

DP-FedAdamW: An Efficient Optimizer for Differentially Private Federated Large Models

Jin Liu¹ Ning Xi^{1*} Yinbin Miao¹ Junkang Liu²

¹ Xidian University ² Tianjin University

{jinliu9787, junkangliukk}@gmail.com; {nxi, ybmiao}@xidian.edu.cn

Abstract

Balancing convergence efficiency and robustness under Differential Privacy (DP) is a central challenge in Federated Learning (FL). Although AdamW accelerates training and fine-tuning in large-scale models, we find that directly applying it to Differentially Private FL (DPFL) suffers from three major issues: (i) data heterogeneity and privacy noise jointly amplify the variance of the second-moment estimator, (ii) DP perturbations bias the second-moment estimator, and (iii) DP amplifies AdamW’s sensitivity to local overfitting, worsening client drift. We propose DP-FedAdamW, the first AdamW-based optimizer for DPFL. It restores the effectiveness of AdamW under DP by stabilizing second-moment variance, removing DP-induced bias, and aligning local updates with the global descent direction to curb client drift. Theoretically, we establish an unbiased second-moment estimator and prove a linearly accelerated convergence rate without any heterogeneity assumption, while providing tighter (ϵ, δ) -DP guarantees. Our empirical results demonstrate the strong performance of DP-FedAdamW across language and vision Transformers, as well as ResNet-18. On Tiny-ImageNet (Swin-Base, $\epsilon=1$), DP-FedAdamW outperforms the state-of-the-art (SOTA) by 5.83%. The code is available at <https://github.com/junkangLiu0/DP-FedAdamW>.

1. Introduction

As data volumes surge and privacy concerns grow, centralized learning methods face increasing limitations in terms of security and collaboration. Federated Learning (FL) [40] enables decentralized training without sharing raw data, reducing privacy risks [36]. However, FL model updates remain vulnerable to attacks [9, 43, 62, 64]. Differential Privacy (DP) [7] provides formal protection by adding noise, but this increases gradient variance and hinders convergence [71], especially under non-IID data.

The emergence of large-scale models such as Swin

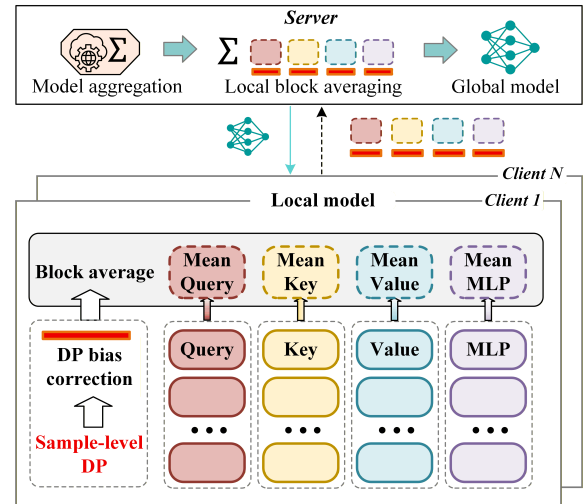


Figure 1. Illustration of our DP-FedAdamW. It aggregates the mean of block-wise bias-corrected second-moment estimates and performs local-global alignment to stabilize DPFL optimization.

Transformer [38], BERT [16], and GPT [2] has become dominant in vision and language tasks. These deep, parameter-heavy models are highly sensitive to optimization settings, which not only affect training stability but also overall performance. Most DPFL algorithms are built on SGD and adopt the DPSGD [1] mechanism for preserving privacy [13, 14, 44]. While effective for small-scale networks, DPSGD becomes inefficient for large Transformer models as the combination of gradient clipping and injected noise severely distorts gradients, amplifying instability, and slowing convergence.

In pursuit of more efficient optimizers, adaptive methods such as AdamW [39] have become state-of-the-art (SOTA) in deep learning. By decoupling weight decay from adaptive moment estimation, AdamW provides enhanced stability and generalization, making it a compelling choice for various machine learning tasks. These advantages make AdamW a promising candidate for DPFL, where privacy and robustness are crucial. However, our experiments reveal that AdamW does not always retain its usual benefits

*Corresponding author.

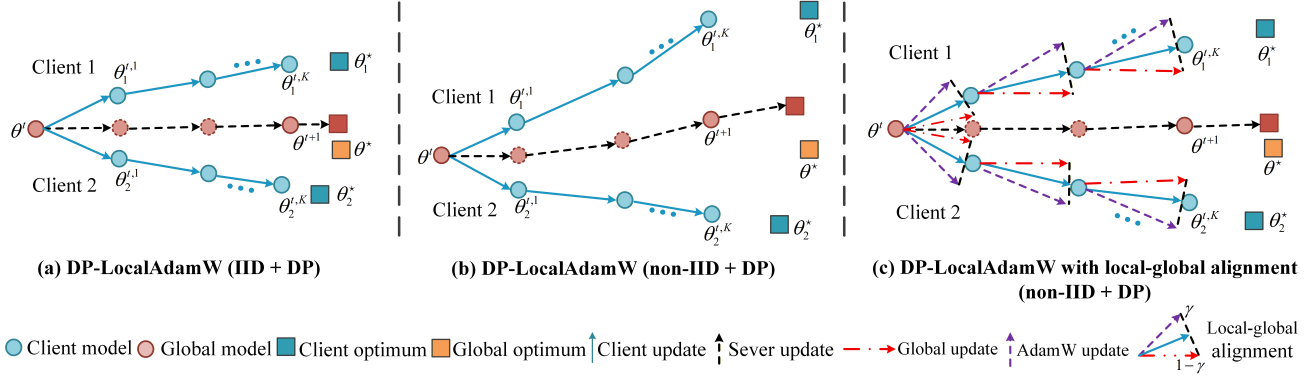


Figure 2. An illustration of local update in DP-LocalAdamW, which corrects client drift through global update guidance.

in the DPFL setting. In some cases, its performance is comparable to, or even worse than, that of SGD. As illustrated in Table 2, when AdamW is applied locally without any modifications (referred to as DP-LocalAdamW), its performance on the Swin-Tiny model (CIFAR-100, Dirichlet distribution $\alpha=0.1$) is worse than DP-FedAvg-LS. These findings suggest that AdamW struggles with the challenges posed by privacy constraints and federated settings.

Challenges. We identify three main challenges that contribute to the unexpected results. **(i) Second-moment estimator variance amplification.** Non-IID client data and DP noise jointly inflate the variance of the second-moment estimator, leading to unstable adaptive scaling in AdamW. **(ii) Bias in second-moment estimator.** Gradient clipping and noise injection introduce a systematic bias in the second-moment estimator of AdamW. **(iii) Client drift.** Under non-IID data, DP clipping and noise amplify AdamW’s sensitivity to local overfitting, further worsening client drift. All challenges are explored in more detail in Section 4. These findings raise a key question: **How can AdamW remain effectiveness in the presence of differential privacy and federated heterogeneity?**

To answer this, we propose DP-FedAdamW, a new differentially private adaptive optimizer that restores the effectiveness of AdamW under non-IID data and privacy noise. DP-FedAdamW (i) aggregates second-moment estimates in a block-wise manner to stabilize variance and improve communication efficiency, (ii) eliminates DP-induced bias through an unbiased second-moment correction, and (iii) enforces local-global update alignment to mitigate client drift, achieving more stable and accurate convergence.

Our contributions are summarized as follows:

- **Empirical importance of AdamW and challenges in DPFL.** We empirically demonstrate the effectiveness of AdamW in DPFL. Our analysis reveals three key challenges: high variance in second-moment estimator, bias in second-moment estimator, and high client drift.
- **DP-FedAdamW: A principled adaptive optimizer for DPFL.** We propose DP-FedAdamW, a communication-

efficient DPFL algorithm. It integrates global update estimation into local updates to suppress client drift, aggregates block-wise means of second-moment statistics across clients to stabilize variance, and introduces a bias-corrected mechanism to eliminate second-moment estimation bias caused by DP.

- **Theoretical and empirical guarantees.** We derive the first convergence guarantee for a DPFL adaptive optimizer that overcomes gradient heterogeneity, achieving a linear speedup rate of $\mathcal{O}(\sqrt{(L\Delta\sigma_l^2)/(SKT\tau^2)} + (L\Delta)/T + \sigma^2G_g^2/s^2R^2)$. Experiments on vision and language benchmarks demonstrate that DP-FedAdamW achieves a better privacy-utility trade-off and consistently outperforms SOTA DPFL baselines in convergence.

2. Related work

Differentially Private Federated Learning (DPFL). We adopt sample-level DP for FL. The most prevalent method, DP-FedAvg [44], applies per-sample clipping and Gaussian noise locally before aggregation. DP-SCAFFOLD [44] introduces control variates to alleviate client drift under privacy constraints. DP-FedAvg-LS [26] improves performance using Laplacian smoothing. Beyond SGD, DP-FedSAM [54] uses Sharpness-Aware Minimization (SAM [12]) to find flatter minima, improving robustness to DP noise. However, most methods still rely on SGD-style local updates, sometimes combined with SAM-like strategies, while adaptive optimization remains underexplored [3–6, 10, 11, 20, 23–25, 27, 28, 37, 41, 45–49, 52, 53, 55, 56, 58, 68–70]. In large-scale model optimization, AdamW has been shown to outperform SGD significantly in both convergence speed and final performance.

Adaptive optimization. Adaptive optimizers, such as Adam [17], AdamW [39], AMSGrad [50] adapt per-parameter step sizes from gradient history, achieving faster convergence than vanilla SGD and becoming standard for large models (e.g., Transformers). Among these, AdamW applies weight decay directly on parameters, preventing undesired interaction between regularization and moment

Table 1. Comparison with existing DPFL optimizers.

Method	DP-FedAvg	DP-SCAFFOLD	DP-FedSAM	DP-LocalAdamW	DP-FedAdamW
Optimizer Type	SGD	SGD	SAM	AdamW	AdamW
Scalability to Large Models	Limited	Limited	Moderate	Moderate	High
Client Drift	Strong	Weakened	Slightly weakened	Slightly weakened	Weakened
Communication Cost	1×	2×	2×	1×	1×

adaptation, thus improving convergence and generalization [29, 30, 63, 66, 67]. Adaptive optimization has also been explored in private settings [21].

Adaptive optimization in Federated Learning. A line of work introduces adaptive updates into FL. FedOpt [51] introduces server-side adaptivity, instantiating Adam and Yogi at the aggregator. FAFED [61] stabilizes training by aggregating clients’ first- and second-moment estimates of Adam. FedLADA [57] aggregates the second-moment estimate to reduce communication cost. However, these studies do not consider the effect of DP noise [31–33, 72–74].

Our contributions. To the best of our knowledge, adaptive optimization has not yet been systematically explored for large-model DPFL. Existing DPFL optimizers still struggle to train large models in a stable and efficient manner. To address this gap, we propose a method that stabilizes second-moment statistics via block-wise aggregation, removes DP-induced bias in the second-moment, and aligns local updates with an estimated global descent direction in a communication-efficient way. We summarize representative DPFL methods and their characteristics in Table 1.

3. DPFL framework

Consider a general DPFL system with N clients. The FL objective is to minimize the following population risk:

$$f(\theta) = \frac{1}{N} \sum_{i=1}^N f_i(\theta), f_i(\theta) := \mathbb{E}_{\xi_i \sim \mathcal{P}_i} [F_i(\theta; \xi_i)], \quad (1)$$

where f_i is the loss function of client i . ξ_i is drawn from distribution $\mathcal{P}_i$ in client i , and $\mathbb{E}[\cdot]$ denotes expectation with respect to the sample ξ_i .

In communication round t , the server selects S clients and distributes the global model θ^t . Client i computes the gradient for each sample j in local k iterations (with a mini-batch $\mathcal{D}_i^k$ of size $\lfloor sR \rfloor$, s is data sampling rate, R is number of client samples): $g_{ij} = \nabla f_i(\theta_i^{t,k}; \xi_{ij})$, and performs gradient clipping and Gaussian noise addition to satisfy DP:

$$\bar{g}_{ij} = g_{ij} / \max(1, \|g_{ij}\|_2 / C), \quad (2)$$

$$\tilde{g}_i^{t,k} \leftarrow \frac{1}{sR} \sum_{j \in \mathcal{D}_i^k} \bar{g}_{ij} + \frac{C}{sR} \mathcal{N}(0, \sigma^2 C^2 I) \quad (3)$$

where C is the clipping threshold, σ is the noise factor. After K local steps, the server updates the global model θ^{t+1} by adding the averaged local update to θ^t .

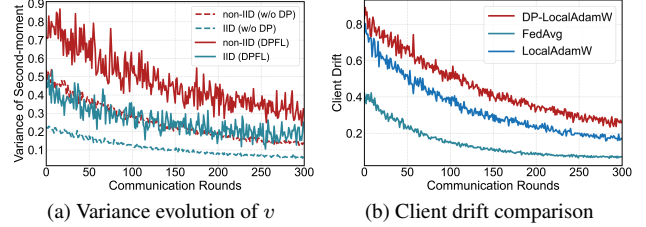


Figure 3. Training on CIFAR-100, Swin-Tiny, $\sigma=1$, $\alpha=0.1$. (a) Non-IID (DPFL) causes high variance in second-moment estimator across clients of DP-LocalAdamW. (b) DP-LocalAdamW suffers from more severe client drift than FedAvg and LocalAdamW.

4. Motivation and challenges

Prior work shows that AdamW can substantially improve convergence and generalization. However, our key observation is that, in DPFL, AdamW fails to exhibit its typical advantages. In fact, in transformer-based models, AdamW sometimes performs worse than SGD, indicating the benefits of AdamW are diminished when privacy constraints and the FL setup introduce additional complexities.

Challenge 1: Variance amplification of the second-moment estimator under non-IID and DP. In FL, non-IID data disperses client gradients, while DP noise further increases the variance of noisy gradients. This high variance accumulates in AdamW’s exponentially weighted moving averages of the first-moment $\mathbf{m}$ and second-moment $\mathbf{v}$:

$$\begin{aligned} \mathbf{m}_i^{t,k} &\leftarrow \beta_1 \mathbf{m}_i^{t,k-1} + (1 - \beta_1) \tilde{g}_i^{t,k}, \\ \mathbf{v}_i^{t,k} &\leftarrow \beta_2 \mathbf{v}_i^{t,k-1} + (1 - \beta_2) \tilde{g}_i^{t,k} \odot \tilde{g}_i^{t,k}, \end{aligned} \quad (4)$$

where $\beta_1=0.9$, $\beta_2=0.999$ are the exponential decay rates. $\odot$ represents the elementwise (Hadamard) product.

$$\text{Var}(\mathbf{v}^t) \approx \frac{(1 - \beta_2)^2}{1 - \beta_2^2} \text{Var}(\tilde{g}^t \odot \tilde{g}^t). \quad (5)$$

Since $\tilde{g}^t \odot \tilde{g}^t$ magnifies noise and $\beta_2 \approx 1$ induces slow averaging, the variance of $\mathbf{v}$ grows rapidly and dominates the optimization noise under DP and heterogeneity. Figure 3(a) shows the variance of $\mathbf{v}$ under non-IID+DP remains higher and converges much slower than in IID or non-DP cases.

Challenge 2: Systematic bias in the second-moment estimator introduced by DP. Under DP clipping and noise, directly using the perturbed gradient to estimate the second-moment produces a systematic bias. This bias is additive and distinct from the initialization bias addressed by

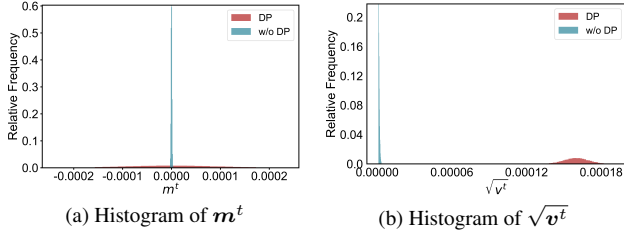


Figure 4. Histogram for DP-LocalAdamW, CIFAR-10, Swin-Tiny, $\sigma=1$, $\alpha=0.1$. (a) The distribution centers of m^t are aligned with or without DP, but the variance is larger with DP. (b) The distribution of $\sqrt{v^t}$ shows a significant difference, with the center of the distribution shifting approximately by $\sqrt{\sigma^2 C^2 / (sR)^2}$.

Adam’s exponential moving average, so that correction does not remove it in Figure 4. Recalling the definition of $v_i^{t,k}$ in Eqs. 4, 2, and 3, we have

$$\begin{aligned}\mathbb{E}[\tilde{g}_i^{t,k} \odot \tilde{g}_i^{t,k}] &= \mathbb{E}[\bar{g}_i^{t,k} \odot \bar{g}_i^{t,k}] + \sigma^2 C^2 / (sR)^2 I, \\ \mathbb{E}[v_i^{t,k}] &= \mathbb{E}[v_i^{t,k}]_{\text{w/o DP}} + (1 - \beta_2^k) \sigma^2 C^2 / (sR)^2 I.\end{aligned}\quad (6)$$

Challenge 3: Client drift caused by local adaptivity and DP mechanism. In FL, each client optimizes its local objective $f(\cdot)$, creating a gap between local and global optima. AdamW accelerates this by pushing clients toward their local optima, increasing client drift. In DPFL, gradient clipping and noise injection reduce effective sample sizes, making updates noisier and amplifying overfitting. This exacerbates client drift, hindering global model convergence. Figure 3(b) shows that DP-LocalAdamW results in higher client drift compared to FedAvg and LocalAdamW, where LocalAdamW refers to Federated AdamW without DP.

5. Methodology

To restore the effectiveness of AdamW in DPFL, we propose Differentially Private Federated AdamW (DP-FedAdamW). DP-FedAdamW is built upon a unified perspective: *the moment statistics of AdamW become distorted under non-IID data and DP*. Our design corrects this distortion along three complementary axes: parameter-block aggregation to stabilize second-moment variance, explicit removal of DP-induced bias in the second-moment, and local-global update alignment to reduce client drift.

5.1. Second-moment aggregation

To address **Challenge 1**, we stabilize the training process by aggregating the second-moment estimates of AdamW. Directly aggregating per-parameter second moments would require communicating an additional vector with the same dimensionality as the model parameters, nearly doubling the communication overhead. In addition, under DP noise, coordinate-wise second-moment estimates may fluctuate

Algorithm 1 DP-FedAdamW

```

1: Initial model  $\theta^0$ ; rounds  $T$ ; local steps  $K$ ; step size
    $\eta$ ; weight decay  $\lambda$ ; AdamW  $(\beta_1, \beta_2, \tau)$ ; clip norm  $C$ ;
   noise multiplier  $\sigma$ ;  $S$  clients per round
2: for  $t = 1$  to  $T$  do
3:   for selected clients  $i = 1, \dots, S$  in parallel do
4:      $\theta_i^{t,0} \leftarrow \theta^t$ ;  $m_i^{t,0} \leftarrow 0$ ;  $v_i^{t,0} \leftarrow \bar{v}^t$ 
5:     for  $k = 1$  to  $K$  do
6:       sample  $\mathcal{D}_i^k \subset \mathcal{D}_i$  of size  $\lfloor sR \rfloor$ 
7:       for each sample  $j \in \mathcal{D}_i^k$  do
8:          $g_{ij} \leftarrow \nabla f_i(\theta_i^{t,k}; \xi_{ij})$ 
9:          $\bar{g}_{ij} \leftarrow g_{ij} / \max(1, \|g_{ij}\|_2 / C)$ 
10:      end for
11:       $\tilde{g}_i^{t,k} \leftarrow \frac{1}{sR} \sum_{j \in \mathcal{D}_i^k} \bar{g}_{ij} + \frac{C}{sR} \mathcal{N}(0, \sigma^2 C^2 I)$ 
12:       $m_i^{t,k} \leftarrow \beta_1 m_i^{t,k-1} + (1 - \beta_1) \tilde{g}_i^{t,k}$ 
13:       $v_i^{t,k} \leftarrow \beta_2 v_i^{t,k-1} + (1 - \beta_2) \tilde{g}_i^{t,k} \odot \tilde{g}_i^{t,k}$ 
14:       $\hat{m}_i^{t,k} \leftarrow m_i^{t,k} / (1 - \beta_1^k)$ 
15:       $\hat{v}_i^{t,k} \leftarrow v_i^{t,k} / (1 - \beta_2^k)$ 
16:       $\vartheta_i^{t,k} \leftarrow 1 / (\sqrt{\hat{v}_i^{t,k} - (\frac{\sigma C}{sR})^2} + \tau)$ 
17:       $\theta_i^{t,k+1} \leftarrow \theta_i^{t,k} - \eta(\hat{m}_i^{t,k} \odot \vartheta_i^{t,k} + \gamma \Delta_G^t - \lambda \theta_i^{t,k})$ 
18:    end for
19:    Send  $(\theta_i^{t,K} - \theta_i^{t,0}, \bar{v}_i = \text{Block\_mean}(v_i^{t,K}))$ 
20:  end for
21:   $\Delta_G^{t+1} = \frac{-1}{SK\eta} \sum_{i=1}^S (\theta_i^{t,K} - \theta_i^{t,0})$ 
22:   $\theta^{t+1} = \theta^t + \frac{1}{S} \sum_{i=1}^S (\theta_i^{t,K} - \theta_i^{t,0})$ 
23:   $\bar{v}^{t+1} = \frac{1}{S} \sum_{i=1}^S \bar{v}_i$ 
24:  Broadcast  $(\theta^{t+1}, \bar{v}^{t+1}, \Delta_G^{t+1})$  to clients
25: end for

```

significantly, resulting in an unstable preconditioner and noisy adaptive scaling. We therefore adopt a block-wise aggregation scheme that communicates only *one* statistic per parameter block.

We adopt a parameter-block averaging strategy, where each block shares a single learning rate to capture local curvature effectively. Based on this idea, we partition the parameters into B blocks and compute the mean of second-moments within each block (line 19 in Algorithm 1):

$$\bar{v}_b = \frac{1}{|v_b|} \sum_{i \in v_b} v_i, \quad b = 1, \dots, B. \quad (7)$$

Block-wise averaging strategy. We align the blocks with the model architecture. Attention-related parameters are partitioned at the granularity of attention heads, while the remaining modules are partitioned at the granularity of whole layers. For each parameter block, we compute the mean of its second-moment estimates and transmit only this block-wise statistic:

- **Head-wise Q, K, V.** According to the number of attention heads in the model, each head-specific slice of the query (Q), key (K), and value (V) weight matrices is treated as an independent parameter block. For every such Q/K/V block, we compute and transmit the mean of its second-moment estimates.
- **attn.proj and MLP layers.** Each attention output projection layer (attn.proj) and each MLP layer are regarded as a single parameter block. For each layer, we aggregate the second-moment estimates of all parameters within the layer and transmit the block-wise mean.
- **Embedding and output layers.** Similarly, each Embedding layer and each output (e.g., classification) layer is treated as one parameter block. For these blocks, we also compute and transmit the mean of the corresponding second-moment estimates.

CNNs (e.g., ResNet). For convolutional networks, each convolutional layer or residual block is defined as one block. This reduces communication cost from per-parameter second-moments to B block-level values while preserving adaptive behavior across network components.

5.2. Unbiased second-moment correction

To address **Challenge 2**, we propose a bias-corrected mechanism (BC). The correction term subtracts the additive inflation caused by DP Gaussian noise from the second-moment estimate of DP-LocalAdamW, thereby improving adaptive scaling. Here, “unbiased” is with respect to the clipped-gradient second-moment. The update rule is given by (line 16 in Algorithm 1):

$$\vartheta_i^{t,k} = 1 / \left(\sqrt{\hat{v}_i^{t,k} - \left(\frac{\sigma C}{sR} \right)^2} + \tau \right) \quad (8)$$

Here, $\hat{v}_i^{t,k}$ is the exponential moving average of the noisy gradient second moment for parameter i , σ is the noise multiplier, C is the clipping norm, and sR is the batch size. The term $\frac{\sigma C}{sR}$ corresponds to the variance of the Gaussian noise added for differential privacy, and subtracting it removes the constant bias introduced by DP noise in the second-moment estimate. τ is the AdamW numerical stability term.

5.3. Local-global alignment

To mitigate **Challenge 3**, we augment the local AdamW update with an explicit alignment toward the global descent direction. Concretely, we modify the local update rule (line 17 in Algorithm 1) as

$$\theta_i^{t,k+1} = \theta_i^{t,k} - \eta \left(\hat{m}_i^{t,k} \odot \vartheta_i^{t,k} - \lambda \theta_i^{t,k} + \gamma \Delta_G^t \right), \quad (9)$$

where $\Delta_G^t = \frac{-1}{SK\eta} \sum_{i=1}^S (\theta_i^{t,K} - \theta_i^{t,0})$ is an empirical estimate of the global update over S participating clients.

Under non-IID data, the adaptive local gradient $\hat{m}_i^{t,k} \odot \vartheta_i^{t,k}$ is biased toward each client’s own objective. DP clipping distorts the magnitude and direction of these gradients,

and DP noise adds high-variance perturbations, pushing the local parameters $\theta_i^{t,k}$ toward different client-specific optima and amplifying the client drift identified in **Challenge 3**. The alignment term $\gamma \Delta_G^t$ softly regularizes each local step toward the global descent direction. It is nearly inactive when a client’s update already follows the global trend, but when non-IID data and DP operations drive the local direction away, the alignment term counteracts this deviation and steers the trajectory back toward the global path. As illustrated in Figure 2, this local-global alignment tightens the spread of client models, reduces cross-client update variance at aggregation, and yields a smoother, more stable global optimization trajectory under strong non-IID and DP.

6. Theoretical analysis

6.1. Convergence analysis

In this section, we provide the convergence theoretical analysis of the proposed DP-FedAdamW algorithm. We first introduce several standard assumptions for the non-convex function f .

Assumption 1 (Smoothness). *The non-convex f_i is a L -smooth function for all $i \in [m]$, i.e., $\|\nabla f_i(\theta_1) - \nabla f_i(\theta_2)\| \leq L\|\theta_1 - \theta_2\|$, for all $\theta_1, \theta_2 \in \mathbb{R}^d$.*

Assumption 2 (Bounded Stochastic Gradient). *$g_i^t = \nabla f_i(\theta_i^t, \xi_i^t)$ computed using a sampled mini-batch data ξ_i^t in the local client i is an unbiased estimator of ∇f_i with bounded variance, i.e., $\mathbb{E}_{\xi_i^t}[g_i^t] = \nabla f_i(\theta_i^t)$ and $\mathbb{E}_{\xi_i^t}\|g_i^t - \nabla f_i(\theta_i^t)\|^2 \leq \sigma_l^2$, for all $\theta_i^t \in \mathbb{R}^d$.*

Assumption 3 (Bounded Stochastic Gradient II). *Each element of the stochastic gradient g_i^t is bounded, i.e., $\|g_i^t\|_\infty = \|f_i(\theta_i^t, \xi_i^t)\|_\infty \leq G_g$, for all $\theta_i^t \in \mathbb{R}^d$ and any sampled mini-batch data ξ_i^t .*

Theorem 1 (Convergence for non-convex functions). *Under Assumptions 1, 2, and 3, if we take $g^0 = 0$, then DP-FedAdamW converges as follows*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(\theta^t)\|^2] \lesssim \mathcal{O} \left(\sqrt{\frac{L\Delta\sigma_l^2}{SKT\tau^2}} + \frac{L\Delta}{T} + \frac{\sigma^2 G_g^2}{s^2 R^2} \right). \quad (10)$$

Here $G_0 := \frac{1}{N} \sum_{i=1}^N \|\nabla f_i(\theta^0)\|^2$, $\Delta = f(\theta^0) - f^*$, S is the number of participating clients per round, σ the DP noise level, σ_l the stochastic gradient noise, K the number of local iterations, and T the total communication rounds.

The proof is provided in Appendix D.1. The convergence rate of DP-FedAdamW is faster than that of DP-LocalAdamW’s $\mathcal{O} \left(\sqrt{\frac{L(\sigma_l^2 + \sigma_g^2)}{SKT\tau^2}} + \frac{L\Delta}{T} + \frac{\sigma^2 G_g^2}{s^2 R^2} \right)$, and we do not need bounded heterogeneity assumption. This is due to the suppression of local drift by the global

update estimation Δ_G^t . Unlike prior adaptive FL optimizers [57] which rely on a heterogeneity assumption, our convergence result holds without such restrictions, making it more widely applicable.

6.2. Privacy analysis

Differential privacy (DP) [8] provides a formal framework for quantifying disclosure risk. We perform privacy accounting under Rényi DP (RDP) [42], which gives tight bounds under composition and subsampling.

Definition 1. (Sample-level DP). For any two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ in a single sample or record (either by addition or removal), a randomized mechanism $\mathcal{M} : \mathcal{X}^n \rightarrow \mathcal{Y}$ satisfies (ε, δ) -DP at the sample level if, for every possible output set $O \subseteq \mathbb{Y}$, the following holds:

$$\Pr[\mathcal{M}(\mathcal{D}) \in O] \leq e^\varepsilon \Pr[\mathcal{M}(\mathcal{D}') \in O] + \delta, \quad (11)$$

where ε is the privacy budget, δ is the failure probability.

Definition 2. (Rényi DP, [42]). For any order $\zeta \in (1, \infty)$, $\mathcal{M}$ satisfies (ζ, ε_R) -RDP if for any two neighboring datasets $\mathcal{D}$, $\mathcal{D}'$ that differ by a single sample, it has:

$$D_\zeta[\mathcal{M}(\mathcal{D}) \parallel \mathcal{M}(\mathcal{D}')] := \frac{1}{\zeta-1} \log \mathbb{E} \left[\left(\frac{p_{\mathcal{M}(\mathcal{D})}(Y)}{p_{\mathcal{M}(\mathcal{D}')} (Y)} \right)^\zeta \right] \leq \varepsilon_R, \quad (12)$$

The expectation $\mathbb{E}$ is taken over the output $Y \sim \mathcal{M}(\mathcal{D}')$.

We adopt standard RDP accounting following [44], the cumulative privacy over K local steps and T rounds is:

Theorem 2 (Privacy guarantee). For DP-FedAdamW, θ^T is (ε, δ) -DP towards a third party

$$\varepsilon = \mathcal{O} \left(\frac{s \sqrt{TK \log(2/\delta) \log(2T/\delta)}}{\sigma} \right), \quad (13)$$

Towards the server, the accumulated privacy guarantee is $(\varepsilon_s, \delta_s)$ -DP

$$\varepsilon_s = \varepsilon \sqrt{\frac{N}{l}}, \delta_s = \frac{\delta}{2} \left(\frac{1}{l} + 1 \right), \quad (14)$$

where N is the number of clients, l is the client sampling rate, and s is the client data sampling rate.

7. Experiments

7.1. Experimental setup

DPFL algorithms. We conduct experiments on six DPFL algorithms: DP-FedAvg [44], DP-SCAFFOLD [44], DP-FedAvg-LS [26], DP-FedSAM [54], DP-LocalAdamW and our DP-FedAdamW. All methods use the same noise multiplier σ for fair comparison. Under the same clipping norm and communication rounds, DP-FedSAM can consume approximately twice the privacy budget of other methods,

since it injects independent DP noise into two gradient evaluations at each local step. All methods are implemented in PyTorch and run on NVIDIA RTX 4090 GPUs.

Datasets and data partition. We evaluate DP-FedAdamW on seven datasets covering both vision and language tasks. For image classification, we adopt CIFAR-10, CIFAR-100 [18], and Tiny-ImageNet [19]. For natural language understanding, we consider four GLUE [60] benchmarks: SST-2, QQP, QNLI, and MNLI. We simulate non-IID client data using a Dirichlet partition [15], where each client's distribution is sampled from $\text{Dir}(\alpha)$ with $\alpha \in \{0.1, 0.6, 0.8\}$.

Model architectures. We study three representative architecture families. (1) convolutional neural networks (CNNs): GNResNet-18 evaluated on CIFAR-10/100. (2) vision transformers: ViT-Base for CIFAR-10/100 and Swin-Tiny/Base for CIFAR-10/100 and Tiny-ImageNet. (3) language transformer: RoBERTa-Base assessed on GLUE.

Configurations. For DP-FedAvg, DP-SCAFFOLD, DP-FedAvg-LS and DP-FedSAM, the learning rate η is selected from $\{10^{-2}, 3 \times 10^{-2}, 5 \times 10^{-2}, 10^{-1}, 3 \times 10^{-1}\}$ with a weight decay of 0.001. For DP-LocalAdamW and DP-FedAdamW, η is selected from $\{10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 8 \times 10^{-4}, 10^{-3}\}$ with weight decay 0.01 or 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We apply the cosine learning rate decay and set $\gamma=0.5$ for DP-FedAdamW, the weight decay $\lambda=0.01$. For ResNet-18, rounds $T=300$, batch size 50, local step $K=50$. For Transformers, rounds $T=100$, batch size 16, local step $K=20$. The noise multiplier is $\sigma=1$. The privacy failure probability is set to $\delta = 10^{-5}$ for CIFAR10/100 and $\delta = 10^{-6}$ for Tiny-ImageNet. Please refer to Appendix B.4 for more details of the experimental settings. All results are averaged over 5 runs with std reported.

7.2. Performance on vision datasets

ResNet-18 on CIFAR-10/100. Table 2 summarizes the results on CIFAR-10/100 under different heterogeneity levels ($\alpha=0.6$ and $\alpha=0.1$). DP-FedAdamW achieves the best performance across all settings, reaching 74.81% and 64.65% on CIFAR-10, and 39.56% and 33.55% on CIFAR-100. When the data heterogeneity is severe ($\alpha=0.1$), DP-FedAdamW surpasses the SOTA baselines (DP-FedAvg-LS for CIFAR-10 and DP-LocalAdamW for CIFAR-100) by 3.81% and 4.85%, respectively.

ViT-Base on CIFAR-10/100. We fine-tune ViT-Base on CIFAR-10/100, and the results are reported in Table 2 under $\alpha=0.6$ and $\alpha=0.1$. DP-FedAdamW consistently achieves the best performance. On CIFAR-100 with $\alpha=0.1$, it outperforms the strongest baseline DP-FedSAM by 1.29%. Under sample-level DPFL, DP-FedSAM operates under a much weaker privacy guarantee, as it uses approximately twice the privacy budget of the other methods.

Swin-Tiny/Base on CIFAR-10/100/Tiny-ImageNet. We fine-tune Swin-Tiny and Swin-Base on CIFAR-10, CIFAR-

Table 2. Averaged test accuracy (%) and privacy budget comparison on CIFAR-10 and CIFAR-100 using ResNet-18 and ViT-Base. Assuming DP-FedAvg is (ϵ, δ) -DP, all other methods share this budget except DP-FedSAM, which is approximately $(2\epsilon, \delta)$ -DP.

Method	CIFAR-10 (ResNet-18)		CIFAR-100 (ResNet-18)		CIFAR-10 (ViT-Base)		CIFAR-100 (ViT-Base)		Privacy Budget
	$\alpha=0.6$	$\alpha=0.1$	$\alpha=0.6$	$\alpha=0.1$	$\alpha=0.6$	$\alpha=0.1$	$\alpha=0.6$	$\alpha=0.1$	
DP-FedAvg	71.69 $\pm$ 0.37	60.38 $\pm$ 0.21	31.13 $\pm$ 0.57	25.04 $\pm$ 0.72	90.51 $\pm$ 0.37	78.85 $\pm$ 0.21	45.86 $\pm$ 0.57	13.41 $\pm$ 0.72	(ϵ, δ)
DP-SCAFFOLD	71.79 $\pm$ 0.44	59.04 $\pm$ 0.39	31.52 $\pm$ 0.36	24.94 $\pm$ 0.58	90.88 $\pm$ 0.52	80.69 $\pm$ 0.64	62.45 $\pm$ 0.57	64.69 $\pm$ 0.76	(ϵ, δ)
DP-FedAvg-LS	72.27 $\pm$ 0.57	60.84 $\pm$ 0.45	35.91 $\pm$ 0.54	28.68 $\pm$ 0.52	91.48 $\pm$ 0.47	88.48 $\pm$ 0.50	67.76 $\pm$ 0.58	65.85 $\pm$ 0.63	(ϵ, δ)
DP-FedSAM	67.39 $\pm$ 0.41	57.04 $\pm$ 0.46	30.36 $\pm$ 0.50	24.03 $\pm$ 0.66	91.76 $\pm$ 0.55	89.35 $\pm$ 0.65	70.24 $\pm$ 0.63	66.27 $\pm$ 0.65	$(2\epsilon, \delta)$
DP-LocalAdamW	71.73 $\pm$ 0.47	60.55 $\pm$ 0.53	34.92 $\pm$ 0.69	28.70 $\pm$ 0.81	91.16 $\pm$ 0.47	89.85 $\pm$ 0.53	68.21 $\pm$ 0.69	65.24 $\pm$ 0.81	(ϵ, δ)
DP-FedAdamW	74.81$\pm$0.19	64.65$\pm$0.24	39.56$\pm$0.42	33.55$\pm$0.50	92.02$\pm$0.49	90.30$\pm$0.24	71.59$\pm$0.42	67.56$\pm$0.50	(ϵ, δ)

Table 3. Averaged test accuracy (%) and privacy budget comparison on CIFAR-100 and Tiny-ImageNet using Swin-Tiny and Swin-Base.

Method	CIFAR-100 (Swin-T)		CIFAR-100 (Swin-B)		Tiny-ImageNet (Swin-T)		Tiny-ImageNet (Swin-B)		Privacy Budget
	$\alpha=0.6$	$\alpha=0.1$	$\alpha=0.6$	$\alpha=0.1$	$\alpha=0.6$	$\alpha=0.1$	$\alpha=0.6$	$\alpha=0.1$	
DP-FedAvg	36.96 $\pm$ 0.50	28.07 $\pm$ 0.31	65.61 $\pm$ 0.27	60.77 $\pm$ 0.32	18.82 $\pm$ 0.94	10.12 $\pm$ 1.36	52.04 $\pm$ 0.66	36.19 $\pm$ 0.74	(ϵ, δ)
DP-SCAFFOLD	41.68 $\pm$ 0.64	33.52 $\pm$ 0.65	68.03 $\pm$ 0.55	65.34 $\pm$ 0.49	21.91 $\pm$ 0.78	14.54 $\pm$ 0.72	53.78 $\pm$ 0.79	43.55 $\pm$ 0.60	(ϵ, δ)
DP-FedAvg-LS	41.95 $\pm$ 0.53	36.72 $\pm$ 0.65	68.80 $\pm$ 0.52	66.08 $\pm$ 0.56	23.03 $\pm$ 0.64	15.99 $\pm$ 0.68	55.83 $\pm$ 0.66	47.46 $\pm$ 0.57	(ϵ, δ)
DP-FedSAM	44.19 $\pm$ 0.48	38.43 $\pm$ 0.62	70.44 $\pm$ 0.57	66.28 $\pm$ 0.64	22.75 $\pm$ 0.65	16.86 $\pm$ 0.77	56.12 $\pm$ 0.74	48.08 $\pm$ 0.62	$(2\epsilon, \delta)$
DP-LocalAdamW	42.04 $\pm$ 0.63	35.51 $\pm$ 0.58	68.33 $\pm$ 0.33	65.28 $\pm$ 0.39	23.41 $\pm$ 0.87	15.81 $\pm$ 0.73	54.58 $\pm$ 0.60	46.52 $\pm$ 0.64	(ϵ, δ)
DP-FedAdamW	46.77$\pm$0.38	39.36$\pm$0.55	71.58$\pm$0.30	67.45$\pm$0.26	23.96$\pm$0.50	19.19$\pm$0.55	58.44$\pm$0.54	50.76$\pm$0.51	(ϵ, δ)

Table 4. Averaged test accuracy (%) on RoBERTa-Base.

Method	SST-2	QQP	QNLI	MNLI
DP-FedAvg	49.08 $\pm$ 0.12	63.18 $\pm$ 0.50	49.46 $\pm$ 0.44	35.45 $\pm$ 0.20
DP-SCAFFOLD	52.35 $\pm$ 0.18	66.87 $\pm$ 0.42	52.62 $\pm$ 0.52	38.74 $\pm$ 0.32
DP-FedAvg-LS	68.38 $\pm$ 0.50	72.59 $\pm$ 0.48	69.22 $\pm$ 0.68	50.46 $\pm$ 0.55
DP-FedSAM	70.95 $\pm$ 0.62	75.66 $\pm$ 0.53	70.54 $\pm$ 0.63	56.88 $\pm$ 0.56
DP-LocalAdamW	91.04 $\pm$ 0.53	82.33 $\pm$ 0.59	86.07 $\pm$ 0.47	75.20 $\pm$ 0.44
DP-FedAdamW	91.17$\pm$0.32	83.34$\pm$0.39	86.33$\pm$0.34	78.68$\pm$0.38

100, and Tiny-ImageNet, and report the results in Table 3 and Table 10 in Appendix B.1. DP-FedAdamW consistently achieves the best accuracy across all settings. On CIFAR-100 with Swin-Tiny at $\alpha=0.6$, it reaches 46.77%, outperforming DP-FedSAM by 2.58%. On Tiny-ImageNet with Swin-Base at $\alpha=0.1$, it achieves 50.76%, exceeding DP-FedSAM by 2.68%. Figure 5(c,d) illustrates the advantage of DP-FedAdamW on Swin-Tiny.

7.3. Performance on language datasets

RoBERTa-Base on GLUE. We further evaluate our method on four GLUE tasks (SST-2, QQP, QNLI, MNLI) by fine-tuning RoBERTa-Base under $\alpha=0.8$, and report the results in Table 4. DP-FedAdamW attains the best accuracy on all tasks, reaching 91.17%, 83.34%, 86.33%, and 78.68%, respectively. On MNLI, DP-FedAdamW outperforms the SOTA DP-LocalAdamW by 3.48%, showing that the benefits of DP-FedAdamW extend from vision benchmarks to language understanding in the DPFL setting.

7.4. Discussion of AdamW under DP in FL

Performance under different privacy budgets ϵ . Table 5 reports the results of Swin-Base on CIFAR-10, CIFAR-

100, and Tiny-ImageNet under $\epsilon=1, 2, 3$ with $\alpha=0.1$. As expected, all methods benefit from larger ϵ (weaker privacy). The performance gap between DP-FedAdamW and other methods is most pronounced when ϵ is small, i.e., under stronger privacy constraints. For instance, on CIFAR-10, DP-FedAdamW achieves 77.50% at $\epsilon=1$, surpassing the best baseline by 5.93%. These results show that DP-FedAdamW achieves a superior utility–privacy trade-off.

Noise multiplier σ . Results are provided in Appendix B.1.

7.5. Ablation study

Effect of each component. DP-FedAdamW differs from DP-LocalAdamW in three aspects: (i) block-wise second-moment aggregation (Agg); (ii) DP bias correction on the second-moment (BC); (iii) local–global alignment term (Align). We construct several variants by removing one component at a time, while keeping all other settings fixed.

Table 6 shows that each component yields a gain, and their combination achieves the best performance. Bias correction (BC) improves robustness to DP noise; local-global alignment stabilizes optimization under non-IID data and DP; and block-wise aggregation provides an additional improvement with negligible communication overhead.

Impact of aggregation strategy. Table 7 shows that our block-wise averaging strategy, Agg-mean- v , achieves the best balance between accuracy and communication cost. Agg-mean- v achieves comparable gains with only $\mathcal{O}(B)$ communication, where B is the number of blocks, demonstrating its scalability and ability to stabilize updates even under strict communication budgets.

Effect of the alignment coefficient γ . Table 8 varies γ on Swin-Base ($\alpha=0.1$). Increasing γ from 0.0 to 0.5 im-

Table 5. Effect of privacy budgets (ϵ) on test accuracy (%) using Swin-Base, $\alpha = 0.1$.

Method	CIFAR-10			CIFAR-100			Tiny-ImageNet		
	$\epsilon=1$	$\epsilon=2$	$\epsilon=3$	$\epsilon=1$	$\epsilon=2$	$\epsilon=3$	$\epsilon=1$	$\epsilon=2$	$\epsilon=3$
DP-FedAvg	66.52 $\pm$ 0.89	72.63 $\pm$ 0.55	88.02 $\pm$ 0.76	19.66 $\pm$ 1.52	42.98 $\pm$ 0.60	57.73 $\pm$ 0.55	20.65 $\pm$ 0.77	37.48 $\pm$ 0.54	45.32 $\pm$ 0.47
DP-SCAFFOLD	66.39 $\pm$ 0.68	72.81 $\pm$ 0.53	88.07 $\pm$ 0.66	22.38 $\pm$ 0.67	44.52 $\pm$ 0.68	59.85 $\pm$ 0.52	20.93 $\pm$ 0.66	38.40 $\pm$ 0.51	47.06 $\pm$ 0.30
DP-FedAvg-LS	71.57 $\pm$ 1.02	75.02 $\pm$ 0.59	88.96 $\pm$ 0.54	25.77 $\pm$ 0.61	49.54 $\pm$ 0.59	62.46 $\pm$ 0.53	26.89 $\pm$ 0.46	46.57 $\pm$ 0.44	48.83 $\pm$ 0.21
DP-FedSAM	53.51 $\pm$ 1.39	59.33 $\pm$ 0.86	76.82 $\pm$ 0.97	12.33 $\pm$ 0.64	28.32 $\pm$ 0.49	46.93 $\pm$ 0.50	10.88 $\pm$ 0.95	26.76 $\pm$ 0.78	41.69 $\pm$ 0.45
DP-LocalAdamW	66.57 $\pm$ 1.35	73.04 $\pm$ 0.58	88.83 $\pm$ 0.71	24.69 $\pm$ 0.88	47.26 $\pm$ 0.59	62.18 $\pm$ 0.61	28.40 $\pm$ 0.72	46.62 $\pm$ 0.37	48.55 $\pm$ 0.28
DP-FedAdamW	77.50$\pm$0.54	80.37$\pm$0.58	89.98$\pm$0.55	30.06$\pm$0.41	52.11$\pm$0.54	65.47$\pm$0.42	34.23$\pm$0.50	50.95$\pm$0.26	52.53$\pm$0.22

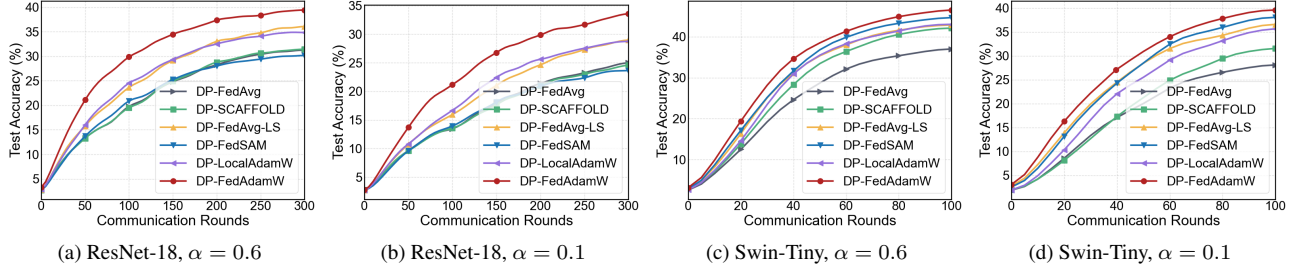


Figure 5. Test accuracy (%) on CIFAR-100 using ResNet-18 and Swin-Tiny under the Dirichlet $\alpha = 0.6$ and $\alpha = 0.1$ settings.

Table 6. Ablation on the main components of DP-FedAdamW on Swin-Base. We report test accuracy (%) on CIFAR-100 and Tiny-ImageNet under non-IID setting ($\alpha = 0.1$). “Agg” denotes block-wise second-moment aggregation, “BC” the DP bias correction, and “Align” the local–global alignment term.

Method	CIFAR-100	Tiny-ImageNet
DP-LocalAdamW	65.28	46.52
w/o Agg	66.41	47.35
w/o BC	67.02	48.11
w/o Align	66.37	47.82
DP-FedAdamW (full)	67.42	50.73

Table 7. Ablation study of moment aggregation strategies of DP-LocalAdamW on CIFAR-100 with Swin-Base under $\alpha=0.1$.

Strategy	CIFAR-100	Tiny-ImageNet	Comm(↑)
NoAgg	66.21 $\pm$ 0.11	49.12 $\pm$ 0.12	5.7M
Agg- $\mathbf{v}$	67.52 $\pm$ 0.12	50.88 $\pm$ 0.12	11.4M
Agg-mean- $\mathbf{v}$	67.45$\pm$0.10	50.76$\pm$0.10	5.7M

Table 8. Impact of γ using Swin-Base ($\alpha=0.1$).

γ	0.00	0.25	0.50	0.75	1.00
CIFAR-100 (%)	66.37	67.01	67.45	67.10	66.52
Tiny-ImageNet (%)	47.82	49.36	50.76	50.21	49.18

proves accuracy on CIFAR-100 and Tiny-ImageNet, indicating that a moderate alignment term effectively counteracts client drift and stabilizes optimization under DP noise. Further increasing γ beyond 0.5 leads to a slight degradation, as overly strong alignment over-constrains local updates. We therefore set $\gamma=0.5$ as the default.

Table 9. Comparison with DP variants of federated adaptive optimizers under the same DP settings (Swin-Base, $\alpha=0.1$, $\sigma=1$).

Method	CIFAR-100	Tiny-ImageNet
DP-FedOpt(Adam) [51]	64.54 $\pm$ 0.56	44.81 $\pm$ 0.62
DP-FAFED [61]	65.75 $\pm$ 0.54	46.10 $\pm$ 0.67
DP-FedLADA [57]	66.03 $\pm$ 0.61	47.85 $\pm$ 0.57
DP-FedAdamW (ours)	67.45$\pm$0.26	50.76$\pm$0.51

Comparison with DP variants of federated adaptive optimizers. As shown in Table 9, on Swin-Base with Dirichlet $\alpha=0.1$, DP noise $\sigma=1$, simply adding DP to existing federated adaptive optimizers yields limited gains. DP-FAFED and DP-FedLADA slightly outperform DP-FedOpt(Adam) on both CIFAR-100 and Tiny-ImageNet. Our DP-FedAdamW consistently achieves the best performance, with roughly 1.42% on CIFAR-100 and 2.91% on Tiny-ImageNet compared to the SOTA DP baseline.

8. Conclusion

In this work, we revisited AdamW for differentially private federated learning and revealed that a naïve adoption leads to amplified second-moment variance, DP-induced bias, and pronounced client drift under small-batch local training. Building on these insights, we introduced DP-FedAdamW, which integrates block-wise second-moment aggregation, explicit DP bias correction, and a local–global alignment mechanism, with non-convex convergence and sample-level privacy guarantees. Experiments on vision and language benchmarks demonstrate DP-FedAdamW consistently surpasses SOTA DPFL baselines under the same privacy budgets, especially in highly non-IID regimes.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. U24A20243, No. 62232013), the National Key Research and Development Program of China (No. 2024YFB3108700), the Fundamental Research Funds for the Central Universities (No. YJSJ26003), and the Program of China Scholarship Council (CSC).

References

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 1
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [3] Weixin An, Yuanyuan Liu, Fanhua Shang, and Hongying Liu. Robust and faster zeroth-order minimax optimization: complexity and applications. *Advances in Neural Information Processing Systems*, 37:37050–37069, 2024. 2
- [4] Weixin An, Yuanyuan Liu, Fanhua Shang, Hongying Liu, and Licheng Jiao. Des-inspired accelerated unfolded linearized admm networks for inverse problems. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3): 5319–5333, 2024.
- [5] Qiao Ding, Nanyu Li, Heng Ding, Jian Wang, Tao Li, Yongqing Chen, Yantuan Xian, and Junyang Chen. Traffic prediction and load balancing routing algorithm based on deep q-network for sd-iot. *Advanced Engineering Informatics*, 68:103596, 2025.
- [6] Qiao Ding, Heng Ding, Jian Wang, Yantuan Xian, Tao Li, Nanyu Li, Tao Fang, and Junyang Chen. Enhancing news classification: Domain-specific guided pretraining based on adaptive selective masking. *Knowledge-Based Systems*, page 115516, 2026. 2
- [7] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006. 1
- [8] 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. 6
- [9] Mingyuan Fan, Cen Chen, Chengyu Wang, Xiaodan Li, and Wenmeng Zhou. Refiner: Data refining against gradient leakage attacks in federated learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3005–3024, 2025. 1
- [10] Chen Feng, Ziquan Liu, Zhuo Zhi, Ilija Bogunovic, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. PROSAC: Provably safe certification for machine learning models under adversarial attacks. In *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI) [Oral]*, 2025. 2
- [11] Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. 2
- [12] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021. 2
- [13] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017. 1
- [14] Xiaolan Gu, Ming Li, and Li Xiong. {DP-BREM};{Differentially-Private} and {Byzantine-Robust} federated learning with client momentum. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3065–3082, 2025. 1
- [15] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 6
- [16] Jacob Devlin Ming-Wei Chang Kenton, Lee Kristina Toutanova, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*. Minneapolis, Minnesota, 2019. 1
- [17] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2
- [18] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. Technical Report. 6, 3
- [19] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge, 2015. Stanford CS231N. 6, 3
- [20] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 2, 1
- [21] Tian Li, Manzil Zaheer, Sashank Reddi, and Virginia Smith. Private adaptive optimization with side information. In *International Conference on Machine Learning*, pages 13086–13105. PMLR, 2022. 3, 1
- [22] Tian Li, Manzil Zaheer, Ken Liu, Sashank J. Reddi, Hugh Brendan McMahan, and Virginia Smith. Differentially private adaptive optimization with delayed preconditioners. In *The Eleventh International Conference on Learning Representations*, 2023. 1
- [23] Yuqi Li, Kai Li, Xin Yin, Zhifei Yang, Junhao Dong, Zeyu Dong, Chuanguang Yang, Yingli Tian, and Yao Lu. Sep-prune: Structured pruning for efficient deep speech separation. *arXiv preprint arXiv:2505.12079*, 2025. 2
- [24] Yuqi Li, Chuanguang Yang, Hansheng Zeng, Zeyu Dong, Zhulin An, Yongjun Xu, Yingli Tian, and Hao Wu. Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2025.
- [25] Yuqi Li, Siwei Meng, Chuanguang Yang, Weilun Feng, Junming Liu, Zhulin An, Yikai Wang, and Yingli Tian. A comprehensive survey of interaction techniques in 3d scene generation. *Authorea Preprints*, 2026. 2

- [26] Zhicong Liang, Bao Wang, Quanquan Gu, Stanley Osher, and Yuan Yao. Differentially private federated learning with laplacian smoothing. *Applied and Computational Harmonic Analysis*, 72:101660, 2024. 2, 6
- [27] Bangyan Liao, Zhenjun Zhao, Haoang Li, Yi Zhou, Yingping Zeng, Hao Li, and Peidong Liu. Convex relaxation for robust vanishing point estimation in manhattan world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15823–15832, 2025. 2
- [28] Dingyuan Liu, Qiannan Shen, and Jiacy Liu. The health-wealth gradient in labor markets: Integrating health, insurance, and social metrics to predict employment density. *Computation*, 14(1):22, 2026. 2
- [29] Junkang Liu, Fanhua Shang, Yuanyuan Liu, Hongying Liu, Yuangang Li, and YunXiang Gong. Fedbcgd: Communication-efficient accelerated block coordinate gradient descent for federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2955–2963, 2024. 3, 1
- [30] Junkang Liu, Yuanyuan Liu, Fanhua Shang, Hongying Liu, Jin Liu, and Wei Feng. Improving generalization in federated learning with highly heterogeneous data via momentum-based stochastic controlled weight averaging. In *Forty-second International Conference on Machine Learning*, 2025. 3
- [31] Junkang Liu, Fanhua Shang, Yuxuan Tian, Hongying Liu, and Yuanyuan Liu. Consistency of local and global flatness for federated learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3875–3883, 2025. 3, 1
- [32] Junkang Liu, Fanhua Shang, Junchao Zhou, Hongying Liu, Yuanyuan Liu, and Jin Liu. Fedmuon: Accelerating federated learning with matrix orthogonalization. *arXiv preprint arXiv:2510.27403*, 2025. 1
- [33] Junkang Liu, Fanhua Shang, Kewen Zhu, Hongying Liu, Yuanyuan Liu, and Jin Liu. Fedadamw: A communication-efficient optimizer with convergence and generalization guarantees for federated large models. *arXiv preprint arXiv:2510.27486*, 2025. 3, 1
- [34] Junkang Liu, Yuxuan Tian, Fanhua Shang, Yuanyuan Liu, Hongying Liu, Junchao Zhou, and Daorui Ding. Dp-fedpgn: Finding global flat minima for differentially private federated learning via penalizing gradient norm. *arXiv preprint arXiv:2510.27504*, 2025. 1
- [35] Junkang Liu, Fanhua Shang, Hongying Liu, Jin Liu, Weixin An, and Yuanyuan Liu. Taming preconditioner drift: Unlocking the potential of second-order optimizers for federated learning on non-iid data, 2026. 1
- [36] Jin Liu, Ning Xi, Yinbin Miao, and Junkang Liu. Rethinking loRA for privacy-preserving federated learning in large models. In *The Fourteenth International Conference on Learning Representations*, 2026. 1
- [37] Yuanyuan Liu, Fanhua Shang, Weixin An, Hongying Liu, and Zhouchen Lin. Kill a bird with two stones: Closing the convergence gaps in non-strongly convex optimization by directly accelerated svrg with double compensation and snapshots. In *International Conference on Machine Learning*, pages 14008–14035. PMLR, 2022. 2
- [38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1
- [39] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5):5, 2017. 1, 2
- [40] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1
- [41] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):76–87, 2025. 2
- [42] Ilya Mironov. Rényi differential privacy. In *Proc. IEEE computer security foundations symposium (CSF)*, pages 263–275, 2017. 6
- [43] Xutong Mu, Ke Cheng, Yulong Shen, Xiaoxiao Li, Zhao Chang, Tao Zhang, and Xindi Ma. Feddmc: Efficient and robust federated learning via detecting malicious clients. *IEEE Transactions on Dependable and Secure Computing*, 21(6):5259–5274, 2024. 1
- [44] Maxence Noble, Aurélien Bellet, and Aymeric Dieuleveut. Differentially private federated learning on heterogeneous data. In *International conference on artificial intelligence and statistics*, pages 10110–10145. PMLR, 2022. 1, 2, 6
- [45] Zihao Peng, Boyuan Li, Le Li, Shengbo Chen, Guanghui Wang, Hong Rao, and Cong Shen. Performance optimization for noise interference privacy protection in federated learning. *IEEE Transactions on Cognitive Communications and Networking*, 9(5):1322–1339, 2023. 2
- [46] Zihao Peng, Nan Zou, Jiandian Zeng, Boyuan Li, Shengbo Chen, Tian Wang, and Weijia Jia. Dynamic grouping and aggregation weight optimization for hierarchical federated learning with quantization. *IEEE Transactions on Networking*, pages 1–1, 2026.
- [47] Xin Qi, Meixuan Li, Sijin Zhou, Wei Feng, and Zhuang Qi. Federated learning for science: A survey on the path to a trustworthy collaboration ecosystem. *Authorea Preprints*, 2025.
- [48] Xin Qi, Tao Xu, Chengrun Dang, Zhuang Qi, Lei Meng, and Han Yu. Federated learning in oncology: bridging artificial intelligence innovation and privacy protection. *Information Fusion*, page 104154, 2026.
- [49] Zhuang Qi, Lei Meng, Zitan Chen, Han Hu, Hui Lin, and Xiangxu Meng. Cross-silo prototypical calibration for federated learning with non-iid data. In *Proceedings of the 31st ACM international conference on multimedia*, pages 3099–3107, 2023. 2
- [50] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. 2
- [51] Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and

- Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. 3, 8
- [52] Qiannan Shen and Jing Zhang. Ai-enhanced disaster risk prediction with explainable shap analysis: A multi-class classification approach using xgboost. 2025. Preprint, Version 1, posted December 31, 2025. 2
- [53] Qiannan Shen and Jing Zhang. Mftformer: Meteorological-frequency-temporal transformer with block-aligned fusion for traffic flow prediction. *Research Square*, 2026. Preprint, doi:10.21203/rs.3.rs-8770196/v1. 2
- [54] Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, and Dacheng Tao. Make landscape flatter in differentially private federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24552–24562, 2023. 2, 6
- [55] Wenxi Sun, Zhichun Qi, and Qiannan Shen. High-recall deep learning: A gated recurrent unit approach to bank account fraud detection on imbalanced data. pages 207–212, 2025. 2
- [56] Wenxi Sun, Qiannan Shen, Yijun Gao, Qinkai Mao, Tong-song Qi, and Shuo Xu. Objective over architecture: Fraud detection under extreme imbalance in bank account opening. *Computation*, 13(12):290, 2025. 2
- [57] Yan Sun, Li Shen, Hao Sun, Liang Ding, and Dacheng Tao. Efficient federated learning via local adaptive amended optimizer with linear speedup. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14453–14464, 2023. 3, 6, 8
- [58] Zhonglin Sun, Chen Feng, Ioannis Patras, and Georgios Tzimiropoulos. LAFS: Landmark-based Facial Self-supervised Learning for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [59] Qiaoyue Tang, Frederick Shpilevskiy, and Mathias Lécuyer. Dp-adambc: Your dp-adam is actually dp-sgd (unless you apply bias correction). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15276–15283, 2024. 1
- [60] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018. 6, 3
- [61] Xidong Wu, Feihu Huang, Zhengmian Hu, and Heng Huang. Faster adaptive federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10379–10387, 2023. 3, 8
- [62] Hongyong Xiao, Xutong Mu, and Ke Cheng. Fedrma: A robust federated learning resistant to multiple poisoning attacks. *Journal of Networking and Network Applications*, 4(1):31–38, 2024. 1
- [63] Shuo Xie and Zhiyuan Li. Implicit bias of adamw: ℓ_∞ -norm constrained optimization. In *International Conference on Machine Learning*, pages 54488–54510. PMLR, 2024. 3
- [64] Xiangrui Xu, Zhize Li, Yufei Han, Bin Wang, Jiqiang Liu, and Wei Wang. From risk to resilience: Towards assessing and mitigating the risk of data reconstruction attacks in federated learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3141–3160, 2025. 1
- [65] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. {PrivateFL}: Accurate, differentially private federated learning via personalized data transformation. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1595–1612, 2023. 1
- [66] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresight-drive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025. 3
- [67] Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548*, 2025. 3
- [68] Zhenjun Zhao. Balf: Simple and efficient blur aware local feature detector. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3362–3372, 2024. 2
- [69] Zhenjun Zhao and Ben M Chen. Benchmark for evaluating initialization of visual-inertial odometry. In *2023 42nd Chinese Control Conference (CCC)*, pages 3935–3940. IEEE, 2023.
- [70] Zhenjun Zhao, Heng Yang, Bangyan Liao, Yingping Zeng, Shaocheng Yan, Yingdong Gu, Peidong Liu, Yi Zhou, Haoang Li, and Javier Civera. Advances in global solvers for 3d vision. *arXiv preprint arXiv:2602.14662*, 2026. 2
- [71] Lele Zheng, Xiang Wang, Tao Zhang, Yang Cao, Ke Cheng, and Yulong Shen. Differentially private subspace fine-tuning for large language models. *arXiv preprint arXiv:2601.11113*, 2026. 1
- [72] Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. Odtrack: Online dense temporal token learning for visual tracking. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7588–7596, 2024. 3
- [73] Yaozong Zheng, Bineng Zhong, Qihua Liang, Ning Li, and Shuxiang Song. Decoupled spatio-temporal consistency learning for self-supervised tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10635–10643, 2025.
- [74] Yaozong Zheng, Bineng Zhong, Qihua Liang, Shengping Zhang, Guorong Li, Xianxian Li, and Rongrong Ji. Towards universal modal tracking with online dense temporal token learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3