Supplementary Material for
" Decentralized Directed Collaboration for Personalized Federated Learning "

In this part, we provide supplementary materials including more introduction to the related works, experimental details and results, and the proof of the main theorem.

- **Appendix** A: More details in the related works.
- **Appendix** B: More details in the client selection.
- **Appendix** C: More details in the experiments.
- **Appendix** D: Proof of the theoretical analysis.

A. More Details in the Related Works

Decentralized/Distributed Training. Decentralized/Distributed Training, which allows edge clients to communicate with each other in a peer-to-peer manner, is an encouraging field that shares several benefits: (1) guarantees collaborative learning through local computation and the exchange of model parameters; (2) is low for feeding the models of adjacent clients, generating a more intelligent private model; (3) avoids central failure in the collaborative system. Thus, Decentralized/Distributed Training has been applied in many fields[5]: (1) Healthcare [45], favoring the decentralization of clinical records and collaborative diagnosis; (2) Mobile Services [61], decreasing response times and increasing the bandwidth of constraints devices; (3) Vehicles [71], ensuring high mobility and local storage management.

Since the prototype of DFL (fully decentralized federated learning [28]) was proposed, it has been a promising approach to save communication costs as the compromise of CFL. By combining SGD and gossip, early work achieved decentralized training and convergence in [6]. D-PSGD [38] is the classic decentralized parallel SGD method. FastMix [68] investigates the advantage of increasing the frequency of local communications within a network topology, which establishes the optimal computational complexity and near-optimal communication complexity. DeEPCA [67] integrates FastMix into a decentralized PCA algorithm to accelerate the training process. DeLi-CoCo [18] performs multiple compression gossip steps in each iteration for fast convergence with arbitrary communication compression. Network-DANE [30] uses multiple gossip steps and generalizes DANE to decentralized scenarios. QG-DSGDm [40] modifies the momentum term of decentralized SGD (DSGD) to be adaptive to heterogeneous data, while SkewScout [21] replaces batch norm with layer norm. Meta-L2C [36] dynamically updates the mixing weights based on meta-learning and learns a sparse topology to reduce communication costs. The work in [73] provides the topology-aware generalization analysis for DSGD, they explore the impact of various communication topologies on the generalizability.

B. More details in the client selection

Push sum based directed distributed averaging. The initial Push sum algorithm [26] considers the averaged consensus $1/n\sum_{i=1}^{n} y_i^0$ of all clients. Let $y_i^0 \in \mathbb{R}^d$ be a vector at client i and typical gossip iterations forms $y_i^{t+1} = \sum_{j=1}^{n} p_{i,j}^t y_j^t$, where $P^t \in \mathbb{R}^{n\times n}$ is the mixing matrix. Inspired by the Markov chains [51], the mixing matrices P^t are designed to be column stochastic (each column must sum to 1). So the gossip iterations converge to a limit $y_i^\infty = \pi_i \sum_{j=1}^{n} y_j^0$, where π is the ergodic limit of the chain. When the matrices P^t are symmetric, it is straightforward to satisfy $\pi_i = 1/n$ by defining P^t doubly-stochastic (each row and each column must sum to 1). However, symmetric P^t are hard to meet due to the unstable communication in reality. The Push sum algorithm adds one additional scalar parameter w_i^t to achieve $\pi_i = 1/n$ under the column-stochastic and asymmetric mixing matrices P^t. The parameter is initialized to $w_i^0 = 1$ for all i and updated using the same linear iteration, $w_i^{t+1} = \sum_{j=1}^{n} p_{i,j}^t w_j^t$. It recovers the average of the initial vectors by computing the de-biased ratio y_i^∞ / w_i^∞, and the scalar parameters converge to $w_i^\infty = \pi_i \sum_{j=1}^{n} w_j^0$.

Directed random graph. We transfer the mixing matrices from column stochastic (all columns sum to 1) to row stochastic (all rows sum to 1), meaning that the clients can actively select the information they need rather than passively accept, which is more beneficial for directed collaboration in the DPFL problem. In the experiments, each client pulls the shared parameters from its in-neighbors $j \in \mathcal{N}_{i,t}^{in}$, and "pulls a message" from itself as well. Recall that each client i can choose its mixing weights (ith row of P^t) independently of the other clients. So in order to provide more flexible collaboration and closer ties for clients, we randomly choose the in-neighbors under the communication bandwidth limitation. We use uniform mixing weights for the pulled models here, meaning that clients assign uniform model weights to all neighbors. So assuming that

each client can pull models with up to n neighbors, each row P_i^t of P^t has exactly $n+1$ non-zero entries, both of which are equal to $1/(n+1)$. Thus, we get that

$$p_{i,j}^t = \begin{cases} 1/(n+1), & j \in \mathcal{N}_{i,t}^{in}; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Undirected random graph. For the undirected DPFL methods (i.e. DFedAvgM and Dis-PFL), we use a time-varying and undirected random graph to represent the inter-client connectivity. Clients randomly choose their in-neighbors to pull the shared models and push a message in return. We adopt these graphs to be consistent with the experimental setup used in [14, 52, 56]. So the mixing matrics in the undirected graph is a symmetric doubly-stochastic (each row and each column must sum to 1), which satisfies $p_{i,j}^t = p_{j,i}^t$ in Formula (6). Notably, the model communication bandwidth of in-neighbors in DPFL is strictly limited as the same as the busiest server in CPFL.

C. More details in the experiments

In this section, we provide more details of our experiments including datasets, baselines, and more extensive experimental results to compare the performance of the proposed DFedPGP against other baselines on the Tiny-ImageNet dataset. All our experiments are trained and tested on a single Nvidia RTX3090 GPU under the environment of Python 3.8.5, PyTorch 1.11.1, CUDA 11.6, and CUDNN 8.0.

C.1. More Details about Baselines

Local is the simplest method for personalized learning. It only trains the personalized model on the local data and does not communicate with other clients. For the fair competition, we train 5 epochs locally in each round.

FedAvg [41] is the most commonly discussed method in FL. It selects partial clients to perform local training on each dataset and then aggregates the trained models to update the global model. Actually, the local model in FedAvg is also the comparable personalized model for each client.

FedPer [1] proposes a model decoupling approach for PFL, with a consensus representation and many local classifiers, to combat the ill effects of statistical heterogeneity. We set the linear layer as the personalized layer and the rest model as the base layer. It follows FedAvg's training paradigm but only passes the base layer to the server and keeps the personalized layer locally.

FedRep [13] also proposes a personalized model decoupling framework like FedPer, but it fixes one part when updating the other. We follow the official implementation[2] to train the head for 10 epochs with the body fixed, and then train the body for 5 epochs with the head fixed.

FedBABU [46] is also a model decoupling method that achieves good personalization via fine-tuning from a good shared representation base layer. Different from FedPer and FedRep, FedBABU only updates the base layer with the personalized layer fixed and finally fine-tunes the whole model. Following the official implementation[3], it fine-tunes 5 times in our experiments.

Ditto [37] achieves personalization via a trade-off between the global model and local objectives. It totally trains two models on the local datasets, one for the global model (similarly aggregated as in FedAvg) with its local empirical risk, and one for the personal model (kept locally) with both empirical risk and the proximal term towards the global model. We set the regularization parameters λ as 0.75.

DFedAvgM [56] is the decentralized FedAvg with momentum, in which clients only connect with their neighbors by an undirected graph. For each client, it first initials the local model with the received models then updates it on the local datasets with a local stochastic gradient.

OSGP [2] is the directed version of DFedAvg, which allows clients to send the local models to their out-neighbors by a directed graph. It is regarded as a representative of a personalized baseline over directed communication.

Dis-PFL [14] employs personalized sparse masks to customize sparse local models in the PFL setting. Each client first initials the local model with the personalized sparse masks and updates it with empirical risk. Then filter out the parameter weights that have little influence on the gradient through cosine annealing pruning to obtain a new mask. Following the official implementation[4], the sparsity of the local model is set to 0.5 for all clients.

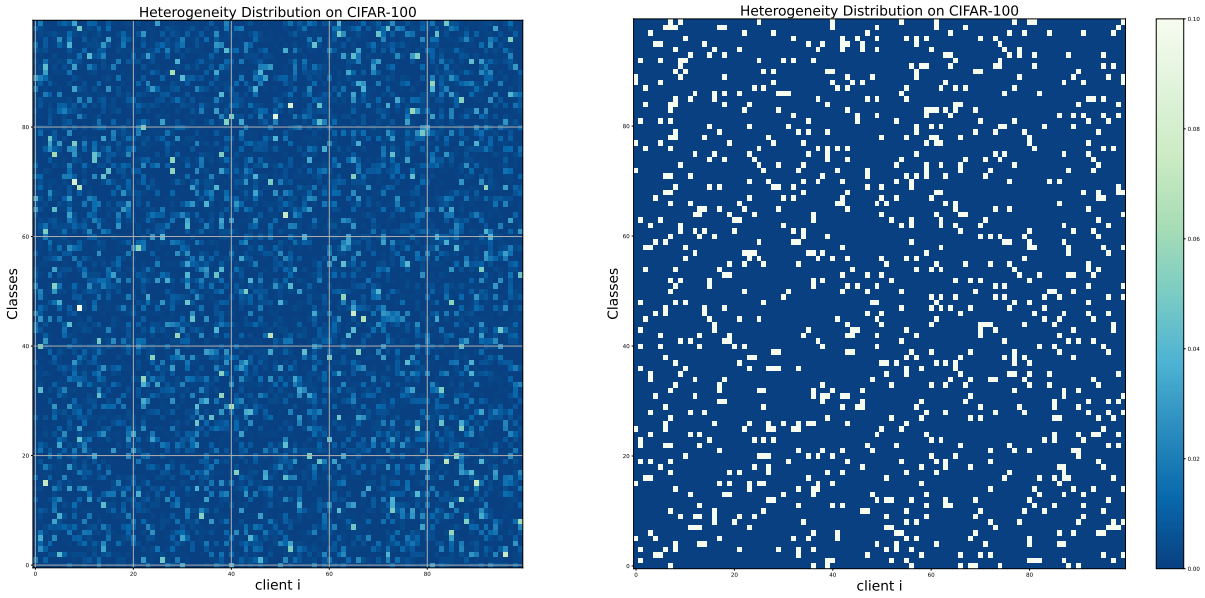
[2] https://github.com/lgcollins/FedRep
[3] https://github.com/jhoon-oh/FedBABU
[4] https://github.com/rong-dai/DisPFL

C.2. Datasets and Data Partition

CIFAR-10/100 and Tiny-ImageNet are three basic datasets in the computer version study. As shown in Table 5, they are all colorful images with different classes and different resolutions. We use two non-IID partition methods to split the training data in our implementation. One is based on Dirichlet distribution on the label ratios to ensure data heterogeneity among clients. The Dirichlet distribution defines the local dataset to obey a Dirichlet distribution (see in Figure 4a), where a smaller α means higher heterogeneity. Another assigns each client a limited number of categories, called Pathological distribution. Pathological distribution defines the local dataset to obey a uniform distribution of active categories c (see in Figure 4b), where fewer categories mean higher heterogeneity. The distribution of the test datasets is the same as in training datasets. We run 500 communication rounds for CIFAR-10, CIFAR-100, and 300 rounds for Tiny-ImageNet.

Table 5. The details on the CIFAR-10 and CIFAR-100 datasets.

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
Tiny-ImageNet	100,000	10,000	200	3×64×64



(a) Dirichlet $\alpha = 0.3$ on CIFAR-100.

(b) Pathological $c = 10$ on CIFAR-100.

Figure 4. Heat-map of the Dirichlet split and Pathological split.

C.3. More Experiments Results on Tiny ImageNet

Comparison with the baselines. In Table 6 and Figure 5, we compare DFedPGP with other baselines on the Tiny-ImageNet with different data distributions. The comparison shows that the proposed method has a competitive performance, especially under higher heterogeneity, e.g. Pathological-10. Specifically in the Pathological-10 setting, DFedPGP achieves 49.16%, at least 1.81% and 7.08% improvement from the CFL methods and DFL methods. However, in the Dirichlet-0.3 setting, almost all the partial model personalized methods (i.e. FedPer, FedRep, DFedPGP except FedBABU) face a severe performance degradation compared with the full model personalized methods (i.e. FedAvg, DFedAvgM, OSGP). This may account for the low classification ability in partial model personalized methods without aggregation with neighbors in the multiple-image classification tasks, especially in the long-tail data distribution scenario (i.e. Dirichelet-0.3). The original intention of our design is to build a great personalized model through partial model personalization training and directed collaboration with neighbors. So when the heterogeneity increases, our algorithms have a significant improvement.

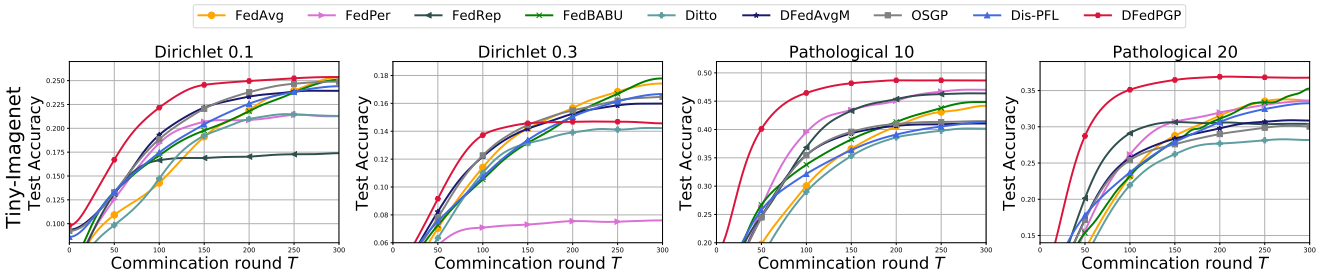


Figure 5. Test accuracy on Tiny-ImageNet with heterogenous data partitions.

Table 6. Test accuracy (%) on Tiny-ImageNet in both Dirichlet and Pathological distribution settings on Tiny-ImageNet.

Algorithm	Tiny-ImageNet			
	Dirichlet		Pathological	
	$\alpha = 0.1$	$\alpha = 0.3$	c = 10	c = 20
Local	$12.13_{\pm.13}$	$5.42_{\pm.21}$	$28.49_{\pm.16}$	$16.72_{\pm.34}$
FedAvg	$25.55_{\pm.02}$	$17.58_{\pm.25}$	$44.56_{\pm.39}$	$34.10_{\pm.59}$
FedPer	$21.64_{\pm.72}$	$7.71_{\pm.08}$	$47.35_{\pm.03}$	$33.68_{\pm.33}$
FedRep	$17.54_{\pm.79}$	$5.78_{\pm.05}$	$46.76_{\pm.73}$	$31.15_{\pm.54}$
FedBABU	$25.59_{\pm.08}$	$\mathbf{18.18}_{\pm.06}$	$46.53_{\pm.20}$	$37.01_{\pm.31}$
Ditto	$21.71_{\pm.66}$	$14.47_{\pm.14}$	$40.65_{\pm.15}$	$28.74_{\pm.38}$
DFedAvgM	$24.42_{\pm.74}$	$16.51_{\pm.68}$	$41.94_{\pm.37}$	$31.50_{\pm.46}$
OSGP	$25.29_{\pm.26}$	$17.07_{\pm.17}$	$42.08_{\pm.43}$	$30.58_{\pm.51}$
Dis-PFL	$24.71_{\pm.18}$	$16.94_{\pm.36}$	$41.93_{\pm.12}$	$33.57_{\pm.62}$
DFedPGP	$\mathbf{25.71}_{\pm.20}$	$14.94_{\pm.44}$	$\mathbf{49.16}_{\pm.19}$	$\mathbf{37.25}_{\pm.27}$

Table 7. The required communication rounds when achieving the target accuracy (%) on Tiny-ImageNet.

Algorithm	Tiny-ImageNet							
	Dirichlet-0.1		Dirichlet-0.3		Pathological-10		Pathological-20	
	acc@20	speedup	acc@14	speedup	acc@40	speedup	acc@30	speedup
FedAvg	160	1.11 ×	144	1.47 ×	192	1.36 ×	172	1.50 ×
FedPer	123	1.45 ×	-	-	103	2.53 ×	134	1.93 ×
FedRep	-	-	-	-	116	2.25 ×	117	2.21 ×
FedBABU	156	1.14 ×	174	1.22 ×	178	1.47 ×	181	1.43 ×
Ditto	178	1.00 ×	212	1.00 ×	261	1.00 ×	-	-
DFedAvgM	110	1.62 ×	141	1.50 ×	173	1.51 ×	210	1.23 ×
OSGP	115	1.55 ×	136	1.56 ×	160	1.63 ×	258	1.00 ×
Dis-PFL	143	1.24 ×	166	1.28 ×	227	1.15 ×	188	1.37 ×
DFedPGP	**74**	**2.41 ×**	**108**	**1.96 ×**	**54**	**4.83 ×**	**53**	**4.87 ×**

Convergence speed. We show the convergence speed of DFedPGP in Table 7 and Figure 5 by reporting the number of rounds required to achieve the target personalized accuracy (acc@) on Tiny-ImageNet. We set the algorithm that takes the most rounds to reach the target accuracy as "1.00×", and find that the proposed DFedPGP achieves the fastest convergence speed on average (3.51× on average) among the SOTA PFL algorithms. Direct communication guarantees flexible choice of neighbors and closer ties between clients, which speeds up personalized convergence and achieves higher personalized performance for each client. Also, the partial model personalization and alternate updating mode will both bring a comparable gain to the convergence speed from the difference between DFedPGP and OSGP. Thus, our methods can efficiently train the

personalized model under different data heterogeneity.

C.4. More Details about hyperparameters selection

Here we detail the hyperparameter selection in our experiments. We fix the total communication rounds T, mini-batch size and weight decay for all the benchmarks and our proposed DFedPGP. The other selections are stated as follows.

Table 8. General hyperparameters introductions.

Hyperparameter	CIFAR-10/100, Tiny-ImageNet	Best Selection
Communication Round	500	-
Batch Size	128	-
Weight Decay	5e-4	-
Momentum	0.9	-
Learning Rate Decay	0.9	-
Local Interval	[1, 3, 5, 8]	5
Local Learning Rate	[0.01, 0.1, 0.5, 1]	0.1

D. Proof of Theoretical Analysis

D.1. Preliminary Lemmas

Lemma 1 (Local update for personalized model v_i in DFedPGP, Lemma 23 [47])**.** *Consider F which is L-smoothness and fix $v^0 \in \mathbb{R}^d$. Define the sequence (v^k) of iterates produced by stochastic gradient descent with a fixed learning rate $\eta_v \leq 1/(2K_vL_v)$ starting from v^0, we have the bound*

$$\mathbb{E}\|v^{K_v-1} - v^0\|^2 \leq 16\eta_v^2 K_v^2 \mathbb{E}\|\nabla F(v^0)\|^2 + 8\eta_v^2 K_v^2 \sigma_v^2 \,.$$

Proof.

$$\begin{aligned}
\mathbb{E}\|v_i^{t,k+1} - v_i^{t,0}\|^2 &= \mathbb{E}\Big\|v_i^{t,k} - \eta_v \nabla_v F_i(z_i^t, v_i^{t,k}; \xi_i) - v_i^{t,0}\Big\|^2 \\
&\overset{a)}{\leq} \Big(1 + \frac{1}{K_v - 1}\Big)\mathbb{E}\|v_i^{t,k} - v_i^{t,0}\|^2 + K_v\eta_v^2 \mathbb{E}\Big\|\nabla_v F_i(z_i^t, v_i^{t,k}; \xi_i) - \nabla_v F_i(z_i^t, v_i^t) + \nabla_v F_i(z_i^t, v_i^t)\Big\| \\
&\leq \Big(1 + \frac{1}{K_v - 1}\Big)\mathbb{E}\Big\|v_i^{t,k} - v_i^{t,0}\Big\|^2 + K_v\eta_v^2 \Big(\sigma_v^2 + \mathbb{E}\Big\|\nabla_v F_i(z_i^t, v_i^t) - \nabla_v F_i(z_i^t, v_i^{t,0}) + \nabla_v F_i(z_i^t, v_i^{t,0})\Big\|^2\Big) \\
&\overset{b)}{\leq} \Big(1 + \frac{1}{K_v - 1}\Big)\mathbb{E}\|v_i^{t,k} - v_i^{t,0}\|^2 + K_v\eta_v^2\sigma_v^2 + 2K_v\eta_v^2 L_v^2 \|v_i^{t,k} - v_i^{t,0}\|^2 + 2K_v\eta_v^2 \big\|\nabla_v F_i(z_i^t, v_i^{t,0})\big\|^2 \\
&\leq \Big(1 + \frac{1}{K_v - 1} + 2K_v\eta_v^2 L_v^2\Big)\mathbb{E}\|v_i^{t,k} - v_i^{t,0}\|^2 + K_v\eta_v^2\sigma_v^2 + 2K_v\eta_v^2 \|\nabla_v F_i(z_i^t, v_i^{t,0})\|^2 \\
&\overset{c)}{\leq} \Big(1 + \frac{2}{K_v - 1}\Big)\mathbb{E}\|v_i^{t,k} - v_i^{t,0}\|^2 + K_v\eta_v^2\sigma_v^2 + 2K_v\eta_v^2 \|\nabla_v F_i(z_i^t, v_i^{t,0})\|^2.
\end{aligned} \tag{7}$$

where we used a) the inequality $2\alpha\beta \leq \alpha/K + K\beta$ for reals α, β, K; b) L-smoothness of F, and c) the condition on the learning rate $\eta_v \leq 1/(2K_vL_v)$. Let $A = K_v\eta_v^2\sigma_v^2 + 2K_v\eta_v^2\|\nabla_v F_i(z_i^t, v_i^{t,0})\|^2$. Unrolling the inequality and summing up the series gives for all $k \leq K_v - 1$:

$$\begin{aligned}
\mathbb{E}\|v_i^{t,k+1} - v_i^{t,0}\|^2 &\leq \Big(1 + \frac{2}{K_v - 1}\Big)\mathbb{E}\|v_i^{t,k} - v_i^{t,0}\|^2 + A \\
&\leq A\sum_{k=0}^{K_v-1}\Big(1 + \frac{2}{k-1}\Big)^k \leq \frac{A}{2}(K_v - 1)\sum_{k=0}^{K_v-1}\Big(1 + \frac{2}{K_v - 1}\Big)^k \\
&\leq \frac{A}{2}(K_v - 1)\Big(1 + \frac{2}{K_v - 1}\Big)^{K_v-1}.
\end{aligned} \tag{8}$$

Using the bound $(1+2/K_v-1)^{K_v-1} \leq e^2 < 8$ for $K_v > 1$, we have:

$$\mathbb{E}\|v_i^{K_v-1} - v_i^0\|^2 \leq 4A(K_v-1) \leq 16\eta_v^2 K_v^2 \mathbb{E}\left\|\nabla F(v^0)\right\|^2 + 8\eta_v^2 K_v^2 \sigma_v^2. \tag{9}$$

□

Lemma 2 (Local update for shared model u_i in DFedPGP)**.** *For all clients $i \in \{1,2,...,m\}$ and local iteration steps $k \in \{0,1,...,K_u-1\}$, assume that assumptions 2-4 hold and define $\nabla_u F_i(z_i^{t,k}, v_i^{t+1}; \xi_i) = \nabla_u F_i(u_i^{t,k}/\mu_i^t, v_i^{t+1}; \xi_i)$, we can get*

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|u_i^{t,k} - u_i^t\right\|^2 \leq 32K_u\eta_u^2\sigma_u^2 + 32K_u\eta_u^2\sigma_g^2 + \frac{32K_u\eta_u^2}{m}\sum_{i=1}^{m}\mathbb{E}\|\nabla f(z_i^t, V^{t+1})\|^2. \tag{10}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left\|u_i^{t,k+1} - u_i^t\right\|^2 &= \mathbb{E}\left\|u_i^{t,k} - \eta_u\nabla_u F_i(z_i^{t,k}, v_i^{t+1}; \xi_i) - u_i^t\right\|^2 \\
&\leq (1+\frac{1}{2K_u-1})\mathbb{E}\|u_i^{t,k} - u_i^t\|^2 + 2K_u\eta_u^2\mathbb{E}\left\|\nabla_u F_i(z_i^{t,k}, v_i^{t+1}; \xi_i)\right\|^2 \\
&\leq (1+\frac{1}{2K_u-1})\mathbb{E}\|u_i^{t,k} - u_i^t\|^2 + 2K_u\eta_u^2\mathbb{E}\Big\|\nabla_u F_i(z_i^{t,k}, v_i^{t+1}; \xi_i) - \nabla_u F_i(z_i^{t,k}, v_i^{t+1}) \\
&\quad + \nabla_u F_i(z_i^{t,k}, v_i^{t+1}) - \nabla_u F(z_i^{t,k}, V^{t+1}) + \nabla_u F(z_i^{t,k}, V^{t+1}) - \nabla_u F(z_i^t, V^{t+1}) + \nabla_u F(z_i^t, V^{t+1})\Big\|^2 \\
&\leq \Big(1+\frac{1}{2K_u-1}\Big)\mathbb{E}\|u_i^{t,k} - u_i^t\|^2 + 8K_u\eta_u^2\Big(\sigma_u^2 + \sigma_g^2 + L_u^2\mathbb{E}\|z_i^{t,k} - z_i^t\|^2 + \mathbb{E}\|\nabla_u F(z_i^t, V^{t+1})\|^2\Big).
\end{aligned} \tag{11}$$

where we use Assumption 3, 4 and L-smoothness in the last inequation.

In addition, according to line 11 of Algorithm 1, we can obtain $\mathbb{E}\|z_i^{t,k} - z_i^t\|^2 = \frac{1}{\|\mu_i^t\|^2}\mathbb{E}\|u_i^{t,k} - u_i^t\|^2$. According to Property 2.1 by [57], there exists $\delta > 0$ that satisfies $\|\mu_i^t\| > \delta$. Therefore, we can get $\mathbb{E}\|z_i^{t,k} - z_i^t\|^2 \leq \frac{1}{\delta^2}\mathbb{E}\|u_i^{t,k} - u_i^t\|^2$. Assume the learning rate $0 < \eta_u < \frac{\delta}{8L_uK_u}$, then we have

$$\begin{aligned}
\mathbb{E}\|u_i^{t,k+1} - u_i^t\|^2 &\leq \Big(1+\frac{1}{2K_u-1}+\frac{8K_uL_u^2\eta_u^2}{\delta^2}\Big)\mathbb{E}\|u_i^{t,k} - u_i^t\|^2 + 8K_u\eta_u^2\sigma_u^2 + 8K_u\eta_u^2\sigma_g^2 + 8K_u\eta_u^2\mathbb{E}\|\nabla f(z_i^t, V^{t+1})\|^2 \\
&\leq \Big(1+\frac{1}{K_u-1}\Big)\mathbb{E}\|u_i^{t,k} - u_i^t\|^2 + 8K_u\eta_u^2\sigma_u^2 + 8K_u\eta_u^2\sigma_g^2 + 8K_u\eta_u^2\mathbb{E}\|\nabla f(z_i^t, V^{t+1})\|^2 \\
&\leq \sum_{k=0}^{K_u-1}\Big(1+\frac{1}{K_u-1}\Big)^k\big(8K_u\eta_u^2\sigma_u^2 + 8K_u\eta_u^2\sigma_g^2 + 8K_u\eta_u^2\mathbb{E}\|\nabla f(z_i^t, V^{t+1})\|^2\big) \\
&\leq (K_u-1)\Big((1+\frac{1}{K_u-1})^{K_u} - 1\Big) \times \Big(8K_u\eta_u^2\sigma_u^2 + 8K_u\eta_u^2\sigma_g^2 + 8K_u\eta_u^2\mathbb{E}\|\nabla f(z_i^t, V^{t+1})\|^2\Big) \\
&\leq 32K_u\eta_u^2\sigma_u^2 + 32K_u\eta_u^2\sigma_g^2 + 32K_u\eta_u^2\mathbb{E}\|\nabla f(z_i^t, V^{t+1})\|^2.
\end{aligned} \tag{12}$$

where we use the inequality $(1+\frac{1}{K_u-1})^{K_u} \leq 5$ holds for any $K_u > 1$ in the last equation. Summing up from $i=1$ to m, then we complete the proof.

□

Lemma 3 (Mixing connectivity [3])**.** *Suppose the time-varying communication topology is strongly connected. It holds for $\forall i \in \{1,\cdots,m\}$ and $t \geq 0$ that*

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|\bar{u}^t - z_i^t\|^2 \leq \frac{8K_u^2\eta_u^2C^2}{(1-q)^2K_u - 8K_u^2\eta_u^2L_u^2C^2}\Big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}[\Delta_{\bar{u}}^t]\Big). \tag{13}$$

Proof. Suppose that Assumption 1 holds. Let $\lambda = 1 - nD^{-(K_u+1)\Delta B}$ and let $q = \lambda^{1/((K_u+1)\Delta B+1)}$. Then there exists a constant C, it satisfies

$$C < \frac{2\sqrt{d}D^{(K_u+1)\Delta B}}{\lambda^{\frac{(K_u+1)\Delta B+2}{(K_u+1)\Delta B+1}}}. \tag{14}$$

where d is the dimension of $\bar{u}^t$, z_i^t, and u_i^0, such that, for all $i = 1, 2, \ldots, m$ (non-virtual nodes) and $t \geq 0$,

$$\|\bar{u}^t - z_i^t\| \leq Cq^t\|u_i^0\| + \eta_u C \sum_{j=1}^{t} q^{t-j} \| \sum_{k=0}^{K_u-1} \nabla_u F_i(z_i^{t,k}, v_i^{j+1}; \xi_i)\|. \tag{15}$$

To unfold the stochastic gradient item, we get

$$\begin{aligned}
\mathbb{E}\Big\|\nabla_u F_i(z_i^{t,k}, v_i^{j+1}; \xi_i)\Big\|^2 &\leq \Big\|\nabla_u F_i(z_i^{t,k}, v_i^{t+1}; \xi_i) - \nabla_u F_i(z_i^{t,k}, v_i^{t+1}) + \nabla_u F_i(z_i^{t,k}, v_i^{t+1}) - \nabla_u F(z_i^{t,k}, V^{t+1}) \\
&\quad + \nabla_u F(z_i^{t,k}, V^{t+1}) - \nabla_u F(z_i^t, V^{t+1}) + \nabla_u F(z_i^t, V^{t+1})\Big\|^2 \\
&\leq 4\sigma_u^2 + 4\sigma_g^2 + 4L_u^2\mathbb{E}\|z_i^{t,k} - z_i^t\|^2 + 4\mathbb{E}\|\nabla_u F(z_i^t, V^{t+1})\|^2 \\
&\leq 4\sigma_u^2 + 4\sigma_g^2 + \frac{4L_u^2}{\delta^2}\mathbb{E}\|u_i^{t,k} - u_i^t\|^2 + 4\mathbb{E}\|\nabla_u F(z_i^t, V^{t+1})\|^2 \\
&\overset{a)}{\leq} 4\sigma_u^2 + 4\sigma_g^2 + \frac{128K_u L_u^2\eta_u^2}{\delta^2}\Big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}\|\nabla f(z_i^t, V^{t+1}))\|^2\Big) + 4\mathbb{E}\|\nabla_u F(z_i^t, V^{t+1})\|^2 \\
&\leq 4\Big(1 + \frac{32K_u L_u^2\eta_u^2}{\delta^2}\Big)\Big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}\|\nabla_u F(z_i^t, V^{t+1})\|^2\Big).
\end{aligned} \tag{16}$$

where a) uses Lemma 2. Focusing on the last term we have:

$$\begin{aligned}
\mathbb{E}\|\nabla_u F(z_i^t, V^{t+1})\|^2 &\leq \mathbb{E}\|\nabla_u F(z_i^t, V^{t+1}) - \nabla_u F(\bar{u}^t, V^{t+1}) + \nabla_u F(\bar{u}^t, V^{t+1})\|^2 \\
&\leq L_u^2\mathbb{E}\|\bar{u}^t - z_i^t\|^2 + \mathbb{E}[\Delta_{\bar{u}}^t].
\end{aligned} \tag{17}$$

Substituting Formula (17) and (16) into (15), then squaring both sides and taking expectations, we have

$$\begin{aligned}
\mathbb{E}\|\bar{u}^t - z_i^t\|^2 &\leq \big(Cq^t\|u_i^0\| + \eta_u C \sum_{j=1}^{t} q^{t-j}\mathbb{E}\| \sum_{k=0}^{K_u-1} \nabla_u F_i(z_i^{t,k}, v_i^{j+1}; \xi_i)\|\big)^2 \\
&\overset{a)}{\leq} 2C^2q^{2t}\|u_i^0\|^2 + 2\eta_u^2C^2\big(\sum_{j=1}^{t} q^{t-j}\mathbb{E}\| \sum_{k=0}^{K_u-1} \nabla_u F_i(z_i^{t,k}, v_i^{j+1}; \xi_i)\|\big)^2 \\
&\leq 2C^2q^{2t}\|u_i^0\|^2 + \frac{2K_u^2\eta_u^2C^2}{(1-q)^2}\mathbb{E}\|\nabla_u F_i(z_i^{t,k}, v_i^{j+1}; \xi_i)\|^2 \\
&\leq 2C^2q^{2t}\|u_i^0\|^2 + \frac{8K_u^2\eta_u^2C^2}{(1-q)^2}\Big(1 + \frac{32K_u L_u^2\eta_u^2}{\delta^2}\Big)\Big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}\|\nabla_u F(z_i^t, V^{t+1})\|^2\Big) \\
&\leq 2C^2q^{2t}\|u_i^0\|^2 + \frac{8K_u^2\eta_u^2C^2}{(1-q)^2}\Big(1 + \frac{32K_u L_u^2\eta_u^2}{\delta^2}\Big)\Big(\sigma_u^2 + \sigma_g^2 + L_u^2\mathbb{E}\|\bar{u}^t - z_i^t\|^2 + \mathbb{E}[\Delta_{\bar{u}}^t]\Big).
\end{aligned} \tag{18}$$

where a) uses $< x, y > \leq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2$.

Move $\mathbb{E}\|\bar{u}^t - z_i^t\|^2$ to the left side of the inequality and assume $\|u_i^0\| = 0$ and $0 < \eta_u < \frac{\delta}{4\sqrt{2}K_u L_u}$, then we have

$$\mathbb{E}\|\bar{u}^t - z_i^t\|^2 \leq \frac{8K_u^2\eta_u^2C^2(K_u+1)}{(1-q)^2K_u - 8K_u^2\eta_u^2L_u^2C^2(K_u+1)}\Big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}[\Delta_{\bar{u}}^t]\Big). \tag{19}$$

Summing up from $i = 1$ to m, then we complete the proof. □

D.2. Proof of Convergence Analysis

Proof Outline and the Challenge of Dependent Random Variables. We start with

$$F\left(\bar{u}^{t+1}, V^{t+1}\right) - F\left(\bar{u}^{t}, V^{t}\right) = F\left(\bar{u}^{t}, V^{t+1}\right) - F\left(\bar{u}^{t}, V^{t}\right) + F\left(\bar{u}^{t+1}, V^{t+1}\right) - F\left(\bar{u}^{t}, V^{t+1}\right). \tag{20}$$

The first line corresponds to the effect of the v-step and the second line to the u-step. The former is

$$\begin{aligned} F\left(\bar{u}^{t}, V^{t+1}\right) - F\left(\bar{u}^{t}, V^{t}\right) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\Big[F_i(\bar{u}^{t}, v_i^{t+1}) - F_i(\bar{u}^{t}, v_i^{t})\Big] \\ &\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\Big[\Big\langle \nabla_v F_i\left(\bar{u}^{t}, v_i^{t}\right), v_i^{t+1} - v_i^{t}\Big\rangle + \frac{L_v}{2}\|v_i^{t+1} - v_i^{t}\|^2\Big]. \end{aligned} \tag{21}$$

It is easy to handle with standard techniques that rely on the smoothness of $F\left(u^{t}, \cdot\right)$. The latter is more challenging. In particular, the smoothness bound for the u-step gives us

$$F\left(\bar{u}^{t+1}, V^{t+1}\right) - F\left(\bar{u}^{t}, V^{t+1}\right) \leq \left\langle \nabla_u F\left(\bar{u}^{t}, V^{t+1}\right), \bar{u}^{t+1} - \bar{u}^{t}\right\rangle + \frac{L_u}{2}\|\bar{u}^{t+1} - \bar{u}^{t}\|^2. \tag{22}$$

D.2.1 Proof of Convergence Analysis for DFedPGP

Analysis of the u-Step.

$$\begin{aligned} &\mathbb{E}\Big[F\left(\bar{u}^{t+1}, V^{t+1}\right) - F\left(\bar{u}^{t}, V^{t+1}\right)\Big] \leq \left\langle \nabla_u F\left(\bar{u}^{t}, V^{t+1}\right), \bar{u}^{t+1} - \bar{u}^{t}\right\rangle + \frac{L_u}{2}\mathbb{E}\|\bar{u}^{t+1} - \bar{u}^{t}\|^2 \\ &\leq \frac{-\eta_u}{m}\sum_{i=1}^{m}\mathbb{E}\Big\langle \nabla_u F\left(\bar{u}^{t}, V^{t+1}\right), \sum_{k=0}^{K_u-1}\nabla_u F\left(z_i^{t,k}, v_i^{t+1}; \xi_i\right)\Big\rangle + \frac{L_u}{2}\mathbb{E}\|\bar{u}^{t+1} - \bar{u}^{t}\|^2 \\ &\leq -\eta_u K_u \mathbb{E}[\Delta_{\bar{u}}^{t}] + \frac{\eta_u}{m}\sum_{i=1}^{m}\sum_{k=0}^{K_u-1}\mathbb{E}\Big\langle \nabla_u F\left(\bar{u}^{t}, V^{t+1}\right), \nabla F\left(\bar{u}^{t}, v_i^{t+1}\right) - \nabla_u F\left(z_i^{t,k}, v_i^{t+1}; \xi_i\right)\Big\rangle + \frac{L_u}{2}\mathbb{E}\|\bar{u}^{t+1} - \bar{u}^{t}\|^2 \\ &\overset{a)}{\leq} \frac{-\eta_u K_u}{2}\mathbb{E}[\Delta_{\bar{u}}^{t}] + \underbrace{\frac{\eta_u L_u^2}{2m}\sum_{i=1}^{m}\sum_{k=0}^{K_u-1}\mathbb{E}\|z_i^{t,k} - \bar{u}^{t}\|^2}_{\mathcal{T}_{1,u}} + \underbrace{\frac{L_u}{2}\mathbb{E}\|\bar{u}^{t+1} - \bar{u}^{t}\|^2}_{\mathcal{T}_{2,u}}. \end{aligned} \tag{23}$$

Where a) uses $\mathbb{E}\left[\nabla_u F(z_i^{t,k}, v_i^{t+1}; \xi_i)\right] = \nabla_u F\left(z_i^{t,k}, v_i^{t+1}\right)$ and $\langle x, y\rangle \leq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2$ for vectors x, y followed by L-smoothness.

For $\mathcal{T}_{1,u}$, we can use Lemma 3 and set $AA = \frac{8K_u^2\eta_u^2C^2(K_u+1)}{(1-q)^2K_u - 8K_u^2L_u^2\eta_u^2C^2(K_u+1)}$, then we have:

$$\mathcal{T}_{1,u} \leq \frac{K_u L_u^2 \eta_u}{2}\Big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}[\Delta_{\bar{u}}^{t}]\Big)AA. \tag{24}$$

Meanwhile, for $\mathcal{T}_{2,u}$,

$$\begin{aligned}
\mathcal{T}_{2,u} &\leq \frac{\eta_u^2 L_u}{2m}\sum_{i=1}^{m}\sum_{k=0}^{K_u-1}\left\|\nabla_u F\left(z_i^{t,k}, v_i^{t+1};\xi_i\right)\right\|^2 \\
&\overset{a)}{\leq} \frac{\eta_u^2 L_u}{2m}\sum_{i=1}^{m}\sum_{k=0}^{K_u-1}\Big\|\nabla_u F\left(z_i^{t,k}, v_i^{t+1};\xi_i\right) - \nabla_u F\left(z_i^{t,k}, v_i^{t+1}\right) + \nabla_u F\left(z_i^{t,k}, v_i^{t+1}\right) - \nabla_u F\left(z_i^t, v_i^{t+1}\right) \\
&\quad + \nabla_u F\left(z_i^t, v_i^{t+1}\right) + \nabla_u F\left(z_i^t, V^{t+1}\right) + \nabla_u F\left(z_i^t, V^{t+1}\right) - \nabla_u F\left(\bar{u}^t, V^{t+1}\right) + \nabla_u F\left(\bar{u}^t, V^{t+1}\right)\Big\|^2 \\
&\leq \frac{5}{2}\eta_u^2 K_u L_u\Big(\sigma_u^2 + \frac{L_u^2}{m\delta^2}\sum_{i=1}^{m}\mathbb{E}\|u_i^{t,k} - u_i^t\|^2 + \sigma_g^2 + \frac{L_u^2}{m}\sum_{i=1}^{m}\mathbb{E}\|z_i^t - \bar{u}^t\|^2 + \mathbb{E}[\Delta_{\bar{u}}^t]\Big) \\
&\leq \frac{5}{2}K_u L_u\eta_u^2\Big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}[\Delta_{\bar{u}}^t] + \frac{32K_u L_u^2\eta_u^2}{\delta^2}\Big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}[\Delta_{\bar{u}}^t]\Big)\Big(L_u^2 AA + 1\Big) + L_u^2\Big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}[\Delta_{\bar{u}}^t]\Big)AA\Big) \\
&\leq \frac{5}{2}K_u L_u\eta_u^2\big[1 + \frac{32K_u L_u^2\eta_u^2}{\delta^2}(L_u^2 AA + 1) + AA\big]\big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}[\Delta_{\bar{u}}^t]\big).
\end{aligned} \tag{25}$$

where we use Assumption 3, 4 and L-Smoothness in a). Based on the analysis above, we have:

$$\begin{aligned}
\mathbb{E}\Big[F\left(\bar{u}^{t+1}, V^{t+1}\right) - F\left(\bar{u}^t, V^{t+1}\right)\Big] &\leq \frac{K_u\eta_u}{2}\mathbb{E}[\Delta_{\bar{u}}^t] + \mathcal{T}_{1,u} + \mathcal{T}_{2,u} \\
&\leq \big(\frac{-\eta_u K_u}{2} + \frac{K_u L_u^2\eta_u}{2}AA + \frac{5K_u L_u\eta_u^2}{2}\big[1 + \frac{32K_u L_u^2\eta_u^2}{\delta^2}(L_u^2 AA + 1) + L_u^2 AA\big]\big)\mathbb{E}[\Delta_{\bar{u}}^t] \\
&\quad + \big(\frac{k_u L_u^2\eta_u}{2}AA + \frac{5K_u L_u\eta_u^2}{2}\big[1 + \frac{32K_u L_u^2\eta_u^2}{\delta^2}(L_u^2 AA + 1) + L_u^2 AA\big]\big)\big(\sigma_u^2 + \sigma_g^2\big).
\end{aligned} \tag{26}$$

Analysis of the v-Step.

$$\mathbb{E}\Big[F\left(\bar{u}^t, V^{t+1}\right) - F\left(\bar{u}^t, V^t\right)\Big] \leq \underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\Big\langle\nabla_v F_i\left(\bar{u}^t, v_i^t\right), v_i^{t+1} - v_i^t\Big\rangle}_{\mathcal{T}_{1,v}} + \underbrace{\frac{L_v}{2m}\sum_{i=1}^{m}\mathbb{E}\|v_i^{t+1} - v_i^t\|^2}_{\mathcal{T}_{2,v}}. \tag{27}$$

For $\mathcal{T}_{1,v}$,

$$\begin{aligned}
\mathcal{T}_{1,v} &\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\Big\langle\nabla_v F_i\left(\bar{u}^t, v_i^t\right) - \nabla_v F_i\left(z_i^t, v_i^t\right) + \nabla_v F_i\left(z_i^t, v_i^t\right), -\eta_v\sum_{k=0}^{K_v-1}\mathbb{E}\nabla_v F_i(u_i^t, v_i^t;\xi_i)\Big\rangle \\
&\overset{a)}{\leq} \frac{-\eta_v K_v}{m}\sum_{i=1}^{m}\mathbb{E}\|\nabla_v F_i(u_i^t, v_i^t)\|^2 + \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\Big\langle\nabla_v F_i\left(\bar{u}^t, v_i^t\right) - \nabla_v F_i\left(z_i^t, v_i^t\right), v_i^{t+1} - v_i^t\Big\rangle \\
&\overset{b)}{\leq} -\eta_v K_v\mathbb{E}[\Delta_v^t] + \underbrace{\frac{L_{vu}^2}{2m}\sum_{i=1}^{m}\mathbb{E}\|\bar{u}^t - z_i^t\|^2}_{\mathcal{T}_{3,v}} + \underbrace{\frac{1}{2m}\sum_{i=1}^{m}\mathbb{E}\|v_i^{t+1} - v_i^t\|^2}_{\frac{1}{L_v}\mathcal{T}_{2,v}}.
\end{aligned} \tag{28}$$

where a) and b) is get from the unbiased expectation property of $\nabla_v F_i(u_i^t, v_i^t;\xi_i)$ and $< x, y > \leq \frac{1}{2}(\|x\|^2 + \|y\|^2)$, respectively.

For $\mathcal{T}_{2,v}$, according to Lemma 1, we have

$$\begin{aligned}
\mathcal{T}_{2,v} &\leq \frac{L_v}{2}\Big(\frac{16\eta_v^2 K_v^2}{m}\sum_{i=1}^{m}\mathbb{E}\|\nabla_v F_i(u_i^t, v_i^t)\|^2 + 8\eta_v^2 K_v^2\sigma_v^2\Big) \\
&\leq 8L_v\eta_v^2 K_v^2\mathbb{E}[\Delta_v^t] + 4L_v\eta_v^2 K_v^2\sigma_v^2.
\end{aligned} \tag{29}$$

For $\mathcal{T}_{3,v}$, according to Lemma 3, we have

$$\frac{L_{vu}^2}{2m}\sum_{i=1}^{m}\mathbb{E}\|\bar{u}^t - z_i^t\|^2 \leq \frac{L_{vu}^2}{2}\big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}[\Delta_{\bar{u}}^t]\big)AA. \tag{30}$$

After that, summing Formula (28), (29) and (30), we have

$$\begin{aligned}\mathbb{E}\Big[F\left(\bar{u}^t, V^{t+1}\right) - F\left(\bar{u}^t, V^t\right)\Big] \leq \Big(-\eta_v K_v + 8\eta_v^2 K_v^2 L_v + 8\eta_v^2 K_v^2\Big)\mathbb{E}[\Delta_v^t] + 4\eta_v^2 K_v^2 \sigma_v^2 L_v^2(1+L_v) \\ + \frac{L_{vu}^2}{2}\big(\sigma_u^2 + \sigma_g^2 + \mathbb{E}[\Delta_{\bar{u}}^t]\big)AA.\end{aligned} \tag{31}$$

Obtaining the Final Convergence Bound.

$$\begin{aligned}&\mathbb{E}\Big[F\left(\bar{u}^{t+1}, V^{t+1}\right) - F\left(\bar{u}^t, V^t\right)\Big] = \mathbb{E}\Big[F\left(\bar{u}^t, V^{t+1}\right) - F\left(\bar{u}^t, V^t\right) + F\left(\bar{u}^{t+1}, V^{t+1}\right) - F\left(\bar{u}^t, V^{t+1}\right)\Big] \\ &\leq \big(\frac{-\eta_u K_u}{2} + \frac{K_u L_u^2 \eta_u}{2}AA + \frac{5K_u L_u \eta_u^2}{2}\big[1 + \frac{32K_u L_u^2 \eta_u^2}{\delta^2}(L_u^2 AA + 1) + L_u^2 AA\big] + \frac{L_{vu}^2}{2}AA\big)\mathbb{E}[\Delta_{\bar{u}}^t] \\ &+ \Big(-\eta_v K_v + 8\eta_v^2 K_v^2 L_v + 8\eta_v^2 K_v^2\Big)\mathbb{E}[\Delta_v^t] + 4\eta_v^2 K_v^2 L_v^2 \sigma_v^2 (1+L_v) \\ &+ \big(\frac{K_u L_u^2 \eta_u}{2}AA + \frac{5K_u L_u \eta_u^2}{2}\big[1 + \frac{32K_u L_u^2 \eta_u^2}{\delta^2}(L_u^2 AA + 1) + L_u^2 AA\big] + \frac{L_{vu}^2}{2}AA\big)\Big(\sigma_u^2 + \sigma_g^2\Big).\end{aligned} \tag{32}$$

Summing from $t = 1$ to T, assume the local learning rates satisfy $\eta_u = \mathcal{O}(1/L_u K_u\sqrt{T}), \eta_v = \mathcal{O}(1/L_v K_v\sqrt{T})$, F^* is denoted as the minimal value of F, i.e., $F(\bar{u}, V) \geq F^*$ for all $\bar{u} \in \mathbb{R}^d$, and $V = (v_1, \ldots, v_m) \in \mathbb{R}^{d_1+\ldots+d_m}$. Assume $C^2 \ll (1-q)^2 T$, then unfold AA, we can generate

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T}\big(\frac{1}{L_u}\mathbb{E}[\Delta_{\bar{u}}^t] + \frac{1}{L_v}\mathbb{E}[\Delta_v^t]\big) \leq \mathcal{O}\Big(\frac{F(\bar{u}^1, V^1) - F^*}{\sqrt{T}} + \frac{(1+L_v)\sigma_v^2}{\sqrt{T}} + \big(\sigma_u^2 + \sigma_g^2\big)\big(\frac{C^2}{(1-q)^2 L_u T} \\ + \frac{1}{K_u L_u\sqrt{T}} + \frac{1}{K_u L_u \delta^2 T^{3/2}} + \frac{C^2}{K_u L_u (1-q)^2 T^{3/2}} + \frac{L_{vu}^2 C^2}{(1-q)^2 L_u^2\sqrt{T}}\big).\end{aligned} \tag{33}$$

Combining $\chi := \max\{L_{uv}, L_{vu}\}/\sqrt{L_u L_v}$ in Assumption 2 and assume that

$$\begin{aligned}\sigma_1^2 &= (1+L_v)\sigma_v^2 + \big(\frac{1}{K_u L_u} + \frac{L_v \chi^2 C^2}{(1-q)^2 L_u}\big)\big(\sigma_u^2 + \sigma_g^2\big), \\ \sigma_2^2 &= \frac{C^2}{(1-q)^2 L_u}\big(\sigma_u^2 + \sigma_g^2\big), \\ \sigma_3^2 &= \big(\frac{1}{K_u L_u \delta^2} + \frac{C^2}{(1-q)^2 K_u L_u}\big)\big(\sigma_u^2 + \sigma_g^2\big).\end{aligned} \tag{34}$$

Then, we have the final convergence bound:

$$\frac{1}{T}\sum_{i=1}^{T}\big(\frac{1}{L_u}\mathbb{E}[\Delta_{\bar{u}}^t] + \frac{1}{L_v}\mathbb{E}[\Delta_v^t]\big) \leq \mathcal{O}\Big(\frac{F(\bar{u}^1, V^1) - F^*}{\sqrt{T}} + \frac{\sigma_1^2}{\sqrt{T}} + \frac{\sigma_2^2}{T} + \frac{\sigma_3^2}{\sqrt{T^3}}\Big). \tag{35}$$