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

Supplementary Material

A. More Related Work

Adaptive optimization in centralized DP. Recent studies have investigated adaptive optimization under centralized differential privacy. AdaDPS [21] improves private adaptive optimization by exploiting non-sensitive side information for preconditioning. DP^2-RMSProp [22] instead uses delayed preconditioners to preserve adaptivity without auxiliary data. DP-AdamBC [59] further identifies the bias introduced by DP noise in Adam's second-moment estimation and corrects it to restore Adam-style adaptive updates.

Optimization in FL. FedBCGD [29] proposes an accelerated block coordinate gradient descent framework for FL. FedAdamW [33] introduces a communication-efficient AdamW-style optimizer tailored for federated large models. FedNSAM [31] studies the consistency relationship between local and global flatness in FL. FedMuon [32] accelerates federated optimization via matrix orthogonalization. FedPAC [35] mitigates preconditioner drift to unlock the potential of second-order optimizers. FedProx [20] addresses heterogeneity in federated learning by introducing a proximal term into the local objective to limit client drift. In contrast, our DP-FedAdamW performs local-global alignment in the update rule rather than through a proximal objective.

Other DPFL methods. PrivateFL [65] improves DPFL under data heterogeneity via personalized data transformation, instead of changing the optimization algorithm. DP-FedPGN [34] studies client-level DPFL and improves generalization by introducing a global gradient norm penalty to seek flatter minima under privacy constraints. Different from client-level DP, which protects the participation of an entire client in training, sample-level DP protects the contribution of each individual data sample.

B. More Implementation Details

B.1. More Results

Swin-Tiny/Base on CIFAR-10. Table 10 reports the averaged test accuracy on CIFAR-10 for six different DPFL methods evaluated on Swin-Tiny and Swin-Base, under two data heterogeneity levels (Dirichlet α=0.6 and α=0.1). Overall, DP-FedAdamW consistently achieves the best performance, while Swin-Base is markedly more accurate than Swin-Tiny and all methods degrade when the data becomes more heterogeneous (smaller α). For Swin-Base with α=0.1, DP-FedAdamW attains 90.68% test accuracy, outperforming DP-FedSAM (89.82%) and DP-FedAvg (88.23%), even though DP-FedSAM needs roughly twice the privacy budget.

Table 10. Averaged test accuracy (%) comparison on CIFAR-10. Assuming DP-FedAvg is (ε, δ)-DP, all other methods share this budget except DP-FedSAM, which is approximately $(2\varepsilon, \delta)$-DP.

Method	Swin-Tiny		Swin-Base		Privacy Budget
	α=0.6	α=0.1	α=0.6	α=0.1	
DP-FedAvg	$82.93_{\pm 0.53}$	$78.24_{\pm 0.59}$	$89.56_{\pm 0.47}$	$88.23_{\pm 0.44}$	(ε, δ)
DP-SCAFFOLD	$82.85_{\pm 0.50}$	$78.53_{\pm 0.54}$	$90.07_{\pm 0.48}$	$89.08_{\pm 0.62}$	(ε, δ)
DP-FedAvg-LS	$83.94_{\pm 0.62}$	$79.67_{\pm 0.68}$	$90.97_{\pm 0.52}$	$90.14_{\pm 0.56}$	(ε, δ)
DP-FedSAM	$84.05_{\pm 0.51}$	$80.36_{\pm 0.57}$	$91.25_{\pm 0.54}$	$89.82_{\pm 0.60}$	$(2\varepsilon, \delta)$
DP-LocalAdamW	$83.73_{\pm 0.58}$	$79.28_{\pm 0.64}$	$90.84_{\pm 0.39}$	$89.08_{\pm 0.41}$	(ε, δ)
DP-FedAdamW	$84.23_{\pm 0.32}$	$81.23_{\pm 0.39}$	$91.79_{\pm 0.34}$	$90.68_{\pm 0.38}$	(ε, δ)

Noise multiplier results. To further analyze the impact of privacy noise on algorithm performance, we evaluate different noise multipliers $\sigma \in \{0.5, 1.0, 1.5, 2.0\}$ on CIFAR-100 using the Swin-Base model under Dirichlet α=0.6 and α=0.1. As shown in Table 11, increasing the noise multiplier consistently degrades test accuracy for all methods, while DP-FedAdamW maintains the best performance across all σ. Although σ=0.5 yields the highest accuracy, it also consumes a much larger privacy budget. In contrast, setting σ=1 significantly strengthens the formal privacy guarantee while only causing a small drop in accuracy, whereas larger noise levels ($\sigma \geq 1.5$) lead to a much more pronounced loss in utility. Therefore, we adopt σ=1 as the default noise multiplier in our main experiments, as it offers a favorable privacy–utility trade-off.

Table 11. Test accuracy (%) on CIFAR-100 with Swin-Base under different noise multipliers σ.

Method	$\alpha = 0.6$				$\alpha = 0.1$			
	σ=0.5	σ=1	σ=1.5	σ=2	σ=0.5	σ=1	σ=1.5	σ=2
DP-FedAvg	$67.23_{\pm 0.33}$	$65.61_{\pm 0.27}$	$63.78_{\pm 0.34}$	$60.22_{\pm 0.64}$	$62.49_{\pm 0.26}$	$60.77_{\pm 0.32}$	$57.85_{\pm 0.48}$	$54.34_{\pm 0.79}$
DP-SCAFFOLD	$70.41_{\pm 0.25}$	$68.03_{\pm 0.55}$	$64.64_{\pm 0.48}$	$63.14_{\pm 0.58}$	$66.72_{\pm 0.38}$	$65.34_{\pm 0.49}$	$61.73_{\pm 0.44}$	$58.23_{\pm 0.66}$
DP-FedAvg-LS	$70.93_{\pm 0.40}$	$68.80_{\pm 0.52}$	$66.52_{\pm 0.43}$	$64.77_{\pm 0.63}$	$68.36_{\pm 0.39}$	$66.08_{\pm 0.56}$	$62.87_{\pm 0.52}$	$59.98_{\pm 0.58}$
DP-FedSAM	$72.55_{\pm 0.22}$	$70.44_{\pm 0.57}$	$67.45_{\pm 0.33}$	$65.75_{\pm 0.69}$	$67.94_{\pm 0.51}$	$66.28_{\pm 0.64}$	$62.21_{\pm 0.60}$	$60.84_{\pm 0.71}$
DP-LocalAdamW	$70.74_{\pm 0.13}$	$68.33_{\pm 0.33}$	$66.46_{\pm 0.53}$	$64.10_{\pm 0.60}$	$67.35_{\pm 0.31}$	$65.28_{\pm 0.39}$	$60.55_{\pm 0.43}$	$58.78_{\pm 0.74}$
DP-FedAdamW	$73.62_{\pm 0.36}$	$71.58_{\pm 0.30}$	$68.52_{\pm 0.45}$	$66.79_{\pm 0.51}$	$69.22_{\pm 0.37}$	$67.45_{\pm 0.26}$	$63.28_{\pm 0.42}$	$61.05_{\pm 0.61}$

Table 12. Comparison with AdaDPS on Swin-Tiny.

Method	**CIFAR-100**(α=0.6)	**CIFAR-100**(α=0.1)	**Tiny-ImageNet**(α=0.6)	**Tiny-ImageNet**(α=0.1)
AdaDPS	44.23	38.15	22.08	15.84
DP-FedAdamW	46.77	39.36	23.96	19.19

Compared with AdaDPS in the federated setting. We additionally compare our method with AdaDPS under the same experimental setting for a fair comparison (Table 12). On CIFAR-100 with Swin-Tiny (α=0.6), AdaDPS achieves 44.23% test accuracy, compared with 46.77% for our method. On Tiny-ImageNet with Swin-Tiny (α=0.1), AdaDPS obtains 15.84%, while our method reaches 19.19%. This indicates that our design is better suited to federated non-IID training, where adaptive moments need to be stabilized under heterogeneity and partial participation.

Communication overhead of Agg-`mean`-$\boldsymbol{v}$. Instead of transmitting a full coordinate-wise second-moment vector, Agg-`mean`-$\boldsymbol{v}$ communicates only one scalar for each parameter block. Therefore, if the model has d parameters and is partitioned into B blocks, its additional one-way communication overhead is B scalars per transmission, rather than d scalars. If the aggregated block-wise statistics are also broadcast back to clients, the total extra communication becomes $2B$ scalars per round per client. Here, B is determined by our implementation-level architecture-aligned block partition: for Transformer-like models, each head-specific slice of `Q`, `K`, and `V` forms one block, while each `attn.proj` layer, each `MLP` layer, the `Embedding` layer, and the `output` layer each form one block; for CNNs, each convolutional layer or residual block is treated as one block. In practice, different datasets may slightly change the total parameter count d through dataset-specific output heads (e.g., different numbers of classes), while the computation rule of B remains unchanged unless the model architecture or the block definition is modified. Table 13 reports the corresponding B values and additional one-way payloads for Swin-Base, ViT-Base, and RoBERTa-base under this default block partition.

Table 13. Additional one-way communication overhead of Agg-`mean`-$\boldsymbol{v}$. Here, d is the number of model parameters, B is the number of parameter blocks under our default implementation-level block partition, the extra payload is reported under FP32 precision, and the overhead ratio is defined as B/d.

Model	d	B	Extra payload	Overhead ratio
ViT-Base	$\approx$ 86M	458	$\approx$ 1.8 KB	$\approx 5.3 \times 10^{-6}$
RoBERTa-base	$\approx$ 125M	458	$\approx$ 1.8 KB	$\approx 3.7 \times 10^{-6}$
Swin-Base	$\approx$ 88M	1178	$\approx$ 4.6 KB	$\approx 1.3 \times 10^{-5}$

Sensitivity to AdamW hyperparameters (β_1, β_2). We further study the sensitivity of DP-LocalAdamW and DP-FedAdamW to the AdamW momentum hyperparameters (β_1, β_2). In all main experiments, we use the standard AdamW defaults, i.e., β_1=0.9 and β_2=0.999. Table 14 reports the ablation results on CIFAR-100 with Swin-Base under α=0.1. We observe that both methods are relatively stable across different (β_1, β_2) choices, while the default setting $(0.9, 0.999)$ achieves the best performance overall.

Table 14. Sensitivity study of (β_1, β_2) on CIFAR-100 with Swin-Base (α=0.1).

β_1	β_2	DP-LocalAdamW	DP-FedAdamW
0.9	0.99	$64.77_{\pm 0.62}$	$66.84_{\pm 0.50}$
0.9	0.999	$65.28_{\pm 0.39}$	$67.45_{\pm 0.26}$
0.95	0.99	$64.32_{\pm 0.32}$	$66.98_{\pm 0.23}$
0.95	0.999	$65.16_{\pm 0.54}$	$67.11_{\pm 0.48}$

B.2. Datasets

Vision datasets. We evaluate our methods on three widely used image classification benchmarks: CIFAR-10, CIFAR-100, and Tiny-ImageNet in Table 15. CIFAR-10 and CIFAR-100 [18] each contain $60{,}000$ natural images of size 32×32, split into $50{,}000$ training and $10{,}000$ test examples. CIFAR-10 comprises 10 coarse-grained object categories, whereas CIFAR-100 has finer labeling with 100 fine-grained categories organized into 20 superclasses. Tiny-ImageNet [19] is a downsampled subset of ImageNet with 200 classes, where each class contains 500 training images and 50 validation images of size 64×64.

Table 15. Summary of the vision datasets used in our experiments.

Dataset	#Classes	Image Size	Train / Test
CIFAR-10 [18]	10	32×32	50,000 / 10,000
CIFAR-100 [18]	100	32×32	50,000 / 10,000
Tiny-ImageNet [19]	200	64×64	100,000 / 10,000

Language datasets. For natural language understanding, we experiment on four tasks from the GLUE [60] benchmark: SST-2, QQP, QNLI, and MNLI (Table 16. SST-2 is a binary sentiment classification task over movie reviews. QQP is a paraphrase identification task that determines whether two Quora questions express the same semantic intent. QNLI is a binary classification task that predicts whether a candidate sentence contains the answer to a given question. MNLI is a three-way natural language inference benchmark (entailment, contradiction, neutral) covering multiple genres of text. We follow the standard GLUE data splits and evaluation protocols.

Table 16. Summary of GLUE datasets used in our experiments.

Dataset	Task Type	#Classes	Train/Test
SST-2	Sentiment Classification (binary)	2	67,349 / 1,821
QQP	Duplicate Question Detection (binary)	2	363,846 / 390,965
QNLI	Question-Answer NLI (binary)	2	104,743 / 5,463
MNLI	Natural Language Inference (entailment)	3	392,702 / 9,815

B.3. Models

We study three representative architecture families, covering five specific models in total, as summarized in Table 17.

(1) **ResNet-18.** We adopt a ResNet-18 with Group Normalization (GN), where all Batch Normalization layers are replaced by GN to avoid dependence on batch statistics, which is more suitable for federated and differentially private training. To adapt CIFAR-10/100, we follow standard practices and modify the stem by replacing the original 7×7 convolution with a 3×3 kernel and removing the initial downsampling layers (i.e., the stride-2 convolution and max-pooling layer).

(2) **ViT-Base.** We use the standard Vision Transformer Base (ViT-Base) architecture, which splits an input image into non-overlapping 16×16 patches and linearly embeds them into a sequence of tokens with a prepended class token. The token sequence is processed by 12 Transformer encoder layers with multi-head self-attention and MLP blocks, and the final class token representation is fed to a linear classifier to produce the output logits.

(3) **Swin-Tiny.** Swin-Tiny is a hierarchical Vision Transformer that uses shifted window self-attention. The model consists of four stages with depths $[2, 2, 6, 2]$, embedding dimensions $[96, 192, 384, 768]$, and attention heads $[3, 6, 12, 24]$. Images are partitioned into 4×4 patches, and each stage applies window-based attention with window size 7 and an MLP ratio of 4, using LayerNorm and relative positional bias. DropPath regularization with depth-scaled rates is used, and the final features are aggregated by global pooling and passed to a linear classifier.

(4) **Swin-Base.** Swin-Base is a larger variant of Swin-Tiny, sharing the same hierarchical shifted-window design. It is configured with four stages of depths $[2, 2, 18, 2]$, embedding dimensions $[128, 256, 512, 1024]$, and attention heads $[4, 8, 16, 32]$, using a patch size of 4×4, window size 7, and MLP ratio 4. Similar to Swin-Tiny, the final stage features are pooled globally and fed into a linear classifier to generate predictions.

(5) **RoBERTa-Base.** We use the RoBERTa-Base architecture, which follows the standard Transformer encoder design with 12 layers, 12 attention heads, and a hidden size of 768. It employs byte-pair encoding (BPE) tokenization and is pre-trained with a masked language modeling objective on large-scale corpora using dynamic masking. The final representation corresponding to the classification token is fed into a task-specific linear classifier.

Table 17. Overview of the model architectures used in our experiments. "Depth" denotes the total number of layers (blocks for ResNet/Swin, encoder layers for ViT/RoBERTa), and "Stages" denotes the number of blocks per stage for hierarchical models.

Model	Family	Depth	Stages	Width / Dim	Heads
ResNet-18 (GN)	CNN	18	$[2, 2, 2, 2]$	$[64, 128, 256, 512]$	–
ViT-Base	ViT Transformer	12	–	768 hidden, 3072 MLP	12 / layer
Swin-Tiny	Swin Transformer	12	$[2, 2, 6, 2]$	$[96, 192, 384, 768]$	$[3, 6, 12, 24]$
Swin-Base	Swin Transformer	24	$[2, 2, 18, 2]$	$[128, 256, 512, 1024]$	$[4, 8, 16, 32]$
RoBERTa-Base	NLP Transformer	12	–	768 hidden, 3072 FFN	12 / layer

Algorithm 2 DP-LocalAdamW

Initial model $\boldsymbol{\theta}^0$; rounds T; local steps K; step size η; weight decay λ; AdamW (β_1, β_2, τ); clip norm C; noise multiplier σ; S clients per round
for $t = 1$ **to** T **do**
  for selected clients $i = 1, \dots, S$ in parallel **do**
    $\boldsymbol{\theta}_i^{t,0} \leftarrow \boldsymbol{\theta}^t$; $\boldsymbol{m}_i^{t,0} \leftarrow \mathbf{0}$; $\boldsymbol{v}_i^{t,0} \leftarrow \mathbf{0}$
    for $k = 1$ **to** K **do**
      sample $\mathcal{D}_i^k \subset \mathcal{D}_i$ of size $\lfloor sR \rfloor$
      for each sample $j \in \mathcal{D}_i^k$ **do**
        $g_{ij} \leftarrow \nabla f_i(\boldsymbol{\theta}_i^{t,k}; \xi_{ij})$
        $\bar{g}_{ij} \leftarrow g_{ij} / \max(1, \|g_{ij}\|_2 / C)$
      end for
      $\tilde{g}_i^{t,k} \leftarrow \frac{1}{sR} \sum_{j \in \mathcal{D}_i^k} \bar{g}_{ij} + \frac{C}{sR} \mathcal{N}\big(0, \sigma^2 C^2 I\big)$
      $\boldsymbol{m}_i^{t,k} \leftarrow \beta_1 \boldsymbol{m}_i^{t,k-1} + (1 - \beta_1)\, \tilde{g}_i^{t,k}$
      $\boldsymbol{v}_i^{t,k} \leftarrow \beta_2 \boldsymbol{v}_i^{t,k-1} + (1 - \beta_2)\, \tilde{g}_i^{t,k} \odot \tilde{g}_i^{t,k}$
      $\hat{\boldsymbol{m}}_i^{t,k} \leftarrow \boldsymbol{m}_i^{t,k} / (1 - \beta_1^k)$
      $\hat{\boldsymbol{v}}_i^{t,k} \leftarrow \boldsymbol{v}_i^{t,k} / (1 - \beta_2^k)$
      $\boldsymbol{\theta}_i^{t,k+1} \leftarrow \boldsymbol{\theta}_i^{t,k} - \eta\, (\hat{\boldsymbol{m}}_i^{t,k} / (\sqrt{\hat{\boldsymbol{v}}_i^{t,k}} + \tau) - \lambda \boldsymbol{\theta}_i^{t,k})$
    end for
    Clients send $(\boldsymbol{\theta}_i^{t,K} - \boldsymbol{\theta}_i^{t,0})$ to server
  end for
  $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t + \frac{1}{S} \sum_{i=1}^{S} (\boldsymbol{\theta}_i^{t,K} - \boldsymbol{\theta}_i^{t,0})$
  Broadcast $\boldsymbol{\theta}^{t+1}$ to clients
end for

B.4. Complete Configuration

Federated setup. In all experiments, we fine-tune the models using parameter-efficient Low-Rank Adaptation (**LoRA**). We then detail the federated learning configurations adopted in our experiments. Table 18 specifies the settings for all vision benchmarks, including the number of clients N, communication rounds T, local update steps K, client participation rate l, local batch size sR, DP noise multiplier σ, failure probability δ, and LoRA hyperparameters (r, α_L). Table 20 reports the corresponding configuration for the GLUE benchmarks with RoBERTa-Base and LoRA, using the same federated parameters while additionally documenting the maximum sequence length and dropout rate.

Other setup. DP-FedAvg, DP-SCAFFOLD, DP-FedAvg-LS, and DP-FedSAM adopt learning rates selected from $\{10^{-2}, 3{\times}10^{-2}, 5{\times}10^{-2}, 10^{-1}, 3{\times}10^{-1}\}$ with a fixed weight decay of 0.001. For DP-FedSAM, the SAM perturbation ρ is selected via grid search over $\{0.01, 0.05, 0.1, 0.5\}$. For DP-LocalAdamW and DP-FedAdamW, the learning rate is chosen from $\{10^{-4}, 2{\times}10^{-4}, 3{\times}10^{-4}, 5{\times}10^{-4}, 8{\times}10^{-4}, 10^{-3}\}$, combined with weight decay in $\{0.01, 0.001\}$ and standard AdamW parameters (β_1=0.9, β_2=0.999). We employ cosine learning-rate decay throughout, and DP-FedAdamW additionally uses parameter γ=0.5 and weight decay λ=0.01. For all visual fine-tuning tasks, we initialize from official ImageNet-22K pre-trained weights to ensure consistency across methods. Across all tasks, the gradient clipping threshold is chosen by grid search over $\{0.05, 0.1, 0.2, 0.3, 0.5\}$, and we find that larger clipping values can trigger gradient explosion. To obtain a simple and robust configuration across all benchmarks, we therefore adopt a unified clipping threshold of C=0.1.

Table 18. Federated configurations for vision models across CIFAR-10/100 and Tiny-ImageNet. We report the number of clients N, communication rounds T, local steps K, client participation rate l, local batch size sR, DP noise multiplier σ, failure probability δ, and LoRA hyperparameters (r, α_L) and dropout rate.

Dataset	**Model**	N	T	K	l	sR	σ	δ	r **(LoRA)**	α_L **(LoRA)**	**Dropout**
CIFAR-10	ResNet-18 (GN)	50	300	50	0.2	50	1.0	10^{-5}	16	32	0.1
	ViT-Base	50	100	20	0.1	16	1.0	10^{-5}	16	32	0.1
	Swin-Tiny/Base	50	100	20	0.1	16	1.0	10^{-5}	16	32	0.1
CIFAR-100	ResNet-18 (GN)	50	300	50	0.2	50	1.0	10^{-5}	16	32	0.1
	ViT-Base	50	100	20	0.1	16	1.0	10^{-5}	16	32	0.1
	Swin-Tiny/Base	50	100	20	0.1	16	1.0	10^{-5}	16	32	0.1
Tiny-ImageNet	Swin-Tiny/Base	50	100	20	0.1	16	1.0	10^{-6}	16	32	0.1

Table 19. Federated configurations for RoBERTa-Base with LoRA on GLUE tasks. We report the number of clients N, communication rounds T, local steps K, client participation rate l, local batch size sR, DP noise multiplier σ, failure probability δ, and LoRA hyperparameters (r, α_L), as well as the maximum sequence length and dropout rate.

Dataset	**Model**	N	T	K	l	sR	σ	δ	r **(LoRA)**	α_L **(LoRA)**	**Length_max**	**Dropout**
SST-2	RoBERTa-Base	4	100	20	1.0	16	1.0	10^{-5}	16	32	128	0.1
QQP		4	100	20	1.0	16	1.0	10^{-6}	16	32	128	0.1
QNLI		4	100	20	1.0	16	1.0	10^{-5}	16	32	128	0.1
MNLI		4	100	20	1.0	16	1.0	10^{-6}	16	32	128	0.1

Table 20. Hyperparameter configuration of ResNet-18 (CIFAR-10/100) across different algorithms.

Method	**Local Optimizer**	η	β_1	β_2	**Weight Decay**
DP-FedAvg	SGD	10^{-1}	-	-	0.001
DP-SCAFFOLD	SGD	10^{-1}	-	-	0.001
DP-FedAvg-LS	SGD	10^{-1}	-	-	0.001
DP-FedSAM	SAM	10^{-1}	-	-	0.001
DP-LocalAdamW	AdamW	3×10^{-4}	0.9	0.999	0.01
DP-FedAdamW	AdamW	3×10^{-4}	0.9	0.999	0.01

C. DP-LocalAdamW Algorithm

For completeness, we provide in Algorithm 2 the full local training procedure of **DP-LocalAdamW**.

D. Theoretical Analysis Details

D.1. Proof of Theorem 1 and Convergence Analysis

Assumption A.1 (Smoothness)**.** *The non-convex f_i is a L-smooth function for all $i \in [m]$, i.e., $\|\nabla f_i(\boldsymbol{\theta}_1) - \nabla f_i(\boldsymbol{\theta}_2)\| \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$.*

Assumption A.2 (Bounded Stochastic Gradient)**.** *$\boldsymbol{g}_i^t = \nabla f_i(\boldsymbol{\theta}_i^t, \xi_i^t)$ computed by 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., $\mathbb{E}_{\xi_i^t}[\boldsymbol{g}_i^t] = \nabla f_i(\boldsymbol{\theta}_i^t)$ and $\mathbb{E}_{\xi_i^t}\|g_i^t - \nabla f_i(\boldsymbol{\theta}_i^t)\|^2 \leq \sigma_l^2$, for all $\boldsymbol{\theta}_i^t \in \mathbb{R}^d$.*

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

Assumption A.4 (Bounded Heterogeneity)**.** *The gradient dissimilarity between clients is bounded: $\|\nabla f_i(\boldsymbol{\theta}_1) - \nabla f_i(\boldsymbol{\theta}_2)\|^2 \leq \sigma_g^2$, for all $\boldsymbol{\theta} \in \mathbb{R}^d$.*

In this section, we give the theoretical analysis of our proposed DP-FedAdamW algorithm. Firstly we state some standard assumptions for the non-convex function F.

We use the common gradient bound condition in our proof. Then we can upper bound G_ϑ as:

$$G_\vartheta = \left\{\|\vartheta_i^{t,k}\|_\infty\right\} = \left\{\|\frac{1}{\sqrt{(1-\beta_1^k)^2/(1-\beta_2^t)\boldsymbol{v}_i^{t,k}}+\tau}\|_\infty\right\}$$
$$= \left\{\|\frac{1}{\sqrt{(1-\beta_1^k)^2/(1-\beta_2^t)(\beta_2\boldsymbol{v}_i^{t,k-1}+(1-\beta_2)\,\boldsymbol{g}_i^{t,k}\cdot\boldsymbol{g}_i^{t,k})}+\tau}\|_\infty\right\}.$$

$\sqrt{(1-\beta_1^k)^2/(1-\beta_2^t)(\beta_2\boldsymbol{v}_i^{t,k-1}+(1-\beta_2)\,\boldsymbol{g}_i^{t,k}\cdot\boldsymbol{g}_i^{t,k}}$ is bounded as:

$$\|\sqrt{(1-\beta_1^k)^2/(1-\beta_2^t)(\beta_2\boldsymbol{v}_i^{t,k-1}+(1-\beta_2)\,\boldsymbol{g}_i^{t,k}\cdot\boldsymbol{g}_i^{t,k}}\|_\infty \le G_g. \tag{15}$$

Thus we bound G_ϑ as $\frac{1}{G_g} \le G_\vartheta \le \frac{1}{\tau}$.

D.2. Main Lemmas

Lemma 1. *Suppose $\{X_1,\cdots,X_\tau\}\subset\mathbb{R}^d$ be random variables that are potentially dependent. If their marginal means and variances satisfy $\mathbb{E}[X_i]=\mu_i$ and $\mathbb{E}\left[\|X_i-\ \mu_i\|^2\right]\le\sigma^2$, then it holds that*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{\tau}X_i\right\|^2\right]\le\left\|\sum_{i=1}^{\tau}\mu_i\right\|^2+\tau^2\sigma^2.$$

If they are correlated in the Markov way such that $\mathbb{E}[X_i\mid X_{i-1},\cdots X_1]=\mu_i$ and $\mathbb{E}\left[\|X_i-\mu_i\|^2\mid\ \mu_i\right]\le\sigma^2$, i.e., the variables $\{X_i-\mu_i\}$ form a martingale. Then the following tighter bound holds:

$$\mathbb{E}\left[\left\|\sum_{i=1}^{\tau}X_i\right\|^2\right]\le 2\mathbb{E}\left[\left\|\sum_{i=1}^{\tau}\mu_i\right\|^2\right]+2\tau\sigma^2.$$

Lemma 2. *Given vectors $v_1,\cdots,v_N\in\mathbb{R}^d$ and $\bar{v}=\frac{1}{N}\sum_{i=1}^N v_i$, if we sample $\mathcal{S}\subset\{1,\cdots,N\}$ uniformly randomly such that $|\mathcal{S}|=S$, then it holds that*

$$\mathbb{E}\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}}v_i\right\|^2\right]=\|\bar{v}\|^2+\frac{N-S}{S(N-1)}\frac{1}{N}\sum_{i=1}^N\|v_i-\bar{v}\|^2.$$

Proof. Letting $\mathbb{I}\{i \in \mathcal{S}\}$ be the indicator for the event $i \in \mathcal{S}_r$, we prove this lemma by direct calculation as follows:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}} v_i\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{S}\sum_{i=1}^{N} v_i\mathbb{I}\{i\in\mathcal{S}\}\right\|^2\right] \\
&= \frac{1}{S^2}\mathbb{E}\left[\left(\sum_i \|v_i\|^2 \mathbb{I}\{i\in\mathcal{S}\} + 2\sum_{i<j} v_i^\top v_j \mathbb{I}\{i,j\in\mathcal{S}\}\right)\right] \\
&= \frac{1}{SN}\sum_{i=1}^{N}\|v_i\|^2 + \frac{1}{S^2}\frac{S(S-1)}{N(N-1)} 2\sum_{i<j} v_i^\top v_j \\
&= \frac{1}{SN}\sum_{i=1}^{N}\|v_i\|^2 + \frac{1}{S^2}\frac{S(S-1)}{N(N-1)}\left(\left\|\sum_{i=1}^{N} v_i\right\|^2 - \sum_{i=1}^{N}\|v_i\|^2\right) \\
&= \frac{N-S}{S(N-1)}\frac{1}{N}\sum_{i=1}^{N}\|v_i\|^2 + \frac{N(S-1)}{S(N-1)}\|\bar{v}\|^2 \\
&= \frac{N-S}{S(N-1)}\frac{1}{N}\sum_{i=1}^{N}\|v_i-\bar{v}\|^2 + \|\bar{v}\|^2.
\end{aligned}$$

□

D.3. Basic Assumptions and Notations

Let $\mathcal{F}^0 = \emptyset$ and $\mathcal{F}_i^{t,k} := \sigma\left(\left\{\boldsymbol{\theta}_i^{t,j}\right\}_{0\le j\le k} \cup \mathcal{F}^t\right)$ and $\mathcal{F}^{t+1} := \sigma\left(\cup_i \mathcal{F}_i^{t,K}\right)$ for all $t \ge 0$ where $\sigma(\cdot)$ indicates the σ-algebra. Let $\mathbb{E}_t[\cdot] := \overline{\mathbb{E}}\left[\cdot \mid \mathcal{F}^t\right]$ be the expectation, conditioned on the filtration $\mathcal{F}^t$, with respect to the random variables $\left\{\mathcal{S}^t, \left\{\xi_i^{t,k}\right\}_{1\le i\le N, 0\le k<K}\right\}$ in the t-th iteration. We also use $\mathbb{E}[\cdot]$ to denote the global expectation over all randomness in algorithms. Through out the proofs, we use $\sum_i$ to represent the sum over $i \in \{1,\dots,N\}$, while $\sum_{i\in\mathcal{S}^t}$ denotes the sum over $i \in \mathcal{S}^t$. Similarly, we use $\sum_k$ to represent the sum of $k \in \{0,\dots,K-1\}$. For all $t \ge 0$, we define the following auxiliary variables to facilitate proofs:

$$\begin{aligned}
\mathcal{E}_t &:= \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^t\right) \odot \frac{1}{SK}\sum_{i\in S^t}\sum_{k=1}^{K}\vartheta_i^{t,k} - \Delta_G^{t+1}\right\|^2\right], \\
U_t &:= \frac{1}{NK}\sum_i\sum_k \mathbb{E}\left[\left\|\boldsymbol{\theta}_i^{t,k} - \boldsymbol{\theta}^t\right\|\right]^2, \\
\xi_i^{t,k} &:= \mathbb{E}\left[\boldsymbol{\theta}_i^{t,k+1} - \boldsymbol{\theta}_i^{t,k} \mid \mathcal{F}_i^{t,k}\right], \\
\Xi_t &:= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\xi_i^{t,0}\right\|^2\right].
\end{aligned}$$

Throughout the Appendix, we let $\Delta := f\left(\boldsymbol{\theta}^0\right) - f^\star, G_0 := \frac{1}{N}\sum_i \left\|\nabla f_i\left(\boldsymbol{\theta}^0\right)\right\|^2, \boldsymbol{\theta}^{-1} := \boldsymbol{\theta}^0$ and $\mathcal{E}_{-1} := \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^0\right) - g^0\right\|^2\right]$. We will use the following foundational lemma for all our algorithms.

D.4. Proof of Theorem 1

Lemma 3. *Under Assumption 1 , if $\gamma L \le \frac{1}{24}$, the following holds all $t \ge 0$:*

$$\mathbb{E}\left[f\left(\boldsymbol{\theta}^{t+1}\right)\right] \le \mathbb{E}\left[f\left(\boldsymbol{\theta}^t\right)\right] - \frac{11\gamma}{24}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\|^2\right] + \frac{13\gamma}{24}\mathcal{E}_t + \frac{13\gamma}{24}\frac{\sigma^2 G_g^2}{s^2 R^2}.$$

Proof. Since f is L-smooth, we have

$$\begin{aligned}
f\left(\boldsymbol{\theta}^{t+1}\right) &\leq f\left(\boldsymbol{\theta}^{t}\right)+\left\langle \nabla f\left(\boldsymbol{\theta}^{t}\right), \boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^{t}\right\rangle+\frac{L}{2}\left\|\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^{t}\right\|^{2} \\
&\leq f\left(\boldsymbol{\theta}^{t}\right)+\gamma\left\langle \nabla f\left(\boldsymbol{\theta}^{t}\right), \Delta_{G}^{t+1}\right\rangle+\frac{L\gamma^{2}}{2}\left\|\Delta_{G}^{t+1}\right\|^{2} \\
&= f\left(\boldsymbol{\theta}^{t}\right)-\gamma\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \vartheta_{i}^{t, k}\right\|^{2}+\gamma\left\langle \nabla f\left(\boldsymbol{\theta}^{t}\right), \nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \vartheta_{i}^{t, k}-\Delta_{G}^{t+1}\right\rangle \\
&+\frac{L\gamma^{2}}{2}\left\|\Delta_{G}^{t+1}\right\|^{2}.
\end{aligned}$$

Since $\boldsymbol{\theta}^{t+1}=\boldsymbol{\theta}^{t}-\gamma\Delta_{G}^{t+1}$, using Young's inequality, we further have

$$\begin{aligned}
&\mathbb{E} f\left(\boldsymbol{\theta}^{t+1}\right) \\
&\leq f\left(\boldsymbol{\theta}^{t}\right)-(\gamma G_{g}^{2}-\frac{\gamma}{2})\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right)\right\|^{2}+\frac{\gamma}{2}\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \vartheta_{i}^{t, k}-\Delta_{G}^{t+1}\right\|^{2} \\
&+L\gamma^{2}\left(\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \vartheta_{i}^{t, k}\right\|^{2}+\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \vartheta_{i}^{t, k}-\Delta_{G}^{t+1}\right\|^{2}\right) \\
&\leq f\left(\boldsymbol{\theta}^{t}\right)-(\gamma G_{g}^{2}-\frac{\gamma}{2}-L\gamma^{2}G_{g}^{2})\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right)\right\|^{2}+\frac{13\gamma}{24}\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \vartheta_{i}^{t, k}-\Delta_{G}^{t+1}\right\|^{2}+\frac{13\gamma}{24}\frac{\sigma^{2}G_{g}^{2}}{s^{2}R^{2}} \\
&\leq f\left(\boldsymbol{\theta}^{t}\right)-\frac{11\gamma}{24}\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right)\right\|^{2}+\frac{13\gamma}{24}\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \vartheta_{i}^{t, k}-\Delta_{G}^{t+1}\right\|^{2}+\frac{13\gamma}{24}\frac{\sigma^{2}G_{g}^{2}}{s^{2}R^{2}},
\end{aligned}$$

where the last inequality is due to $\gamma L \leq \frac{1}{24}$, $-(\gamma G_{g}^{2}-\frac{\gamma}{2}-L\gamma^{2}G_{g}^{2}) \leq -\frac{11\gamma}{24}$. Taking the global expectation, we finish the proof. □

Lemma 4. *If $\gamma L \leq \frac{\gamma}{6}$, the following holds for $t \geq 1$:*

$$\mathcal{E}_{t} \leq\left(1-\frac{8\gamma}{9}\right)\mathcal{E}_{t-1}+\frac{4\gamma^{2}L^{2}}{\gamma}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^{t-1}\right)\right\|^{2}\right]+\frac{2\gamma^{2}\sigma_{l}^{2}}{SK\tau^{2}}+4\frac{\gamma}{\tau^{2}}L^{2}U_{t}.$$

Additionally, it holds for $t=0$ that

$$\mathcal{E}_{0} \leq(1-\gamma)\mathcal{E}_{-1}+\frac{2\gamma^{2}\sigma_{l}^{2}}{SK}+4\gamma L^{2}U_{0}.$$

Proof. For $t>1$,

$$\begin{aligned}
\mathcal{E}_{t} =&\mathbb{E}\left[\left\|\frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \vartheta_{i}^{t, k}-\Delta_{G}^{t+1}\right\|^{2}\right] \\
=&\mathbb{E}\left[\left\|(1-\gamma)\left(\frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \vartheta_{i}^{t, k}-\Delta_{G}^{t}\right)+\gamma\left(\frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \vartheta_{i}^{t, k}-\frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} g_{i}^{t, k} \odot \vartheta_{i}^{t, k}\right)\right\|^{2}\right] \\
\leq&\mathbb{E}\left[\left\|(1-\gamma)\left(\frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \vartheta_{i}^{t, k}-\Delta_{G}^{t}\right)\right\|^{2}\right]+\gamma^{2}\frac{1}{\tau^{2}}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^{t}\right)-\frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} g_{i}^{t, k}\right\|^{2}\right] \\
&+2\gamma\mathbb{E}\left[\left\langle(1-\gamma)\left(\frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \vartheta_{i}^{t, k}-\Delta_{G}^{t}\right), \frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} \nabla f\left(\boldsymbol{\theta}^{t}\right) \odot \vartheta_{i}^{t, k}-\frac{1}{SK} \sum_{i \in S^{t}} \sum_{k=1}^{K} g_{i}^{t, k} \odot \vartheta_{i}^{t, k}\right\rangle\right].
\end{aligned}$$

Note that $\left\{\nabla F\left(\boldsymbol{\theta}_i^{t,k};\xi_i^{t,k}\right)\right\}_{0\le k<K}$ are sequentially correlated. Applying the AM-GM inequality and Lemma 1, we have

$$\mathcal{E}_t \le \left(1+\frac{\gamma}{2}\right)\mathbb{E}\left[\left\|(1-\gamma)\left(\nabla f\left(\boldsymbol{\theta}^t\right)-\Delta_G^t\right)\right\|^2\right]+2\frac{\gamma}{\tau^2}L^2U_t+2\frac{\gamma^2}{\tau^2}\left(\frac{\sigma_l^2}{SK}+L^2U_t\right).$$

Using the AM-GM inequality again and Assumption 1, we have

$$\begin{aligned}\mathcal{E}_t &\le (1-\gamma)^2\left(1+\frac{\gamma}{2}\right)\left[\left(1+\frac{\gamma}{2}\right)\mathcal{E}_{t-1}+\left(1+\frac{2}{\gamma}\right)L^2\mathbb{E}\left[\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\right\|^2\right]\right]+\frac{2\gamma^2\sigma_l^2}{SK\tau^2}+4\frac{\gamma}{\tau^2}L^2U_t\\ &\le (1-\gamma)\mathcal{E}_{t-1}+\frac{2}{\gamma}L^2\mathbb{E}\left[\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\right\|^2\right]+\frac{2\gamma^2\sigma_l^2}{\tau^2SK}+4\frac{\gamma}{\tau^2}L^2U_t\\ &\le \left(1-\frac{8\gamma}{9}\right)\mathcal{E}_{t-1}+4\frac{\gamma^2L^2}{\gamma}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2\right]+\frac{2\gamma^2\sigma_l^2}{\tau^2SK}+4\frac{\gamma}{\tau^2}L^2U_t,\end{aligned}$$

where we plug in $\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\right\|^2\le 2\gamma^2\left(\left\|\nabla f\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2+\left\|\Delta_G^t-\nabla f\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2\right)$ and use $\gamma L\le\frac{\gamma}{6}$ in the last inequality. Similarly for $t=0$,

$$\begin{aligned}\mathcal{E}_0 &\le \left(1+\frac{\gamma}{2}\right)\mathbb{E}\left[\left\|(1-\gamma)\left(\nabla f\left(\boldsymbol{\theta}^0\right)-g^0\right)\right\|^2\right]+2\frac{\gamma}{\tau^2}L^2U_0+2\frac{\gamma^2}{\tau^2}\left(\frac{\sigma_l^2}{SK}+L^2U_0\right)\\ &\le (1-\gamma)\mathcal{E}_{-1}+\frac{2\gamma^2\sigma_l^2}{SK\tau^2}+4\frac{\gamma}{\tau^2}L^2U_0.\end{aligned}$$

□

Lemma 5. *If $\eta LK\le\frac{1}{\gamma}$, the following holds for $t\ge 0$:*

$$U_t \le 2eK^2\Xi_t+K\eta^2\gamma^2\frac{1}{\tau^2}\sigma_l^2\left(1+2K^3L^2\eta^2\gamma^2\right).$$

Proof. Recall that $\xi_i^{t,k}:=\mathbb{E}\left[\boldsymbol{\theta}_i^{t,k+1}-\boldsymbol{\theta}_i^{t,k}\mid\mathcal{F}_i^{t,k}\right]=-\eta\left((1-\gamma)\Delta_G^t+\gamma\nabla f_i\left(\boldsymbol{\theta}_i^{t,k}\right)\odot\vartheta_i^{t,k}\right)$. Then we have

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_i^{t,j}-\xi_i^{t,j-1}\right\|^2\right] &\le \frac{1}{\tau^2}\eta^2L^2\gamma^2\mathbb{E}\left[\left\|\boldsymbol{\theta}_i^{t,j}-\boldsymbol{\theta}_i^{t,j-1}\right\|^2\right]\\ &\le \frac{1}{\tau^2}\eta^2L^2\gamma^2\left(\eta^2\gamma^2\sigma_l^2+\mathbb{E}\left[\left\|\xi_i^{t,j-1}\right\|^2\right)\right.\end{aligned}$$

For any $1\le j\le k-1\le K-2$, using $\eta L\le\frac{1}{\gamma K}\le\frac{1}{\gamma(k+1)}$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_i^{t,j}\right\|^2\right] &\le \left(1+\frac{1}{k}\right)\mathbb{E}\left[\left\|\xi_i^{t,j-1}\right\|^2\right]+(1+k)\mathbb{E}\left[\left\|\xi_i^{t,j}-\xi_i^{t,j-1}\right\|^2\right]\\ &\le \left(1+\frac{2}{k}\right)\mathbb{E}\left[\left\|\xi_i^{t,j-1}\right\|^2\right]+(k+1)\frac{1}{\tau^2}L^2\eta^4\gamma^4\sigma_l^2\\ &\le e^2\mathbb{E}\left[\left\|\xi_i^{t,0}\right\|^2\right]+4\frac{1}{\tau^2}k^2L^2\eta^4\gamma^4\sigma_l^2.\end{aligned}$$

where the last inequality is by unrolling the recursive bound and using $\left(1+\frac{2}{k}\right)^k\le e^2$. By Lemma 1 , it holds that for $k\ge 2$,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\theta}_i^{t,k}-\boldsymbol{\theta}^t\right\|^2\right] &\le 2\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\xi_i^{t,j}\right\|^2\right]+2\frac{1}{\tau^2}k\eta^2\gamma^2\sigma_l^2\\ &\le 2k\sum_{j=0}^{k-1}\mathbb{E}\left[\left\|\xi_i^{t,k}\right\|^2\right]+2\frac{1}{\tau^2}k\eta^2\gamma^2\sigma_l^2\\ &\le 2e^2k^2\mathbb{E}\left[\left\|\xi_i^{t,0}\right\|^2\right]+2\frac{1}{\tau^2}k\eta^2\gamma^2\sigma_l^2\left(1+4k^3L^2\eta^2\gamma^2\right).\end{aligned}$$

This is also valid for $k = 0, 1$. Summing up over i and k, we finish the proof. □

Lemma 6. *If* $288e(\eta KL)^2\left((1-\gamma)^2 + e(\gamma\gamma LR)^2\right) \leq 1$*, then it holds for* $t \geq 0$ *that*

$$\sum_{t=0}^{R-1} \Xi_t \leq \frac{1}{72eK^2L^2} \sum_{t=-1}^{R-2} \left(\mathcal{E}_t + \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\|^2\right]\right) + 2\eta^2\gamma^2\frac{1}{\tau^2}eRG_0.$$

Proof. Note that $\xi_i^{t,0} = -\eta\left((1-\gamma)\Delta_G^t + \gamma\nabla f_i\left(\boldsymbol{\theta}^t\right)\odot\vartheta_i^{t,k}\right)$,

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\xi_i^{t,0}\right\|^2 \leq 2\eta^2\left((1-\gamma)^2\left\|\Delta_G^t\right\|^2 + \gamma^2\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i\left(\boldsymbol{\theta}^t\right)\right\|^2\right).$$

Using Young's inequality, we have for any $q > 0$ that

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f_i\left(\boldsymbol{\theta}^t\right)\right\|^2\right] &\leq (1+q)\mathbb{E}\left[\left\|\nabla f_i\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2\right] + \left(1+q^{-1}\right)L^2\mathbb{E}\left[\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\right\|^2\right] \\
&\leq (1+q)\mathbb{E}\left[\left\|\nabla f_i\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2\right] + 2\left(1+q^{-1}\right)\gamma^2L^2\left(\mathcal{E}_{t-1} + \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2\right]\right) \\
&\leq (1+q)^t\mathbb{E}\left[\left\|\nabla f_i\left(\boldsymbol{\theta}^0\right)\right\|^2\right] + \frac{2}{q}\gamma^2L^2\sum_{j=0}^{t-1}\left(\mathcal{E}_j + \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^j\right)\right\|^2\right)(1+q)^{t-j}\right. .
\end{aligned}$$

Take $q = \frac{1}{t}$ and we have

$$\mathbb{E}\left[\left\|\nabla f_i\left(\boldsymbol{\theta}^t\right)\right\|^2\right] \leq e\mathbb{E}\left[\left\|\nabla f_i\left(\boldsymbol{\theta}^0\right)\right\|^2\right] + 2e(t+1)\gamma^2L^2\sum_{j=0}^{t-1}\left(\mathcal{E}_j + \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^j\right)\right\|^2\right)\right. . \tag{16}$$

Note that this inequality is valid for $t = 0$. Therefore, using (16), we have

$$\begin{aligned}
\sum_{t=0}^{R-1}\Xi_t &\leq \sum_{t=0}^{R-1} 2\eta^2\mathbb{E}\left[(1-\gamma)^2\left\|\Delta_G^t\right\|^2 + \gamma^2\frac{1}{\tau^2}\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i\left(\boldsymbol{\theta}^t\right)\right\|^2\right] \\
&\leq \sum_{t=0}^{R-1} 2\eta^2\left(2(1-\gamma)^2\left(\mathcal{E}_{t-1} + \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2\right]\right) + \gamma^2\frac{1}{\tau^2}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla f_i\left(\boldsymbol{\theta}^t\right)\right\|^2\right]\right) \\
&\leq \sum_{t=0}^{R-1} 4\eta^2(1-\gamma)^2\left(\mathcal{E}_{t-1} + \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2\right]\right) \\
&\quad + 2\eta^2\gamma^2\frac{1}{\tau^2}\sum_{t=0}^{R-1}\left(\frac{e}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla f_i\left(\boldsymbol{\theta}^0\right)\right\|^2\right] + 2e(t+1)(\gamma L)^2\sum_{j=0}^{t-1}\left(\mathcal{E}_j + \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^j\right)\right\|^2\right]\right)\right) \\
&\leq 4\eta^2(1-\gamma)^2\sum_{t=0}^{R-1}\left(\mathcal{E}_{t-1} + \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2\right]\right) \\
&\quad + 2\eta^2\gamma^2\frac{1}{\tau^2}\left(eRG_0 + 2e(\gamma LR)^2\sum_{t=0}^{R-2}\left(\mathcal{E}_t + \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\|^2\right]\right)\right).
\end{aligned}$$

Rearranging the equation and applying the upper bound of η, we finish the proof. □

Theorem 3 (Convergence for non-convex functions)**.** *Under Assumption 1 and 2 and 3 , if we take* $g^0 = 0$,

$$\gamma = \min\left\{, \sqrt{\frac{SKL\Delta\tau^2}{\sigma_l^2 R}}\right\} \text{ for any constant } c \in (0,1], \quad \gamma = \min\left\{\frac{1}{24LG_g}, \frac{\gamma}{6L}\right\},$$

$$\eta KL \lesssim \min\left\{1, \frac{1}{\gamma\gamma LR}, \left(\frac{L\Delta}{G_0\gamma^3R}\right)^{1/2}, \frac{1}{(\gamma N)^{1/2}}, \frac{1}{(\gamma^3NK)^{1/4}}\right\}.$$

then DP-FedAdamW converges as

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

Here $G_0 := \frac{1}{N}\sum_{i=1}^N \left\|\nabla f_i\left(\boldsymbol{\theta}^0\right)\right\|^2$.

Proof. Combining Lemma 3 and 4, we have

$$\begin{aligned}\mathcal{E}_t \leq& \left(1-\frac{8\gamma}{9}\right)\mathcal{E}_{t-1} + 4\frac{(\gamma L)^2}{\gamma}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^{t-1}\right)\right\|^2\right] + \frac{2\gamma^2\sigma_l^2}{SK\tau^2}\\ &+ 4\frac{\gamma}{\tau^2}L^2\left(2eK^2\Xi_t + K\eta^2\gamma^2\frac{1}{\tau^2}\sigma_l^2\left(1+2K^3L^2\eta^2\gamma^2\right),\right.\end{aligned}$$

and

$$\mathcal{E}_0 \leq (1-\gamma)\mathcal{E}_{-1} + \frac{2\gamma^2\sigma_l^2}{SK\tau^2} + 4\gamma\frac{1}{\tau^2}L^2\left(2eK^2\Xi_0 + K\eta^2\gamma^2\frac{1}{\tau^2}\sigma_l^2\left(1+2K^3L^2\eta^2\gamma^2\right)\right).$$

Summing over t from 0 to $R-1$ and applying Lemma 6,

$$\begin{aligned}\sum_{t=0}^{R-1}\mathcal{E}_t \leq& \left(1-\frac{8\gamma}{9}\right)\sum_{t=-1}^{R-2}\mathcal{E}_t + 4\frac{(\gamma L)^2}{\gamma}\sum_{t=0}^{R-2}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\|^2\right] + 2\frac{\gamma^2\sigma_l^2}{SK\tau^2}R\\ &+ 4\gamma\frac{1}{\tau^2}L^2\left(2eK^2\sum_{t=0}^{R-1}\Xi_t + RK\eta^2\gamma^2\frac{1}{\tau^2}\sigma_l^2\left(1+2K^3L^2\eta^2\gamma^2\right)\right)\\ \leq& \left(1-\frac{7\gamma}{9}\right)\sum_{t=-1}^{R-2}\mathcal{E}_t + \left(4\frac{(\gamma L)^2}{\gamma}+\frac{\gamma}{9}\right)\sum_{t=-1}^{R-2}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\|^2\right] + 16\gamma^3\frac{1}{\tau^4}(e\eta KL)^2RG_0\\ &+ \frac{2\gamma^2\sigma_l^2}{SK\tau^2}R + 4\gamma^3\frac{1}{\tau^4}(\eta KL)^2\left(\frac{1}{K}+2(\eta KL\gamma)^2\right)\sigma_l^2R\\ \leq& \left(1-\frac{7\gamma}{9}\right)\sum_{t=-1}^{R-2}\mathcal{E}_t + \frac{2\gamma}{9}\sum_{t=-1}^{R-2}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\|^2\right] + 16\gamma^3\frac{1}{\tau^4}(e\eta KL)^2RG_0 + \frac{4\gamma^2\sigma_l^2}{SK\tau^2}R.\end{aligned}$$

Here in the last inequality we apply

$$4\gamma\frac{1}{\tau^4}(\eta KL)^2\left(\frac{1}{K}+2(\eta KL\gamma)^2\right) \leq \frac{2}{NK} \quad \text{and} \quad \gamma L \leq \frac{\gamma}{6}.$$

Therefore,

$$\sum_{t=0}^{R-1}\mathcal{E}_t \leq \frac{9}{7\gamma}\mathcal{E}_{-1} + \frac{2}{7}\mathbb{E}\left[\sum_{t=-1}^{R-2}\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\|^2\right] + \frac{144}{7}\frac{1}{\tau^4}(e\gamma\eta KL)^2G_0R + \frac{36\gamma\sigma_l^2}{7SK\tau^2}R.$$

Combine this inequality with Lemma 3 and we get

$$\frac{1}{\gamma}\mathbb{E}\left[f\left(\boldsymbol{\theta}^t\right)-f\left(\boldsymbol{\theta}^0\right)\right] \leq -\frac{1}{7}\sum_{t=0}^{R-1}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\|^2\right] + \frac{39}{56\gamma}\mathcal{E}_{-1} + \frac{78}{7}\frac{1}{\tau^4}(e\gamma\eta KL)^2G_0R + \frac{39\gamma\sigma_l^2}{14SK\tau^2}R.$$

Finally, noticing that $g^0=0$ implies $\mathcal{E}_{-1} \leq 2L\left(f\left(\boldsymbol{\theta}^0\right)-f^*\right) = 2L\Delta$, we obtain

$$\begin{aligned}\frac{1}{R}\sum_{t=0}^{R-1}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\|^2\right] &\lesssim \frac{L\Delta}{\gamma LR} + \frac{\mathcal{E}_{-1}}{\gamma T} + (\gamma\eta KL)^2\frac{1}{\tau^4}G_0 + \frac{\gamma\sigma_l^2}{SK\tau^2} + \frac{\sigma^2G_g^2}{s^2T^2}\\ &\lesssim \frac{L\Delta}{T} + \frac{L\Delta}{\gamma T} + \frac{\gamma\sigma_l^2}{SK\tau^2} + (\gamma\eta KL)^2G_0\frac{1}{\tau^4} + \frac{\sigma^2G_g^2}{s^2T^2}\\ &\lesssim \frac{L\Delta}{T} + \sqrt{\frac{L\Delta\sigma_l^2}{SKT\tau^2}} + \frac{\sigma^2G_g^2}{s^2T^2}.\end{aligned}$$

□

E. Motivation for Block-wise Second-Moment Aggregation

This section explains the rationale behind the block-wise second-moment aggregation used in DP-FedAdamW. As discussed in Challenge 1, under non-IID data and DP perturbation, the second-moment estimator in AdamW can suffer from substantial variance amplification. Our block-wise design is introduced to mitigate this effect, while also reducing the communication overhead of second-moment transmission.

In AdamW, the second moment is updated using squared noisy gradients:

$$v_i^{t,k} \leftarrow \beta_2 v_i^{t,k-1} + (1-\beta_2)\,\tilde{g}_i^{t,k} \odot \tilde{g}_i^{t,k}. \tag{17}$$

Under non-IID data, client gradients can already be highly dispersed. DP perturbation further increases the variance of the noisy gradients. Since the second-moment update depends on the squared gradient, this noise can be amplified in the estimation of v. Moreover, because β_2 is typically close to 1, the exponential moving average evolves slowly, so the high variance in the second-moment estimator may persist across iterations and adversely affect adaptive scaling.

To alleviate this issue, we replace fully coordinate-wise second-moment aggregation with a block-wise approximation. Specifically, we partition the model parameters into B blocks and communicate only one aggregated second-moment statistic for each block. Let $\boldsymbol{v}$ denote the second-moment estimate, and let $\boldsymbol{v}_b$ denote the subset of entries belonging to block b. We compute the block-wise mean as

$$\bar{v}_b = \frac{1}{|\boldsymbol{v}_b|} \sum\nolimits_{i \in \boldsymbol{v}_b} v_i, \qquad b = 1, \ldots, B. \tag{18}$$

This design is motivated by two considerations. First, averaging within a block can smooth noisy coordinate-wise second-moment estimates and thus reduce estimator variance. Second, in AdamW, the second moment mainly serves to control the scale of parameter updates. Therefore, a block-wise aggregated statistic can still provide useful adaptive scaling without requiring full coordinate-wise second-moment communication. In addition, this design reduces the communication cost of second-moment aggregation from transmitting a full vector in $\mathbb{R}^d$ to transmitting only B scalars.

Besides mitigating the variance amplification of the second-moment estimator, our block-wise design is also motivated by the intuition that different architectural modules may exhibit different optimization scales. Therefore, instead of enforcing fully coordinate-wise second-moment adaptation, we use an architecture-aligned block partition to preserve coarse-grained adaptivity across major network components.

Architecture-aligned block partition. We adopt an architecture-aligned partition of model parameters.

- **Head-wise `Q`, `K`, `V`.** In multi-head attention, each head-specific slice of the `query` (`Q`), `key` (`K`), and `value` (`V`) weight matrices is treated as one parameter block. This preserves separation across different attention heads while sharing one aggregated second-moment statistic within each head.
- **`attn.proj` and `MLP` layers.** Each attention output projection layer (`attn.proj`) and each `MLP` layer is treated as a single parameter block. This provides a simple layer-wise grouping with low communication overhead.
- **`Embedding` and `output` layers.** Each `Embedding` layer and each `output` layer (e.g., the classification head) is treated as one parameter block. This again avoids per-parameter second-moment communication while maintaining a simple module-level adaptive statistic.

CNNs (e.g., ResNet). For convolutional networks, each convolutional layer or residual block is treated as one parameter block. This yields a natural module-level partition and reduces the communication cost of second-moment aggregation from per-parameter statistics to B block-level values.

Overall, our block-wise strategy is a structured approximation to coordinate-wise second-moment aggregation. Its primary goal is to mitigate the variance amplification of the second-moment estimator under non-IID data and DP perturbation, while communication reduction is an additional practical benefit.