equiv \nabla_x f(x,\xi)$ the gradient with respect to x computed on a single sample ξ, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$, and T denote the total number of iterations of the optimization algorithm. We then introduce the following assumptions used in our work:

Assumption 1 (L-Smoothness)**.** *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be L-smooth if it satisfies, for all $x, y \in \mathbb{R}^d$:*

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

Assumption 2 (Bounded Variance)**.** *The per-sample gradient has bounded variance, i.e.,*

$$\mathbb{E}\left[\|\nabla f^{(\xi)}(x) - \nabla f(x)\|^2\right] \le \sigma_{SGD}^2, \quad \forall x \in \mathbb{R}^d.$$

Here, the expectation is taken with respect to the sampling of ξ and σ_{SGD} is a constant representing the variance bound.

Assumption 3 (Bounded Gradient)**.** *The per-sample gradient has a bounded norm, i.e.,*

$$\|\nabla f^{(\xi)}(x)\| \leq G, \quad \forall x \in \mathbb{R}^d,\ \xi \in D,$$

where G is a positive constant.

Assumption 4 (Gradient Auto-Correlation)**.** *For all $t \in \{0, \ldots, T-1\}$, there exist sequences $\{c_r\}$ and $\{c_{-r}\}$ with $c_r \geq 0$, and $\forall r \geq 0$, such that*

$$\langle \nabla f(x_t), \nabla f(x_{t-r}) \rangle \geq c_r \|\nabla f(x_t)\|^2 + c_{-r} \|\nabla f(x_{t-r})\|^2.$$

Assumption 5 (Independent Sampling Noise)**.** *Let $\zeta_i^{(\xi)} = \nabla f^{(\xi)}(x_i) - \nabla f(x_i)$ represent the sampling noise from the sample ξ in the i-th iteration. If $i \neq j$, then the following condition holds:*

$$\mathbb{E}\left[\left(\zeta_i^{(\xi)}\right)^T \zeta_j^{(\xi)}\right] = 0.$$

Assumption 1 is a widely adopted smoothness condition in non-convex optimization (Zaheer et al. 2018). Assumption 2 is standard in the analysis of gradient clipping (Gorbunov, Danilova, and Gasnikov 2020). Assumption 3 is commonly used in the DPSGD setting to control the additional bias introduced by clipping (Zhang et al. 2024b). Assumptions 4 and 5 are proposed and validated in (Zhang et al. 2024a) and (Xiao et al. 2023), respectively.

Differential Privacy (DP)

DP (Dwork, Roth et al. 2014) provides a privacy guarantee such that the outputs of a mechanism cannot be distinguished by the inclusion or exclusion of any single record in a dataset. Formally, DP is defined as follows:

Definition 1 (Differential Privacy (DP) (Dwork et al. 2006))**.** *A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{R}^d$ is (ϵ, δ)-DP if for all neighboring datasets D and D', and for any output set $\mathcal{S} \subseteq \mathcal{R}^d$, we have*

$$Pr[\mathcal{M}(D) \in \mathcal{S}] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in \mathcal{S}] + \delta, \qquad (2)$$

where $\delta \in [0, 1]$ denotes a failure probability.

When $\delta = 0$, the mechanism $\mathcal{M}$ is said to satisfy *pure DP*; if $\delta > 0$, it satisfies *approximate DP*.

The Gaussian mechanism is widely used to achieve the DP guarantee. The definition of global sensitivity and the Gaussian mechanism are defined as follows:

Definition 2 (Global Sensitivity)**.** *Let $\mathcal{H} : \mathcal{D} \to \mathcal{R}^d$ be a function that maps datasets to d-dimensional vectors. The global sensitivity of $\mathcal{H}$ is defined as:*

$$\Delta\mathcal{H} = \max_{D \sim D'} \|\mathcal{H}(D) - \mathcal{H}(D')\|.$$

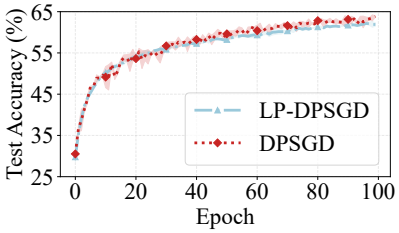
Definition 3 (Gaussian Mechanism (Dwork, Roth et al. 2014))**.** *For a function $\mathcal{H} : \mathcal{D} \to \mathcal{R}^d$ with ℓ_2 global sensitivity $\Delta\mathcal{H}$, the Gaussian mechanism is defined as*

$$\mathcal{M}(D) = \mathcal{H}(D) + \mathcal{N}(0, \sigma_{DP}^2 I_d),$$

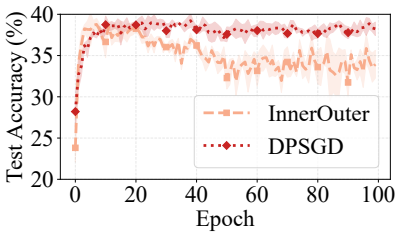
where $\mathcal{N}(0, \sigma_{DP}^2 I_d)$ denotes the d-dimensional multivariate Gaussian distribution with mean zero and covariance matrix $\sigma_{DP}^2 I_d$. The noise parameter is set to

$$\sigma_{DP} = \frac{\Delta\mathcal{H}\sqrt{2\ln(1.25/\delta)}}{\epsilon},$$

which ensures that $\mathcal{M}$ satisfies (ϵ, δ)-DP.



(a) DPSGD and LP-DPSGD with $\epsilon = 8$.



(b) DPSGD and InnerOuter with $\epsilon = 1$.

Figure 1: Test accuracy (%) comparison of DPSGD and two existing methods on CIFAR-10 with a 5-Layer CNN over 100 epochs under different privacy budgets (ϵ).

Our approach is also built upon post-processing, one fundamental DP properties:

Lemma 1 (Post-Processing (Dwork et al. 2006))**.** *Let $\mathcal{M} : \mathcal{D} \to \mathcal{R}^d$ be an (ϵ, δ)-DP mechanism and let $\mathcal{H} : \mathcal{R}^d \to \mathcal{R}^d$ be any deterministic or randomized function. Then the composition $\mathcal{H} \circ \mathcal{M}$ satisfies (ϵ, δ)-DP.*

Motivation

Reducing DP Noise. As mentioned in Section , different works have been proposed to mitigate the effects of DP noise. The state-of-the-art work by Zhang et al. (Zhang et al. 2024a), named *LP-DPSGD*, was proposed to preserve the integrity of the true gradient signals while mitigating the impact of DP noise. LP-DPSGD employs a low-pass filter to process gradients, effectively retaining low-frequency gradient signals while suppressing high-frequency noise, there by improving the signal-to-noise ratio of the gradients.

Although LP-DPSGD reduces the impact of DP noise, it introduces an increased bias as a trade-off. As shown by Koloskova et al. (Koloskova, Hendrikx, and Stich 2023), if the true gradient is sufficiently large, it can offset the sampling variance σ_{SGD}, thus preventing clipping bias. However, when incorporating a low-pass filter, this clipping bias cannot be eliminated, as the true gradients from different training iterations could be (negatively) correlated or even completely uncorrelated.

As can be seen in Figure 1 (a), the performance of LP-DPSGD is even worse than that of vanilla DPSGD. This is because the impact of clipping bias outweighs that of DP noise in this scenario. As a result, while LP-DPSGD effectively suppresses DP noise, the additional bias introduced by the low-pass filter undermines the overall performance. Further details on our analysis are provided in Section .

Mitigating Clipping Bias. Xiao *et al.* (Xiao et al. 2023) found that the clipping bias term is proportional to the sampling variance σ_{SGD}. The authors proposed the ***DPSGD with Inner-Outer Momentum*** (abbreviated to ***InnerOuter*** for simplicity) approach as a way of reducing clipping bias. The inner momentum smooths the gradients at the sample level by averaging the gradients in the previous training iterations before clipping, effectively reducing the impact of sampling noise. The outer momentum aggregates the clipped gradients of all samples in a batch, adds DP noise, and applies a second round of smoothing at the batch level.

However, InnerOuter does not perform well when the DP noise is large. The accumulation of historical gradients through the outer momentum operation cannot separate the true gradient and DP noise signals. This accumulates more DP noise which leads to inaccurate gradient estimates in training iterations. As seen in Figure 1 (b), InnerOuter is not effective where DP noise dominates the signal. Furthermore, the authors did not provide theoretical proofs leaving InnerOuter's convergence insufficiently validated.

Our Proposed Approach

Now we provide details of DP-PMLF. Our approach is built on two complementary ideas. **(1) Per-sample Momentum:** By maintaining a momentum term for each sample, we average historical gradients over a window of k iterations. This *per-sample momentum* reduces sampling variance and mitigates clipping bias by smoothing out fluctuations before the clipping step. **(2) Low-pass Filter:** DP noise is evenly distributed among all frequency components, while true gradient signals concentrate in low frequencies. By applying a linear low-pass filter to the aggregated noisy momentum, we suppress high-frequency noise while preserving the true low-frequency components.

Together, these ideas balance the trade-off between reducing noise and controlling clipping bias. The per-sample momentum provides a more stable gradient estimate prior to clipping, and the low-pass filter further cleans the aggregated signal without consuming additional privacy budget.

Our proposed method is outlined in Algorithm 1. The algorithm proceeds as follows:

Per-sample Momentum Calculation (lines 1 and 5): For each sample ξ, we compute a momentum term by averaging its gradients over the previous k iterations using exponential decay weights. The momentum term

$$v_t^{(\xi)} = \sum_{i=t-k+1}^{t} \hat{\beta}^{\,t-i} \nabla f^{(\xi)}(x_i).$$

Here, $\hat{\beta}^{\,t-i} = \frac{\beta^{t-i}}{c_\beta}$ and $c_\beta = \sum_{i=t-k+1}^{t} \beta^{t-i}$. c_β is the normalization constant that ensures that the momentum coefficients sum to one, preventing excessive accumulation of DP noise.

Momentum Clipping and Noise Addition (lines 6 and 8): Each sample's momentum $v_t^{(\xi)}$ is clipped to a threshold C, which bounds the global sensitivity $\tilde{v}_t^{(\xi)} = \text{clip}(v_t^{(\xi)}, C)$. The aggregated momentum is then computed and Gaussian noise with scale σ_{DP} is added to each dimension of the average clipped momentum to satisfy DP guarantees:

$$\bar{v}_t = \frac{1}{B} \sum_{\xi \in \mathcal{B}_t} \tilde{v}_t^{(\xi)} + w_t, \quad \text{with } w_t \sim \mathcal{N}(0, \sigma_{DP}^2 I_d).$$

Algorithm 1: DP-PMLF

Require: dataset D, initial model parameters x_0, learning rate η, momentum length k, filter parameters $\{a_r\}_{r=1}^{n_a}$, $\{b_r\}_{r=0}^{n_b}$, clipping threshold C, noise scale σ_{DP}, batch size B, iteration number T, per-sample momentum factor β

1: $c_\beta = sum(\sum_{i=t-k+1}^{t} \beta^{t-i})$
2: **for** $t = 0$ to $T-1$ **do**
3: Sample minibatch $\mathcal{B}_t$ of size B from D
4: **for** $\xi \in \mathcal{B}_t$ **do**
5: $v_t^{(\xi)} = \sum_{i=t-k+1}^{t} \hat{\beta}^{t-i} \nabla f^{(\xi)}(x_i)$, where $\hat{\beta}^{t-i} = \frac{\beta^{t-i}}{c_\beta}$
6: $\tilde{v}_t^{(\xi)} = \text{clip}(v_t^{(\xi)}, C)$
7: **end for**
8: $\bar{v}_t = \frac{1}{B} \sum_{\xi \in \mathcal{B}_t} \tilde{v}_t^{(\xi)} + w_t$, where $w_t \sim \mathcal{N}(0, \sigma_{DP}^2 I_d)$
9: $m_t = -\sum_{r=1}^{n_a} a_r m_{t-r} + \sum_{r=0}^{n_b} b_r \bar{v}_{t-r}$
10: $c_{b,t} = 1, c_{m,t} = -\sum_{r=1}^{n_a} a_r c_{m,t-r} + \sum_{r=0}^{n_b} b_r c_{b,t-r}$
11: $\hat{m}_t = m_t / c_{m,t}$
12: $x_{t+1} = x_t - \eta \hat{m}_t$
13: **end for**
14: **return** x_T

Low-pass Filtering and Bias Correction (lines 9 - 11): We apply a linear low-pass filter with coefficients $\{a_r\}_{r=1}^{n_a}$ and $\{b_r\}_{r=0}^{n_b}$ to the aggregated noisy momentum:

$$m_t = -\sum_{r=1}^{n_a} a_r \, m_{t-r} + \sum_{r=0}^{n_b} b_r \, \bar{v}_{t-r},$$

where m_t is the filtered output at time t, $\bar{v}_{t-r}$ represents the aggregated noisy momentum at time $t-r$, $\{a_r\}$ and $\{b_r\}$ are the filter coefficients and n_a and n_b determine the filter order. To ensure that after filtering the mean of the signal remains unchanged (Winder 2002), the design of the filter coefficients should satisfy the following constraint:

$$-\sum_{r=1}^{n_a} a_r \; + \sum_{r=0}^{n_b} b_r = 1. \tag{3}$$

We calculate an initialization bias correction term via

$$c_{m,t} = -\sum_{r=1}^{n_a} a_r \, c_{m,t-r} + \sum_{r=0}^{n_b} b_r c_{b,t-r}.$$

We normalize m_t to correct the initialization bias for the filter's effect $\hat{m}_t = m_t / c_{m,t}$. This step smooths the signal, suppressing high-frequency DP noise while retaining the low-frequency, true momentum components.

Model Update (line 12): Finally, the model parameters are updated using the corrected momentum $x_{t+1} = x_t - \eta \, \hat{m}_t$.

Theoretical Analysis

We now present key lemmas that underpin our convergence analysis.

Lemma 2 (Effectiveness of Low-pass Filter)**.**

$$\hat{m}_t = \sum_{r=0}^{t} \hat{\kappa}_r \, \bar{v}_{t-r}, \text{ with}$$

$$\hat{\kappa}_r = \frac{\kappa_r}{\sum_{r=0}^{t} \kappa_r} \quad \text{and} \quad \kappa_r = \sum_{r_2=0}^{\min(n_b,r)} b_{r_2} \sum_{r_1=1}^{n_a} z_{a,r_1} \left(p_{a,r_1}\right)^{r-r_2}.$$

We begin by analyzing how the low-pass filter aggregates historical momentum information while attenuating high-frequency DP noise. This is crucial because, as shown in Section 3.3, the true gradient signal concentrates in the low-frequency regime while DP noise is spectrally flat. Thus, the low-pass filter suppresses the high-frequency components, as indicated by the decay properties of $\{\hat{\kappa}_r\}$. The proof of this lemma is available in the appendix.

Lemma 3 (Bounded Momentum Variance)**.** *Under Assumptions 1, 2, and 5, if the step size satisfies*

$$\eta \leq \sqrt{\frac{\sigma_{SGD}^2}{L^2 k^3 (C^2 + d\sigma_{DP}^2)}},$$

then

$$\mathbb{E}\left[\left\|v_t^{(\xi)} - \nabla f(x_t)\right\|^2\right] \leq \mathcal{O}\left(\frac{\sigma_{SGD}^2}{\rho^2}\right),$$

where

$$\rho = \sqrt{\frac{(1+\beta)(1-\beta^k)}{(1-\beta)(1+\beta^k)}}.$$

Proof Sketch. We decompose the error into the variance of the sampling noise $\mathbb{E}\left[\left\|\nabla f^{(\xi)}(x_i) - \nabla f(x_i)\right\|^2\right]$, and the error due to the drift between $\nabla f(x_i)$ and $\nabla f(x_t)$. The former is directly controlled by Assumption 2 and the independence in Assumption 5, while the latter is bounded via the L-smoothness condition (Assumption 1). The weighted averaging in the per-sample momentum reduces the overall variance by the factor ρ^2. Detailed derivations are provided in the appendix. □

Remark. *The factor ρ^2 increases with both the per-sample momentum factor β and the momentum length k. When β reaches its maximum value of 1, the exponential decay reduces to an equal-weight average, and ρ^2 approaches k. Theoretically, larger values of β and k are preferable; however, in practice, a large β may cause training to rely overly on historical information, potentially slowing convergence.*

Convergence Analysis We now combine the above lemmas to establish the convergence rate of Algorithm 1. The analysis builds on a standard descent lemma for L-smooth functions and is augmented by our representation of the low-pass filtered momentum and the variance reduction effect.

Step 1: Descent Lemma. By L-smoothness from Assumption 1, we have

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

Step 2: Decomposition of Gradient. Using the representation from Lemma 2, we write

$$\langle \nabla f(x_t), \hat{m}_t \rangle = \sum_{r=0}^{t} \hat{\kappa}_r \, \langle \nabla f(x_t), \bar{v}_{t-r} \rangle.$$

We decompose each inner product into two parts:

1. The correlation between the current gradient and the historical (unclipped) momentum, which under Assumption 4 can be bounded in terms of $\|\nabla f(x_t)\|^2$ and $\|\nabla f(x_{t-r})\|^2$.
2. A bias term arising from clipping, i.e., the difference $\mathbb{E}[\tilde{v}_{t-r}^{(\xi)} - v_{t-r}^{(\xi)}]$, which is bounded by $\mathcal{O}\left(G + \frac{\sigma_{SGD}}{\rho}\right)$ using Assumption 3 and Lemma 3.

Step 3: Final Convergence Bound. After careful estimation of the descent term and the term $\mathbb{E}\left[\|\hat{m}_t\|^2\right]$ (which also incorporates the effect of the DP noise with variance $d\,\sigma_{DP}^2$), telescoping over T iterations and taking averages yields the following convergence guarantee:

Theorem 1 (Convergence Bound)**.** *Under Assumptions 1–5, if Algorithm 1 is running for T iterations with step size*

$$\eta \leq \sqrt{\frac{\sigma_{SGD}^2}{L^2 k^3 (C^2 + d\,\sigma_{DP}^2)}},$$

then $\mathbb{E}\left[\|\nabla f(x_t)\|^2\right]$ is upper bounded by

$$\mathcal{O}\left(\frac{f(x_0) - f^*}{\eta T} + L\eta C^2 + \frac{L\eta}{\Gamma_{DP}} \, d\,\sigma_{DP}^2 + \left(\frac{G^2}{\Gamma_{SGD}} + \frac{\sigma_{SGD}^2}{\rho^2 \, \Gamma_{SGD}}\right)\right),$$

where $\hat{c}_r = \sum_{i=t-r-k+1}^{t-r} \hat{\beta}^{t-r-i} c_{t-i}$, $\rho = \sqrt{\frac{(1+\beta)(1-\beta^k)}{(1-\beta)(1+\beta^k)}}$, and $f^* = min_x f(x)$. $\Gamma_{DP} = \frac{\sum_{t=0}^{T-1} \sum_{r=0}^{t} \hat{\kappa}_r \hat{c}_r}{\sum_{t=0}^{T-1} \sum_{r=0}^{t} \hat{\kappa}_r^2}$ and $\Gamma_{SGD} = \frac{\sum_{t=0}^{T-1} \sum_{r=0}^{t} \hat{\kappa}_r \hat{c}_r}{\sum_{t=0}^{T-1} \sum_{r=0}^{t} \frac{\hat{\kappa}_r}{\hat{c}_r}}$ are two ratios introduce by low-pass filtering.

Remark. *The first term represents the optimization error decreasing with the number of iterations. The second term $L\eta C^2$ reflects the error introduced by gradient clipping. The third term captures the impact of the DP noise, where the aggregated effect is modulated by the filter coefficients. The final term aggregates the residual bias due to low-pass filtering and the reduced clipping bias from per-sample momentum. Our approach reduces the clipping bias term by introducing the variance reduction factor ρ.*

Corollary 1. *If the per-sample gradient norm is bounded by $G = \mathcal{O}\left(\frac{\sigma_{SGD}}{\rho}\right)$, then the bound in Theorem 1 simplifies to*

$$\mathcal{O}\left(\frac{f(x_0) - f^*}{\eta T} + L\eta C^2 + \frac{L\eta \, d\,\sigma_{DP}^2}{\Gamma_{DP}} + \frac{\sigma_{SGD}^2}{\rho^2 \, \Gamma_{SGD}}\right).$$

Remark. *A careful inspection of our convergence bound reveals that, by choosing β, k, and low-pass filter coefficients appropriately, our approach reduces the clipping bias term by introducing the variance reduction factor ρ and mitigates the effect of DP noise via the low-pass filter compared to vanilla DPSGD (Abadi et al. 2016).*

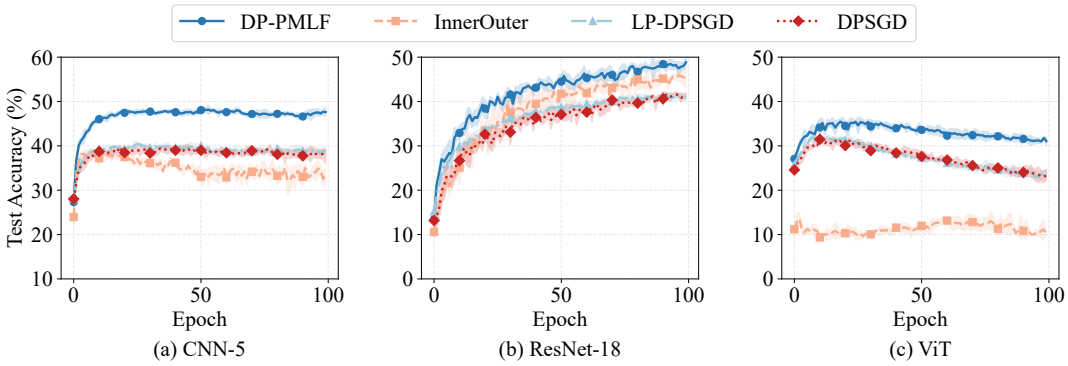
Privacy Analysis We establish that our approach satisfies (ϵ, δ)-differential privacy.

Theorem 2 (Differential Privacy Guarantee)**.** *There exist absolute constants c_1 and c_2 such that, given sampling probability $q = B/n$ and T iterations, for any $\epsilon < c_1 q^2 T$, Algorithm 1 is (ϵ, δ)-differentially private for any $\delta > 0$ if the noise scale (σ_{DP}) satisfies*

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

Method	MNIST		Fashion-MNIST		CIFAR-10		CIFAR-100	
	ϵ=1	ϵ=8	ϵ=1	ϵ=8	ϵ=1	ϵ=8	ϵ=1	ϵ=8
DPSGD	89.00 ± 0.06	88.95 ± 0.01	78.96 ± 0.05	79.04 ± 0.04	35.74 ± 0.26	47.74 ± 1.20	7.52 ± 0.49	18.27 ± 0.48
LP-DPSGD	88.99 ± 0.06	88.96 ± 0.01	79.03 ± 0.10	79.02 ± 0.10	35.84 ± 0.63	48.37 ± 0.36	7.55 ± 0.26	18.52 ± 0.27
InnerOuter	92.15 ± 0.15	**92.43 ± 0.06**	80.50 ±2.28	81.18 ± 1.56	11.55 ± 1.07	33.53 ± 0.52	1.13 ± 0.20	13.93 ± 0.40
DP-PMLF	**92.16 ± 0.05**	92.39 ± 0.07	**80.65 ± 1.17**	**81.93 ± 0.83**	**40.96 ± 1.18**	**51.47 ± 0.33**	**11.40 ± 0.21**	**23.15 ± 0.52**

Table 1: Test accuracy (%) comparison across datasets on ViT with fixed epoch (Epoch = 25 for MNIST and Fashion-MNIST, Epoch = 50 for CIFAR-10 and CIFAR-100) and different privacy budgets ϵ = 1 and 8.



(a) CNN-5 (b) ResNet-18 (c) ViT

Figure 2: Test accuracy (%) comparison across different models on CIFAR-10 with fixed privacy budget $\epsilon = 1$.

Proof. Let D and D' be any two neighbouring datasets where D' contains exactly one additional sample ξ compared to D. The global sensitivity of the clipped per-sample momentum satisfies $\|\text{clip}(v_t^{(\xi)}, C)\| \leq C$. Thus, by utilizing the Gaussian mechanism in Definition 3 and privacy amplification by subsampling (Balle, Barthe, and Gaboardi 2018), $\bar{v}_t$ in each training iteration is protected by $(\mathcal{O}(q/\sigma_{DP}), \delta/T)$-DP. Given that the subsequent low-pass filter is applied as a post-processing step (Lemma 1) in each training iteration, $\hat{m}_t$ is also protected by $(\mathcal{O}(q/\sigma_{DP}), \delta/T)$-DP. The moments accountant method (Abadi et al. 2016) then implies that, over T iterations, the overall privacy guarantee is (ϵ, δ)-DP provided if the stated condition on σ_{DP} holds. □

Experiments

Next we evaluate DP-PMLF through comprehensive experiments. Due to page limitation, details of the experimental setting and additional results are given in the appendix.

Experiment Setting

Dataset. We evaluate our approach on four image classification datasets, including MNIST (Deng 2012), Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017), CIFAR-10 (Krizhevsky and Hinton 2009), and CIFAR-100 (Krizhevsky and Hinton 2009), and four sentence classification datasets, including MNLI, QNLI, QQP, and SST-2 from the GLUE benchmark (Wang et al. 2018).

Baselines. We compare the test accuracy of DP-PMLF with vanilla DPSGD (Abadi et al. 2016) and two state-of-the-art methods introduced in Section : LP-DPSGD (Zhang et al. 2024a) and InnerOuter (Xiao et al. 2023).

Models. We utilized three models for image classification tasks: a 5-layer CNN (Zhang et al. 2024a), ResNet-18 (He et al. 2016), and the Vision Transformer (ViT) (Dosovitskiy et al. 2021). These models are initialized with random weights without pretraining. For sentence classification tasks, we fine-tune a pre-trained RoBERTa-base model (Liu et al. 2019).

Hyper-parameters. The parameter choices are detailed in the appendix and based on settings commonly used in the literature. All experiments are repeated five times, with the mean and standard deviation reported.

Privacy-Utility Trade-off

Image Classification We compare the performance of our method against baselines across different models and datasets with varying privacy budgets ϵ. We report the test accuracy for $\epsilon = 1$ and $\epsilon = 8$ for the ViT model in Table 1. As can be seen, DP-PMLF maintains its leading performance. For instance, on Fashion-MNIST, it achieves accuracies of about 80.65% at $\epsilon = 1$ and 81.93% at $\epsilon = 8$. On CIFAR-100, our approach maintains a 4–5% margin over the next-best baselines across both privacy budgets.

Under a high DP noise regime ($\epsilon = 1$), as shown in Table 1, the InnerOuter method shows a degradation in performance. This is because the InnerOuter method lacks normalization and suffers from excessive noise accumulation. In contrast, DP-PMLF utilizes both normalization and a low-pass filter allowing our approach to better control and filter DP noise, thereby achieving superior results.

Figure 2 presents the test accuracy on CIFAR-10 under $\epsilon = 1$ for three model architectures: CNN-5, Resnet-18, and ViT. In all cases, DP-PMLF consistently surpasses the baseline methods. For example, with CNN-5, DP-PMLF attains

Method	MNLI		QNLI		QQP		SST-2	
	ϵ=1	ϵ=8	ϵ=1	ϵ=8	ϵ=1	ϵ=8	ϵ=1	ϵ=8
DPSGD	51.36 ± 0.66	72.00 ± 0.23	65.59 ± 0.66	85.47 ± 0.78	71.20 ± 0.97	80.38 ± 0.37	76.19 ± 1.15	**90.83 ± 0.38**
LP-DPSGD	52.75 ± 0.62	71.48 ± 0.23	66.34 ± 0.94	85.44 ± 0.45	71.61 ± 0.71	80.55 ± 0.40	76.46 ± 0.24	89.24 ± 0.66
InnerOuter	48.45 ± 0.89	70.35 ± 0.38	69.46 ± 0.87	86.07 ± 0.58	70.27 ± 0.64	83.18 ± 0.34	76.49 ± 0.63	89.08 ± 0.43
DP-PMLF	**56.81 ± 0.74**	**75.56 ± 0.42**	**72.38 ± 0.62**	**86.96 ± 0.69**	**75.55 ± 1.16**	**83.42 ± 0.52**	**78.07 ± 0.96**	90.39 ± 1.03

Table 2: Test accuracy (%) comparison across GLUE benchmark subsets with different privacy budgets $\epsilon = 1, 8$.

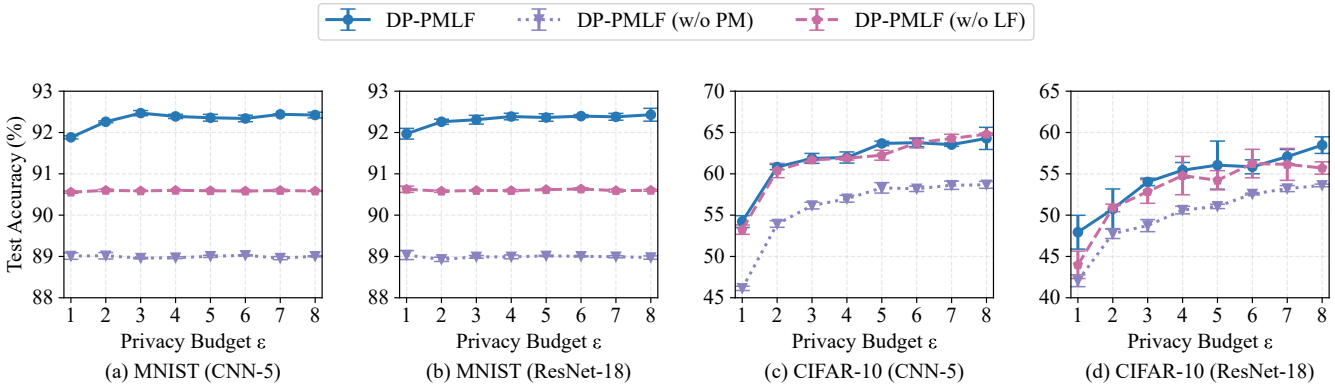


Figure 3: Test accuracy (%) for DP-PMLF and its two variants, which are DP-PMLF without Per-sample Momentum (DP-PMLF (w/o PM)) and DP-PMLF without Low-pass Filter (DP-PMLF (w/o LF)).

approximately 47% accuracy, exceeding the best baseline by around 9%. Similarly, with ResNet-18, DP-PMLF reaches nearly 50%, which is 1–2% higher than the strongest competitor. Finally, when using ViT, DP-PMLF achieves about 31% accuracy, compared to only 23% for the best baseline.

Sentence Classification To further assess the performance of our approach, we extend our evaluation to sentence classification tasks using four datasets from the GLUE benchmark, with results presented in Table 2. These experiments further demonstrate the effectiveness of DP-PMLF, which consistently shows a significant performance improvement over other baselines. When $\epsilon = 1$, our method surpasses the baselines by over 4% on MNLI and nearly 3% on QNLI. Although the performance gap decreases under a more relaxed privacy budget of $\epsilon = 8$, DP-PMLF still outperforms or remains highly competitive with the baseline methods. These results show the effectiveness of our approach for sentence classification tasks in a differential privacy setting.

Ablation Studies

We evaluated on two core components of DP-PMLF: per-sample momentum and the low-pass filter. Specifically, we denote DP-PMLF without per-sample momentum as *DP-PMLF (w/o PM)* and DP-PMLF without the low-pass filter as *DP-PMLF (w/o LF)*. We used MNIST and CIFAR-10 on the CNN-5 and Resnet-18 models with different privacy budget (ϵ) values ranging from 1 to 8. Figure 3 shows that DP-PMLF consistently outperforms DP-PMLF (w/o PM) for different ϵ values. This is because per-sample momentum effectively reduces clipping bias and thus narrows the neighborhood around the optimal convergence point.

Further, compared to DP-PMLF (w/o LF), our approach exhibits superior performance on MNIST by leveraging historical gradients. This refines the gradient descent direction and potentially accelerates convergence. For CIFAR-10, we adopt more complicated filter coefficients to enhance gradient signal smoothing and mitigate DP noise. This is advantageous when DP noise is large, e.g., when $\epsilon \leq 6$. However, when DP noise is relatively small ($\epsilon > 6$), excessive smoothing may lead to the loss of true gradient information, causing DP-PMLF to perform slightly worse than DP-PMLF (w/o LF). This is evident in Figure 3(c) where our approach is achieving approximately 0.5–0.7% less test accuracy than DP-PMLF (w/o LF) when $\epsilon > 6$.

Conclusion and Future Work

In this work, we propose a novel DPSGD variant that incorporates per-sample momentum and a low-pass filter to simultaneously reduce the effect of DP noise and clipping bias. We provide a theoretical proof of an improved convergence rate associated with a formal DP guarantee. Our experimental results show that our approach achieves higher utility in image and sentence classifications compared to the state-of-the-art DPSGD variants. In future work, we will investigate how to analyze our approach under some general assumptions such as non-convex Polyak-Łojasiewicz conditions (Karimi, Nutini, and Schmidt 2016), and (L_0, L_1)-smoothness (Zhang et al. 2020). Furthermore, we will develop an adaptive method to optimize the selection of the hyper-parameters in per-sample momentum and low-pass filtering. Finally, we will investigate how to apply this method to different domain applications, such as natural language processing tasks and reinforcement learning.

Acknowledgments

This research is supported by CSIRO Data61 Scholarship Program.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In *SIGSAC Conference on Computer and Communications Security*, CCS '16, 308–318. New York, NY, USA: Association for Computing Machinery. ISBN 9781450341394.

Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D. S.; Karthikesalingam, A.; King, D.; Ashrafian, H.; and Darzi, A. 2021. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*, 4(1): 65.

Amid, E.; Ganesh, A.; Mathews, R.; Ramaswamy, S.; Song, S.; Steinke, T.; Suriyakumar, V. M.; Thakkar, O.; and Thakurta, A. 2022. Public Data-Assisted Mirror Descent for Private Model Training. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, 517–535. Baltimore, Maryland, USA: PMLR.

Andrew, G.; Thakkar, O.; McMahan, B.; and Ramaswamy, S. 2021. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34: 17455–17466.

Asi, H.; Duchi, J.; Fallah, A.; Javidbakht, O.; and Talwar, K. 2021. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, 383–392. PMLR, Virtual: PMLR.

Bachute, M. R.; and Subhedar, J. M. 2021. Autonomous driving architectures: insights of machine learning and deep learning algorithms. *Machine Learning with Applications*, 6: 100164.

Balle, B.; Barthe, G.; and Gaboardi, M. 2018. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31.

Béthune, L.; Massena, T.; Boissin, T.; Bellet, A.; Mamalet, F.; Prudent, Y.; Friedrich, C.; Serrurier, M.; and Vigouroux, D. 2024. DP-SGD Without Clipping: The Lipschitz Neural Network Way. In *International Conference on Learning Representations (ICLR)*. Vienna, Austria: ICLR.

Boulemtafes, A.; Derhab, A.; and Challal, Y. 2020. A review of privacy-preserving techniques for deep learning. *Neurocomputing*, 384: 21–45.

Bu, Z.; Wang, Y.-X.; Zha, S.; and Karypis, G. 2024. Automatic clipping: Differentially private deep learning made easier and stronger. *Advances in Neural Information Processing Systems*, 36.

Chen, L.; Yue, D.; Ding, X.; Wang, Z.; Choo, K.-K. R.; and Jin, H. 2023a. Differentially private deep learning with dynamic privacy budget allocation and adaptive optimization. *IEEE Transactions on Information Forensics and Security*, 18(1): 4422–4435.

Chen, X.; Wang, X.; Zhang, K.; Fung, K.-M.; Thai, T. C.; Moore, K.; Mannel, R. S.; Liu, H.; Zheng, B.; and Qiu, Y. 2022. Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis*, 79: 102444.

Chen, X.; Wu, S. Z.; and Hong, M. 2020. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33: 13773–13782.

Chen, X.; Yao, L.; McAuley, J.; Zhou, G.; and Wang, X. 2023b. Deep reinforcement learning in recommender systems: A survey and new perspectives. *Knowledge-Based Systems*, 264: 110335.

Choquette-Choo, C. A.; Tramer, F.; Carlini, N.; and Papernot, N. 2021. Label-only membership inference attacks. In *International conference on machine learning*, 1964–1974. PMLR, Virtual: PMLR.

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; and Naor, M. 2006. Our data, ourselves: Privacy via distributed noise generation. In *International Conference on the Theory and Applications of Cryptographic Techniques*, 486–503. Lyon, France: Springer.

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

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

Fu, Z.; Niu, X.; and Maher, M. L. 2023. Deep learning models for serendipity recommendations: a survey and new perspectives. *ACM Computing Surveys*, 56(1): 1–26.

Golatkar, A.; Achille, A.; Wang, Y.-X.; Roth, A.; Kearns, M.; and Soatto, S. 2022. Mixed differential privacy in computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8376–8386. Louisiana, USA: IEEE.

Gorbunov, E.; Danilova, M.; and Gasnikov, A. 2020. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33: 15042–15053.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Karimi, H.; Nutini, J.; and Schmidt, M. 2016. Linear convergence of gradient and proximal-gradient methods under the

polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, 795–811. Springer.

Koloskova, A.; Hendrikx, H.; and Stich, S. U. 2023. Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, 17343–17363. PMLR, Hawai: PMLR.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Lee, J.; and Kifer, D. 2018. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. In *SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1656–1665. London, UK: ACM.

Li, T.; Zaheer, M.; Reddi, S.; and Smith, V. 2022. Private adaptive optimization with side information. In *International Conference on Machine Learning*, 13086–13105. PMLR, Hawai: PMLR.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Nguyen, N.-B.; Chandrasegaran, K.; Abdollahzadeh, M.; and Cheung, N.-M. 2023. Re-thinking model inversion attacks against deep neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16384–16393. Vancouver, Canada: IEEE.

Olatunji, I. E.; Nejdl, W.; and Khosla, M. 2021. Membership inference attack on graph neural networks. In *International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 11–20. IEEE, Virtual: IEEE.

Papernot, N.; Thakurta, A.; Song, S.; Chien, S.; and Erlingsson, Ú. 2021. Tempered sigmoid activations for deep learning with differential privacy. In *the AAAI Conference on Artificial Intelligence*, 9312–9321. Virtual: AAAI.

Shamsabadi, A. S.; and Papernot, N. 2023. Losing less: A loss for differentially private deep learning. In *Privacy Enhancing Technologies (PETs)*, 307–320. Lausanne, Switzerland: PoPETs.

Tanuwidjaja, H. C.; Choi, R.; Baek, S.; and Kim, K. 2020. Privacy-preserving deep learning on machine learning as a service—a comprehensive survey. *IEEE Access*, 8: 167425–167447.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, K.-C.; Fu, Y.; Li, K.; Khisti, A.; Zemel, R.; and Makhzani, A. 2021a. Variational model inversion attacks. *Advances in Neural Information Processing Systems*, 34: 9706–9719.

Wang, W.; Wang, T.; Wang, L.; Luo, N.; Zhou, P.; Song, D.; and Jia, R. 2021b. DPlis: Boosting Utility of Differentially Private Deep Learning via Randomized Smoothing. *Privacy Enhancing Technology*, 2021(4): 163–183.

Winder, S. 2002. *Analog and digital filter design*. Elsevier.

Xia, T.; Shen, S.; Yao, S.; Fu, X.; Xu, K.; Xu, X.; and Fu, X. 2023. Differentially private learning with per-sample adaptive clipping. In *the AAAI Conference on Artificial Intelligence*, volume 37, 10444–10452. Washington, DC, USA: AAAI.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 1–6.

Xiao, H.; Xiang, Z.; Wang, D.; and Devadas, S. 2023. A theory to instruct differentially-private learning via clipping bias reduction. In *IEEE Symposium on Security and Privacy (SP)*, 2170–2189. IEEE, San Francisco, CA, USA: IEEE.

Yu, D.; Naik, S.; Backurs, A.; Gopi, S.; Inan, H. A.; Kamath, G.; Kulkarni, J.; Lee, Y. T.; Manoel, A.; Wutschitz, L.; Yekhanin, S.; and Zhang, H. 2022. Differentially Private Fine-tuning of Language Models. In *International Conference on Learning Representations (ICLR)*. Virtual: ICLR.

Yu, D.; Zhang, H.; Chen, W.; and Liu, T. 2021a. Do not Let Privacy Overbill Utility: Gradient Embedding Perturbation for Private Learning. In *International Conference on Learning Representations (ICLR)*. Virtual: ICLR.

Yu, D.; Zhang, H.; Chen, W.; Yin, J.; and Liu, T.-Y. 2021b. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, 12208–12218. PMLR, Virtual: PMLR.

Yu, L.; Liu, L.; Pu, C.; Gursoy, M. E.; and Truex, S. 2019. Differentially private model publishing for deep learning. In *IEEE symposium on security and privacy (SP)*, 332–349. IEEE, San Francisco, CA, USA: IEEE.

Zaheer, M.; Reddi, S.; Sachan, D.; Kale, S.; and Kumar, S. 2018. Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 31.

Zhang, J.; He, T.; Sra, S.; and Jadbabaie, A. 2020. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In *International Conference on Learning Representations (ICLR)*, 1–12. Virtual: ICLR.

Zhang, X.; Bu, Z.; Hong, M.; and Razaviyayn, M. 2024a. DOPPLER: Differentially Private Optimizers with Low-pass Filter for Privacy Noise Reduction. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 1–26. Vancouver, BC, Canada: Curran Associates.

Zhang, X.; Bu, Z.; Wu, S.; and Hong, M. 2024b. Differentially Private SGD Without Clipping Bias: An Error-Feedback Approach. In *International Conference on Learning Representations (ICLR)*, 1–27. Vienna, Austria: ICLR.

Zhao, X.; Zhang, W.; Xiao, X.; and Lim, B. 2021. Exploiting explanations for model inversion attacks. In *IEEE/CVF international conference on computer vision*, 682–692. Montreal, BC, Canada: IEEE.

Zhou, Y.; Wu, S.; and Banerjee, A. 2021. Bypassing the Ambient Dimension: Private SGD with Gradient Subspace Identification. In *International Conference on Learning Representations (ICLR)*, 1–28. Virtual: ICLR.