In this part, we will introduce the gradient perturbation, the proofs of the major equations and the theorem, and some extra experiments. In Section A, we provide a explanation for understanding the gradient perturbation step in our proposed FedSpeed. In Section C, we provide the full proofs of the major equation in the text, some main lemmas and the theorem. In Section B, we provide the details of the implementation of the experiments including the setups, dataset, hyper-parameters and some extra experiments.

A GRADIENT PERTURBATION

A.1 UNDERSTANDING OF GRADIENT PERTURBATION

We propose the gradient perturbation in the local training stage instead of the traditional stochastic gradient, which merges an extra gradient ascent step to the vanilla gradient by a hyper-parameter α. While its ascent step usually approximates the worst point in the neighbourhood. This has been studied in many previous works, e.g. for the form of extra gradient and the sharpness aware minimization. In our studies, we perform the extra gradient ascent step instead of the descent step in extra gradient method. It also could be considered as a variant of the sharpness aware minimization method via weighted averaging the ascent step gradient and the vanilla gradient, instead of the normalized gradient. Here we illustrate the implicit of this quasi-gradient $\tilde{\mathbf{g}}$ in our proposed FedSpeed and explain the positive efficiency for the local training from the perspective of objective functions.

Firstly we consider to minimize the non-convex problem $\mathcal{L}_p(\mathbf{x})$. To approach the stationary point of $\mathcal{L}_p$, we can simply introduce a penalized gradient term as a extra loss in $\mathcal{L}_p$, which is to solve the problem $\min_{\mathbf{x}}\{\mathcal{L}(\mathbf{x}) \triangleq \mathcal{L}_p(\mathbf{x}) + \frac{\beta}{2}\|\nabla\mathcal{L}_p(\mathbf{x})\|^2\}$. The final optimization target is consistent with the vanilla target, while penalizing gradient term can approach a flatten minimal empirically. We compute the gradient form as follows:

$$\nabla\mathcal{L}(\mathbf{x}) = \nabla\mathcal{L}_p(\mathbf{x}) + \frac{\beta}{2}\nabla\|\nabla\mathcal{L}_p(\mathbf{x})\|^2 = \nabla\mathcal{L}_p(\mathbf{x}) + \beta\nabla^2\mathcal{L}_p(\mathbf{x})\cdot\nabla\mathcal{L}_p(\mathbf{x}). \tag{1}$$

The update in Equation (1) contains second-order Hessian information, which involves a huge amount of parameters for calculation. To further simplify the updates, we consider an approximation for the gradient form. We expand the function $\mathcal{L}_p$ via Taylor expansion as:

$$\mathcal{L}_p(\mathbf{x}+\Delta) = \mathcal{L}_p(\mathbf{x}) + \nabla\mathcal{L}_p(\mathbf{x})\Delta + \frac{1}{2}\Delta^T\nabla^2\mathcal{L}_p(\mathbf{x})\Delta + \mathcal{R}_\Delta,$$

where $\mathcal{R}_\Delta = \mathcal{O}(\|\Delta\|^2)$ is the infinitesimal to $\|\Delta\|^2$, which is directly omitted in our approximation.

Thus we have the gradient form on Δ as:

$$\nabla\mathcal{L}_p(\mathbf{x}+\Delta) \approx \nabla\mathcal{L}_p(\mathbf{x}) + \nabla^2\mathcal{L}_p(\mathbf{x})\Delta.$$

$\mathcal{R}_\Delta$ is relevant to Δ. We set the $\Delta = \rho\nabla\mathcal{L}_p(\mathbf{x})$ and then we have:

$$\nabla^2\mathcal{L}_p(\mathbf{x})\nabla\mathcal{L}_p(\mathbf{x}) \approx \frac{1}{\rho}\big(\nabla\mathcal{L}_p\big(\mathbf{x}+\rho\nabla\mathcal{L}_p(\mathbf{x})\big) - \nabla\mathcal{L}_p(\mathbf{x})\big). \tag{2}$$

Thus we connect Equation (1) and Equation (2), we have:

$$\begin{aligned}
\nabla\mathcal{L}(\mathbf{x}) &= \nabla\mathcal{L}_p(\mathbf{x}) + \beta\nabla^2\mathcal{L}_p(\mathbf{x})\cdot\nabla\mathcal{L}_p(\mathbf{x}) \\
&\approx \nabla\mathcal{L}_p(\mathbf{x}) + \frac{\beta}{\rho}\big(\nabla\mathcal{L}_p\big(\mathbf{x}+\rho\nabla\mathcal{L}_p(\mathbf{x})\big) - \nabla\mathcal{L}_p(\mathbf{x})\big) \\
&= \big(1-\frac{\beta}{\rho}\big)\nabla\mathcal{L}_p(\mathbf{x}) + \frac{\beta}{\rho}\nabla\mathcal{L}_p\big(\mathbf{x}+\rho\nabla\mathcal{L}_p(\mathbf{x})\big) \\
&= (1-\alpha)\nabla\mathcal{L}_p(\mathbf{x}) + \alpha\nabla\mathcal{L}_p\big(\mathbf{x}+\rho\nabla\mathcal{L}_p(\mathbf{x})\big).
\end{aligned}$$

Here we can see that the balance weight α in our proposed method is actually the ratio of the gradient penalized weight β and the gradient ascent step size ρ. To fix the step size ρ, increasing α

means increasing the gradient penalized weight β, which facilitates searching for a flatten stationary point to improve the generalization performance. While the second term of $\nabla\mathcal{L}(\mathbf{x})$ can not be directly computed for its nested form, we approximate the second term with the chain rule as follows:

$$\nabla\mathcal{L}_p\big(\mathbf{x}+\rho\nabla\mathcal{L}_p(\mathbf{x})\big) \approx \nabla\mathcal{L}_p(\theta)|_{\theta=\mathbf{x}+\rho\nabla\mathcal{L}_p(\mathbf{x})}.$$

Finally we have:

$$\nabla\mathcal{L}(\mathbf{x}) \approx (1-\alpha)\nabla\mathcal{L}_p(\mathbf{x}) + \alpha\nabla\mathcal{L}_p(\theta)|_{\theta=\mathbf{x}+\rho\nabla\mathcal{L}_p(\mathbf{x})}. \tag{3}$$

The Equation (3) provides an understanding for the weighted quasi gradient $\tilde{\mathbf{g}}$ on the local training stage in our proposed FedSpeed. We select an appropriate $0 \leq \beta \leq \rho$ to satisfy the update of perturbation gradient. It executes a gradient ascent step firstly with the step size ρ to $\check{\mathbf{x}}$. Then it generates the stochastic gradient by the same sampled mini-batch data as the ascent step at $\check{\mathbf{x}}$. The quasi-gradient is merged as Equation (3) to execute the gradient descent step.

This is just a simple approximation for the gradient perturbation to help for understanding the implicit of the quasi-gradient and its performance in the training stage. Actually the error of the approximation depends a lot on ρ. The smaller ρ, the higher the accuracy of this estimation, but the smaller ρ, the less efficient the optimizer performs.

B EXPERIMENTS

B.1 SETUPS

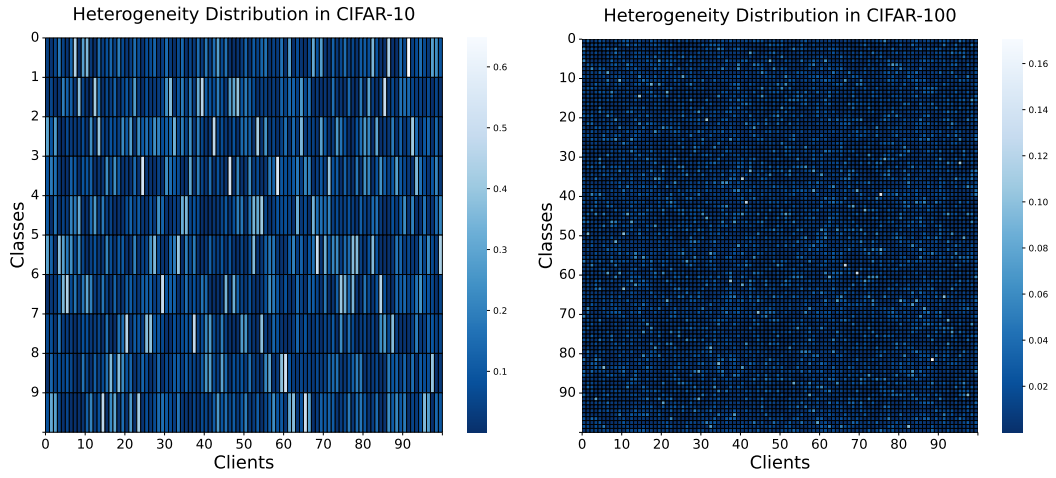
Table 1: Dataset introductions.

Dataset	Training Data	Test Data	Class	Size
CIFAR-10	50,000	10,000	10	3×32×32
CIFAR-100	50,000	10,000	100	3×32×32
TinyImagenet	100,000	10,000	200	3×64×64

Dataset and Backbones. Extensive experiments are tested on CIFAR-10/100 dataset. We test on the two different settings as 10% participation of total 100 clients and 2% participation of total 500 clients. CIFAR-10 dataset contains 50,000 training data and 10,000 test data in 10 classes. Each data sample is a 3×32×32 color image. CIFAR-100 Krizhevsky et al. (2009) includes 50,000 training data and 10,000 test data in 100 classes as 500 training samples per class. TinyImagenet involves 100,000 training images and 10,000 test images in 200 classes for 3×64×64 color images, as shown in Table 1. To fairly compare with the other baselines, we train and test the performance on the standard ResNet-18 He et al. (2016) backbone with the 7×7 filter size in the first convolution layer as implemented in the previous works, e.g. for Karimireddy et al. (2020); Durmus et al. (2021); Xu et al. (2021). We follow the Hsieh et al. (2020) to replace the batch normalization layer with group normalization layer Wu & He (2018), which can be aggregated directly by averaging. These are all common setups in many previous works.

Dataset Partitions. To fairly compare with the other baselines, we follow the Hsu et al. (2019) to introduce the heterogeneity via splitting the total dataset by sampling the label ratios from the Dirichlet distribution. An additional parameter is used to control the level of the heterogeneity of the entire data partition. In order to visualize the distribution of heterogeneous data, we make the heat maps of the label distribution in different dataset, as shown in Figure 1. Since the heat map of 500 clients cannot be displayed normally, we show 100 clients case. It could be seen that for heterogeneity weight equals to 0.6, about 10% to 20% of the categories dominate on each client, which is white block in the Figure 1. The IID dataset is totally averaged in each client.

Data Argumentation. For CIFAR-10/100, we follow the implementation in the Karimireddy et al. (2020); Durmus et al. (2021) to normalize the pixel value within a specific mean and std value in our code, which are [0.491, 0.482, 0.447] for mean, [0.247, 0.243, 0.262] for std and [0.5071, 0.4867, 0.4408] for mean, [0.2675, 0.2565, 0.2761] for std. We randomly flip the training samples and randomly crop the images enlarged with the padding equal to 4. For TinyImagenet, the same argumentation is applied except for the padding equal to 8.



(a) CIFAR-10. (b) CIFAR-100.

Figure 1: Heat maps for different dataset under heterogeneity weight equals to 0.6 for Dirichlet distribution.

Baselines. FedAvg McMahan et al. (2017) is proposed as the basic framework in the federated learning. And FedOpt improves it as a two-stage optimizer with a local and global optimizer update alternatively. Yang et al. (2021) proves a specific η_g (not the average weight) can achieve faster convergence (non-dominant term). FedAdam Reddi et al. (2020) utilizes a adaptive optimizer on the global server and SGD optimizer on the clients, which average the averaged local gradients as a quasi-gradient for global server to implement the adaptive update. SCAFFOLD Karimireddy et al. (2020) applies the variance reduction technique , i.e. SVRG, to approximate the global gradient as the averaged local gradients and transfer an extra variable to the client per round. This implementation can accelerate the convergence rate of the non-dominant term theoretically and achieve a high performance empirically. FedCM Xu et al. (2021) proposes a client-level momentum to merge the global update as a momentum buffer to the local updates, which extremely reduces the local consistency. Though it introduces a unpredictable biases into the local updates, it achieves the SOTA performance ahead of other methods. FedProx Sahu et al. (2018) implements the prox-point optimizer into the FL framework on local updates with a regularization prox-term regularizer. It limits the local updates towards the initial point at the start of each local stage. Many previous works have analyzed its advantages and weaknesses. Durmus et al. (2021); Wang et al. (2022); Gong et al. (2022) use different variants of primal-dual method into FL and achieve nice satisfactory in the FL framework. It does not need a heterogeneity bounded assumption theoretically, which requires a high local convergence guarantees. Our proposed FedSpeed achieve the same convergence rate without assuming the local exact solution and we provide the local interval bound to achieve this faster convergence. Both theoretical analysis and empirical results verifies the performance of our proposed FedSpeed.

B.2 Experiments

B.3 Hyer-parameters

Hyper-parameters Selections. We fix the local learning rate as 0.1 and global learning rate as 1.0 for average, except for the FedAdam which is applied 0.1. The penalized weight of prox-term in FedProx, FedDyn, FedADMM and FedSpeed is selected from the [0.001, 0.01, 0.1, 0.5]. The learning rate decay is fixed as 0.998 expect for the FedDyn, FedADMM and FedSpeed is selected from [0.998, 0.999, 0.9995, 0.99995]. The perturbation weight is selected from [0, 0.5, 0.75, 0.875, 0.9375, 1]. The batchsize is selected from [20, 50]. The local interval K is selected from [1, 2, 5, 10, 20]. For the specific parameters in FedAdam, the momentum weight is set as 0.1 and the second order momentum weight is set as 0.01. The minimal value is set as 0.001 to prevent the calculation of dividing by 0. The client-level momentum weight of FedCM is set as 0.1.

Table 2: Communication rounds required to achieve the target accuracy. On CIFAR-10/100 it trains 1,500 rounds and on TinyImagenet it trains 3,000 rounds. "-" means the test accuracy can not achieve the target accuracy within the fixed training rounds. **DIR** represents for the Dirichlet distribution with the heterogeneity weight equal to 0.6. Local interval K is set as 5 on CIFAR-10 (100-10%) and 2 on others. Other hyper-parameters are introduced above.

Dataset	CIFAR-10 (100-10%)				CIFAR-10 (500-2%)			
Heterogeneity	IID.		DIR.		IID.		DIR.	
Target Acc. (%)	80.0	85.0	80.0	85.0	75.0	82.5	75.0	82.5
FedAvg	344	-	472	-	772	-	1357	-
FedProx	338	-	465	-	720	-	1151	-
FedAdam	324	1343	689	-	613	1476	878	-
SCAFFOLD	207	654	272	-	628	-	967	-
FedCM	**109**	620	192	1092	**325**	**1160**	**449**	**1399**
FedDyn	**121**	**400**	166	-	547	-	673	-
FedADMM	169	917	**174**	**756**	505	1440	687	-
FedSpeed	136	**280**	**169**	**380**	**495**	**926**	**662**	**1148**

Dataset	CIFAR-100 (500-2%)				TinyImagenet (500-2%)			
Heterogeneity	IID.		DIR.		IID.		DIR.	
Target Acc.	40.0	50.0	40.0	50.0	33.0	40.0	33.0	40.0
FedAvg	1013	-	-	-	1615	-	-	-
FedProx	957	-	-	-	1588	-	-	-
FedAdam	614	1277	847	-	1151	2495	1584	-
SCAFFOLD	720	-	784	-	949	-	1187	-
FedCM	**505**	**1150**	**526**	**1336**	**661**	**1360**	**817**	**1843**
FedDyn	661	-	703	-	1419	-	2559	-
FedADMM	687	-	715	-	921	-	2711	-
FedSpeed	**522**	**973**	**541**	**1038**	**684**	**1373**	**962**	**1885**

B.3.1 BEST PERFORMING HYPER-PARAMETERS.

For fair comparison, the learning rate is fixed for all the methods.

For CIFAR-10 dataset, we select the batchsize as 50 for 100 clients and 20 for 500 clients. The total dataset is 50,000 and there are 100 images under a single client if it is set as 500 clients. Thus we decay it to 20 for 5 iterations per local epoch. The local epochs is set as 5, the same as the experiments of Karimireddy et al. (2020); Durmus et al. (2021); Xu et al. (2021) etc. and their performance is matching. We select the local interval K as 5. The prox-term weight is selected as 0.1. The learning rate decay is selected as 0.9995 for prox-term based methods. We train the total dataset for 1,500 communication rounds.

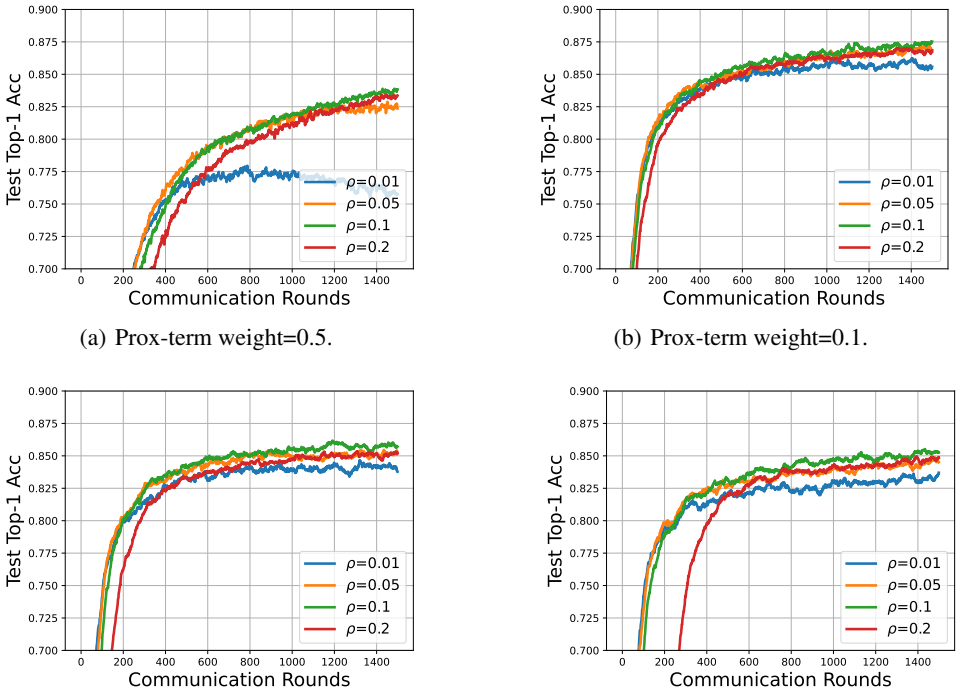
For CIFAR-100 dataset, we select the 500 clients with 2% participation ratio in the experiments. Thus for each hyper-parameters we fine-tune a little. The batchsize is selected as 20 to avoid too little iterations per local epoch. The local epochs is set as 2 for the final results comparison. The ablation study on local interval K indicates that our proposed FedSpeed outperforms significantly than other methods when K is large. Thus to compare the performance more clearly, we select the 2 as the local

epochs. We decay the prox-term weight as 0.01 for prox-term based methods. The learning rate decay is selected as 0.99995 for prox-based methods. We train 1,500 rounds and then test the performance.

For TinyImagenet dataset, the most selections are the same as for the CIFAR-100 dataset. The prox-term weight is selected as 0.1 and the learning rate decay is selected as 0.9995. Total 3,000 communication rounds are implemented in the training stage.

B.3.2 SPEED COMPARISON.

Table 2 shows the communication rounds required to achieve the target test accuracy. At the beginning of training, FedCM performs faster than others and usually achieve a high accuracy finally. FedSpeed is faster in the middle and late stages of training. We bold the data for the top-2 in each test and generally FedCM and FedSpeed significantly performs well on the training speed.



(a) Prox-term weight=0.5.

(b) Prox-term weight=0.1.

(c) Prox-term weight=0.01.

(d) Prox-term weight=0.001.

Figure 2: Performance of different ascent step size ρ under different prox-term weights of $[0.001, 0.01, 0.1, 0.5]$.

Figure 2 shows the performance of different learning rate decay and prox-term weight for FedSpeed.

B.3.3 TIME COST

Table 3: Training wall-clock time comparison.

α_1	Times (s/Round)	Rounds	Total (s)	Cost Ratio
FedAvg	10.44	-	-	-
FedProx	11.33	-	-	-
FedAdam	14.74	1343	19795.8	4.31×
SCAFFOLD	14.34	654	9378.3	2.03×
FedCM	13.22	622	8222.8	1.78×
FedDyn	14.11	400	5644.0	1.22×
FedSpeed	16.42	281	4614.0	1×

We test the time on the A100-SXM4-40GB GPU and show the performance in the Table B.3.3. Experimental setups are the same as the CIFAR-10 10% participation among total 100 clients on the DIR-0.6 dataset. The rounds in the table are the communication rounds required that the test accuracy achieves accuracy 85%. "-" means it can not achieve the target accuracy.

FedSpeed is slower due to the requirement of computing an extra gradient. So it gets slower in one single update, approximately $1.57\times$ wall-clock time costs than FedAvg. But its convergence process is very fast. For the final convergence speed, FedSpeed still has a considerable advantage over other algorithms. The issue is possibly one of the improvements for FedSpeed in the future. For example, introduces a single-call gradient method to save half the costs during backpropagation. We are also currently trying to introduce new module to save the cost.

B.3.4 DIFFERENT HETEROGENEITY.

Table 4: Comparison on different heterogeneous dataset.

α_1	IID	Dir-0.6	Dir-0.3	Drops (i.i.d. $>$ Dir-0.6)	Drops (Dir-0.6 $>$ Dir-0.3)
FedAvg	77.01	75.21	71.96	1.80	3.25
FedAdam	82.92	80.55	76.87	2.37	3.68
SCAFFOLD	80.11	77.71	74.34	2.40	3.37
FedCM	84.20	83.48	81.02	0.72	2.46
FedDyn	83.36	80.57	77.33	2.79	3.24
FedSpeed	85.80	84.79	82.68	1.01	2.11

We test on the Dir-0.3 setups on CIFAR-10 and show the results as Table B.3.4, the other settings are the same as the test in the text. The (i.i.d. $>$ Dir-0.6) is the difference between the IID dataset and the Dir-0.6 dataset and (Dir-0.6 $>$ Dir-0.3) is the difference between the Dir-0.6 dataset and the DIR-0.3 dataset. FedSpeed can outperform the others on the Dir-0.3 setups whose heterogeneity is much stronger than Dir-0.6 setups. the heterogeneity becomes stronger, FedSpeed can still maintain a stable generalization performance. The correction term helps to correct the biases during the local training, while the gradient perturbation term helps to resist the local over-fitting on the heterogeneous dataset. FedSpeed can benefit from avoiding falling into the biased optima.

B.3.5 ABLATION STUDIES

Table 5: Comparison on different heterogeneous dataset.

Prox-term	Prox-correction term	Gradient perturbation	Accuracy ($\%$)
-	-	-	81.92
$\surd$	-	-	82.24
$\surd$	$\surd$	-	83.94
$\surd$	-	$\surd$	83.88
$\surd$	$\surd$	$\surd$	85.70

From the practical training point of view, compared with the vanilla FedAvg, FedSpeed adds three main modules: (1) prox-term, (2) prox-correction term, and (3) gradient perturbation. We test the performance of 500 communication rounds of the different combination of the modules above on the CIFAR-10 with the settings of 10% participating ratio of total 100 clients. The TableB.3.5 shows their performance.

From the table above, we can clearly see the performance of different modules. The prox-term is proposed by the FedProx. But due to some issues we point out in our paper, this term has also a negative impact on the performance in FL. When the prox-correction term is introduced in, it improves the performance from 82.24% to 83.94%. When the gradient perturbation is introduced in, it improves the performance from 82.24% to 83.88%. While FedSpeed applies them together and achieves a 3.46% improvement.

Different performance of these modules:

As introduced in our paper, the prox-term simply performs as a balance between the local and global solutions, and there still exists the non-vanishing inconsistent biases among the local solutions, i.e., the local solutions are still largely deviated from each other, implying that local inconsistency is still not eliminated. Thus we utilize the prox-correction term to correct the inconsistent biases during the local training. About the function of gradient perturbation, we refer to a theoretical explanation in the main text, and its proof is provided in the supplementary material due to the space limitations. This perturbation is similar to utilize a penalized gradient term to the objective function during local optimization process. The additional penalty will bring better properties to the local state, e.g. for flattened minimal and smoothness. For federated learning, the smoother the local minima is, the more flatness the model merged on the server will be. FedSpeed benefits from these two modules to improve the performance and achieves the SOTA results.

C Proofs for Analysis

In this part we will demonstrate the proofs of all formula mentioned in this paper. Each formula is presented in the form of a lemma.

C.1 Proof of Equation (2)

Equation (2) shows the update in the total local training stage.

Lemma C.1 *For $\forall\, \mathbf{x}_{i,k}^t \in \mathbb{R}^d$ and $i \in \mathcal{S}^t$, we denote $\delta_{i,k}^t = \mathbf{x}_{i,k}^t - \mathbf{x}_{i,k-1}^t$ with setting $\delta_{i,0}^t = 0$, and $\Delta_{i,K}^t = \sum_{k=0}^K \delta_{i,k}^t = \mathbf{x}_{i,K}^t - \mathbf{x}_{i,0}^t$, under the update rule in Algorithm Algorithm 1, we have:*

$$\Delta_{i,K}^t = -\lambda\gamma \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t + \gamma\lambda \hat{\mathbf{g}}_i^{t-1}, \tag{4}$$

where $\sum_{k=0}^{K-1} \gamma_k = \sum_{k=0}^{K-1} \frac{\eta_l}{\lambda}\big(1-\frac{\eta_l}{\lambda}\big)^{K-1-k} = \gamma = 1-(1-\frac{\eta_l}{\lambda})^K$.

Proof 1 *According to the update rule of Line.11 in Algorithm Algorithm 1, we have:*

$$\begin{aligned}\delta_k &= \Delta_{i,k}^t - \Delta_{i,k-1}^t = \mathbf{x}_{i,k}^t - \mathbf{x}_{i,k-1}^t \\ &= -\eta_l\big(\tilde{\mathbf{g}}_{i,k-1}^t - \hat{\mathbf{g}}_i^{t-1} + \frac{1}{\lambda}(\mathbf{x}_{i,k-1}^t - \mathbf{x}_{i,0}^t)\big) = -\eta_l(\tilde{\mathbf{g}}_{i,k-1}^t - \hat{\mathbf{g}}_i^{t-1} + \frac{1}{\lambda}\Delta_{i,k-1}^t).\end{aligned}$$

Then We can formulate the iterative relationship of $\Delta_{i,k}^t$ as:

$$\Delta_{i,k}^t = \Delta_{i,k-1}^t - \eta_l(\tilde{\mathbf{g}}_{i,k-1}^t - \hat{\mathbf{g}}_i^{t-1} + \frac{1}{\lambda}\Delta_{i,k-1}^t) = (1-\frac{\eta_l}{\lambda})\Delta_{i,k-1}^t - \eta_l(\tilde{\mathbf{g}}_{i,k-1}^t - \hat{\mathbf{g}}_i^{t-1}).$$

Taking the iteration on k and we have:

$$\begin{aligned}\mathbf{x}_{i,K}^t - \mathbf{x}_{i,0}^t = \Delta_{i,K}^t &= (1-\frac{\eta_l}{\lambda})^K \Delta_{i,0}^t - \eta_l \sum_{k=0}^{K-1} (1-\frac{\eta_l}{\lambda})^{K-1-k}(\tilde{\mathbf{g}}_{i,k}^t - \hat{\mathbf{g}}_i^{t-1}) \\ &\overset{(a)}{=} -\eta_l \sum_{k=0}^{K-1} (1-\frac{\eta_l}{\lambda})^{K-1-k}(\tilde{\mathbf{g}}_{i,k}^t - \hat{\mathbf{g}}_i^{t-1}) \\ &= -\lambda \sum_{k=0}^{K-1} \frac{\eta_l}{\lambda}(1-\frac{\eta_l}{\lambda})^{K-1-k}(\tilde{\mathbf{g}}_{i,k}^t - \hat{\mathbf{g}}_i^{t-1}) \\ &= -\lambda \sum_{k=0}^{K-1} \frac{\eta_l}{\lambda}(1-\frac{\eta_l}{\lambda})^{K-1-k}\tilde{\mathbf{g}}_{i,k}^t + \big(1-(1-\frac{\eta_l}{\lambda})^K\big)\lambda\hat{\mathbf{g}}_i^{t-1} \\ &= -\lambda\gamma \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t + \gamma\lambda\hat{\mathbf{g}}_i^{t-1}.\end{aligned}$$

(a) applies $\Delta_{i,0}^t = \delta_{i,0}^t = 0$.

C.2 PROOF OF EQUATION (3)

Equation (3) shows the update of the prox-correction term, which utilizes the weighted sum of the previous local offsets as a bias controller for eliminating the non-vanishing bias resulting from the prox-term.

Lemma C.2 *Under the update rule in Algorithm Algorithm 1, we have:*

$$\hat{\mathbf{g}}_i^t = (1-\gamma)\hat{\mathbf{g}}_i^{t-1} + \gamma \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t. \tag{5}$$

where $\sum_{k=0}^{K-1} \gamma_k = \sum_{k=0}^{K-1} \frac{\eta_l}{\lambda}\big(1-\frac{\eta_l}{\lambda}\big)^{K-1-k} = \gamma = 1-(1-\frac{\eta_l}{\lambda})^K$.

Proof 2 *According to the update rule of Line.13 in Algorithm Algorithm 1, we have:*

$$\begin{aligned}
\hat{\mathbf{g}}_i^t &= \hat{\mathbf{g}}_i^{t-1} - \frac{1}{\lambda}(\mathbf{x}_{i,K}^t - \mathbf{x}_{i,0}^t) \\
&\overset{(a)}{=} \hat{\mathbf{g}}_i^{t-1} + \frac{\eta_l}{\lambda} \sum_{k=0}^{K-1} \big(1-\frac{\eta_l}{\lambda}\big)^{K-1-k} (\tilde{\mathbf{g}}_{i,k}^t - \hat{\mathbf{g}}_i^{t-1}) \\
&= \hat{\mathbf{g}}_i^{t-1} + \frac{\eta_l}{\lambda} \sum_{k=0}^{K-1} \big(1-\frac{\eta_l}{\lambda}\big)^{K-1-k} \tilde{\mathbf{g}}_{i,k}^t - \frac{\eta_l}{\lambda}\Big(\sum_{k=0}^{K-1} \big(1-\frac{\eta_l}{\lambda}\big)^{K-1-k}\Big) \hat{\mathbf{g}}_i^{t-1} \\
&= \hat{\mathbf{g}}_i^{t-1} + \frac{\eta_l}{\lambda} \sum_{k=0}^{K-1} \big(1-\frac{\eta_l}{\lambda}\big)^{K-1-k} \tilde{\mathbf{g}}_{i,k}^t - \frac{\eta_l}{\lambda} \frac{1-(1-\frac{\eta_l}{\lambda})^K}{\frac{\eta_l}{\lambda}} \hat{\mathbf{g}}_i^{t-1} \\
&= (1-\frac{\eta_l}{\lambda})^K \hat{\mathbf{g}}_i^{t-1} + \frac{\eta_l}{\lambda} \sum_{k=0}^{K-1} \big(1-\frac{\eta_l}{\lambda}\big)^{K-1-k} \tilde{\mathbf{g}}_{i,k}^t \\
&= (1-\gamma)\hat{\mathbf{g}}_i^{t-1} + \gamma \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t.
\end{aligned}$$

(a) applies the Lemma C.1.

C.3 PROOF OF EQUATION (4) AND (5)

Lemma C.3 *Considering the* $\mathbf{u}^{t+1} = \frac{1}{m}\sum_{i\in[m]} \mathbf{x}_{i,K}^t$ *is the mean averaged parameters among the last iteration of local clients at time* t*, the auxiliary sequence* $\big\{\mathbf{z}^t = \mathbf{u}^t + \frac{1-\gamma}{\gamma}(\mathbf{u}^t - \mathbf{u}^{t-1})\big\}_{t>0}$ *satisfies the update rule as:*

$$\mathbf{z}^{t+1} = \mathbf{z}^t - \lambda \frac{1}{m} \sum_{i\in[m]} \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t. \tag{6}$$

Proof 3 *Firstly, according to the lemma C.1 and Line.14 and Line.16 in Algorithm 1, we have:*

$$\begin{aligned}
\mathbf{u}^{t+1} - \mathbf{u}^t &= \frac{1}{m} \sum_{i\in[m]} (\mathbf{x}_{i,K}^t - \mathbf{x}_{i,K}^{t-1}) \\
&= \frac{1}{m} \sum_{i\in[m]} (\mathbf{x}_{i,K}^t - \mathbf{x}_{i,0}^t - \lambda \hat{\mathbf{g}}_i^{t-1}) \\
&= \frac{1}{m} \sum_{i\in[m]} (-\lambda\gamma \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t + \lambda\gamma \hat{\mathbf{g}}_i^t - \lambda \hat{\mathbf{g}}_i^{t-1})
\end{aligned}$$

$$= -\lambda \frac{1}{m} \sum_{i\in[m]} \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \left(\gamma \tilde{\mathbf{g}}_{i,k}^t + (1-\gamma)\hat{\mathbf{g}}_i^{t-1}\right).$$

This could be considered as a momentum-like term with the coefficient of γ. Here we define a virtual observation sequence $\{\mathbf{u}^t\}$ and its update rule is:

$$\mathbf{u}_{i,k+1}^t = \mathbf{u}_{i,k}^t - \lambda \frac{\gamma_k}{\gamma} \left(\gamma \tilde{\mathbf{g}}_{i,k}^t + (1-\gamma)\hat{\mathbf{g}}_i^{t-1}\right),$$
$$\mathbf{u}_{i,0}^{t+1} = \mathbf{u}^{t+1} = \frac{1}{m} \sum_{i\in[m]} \mathbf{u}_{i,K}^t.$$

According to the lemma C.2 and above update rule, we can get that:

$$\begin{aligned}
\hat{\mathbf{g}}_i^t &= (1-\gamma)\hat{\mathbf{g}}_i^{t-1} + \gamma \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t \\
&= -\frac{1}{\lambda}(\mathbf{u}_{i,K}^t - \mathbf{u}_{i,0}^t) - \gamma \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t + \gamma \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t = -\frac{1}{\lambda}(\mathbf{u}_{i,K}^t - \mathbf{u}_{i,0}^t).
\end{aligned}$$

This function indicates that the virtual sequence $\mathbf{u}^t$ could be considered as a momentum-based update method with a global correction term to guide the local update, and the correction term is calculated from the offset of the virtual observation sequence during the training process at round t.

Then we expand the the auxiliary sequence $\mathbf{z}^t$ as:

$$\begin{aligned}
\mathbf{z}^{t+1} - \mathbf{z}^t &= (\mathbf{u}^{t+1} - \mathbf{u}^t) + \frac{1-\gamma}{\gamma}(\mathbf{u}^{t+1} - \mathbf{u}^t) - \frac{1-\gamma}{\gamma}(\mathbf{u}^t - \mathbf{u}^{t-1}) \\
&= \frac{1}{\gamma}(\mathbf{u}^{t+1} - \mathbf{u}^t) - \frac{1-\gamma}{\gamma}(\mathbf{u}^t - \mathbf{u}^{t-1}) \\
&= -\lambda \frac{1}{m} \sum_{i\in[m]} \left(\left(\sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t \right) + \frac{1-\gamma}{\gamma} \hat{\mathbf{g}}_i^{t-1} \right) - \frac{1-\gamma}{\gamma}(\mathbf{u}^t - \mathbf{u}^{t-1}) \\
&= -\lambda \frac{1}{m} \sum_{i\in[m]} \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t - \frac{1-\gamma}{\gamma} \frac{1}{m} \sum_{i\in[m]} \lambda \hat{\mathbf{g}}_i^{t-1} - \frac{1-\gamma}{\gamma}(\mathbf{u}^t - \mathbf{u}^{t-1}) \\
&= -\lambda \frac{1}{m} \sum_{i\in[m]} \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t - \frac{1-\gamma}{\gamma} \frac{1}{m} \sum_{i\in[m]} (\mathbf{u}^t - \mathbf{u}^{t-1} + \lambda \hat{\mathbf{g}}_i^{t-1}) \\
&= -\lambda \frac{1}{m} \sum_{i\in[m]} \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t - \frac{1-\gamma}{\gamma} \frac{1}{m} \sum_{i\in[m]} (\mathbf{x}_{i,K}^{t-1} - \mathbf{x}_{i,K}^{t-2} + \lambda \hat{\mathbf{g}}_i^{t-1}) \\
&= -\lambda \frac{1}{m} \sum_{i\in[m]} \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t - \frac{1-\gamma}{\gamma} \frac{1}{m} \sum_{i\in[m]} (\mathbf{x}_{i,K}^{t-1} - \mathbf{x}_{i,0}^{t-1} + \lambda \hat{\mathbf{g}}_i^{t-1} - \lambda \hat{\mathbf{g}}_i^{t-2}) \\
&= -\lambda \frac{1}{m} \sum_{i\in[m]} \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma} \tilde{\mathbf{g}}_{i,k}^t.
\end{aligned}$$

C.4 Proof of Theorem 4.5

Firstly we state some important lemmas applied in the proof.

Lemma C.4 *(Bounded global update) The global update $\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^t$ holds the upper bound of:*

$$\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 \leq \frac{1}{\gamma}\left(\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t}\|^2\right) + \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2.$$

Proof 4 *According to the lemma C.2,we have:*

$$\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^t = (1-\gamma)\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1} + \gamma\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t.$$

Take the L2-norm and we have:

$$\begin{aligned}\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^t\|^2 &= \|(1-\gamma)\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1} + \gamma\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2 \\ &\leq (1-\gamma)\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 + \gamma\|\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2.\end{aligned}$$

Thus we have the following recursion,

$$\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 \leq \frac{1}{\gamma}\left(\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t}\|^2\right) + \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2.$$

Lemma C.5 *(Bounded local update) The local update $\hat{\mathbf{g}}_i^t$ holds the upper bound of:*

$$\begin{aligned}\frac{1}{m}\sum_{i\in[m]}\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2 \leq \frac{P}{\gamma}\frac{1}{m}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\hat{\mathbf{g}}_i^{t}\|^2\right) + \frac{24PL^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t - \mathbf{x}^t\|^2 \\ + 12P\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + P(12\sigma_g^2 + \sigma_l^2),\end{aligned}$$

where $\frac{1}{P} = 1 - \frac{24\lambda^2L^2(1-2\gamma)^2}{\gamma^2}$.

Proof 5 *According to the lemmaC.2, we have:*

$$\hat{\mathbf{g}}_i^t = (1-\gamma)\hat{\mathbf{g}}_i^{t-1} + \gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t.$$

Take the L2-norm and we have:

$$\begin{aligned}\|\hat{\mathbf{g}}_i^t\|^2 &= \|(1-\gamma)\hat{\mathbf{g}}_i^{t-1} + \gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2 \\ &\overset{(a)}{\leq} (1-\gamma)\|\hat{\mathbf{g}}_i^{t-1}\|^2 + \gamma\|\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2 \\ &\overset{(b)}{\leq} (1-\gamma)\|\hat{\mathbf{g}}_i^{t-1}\|^2 + \gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\|\tilde{\mathbf{g}}_{i,k}^t\|^2 \\ &= (1-\gamma)\|\hat{\mathbf{g}}_i^{t-1}\|^2 + \sum_{k=0}^{K-1}\gamma_k\|\tilde{\mathbf{g}}_{i,k}^t\|^2.\end{aligned}$$

(a) and (b) apply the Jensen inequality.
Thus we have the following recursion:

$$\frac{1}{m}\sum_{i\in[m]}\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2 \leq \frac{1}{\gamma}\frac{1}{m}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\hat{\mathbf{g}}_i^{t}\|^2\right) + \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\tilde{\mathbf{g}}_{i,k}^t\|^2.$$

Here we provide a loose upper bound as a constant for the quasi-stochastic gradient:

$$\begin{aligned}
&\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\tilde{\mathbf{g}}_{i,k}^t\|^2 \\
&= \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|(1-\alpha)\mathbf{g}_{i,k,1}^t + \alpha\mathbf{g}_{i,k,2}^t\|^2 \\
&= \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{g}_{i,k,1}^t + \alpha(\mathbf{g}_{i,k,2}^t - \mathbf{g}_{i,k,1}^t)\|^2 \\
&\leq \frac{2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\left(\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)\|^2 + \alpha^2\mathbb{E}_t\|\nabla F_i(\check{\mathbf{x}}_{i,k}^t) - \nabla F_i(\mathbf{x}_{i,k}^t)\|^2\right) + \sigma_l^2 \\
&\leq \frac{2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\left(\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)\|^2 + \alpha^2L^2\rho^2\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)\|^2\right) + \sigma_l^2 \\
&\leq \frac{4}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t) - \nabla F_i(\mathbf{z}^t) + \nabla F_i(\mathbf{z}^t) - \nabla F(\mathbf{z}^t) + \nabla F(\mathbf{z}^t)\|^2 + \sigma_l^2 \\
&\leq \frac{12L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t - \mathbf{z}^t\|^2 + 12\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + (12\sigma_g^2 + \sigma_l^2) \\
&\leq \frac{12L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t - \mathbf{x}^t + \mathbf{x}^t - \mathbf{u}^t + \mathbf{u}^t - \mathbf{z}^t\|^2 \\
&\quad + 12\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + (12\sigma_g^2 + \sigma_l^2) \\
&\leq \frac{24L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t - \mathbf{x}^t\|^2 + 24L^2\|\mathbf{x}^t - \mathbf{u}^t + \mathbf{u}^t - \mathbf{z}^t\|^2 + (12\sigma_g^2 + \sigma_l^2) \\
&\quad + 12\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 \\
&\leq \frac{24L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t - \mathbf{x}^t\|^2 + \frac{24L^2\lambda^2(1-2\gamma)^2}{\gamma^2}\frac{1}{m}\sum_i\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2 \\
&\quad + 12\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + (12\sigma_g^2 + \sigma_l^2).
\end{aligned}$$

We applies the Jensen inequality, the basic inequality $\|\sum_{i=1}^n \mathbf{a}_i\|^2 \leq n\sum_{i=1}^n\|\mathbf{a}_i\|^2$, and the upper bound of $\rho \leq \frac{1}{\alpha L}$. Combining the above inequalities, let $\frac{1}{P} = 1 - \frac{24L^2\lambda^2(1-2\gamma^2)}{\gamma^2}$ is the constant, we have:

$$\begin{aligned}
\frac{1}{m}\sum_{i\in[m]}\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2 \leq \frac{P}{\gamma}\frac{1}{m}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\hat{\mathbf{g}}_i^{t}\|^2\right) + \frac{24PL^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t - \mathbf{x}^t\|^2 \\
+ 12P\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + P(12\sigma_g^2 + \sigma_l^2).
\end{aligned}$$

C.4.1 L-SMOOTHNESS OF THE FUNCTION F

For the general non-convex case, according to the Assumptions and the smoothness of F, we take the conditional expectation at round $t+1$ and expand the $F(\mathbf{z}^{t+1})$ as:

$$\begin{aligned}
\mathbb{E}_t[F(\mathbf{z}^{t+1})] &\leq F(\mathbf{z}^t) + \mathbb{E}_t\langle \nabla F(\mathbf{z}^t), \mathbf{z}^{t+1} - \mathbf{z}^t\rangle + \frac{L}{2}\mathbb{E}_t\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2 \\
&= F(\mathbf{z}^t) + \langle \nabla F(\mathbf{z}^t), \mathbb{E}_t[\mathbf{z}^{t+1}] - \mathbf{z}^t\rangle + \frac{L}{2}\mathbb{E}_t\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2 \\
&= F(\mathbf{z}^t) + \mathbb{E}_t\langle \nabla F(\mathbf{z}^t), -\lambda \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\rangle + \frac{L}{2}\mathbb{E}_t\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2 \\
&= F(\mathbf{z}^t) - \lambda\mathbb{E}_t\langle \nabla F(\mathbf{z}^t), \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t - \nabla F(\mathbf{z}^t) + \nabla F(\mathbf{z}^t)\rangle \\
&\quad + \frac{L}{2}\mathbb{E}_t\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2 \\
&= F(\mathbf{z}^t) - \lambda\|\nabla F(\mathbf{z}^t)\|^2 \underbrace{-\lambda\mathbb{E}_t\langle \nabla F(\mathbf{z}^t), \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t - \nabla F(\mathbf{z}^t)\rangle}_{\mathbf{R1}} \\
&\quad + \frac{L}{2}\underbrace{\mathbb{E}_t\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2}_{\mathbf{R2}}.
\end{aligned}$$

C.4.2 BOUNDED R1

Note that $\mathbf{R1}$ can be bounded as:

$$\begin{aligned}
\mathbf{R1} &= -\lambda\mathbb{E}_t\langle \nabla F(\mathbf{z}^t), \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t - \nabla F(\mathbf{z}^t)\rangle \\
&\overset{(a)}{=} -\lambda\mathbb{E}_t\langle \nabla F(\mathbf{z}^t), \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t - \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\nabla F_i(\mathbf{z}^t)\rangle \\
&\overset{(b)}{=} \frac{\lambda}{2}\|\nabla F(\mathbf{z}^t)\|^2 + \frac{\lambda}{2}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\big(\mathbb{E}\tilde{\mathbf{g}}_{i,k}^t - \nabla F_i(\mathbf{z}^t)\big)\|^2 - \frac{\lambda}{2m^2}\mathbb{E}_t\|\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}\tilde{\mathbf{g}}_{i,k}^t\|^2 \\
&\overset{(c)}{\leq} \frac{\lambda}{2}\|\nabla F(\mathbf{z}^t)\|^2 + \underbrace{\frac{\lambda}{2}\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbb{E}\tilde{\mathbf{g}}_{i,k}^t - \nabla F_i(\mathbf{z}^t)\|^2}_{\mathbf{R1.a}} - \frac{\lambda}{2m^2}\mathbb{E}_t\|\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}\tilde{\mathbf{g}}_{i,k}^t\|^2.
\end{aligned}$$

(a) applies the fact that $\frac{1}{m}\sum_{i\in[m]}\nabla F_i(\mathbf{z}^t) = \nabla F(\mathbf{z}^t)$. (b) applies $-\langle \mathbf{x}, \mathbf{y}\rangle = \frac{1}{2}\big(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}+\mathbf{y}\|^2\big)$. (c) applies the Jensen's inequality and the fact that $\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma} = 1$.
According to the update rule we have:

$$\begin{aligned}
\mathbb{E}\tilde{\mathbf{g}}_{i,k}^t &= (1-\alpha)\mathbb{E}\left[\mathbf{g}_{i,k,1}^t\right] + \alpha\mathbb{E}\left[\mathbf{g}_{i,k,2}^t\right] = (1-\alpha)\mathbb{E}\left[\nabla F_i(\mathbf{x}_{i,k}^t; \varepsilon_{i,k}^t)\right] + \alpha\mathbb{E}\left[\nabla F_i(\breve{\mathbf{x}}_{i,k}^t; \varepsilon_{i,k}^t)\right] \\
&= (1-\alpha)\nabla F_i(\mathbf{x}_{i,k}^t) + \alpha\nabla F_i(\breve{\mathbf{x}}_{i,k}^t) = (1-\alpha)\nabla F_i(\mathbf{x}_{i,k}^t) + \alpha\nabla F_i(\mathbf{x}_{i,k}^t + \rho\mathbf{g}_{i,k,1}^t).
\end{aligned}$$

Let $\rho \leq \frac{1}{\sqrt{3}\alpha L}$, thus we could bound the term $\mathbf{R1.a}$ as follows:

$$\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbb{E}\tilde{\mathbf{g}}_{i,k}^t - \nabla F_i(\mathbf{z}^t)\|^2$$

$$
\begin{aligned}
&= \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|(1-\alpha)\nabla F_i(\mathbf{x}_{i,k}^t)+\alpha\nabla F_i(\mathbf{x}_{i,k}^t+\rho\mathbf{g}_{i,k,1}^t)-\nabla F_i(\mathbf{z}^t)\|^2\\
&= \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)-\nabla F_i(\mathbf{z}^t)+\alpha\Big(\nabla F_i(\mathbf{x}_{i,k}^t+\rho\mathbf{g}_{i,k,1}^t)-\nabla F_i(\mathbf{x}_{i,k}^t)\Big)\|^2\\
&\le \frac{2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)-\nabla F_i(\mathbf{z}^t)\|^2+\frac{2\alpha^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\check{\mathbf{x}}_{i,k}^t)-\nabla F_i(\mathbf{x}_{i,k}^t)\|^2\\
&\le \frac{2L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{z}^t\|^2+\frac{2\alpha^2L^2\rho^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{g}_{i,k,1}^t\|^2\\
&= \frac{2L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t+\mathbf{x}^t-\mathbf{u}^t+\mathbf{u}^t-\mathbf{z}^t\|^2+\frac{2\alpha^2L^2\rho^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{g}_{i,k,1}^t\|^2\\
&\le \frac{4L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2+\frac{4L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|(\mathbf{x}^t-\mathbf{u}^t)+(\mathbf{u}^t-\mathbf{z}^t)\|^2\\
&\quad+\frac{2\alpha^2L^2\rho^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{g}_{i,k,1}^t-\nabla F_i(\mathbf{x}_{i,k}^t)\|^2+\frac{2\alpha^2L^2\rho^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)\|^2\\
&\le \frac{4L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2+\frac{2\alpha^2L^2\rho^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)\|^2+2\alpha^2L^2\rho^2\sigma_l^2\\
&\quad+4L^2\mathbb{E}_t\|(\mathbf{x}^t-\mathbf{u}^t)+(\mathbf{u}^t-\mathbf{z}^t)\|^2\\
&= \frac{4L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2+\frac{2\alpha^2L^2\rho^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)\|^2+2\alpha^2L^2\rho^2\sigma_l^2\\
&\quad+4L^2\mathbb{E}_t\|-\frac{1}{m}\sum_{i\in[m]}\lambda\hat{\mathbf{g}}_i^{t-1}+\frac{\gamma-1}{\gamma}(\mathbf{u}^t-\mathbf{u}^{t-1})\|^2\\
&= \frac{4L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2+\frac{2\alpha^2L^2\rho^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)\|^2+2\alpha^2L^2\rho^2\sigma_l^2\\
&\quad+4L^2\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\Big((\mathbf{u}^t-\mathbf{u}^{t-1}+\lambda\hat{\mathbf{g}}_i^{t-1})-\frac{1}{\gamma}(\mathbf{u}^t-\mathbf{u}^{t-1}+\lambda\hat{\mathbf{g}}_i^{t-1})+(\frac{1-2\gamma}{\gamma})\lambda\hat{\mathbf{g}}_i^{t-1}\Big)\|^2\\
&= \frac{4L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2+\frac{2\alpha^2L^2\rho^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)\|^2+2\alpha^2L^2\rho^2\sigma_l^2\\
&\quad+\frac{4\lambda^2L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2\\
&= \frac{4L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2+\frac{4\lambda^2L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2+2\alpha^2L^2\rho^2\sigma_l^2\\
&\quad+\frac{2\alpha^2L^2\rho^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k}^t)-\nabla F_i(\mathbf{z}^t)+\nabla F_i(\mathbf{z}^t)-\nabla F(\mathbf{z}^t)+\nabla F(\mathbf{z}^t)\|^2\\
&\overset{(a)}{\le} \frac{4L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2+\frac{4\lambda^2L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2+2\alpha^2L^2\rho^2\sigma_l^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{2L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{z}^t\|^2+6\alpha^2L^2\rho^2\sigma_g^2+6\alpha^2L^2\rho^2\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2\\
&\le\frac{8L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2+\frac{8\lambda^2L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2+2\alpha^2L^2\rho^2\sigma_l^2\\
&\quad+6\alpha^2L^2\rho^2\sigma_g^2+6\alpha^2L^2\rho^2\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2.\\
&\overset{(b)}{\le}\frac{8L^2}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2+\frac{8\lambda^2L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2-\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t}\|^2\right)\\
&\quad+\frac{8\lambda^2L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2+2\alpha^2L^2\rho^2\sigma_l^2+6\alpha^2L^2\rho^2\sigma_g^2+6\alpha^2L^2\rho^2\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2.
\end{aligned}
$$

(a) applies the bound of ρ as $\rho\le\frac{1}{\sqrt{3}\alpha L}$. (b) applies the lemma C.4. These others use the fact $\mathbb{E}[x-\mathbb{E}[x]]^2=\mathbb{E}[x^2]-[\mathbb{E}[x]]^2$ and $\|\mathbf{x}+\mathbf{y}\|^2\le(1+a)\|\mathbf{x}\|^2+(1+\frac{1}{a})\|\mathbf{y}\|^2$.

We denote $\mathbf{c}^t=\frac{1}{m}\sum_{i\in m}\sum_{k=0}^{K-1}(\gamma_k/\gamma)\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2$ term as the local offset after k iterations updates, we firstly consider the $\mathbf{c}_k^t=\frac{1}{m}\sum_{i\in m}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2$ and it can be bounded as:

$$
\begin{aligned}
\mathbf{c}_k^t&=\frac{1}{m}\sum_{i\in[m]}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}^t\|^2=\frac{1}{m}\sum_{i\in[m]}\mathbb{E}_t\|\mathbf{x}_{i,k}^t-\mathbf{x}_{i,k-1}^t+\mathbf{x}_{i,k-1}^t-\mathbf{x}_{i,0}^t\|^2\\
&=\frac{1}{m}\sum_{i\in[m]}\mathbb{E}_t\|-\eta_l(\tilde{\mathbf{g}}_{i,k-1}^t-\hat{\mathbf{g}}_i^{t-1})+(1-\frac{\eta_l}{\lambda})(\mathbf{x}_{i,k-1}^t-\mathbf{x}_{i,0}^t)\|^2\\
&\le(1+a)(1-\frac{\eta_l}{\lambda})^2\frac{1}{m}\sum_{i\in[m]}\mathbb{E}_t\|\mathbf{x}_{i,k-1}^t-\mathbf{x}_{i,0}^t\|^2+(1+\frac{1}{a})\frac{\eta_l^2}{m}\sum_{i\in[m]}\mathbb{E}_t\|\tilde{\mathbf{g}}_{i,k-1}^t-\hat{\mathbf{g}}_i^{t-1}\|^2\\
&=(1+a)(1-\frac{\eta_l}{\lambda})^2\mathbf{c}_{k-1}^t+(1+\frac{1}{a})\frac{\eta_l^2}{m}\sum_{i\in[m]}\mathbb{E}_t\|(1-\alpha)\mathbf{g}_{i,k-1,1}^t+\alpha\mathbf{g}_{i,k-1,2}^t-\hat{\mathbf{g}}_i^{t-1}\|^2\\
&=(1+\frac{1}{a})\frac{\eta_l^2}{m}\sum_{i\in[m]}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k-1}^t)-\hat{\mathbf{g}}_i^{t-1}+\alpha(\nabla F_i(\check{\mathbf{x}}_{i,k-1}^t)-\nabla F_i(\mathbf{x}_{i,k-1}^t))\|^2\\
&\quad+(1+\frac{1}{a})\eta_l^2\sigma_l^2+(1+a)(1-\frac{\eta_l}{\lambda})^2\mathbf{c}_{k-1}^t\\
&\le(1+\frac{1}{a})\frac{3\eta_l^2}{m}\sum_{i\in[m]}\left(\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k-1}^t)\|^2+\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2+\alpha^2L^2\rho^2\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k-1}^t)\|^2\right)\\
&\quad+(1+\frac{1}{a})\eta_l^2\sigma_l^2+(1+a)(1-\frac{\eta_l}{\lambda})^2\mathbf{c}_{k-1}^t\\
&\le(1+\frac{1}{a})\frac{4\eta_l^2}{m}\sum_{i\in[m]}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k-1}^t)\|^2+(1+\frac{1}{a})\frac{3\eta_l^2}{m}\sum_{i\in[m]}\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2+(1+\frac{1}{a})\eta_l^2\sigma_l^2\\
&\quad+(1+a)(1-\frac{\eta_l}{\lambda})^2\mathbf{c}_{k-1}^t\\
&\le(1+\frac{1}{a})\frac{4\eta_l^2}{m}\sum_{i\in[m]}\mathbb{E}_t\|\nabla F_i(\mathbf{x}_{i,k-1}^t)-\nabla F_i(\mathbf{x}^t)+\nabla F_i(\mathbf{x}^t)-\nabla F_i(\mathbf{z}^t)+\nabla F_i(\mathbf{z}^t)-\nabla F(\mathbf{z}^t)\\
&\quad+\nabla F(\mathbf{z}^t)\|^2+(1+\frac{1}{a})\frac{3\eta_l^2}{m}\sum_{i\in[m]}\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2+(1+\frac{1}{a})\eta_l^2\sigma_l^2+(1+a)(1-\frac{\eta_l}{\lambda})^2\mathbf{c}_{k-1}^t\\
&\le(1+\frac{1}{a})\frac{16\eta_l^2L^2}{m}\sum_{i\in[m]}\mathbb{E}_t\|\mathbf{x}_{i,k-1}^t-\mathbf{x}^t\|^2+(1+\frac{1}{a})16\eta_l^2L^2\|\mathbf{x}^t-\mathbf{z}^t\|^2+(1+\frac{1}{a})\eta_l^2(16\sigma_g^2+\sigma_l^2)
\end{aligned}
$$

$$
\begin{aligned}
&+(1+\frac{1}{a})16\eta_l^2\|\nabla F(\mathbf{z}^t)\|^2+(1+\frac{1}{a})\frac{3\eta_l^2}{m}\sum_{i\in[m]}\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2+(1+a)(1-\frac{\eta_l}{\lambda})^2\mathbf{c}_{k-1}^t\\
\leq&\left[(1+a)(1-\frac{\eta_l}{\lambda})^2+(1+\frac{1}{a})16\eta_l^2L^2\right]\mathbf{c}_{k-1}^t+(1+\frac{1}{a})\eta_l^2(16\sigma_g^2+\sigma_l^2)\\
&+(1+\frac{1}{a})16\eta_l^2\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2+(1+\frac{1}{a})\eta_l^2\left[3+\frac{16\lambda^2L^2(1-2\gamma)^2}{\gamma^2}\right]\frac{1}{m}\sum_{i\in[m]}\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2\\
=&\left[(1+a)(1-\frac{\eta_l}{\lambda})^2+(1+\frac{1}{a})16\eta_l^2L^2\right]\mathbf{c}_{k-1}^t+(1+\frac{1}{a})\eta_l^2(16\sigma_g^2+\sigma_l^2)\\
&+(1+\frac{1}{a})\eta_l^2L^2\left(88P-16\right)\mathbf{c}^t+(1+\frac{1}{a})\frac{2\eta_l^2(P-1)}{3}(12\sigma_g^2+\sigma_l^2)\\
&+(1+\frac{1}{a})16\eta_l^2\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2+(1+\frac{1}{a})\eta_l^2(44P-8)\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2\\
&+(1+\frac{1}{a})\frac{2\eta_l^2(P-1)}{3\gamma}\frac{1}{m}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2-\mathbb{E}_t\|\hat{\mathbf{g}}_i^t\|^2\right)
\end{aligned}
$$

When P satisfies the condition of $P\leq 2$, which means $\frac{1}{P}=1-\frac{24\lambda^2L^2(1-2\gamma)^2}{\gamma^2}\geq\frac{1}{2}$, then we have the constant of $\frac{2(P-1)}{3}\leq\frac{2}{3}<1$, let the last $12\sigma_g^2$ enlarged to $16\sigma_g^2$ for convenience, we have:

$$
\begin{aligned}
\mathbf{c}_k^t\leq&\left[(1+a)(1-\frac{\eta_l}{\lambda})^2+(1+\frac{1}{a})16\eta_l^2L^2\right]\mathbf{c}_{k-1}^t+2(1+\frac{1}{a})\eta_l^2(16\sigma_g^2+\sigma_l^2)+160(1+\frac{1}{a})\eta_l^2L^2\mathbf{c}^t\\
&96(1+\frac{1}{a})\eta_l^2\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2+2(1+\frac{1}{a})\frac{\eta_l^2}{\gamma}\frac{1}{m}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2-\mathbb{E}_t\|\hat{\mathbf{g}}_i^t\|^2\right).
\end{aligned}
$$

Here we get the recursion formula between the $\mathbf{c}_k^t$ and $\mathbf{c}_{k-1}^t$. Actually we need to upper bound the $\mathbf{c}^t=\sum_{k=0}^{K-1}(\gamma_k/\gamma)\mathbf{c}_k^t$, thus let the weight satisfies that:

$$
(1+a)(1-\frac{\eta_l}{\lambda})^2+(1+\frac{1}{a})16\eta_l^2L^2\leq\frac{\gamma_{K-2}}{\gamma_{K-1}}=\frac{\gamma_{K-3}}{\gamma_{K-2}}=\cdots=\frac{\gamma_1}{\gamma_0}=1-\frac{\eta_l}{\lambda},
$$

let $\eta_l\leq\lambda$ and thus we have:

$$
\begin{aligned}
\mathbf{c}^t=&\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbf{c}_k^t\\
&\leq 2(1+\frac{1}{a})\frac{\eta_l^2}{\gamma}\sum_{k'=0}^{K-1}\left(\sum_{k=0}^{k'-1}\gamma_k\right)\Big(16\sigma_g^2+\sigma_l^2+48\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2+80L^2\mathbf{c}^t\\
&\quad+\frac{1}{m\gamma}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2-\mathbb{E}_t\|\hat{\mathbf{g}}_i^t\|^2\right)\Big)\\
&\overset{(a)}{\leq} 2(1+\frac{1}{a})\eta_l^2\sum_{k'=0}^{K-1}\left(\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\right)\Big(16\sigma_g^2+\sigma_l^2+48\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2+80L^2\mathbf{c}^t\\
&\quad+\frac{1}{m\gamma}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2-\mathbb{E}_t\|\hat{\mathbf{g}}_i^t\|^2\right)\Big)\\
&=2(1+\frac{1}{a})\eta_l^2K\left(16\sigma_g^2+\sigma_l^2+48\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2+\frac{1}{m\gamma}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2-\mathbb{E}_t\|\hat{\mathbf{g}}_i^t\|^2\right)\right)
\end{aligned}
$$

$$+ 160(1+\frac{1}{a})\eta_l^2 L^2 K \mathbf{c}^t.$$

(a) enlarge the sum from $k^{'}$ to $K-1$ where $k^{'} \leq K-1$.
Let η_l satisfies the upper bound of $\eta_l \leq \frac{1}{\sqrt{320(1+1/a)KL}}$ for convenience, we can bound the $\mathbf{c}^t$ as:

$$\mathbf{c}^t = 4(1+\frac{1}{a})\eta_l^2 K \left(16\sigma_g^2 + \sigma_l^2 + 48\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + \frac{1}{m\gamma}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\hat{\mathbf{g}}_i^t\|^2\right)\right).$$

Let the a satisfies $a=1$ for convenience, we summarize the extra terms above and bound the term $\mathbf{R1}.\mathbf{a}$ as:

$$\begin{aligned}
\mathbf{R1}.\mathbf{a} &= \frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\mathbb{E}[\tilde{\mathbf{g}}_{i,k}^t] - \nabla F_i(\mathbf{z}^t)\|^2 \\
&\leq 8L^2\mathbf{c}^t + \frac{8\lambda^2L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^t\|^2\right) + 2\alpha^2L^2\rho^2\sigma_l^2 \\
&\quad + \frac{8\lambda^2L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2 + 6\alpha^2L^2\rho^2\sigma_g^2 + 6\alpha^2L^2\rho^2\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 \\
&\leq \frac{8\lambda^2L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^t\|^2\right) + 2\alpha^2L^2\rho^2\sigma_l^2 + 6\alpha^2L^2\rho^2\sigma_g^2 \\
&\quad + \frac{8\lambda^2L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2 + \frac{64\eta_l^2L^2K}{m\gamma}\sum_{i\in[m]}\left(\mathbb{E}_t\|\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\hat{\mathbf{g}}_i^t\|^2\right) \\
&\quad + 3072\eta_l^2L^2K\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + 6\alpha^2L^2\rho^2\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + 64\eta_l^2L^2K(16\sigma_g^2+\sigma_l^2).
\end{aligned}$$

thus we can bound the $\mathbf{R1}$ as follow:

$$\begin{aligned}
\mathbf{R1} &\leq \frac{\lambda}{2}\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + \frac{\lambda}{2}\mathbf{R1}.\mathbf{a} - \frac{\lambda}{2m^2}\mathbb{E}_t\|\sum_{i\in[m]}\sum_{k}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}[\tilde{\mathbf{g}}_{i,k}^t]\|^2 \\
&\leq \left(\frac{\lambda}{2} + 3\lambda\alpha^2L^2\rho^2 + 1536\lambda\eta_l^2L^2K\right)\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 + \frac{32\lambda\eta_lL^2K}{\gamma m}\sum_{i\in[m]}\left(\mathbb{E}\|\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}\|\hat{\mathbf{g}}_i^t\|^2\right) \\
&\quad + \frac{4\lambda^3L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^t\|^2\right) + \lambda\alpha^2L^2\rho^2(3\sigma_g^2+\sigma_l^2) \\
&\quad + \frac{4\lambda^3L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2 + 32\lambda\eta_l^2L^2K(16\sigma_g^2+\sigma_l^2).
\end{aligned}$$

We notice that $\mathbf{R1}$ contains the same term with a negative weight, thus we can set another constrains for λ to eliminate this term. We will prove it in the next part.

C.4.3 Bounded Global Gradient

As we have bounded the term $\mathbf{R1}$ and $\mathbf{R2}$, according to the smoothness inequality, we combine the inequalities above and get the inequality:

$$\mathbb{E}_t[F(\mathbf{z}^{t+1})] \leq F(\mathbf{z}^t) - \lambda\|\nabla F(\mathbf{z}^t)\|^2 + \mathbf{R1} + \frac{L}{2}\mathbf{R2}$$

$$
\begin{aligned}
&= F(\mathbf{z}^t) - \left(\frac{\lambda}{2} - 3\lambda\alpha^2L^2\rho^2 - 1536\lambda\eta_l^2L^2K\right)\|\nabla F(\mathbf{z}^t)\|^2 + \lambda\alpha^2L^2\rho^2(3\sigma_g^2+\sigma_l^2) \\
&\quad + \Big(\frac{4\lambda^3L^2(1-2\gamma)^2}{\gamma^2} + \frac{\lambda^2L}{2m^2} - \frac{\lambda}{2m^2}\Big)\mathbb{E}_t\|\sum_{i\in[m]}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{\mathbf{g}}_{i,k}^t\|^2 \\
&\quad + \frac{32\lambda\eta_lL^2K}{\gamma m}\sum_{i\in[m]}\big(\mathbb{E}\|\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}\|\hat{\mathbf{g}}_i^t\|^2\big) + 32\lambda\eta_l^2L^2K(16\sigma_g^2+\sigma_l^2) \\
&\quad + \frac{4\lambda^3L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^t\|^2\right).
\end{aligned}
$$

We follow as Yang et al. (2021) to set λ that it satisfies $\frac{4\lambda^3L^2(1-2\gamma)^2}{\gamma^2} + \frac{\lambda^2L}{2m^2} - \frac{\lambda}{2m^2} \leq 0$, which is easy to verified that λ has a upper bound for the quadratic inequality. Thus, the stochastic gradient term is diminished by this λ. We denote the constant $\lambda\kappa = \frac{\lambda}{2} - 3\lambda\alpha^2L^2\rho^2 - 1536\lambda\eta_l^2L^2K$ and take the full expectation on the bounded global gradient as:

$$
\begin{aligned}
\lambda\kappa\mathbb{E}\|\nabla F(\mathbf{z}^t)\|^2 \leq \big(\mathbb{E}F(\mathbf{z}^t) - \mathbb{E}F(\mathbf{z}^{t+1})\big) &+ \frac{32\lambda\eta_lL^2K}{\gamma m}\sum_{i\in[m]}\big(\mathbb{E}\|\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}\|\hat{\mathbf{g}}_i^t\|^2\big) \\
&+ \frac{4\lambda^3L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^{t-1}\|^2 - \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^t\|^2\right) \\
&+ 32\lambda\eta_l^2L^2K(16\sigma_g^2+\sigma_l^2) + \lambda\alpha^2L^2\rho^2(3\sigma_g^2+\sigma_l^2).
\end{aligned}
$$

Take the full expectation and telescope sum on the inequality above and applying the fact that $F^* \leq F(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$, we have:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}_t\|\nabla F(\mathbf{z}^t)\|^2 &\leq \frac{1}{\lambda\kappa T}\big(F(\mathbf{z}^1) - \mathbb{E}_t[F(\mathbf{z}^T)]\big) + \frac{32\eta_lL^2K}{\kappa\gamma mT}\sum_{i\in[m]}\big(\mathbb{E}\|\hat{\mathbf{g}}_i^0\|^2 - \mathbb{E}\|\hat{\mathbf{g}}_i^t\|^2\big) \\
&\quad + \frac{4\lambda^2L^2(1-2\gamma)^2}{\kappa\gamma^3T}\left(\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^0\|^2 - \mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^t\|^2\right) \\
&\quad + \frac{1}{\kappa}\big(32\lambda\eta_l^2L^2K(16\sigma_g^2+\sigma_l^2) + \lambda\alpha^2L^2\rho^2(3\sigma_g^2+\sigma_l^2)\big) \\
&\leq \frac{1}{\lambda\kappa T}\big(F(\mathbf{z}^0) - F^*\big) + \frac{32\eta_lL^2K}{\kappa\gamma mT}\sum_{i\in[m]}\mathbb{E}\|\hat{\mathbf{g}}_i^0\|^2 \\
&\quad + \frac{4\lambda^2L^2(1-2\gamma)^2}{\kappa\gamma^3T}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^0\|^2 \\
&\quad + \frac{1}{\kappa}\big(32\lambda\eta_l^2L^2K(16\sigma_g^2+\sigma_l^2) + \lambda\alpha^2L^2\rho^2(3\sigma_g^2+\sigma_l^2)\big)
\end{aligned}
$$

Here we summarize the conditions and some constrains in the above conclusion. Firstly we should note that $\gamma = 1 - (1 - \frac{\eta_l}{\lambda})^K < 1$ when $\eta_l \leq 2\lambda$. Thus we have $1/\gamma > 1$. When K satisfies that $K \geq \frac{\lambda}{\eta_l}$, $(1-\frac{\eta_l}{\lambda})^K \leq e^{-\frac{\eta_l}{\lambda}K} \leq e^{-1}$, and then $\gamma > 1 - e^{-1}$ and $1/\gamma < \frac{e}{e-1} < 2$. To let $\kappa = \frac{1}{2} - 3\alpha^2L^2\rho^2 - 1536\eta_l^2L^2K > 0$ hold, ρ and η_l satisfy that $\rho < \frac{1}{\sqrt{6}\alpha L}$ and $\eta_l < \frac{1}{32\sqrt{3KL}}$.

$$
\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla F(\mathbf{z}^t)\|^2 \leq \frac{2(F(\mathbf{z}^1) - F^*)}{\lambda\kappa T} + \frac{64\eta_lL^2K}{\kappa T}\frac{1}{m}\sum_{i\in[m]}\mathbb{E}\|\hat{\mathbf{g}}_i^0\|^2 + \frac{32\lambda^2L^2}{\kappa T}\mathbb{E}_t\|\frac{1}{m}\sum_{i\in[m]}\hat{\mathbf{g}}_i^0\|^2
$$

$$+\frac{1}{\kappa}\left(32\lambda\eta_l^2L^2K(16\sigma_g^2+\sigma_l^2)+\lambda\alpha^2L^2\rho^2(3\sigma_g^2+\sigma_l^2)\right).$$

REFERENCES

Alp Emre Durmus, Zhao Yue, Matas Ramon, Mattina Matthew, Whatmough Paul, and Saligrama Venkatesh. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

Yonghai Gong, Yichuan Li, and Nikolaos M Freris. Fedadmm: A robust federated deep learning framework with adaptivity to system heterogeneity. *arXiv preprint arXiv:2204.03529*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.

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.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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*, pp. 1273–1282. PMLR, 2017.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3:3, 2018.

Han Wang, Siddartha Marella, and James Anderson. Fedadmm: A federated primal-dual algorithm allowing partial participation. *arXiv preprint arXiv:2203.15104*, 2022.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. Fedcm: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.