

FEDERATED LEARNING IN STREAMING SUBSPACE

Anonymous authors

Paper under double-blind review

We provide more details about our work and results in the appendix. Below is the content of the appendix:

- Appendix A: More detailed discussion of related work.
- Appendix B: More comparative experimental results and ablation experiments.
- Appendix C: Proof and discussion of the theoretical analysis of FLSS.
- Appendix D: The hyperparameters of baseline algorithms.
- Appendix E: Further discussion on our proposed strategy FLSS.

A RELATED WORK

A.1 FEDERATED LEARNING

Federated learning is a distributed machine learning framework through iterative communication and computation between servers and clients. FedAvg McMahan et al. (2017) is a well-known FL method and the basic framework of many FL methods. We first introduce its main steps: (1) Server sends the current global model to clients; (2) The clients initialize the current global model as its own local model; (3) The clients train the local model on its own private data and send the trained local models to the server; (4) The server receives the client models and aggregates them to obtain the global model, and then resends it to the clients. However, the above solutions often face the problems of high communication and poor performance in heterogeneous scenarios. Therefore, a lot of work have been carried out to solve the above problems.

Traditional Federated Learning. Federated learning algorithms designed to enhance performance in heterogeneous environments can be divided into four different types Zhang et al. (2023): regularization-based FL Li et al. (2020); Acar et al. (2021); Kim et al. (2022), update correction-based FL Karimireddy et al. (2020); Gao et al. (2022); Niu & Deng (2022), model split-based FL Li et al. (2021); Jiang et al. (2022), and knowledge distillation-based FL Zhu et al. (2021); Lee et al. (2022); Gong et al. (2022); Huang et al. (2022). In the field of regularization-based FL, FedProx Li et al. (2020) introduces a proximal term to reduce the Euclidean distance between the global model and the local model, while FedDyn Acar et al. (2021) adopts dynamic regularization to align the local optimal point with the minimum value of the global empirical loss. For FL based on update correction, methods such as SCAFFOLD Karimireddy et al. (2020) and FedDC Gao et al. (2022) employ global gradient calibration to mitigate local model drift. However, these methods require the transmission of twice the message size required by FedAvg McMahan et al. (2017). In model split-based FL, MOON Li et al. (2021) enhances the consistency between local and global model representations by adding a contrastive learning loss. Meanwhile, in knowledge distillation-based FL, FedGen Zhu et al. (2021) utilizes a generator trained on the server to absorb local insights and utilizes the synthesized knowledge as an inductive bias to guide the local training process. Furthermore, FedNTD Lee et al. (2022) uses local non-true distillation to solve the problem of forgetting global information during local training.

Communication-efficient Federated Learning. To address the challenge of communication overhead in federated learning, many frameworks for gradient compression techniques have been proposed. Fetchsgd Rothchild et al. (2020) utilizes sketching techniques to effectively compress local gradients. Signsgd+EF Karimireddy et al. (2019) combines error feedback with 1-bit quantization, which reduces communication costs and improves the generalization ability of Signsgd. Furthermore, STC Sattler et al. (2019) is specifically designed for federated learning, combining top-k sparsity and quantization techniques to optimize data transfer. Similarly, DGC Lin et al. (2018) utilizes sparsification to preserve important gradients while minimizing bandwidth in distributed training

environments. LBG M Azam et al. (2021) utilizes the low-rank characteristics of gradient space to reduce communication requirements; however, it does not fully consider the relationship between early global model information and local model updates. Although these methods have proven their feasibility in reducing communication load, their effectiveness is often limited in heterogeneous environments. This limitation is due to the stochastic nature of the compression framework and the fact that the client does not have complete information about the global model.

A.2 TRAINING IN TINY SUBSPACE

Many studies have emphasized the inherent low-dimensional characteristics of neural networks Tuddenham et al. (2020); Vinyals & Povey (2012); Gressmann et al. (2020). A seminal study in Li et al. (2018); Gur-Ari et al. (2018) reveals that training a neural network within a randomly chosen subspace helps to achieve parameter compression, though the final accuracy may not be as high as that in the original space. The following work Gressmann et al. (2020) improved the training of fixed random subspaces by considering different layers of the network and re-drawing the random subspace at each step. Different from random subspaces, Li *et al.* Li et al. (2022a;b) successfully extracted a subspace that approximates the entire parameter trajectory by performing principal component analysis on a pre-trained neural network. Efficient dimensionality reduction is achieved by limiting the training process to this subspace.

However, although the above subspace contains model information to a certain extent, it is essentially limited to the early stage of pre-training. Often it takes multiple epochs to reach its full potential. In contrast, streaming subspace adapts to data changes by continuously updating the subspace to dynamically capture real-time model information.

B ADDITIONAL EXPERIMENTS

B.1 EFFECT OF PROJECTED OBJECTS

Table 1: We tested the impact of using streaming subspace on model updates or gradients respectively on algorithm performance.

Method	Proj(\mathbf{g}_t^k)		Proj($\nabla F_k \mathbf{w}_{t,i}^k$)			Proj(\mathbf{g}_t^k) + Proj($\nabla F_k(\mathbf{w}_{t,i}^k)$)		
	$sr = 1$	$sr = 3$	$sr = 1$	$sr = 3$	$sr = 10$	$sr = 1$	$sr = 3$	$sr = 10$
FedAvg+FLSS	57.34	60.02	55.98	57.58	56.58	56.10	57.01	56.28

We apply the streaming subspace strategy to different locations of FedAvg McMahan et al. (2017), including the model updates and gradients of the local model. Then we tested the performance of using ResNet-18 at different scaled (sr) learning rates in Cifar10, as shown in Tab. 1. Using the streaming subspace strategy for model updates can improve communication efficiency and performance through little additional computational overhead.

B.2 HETEROGENEITY

To further demonstrate the performance of the FLSS-equipped algorithms on different datasets, we conduct additional experiments in heterogeneous scenarios, as shown in Tab. 2. From the results, we can see that the algorithms with FLSS outperform FedAvg McMahan et al. (2017) and Signsgd+EF Karimireddy et al. (2019), which suggests that the FLSS strategy can effectively utilize the early knowledge of the global model to achieve better performance.

B.3 CONVERGENCE

We present the loss throughout the training process in Fig. 1a. The experimental results confirm that the FLSS-equipped FL algorithm converges. Notably, the FLSS loss slightly increases at 200 rounds before continuing to decrease. This indicates that the model needs several rounds to adapt to the

Table 2: Test accuracy on different datasets under Dirichlet distribution. Cifar100* represents using ResNet-18 on Cifar100.

Method	FMNIST			Cifar100*			TINY		
	$\beta=0.1$	$\beta=0.5$	$\beta=1$	$\beta=0.1$	$\beta=0.5$	$\beta=1$	$\beta=0.1$	$\beta=0.5$	$\beta=1$
FedAvg	85.06	91.02	91.18	23.04	26.42	27.43	14.26	15.61	18.47
FedProx	84.06	91.03	91.14	23.00	26.25	27.02	14.12	15.63	18.35
Moon	85.03	91.18	91.29	23.08	26.54	27.10	15.21	15.72	18.51
FedDyn	83.11	90.68	90.92	24.41	28.65	29.09	15.63	—	19.07
FedGen	84.27	91.17	91.25	23.42	26.24	27.85	15.44	15.80	18.72
FedNTD	84.96	91.15	91.36	22.84	26.51	27.15	15.43	15.77	18.39
FedAvg+FLSS	86.25	91.64	91.29	24.31	29.32	30.68	17.01	17.84	19.00
Fetchsgd	79.79	90.67	90.56	21.99	24.43	25.35	14.12	14.46	16.16
Signsgd+EF	80.79	90.76	90.37	22.57	25.99	26.01	14.02	14.20	16.86
STC	80.40	85.33	85.78	22.38	25.92	26.42	13.86	14.59	15.93
Sign+EF+FLSS	81.32	91.05	91.02	23.01	27.59	26.17	13.20	15.04	17.32

streaming subspace intervention. The loss reduction shows that constraining local model updates to subspace can continue to train and converge.

To verify the low-rank characteristics of different network update spaces, we compute the Singular Values (SV) of the other networks, as shown in Fig. 1b, Fig. 1c, and Fig. 1d. We observe that smaller networks, with larger percentages of the first few principal components, require fewer subspace degrees of freedom to approximate the update trajectory. In contrast, larger networks, such as ResNet-50, need more orthogonal bases to approximate their update trajectory.

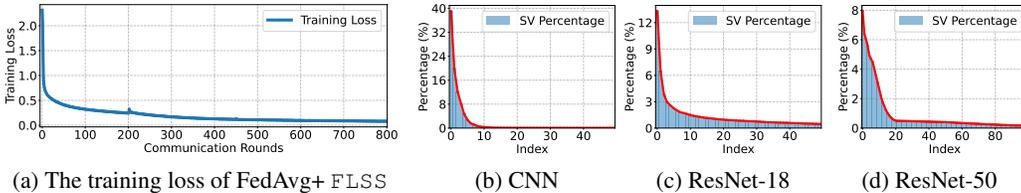


Figure 1: (a) is the training loss curve of FedAvg+FLSS in FMNIST. (b) to (d) are the singular value distributions of the global model update trajectory of CNN in Cifar10, ResNet-18 in Cifar10, and ResNet-50 in Cifar100, respectively.

B.4 FEATURES VISUALIZATION

We visualize the feature representations of the different algorithms in FMNIST using t-SNE Van der Maaten & Hinton (2008) in Fig. 2. The feature representations extracted by FedAvg+FLSS become more and more distinct with iterative updates of the algorithm. Based on Fig. 2b and Fig. 2f, it can be seen that the FL algorithm equipped with FLSS ends up with more distinguishable features than those extracted by FedAvg.

B.5 DIFFERENT LOCAL EPOCHS

Increasing local epochs results in higher computational costs but reduces the number of communication rounds. We evaluate the performance of CNN and ResNet-18 over 400 rounds on Cifar10 with $\beta = 0.5$, as shown in Tab. 3. Across different local epochs settings, FLSS performs better than most baselines. Notably, FLSS shows more significant performance improvement with fewer local epochs. Specifically, with 1 local epoch, FedAvg combined with FLSS achieves improvements of 2.43% and 3.41%, respectively.

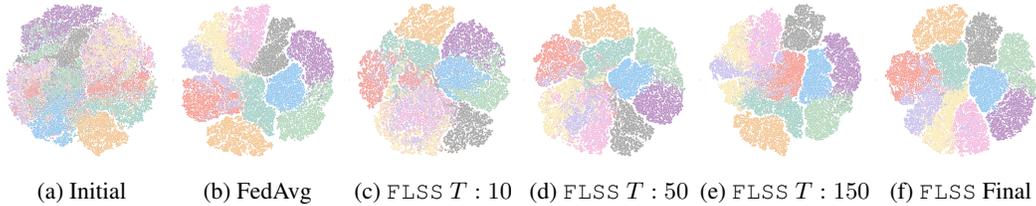


Figure 2: t-SNE visualization of features extracted by the CNN model at different times on FMNIST. FLSS Final and FedAvg denote the final features with and without FLSS, respectively. T denotes the number of communication rounds.

Table 3: The impact of different client local training epochs on the performance of different algorithms.

Method	CNN	Local Epochs			ResNet	Local Epochs		
	Com.cost	1	5	10	Com.cost	1	5	10
FedAvg	0.17 G	59.88	65.29	65.31	2.23 G	48.51	56.17	58.88
FedProx	0.17 G	59.77	65.38	65.11	2.23 G	48.44	56.03	58.27
Moon	0.17 G	59.93	65.33	65.26	2.23 G	48.74	56.28	59.24
FedGen	—	60.03	65.61	65.12	—	49.23	56.24	59.04
FedNTD	0.17 G	60.42	65.52	65.15	2.23 G	49.64	56.42	59.06
FedAvg+FLSS	35.14 M	62.31	67.19	65.53	0.45 G	51.92	57.34	58.93
Fetchsgd	43.92 M	52.20	59.83	58.57	0.56 G	—	54.18	56.04
Signsgd+EF	5.49 M	59.21	64.27	63.95	69.88 M	47.22	54.23	55.68
STC	5.49 M	59.55	59.02	62.38	69.88 M	47.91	55.03	55.35
Sign+EF+FLSS	1.11 M	59.01	64.69	64.06	14.0 M	49.02	54.77	56.81

B.6 COMPARISON WITH OTHER BASELINES

Table 4: Average test accuracy and communication cost of different algorithms under varying degrees of heterogeneity.

Method	Cifar10			Cifar100		
	$\beta=0.1$	$\beta=0.5$	$\beta=1$	$\beta=0.1$	$\beta=0.5$	$\beta=1$
LBGM	56.37(0.21G)	64.42(0.20G)	65.15(0.20G)	27.70(0.27G)	28.81(0.27G)	30.01(0.26 G)
FedAvg+FLSS	59.44(0.21G)	67.19(0.21G)	69.82(0.21G)	28.41(0.22G)	31.47(0.22G)	31.61(0.22G)

We compared the performance and total communication cost of 400 rounds between LBGM and FLSS, as shown in Tab. 4. LBGM is a low-rank method based on gradient space, focusing on local training trajectories, while FLSS targets low-rank properties of the global model. FLSS projects local updates onto the global subspace to filter out harmful components. Additionally, LBGM emphasizes a single gradient direction early in training, while FLSS uses all early training information, unifying them into a low-rank subspace.

C CONVERGENCE OF FLSS

C.1 NOTATION

We defined the local model update at device k as $\mathbf{g}_t^k = \mathbf{w}_{t+1}^k - \mathbf{w}_t$. We define the low-dimensional trajectory of the local model updated on device k within subspace \mathbf{P} as $\hat{\mathbf{g}}_t^k$, $\hat{\mathbf{g}}_t^k = \text{Proj}_{\mathbf{P}}(\text{Proj}_{\mathbf{P}^\perp}(\mathbf{w}_{t+1}^k - \mathbf{w}_t))$. The global model parameters updated as $\mathbf{w}_{t+1} = \mathbf{w}_t +$

$\frac{1}{M} \sum_{k \in \mathcal{M}_t} \hat{\mathbf{g}}_t^k$, we define the following auxiliary variable: $\mathbf{v}_{t+1} = \mathbf{w}_t + \frac{1}{N} \sum_{k=1}^N \hat{\mathbf{g}}_t^k$. We have

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 &= \|\mathbf{w}_{t+1} - \mathbf{v}_{t+1} + \mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \\ &= \|\mathbf{w}_{t+1} - \mathbf{v}_{t+1}\|_2^2 + \|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 + 2\langle \mathbf{w}_{t+1} - \mathbf{v}_{t+1}, \mathbf{v}_{t+1} - \mathbf{w}^* \rangle. \end{aligned} \quad (1)$$

In the following, we bound the average of the terms on the right hand side (RHS).

C.2 KEY LEMMAS

Lemma C.2.1 *Suppose that Assumptions 3.4.1 to 3.4.4 hold, the difference between \mathbf{w}_{t+1} and \mathbf{v}_{t+1} can be bounded*

$$\mathbb{E} \left[\|\mathbf{w}_{t+1} - \mathbf{v}_{t+1}\|_2^2 \right] \leq \frac{2\eta_t^2 \tau^2 G^2}{M} + \frac{2 \sum_{r=R+1}^D \sigma_r^2}{M \sum_{r=1}^D \sigma_r^2} \eta_t^2 \tau^2 G^2.$$

Proof. See Appendix C.7.

Lemma C.2.2 *Suppose that Assumptions 3.4.1 to 3.4.4 hold, the upper bound of $\mathbb{E} \left[\|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \right]$ is as follows*

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \right] &\leq (1 - \mu\eta_t\tau(1 - \eta_t)) \mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right] \\ &\quad + (2 + \mu) \eta_t^2 G^2 \frac{\tau(\tau+1)(2\tau+1)}{6} + (2L\eta_t^2\tau^2 + 4L\eta_t^2\tau)\Gamma. \end{aligned}$$

Proof. See Appendix C.8.

Lemma C.2.3 *Let $\mathbb{E}_{\mathcal{M}_t}$ denote expectation over the device scheduling randomness at the global iteration t . We have $\mathbb{E}_{\mathcal{M}_t} [\mathbf{w}_{t+1}] = \mathbf{v}_{t+1}$, from which it follows that*

$$\mathbb{E}_{\mathcal{M}_t} [\langle \mathbf{w}_{t+1} - \mathbf{v}_{t+1}, \mathbf{v}_{t+1} - \mathbf{w}^* \rangle] = 0.$$

Proof. Due to the randomness of the device scheduling policy and the scheduling update of each device appears $\binom{N-1}{M-1}$ times, it follows that

$$\mathbb{E}_{\mathcal{M}_t} \left[\frac{1}{M} \sum_{k \in \mathcal{M}_t} \hat{\mathbf{g}}_t^k \right] = \frac{\binom{N-1}{M-1}}{M \binom{N}{M}} \sum_{k=1}^N \hat{\mathbf{g}}_t^k = \frac{1}{N} \sum_{k=1}^N \hat{\mathbf{g}}_t^k. \quad (2)$$

C.3 THEOREMS

Theorem C.3.1 *Suppose that Assumptions 3.4.1 to 3.4.4 hold and a learning rate η_t such that $0 < \eta_t \leq \min\{\frac{1}{\mu B}, \frac{1}{L(\tau+1)}\}$ is chose, we have*

$$\mathbb{E} \left[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \right] \leq (1 - \mu\eta_t B) \mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right] + \eta_t^2 C, \quad (3)$$

where

$$B = \tau - \frac{\tau}{L(\tau+1)}, C = (2 + \mu) G^2 \frac{2\tau^3 + 3\tau^2 + \tau}{6} + (2L\tau^2 + 4L\tau)\Gamma + \frac{2(1 + \rho_R)\tau^2 G^2}{M}. \quad (4)$$

Proof. See Appendix C.9.

C.4 COROLLARIES

Corollary C.4.1 *Suppose that Assumptions 3.4.1 to 3.4.4 hold with $\mu \geq 0$, a constant learning rate $\eta > 0$ such that $\eta \leq \frac{1}{L(\tau+1)}$, we have*

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \frac{L}{2} (1 - \mu\eta B)^T \mathbb{E} \left[\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2 \right] + \frac{L}{2} \sum_{t=1}^T \eta^2 (1 - \mu\eta B)^{T-t} C. \quad (5)$$

Proof. See Appendix C.10.1.

Corollary C.4.2 *Let Assumptions 3.4.1 to 3.4.4 hold and L, μ, G, ρ_R be defined therein. Choose $\kappa = \frac{L}{\mu}$, $\gamma = \frac{2L^2(\tau+1)^2}{\mu\tau(L(\tau+1)-1)} - 1$, and the learning rate $\eta_t = \frac{2}{\mu B(\gamma+t)}$. Then FLSS satisfies*

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \frac{\kappa}{\gamma + T - 1} \left(\frac{2C}{\mu B^2} + \frac{\mu(\gamma + 1)}{2} \mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2] \right). \quad (6)$$

Proof. See Appendix C.10.2.

C.5 DISCUSSION ON ASSUMPTION 3.4.4

We define \mathcal{G} to be the set consisting of global model updates and $\mathbf{G} \in \mathbb{R}^{D \times J}$ to be the matrix consisting of the set of global model updates \mathcal{G} . In the experiments, $\mathbf{G} \in \mathbb{R}^{D \times J}$ can be obtained by sampling the global updates. A truncated singular value decomposition of \mathbf{G} of rank R yields \mathbf{P} , whose singular values are $\sigma_1, \dots, \sigma_D$. Based on the linearity property of expectation, we have

$$\mathbb{E}[\text{Proj}(\mathbf{g}_t)] = \mathbb{E}[\mathbf{P}\mathbf{P}^T\mathbf{g}_t] = \mathbf{P}\mathbf{P}^T\mathbb{E}[\mathbf{g}_t] = \text{Proj}(\mathbb{E}[\mathbf{g}_t]). \quad (7)$$

Due to the low-rank character of the global model update space, the last few singular values are small and the corresponding dimensions are almost null space. In Eq. (7), it is assumed that the expectation of the global model update $\mathbb{E}[\mathbf{g}_t]$ will be contained within the subspace \mathbf{P} , so $\mathbb{E}[\text{Proj}(\mathbf{g}_t)] = \mathbb{E}[\mathbf{g}_t]$.

C.6 PROOFS OF PROPOSITION 3.4.1

For the global update $\mathbf{g}_t \in \mathcal{G}$, we compute the expectation of the squared projection error:

$$\mathbb{E}_{\mathbf{g}_t \in \mathcal{G}}(\|\mathbf{g}_t - \text{Proj}(\mathbf{g}_t)\|^2) = \frac{1}{J} \sum_{j=1}^J \|\mathbf{G}_j - \mathbf{P}\mathbf{P}^T\mathbf{G}_j\|^2. \quad (8)$$

Using trace properties, we have

$$\frac{1}{J} \sum_{j=1}^J \|\mathbf{G}_j - \mathbf{P}\mathbf{P}^T\mathbf{G}_j\|^2 = \frac{1}{J} \text{tr}(\mathbf{G}^T(\mathbf{I}_D - \mathbf{P}\mathbf{P}^T)\mathbf{G}) = \frac{1}{J} \text{tr}((\mathbf{I}_D - \mathbf{P}\mathbf{P}^T)\mathbf{G}\mathbf{G}^T). \quad (9)$$

Since $\mathbf{I}_D - \mathbf{P}\mathbf{P}^T$ projected $\mathbf{G}\mathbf{G}^T$ to a space orthogonal to the columns of \mathbf{P} , we have

$$\mathbb{E}(\|\mathbf{g}_t - \text{Proj}(\mathbf{g}_t)\|^2) = \frac{1}{J} \sum_{r=R+1}^D \sigma_r^2 \leq \frac{\sum_{r=R+1}^D \sigma_r^2}{\sum_{r=1}^D \sigma_r^2} \eta_t^2 \tau^2 G^2. \quad (10)$$

The last inequality is due to Assumption 3.4.3, $\mathbb{E}[\|\nabla F_k(\mathbf{w}_t^k, \xi_t^k)\|_2^2] \leq G^2$, so that $\mathbb{E}[\|\mathbf{g}_t\|_2^2] \leq \eta_t^2 \tau^2 G^2$ and $\|\mathbf{G}\|_2^2 \leq J\eta_t^2 \tau^2 G^2$.

C.7 PROOF OF LEMMA C.2.1

According to the definitions, $\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{1}{M} \sum_{k \in \mathcal{M}_t} \hat{\mathbf{g}}_t^k$, $\mathbf{v}_{t+1} = \mathbf{w}_t + \frac{1}{N} \sum_{k=1}^N \hat{\mathbf{g}}_t^k$, $i_m \in \mathcal{M}_t$, and $\hat{\mathbf{g}}_t \triangleq \frac{1}{N} \sum_{k=1}^N \hat{\mathbf{g}}_t^k$. Taking the expectation of the first term of Eq. (1), we have

$$\begin{aligned} \mathbb{E}_{\mathcal{M}_t}[\|\mathbf{w}_{t+1} - \mathbf{v}_{t+1}\|_2^2] &= \mathbb{E}_{\mathcal{M}_t} \left[\left\| \frac{1}{M} \sum_{m=1}^M (\hat{\mathbf{g}}_t^{i_m} - \hat{\mathbf{g}}_t) \right\|_2^2 \right] \\ &= \frac{1}{M^2} \mathbb{E}_{\mathcal{M}_t} \left[\sum_{m=1}^M \|\hat{\mathbf{g}}_t^{i_m} - \hat{\mathbf{g}}_t\|_2^2 + \sum_{m=1}^M \sum_{m'=1, m' \neq m}^M \langle \hat{\mathbf{g}}_t^{i_m} - \hat{\mathbf{g}}_t, \hat{\mathbf{g}}_t^{i_{m'}} - \hat{\mathbf{g}}_t \rangle \right]. \end{aligned} \quad (11)$$

Due to the symmetry, it follows that

$$\mathbb{E}_{\mathcal{M}_t} \left[\sum_{m=1}^M \|\hat{\mathbf{g}}_t^{i_m} - \hat{\mathbf{g}}_t\|_2^2 \right] = \frac{\binom{N-1}{M-1}}{\binom{N}{M}} \sum_{k=1}^N \|\hat{\mathbf{g}}_t^k - \hat{\mathbf{g}}_t\|_2^2 = \frac{M}{N} \sum_{k=1}^N \|\hat{\mathbf{g}}_t^k - \hat{\mathbf{g}}_t\|_2^2, \quad (12)$$

where the first equality is because there are $\binom{N}{M}$ choices in selecting M from N clients. For each index $k, k \in [N]$, the number of times is selected is $\binom{N-1}{M-1}$.

$$\begin{aligned} \mathbb{E}_{\mathcal{M}_t} \left[\sum_{m=1}^M \sum_{m'=1, m' \neq m}^M \langle \hat{\mathbf{g}}_t^{i_m} - \hat{\mathbf{g}}_t, \hat{\mathbf{g}}_t^{i_{m'}} - \hat{\mathbf{g}}_t \rangle \right] &= \frac{\binom{N-2}{M-2}}{\binom{N}{M}} \sum_{k=1}^K \sum_{\substack{k'=1 \\ k' \neq k}}^N \langle \hat{\mathbf{g}}_t^k - \hat{\mathbf{g}}_t, \hat{\mathbf{g}}_t^{k'} - \hat{\mathbf{g}}_t \rangle \\ &= -\frac{\binom{N-2}{M-2}}{\binom{N}{M}} \sum_{k=1}^N \|\hat{\mathbf{g}}_t^k - \hat{\mathbf{g}}_t\|_2^2 \leq 0. \end{aligned} \quad (13)$$

where the first equality is because, for each particular index pair (k, k') , $k' \in [N]$, $k \neq k'$, the number of times is selected is $\binom{N-2}{M-2}$, and the second equality is because $\left\| \sum_{k=1}^N (\hat{\mathbf{g}}_t^k - \hat{\mathbf{g}}_t) \right\|_2^2 = 0$. Substituting Eq. (12) and Eq. (13) into Eq. (11) yields

$$\begin{aligned} &\mathbb{E} \left[\|\mathbf{w}_{t+1} - \mathbf{v}_{t+1}\|_2^2 \right] \\ &= \frac{1}{NM} \sum_{k=1}^N \mathbb{E} \left[\|\hat{\mathbf{g}}_t^k - \hat{\mathbf{g}}_t\|_2^2 \right] + \frac{\binom{N-2}{M-2}}{M^2 \binom{N}{M}} \sum_{k=1}^N \sum_{\substack{k'=1 \\ k' \neq k}}^N \langle \hat{\mathbf{g}}_t^k - \hat{\mathbf{g}}_t, \hat{\mathbf{g}}_t^{k'} - \hat{\mathbf{g}}_t \rangle \leq \frac{1}{NM} \sum_{k=1}^N \mathbb{E} \left[\|\hat{\mathbf{g}}_t^k - \hat{\mathbf{g}}_t\|_2^2 \right] \\ &= \frac{1}{NM} \left(\sum_{k=1}^N \mathbb{E} \left[\|\hat{\mathbf{g}}_t^k\|_2^2 \right] - \mathbb{E} \left[\|\hat{\mathbf{g}}_t\|_2^2 \right] \right) \leq \frac{1}{NM} \sum_{k=1}^N \mathbb{E} \left[\|\hat{\mathbf{g}}_t^k\|_2^2 \right] \leq \frac{1}{NM} \sum_{k=1}^N \mathbb{E} \left[\|\mathbf{g}_t^k + \mathbf{e}_t^k\|_2^2 \right] \\ &\leq \frac{2\eta_t^2 \tau^2 G^2}{M} + \frac{2 \sum_{r=R+1}^D \sigma_r^2}{M \sum_{r=1}^D \sigma_r^2} \eta_t^2 \tau^2 G^2. \end{aligned} \quad (14)$$

C.8 PROOF OF LEMMA C.2.2

According to the definition of \mathbf{v}_{t+1} , $\mathbf{v}_{t+1} = \mathbf{w}_t + \frac{1}{N} \sum_{k=1}^N \hat{\mathbf{g}}_t^k$, taking the expectation and expanding the second term of the Eq. (1), we have

$$\mathbb{E} \left[\|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \right] = \underbrace{\mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right]}_{A_1} + \underbrace{\mathbb{E} \left[\left\| \frac{1}{N} \sum_{k=1}^N \hat{\mathbf{g}}_t^k \right\|_2^2 \right]}_{A_2} + 2 \underbrace{\mathbb{E} \left[\left\langle \mathbf{w}_t - \mathbf{w}^*, \frac{1}{N} \sum_{k=1}^N \hat{\mathbf{g}}_t^k \right\rangle \right]}_{A_3}. \quad (15)$$

For A_2 , due to the convexity of $\|\cdot\|_2^2$ and the L -smoothness of $F_k(\cdot)$, $\|\nabla F_k(\mathbf{w}_{t,i}^k, \xi_{t,i}^k)\|_2^2 \leq 2L(F_k(\mathbf{w}_{t,i}^k) - F_k^*)$, we have

$$\begin{aligned} A_2 &\leq \frac{1}{N} \sum_{k=1}^N \mathbb{E} \left[\|\hat{\mathbf{g}}_t^k\|_2^2 \right] \leq \frac{1}{N} \sum_{k=1}^N \mathbb{E} \left[\|\mathbf{g}_t^k\|_2^2 \right] = \frac{\eta_t^2}{N} \sum_{k=1}^N \mathbb{E} \left[\left\| \sum_{i=1}^{\tau} \nabla F_k(\mathbf{w}_{t,i}^k, \xi_{t,i}^k) \right\|_2^2 \right] \\ &\leq \frac{\eta_t^2 \tau}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\|\nabla F_k(\mathbf{w}_{t,i}^k, \xi_{t,i}^k)\|_2^2 \right] \leq \frac{2L\eta_t^2 \tau}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F_k^*]. \end{aligned} \quad (16)$$

For A_3 , according to Assumption 3.4.4 and the definition of $\hat{\mathbf{g}}_t^k$, we can know that $\hat{\mathbf{g}}_t^k = \text{Proj}_{\mathbf{P}}(\text{Proj}_{\mathbf{P}}(\mathbf{g}_t^k))$, and $\text{Proj}_{\mathbf{P}}(\text{Proj}_{\mathbf{P}}(\mathbf{w}_t - \mathbf{w}^*)) = \mathbf{w}_t - \mathbf{w}^* + \epsilon_t$. We have

$$2 \mathbb{E} \left[\left\langle \mathbf{w}_t - \mathbf{w}^*, \frac{1}{N} \sum_{k=1}^N \hat{\mathbf{g}}_t^k \right\rangle \right] = \underbrace{\frac{2\eta_t}{N} \sum_{k=1}^N \mathbb{E} \left[\left\langle \mathbf{w}^* - \mathbf{w}_t, \sum_{i=1}^{\tau} \nabla F_k(\mathbf{w}_{t,i}^k, \xi_{t,i}^k) \right\rangle \right]}_{B_1}. \quad (17)$$

For B_1 , we split $\mathbf{w}^* - \mathbf{w}_t$ into $\mathbf{w}^* - \mathbf{w}_{t,i}^k$ and $\mathbf{w}_{t,i}^k - \mathbf{w}_t$, so B_1 can be split into two items: $C_1 = \frac{2\eta_t}{K} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [\langle \mathbf{w}_{t,i}^k - \mathbf{w}_t, \nabla F_k(\mathbf{w}_{t,i}^k, \xi_{t,i}^k) \rangle]$, $C_2 =$

378 $\frac{2\eta_t}{K} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [\langle \mathbf{w}^* - \mathbf{w}_{t,i}^k, \nabla F_k(\mathbf{w}_{t,i}^k, \xi_{t,i}^k) \rangle]$. So next we calculate the upper bounds
 379 of these two terms respectively. To bound C_1 , we have

$$\begin{aligned} 381 \quad C_1 &\leq \frac{\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\frac{1}{\eta_t} \|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2 + \eta_t \|\nabla F_k(\mathbf{w}_{t,i}^k, \xi_{t,i}^k)\|_2^2 \right] \\ 382 &\leq \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2 \right] + \frac{2L\eta_t^2}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F_k^*], \end{aligned} \quad (18)$$

383 where the first inequality is by Cauchy-Schwarz inequality, and the second inequality is by the
 384 L -smoothness of $F_k(\cdot)$, $\|\nabla F_k(\mathbf{w}_{t,i}^k, \xi_{t,i}^k)\|_2^2 \leq 2L(F_k(\mathbf{w}_{t,i}^k) - F_k^*)$. To bound C_2 , we have

$$\begin{aligned} 385 \quad C_2 &= \frac{2\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [\langle \mathbf{w}^* - \mathbf{w}_{t,i}^k, \nabla F_k(\mathbf{w}_{t,i}^k) \rangle] \\ 386 &\leq \frac{2\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[F_k(\mathbf{w}^*) - F_k(\mathbf{w}_{t,i}^k) - \frac{\mu}{2} \|\mathbf{w}_{t,i}^k - \mathbf{w}^*\|_2^2 \right], \end{aligned} \quad (19)$$

387 where the first equality is by $\mathbb{E}_{\xi} [\nabla F_k(\mathbf{w}_t, \xi_{t,i}^k)] = \nabla F_k(\mathbf{w}_t)$, $\forall i, k, t$ and the first inequality is by
 388 the fact that F_k is μ -strongly convex.

389 For A_3 , substituting Eq. (18) and Eq. (19) into Eq. (17), we have

$$\begin{aligned} 390 \quad 2\mathbb{E} \left[\langle \mathbf{w}_t - \mathbf{w}^*, \frac{1}{N} \sum_{k=1}^N \hat{\mathbf{g}}_t^k \rangle \right] &\leq \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2 \right] + \frac{2L\eta_t^2}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F_k^*] \\ 391 &\quad + \frac{2\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} (F_k(\mathbf{w}^*) - F_k(\mathbf{w}_{t,i}^k)) - \frac{\mu\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\|\mathbf{w}_{t,i}^k - \mathbf{w}^*\|_2^2 \right]. \end{aligned} \quad (20)$$

392 For $\mathbb{E} \left[\|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \right]$, substituting Eq. (20) and Eq. (16) into Eq. (15), we have

$$\begin{aligned} 393 \quad &\mathbb{E} \left[\|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \right] \\ 394 &\leq \mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right] + \frac{2L\eta_t^2\tau}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F_k^*] + \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2 \right] \\ 395 &\quad + \frac{2L\eta_t^2}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F_k^*] + \frac{2\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} (F_k(\mathbf{w}^*) - F_k(\mathbf{w}_{t,i}^k)) \\ 396 &\quad - \frac{\mu\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\|\mathbf{w}_{t,i}^k - \mathbf{w}^*\|_2^2 \right] \\ 397 &= \mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right] - \underbrace{\frac{\mu\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\|\mathbf{w}_{t,i}^k - \mathbf{w}^*\|_2^2 \right]}_{D_1} + \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2 \right] \\ 398 &\quad + \underbrace{\frac{2L\eta_t^2(\tau+1)}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F_k^*] - \frac{2\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} (F_k(\mathbf{w}_{t,i}^k) - F_k(\mathbf{w}^*))}_{D_2}. \end{aligned} \quad (21)$$

399 To bound D_1 , we first calculate the upper bound of $-\|\mathbf{w}_{t,i}^k - \mathbf{w}^*\|_2^2$

$$\begin{aligned} 400 \quad -\|\mathbf{w}_{t,i}^k - \mathbf{w}^*\|_2^2 &= -\|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2 - \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - 2\langle \mathbf{w}_{t,i}^k - \mathbf{w}_t, \mathbf{w}_t - \mathbf{w}^* \rangle \\ 401 &\leq -\|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2 - \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \frac{1}{\eta_t} \|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2 + \eta_t \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \end{aligned}$$

$$= -(1 - \eta_t) \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \left(\frac{1}{\eta_t} - 1 \right) \|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2, \quad (22)$$

where the first inequality is by Cauchy-Schwarz inequality. We next aim to bound D_2 . We define $\gamma_t = 2\eta_t(1 - L\eta_t\tau - L\eta_t)$. Let $\gamma_t \geq 0$, we have $\eta_t \leq \frac{1}{L(\tau+1)}$, $\gamma_t \leq 2\eta_t$. We define $\Gamma = F^* - \frac{1}{N} \sum_{k=1}^N F_k^*$, which is a measure of non-IID degree. Then we have

$$\begin{aligned} D_2 &= \frac{2L\eta_t^2(\tau+1)}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F_k^*] - \frac{2\eta_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} (F_k(\mathbf{w}_{t,i}^k) - F_k(\mathbf{w}^*)) \\ &= \underbrace{-\frac{\gamma_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F^*]}_E + \frac{2L\eta_t^2(\tau+1)}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F^* - F_k^*]. \end{aligned} \quad (23)$$

To bound E , considering $\gamma_t \geq 0$, we need to obtain the lower bound of $\frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F^*]$. Then, we split $F_k(\mathbf{w}_{t,i}^k) - F^*$ into $F_k(\mathbf{w}_{t,i}^k) - F_k(\mathbf{w}_t)$ and $F_k(\mathbf{w}_t) - F^*$, and take the expectations of them respectively. We first calculate the lower bound of $\frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F_k(\mathbf{w}_t)]$:

$$\begin{aligned} \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F_k(\mathbf{w}_t)] &\geq \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [\langle \nabla F_k(\mathbf{w}_t), \mathbf{w}_{t,i}^k - \mathbf{w}_t \rangle] \\ &\geq -\frac{1}{2N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\eta_t \|\nabla F_k(\mathbf{w}_t)\|^2 + \frac{1}{\eta_t} \|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|^2 \right] \\ &\geq -\frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\eta_t L [F_k(\mathbf{w}_t) - F_k^*] + \frac{1}{2\eta_t} \|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|^2 \right]. \end{aligned} \quad (24)$$

Where the first inequality is by the convexity of $F_k(\cdot)$, the second inequality is by Cauchy-Schwarz inequality, and the third inequality is by the L -smoothness of $F_k(\cdot)$, $\|\nabla F_k(\mathbf{w}_t)\|^2 \leq 2L(F_k(\mathbf{w}_t) - F_k^*)$.

According to the above formula, we can obtain the bounds of $-\frac{\gamma_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F^*]$

$$\begin{aligned} &-\frac{\gamma_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F^*] \\ &\leq \frac{\gamma_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\eta_t L (F_k(\mathbf{w}_t) - F_k^*) + \frac{1}{2\eta_t} \|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|^2 - (F(\mathbf{w}_t) - F^*) \right]. \end{aligned} \quad (25)$$

For D_2 , recall the property of γ_t in Eq. (23), $0 \leq \gamma_t \leq 2\eta_t$, substituting Eq. (25) into Eq. (23), we have

$$\begin{aligned} &-\frac{\gamma_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F_k(\mathbf{w}_{t,i}^k) - F^*] + 2L\eta_t^2\tau(\tau+1)\Gamma \\ &\leq \frac{\gamma_t}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\eta_t L (F_k(\mathbf{w}_t) - F_k^*) + \frac{1}{2\eta_t} \|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|^2 - (F(\mathbf{w}_t) - F^*) \right] + 2L\eta_t^2\tau(\tau+1)\Gamma \\ &= \frac{\gamma_t(\eta_t L - 1)}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [F(\mathbf{w}_t) - F^*] + (2L\eta_t^2\tau^2 + 2L\eta_t^2\tau + \gamma_t\eta_t L\tau)\Gamma \\ &\quad + \frac{\gamma_t}{2N\eta_t} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [\|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|^2] \\ &\leq (2L\eta_t^2\tau^2 + 4L\eta_t^2\tau)\Gamma + \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} [\|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|^2]. \end{aligned} \quad (26)$$

So, for $\mathbb{E} \left[\|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \right]$, substituting Eq. (26) and Eq. (22) into Eq. (21), we have

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \right] &\leq (1 - \mu\eta_t\tau(1 - \eta_t))\mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right] \\ &\quad + (2L\eta_t^2\tau^2 + 4L\eta_t^2\tau)\Gamma + \underbrace{\frac{(2 + \mu(1 - \eta_t))}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\|\mathbf{w}_{t,i}^k - \mathbf{w}_t\|_2^2 \right]}_F. \end{aligned} \quad (27)$$

For F , according to the fact $\mathbf{w}_{t,i}^k - \mathbf{w}_t = \sum_{j=1}^i \eta_t \nabla F_k(\mathbf{w}_{t,j}^k, \xi_{t,j}^k)$, and Assumption 3.4.3, the expected squared l_2 -norm of the stochastic gradients is bounded. We have

$$\frac{(2 + \mu(1 - \eta_t))\eta_t^2}{N} \sum_{k=1}^N \sum_{i=1}^{\tau} \mathbb{E} \left[\left\| \sum_{j=1}^i \nabla F_k(\mathbf{w}_{t,j}^k, \xi_{t,j}^k) \right\|_2^2 \right] \leq (2 + \mu - \mu\eta_t)\eta_t^2 G^2 \frac{\tau(\tau + 1)(2\tau + 1)}{6}. \quad (28)$$

So, for the upper bound of $\mathbb{E} \left[\|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \right]$, due to $1 - \eta_t < 1$, substituting Eq. (28) into Eq. (27), we have

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{v}_{t+1} - \mathbf{w}^*\|_2^2 \right] &\leq (1 - \mu\eta_t\tau(1 - \eta_t))\mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right] \\ &\quad + (2 + \mu(1 - \eta_t))\eta_t^2 G^2 \frac{\tau(\tau + 1)(2\tau + 1)}{6} + (2L\eta_t^2\tau^2 + 4L\eta_t^2\tau)\Gamma \\ &\leq (1 - \mu\eta_t\tau(1 - \eta_t))\mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right] \\ &\quad + (2 + \mu)\eta_t^2 G^2 \frac{\tau(\tau + 1)(2\tau + 1)}{6} + (2L\eta_t^2\tau^2 + 4L\eta_t^2\tau)\Gamma. \end{aligned} \quad (29)$$

C.9 PROOFS OF THEOREM C.3.1

According to Lemma C.2.1 to C.2.3, and a learning rate η_t such that $0 < \eta_t \leq \min\{\frac{1}{\mu B}, \frac{1}{L(2\tau+1)}\}$, it can be concluded:

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \right] &\leq (1 - \mu\eta_t\tau(1 - \eta_t))\mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right] + \frac{2 \sum_{r=R+1}^D \sigma_r^2}{M \sum_{r=1}^D \sigma_r^2} \eta_t^2 \tau^2 G^2 \\ &\quad + (2 + \mu)\eta_t^2 G^2 \frac{\tau(\tau + 1)(2\tau + 1)}{6} + (2L\eta_t^2\tau^2 + 4L\eta_t^2\tau)\Gamma + \frac{2\eta_t^2\tau^2 G^2}{M} \\ &\leq (1 - \mu\eta_t B)\mathbb{E} \left[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \right] + \eta_t^2 C, \end{aligned} \quad (30)$$

where

$$B = \tau - \frac{\tau}{L(\tau + 1)}, C = (2 + \mu)G^2 \frac{2\tau^3 + 3\tau^2 + \tau}{6} + (2L\tau^2 + 4L\tau)\Gamma + \frac{2(1 + \rho_R)\tau^2 G^2}{M}. \quad (31)$$

C.10 PROOFS OF COROLLARIES

C.10.1 PROOF OF COROLLARY C.4.1

Assuming that Assumptions 3.4.1 to 3.4.4 hold with $\mu \geq 0$, we consider a constant learning rate η such that $0 < \eta \leq \min\{\frac{1}{\mu B}, \frac{1}{L(\tau+1)}\}$. According to Theorem C.3.1, we have

$$\mathbb{E} \left[\|\mathbf{w}_T - \mathbf{w}^*\|_2^2 \right] \leq (1 - \mu\eta B)^T \mathbb{E} \left[\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2 \right] + \sum_{t=1}^T \eta^2 (1 - \mu\eta B)^{T-t} C. \quad (32)$$

From the L -smoothness of function $F(\cdot)$, $\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \frac{L}{2} \mathbb{E} [\|\mathbf{w}_T - \mathbf{w}^*\|_2^2]$, after T global iterations, we have

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \frac{L}{2} (1 - \mu\eta B)^T \mathbb{E} [\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2] + \frac{L}{2} \sum_{t=1}^T \eta^2 (1 - \mu\eta B)^{T-t} C. \quad (33)$$

C.10.2 PROOF OF COROLLARY C.4.2

Let $\Delta_t = \mathbb{E} [\|\mathbf{w}_t - \mathbf{w}^*\|_2^2]$ and consider a diminishing learning rate, $\eta_t = \frac{\beta}{t+\gamma}$ for some $\beta > \frac{1}{\mu B}$ and $\gamma > 0$ such that $\eta_1 \leq \min\{\frac{1}{\mu B}, \frac{1}{L(\tau+1)}\} = \frac{1}{L(\tau+1)}$. We will prove $\Delta_t \leq \frac{v}{\gamma+t}$ where $v = \max\left\{\frac{\beta^2 C}{\beta\mu B - 1}, (\gamma + 1)\Delta_1\right\}$. We prove it by induction. The definition of v ensures that it holds for $t = 1$. Assuming that it also holds for Δ_t , we draw the conclusion

$$\begin{aligned} \Delta_{t+1} &\leq (1 - \eta_t \mu B) \Delta_t + \eta_t^2 C \leq \left(1 - \frac{\beta\mu B}{t+\gamma}\right) \frac{v}{t+\gamma} + \frac{\beta^2 C}{(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2} v + \left[\frac{\beta^2 C}{(t+\gamma)^2} - \frac{\beta\mu B - 1}{(t+\gamma)^2} v\right] \leq \frac{v}{t+\gamma+1}. \end{aligned} \quad (34)$$

Then by the L -smoothness of $F(\cdot)$, we have $\mathbb{E}[F(\mathbf{w}_t)] - F^* \leq \frac{L}{2} \Delta_t \leq \frac{L}{2} \frac{v}{\gamma+t}$. We choose $\beta = \frac{2}{\mu B}$, $\gamma = \frac{2L^2(\tau+1)^2}{\mu\tau(L(\tau+1)-1)} - 1$, and denote $\kappa = \frac{L}{\mu}$. Therefore, η_t can be further expressed as $\eta_t = \frac{2}{\mu B(\gamma+t)}$. we have

$$v = \max\left\{\frac{\beta^2 C}{\beta\mu B - 1}, (\gamma + 1)\Delta_1\right\} \leq \frac{\beta^2 C}{\beta\mu B - 1} + (\gamma + 1)\Delta_1 = \frac{4C}{\mu^2 B^2} + (\gamma + 1)\Delta_1, \quad (35)$$

and

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \frac{L}{2} \frac{v}{\gamma+t} \leq \frac{\kappa}{\gamma+t} \left(\frac{2C}{\mu B^2} + \frac{\mu(\gamma+1)}{2} \Delta_1\right). \quad (36)$$

D HYPERPARAMETERS USED IN BASELINE ALGORITHMS

Besides the hyperparameter setting provided in the main body, the other hyperparameters are as follows: For FedProx, we set $\mu = 0.01$; for MOON, we set $\tau = 1, \mu = 0.01$; for FedGen, the server epoch is 1000 and the generator learning rate is 0.005; for FedDC, we set $\alpha = 0.5$; for FedDyn, we set $\alpha = 0.5$; for FedNTD, we set $\beta = 0.001, \tau = 1$. For communication-efficient algorithms, we set $\delta = 0.05$ in LBGm; for signSGD and STC, we set their compression ratios as 1/32. Besides, We use the SGD optimizer in all experiments with momentum set to 0.

E FURTHER DISCUSSION

We found that the global model space of federated learning has low rank properties. In fact, due to the scarcity of client data, federated learning algorithms face the risk of overfitting. By restricting the local model to a low dimensional subspace, the degree of freedom in model updates is reduced. This can be used as a basis for many federated learning algorithms to improve their generalization capabilities.

In addition, since the global model update of federated learning can be represented with fewer orthogonal bases, FLSS can also be widely integrated as a compression strategy into various compression frameworks to further reduce the compression rate. Compared with traditional compression schemes, FLSS pays more attention to the distribution of model parameters or gradient space. FLSS can adaptively select appropriate orthogonal bases to represent model updates for different networks and different scenarios. In other words, the compression of FLSS is data-driven and task-relevant.

In fact, many algorithms address the heterogeneity problem by considering local and global consistency. For instance, model parameter consistency is tackled by FedProx, representation consistency by Moon, and logit consistency by FedNTD. In contrast to these approaches, our method emphasizes

594 the directional consistency between global and local updates, constraining the update direction by
595 applying a projection to limit the angle.
596

597 REFERENCES

598
599 Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough,
600 and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint*
601 *arXiv:2111.04263*, 2021.
602

603 Sheikh Shams Azam, Seyyedali Hosseinalipour, Qiang Qiu, and Christopher Brinton. Recycling
604 model updates in federated learning: Are gradient subspaces low-rank? In *International Conference*
605 *on Learning Representations*, 2021.

606 Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated
607 learning with non-iid data via local drift decoupling and correction. In *Proceedings of the*
608 *IEEE/CVF conference on computer vision and pattern recognition*, pp. 10112–10121, 2022.
609

610 Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyang Wu, Terrence Chen, David Doermann,
611 and Arun Innanje. Preserving privacy in federated learning with ensemble cross-domain knowledge
612 distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp.
613 11891–11899, 2022.

614 Frithjof Gressmann, Zach Eaton-Rosen, and Carlo Luschi. Improving neural network training in low
615 dimensional random bases. *Advances in Neural Information Processing Systems*, 33:12140–12150,
616 2020.
617

618 Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv*
619 *preprint arXiv:1812.04754*, 2018.

620 Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated
621 learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
622 pp. 10143–10153, 2022.
623

624 Meirui Jiang, Zirui Wang, and Qi Dou. Harmofi: Harmonizing local and global drifts in federated
625 learning on heterogeneous medical images. In *Proceedings of the AAAI Conference on Artificial*
626 *Intelligence*, volume 36, pp. 1087–1095, 2022.

627 Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback
628 fixes signsgd and other gradient compression schemes. In *International Conference on Machine*
629 *Learning*, pp. 3252–3261. PMLR, 2019.
630

631 Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and
632 Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In
633 *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.
634

635 Jinkyu Kim, Geeho Kim, and Bohyung Han. Multi-level branched regularization for federated
636 learning. In *International Conference on Machine Learning*, pp. 11058–11073. PMLR, 2022.

637 Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. Preservation of the
638 global knowledge by not-true distillation in federated learning. *Advances in Neural Information*
639 *Processing Systems*, 35:38461–38474, 2022.
640

641 Chunyuan Li, Heerad Farkhor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension
642 of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

643 Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of*
644 *the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021.
645

646 Tao Li, Lei Tan, Zhehao Huang, Qinghua Tao, Yipeng Liu, and Xiaolin Huang. Low dimensional
647 trajectory hypothesis is true: Dnns can be trained in tiny subspaces. *IEEE Transactions on Pattern*
Analysis and Machine Intelligence, 45(3):3411–3420, 2022a.

- 648 Tao Li, Yingwen Wu, Sizhe Chen, Kun Fang, and Xiaolin Huang. Subspace adversarial training.
649 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
650 13409–13418, 2022b.
- 651 Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith.
652 Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*,
653 2:429–450, 2020.
- 654 Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. Deep gradient compression: Reducing
655 the communication bandwidth for distributed training. In *International Conference on Learning*
656 *Representations*, 2018.
- 657 Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas.
658 Communication-efficient learning of deep networks from decentralized data. In *Artificial intelli-*
659 *gence and statistics*, pp. 1273–1282. PMLR, 2017.
- 660 Yifan Niu and Weihong Deng. Federated learning for face recognition with gradient correction. In
661 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1999–2007, 2022.
- 662 Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman,
663 Joseph Gonzalez, and Raman Arora. Fetchsgd: Communication-efficient federated learning with
664 sketching. In *International Conference on Machine Learning*, pp. 8253–8265. PMLR, 2020.
- 665 Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and
666 communication-efficient federated learning from non-iid data. *IEEE transactions on neural*
667 *networks and learning systems*, 31(9):3400–3413, 2019.
- 668 Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Quasi-newton’s method in the class
669 gradient defined high-curvature subspace. *arXiv preprint arXiv:2012.01938*, 2020.
- 670 Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine*
671 *learning research*, 9(11), 2008.
- 672 Oriol Vinyals and Daniel Povey. Krylov subspace descent for deep learning. In *Artificial intelligence*
673 *and statistics*, pp. 1261–1268. PMLR, 2012.
- 674 Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, Jian Cao, and Haibing
675 Guan. Gpfl: Simultaneously learning global and personalized feature information for personalized
676 federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*,
677 pp. 5041–5051, 2023.
- 678 Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous
679 federated learning. In *International conference on machine learning*, pp. 12878–12889. PMLR,
680 2021.
- 681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701