Supplementary Material for
" Make Landscape Flatter in Differentially Private Federated Learning "

A. More Implementation Detail

A.1. Dataset

EMNIST [10] is a 62-class image classification dataset. In this paper, we use 20% of the dataset, which includes 88,800 training samples and 14,800 validation examples. Both CIFAR-10 and CIFAR-100 [28] have 60,000 images. In addition, these images are divided into 50,000 training samples and 10,000 validation examples. CIFAR-100 has finer labeling, with 100 unique labels, in comparison to CIFAR-10, having 10 unique labels. Furthermore, we divide these datasets to each client based on Dirichlet allocation over 500 clients by default.

A.2. Configuration

For the EMNIST dataset, we set the mini-batch size to 32 and train with a simple CNN model, which includes two convolutional layers with 5×5 kernels, max pooling, followed by a 512-unit dense layer. For CIFAR-10 and CIFAR-100 datasets, we set the mini-batch size to 50 and train with ResNet-18 [18] architecture. For each algorithm and each dataset, the learning rate is set via grid search on the set $\{10^{-0.5}, 10^{-1}, 10^{-1.5}, 10^{-2}\}$. The weight perturbation ratio ρ is set via grid search on the set $\{0.01, 0.1, 0.3, 0.5, 0.7, 1.0\}$. For all methods using the sparsification technique, the sparsity ratio is set to $p = 0.4$.

B. Additional Experiment on CIFAR-100

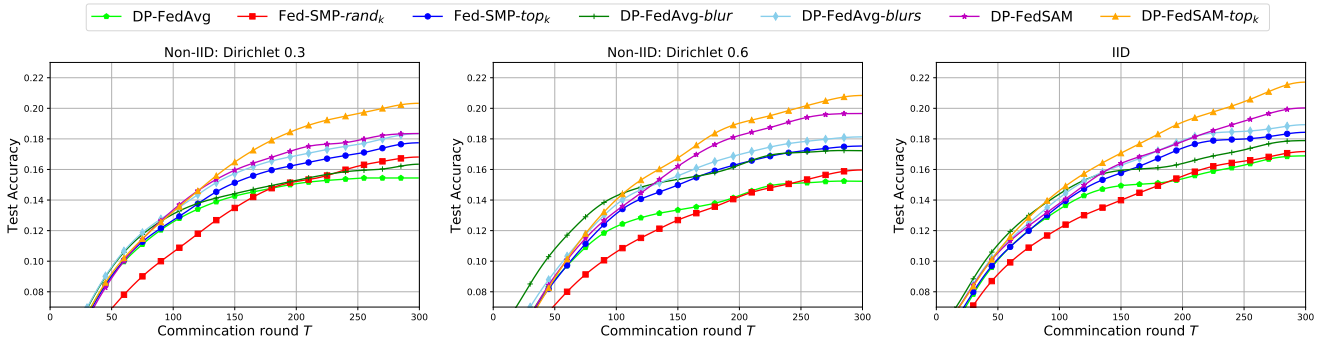
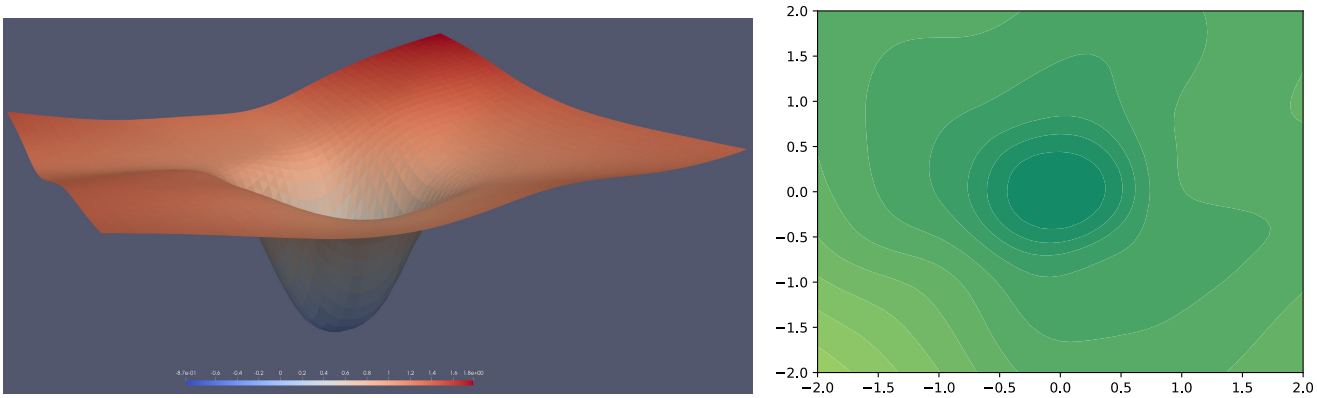


Figure 5. The averaged testing accuracy on *CIFAR-100* dataset under symmetric noise for all compared methods.

Table 4. Averaged training and testing accuracy (%) on *CIFAR-100* in both IID and Non-IID settings under symmetric noise for all compared methods. Note that the performance of the CIFAR-100 dataset is relatively poor across all algorithms due to the more severe impact of DP in complex tasks.

Algorithm	Dirichlet 0.3		Dirichlet 0.6		IID	
	Train	Validation	Train	Validation	Train	Validation
DP-FedAvg	91.14±0.16	16.10±0.71	92.33±0.08	15.92±0.39	94.01±0.10	17.47±0.47
Fed-SMP-rand$_k$	90.70±0.01	17.25±0.16	92.28±0.32	17.50±0.19	94.31±0.02	17.68±0.44
Fed-SMP-top$_k$	92.58±0.24	18.58±0.25	93.51±0.11	18.07±0.09	95.06±0.05	19.09±0.56
DP-FedAvg-blur	91.27±0.01	17.03±0.09	92.33±0.03	17.92±0.01	94.01±0.04	18.47±0.02
DP-FedAvg-blurs	92.98±0.24	18.98±0.25	94.01±0.11	18.27±0.19	95.46±0.05	19.59±0.06
DP-FedSAM	82.19±0.01	18.88±0.31	85.47±0.13	19.09±0.15	87.12±0.37	20.64±0.48
DP-FedSAM-top$_k$	84.49±0.24	**20.85±0.63**	88.23±0.23	**21.24±0.69**	89.86±0.21	**22.30±0.05**



(a) Loss landscape (b) Loss surface contour

Figure 6. Loss landscape and surface contour of DP-FedSAM. Compared with DP-FedAvg in the left of Figure 1 (a) and (b) with the same setting, DP-FedSAM has a flatter landscape with both better generalization ability (flat minima, see Figure 6 (a)) and weight perturbation robustness (see Figure 6 (b)).

B.1. Performance with Compared Baselines

In Table 4 and Figure 5, we evaluate DP-FedSAM and DP-FedSAM-top_k on CIFAR-100 dataset in both settings compared with all baselines from DP-FedAvg to DP-FedAvg-blurs. From all these results, it is clearly seen that our proposed algorithms outperform other baselines under symmetric noise both on accuracy and generalization perspectives. It means that we significantly improve the performance and generate a better trade-off between performance and privacy in DPFL. For instance, in the IID setting, the averaged testing accuracy is 20.64% in DP-FedSAM, where the accuracy gain is 3.17% compared with DP-FedAvg. And the average testing accuracy is 22.30% in DP-FedSAM-top_k, where the accuracy gain is 3.21% compared with Fed-SMP-top_k. That means our algorithms significantly mitigate the performance degradation issue caused by DP.

B.2. Impact of Non-IID levels

Under different participation cases as shown in Table 4, we further prove the robust generalization of the proposed algorithms. Heterogeneous data distribution of local clients is set to various participation levels from IID, Dirichlet 0.6, and Dirichlet 0.3, which makes the training of the global model more difficult. For instance, compared with DP-FedAvg on CIFAR-100, the test accuracy gain in DP-FedSAM is $\{2.78\%, 3.17\%, 3.17\%\}$. Meanwhile, the test accuracy gain in DP-FedSAM-top_k is $\{2.27\%, 3.17\%, 3.21\%\}$ compared with Fed-SMP-top_k. These observations confirm that our algorithms are more robust than baselines in various degrees of heterogeneous data.

C. More details on Discussion for DP with SAM in FL

Loss landscape and contour. To visualize the sharpness of the flat minima and observe robustness to DP noise obtained by DP-FedSAM, we show the loss landscape and surface contour following by the plotting method [30] in Figure 6. It is clear that DP-FedSAM has flatter minima and better robustness to DP noise than DP-FedAvg in the left of Figure 1 (a) and (b), respectively. It indicates that our proposed algorithm achieves better generalization and makes the training process more adaptive to the DPFL setting.

D. Discussion for DP Guarantee in DP-FedSAM with Sparsification

Sparsification is a very common method when considering privacy protection to introduce a large amount of random noise in FL [9, 20, 22]. It retains only the larger weight part of each layer of the local model with a sparsity ratio of k/d (d is the weight scale), and the rest are sparse. The advantage is that the amount of random noise can be reduced (no noise needs to be added to the sparse weight position), so the performance can be improved, which has been thoroughly verified in [9,20,22]. In our methods, SAM needs to perform two gradient calculations and sparsification may lead to some performance degradation because the model is compressed and some information may be lost.

Existing work [22] has verified SGD and top-k sparsification satisfying the Renyi DP. SAM optimizer only adds perturbation on the basis of SGD and affects the model during training. And both SAM and top k sparsification are performed before

the DP process, thereby satisfying the Renyi DP.

E. Main Proof

E.1. Preliminary Lemmas

Lemma 2. *(Lemma B.1, [41]) Under Assumptions 1-2, the updates for any learning rate satisfying $\eta \leq \frac{1}{4KL}$ have the drift due to $\delta_{i,k} - \delta$:*

$$\frac{1}{M}\sum_i \mathbb{E}[\|\delta_{i,k} - \delta\|^2] \leq 2K^2L^2\eta^2\rho^2.$$

Where

$$\delta = \rho\frac{\nabla f(\mathbf{w}^t)}{\|\nabla f(\mathbf{w}^t)\|}, \quad \delta_{i,k} = \rho\frac{\nabla F_i(\mathbf{w}^{t,k}, \xi_i)}{\|\nabla F_i(\mathbf{w}^{t,k}, \xi_i)\|}.$$

Lemma 3. *(lemma B.2, [41]) Under above assumptions, the updates for any learning rate satisfying $\eta_l \leq \frac{1}{10KL}$ have the drift due to $\mathbf{w}^{t,k}(i) - \mathbf{w}^t$:*

$$\frac{1}{M}\sum_i \mathbb{E}[\|\mathbf{w}^{t,k}(i) - \mathbf{w}^t\|^2] \leq 5K\eta^2\Big(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\mathbf{w}^t)\|^2\Big) + 24K^3\eta^4L^4\rho^2.$$

Lemma 4. *The two model parameters conducted by two adjacent datasets which differ only one sample from client i in the communication round t,*

$$\sum_{k=0}^{K-1}\|\mathbf{y}^{t,k}(i) - \mathbf{x}^{t,k}(i)\|_2^2 \leq 2K\max\|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2.$$

Proof. Recall the local update from client i is $\sum_{k=0}^{K-1}\mathbf{w}^{t,k}(i) = \sum_{k=0}^{K-1}\mathbf{w}^{t,k-1}(i) + \Delta_i^t$, (the initial value is assumed as $\mathbf{w}^{t,-1} = \mathbf{w}^{t,0} = \mathbf{w}^t$). Then,

$$\begin{aligned}\sum_{k=0}^{K-1}\|\mathbf{y}^{t,k}(i) - \mathbf{x}^{t,k}(i)\|_2^2 &\leq 2\sum_{k=0}^{K-1}\|\mathbf{y}^{t,k-1}(i) - \mathbf{x}^{t,k-1}(i)\|_2^2 \\ &+ 2\|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2.\end{aligned}$$

The recursion from $\tau = 0$ to k yields

$$\sum_{k=0}^{K-1}\|\mathbf{y}^{t,k}(i) - \mathbf{x}^{t,k}(i)\|_2^2 \overset{a)}{\leq} 2K\max\|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2.$$

Where a) uses the initial value $\mathbf{w}^t(i) = \mathbf{x}^{t,0}(i) = \mathbf{y}^{t,0}(i)$ and $0 < k \leq K$. □

Lemma 5. *Under assumption 1 and 3, the average of local update after the clipping operation from selected clients is*

$$\mathbb{E}\|\frac{1}{m}\sum_{i\in\mathcal{W}^t}\tilde{\Delta}_i^t\|^2 \leq 3K\eta^2(L^2\rho^2 + B^2)$$

Proof.

$$\begin{aligned}\mathbb{E}\|\frac{1}{m}\sum_{i\in\mathcal{W}^t}\tilde{\Delta}_i^t\|^2 \leq \mathbb{E}\|\frac{1}{m}\sum_{i\in\mathcal{W}^t}\sum_{i=0}^{K-1}\eta\tilde{\mathbf{g}}^{t,k}(i)\cdot\alpha_i^t\|^2 \leq \frac{\eta^2}{m}\sum_{i\in\mathcal{W}^t}\sum_{i=0}^{K-1}\mathbb{E}\|\nabla F_i(\mathbf{w}^{t,k}(i) + \delta;\xi_i) - \nabla F_i(\mathbf{w}^{t,K}(i);\xi_i) \\ + \nabla F_i(\mathbf{w}^{t,k}(i);\xi_i) - \nabla F_i(\mathbf{w}^t(i)) + \nabla F_i(\mathbf{w}^t(i))\|^2 \\ \overset{a)}{\leq} 3K\eta^2(L^2\rho^2 + B^2),\end{aligned}$$

where a) uses assumption 1 and 3 and

$$\alpha_i^t := \min\Big(1, \frac{C}{\eta\|\sum_{k=0}^{K-1}\tilde{\mathbf{g}}^{t,k}(i)\|}\Big).$$

□

E.2. Proof of Sensitivity Analysis

Proof of Theorem 1. Recall that the local update before clipping and adding noise on client i is $\Delta_i^t = \mathbf{w}^{t,K}(i) - \mathbf{w}^{t,0}(i)$. Then,

$$
\begin{aligned}
\mathbb{E}\mathcal{S}^2_{\Delta_i^t} &= \max \mathbb{E}\|\Delta_i^t(\mathbf{x}) - \Delta_i^t(\mathbf{y})\|_2^2 \\
&= \mathbb{E}\|\mathbf{x}^{t,K}(i) - \mathbf{x}^{t,0}(i) - (\mathbf{y}^{t,K}(i) - \mathbf{y}^{t,0}(i))\|_2^2 \\
&= \eta^2 \mathbb{E} \sum_{k=0}^{K-1} \|\nabla F_i(\mathbf{x}^{t,k}(i) + \delta_x; \xi_i) - \nabla F_i(\mathbf{y}^{t,k}(i) + \delta_y; \xi_i^{'})\|_2^2 \\
&= \eta^2 L^2 \mathbb{E} \sum_{k=0}^{K-1} \|\mathbf{y}^{t,k}(i) - \mathbf{x}^{t,k}(i) + (\delta_y - \delta_x)\|_2^2 \\
&\overset{a)}{\leq} 2\eta^2 L^2 K \max \|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2 + 2\eta^2 L^2 \rho^2 \mathbb{E} \sum_{k=0}^{K-1} \Big\| \frac{\nabla F_i(\mathbf{y}^{t,k}(i) + \delta_y; \xi_i^{'})}{\|\nabla F_i(\mathbf{y}^{t,k}(i) + \delta_y; \xi_i^{'})\|_2} - \frac{\nabla F_i(\mathbf{y}^{t,k}(i); \xi_i^{'})}{\|\nabla F_i(\mathbf{y}^{t,k}(i); \xi_i^{'})\|_2} \\
&+ \big(\frac{\nabla F_i(\mathbf{x}^{t,k}(i); \xi_i)}{\|\nabla F_i(\mathbf{x}^{t}(i,k); \xi_i)\|_2} - \frac{\nabla F_i(\mathbf{x}^{t,k}(i) + \delta_x; \xi_i)}{\|\nabla F_i(\mathbf{x}^{t,k}(i) + \delta_x; \xi_i)\|_2} \big) + \frac{\nabla F_i(\mathbf{y}^{t,k}(i); \xi_i^{'}))}{\|\nabla F_i(\mathbf{y}^{t,k}(i); \xi_i^{'}))\|_2} - \frac{\nabla F_i(\mathbf{x}^{t,k}(i); \xi_i)}{\|\nabla F_i(\mathbf{x}^{t,k}(i); \xi_i)\|_2} \Big\|_2^2 \\
&\leq 2\eta^2 L^2 K \max \|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2 + 6\eta^2 \rho^2 L^2 \mathbb{E} \sum_{k=0}^{K-1} \Big(4 + \frac{1}{\rho^2} \Big\| \rho \frac{\nabla F_i(\mathbf{y}^{t,k}(i); \xi_i^{'}))}{\|\nabla F_i(\mathbf{y}^{t,k}(i); \xi_i^{'}))\|_2} - \rho \frac{\nabla f(\mathbf{y}^t)}{\|\nabla f(\mathbf{y}^t)\|_2} \\
&+ \big(\rho \frac{\nabla f(\mathbf{x}^t)}{\|\nabla f(\mathbf{x}^t)\|_2} - \rho \frac{\nabla F_i(\mathbf{x}^{t,k}(i); \xi_i)}{\|\nabla F_i(\mathbf{x}^{t,k}(i); \xi_i)\|_2} \big) + \rho \frac{\nabla f(\mathbf{y}^t)}{\|\nabla f(\mathbf{y}^t)\|_2} - \rho \frac{\nabla f(\mathbf{x}^t)}{\|\nabla f(\mathbf{x}^t)\|_2} \Big\|_2^2 \Big) \\
&\overset{b)}{\leq} 2\eta^2 L^2 K \mathcal{S}^2_{\Delta_i^t} + 6\eta^2 \rho^2 K L^2 (4 + 12K^2 L^2 \eta^2 + 6) \\
&\leq \frac{6\eta^2 \rho^2 K L^2 (12K^2 L^2 \eta^2 + 10)}{1 - 2\eta^2 L^2 K}
\end{aligned}
\tag{16}
$$

where a) and b) uses lemma 4 and 2, respectively.

When the local adaptive learning rate satisfies $\eta = \mathcal{O}(1/L\sqrt{KT})$ and the perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$, we have

$$
\mathbb{E}\mathcal{S}^2_{\Delta_i^t} \leq \mathcal{O}\left(\frac{1}{T^2}\right). \tag{17}
$$

□

For comparison, we also present the expected squared sensitivity of local update with SGD in DPFL as follows. It is clearly seen that the upper bound in $\mathbb{E}\mathcal{S}^2_{\Delta_i^t, SAM}$ is tighter than that in $\mathbb{E}\mathcal{S}^2_{\Delta_i^t, SGD}$.

Proof of sensitivity with SGD in FL..

$$
\begin{aligned}
\mathbb{E}\mathcal{S}^2_{\Delta_i^t, SGD} &= \max \mathbb{E}\|\Delta_i^t(\mathbf{x}) - \Delta_i^t(\mathbf{y})\|_2^2 = \eta^2 \mathbb{E} \sum_{i=0}^{K-1} \|\nabla F_i(\mathbf{x}^{t,k}(i); \xi_i) - \nabla F_i(\mathbf{y}^{t,k}(i); \xi_i^{'})\|_2^2 \\
&= \eta^2 \mathbb{E} \sum_{i=0}^{K-1} \|\nabla F_i(\mathbf{x}^{t,k}(i); \xi_i) - \nabla F_i(\mathbf{x}^t(i)) + \nabla F_i(\mathbf{x}^t(i)) - \nabla F_i(\mathbf{y}^t(i)) + \nabla F_i(\mathbf{y}^t(i)) - \nabla F_i(\mathbf{y}^{t,k}(i); \xi_i^{'})\|_2^2 \\
&\overset{a)}{\leq} 3\eta^2 \mathbb{E} \sum_{i=0}^{K-1} (2\sigma_l^2 + L^2 \|y^{t,k}(i) - x^{t,k}(i)\|_2^2) \\
&\overset{b)}{\leq} 6\eta^2 K \sigma_l^2 + 3\eta^2 L^2 K \max \mathbb{E}\|\Delta_i^t(\mathbf{x}) - \Delta_i^t(\mathbf{y})\|_2^2 \\
&\leq \frac{6\eta^2 \sigma_l^2 K}{1 - 3\eta^2 K L^2}.
\end{aligned}
\tag{18}
$$

Where a) and b) uses assumptions 1-2 and lemma 4, respectively. Thus $\mathbb{E}\mathcal{S}^2_{\Delta_i^t,SGD} \leq \mathcal{O}(\frac{\sigma_l^2}{KL^2T})$ when $\eta = \mathcal{O}(1/L\sqrt{KT})$. □

E.3. Proof of Convergence Analysis

Proof of Theorem 3. We define the following notations for convenience:

$$\tilde{\Delta}_i^t = -\eta \sum_{k=0}^{K-1} \tilde{\mathbf{g}}^{t,k}(i) \cdot \alpha_i^t;$$
$$\overline{\Delta_i^t} = -\eta \sum_{k=0}^{K-1} \tilde{\mathbf{g}}^{t,k}(i) \cdot \overline{\alpha}^t,$$

where

$$\alpha_i^t := \min\Big(1, \frac{C}{\eta \| \sum_{k=0}^{K-1} \tilde{\mathbf{g}}^{t,k}(i)\|}\Big),$$
$$\overline{\alpha}^t := \frac{1}{M}\sum_{i=1}^{M} \alpha_i^t,$$
$$\tilde{\alpha}^t := \frac{1}{M}\sum_{i=1}^{M} |\alpha_i^t - \overline{\alpha}^t|.$$

The Lipschitz continuity of ∇f:

$$\begin{aligned}
&\mathbb{E}f(\mathbf{w}^{t+1}) \\
&\leq \mathbb{E}f(\mathbf{w}^t) + \mathbb{E}\Big\langle \nabla f(\mathbf{w}^t), \mathbf{w}^{t+1} - \mathbf{w}^t \Big\rangle + \mathbb{E}\frac{L}{2}\|\mathbf{w}^{t+1} - \mathbf{w}^t\|^2 \\
&= \mathbb{E}f(\mathbf{w}^t) + \mathbb{E}\Big\langle \nabla f(\mathbf{w}^t), \frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t + z_i^t \Big\rangle + \frac{L}{2}\mathbb{E}\Big\| \frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t + z_i^t \Big\|^2 \\
&= \mathbb{E}f(\mathbf{w}^t) + \underbrace{\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t \Big\rangle}_{\text{I}} + \frac{L}{2}\underbrace{\mathbb{E}\Big\langle \|\frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t\|^2 \Big\rangle}_{\text{II}} + \frac{L\sigma^2C^2d}{2m^2},
\end{aligned} \tag{19}$$

where d represents dimension of $\mathbf{w}_i^{t,k}$ and the mean of noise z_i^t is zero. Then, we analyze I and II, respectively.
For I, we have

$$\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t \Big\rangle = \Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M} \tilde{\Delta}_i^t - \overline{\Delta}_i^t \Big\rangle + \Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M} \overline{\Delta}_i^t \Big\rangle. \tag{20}$$

Then we bound the two terms in the above equality, respectively. For the first term, we have

$$\begin{aligned}
&\mathbb{E}\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M} \tilde{\Delta}_i^t - \overline{\Delta}_i^t \Big\rangle \\
&\leq \mathbb{E}\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M}\sum_{k=0}^{K-1} \eta|\alpha_i^t - \overline{\alpha}^t|\tilde{\mathbf{g}}^{t,k}(i) \Big\rangle \\
&\leq \frac{\eta K}{M}\sum_{i=1}^{M} \mathbb{E}|\alpha_i^t - \overline{\alpha}^t|\Big\langle \nabla F_i(\mathbf{w}^t), \tilde{\mathbf{g}}^{t,k}(i) \Big\rangle \\
&\overset{a)}{\leq} \frac{\eta K}{M}\sum_{i=1}^{M} \mathbb{E}|\alpha_i^t - \overline{\alpha}^t|\Big(-\frac{1}{2}(\|\nabla F_i(\mathbf{w}^{t,k})\|^2 + \|F_i(\mathbf{w}^{t,k} + \delta; \xi_i)\|^2) + \frac{1}{2}\|\nabla F_i(\mathbf{w}^{t,k} + \delta; \xi_i) - \nabla F_i(\mathbf{w}^{t,k}; \xi_i)\|^2 \Big) \\
&\overset{b)}{\leq} \eta\tilde{\alpha}^t K(\frac{1}{2}L^2\rho^2 - B^2),
\end{aligned} \tag{21}$$

where $\tilde{\alpha}^t = \frac{1}{M}\sum_{i=1}^{M}|\alpha_i^t - \overline{\alpha}^t|$, a) uses $\langle a,b\rangle = -\frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2 + \frac{1}{2}\|a-b\|^2$ and b) bases on assumption 1,3.
For the second term, we have

$$\begin{aligned}
&\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M}\overline{\Delta}_i^t \Big\rangle \\
&\overset{a)}{\leq} \frac{-\overline{\alpha}^t\eta K}{2}\|\nabla f(\mathbf{w}^t)\|^2 - \frac{\overline{\alpha}^t}{2K}\mathbb{E}\Big\|\frac{1}{\overline{\alpha}^t M}\sum_{i=1}^{M}\overline{\Delta}_i^t\Big\|^2 + \frac{\overline{\alpha}^t}{2}\underbrace{\mathbb{E}\Big\|\sqrt{K}\nabla f(\mathbf{w}^t) - \frac{1}{\overline{\alpha}^t M\sqrt{K}}\sum_{i=1}^{M}\overline{\Delta}_i^t\Big\|^2}_{\text{III}},
\end{aligned} \tag{22}$$

where a) uses $\langle a,b\rangle = -\frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2 + \frac{1}{2}\|a-b\|^2$ and $0<\eta<1$. Next, we bound III as follows:

$$\begin{aligned}
\text{III} &= K\mathbb{E}\Big\|\nabla f(\mathbf{w}^t) + \frac{1}{MK}\sum_{i=1}^{M}\sum_{k=0}^{K-1}\nabla\eta F_i(\mathbf{w}^{t,k}+\delta;\xi_i)\Big\|^2 \\
&\leq \frac{1}{M}\sum_{i=1}^{M}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\eta(F_i(\mathbf{w}^{t,k}+\delta;\xi_i) - \nabla F_i(\mathbf{w}^{t,k};\xi_i)) + \eta(\nabla F_i(\mathbf{w}^{t,k};\xi_i) - \nabla F_i(\mathbf{w}^t)) + (1+\eta)\nabla F_i(\mathbf{w}^t)\Big\|^2 \\
&\overset{a)}{\leq} 3K\eta^2L^2\Big(\rho^2 + \mathbb{E}\|\mathbf{w}^{t,k} - \mathbf{w}^t\|^2 + 2B^2\Big) \\
&\overset{b)}{\leq} 3K\eta^2L^2\Big[\rho^2 + 5K\eta^2\Big(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\mathbf{w}^t)\|^2\Big) + 24K^3\eta^4L^4\rho^2 + B^2\Big],
\end{aligned} \tag{23}$$

where $0<\eta<1$, a) and b uses assumption 1, 3 and lemma 3, respectively.
For II, we uses lemma 5. Then, combining Eq. 12-16, we have

$$\begin{aligned}
\mathbb{E}f(\mathbf{w}^{t+1}) \leq{}& \mathbb{E}f(\mathbf{w}^t) + \eta\tilde{\alpha}_t K(\frac{1}{2}L^2\rho^2 - B^2) - \frac{\overline{\alpha}^t\eta K}{2}\|\nabla f(\mathbf{w}^t)\|^2 - \frac{\eta\overline{\alpha}^t}{2K}\mathbb{E}\Big\|\frac{1}{\eta\overline{\alpha}^t M}\sum_{i=1}^{M}\overline{\Delta}_i^t\Big\|^2 \\
&+ \frac{3\overline{\alpha}^t\eta^2L^2K}{2}\Big[\rho^2 + 5K\eta^2\Big(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\mathbf{w}^t)\|^2\Big) \\
&+ 24K^3\eta^4L^4\rho^2 + B^2\Big] + \frac{3\eta^2KL(L^2\rho^2+B^2)}{2} + \frac{L\sigma^2C^2d}{2m^2}.
\end{aligned} \tag{24}$$

When $\eta \leq \frac{1}{3\sqrt{KL}}$, the inequality is

$$\begin{aligned}
\mathbb{E}f(\mathbf{w}^{t+1}) \leq{}& \mathbb{E}f(\mathbf{w}^t) - \frac{\overline{\alpha}^t\eta K}{2}\mathbb{E}\|\nabla f(\mathbf{w}^t)\|^2 + \frac{\tilde{\alpha}^t\eta KL^2\rho^2}{2} + \frac{3\overline{\alpha}^t\eta^2KL^2\rho^2}{2} - \tilde{\alpha}^t\eta KB^2 \\
&+ \frac{15\overline{\alpha}^tK\eta^4L^2}{2}\Big(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\mathbf{w}^t)\|^2\Big) + 36\eta^6K^4L^6\rho^2 \\
&+ \frac{3\eta^2KL(L^2\rho^2+B^2)}{2} + \frac{L\sigma^2C^2d}{2m^2}.
\end{aligned} \tag{25}$$

Sum over t from 1 to T, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\overline{\alpha}^t\|f(\mathbf{w}^t)\|^2\Big] \leq{}& \frac{2L(f(\mathbf{w}^1)-f^*)}{\sqrt{KT}} + \frac{1}{T}\sum_{t=1}^{T}\tilde{\alpha}^tL^2\rho^2 - 2\tilde{\alpha}^tB^2 + 30\eta^2L^2\frac{1}{T}\sum_{t=1}^{T}\overline{\alpha}^t\Big(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2+6L^2\rho^2)\Big) \\
&+ 72\eta^4K^3L^6\rho^2 + 3\eta L(L^2\rho^2+B^2) + \frac{L\sigma^2C^2d}{\eta m^2K}
\end{aligned} \tag{26}$$

Assume the local adaptive learning rate satisfies $\eta = \mathcal{O}(1/L\sqrt{KT})$, both $\frac{1}{T}\sum_{t=1}^{T}\tilde{\alpha}^t$ and $\frac{1}{T}\sum_{t=1}^{T}\overline{\alpha}^t$ are two important parameters for measuring the impact of clipping. Meanwhile, both $\frac{1}{T}\sum_{t=1}^{T}\tilde{\alpha}^t$ and $\frac{1}{T}\sum_{t=1}^{T}\overline{\alpha}^t$ are also bounded by 1. Then,

our result is

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\overline{\alpha}^t\left\|\nabla f\left(\mathbf{w}^t\right)\right\|^2\right] \leq \underbrace{\mathcal{O}\left(\frac{2L(f(\mathbf{w}^1)-f^*)}{\sqrt{KT}}+\frac{\sigma_l^2L^2\rho}{KT}\right)}_{\text{From FedSAM}} + \underbrace{\underbrace{\mathcal{O}\left(\sum_{t=1}^{T}(\frac{\overline{\alpha}^t\sigma_g^2}{T^2}+\frac{\tilde{\alpha}^tL^2\rho^2}{T})\right)}_{\text{Clipping}} + \underbrace{\mathcal{O}\left(\frac{L^2\sqrt{T}\sigma^2C^2d}{m^2\sqrt{K}}\right)}_{\text{Adding noise}}}_{\text{From operations for DP}}. \tag{27}$$

Assume the perturbation amplitude ρ proportional to the learning rate, e.g., $\rho=\mathcal{O}(\frac{1}{\sqrt{T}})$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\overline{\alpha}^t\left\|\nabla f\left(\mathbf{w}^t\right)\right\|^2\right] \leq \underbrace{\mathcal{O}\left(\frac{2L(f(\mathbf{w}^1)-f^*)}{\sqrt{KT}}+\frac{L^2\sigma_l^2}{KT^2}\right)}_{\text{From FedSAM}} + \underbrace{\underbrace{\mathcal{O}\left(\sum_{t=1}^{T}(\frac{\overline{\alpha}^t\sigma_g^2}{T^2}+\frac{\tilde{\alpha}^tL^2}{T^2})\right)}_{\text{Clipping}} + \underbrace{\mathcal{O}\left(\frac{L^2\sqrt{T}\sigma^2C^2d}{m^2\sqrt{K}}\right)}_{\text{Adding noise}}}_{\text{From operations for DP}}. \tag{28}$$

□