CONVERGENCE OF SPLIT FEDERATED LEARNING VIA A MASKED COMPOSITIONAL OPTIMIZATION LENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Split federated learning (SFL) trains a split neural network across distributed clients by keeping initial layers on-device and offloading the remaining layers to a server. The split point is called the cut layer, where clients transmit intermediate activations to the server and receive backpropagated gradients. While SFL reduces client compute and memory, the resulting client–server coupling makes its optimization behavior poorly understood. This paper develops, to our knowledge, the first *compositional* convergence analysis of SFL. We show that SFL optimizes a single-expectation compositional objective, i.e., a client-side forward map produces cut-layer representations, and a server-side map completes the network and evaluates the loss. Unlike existing federated compositional optimization where the main challenge is tracking an inner expectation, SFL's inner map is a per-minibatch network transform, so the key challenge is controlling the split-induced coupling. We address the challenge and establish nonconvex convergence results for two major SFL variants. For the parallel variant SFL-V1, after T communication rounds, we obtain an $O(1/\sqrt{T})$ rate. For the sequential variant SFL-V2, we obtain an $O(1/\sqrt{T})$ rate with a leading constant that depends on K^2 and split-dependent quantities, where K is the number of clients. Enabled by the compositional formulation, our theory reveals insights on SFL design choices, including (i) how the cut layer affects convergence, and (ii) how SFL-V1 and SFL-V2 scale differently with K. Experiments on regression and classification tasks corroborate our theory.

1 INTRODUCTION

Federated learning (FL) enables clients to train a shared model without sharing raw data McMahan et al. (2017). However, as models become larger, full on-device training can exceed the compute and memory limits of resource-constrained devices Bai et al. (2024). Split learning (SL) alleviates this issue by *splitting* a model at a cut layer Gupta and Raskar (2018). Clients compute early layers, send intermediate activations to the server, and receive backpropagated gradients. Yet, SL typically trains clients sequentially, which limits parallelism and increases training latency Tirana et al. (2024).

Split federated learning (SFL) combines FL's parallel client participation with SL's model splitting. This hybrid design reduces client resource costs while retaining federated scalability Thapa et al. (2022). There are two main variants of SFL. SFL-V1 performs split-model updates in parallel across clients followed by periodic synchronization. SFL-V2 updates the server-side model sequentially across clients, which creates dependencies through the split interface. Some initial works have studied the convergence of SFL, e.g., Han et al. (2024); Lin et al. (2025). These analyses highlight that SFL can exhibit *dual-paced* optimization behavior between clients and server, together with nontrivial interactions with partial participation and data heterogeneity.

In this paper, we propose a complementary viewpoint: **SFL can be understood as a compositional optimization problem with a distinctive non-nested structure.** At a high level, consider a model parameterized by x split at a cut layer into a client-side mapping and a server-side mapping. Let x_c and x_s denote the client-side and server-side parameters, respectively. For client k with sample ξ, the client's forward pass produces the cut-layer representation $u = g^{(k)}(x_c; \xi)$. The server evaluates the loss $f^{(k)}(u, x_s)$ using u and x_s, and it returns the gradient with respect to u for the client to update

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

x_c. This induces the global objective across K clients:

$$\min_{x_c, x_s} \; F(x_c, x_s) := \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\xi} \Big[f^{(k)}\big(g^{(k)}(x_c; \xi),\, x_s\big) \Big] . \tag{1}$$

Technical challenges. It is worth noting that Eq. 1 is a *single-expectation* compositional objective, i.e., the same sample ξ generates both the inner representation and the outer loss. *This is different from existing federated compositional optimization formulations*, which typically involve *nested* expectations such as

$$\min_{x} \; \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\epsilon} \Big[\phi^{(k)}\Big(\mathbb{E}_{\gamma}[\, \psi^{(k)}(x; \gamma)\,],\, \epsilon \Big) \Big] , \tag{2}$$

where $\psi^{(k)}(x; \gamma)$ is an *inner* mapping whose expectation $\mathbb{E}_{\gamma}[\psi^{(k)}(x; \gamma)]$ needs to be estimated, and $\phi^{(k)}(\cdot, \epsilon)$ is an *outer* loss that evaluates this inner expectation under randomness ϵ at client k Zhang et al. (2023; 2024). Such nested structures typically require explicitly tracking the inner expectation via two-timescale updates or variance reduction Wu et al. (2023). In contrast, SFL's inner map $g^{(k)}(x_c; \xi)$ is a per-sample forward pass, and it does *not* average over an inner randomness before the outer loss is evaluated. Hence, techniques designed for Eq. 2 are not directly applicable.

Why does the compositional formulation matter? This perspective makes the cut layer an explicit part of the convergence analysis. In SFL, the cut layer determines the representations sent to the server and the gradients returned to clients. Our formulation captures their properties through smoothness and gradient-moment constants, which allows us to examine how the cut layer affects the convergence results. It also reveals how these constants interact with the number of clients under the different update structures of SFL-V1 and SFL-V2. These results provide a theoretical perspective on two practical design choices: (i) how to choose the cut layer and (ii) which SFL variant to use.

Our main contributions are summarized as follows.

- **Single-expectation compositional formulation for SFL (Sec. 3).** To our best knowledge, this is the first work that casts SFL as a *single-expectation* compositional problem in Eq. 1. Unlike nested federated compositional objectives, the main challenge for SFL is controlling the split-induced coupling between the client-side and server-side subnetworks.
- **Convergence analysis for two major SFL variants (Sec. 4).** Under mild assumptions, we establish nonconvex stationary-point guarantees for two major SFL variants. For the parallel variant (SFL-V1), we obtain an $O(1/\sqrt{T})$ rate after T communication rounds. For the sequential variant (SFL-V2), we obtain an $O(1/\sqrt{T})$ rate with a leading constant that depends on K^2 and split-dependent quantities, where K is the number of clients. Our analysis captures client drift induced by synchronization in both variants and the additional effect of sequential server updates in SFL-V2.
- **Experiments and insights (Sec. 5).** We conduct experiments on regression and classification tasks. The experiments and our theory provide a few insights. For example, both variants generally perform better with later cut layers, and SFL-V2 is more sensitive to the cut layer choice. Moreover, SFL-V1 can outperform SFL-V2 when the number of clients is large.

The rest of the paper is organized as follows. Sec. 2 reviews related work. Sec. 3 provides our compositional formulation for two major variants of SFL. Sec. 4 presents the convergence analysis. Sec. 5 provides empirical validation. Sec. 6 discusses the limitations and concludes the paper.

2 Related Work

Theory for SL and SFL. While most studies on SL/SFL are empirical, there is a recent increase of analytical efforts on SL/SFL convergence. Li and Lyu (2023) provides the first analysis for SL and shows an $O(1/\sqrt{T})$ rate for nonconvex objectives. Their theory also implies that sequential client participation helps mitigate the client drift issue compared with parallel averaging methods, e.g., FedAvg Li and Lyu (2025). Han et al. (2024) gives the first convergence analysis for SFL, yet they did not either characterize the impact of cut layer or the variant-level difference in their theory. Lin et al.

Table 1: Representative convergence results for SL and SFL. In "Assumptions", $\mathcal{S}$ refers to smoothness of the local/global losses; $\mathcal{V}$ refers to a bounded variance of the stochastic gradients; $\mathcal{G}$ refers to a bounded gradient; and $\mathcal{H}$ refers to a bounded heterogeneity between local and global losses. "**Cut-layer impact**" indicates whether the theory explicitly characterizes how the cut layer affects training dynamics, where "Dimension (ZO)" means that the cut layer appears through the client-side parameter dimension in zeroth-order gradient estimation. "**Variant-level distinction**" indicates whether the convergence results clarify the structural difference of SFL-variants, e.g., SFL-V1 and SFL-V2. A more detailed rate comparison involving other factors is given in Appendix G.

Work	Setting	Assumptions	Convergence rate	Cut-layer impact	Variant-level distinction
Li and Lyu (2023; 2025)	Vanilla SL	$\mathcal{S}, \mathcal{V}, \mathcal{H}$	$O(1/\sqrt{T})$	×	–
Han et al. (2024)	SFL	$\mathcal{S}, \mathcal{V}, \mathcal{G}, \mathcal{H}$	$O(1/\sqrt{T})$	×	×
Lin et al. (2025)	SFL (multi-tier)	$\mathcal{S}, \mathcal{V}, \mathcal{G}$	$O(1/\sqrt{T})$	×	×
Nair et al. (2025)	SFL (w. auxiliary models)	$\mathcal{S}, \mathcal{V}, \mathcal{H}$	$O(1/\sqrt{T})$	×	×
Liang et al. (2025)	SFL (zeroth-order)	$\mathcal{S}, \mathcal{V}, \mathcal{H}$	$O(1/\sqrt{T})$	Dimension (ZO)	×
Kou et al. (2026)	SFL (hybrid-order)	$\mathcal{S}, \mathcal{V}, \mathcal{G}$	$O(1/\sqrt{T})$	×	×
This work	SFL	$\mathcal{S}, \mathcal{V}, \mathcal{G}$	$O(1/\sqrt{T})$	✓	✓

(2025) proposes a hierarchical, multi-layer SFL variant together with its convergence theory. Recent works also study beyond-backpropagation variants of SFL. For example, Nair et al. (2025) uses auxiliary models to accelerate SFL. Liang et al. (2025) and Kou et al. (2026) develop zeroth-order and hybrid-order approaches for SFL.

Table 1 summarizes representative convergence results on SL/SFL. Our work complements this literature through a masked compositional analysis of SFL. Our formulation enables characterizing the cut-layer impact through split-dependent quantities. Also, our theory distinguishes the structural difference between SFL-V1 and SFL-V2, which provides insights on algorithm choice.

Theory for stochastic compositional optimization Stochastic compositional optimization typically studies nested objectives, e.g., $\min_x \mathbb{E}_\zeta[f(\mathbb{E}_\xi[g(x;\xi)];\zeta)]$, and their multi-level generalizations Gao (2024). The main difficulty is that the outer gradient depends on an inner expectation, which generally makes plug-in stochastic gradient estimates biased and motivates two-timescale updates or recursive estimators. This challenge becomes more involved in distributed/federated settings, where inner-expectation tracking must be combined with consensus or gradient-tracking mechanisms. Existing work develops distributed and local stochastic compositional methods (Gao and Huang, 2021; Gao et al., 2022), doubly-recursive estimators (Gao, 2024), and federated compositional minimax algorithms with communication-dependent guarantees (Tarzanagh et al., 2022; Zhang et al., 2023; 2024). *Yet these studies focus on nested structures with an inner expectation, whereas SFL has a non-nested structure*.

Related non-nested compositional formulations also arise in vertical federated learning (VFL), e.g., Castiglia et al. (2022); Valdeira et al. (2025). VFL combines representations of complementary feature partitions for overlapping samples. *However, the SFL setting studied here maintains local copies of a common client-side model across horizontally partitioned data.* Our analysis needs to account for periodic averaging, in SFL-V2, the sequential handoff of a shared server-side model.

3 Formulation

3.1 Masked Compositional Formulation

Consider K clients indexed by $k \in \{1, \dots, K\}$. We encode the split of a neural network at a cut layer using a *masking* vector $\omega \in \{0,1\}^n$, applied entrywise to a global parameter vector $x \in \mathbb{R}^n$. In particular, $x_c := \omega \odot x$ corresponds to the client-side parameters, and $x_s := (\mathbf{1} - \omega) \odot x$ corresponds to the server-side parameters. For each client k, let $g^{(k)}(\cdot;\xi)$ denote the client-side forward mapping (the *inner map*) producing the cut-layer representation under minibatch randomness ξ. Let $f^{(k)}(\cdot,\cdot;\zeta)$ denote the server-side mapping (the *outer map*) that takes the cut-layer representations and the server-side parameters as inputs. Here, ζ denotes the server-side randomness induced by the same minibatch ξ. The expectation below is taken over their joint distribution for client k. We also define

$g^{(k)}(\cdot) := \mathbb{E}_\xi[g^{(k)}(\cdot;\xi)]$. Each client k's loss is

$$F^{(k)}(x) := \mathbb{E}_{\xi,\zeta}\big[f^{(k)}(g^{(k)}(x\odot\omega;\xi), x\odot(\mathbf{1}-\omega);\zeta)\big]. \tag{3}$$

Let $F(x) := \frac{1}{K}\sum_{k=1}^K F^{(k)}(x)$ be the global loss. SFL induces the *masked compositional* objective:

$$\min_{x\in\mathbb{R}^n} F(x) := \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{\xi,\zeta}\Big[f^{(k)}\Big(g^{(k)}(x\odot\omega;\xi), x\odot(\mathbf{1}-\omega);\zeta\Big)\Big]. \tag{4}$$

Importantly, although we write the expectation over (ξ,ζ), Eq. 4 is still a *single-expectation* compositional objective. The same minibatch ξ generates both the inner representation and the outer loss. This is structurally different from nested-expectation federated compositional objectives that require tracking an inner expectation Tarzanagh et al. (2022); Gao (2024), e.g., see Eq. 2.

Masked gradient calculation. For any client k and a realized (ξ,ζ), we define the *stochastic* loss:

$$F^{(k)}(x;\xi,\zeta) \;:=\; f^{(k)}\Big(g^{(k)}(x\odot\omega;\xi),\; x\odot(\mathbf{1}-\omega);\;\zeta\Big). \tag{5}$$

Applying the chain rule to the masked composition yields the per-sample gradient estimator:

$$\begin{aligned}\nabla_x F^{(k)}(x;\xi,\zeta) = \omega\odot\Big[\nabla g^{(k)}(x\odot\omega;\xi)\Big]^\top \nabla_g f^{(k)}\Big(g^{(k)}(x\odot\omega;\xi),\; x\odot(\mathbf{1}-\omega);\;\zeta\Big)\\ +\;(\mathbf{1}-\omega)\odot\nabla_{x_s} f^{(k)}\Big(g^{(k)}(x\odot\omega;\xi),\; x\odot(\mathbf{1}-\omega);\;\zeta\Big).\end{aligned} \tag{6}$$

Here, the first term corresponds to backpropagating the cut-layer gradient through the client-side mapping $g^{(k)}$ and is masked to the client-side model by ω, while the second term updates the server-side model masked by $\mathbf{1}-\omega$. This expression makes explicit the *model split-induced coupling*: though the objective in Eq. 4 is a single-expectation composition, the gradients couple the client-side and server-side models through shared $g^{(k)}$ at the cut layer and its associated gradients.

3.2 Algorithmic Formulation of Two SFL Variants

We now formalize two major SFL variants (SFL-V1 and SFL-V2 Thapa et al. (2022)) using our masked compositional formulation. At a high level,

- SFL-V1 executes client-side model training in *parallel*. Each client interacts with its own server-side model and performs local updates for p steps, followed by periodic averaging.
- SFL-V2 introduces *sequential* within-iteration coupling. Specifically, the server-side computation for client k is driven by a mixed iterate that combines the server-side part just updated by client $k-1$ and the current client-side part of client k, forming a ring-like dependency across clients. The client-side models will also be averaged every p steps.

Next, we introduce SFL-V1 and SFL-V2 in detail.

SFL-V1 (parallel local updates with periodic averaging). The pseudocode is in Algorithm 1 in Appendix B. SFL-V1 is a parallel approach, i.e., each client k together with a corresponding server-side model performs a split-SGD update on its current local iterate $x_t^{(k)}$, and all clients synchronize every p steps. For ease of presentation, define $h(x) \;:=\; (\mathbf{1}-\omega)\odot x, \quad z^{(k)} \;:=\; \nabla_x F^{(k)}\big(x;\,\xi,\zeta\big)$. At iteration t, client k and the server computes $u_{t+1}^{(k)} = g^{(k)}(\omega\odot x_t^{(k)};\xi)$, $s_{t+1}^{(k)} = h(x_t^{(k)})$, forms the stochastic gradient $z_t^{(k)}$ via Eq. 6, and performs model updates $x_{t+1}^{(k)} = x_t^{(k)} - \eta\, z_t^{(k)}$, where η is the learning rate. Every p steps, the clients and the server average their models.

SFL-V2 (sequential coupling within an iteration). The pseudocode is in Algorithm 2 in Appendix B. SFL-V2 introduces an additional structural dependency across clients within each iteration. Instead of feeding each client its own current state, SFL-V2 processes the clients in a sequential

order and forms a *mixed* iterate $y_t^{(k)}$ that blends (i) the client-side model of client k, $x_t^{(k)}$, and (ii) the server-side model as it was just left by the previous client, $x_{t+1}^{(k-1)}$. To this end, define

$$y_t^{(k)} \;=\; \omega \odot x_t^{(k)} \;+\; (\mathbf{1}-\omega)\odot x_{t+1}^{(k-1)}, \qquad k=1,\dots,K, \tag{7}$$

with conventions $x_{t+1}^{(0)} := x_t^{(K)}$ and $x_0^{(K)} := x_0$, so that client 1 inherits the server-side model left by client K at the end of the previous iteration. Note that the index $t+1$ in Eq. 7 is deliberate, i.e., within iteration t the clients act in order, so client $k-1$ has produced $x_{t+1}^{(k-1)}$ when client k is served. Then, we run the update using $y_t^{(k)}$ in place of $x_t^{(k)}$. That is, $u_{t+1}^{(k)} = g^{(k)}(\omega \odot y_t^{(k)}; \xi)$, $s_{t+1}^{(k)} = h(y_t^{(k)})$,

$$x_{t+1}^{(k)} \;=\; y_t^{(k)} - \eta\, z_t^{(k)}, \tag{8}$$

where $z_t^{(k)}$ is given by Eq. 6 with x replaced by $y_t^{(k)}$. Equivalently, by the mask structure, Eq. 8 reads $\omega \odot x_{t+1}^{(k)} = \omega \odot x_t^{(k)} - \eta\,\omega \odot z_t^{(k)}$ and $(\mathbf{1}-\omega)\odot x_{t+1}^{(k)} = (\mathbf{1}-\omega)\odot x_{t+1}^{(k-1)} - \eta\,(\mathbf{1}-\omega)\odot z_t^{(k)}$. Each client-side model takes one step per iteration from its own iterate, whereas the single server-side model is handed off from client to client and hence takes K steps per iteration. Similar to SFL-V1, SFL-V2 performs periodic client-side averaging every p steps. Note that there is no server-side averaging as the server only maintains one single server-side model. That is, the server-side model is effectively passed along sequentially via $y_t^{(k)}$ rather than aggregated.

Technical challenges induced by SFL-V1 & SFL-V2. While both variants optimize the same masked compositional objective in Eq. 4, their distinct update structures (i.e., parallel and sequential) lead to different stochastic trajectories. In SFL-V1, the main difficulty is *client drift* between local iterates $x_t^{(k)}$ and their average $\bar{x}_t$ under synchronization. In SFL-V2, besides the client drift issue, the sequential/ring construction in Eq. 7 creates a *temporal dependency chain* across clients within each iteration, because each $y_t^{(k)}$ mixes blocks from $(x_{t-1}^{(k)}, x_t^{(k-1)})$. This coupling requires the analysis to control how errors propagate sequentially through k, which leads to a more intricate recursion.

4 Convergence Analysis

4.1 Assumptions

We start with a few assumptions that are common in the distributed compositional optimization literature. We focus on *non-convex* objectives as SFL typically involves splitting neural networks.

Assumption 4.1 (Smoothness)**.** For any client k, the function $g^{(k)}(\omega \odot \cdot\,; \xi)$ is $L_g(\omega)$-Lipschitz smooth and $f^{(k)}(\cdot\,;\zeta)$ is L_f-Lipschitz smooth. That is, for any x_1, x_2, y_1, y_2, $\|\nabla g^{(k)}(\omega \odot x_1; \xi) - \nabla g^{(k)}(\omega \odot x_2; \xi)\| \le L_g(\omega)\|x_1 - x_2\|$, $\|\nabla f^{(k)}(y_1;\zeta) - \nabla f^{(k)}(y_2;\zeta)\| \le L_f\|y_1 - y_2\|$.

Here, $L_g(\omega)$ quantifies the sensitivity of the *client-side* transformation restricted to the masked coordinates, and it depends explicitly on the split mask ω (i.e., the cut layer).

Assumption 4.2. (Bounded gradients) For any client k, the client-side Jacobian and the second moment of the server-side gradient are bounded. That is, for any x, ξ, $\big\|\nabla g^{(k)}(\omega \odot x; \xi)\big\| \le C_g(\omega), \mathbb{E}_\zeta\Big[\sup_y \big\|\nabla f^{(k)}(y;\zeta)\big\|^2\Big] \le C_f^2$.

Similar assumptions are common in the compositional optimization literature Gao and Huang (2021); Zhang et al. (2023); Gao et al. (2022).

Assumption 4.3. (Bounded variance) Stochastic gradients are unbiased. Also, there exist σ_f, $\sigma_g(\omega)$, $\delta_g(\omega) > 0$ such that for any x, $\mathbb{E}_\zeta\Big[\sup_y \big\|\nabla f^{(k)}(y;\zeta) - \nabla f^{(k)}(y)\big\|^2\Big] \le \sigma_f^2$, $\mathbb{E}_\xi\big\|\nabla g^{(k)}(\omega \odot x;\xi) - \nabla g^{(k)}(\omega \odot x)\big\|^2 \le \sigma_g^2(\omega)$, $\mathbb{E}_\xi\big\|g^{(k)}(\omega \odot x;\xi) - g^{(k)}(\omega \odot x)\big\|^2 \le \delta_g^2(\omega)$.

Here, σ_f and $\sigma_g(\omega)$ bound the stochastic gradient variance, while $\delta_g(\omega)$ bounds the variance in the *smashed data* at the cut layer.

Lemma 4.4. *Under Assumptions 4.1-4.2, the global objective F is L_F-smooth, where*

$$L_F \;=\; C_f L_g(\omega) \;+\; L_f\big(C_g(\omega)+1\big)^2. \tag{9}$$

The proof is in Appendix C. This lemma shows how ω affects the geometry of F. We further define

$$\gamma_g^2(\omega) = K\big[12C_g^2(\omega) + 6\sigma_g^2(\omega)\big], \quad \gamma_f^2 = K\big(12C_f^2 + 6\sigma_f^2\big),$$
$$\sigma_F^2(\omega) = 5C_g^2(\omega)\sigma_f^2 + 5\big(C_g^2(\omega)+1\big)L_f^2\delta_g^2(\omega) + 5C_f^2\sigma_g^2(\omega) + 5\sigma_f^2. \tag{10}$$

4.2 Convergence for SFL-V1

Proposition 4.5. *(SFL-V1: Bounding Client Drift) Let Assumptions 4.1-4.3 hold. If $\eta^2 \le c_1(\omega)$, with $c_1(\omega)$ being a constant, then*

$$\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\left\|\bar{x}_t - x_t^{(k)}\right\|^2 \le 42TKp^2\eta^2(2(C_g^2(\omega)+\sigma_g^2(\omega))+1)\left(\gamma_f^2 + 10L_f^2\gamma_g^2(\omega) + 20KL_f^2\sigma_g^2(\omega)\right)$$
$$+ 140TKp^2\eta^2(C_f^2+\sigma_f^2)\left(2K\sigma_g^2(\omega)+\gamma_g^2(\omega)\right) + 28TK^2p^2\eta^2\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]. \tag{11}$$

The proof is in Appendix D. Proposition 4.5 bounds the accumulated disagreement between local models and their average. Now we are ready to provide the convergence result below.

Theorem 4.6. *(SFL-V1 Convergence) Let Assumptions 4.1-4.3 hold and let $F^* := \inf_x F(x) > -\infty$. If $\eta \leqslant \kappa$, where κ is some constant, then SFL-V1 converges as*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F\left(\bar{x}_t\right)\right\|^2 \le \frac{2(F(\bar{x}_0)-F^*)}{\eta T} + \eta L_F\sigma_F^2(\omega) + 140L_F^2p^2\eta^2(C_f^2+\sigma_f^2)\left(2K\sigma_g^2(\omega)+\gamma_g^2(\omega)\right)$$
$$+ 28KL_F^2p^2\eta^2\sigma_f^2\left(2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right)$$
$$+ 42L_F^2p^2\eta^2\left(2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right)\left(\gamma_f^2 + 10L_f^2\gamma_g^2(\omega) + 20KL_f^2\sigma_g^2(\omega)\right). \tag{12}$$

In addition, if we set $\eta = O(1/\sqrt{T})$, then SFL-V1 converges with an $O(1/\sqrt{T})$ rate.

The proof is in Appendix E. Theorem 4.6 shows that with a stepsize $\eta = O(1/\sqrt{T})$, SFL-V1 converges at an $O(1/\sqrt{T})$ rate. Moreover, we discuss some insights below.

Cut-layer impact: Eq. 12 shows how the compositional formulation connects cut layer to convergence. It determines the client-side representation and the Jacobian through which gradients are backpropagated. In particular, a cut layer with smaller representation and Jacobian variability (i.e., $\delta_g^2(\omega), \sigma_g^2(\omega)$) lead to a tighter convergence. This is also consistent with experiments in Sec. H.5.

Synchronization period: The compositional analysis reveals how cut-layer selection interacts with periodic synchronization. In Eq. 12, the terms proportional to $p^2\eta^2$ capture disagreement accumulated between local models, with coefficients affected by ω. In particular, the synchronization cost grows quadratically with p and is amplified by cut-layer variability. Hence, a more stable cut layer makes training less sensitive to η and allows larger p with less degradation.

4.3 Convergence for SFL-V2

We now consider SFL-V2, where the intermediate mixed iterate $y_t^{(k)}$ introduces sequential coupling across clients within each iteration.

The main technical difficulty for SFL-V2 analysis is that the *mixed iterate* $y_t^{(k)}$ creates an intra-iteration *sequential dependency chain*. That is, the server-side model used by client k is inherited from client $k-1$. So the stochastic gradient of client k depends on the samples of clients $1,\dots,k-1$ in the same iteration, and the server-side blocks stored in $x_t^{(1)},\dots,x_t^{(K)}$ sit at different stages of the chain. As a result, the average $\bar{x}_t$ is not the model that is synchronized, and one cannot use the proof pipeline as from Proposition 4.5 to Theorem 4.6. Instead, we track the synchronized model $\hat{x}_t := \omega \odot \bar{x}_t + (\mathbf{1}-\omega)\odot x_t^{(K)}$. It follows an exact *block-scaled* recursion: the client-side block moves by the averaged gradient, while the server-side block moves by the *sum* of the K sequential gradients. We then control two coupling errors, i.e., the client drift between synchronizations and the server-side lag of client k, which grows with the position k of the client along the chain. Both are plugged into the smoothness argument to bound the terms induced by sequential coupling.

Theorem 4.7. *(SFL-V2 Convergence). Let Assumptions 4.1-4.3 hold and let* $F^* := \inf_x F(x) > -\infty$. *Let* $\bar{x}_t := \frac{1}{K}\sum_{k=1}^{K} x_t^{(k)}$, *and let* $\hat{x}_t := \omega \odot \bar{x}_t + (\mathbf{1} - \omega) \odot x_t^{(K)}$ *denote the global model after iteration* t. *Then, SFL-V2 converges as*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\hat{x}_t)\|^2 \leqslant \frac{2\,(F(x_0) - F^*)}{\eta\, T} + \eta L_F C_f^2\big(C_g^2(\omega) + K^2\big) + \eta^2 L_F^2 C_f^2\Big(p^2 K\, C_g^2(\omega) + \frac{K^3}{3}\Big). \tag{13}$$

In addition, if we set $\eta = O(1/\sqrt{T})$, *then SFL-V2 converges as* $O(1/\sqrt{T})$, *whose leading constant scales with* K^2. *If* η *is constant, the second term in Eq. 13 is a residual that scales with* K^2.

The proof is in Appendix F. We discuss some insights below.

Structural difference to SFL-V1: SFL-V2 has a stationarity guarantee similar in form to SFL-V1, but its dependence on K is different. More specifically, the $O(\eta)$ term in Eq. 13 has the factor K^2 on the server-side block, whereas the corresponding term of SFL-V1, $\eta L_F \sigma_F^2(\omega)$ in Eq. 12, does not grow with K. When η is kept constant, as in our experiments, this term does not vanish with T and determines the residual. The residual of SFL-V2 scales with K^2, while that of SFL-V1 does not. With $\eta = O(1/\sqrt{T})$, both variants converge at an $O(1/\sqrt{T})$ rate, but the leading constant of SFL-V2 is of order K^2. *To our knowledge, this is the first theoretical characterization that distinguishes the convergence behavior of SFL-V1 and SFL-V2, thanks to the compositional formulation.* It also provides guidance for algorithm choice. That is, when the number of clients is large, SFL-V1 can be preferable to SFL-V2. This is consistent with our empirical results in Sec. 5.

Cut-layer impact: The cut layer ω also affects SFL-V2 through the mask-dependent constants in Eq. 13. Different from SFL-V1, these split-dependent constants are multiplied by K^2 in the $O(\eta)$ term, and $L_F^2(\omega)$ is multiplied by K^3 in the $O(\eta^2)$ term, because the sequential handoff amplifies the server-side block. Hence, the same change of the cut layer changes the SFL-V2 bound by a factor that grows with K. Hence, SFL-V2 is expected to be more sensitive to the cut layer than SFL-V1. Our experiment results in Sec. 5 are consistent with this implication.

Summary of theory-guided implications. Our compositional analysis provides three practical guidelines. First, cut layers with smaller split-induced representation/gradient variability improve the constants in both SFL-V1 and SFL-V2. Second, for a large client number K, SFL-V1 is preferable compared to SFL-V2. Third, a cut layer that produces stable smashed representations and gradients makes training more tolerant of infrequent synchronization, so a larger p can be used to save communication with less degradation.

5 Numerical Results

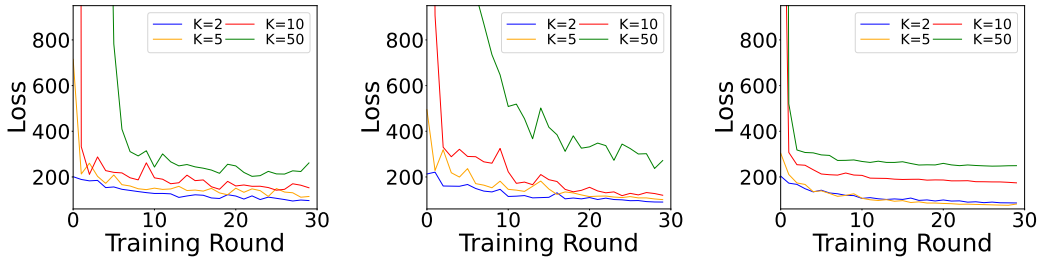
Regression. We use the YearPredictionMSD dataset Bertin-Mahieux (2011) and apply a three-layer MLP, which applies three fully-connected layers with ReLU activations. We consider $L_c \in \{1, 2, 3\}$ with $L_c = i$ meaning splitting the model after each layer. We consider $K \in \{2, 5, 10, 50\}$, and in each setting the data is randomly assigned to clients.

Classification. We use CIFAR-100 Krizhevsky (2009). We use ResNet-18 He et al. (2016) as the model structure and consider $L_c \in \{1, 2, 3, 4\}$, where $L_c = i$ means the model is split after the i-th residual block. We consider $K \in \{2, 5, 10, 50\}$ and study a non-IID scenario using a Dirichlet distribution with heterogeneity $\alpha = 0.1$.

We consider three algorithms, i.e., SFL-V1, SFL-V2 Thapa et al. (2022), and vanilla split learning (SL) Gupta and Raskar (2018). We do not consider FL as it is logically similar to SFL-V1. The details of hyper-parameters are left to Appendix H.1.

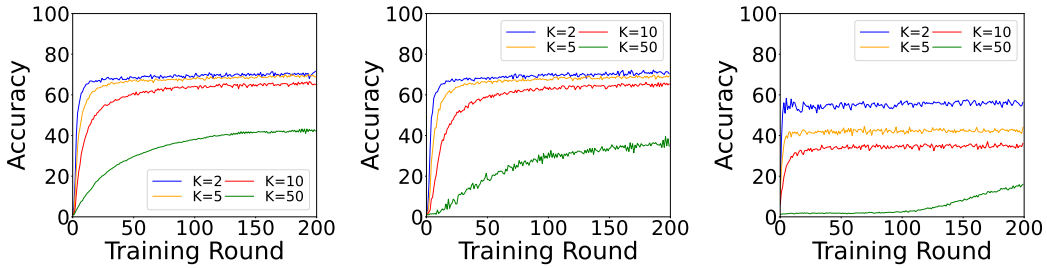
5.1 Experiment Results

Impact of the number of clients K. Figs. 1-2 show the impact of client number K. We can see that increasing K generally worsens performance, and the degradation tends to be more pronounced under heterogeneity. Specifically, on YearPredictMSD (Fig. 1), as K increases, both SFL variants converge to higher MSE plateaus and converge slower. On CIFAR-100 (Fig. 2), the same scaling



(a) SFL-V1 (b) SFL-V2 (c) SL

Figure 1: YearPredictionMSD **loss** (i.e., MSE) versus K under $L_c = 3$.



(a) SFL-V1 (b) SFL-V2 (c) SL

Figure 2: CIFAR-100 **accuracy** (top-1) versus K under $\alpha = 0.1, L_c = 4$.

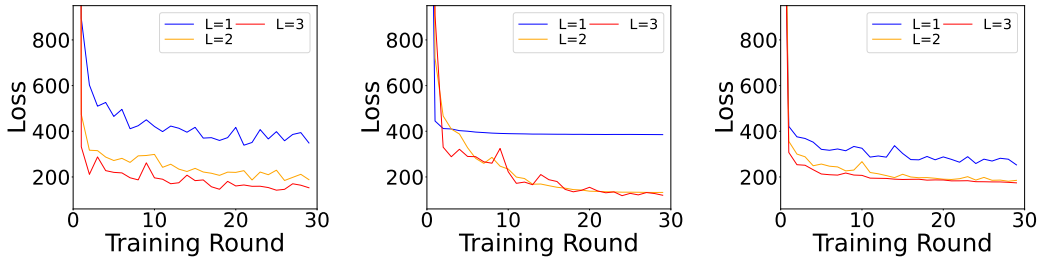
effect appears in accuracy. Under non-IID data, the accuracy drops substantially as K grows (e.g., from around 70% at $K = 2$ to below 40% at $K = 50$ for SFL).

Connection to theory. The above trends are aligned with our theory. In particular, increasing K changes client-side inner maps $g^{(k)}$, which potentially amplifies the stochasticity terms that enter the bounds through split-dependent constants and the resulting smoothness constant L_F in Lemma 4.4. Consequently, even though SFL-V1 admits an $O(1/\sqrt{T})$ stationary rate, its constants worsen as the gradient variance and client drift increase, which is consistent with the observed loss/accuracy deterioration when K becomes larger. More importantly, our SFL-V2 bound contains explicit K^2 and K^3-scaling terms (Theorem 4.7), which reflects the sequential coupling created by the mixed iterate $y_t^{(k)}$. This is consistent with Fig. 2, where SFL-V2 degrades more rapidly with K than SFL-V1.

Impact of the cut layer L_c. Figs. 3-4 show how the choice of cut layer L_c affects model performance. On both YearPredictionMSD (Fig. 3) and CIFAR-100, later splits (i.e., a larger L_c) tend to yield better performance. This observation is consistent with existing studies Han et al. (2024). Importantly, this suggests that L_c is not only a system parameter trading client and server computation, but it also impacts the optimization difficulty of SFL.

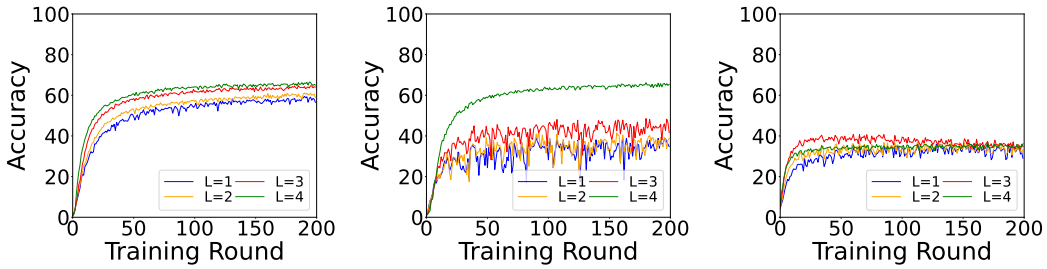
Connection to theory. The cut layer determines the split mask ω, and thereby the split-dependent constants in our bounds, i.e., $C_g(\omega)$, $\sigma_g(\omega)$, $\delta_g(\omega)$ in Assumptions 4.2–4.3 and and $L_F(\omega)$ in Lemma 4.4. An earlier split may induce larger variability in the client-side mapping $g(\cdot;\omega)$ across minibatches, which enlarges these constants. We measure this variability directly in Appendix H.5. For SFL-V2, the dependence is amplified, i.e., $L_F(\omega)$ is multiplied by K^2 in the $\mathcal{O}(\eta)$ term and $L_F^2(\omega)$ by K^3 in the $\mathcal{O}(\eta^2)$ term of Eq. 13. This matches our observation that SFL-V2 is more sensitive to the choice of cut layer.

SFL-V1 vs. SFL-V2 vs. SL. On CIFAR-100, both SFL variants outperform SL, especially when K is large. A likely reason is that SL trains clients *sequentially*, so the server-side model is repeatedly biased toward the most recently visited client distribution, which can induce catastrophic forgetting (Feng et al., 2024). In contrast, SFL-V1/V2 retain FL-style parallelism with periodic aggregation, which reduces sensitivity to any single client. On YearPredictionMSD, where data are randomly assigned to clients, SL performs comparably to SFL, and it is even better at an early cut ($L_c = 1$). This is consistent with the forgetting effect mainly arising under heterogeneous data.



(a) SFL-V1 (b) SFL-V2 (c) SL

Figure 3: YearPredictionMSD **loss** (i.e., MSE) versus L_c.



(a) SFL-V1 (b) SFL-V2 (c) SL

Figure 4: CIFAR-100 **accuracy** (top-1) versus L_c under $\alpha = 0.1$.

Additional experiments. We conduct additional experiments using (1) more datasets (i.e., CIFAR-10, FEMNIST); (2) a deeper model (ResNet-110); (3) different data heterogeneity. See Appendix H for details. Importantly, all these results are also consistent with our theory.

6 Discussion and Conclusion

Our study takes a first step toward understanding SFL through a compositional lens. By casting SFL as a single-expectation compositional problem, we establish, to our knowledge, the first nonconvex convergence guarantees for two major SFL variants. Our theory shows that SFL-V1 attains an $O(1/\sqrt{T})$ stationarity rate while SFL-V2 achieves the same rate with leading constants that scale with K^2. Our theory also provides insights on SFL design choices, including (i) cut-layer selection, (ii) algorithm choice between SFL-V1 and SFL-V2, and (iii) synchronization period. Despite these insights, our work has several limitations that motivate future research.

Limitation 1 (assumptions). Our analysis relies on assumptions such as smoothness and bounded gradients. While these are common in nonconvex stochastic compositional optimization, they might not fully reflect practical training behaviors such as heavy-tailed gradients and sharp non-smooth regimes. A potential next step is to develop analyses that need assumptions weaker than smoothness and allow heavy-tailed noise via robust gradient estimators.

Limitation 2 (practical consideration). As the first attempt, our theory focuses on the canonical SFL approaches and does not explicitly model practical system considerations such as stragglers, asynchronous implementation, or adaptive model splitting. Future research efforts could extend the compositional framework to these system-aware settings by introducing participation and delay models into the update dynamics, and designing split-aware control mechanisms that maintain convergence and accuracy under realistic distributed constraints.

7 AI Disclosure

In this work, we used AI tools for manuscript proofreading/polishing and verification of the proofs, but not for the proof derivation itself. We take sole responsibility for the final content of this work.

REFERENCES

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. *Advances in Neural Information Processing Systems*, 37:14457–14483, 2024.

Thierry Bertin-Mahieux. Year prediction msd, 2011. URL https://archive.ics.uci.edu/dataset/203/yearpredictionmsd.

Timothy J Castiglia, Anirban Das, Shiqiang Wang, and Stacy Patterson. Compressed-vfl: Communication-efficient learning with vertically partitioned data. In *International conference on machine learning*, pages 2738–2766. PMLR, 2022.

Xingyu Feng, Renqi Jia, Chengwen Luo, Victor CM Leung, and Weitao Xu. Slwf: A split learning without forgetting framework for internet of things. *IEEE Internet of Things Journal*, 2024.

Hongchang Gao. A doubly recursive stochastic compositional gradient descent method for federated multi-level compositional optimization. In *International Conference on Machine Learning*, 2024.

Hongchang Gao and Heng Huang. Fast training method for stochastic compositional optimization problems. *Advances in Neural Information Processing Systems*, 34:25334–25345, 2021.

Hongchang Gao, Junyi Li, and Heng Huang. On the convergence of local stochastic compositional gradient descent with momentum. In *International Conference on Machine Learning*, pages 7017–7035. PMLR, 2022.

Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 116:1–8, 2018.

Pengchao Han, Chao Huang, Geng Tian, Ming Tang, and Xin Liu. Convergence analysis of split federated learning on heterogeneous data. *Advances in Neural Information Processing Systems*, 37:103476–103544, 2024.

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

Zhoubin Kou, Zihan Chen, Jing Yang, and Cong Shen. Lean clients, full accuracy: Hybrid zeroth-and first-order split federated learning. *arXiv preprint arXiv:2601.09076*, 2026.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Yipeng Li and Xinchen Lyu. Convergence analysis of sequential federated learning on heterogeneous data. *Advances in Neural Information Processing Systems*, 36:56700–56755, 2023.

Yipeng Li and Xinchen Lyu. Sharp bounds for sequential federated learning on heterogeneous data. *Journal of Machine Learning Research*, 26(70):1–55, 2025.

Dandan Liang, Jianing Zhang, Evan Chen, Zhe Li, Rui Li, and Haibo Yang. Towards straggler-resilient split federated learning: An unbalanced update approach. *Advances in Neural Information Processing Systems*, 2025.

Zheng Lin, Wei Wei, Zhe Chen, Chan-Tong Lam, Xianhao Chen, Yue Gao, and Jun Luo. Hierarchical split federated learning: Convergence analysis and system optimization. *IEEE Transactions on Mobile Computing*, 2025.

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*, pages 1273–1282. PMLR, 2017.

Srijith Nair, Michael Lin, Peizhong Ju, Amirreza Talebi, Elizabeth Serena Bentley, and Jia Liu. Fsl-sage: Accelerating federated split learning via smashed activation gradient estimation. *International Conference on Machine Learning*, 2025.

Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, and Samet Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pages 21146–21179. PMLR, 2022.

Chandra Thapa, Pathum Chamikara Mahawaga Arachchige, Seyit Camtepe, and Lichao Sun. Splitfed: When federated learning meets split learning. In *AAAI conference on artificial intelligence*, volume 36, pages 8485–8493, 2022.

Joana Tirana, Dimitra Tsigkari, George Iosifidis, and Dimitris Chatzopoulos. Workflow optimization for parallel split learning. In *IEEE Conference on Computer Communications*, pages 1331–1340, 2024.

Pedro Valdeira, Shiqiang Wang, and Yuejie Chi. Vertical federated learning with missing features during training and inference. In *International Conference on Learning Representations*, volume 2025, pages 72304–72323, 2025.

Xidong Wu, Jianhui Sun, Zhengmian Hu, Junyi Li, Aidong Zhang, and Heng Huang. Federated conditional stochastic optimization. *Advances in Neural Information Processing Systems*, 36: 5752–5764, 2023.

Xinwen Zhang, Yihan Zhang, Tianbao Yang, Richard Souvenir, and Hongchang Gao. Federated compositional deep auc maximization. *Advances in Neural Information Processing Systems*, 36: 9648–9660, 2023.

Xinwen Zhang, Ali Payani, Myungjin Lee, Richard Souvenir, and Hongchang Gao. A federated stochastic multi-level compositional minimax algorithm for deep auc maximization. In *International Conference on Machine Learning*, 2024.

A ORGANIZATION OF THE APPENDIX

We organize the supplementary materials as follows.

- In Appendix B, we provide the pseudocodes of SFL-V1 and SFL-V2.
- In Appendix C, we prove Lemma 4.4.
- In Appendix D, we prove Proposition 4.5.
- In Appendix E, we prove Theorem 4.6.
- In Appendix F, we prove Theorem 4.7.
- In Appendix G, we provide a detailed rate comparison among representative SL/SFL results.
- In Appendix H, we provide more experiment details and results.
  - In Appendix H.1, we provide details of hyper-parameters.
  - In Appendix H.2, we provide experiment results using more datasets
  - In Appendix H.3, we provide more experiments using a deeper model.
  - In Appendix H.4, we provide experiments with more data heterogeneity levels.
  - In Appendix H.5, we show how the cut layer variability affects SFL's performance.

B PSEUDO-CODES OF SFL-V1 AND SFL-V2

The pseudo-codes of SFL-V1 and SFL-V2 are presented below.

Algorithm 1 SFL-V1: Parallel Updates.

Require: $\eta > 0$, $p > 1$, init. $x_0^{(k)} = x_0, \forall k$.

for $t = 0, 1, \ldots, T-1$ **do**
  for all $k = 1, \ldots, K$ **in parallel do**
    $u_{t+1}^{(k)} = g^{(k)}(\omega \odot x_t^{(k)}; \xi)$;
    $s_{t+1}^{(k)} = h(x_t^{(k)})$;
    Calculate $z_t^{(k)}$ via Eq. 6;
    $x_{t+1}^{(k)} = x_t^{(k)} - \eta\, z_t^{(k)}$;
  end for
  if $\mathrm{mod}(t+1, p) = 0$ **then**
    $\bar{x}_{t+1} \leftarrow \frac{1}{K} \sum_{k'=1}^{K} x_{t+1}^{(k')}$;
    $x_{t+1}^{(k)} \leftarrow \bar{x}_{t+1}$.
  end if
end for

Algorithm 2 SFL-V2: Sequential Updates.

Require: $\eta > 0$, $p > 1$, init. $x_0^{(k)} = x_0, \forall k$; convention $x_{t+1}^{(0)} := x_t^{(K)}$, $x_0^{(K)} := x_0$.

for $t = 0, 1, \ldots, T-1$ **do**
  for $k = 1, \ldots, K$ **do**
    Calculate $y_t^{(k)}$ via Eq. 7;
    $u_{t+1}^{(k)} = g^{(k)}(\omega \odot y_t^{(k)}; \xi)$;
    $s_{t+1}^{(k)} = h(y_t^{(k)})$;
    Calculate $z_t^{(k)}$ via Eq. 6 with x replaced by $y_t^{(k)}$;
    $x_{t+1}^{(k)} = y_t^{(k)} - \eta\, z_t^{(k)}$;
  end for
  if $\mathrm{mod}(t+1, p) = 0$ **then**
    $\bar{x}_{t+1} \leftarrow \frac{1}{K} \sum_{k'=1}^{K} x_{t+1}^{(k')}$;
    $x_{t+1}^{(k)} \leftarrow \omega \odot \bar{x}_{t+1} + (\mathbf{1} - \omega) \odot x_{t+1}^{(K)}$.
  end if
end for

C Proof of Lemma 4.4

Proof. Based on Assumptions 4.1-4.2, we have the fact that the client-side mapping $g^{(k)}(\omega \odot \cdot\,; \xi)$ is Lipschitz in mean square. Let $z_t = y + t(x - y)$ for $t \in [0, 1]$. Then,

$$\begin{aligned}
\mathbb{E}_\xi \big\| g^{(k)}(\omega \odot x; \xi) - g^{(k)}(\omega \odot y; \xi) \big\|^2 &= \mathbb{E}_\xi \left\| \int_0^1 \big[\omega \odot \nabla g^{(k)}(\omega \odot z_t; \xi)\big]^\top (x - y)\, dt \right\|^2 \\
&\le \|x - y\|^2\, \mathbb{E}_\xi \left[\left(\int_0^1 \big\| \nabla g^{(k)}(\omega \odot z_t; \xi) \big\|\, dt \right)^2 \right] \\
&\le \|x - y\|^2 \int_0^1 \mathbb{E}_\xi \big\| \nabla g^{(k)}(\omega \odot z_t; \xi) \big\|^2 dt \\
&\le C_g^2(\omega) \|x - y\|^2.
\end{aligned} \tag{14}$$

Here, first inequality uses the triangle inequality, $\|A^\top v\| \le \|A\|\, \|v\|$, and $\|\omega \odot A\| \le \|A\|$. The second inequality is based on Jensen's inequality, and the third one uses Assumption 4.2. Moreover, by Jensen's inequality and Assumption 4.2, we have $\mathbb{E}_\xi \|\nabla g^{(k)}(\omega \odot x; \xi)\| \le C_g(\omega)$ and $\mathbb{E}_\zeta \|\nabla f^{(k)}(y; \zeta)\| \le C_f$.

Then, for ease of presentation, define $g_x := g^{(k)}(\omega \odot x; \xi)$ and $g_y := g^{(k)}(\omega \odot y; \xi)$, and we have

$$\begin{aligned}
&\|\nabla F(x) - \nabla F(y)\| \\
&= \left\| \frac{1}{K} \sum_{k=1}^K \mathbb{E}\Big[\nabla_x F^{(k)}(x; \xi, \zeta) - \nabla_x F^{(k)}(y; \xi, \zeta) \Big] \right\| \\
&\le \frac{1}{K} \sum_{k=1}^K \mathbb{E} \Big\| \nabla_x F^{(k)}(x; \xi, \zeta) - \nabla_x F^{(k)}(y; \xi, \zeta) \Big\| \\
&= \frac{1}{K} \sum_{k=1}^K \mathbb{E} \Big\| \Big\{ \big[\omega \odot \nabla g^{(k)}(\omega \odot x; \xi)\big]^\top \nabla_g f^{(k)}(g_x, h(x); \zeta) - \big[\omega \odot \nabla g^{(k)}(\omega \odot y; \xi)\big]^\top \nabla_g f^{(k)}(g_x, h(x); \zeta) \\
&\quad + \big[\omega \odot \nabla g^{(k)}(\omega \odot y; \xi)\big]^\top \nabla_g f^{(k)}(g_x, h(x); \zeta) - \big[\omega \odot \nabla g^{(k)}(\omega \odot y; \xi)\big]^\top \nabla_g f^{(k)}(g_y, h(x); \zeta) \\
&\quad + \big[\omega \odot \nabla g^{(k)}(\omega \odot y; \xi)\big]^\top \nabla_g f^{(k)}(g_y, h(x); \zeta) - \big[\omega \odot \nabla g^{(k)}(\omega \odot y; \xi)\big]^\top \nabla_g f^{(k)}(g_y, h(y); \zeta) \\
&\quad + (1 - \omega) \odot \nabla_h f^{(k)}(g_x, h(x); \zeta) - (1 - \omega) \odot \nabla_h f^{(k)}(g_y, h(x); \zeta) \\
&\quad + (1 - \omega) \odot \nabla_h f^{(k)}(g_y, h(x); \zeta) - (1 - \omega) \odot \nabla_h f^{(k)}(g_y, h(y); \zeta) \Big\} \Big\| \\
&\le \frac{1}{K} \sum_{k=1}^K \Big\{ \mathbb{E}\Big[\big\| \nabla g^{(k)}(\omega \odot x; \xi) - \nabla g^{(k)}(\omega \odot y; \xi) \big\|\, \big\| \nabla f^{(k)}(\cdot\,; \zeta) \big\| \Big] + L_f\, \mathbb{E}\Big[\big\| \nabla g^{(k)}(\omega \odot y; \xi) \big\|\, \|g_x - g_y\| \Big] \\
&\quad + L_f\, \mathbb{E}\Big[\big\| \nabla g^{(k)}(\omega \odot y; \xi) \big\| \Big] \|h(x) - h(y)\| + L_f\, \mathbb{E} \|g_x - g_y\| + L_f \|h(x) - h(y)\| \Big\} \\
&\le \frac{1}{K} \sum_{k=1}^K \Big\{ L_g(\omega) \|x - y\| \sqrt{\mathbb{E} \big\| \nabla f^{(k)}(\cdot\,; \zeta) \big\|^2} + L_f \sqrt{\mathbb{E} \big\| \nabla g^{(k)}(\omega \odot y; \xi) \big\|^2} \sqrt{\mathbb{E} \|g_x - g_y\|^2} \\
&\quad + L_f \sqrt{\mathbb{E} \big\| \nabla g^{(k)}(\omega \odot y; \xi) \big\|^2}\, \|x - y\| + L_f \sqrt{\mathbb{E} \|g_x - g_y\|^2} + L_f \|x - y\| \Big\} \\
&\le \Big(C_f L_g(\omega) + C_g^2(\omega) L_f + C_g(\omega) L_f + C_g(\omega) L_f + L_f \Big) \|x - y\| \\
&= \Big(C_f L_g(\omega) + L_f \big(C_g(\omega) + 1 \big)^2 \Big) \|x - y\| \\
&\triangleq L_F \|x - y\|.
\end{aligned} \tag{15}$$

Here, the first inequality is based on Jensen's inequality and the triangle inequality. The second inequality uses the triangle inequality, the facts that $\|A^\top v\| \le \|A\|\, \|v\|$, $\|\omega \odot A\| \le \|A\|$ and

$\|(1-\omega)\odot a\| \le \|a\|$, and Assumption 4.1 applied to the per-sample map $f^{(k)}(\cdot\,;\zeta)$ with the concatenated argument $y=(u,x_s)$, so that each block-wise difference is controlled by L_f times the difference in the perturbed block alone. The third inequality uses Assumption 4.1 applied to the per-sample map $g^{(k)}(\omega\odot\cdot\,;\xi)$, the Cauchy-Schwarz inequality, and Jensen's inequality. The last inequality uses Assumption 4.2 and Eq. 14.

Thus we complete our proof. □

D PROOF OF PROPOSITION 4.5

We use a few lemmas to facilitate the proof of Proposition 4.5. Here is a roadmap.

- We show and prove Lemma D.1 that bounds gradient differences across clients.
- We show and prove Lemma D.2 that bounds the variance of stochastic gradients.
- We show and prove Lemma D.3 that bounds the gradient divergence.
- Finally we prove Proposition 4.5.

D.1 USEFUL LEMMAS

Lemma D.1. *Under Assumptions 4.2-4.3, there exists constants $\gamma_g^2(\omega)$ and γ_f^2 such that for any x, y,*

$$\sum_{k=1}^{K}\left[\left\|\nabla g^{(k)}\left(\omega\odot x\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x\right)\right\|^2\right]\leq\gamma_g^2(\omega),$$
$$\sum_{k=1}^{K}\left[\left\|\nabla f^{(k)}\left(y\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}\left(y\right)\right\|^2\right]\leq\gamma_f^2, \tag{16}$$

where $\gamma_g^2(\omega)$ and γ_f^2 were given in (10).

This lemma helps bound the difference between local and global objectives.

Proof. Throughout, $\xi^{(k)}$ (resp. $\zeta^{(k)}$) denotes the stochastic sample drawn by client k, and all expectations are taken over $\{\xi^{(k)}\}_{k=1}^K$ (resp. $\{\zeta^{(k)}\}_{k=1}^K$). We have

$$\begin{aligned}
&\sum_{k=1}^{K}\left\|\nabla g^{(k)}(\omega\odot x)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x)\right\|^2\\
&=\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}(\omega\odot x)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x)\right\|^2\right]\\
&=\sum_{k=1}^{K}\mathbb{E}\Bigg[\Bigg\|\nabla g^{(k)}(\omega\odot x)-\nabla g^{(k)}(\omega\odot x;\xi^{(k)})\\
&\qquad+\nabla g^{(k)}(\omega\odot x;\xi^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x;\xi^{(k')})\\
&\qquad+\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x;\xi^{(k')})-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x)\Bigg\|^2\Bigg]\\
&\leqslant 3\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}(\omega\odot x)-\nabla g^{(k)}(\omega\odot x;\xi^{(k)})\right\|^2\right]\\
&\quad+3\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}(\omega\odot x;\xi^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x;\xi^{(k')})\right\|^2\right]\\
&\quad+3\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x;\xi^{(k')})-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x)\right\|^2\right]\\
&\leqslant 3K\sigma_g^2(\omega)+6\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}(\omega\odot x;\xi^{(k)})\right\|^2\right]+6\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x;\xi^{(k')})\right\|^2\right]
\end{aligned}$$

$$+\frac{3}{K}\sum_{k=1}^{K}\sum_{k'=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k')}(\omega\odot x;\xi^{(k')})-\nabla g^{(k')}(\omega\odot x)\right\|^2\right]$$
$$\leqslant 3K\sigma_g^2(\omega)+12KC_g^2(\omega)+3K\sigma_g^2(\omega)\ \triangleq\ \gamma_g^2(\omega). \tag{17}$$

Here, the first inequality uses $\|a+b+c\|^2\leqslant 3(\|a\|^2+\|b\|^2+\|c\|^2)$. The second uses $\|a-b\|^2\leqslant 2\|a\|^2+2\|b\|^2$ on the middle term, Jensen's inequality on the inner average of the last term, and Assumption 4.3 on the first term. The last inequality uses Assumption 4.2 (together with Jensen's inequality for the averaged term) and Assumption 4.3.

Similarly, we have the following:

$$\begin{aligned}
&\sum_{k=1}^{K}\left\|\nabla f^{(k)}(y)-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(y)\right\|^2\\
&=\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k)}(y)-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(y)\right\|^2\right]\\
&=\sum_{k=1}^{K}\mathbb{E}\Bigg[\Big\|\nabla f^{(k)}(y)-\nabla f^{(k)}(y;\zeta^{(k)})\\
&\qquad+\ \nabla f^{(k)}(y;\zeta^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(y;\zeta^{(k')})\\
&\qquad+\ \frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(y;\zeta^{(k')})-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(y)\Big\|^2\Bigg]\\
&\leqslant 3\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k)}(y)-\nabla f^{(k)}(y;\zeta^{(k)})\right\|^2\right]\\
&\quad+3\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k)}(y;\zeta^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(y;\zeta^{(k')})\right\|^2\right]\\
&\quad+3\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(y;\zeta^{(k')})-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(y)\right\|^2\right]\\
&\leqslant 3K\sigma_f^2+6\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k)}(y;\zeta^{(k)})\right\|^2\right]+6\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(y;\zeta^{(k')})\right\|^2\right]\\
&\quad+\frac{3}{K}\sum_{k=1}^{K}\sum_{k'=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k')}(y;\zeta^{(k')})-\nabla f^{(k')}(y)\right\|^2\right]\\
&\leqslant 3K\sigma_f^2+12KC_f^2+3K\sigma_f^2\ \triangleq\ \gamma_f^2,
\end{aligned}\tag{18}$$

where the same inequalities are applied, and $\mathbb{E}_\zeta\big[\sup_y\|\nabla f^{(k)}(y;\zeta)\|^2\big]\leqslant C_f^2$ from Assumption 4.2 is used at the fixed point y.

Thus we complete our proof. □

Lemma D.2. *Under Assumptions 4.1–4.3, the local stochastic gradients have bounded variance. That is, for any client k and any x,*

$$\mathbb{E}\left\|\nabla_x F^{(k)}(x;\xi,\zeta)-\nabla F^{(k)}(x)\right\|^2\leq\sigma_F^2(\omega), \tag{19}$$

where $\sigma_F^2(\omega)$ was given in (10).

Proof. For ease of presentation, we define $g_\xi := g^{(k)}(\omega\odot x;\xi)$ and $\bar{g} := g^{(k)}(\omega\odot x) = \mathbb{E}_\xi[g_\xi]$, let $\nabla f^{(k)}(y) := \mathbb{E}_\zeta\big[\nabla f^{(k)}(y;\zeta)\big]$ denote the mean outer gradient, and write $S(\zeta) :=$

$\sup_y \big\|\nabla f^{(k)}(y;\zeta) - \nabla f^{(k)}(y)\big\|$. Throughout, $\|\cdot\|$ is the Euclidean norm, so that on the concatenated argument $y = (u, x_s)$ we have $\|\nabla f^{(k)}\|^2 = \|\nabla_g f^{(k)}\|^2 + \|\nabla_h f^{(k)}\|^2$ and each block is dominated by the full gradient. For a matrix A we have $[\omega \odot A]^\top v = \omega \odot (A^\top v)$ and $\|\omega \odot A\| \le \|A\|$, and, as elsewhere, $\nabla g^{(k)}(\omega \odot x) := \mathbb{E}_\xi[\nabla g^{(k)}(\omega \odot x; \xi)]$.

By Assumption 4.2, the stochastic gradient in Eq. 6 satisfies $\|\nabla_x F^{(k)}(x;\xi,\zeta)\| \le (C_g(\omega) + 1) \sup_y \|\nabla f^{(k)}(y;\zeta)\|$, and we have $\nabla F^{(k)}(x) = \mathbb{E}\big[\nabla_x F^{(k)}(x;\xi,\zeta)\big]$. Finally, define

$$c^{(k)}(x) := \Big[\omega \odot \nabla g^{(k)}(\omega \odot x)\Big]^\top \nabla_g f^{(k)}\big(\bar{g}, h(x)\big) + (1-\omega) \odot \nabla_h f^{(k)}\big(\bar{g}, h(x)\big). \tag{20}$$

We then have

$$\begin{aligned}
&\mathbb{E}\big\|\nabla_x F^{(k)}(x;\xi,\zeta) - \nabla F^{(k)}(x)\big\|^2 \\
&\leqslant \mathbb{E}\Bigg[\Bigg\|\Big[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi)\Big]^\top \nabla_g f^{(k)}\big(g_\xi, h(x);\zeta\big) + (1-\omega) \odot \nabla_h f^{(k)}\big(g_\xi, h(x);\zeta\big) \\
&\qquad\qquad - \Big[\omega \odot \nabla g^{(k)}(\omega \odot x)\Big]^\top \nabla_g f^{(k)}\big(\bar{g}, h(x)\big) - (1-\omega) \odot \nabla_h f^{(k)}\big(\bar{g}, h(x)\big)\Bigg\|^2\Bigg] \\
&= \mathbb{E}\Bigg[\Bigg\|\Big[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi)\Big]^\top \nabla_g f^{(k)}\big(g_\xi, h(x);\zeta\big) \\
&\qquad - \Big[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi)\Big]^\top \nabla_g f^{(k)}\big(g_\xi, h(x)\big) \\
&\qquad + \Big[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi)\Big]^\top \nabla_g f^{(k)}\big(g_\xi, h(x)\big) \\
&\qquad - \Big[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi)\Big]^\top \nabla_g f^{(k)}\big(\bar{g}, h(x)\big) \\
&\qquad + \Big[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi)\Big]^\top \nabla_g f^{(k)}\big(\bar{g}, h(x)\big) \\
&\qquad - \Big[\omega \odot \nabla g^{(k)}(\omega \odot x)\Big]^\top \nabla_g f^{(k)}\big(\bar{g}, h(x)\big) \\
&\qquad + (1-\omega) \odot \nabla_h f^{(k)}\big(g_\xi, h(x);\zeta\big) - (1-\omega) \odot \nabla_h f^{(k)}\big(g_\xi, h(x)\big) \\
&\qquad + (1-\omega) \odot \nabla_h f^{(k)}\big(g_\xi, h(x)\big) - (1-\omega) \odot \nabla_h f^{(k)}\big(\bar{g}, h(x)\big)\Bigg\|^2\Bigg] \\
&\leqslant 5\,\mathbb{E}\Bigg[\Bigg\|\Big[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi)\Big]^\top \Big(\nabla_g f^{(k)}\big(g_\xi, h(x);\zeta\big) - \nabla_g f^{(k)}\big(g_\xi, h(x)\big)\Big)\Bigg\|^2\Bigg] \\
&\quad + 5\,\mathbb{E}\Bigg[\Bigg\|\Big[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi)\Big]^\top \Big(\nabla_g f^{(k)}\big(g_\xi, h(x)\big) - \nabla_g f^{(k)}\big(\bar{g}, h(x)\big)\Big)\Bigg\|^2\Bigg] \\
&\quad + 5\,\mathbb{E}\Bigg[\Bigg\|\Big[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi) - \omega \odot \nabla g^{(k)}(\omega \odot x)\Big]^\top \nabla_g f^{(k)}\big(\bar{g}, h(x)\big)\Bigg\|^2\Bigg] \\
&\quad + 5\,\mathbb{E}\Bigg[\Bigg\|(1-\omega) \odot \Big(\nabla_h f^{(k)}\big(g_\xi, h(x);\zeta\big) - \nabla_h f^{(k)}\big(g_\xi, h(x)\big)\Big)\Bigg\|^2\Bigg] \\
&\quad + 5\,\mathbb{E}\Bigg[\Bigg\|(1-\omega) \odot \Big(\nabla_h f^{(k)}\big(g_\xi, h(x)\big) - \nabla_h f^{(k)}\big(\bar{g}, h(x)\big)\Big)\Bigg\|^2\Bigg] \\
&\leqslant 5C_g^2(\omega)\,\mathbb{E}\big[S^2(\zeta)\big] + 5C_g^2(\omega)L_f^2\,\mathbb{E}\big\|g_\xi - \bar{g}\big\|^2 + 5C_f^2\,\mathbb{E}\big\|\nabla g^{(k)}(\omega \odot x;\xi) - \nabla g^{(k)}(\omega \odot x)\big\|^2 \\
&\quad + 5\,\mathbb{E}\big[S^2(\zeta)\big] + 5L_f^2\,\mathbb{E}\big\|g_\xi - \bar{g}\big\|^2 \\
&\leqslant 5C_g^2(\omega)\sigma_f^2 + 5C_g^2(\omega)L_f^2\delta_g^2(\omega) + 5C_f^2\sigma_g^2(\omega) + 5\sigma_f^2 + 5L_f^2\delta_g^2(\omega) \\
&= 5C_g^2(\omega)\sigma_f^2 + 5\big(C_g^2(\omega)+1\big)L_f^2\delta_g^2(\omega) + 5C_f^2\sigma_g^2(\omega) + 5\sigma_f^2 \;\triangleq\; \sigma_F^2(\omega).
\end{aligned} \tag{21}$$

Here, the first inequality holds because $\mathbb{E}\|X - c\|^2 = \mathbb{E}\|X - \mathbb{E}X\|^2 + \|\mathbb{E}X - c\|^2 \geqslant \mathbb{E}\|X - \mathbb{E}X\|^2$ for any c. The first equality adds and subtracts the intermediate terms $[\omega \odot \nabla g^{(k)}(\omega \odot$

$x;\xi)]^\top \nabla_g f^{(k)}(g_\xi, h(x))$, $[\omega \odot \nabla g^{(k)}(\omega \odot x;\xi)]^\top \nabla_g f^{(k)}(\bar{g}, h(x))$ and $(1-\omega)\odot \nabla_h f^{(k)}(g_\xi, h(x))$. The second inequality uses $\left\|\sum_{i=1}^5 a_i\right\|^2 \leqslant 5\sum_{i=1}^5 \|a_i\|^2$. The third inequality uses $\|A^\top v\| \leqslant \|A\|\,\|v\|$, $\|\omega \odot A\| \leqslant \|A\|$, $\|(1-\omega)\odot a\| \leqslant \|a\|$, and the fact that each block of $\nabla f^{(k)}$ is dominated by the full gradient. On the first and second terms it also uses the realization-wise bound $\|\nabla g^{(k)}(\omega \odot x;\xi)\| \leqslant C_g(\omega)$ from Assumption 4.2. The remaining factor is then at most $S(\zeta)$ on the first and fourth terms. On the second and fifth terms it uses in addition that $\nabla f^{(k)}(\cdot) = \mathbb{E}_\zeta[\nabla f^{(k)}(\cdot\,;\zeta)]$ is L_f-Lipschitz, which follows from Assumption 4.1 and Jensen's inequality, together with $\big(g_\xi, h(x)\big) - \big(\bar{g}, h(x)\big) = (g_\xi - \bar{g}, 0)$. On the third term the second factor is deterministic and satisfies $\|\nabla_g f^{(k)}(\bar{g}, h(x))\| \leqslant \sup_y \|\nabla f^{(k)}(y)\| \leqslant C_f$ by Jensen's inequality and Assumption 4.2. The fourth inequality uses Assumption 4.3.

Thus we complete our proof. □

Lemma D.3. *(SFL-V1: Bounding Gradient Divergence) Let Assumptions 4.1-4.3 hold. Then we have*

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{E}\left[\left\|z_t^{(k)} - \bar{z}_t\right\|^2\right] \leqslant 14(C_f^2+\sigma_f^2)\left(10K\sigma_g^2(\omega) + 5\gamma_g^2(\omega)\right)\\
&+ (14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left(3\gamma_f^2 + 30L_f^2\gamma_g^2(\omega) + 60KL_f^2\sigma_g^2(\omega)\right)\\
&+ 14K\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right] + 140L_g^2(C_f^2+\sigma_f^2)\sum_{k=1}^K \mathbb{E}\left[\left\|x_t^{(k)} - \bar{x}_t\right\|^2\right]\\
&+ 6L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)\sum_{k=1}^K \mathbb{E}\left[\left\|x_{t+1}^{(k)} - \bar{x}_{t+1}\right\|^2\right]
\end{aligned} \tag{22}$$

Lemma D.3 controls the *gradient divergence* i.e., how far local gradient estimators deviate from the average. It shows that as local client models drift away from the global average due to multiple local split updates, their cut-layer representations and backpropagated signals also diverge. With this result, we are ready to bound the client drift.

Proof. Similar to the proof of Lemma D.1, throughout in the k-th summand, ξ and ζ denote the samples $\xi^{(k)}$ and $\zeta^{(k)}$ drawn by client k and $\nabla f^{(k)}(y) := \mathbb{E}_\zeta\left[\nabla f^{(k)}(y;\zeta)\right]$ denotes the mean outer gradient. Recall that $u_{t+1}^{(k)} = g^{(k)}(\omega \odot x_t^{(k)};\xi)$ and $s_{t+1}^{(k)} = h(x_t^{(k)})$.

The following holds:

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{E}\left[\left\|z_t^{(k)} - \bar{z}_t\right\|^2\right]\\
&= \sum_{k=1}^K \mathbb{E}\left[\left\|z_t^{(k)} - \frac{1}{K}\sum_{k'=1}^K z_t^{(k')}\right\|^2\right]\\
&= \sum_{k=1}^K \mathbb{E}\left[\left\|\left[\omega \odot \nabla g^{(k)}\left(\omega \odot x_t^{(k)};\xi\right)\right]^\top \nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)};\zeta\right) + (1-\omega)\odot \nabla_h f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)};\zeta\right)\right.\right.\\
&\quad \left.\left. -\frac{1}{K}\sum_{k'=1}^K \left[\left[\omega \odot \nabla g^{(k')}\left(\omega \odot x_t^{(k')};\xi\right)\right]^\top \nabla_g f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')};\zeta\right) + (1-\omega)\odot \nabla_h f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')};\zeta\right)\right]\right\|^2\right]\\
&= \sum_{k=1}^K \mathbb{E}\left[\left\|\left[\omega \odot \nabla g^{(k)}\left(\omega \odot x_t^{(k)};\xi\right)\right]^\top \nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)};\zeta\right)\right.\right.\\
&\quad - \left[\omega \odot \nabla g^{(k)}\left(\omega \odot x_t^{(k)};\xi\right)\right]^\top \nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right)\\
&\quad \left.\left. + \left[\omega \odot \nabla g^{(k)}\left(\omega \odot x_t^{(k)};\xi\right)\right]^\top \nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{K}\sum_{k'=1}^{K}\left[\omega \odot \nabla g^{(k')}\left(\omega \odot x_t^{(k')};\xi\right)\right]^\top \nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right) \\
&+ \frac{1}{K}\sum_{k'=1}^{K}\left[\omega \odot \nabla g^{(k')}\left(\omega \odot x_t^{(k')};\xi\right)\right]^\top \nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right) \\
&- \frac{1}{K}\sum_{k'=1}^{K}\left[\omega \odot \nabla g^{(k')}\left(\omega \odot x_t^{(k')};\xi\right)\right]^\top \nabla_g f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right) \\
&+ \frac{1}{K}\sum_{k'=1}^{K}\left[\omega \odot \nabla g^{(k')}\left(\omega \odot x_t^{(k')};\xi\right)\right]^\top \nabla_g f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right) \\
&- \frac{1}{K}\sum_{k'=1}^{K}\left[\omega \odot \nabla g^{(k')}\left(\omega \odot x_t^{(k')};\xi\right)\right]^\top \nabla_g f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')};\zeta\right) \\
&+ (1-\omega)\odot \nabla_h f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)};\zeta\right) - (1-\omega)\odot \nabla_h f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right) \\
&+ (1-\omega)\odot \nabla_h f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right) - \frac{1}{K}\sum_{k'=1}^{K}(1-\omega)\odot \nabla_h f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right) \\
&+ \frac{1}{K}\sum_{k'=1}^{K}(1-\omega)\odot \nabla_h f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right) - \frac{1}{K}\sum_{k'=1}^{K}(1-\omega)\odot \nabla_h f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')};\zeta\right)\Bigg\|^2\Bigg] \\
\leqslant &\sum_{k=1}^{K} 7\Bigg\{2(C_f^2+\sigma_f^2)\mathbb{E}\left[\left\|\left[\omega \odot \nabla g^{(k)}\left(\omega \odot x_t^{(k)};\xi\right)\right]^\top - \frac{1}{K}\sum_{k'=1}^{K}\left[\omega \odot \nabla g^{(k')}\left(\omega \odot x_t^{(k')};\xi\right)\right]^\top\right\|^2\right] \\
&+ 2(C_g^2(\omega)+\sigma_g^2(\omega))\sigma_f^2 + 2(C_g^2(\omega)+\sigma_g^2(\omega))\mathbb{E}\left[\left\|\nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_g f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right)\right\|^2\right] \\
&+ 2(C_g^2(\omega)+\sigma_g^2(\omega))\sigma_f^2 + \sigma_f^2 + \mathbb{E}\left[\left\|\nabla_h f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_h f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right)\right\|^2\right] + \sigma_f^2\Bigg\}
\end{aligned}
\tag{23}
$$

Here, the first equality uses the definition of $\bar{z}_t$, and the second one uses Eq. 6. The third equality adds and subtracts the intermediate terms $\left[\omega \odot \nabla g^{(k)}(\omega \odot x_t^{(k)};\xi)\right]^\top \nabla_g f^{(k)}(u_{t+1}^{(k)}, s_{t+1}^{(k)})$, $\frac{1}{K}\sum_{k'=1}^{K}\left[\omega \odot \nabla g^{(k')}(\omega \odot x_t^{(k')};\xi)\right]^\top \nabla_g f^{(k)}(u_{t+1}^{(k)}, s_{t+1}^{(k)})$, $\frac{1}{K}\sum_{k'=1}^{K}\left[\omega \odot \nabla g^{(k')}(\omega \odot x_t^{(k')};\xi)\right]^\top \nabla_g f^{(k')}(u_{t+1}^{(k')}, s_{t+1}^{(k')})$, $(1-\omega)\odot \nabla_h f^{(k)}(u_{t+1}^{(k)}, s_{t+1}^{(k)})$ and $\frac{1}{K}\sum_{k'=1}^{K}(1-\omega)\odot \nabla_h f^{(k')}(u_{t+1}^{(k')}, s_{t+1}^{(k')})$, which splits the difference into 7 groups. The inequality uses $\|\sum_{i=1}^{7} a_i\|^2 \leqslant 7\sum_{i=1}^{7}\|a_i\|^2$ and bounds the seven groups, where we use $\|A^\top v\| \leqslant \|A\|\,\|v\|$, $\|\omega \odot A\| \leqslant \|A\|$, $\|(1-\omega)\odot a\| \leqslant \|a\|$, and the fact that each block of $\nabla f^{(k)}$ is dominated by the full gradient.

Then, collecting the constants and using $\|\omega \odot A\| \leqslant \|A\|$ in the first term,

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|z_t^{(k)} - \bar{z}_t\right\|^2\right] \\
\leqslant &\sum_{k=1}^{K} 7\Bigg\{2(C_f^2+\sigma_f^2)\mathbb{E}\left[\left\|\left[\nabla g^{(k)}\left(\omega \odot x_t^{(k)};\xi\right)\right]^\top - \frac{1}{K}\sum_{k'=1}^{K}\left[\nabla g^{(k')}\left(\omega \odot x_t^{(k')};\xi\right)\right]^\top\right\|^2\right] \\
&+ 2(C_g^2(\omega)+\sigma_g^2(\omega))\mathbb{E}\left[\left\|\nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_g f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right)\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\mathbb{E}\left[\left\|\nabla_h f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla_h f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right)\right\|^2\right]+4(C_g^2(\omega)+\sigma_g^2(\omega))\sigma_f^2+2\sigma_f^2\Bigg\}\\
&=14(C_f^2+\sigma_f^2)\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}\left(\omega\odot x_t^{(k)};\xi\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')};\xi\right)\right\|^2\right]\\
&\quad+14(C_g^2(\omega)+\sigma_g^2(\omega))\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla_g f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right)\right\|^2\right]\\
&\quad+7\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla_h f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla_h f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right)\right\|^2\right]+14K\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]\\
&\leqslant 14(C_f^2+\sigma_f^2)\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}\left(\omega\odot x_t^{(k)};\xi\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')};\xi\right)\right\|^2\right]\\
&\quad+(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left\{\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla_g f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla_g f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right)\right\|^2\right]\right.\\
&\quad\left.+\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla_h f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla_h f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right)\right\|^2\right]\right\}+14K\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]\\
&=14(C_f^2+\sigma_f^2)\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}\left(\omega\odot x_t^{(k)};\xi\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')};\xi\right)\right\|^2\right]\\
&\quad+(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left\{\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k)}\left(u_{t+1}^{(k)}, s_{t+1}^{(k)}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}\left(u_{t+1}^{(k')}, s_{t+1}^{(k')}\right)\right\|^2\right]\right\}\\
&\quad+14K\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]. \qquad (24)
\end{aligned}
$$

Here, the first equality collects the constants, since $\sum_{k=1}^{K}7\big(4(C_g^2(\omega)+\sigma_g^2(\omega))\sigma_f^2+2\sigma_f^2\big)=14K\sigma_f^2\big[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\big]$. The second inequality uses $7\leqslant 14(C_g^2(\omega)+\sigma_g^2(\omega))+7$ and $14(C_g^2(\omega)+\sigma_g^2(\omega))\leqslant 14(C_g^2(\omega)+\sigma_g^2(\omega))+7$. The last equality uses $\|\nabla f^{(k)}\|^2=\|\nabla_g f^{(k)}\|^2+\|\nabla_h f^{(k)}\|^2$, where the average over k' splits block by block.

Now what remains is to estimate the two terms on the right side of Eq. 24.

1. We estimate the first term on the right side of Eq. 24.

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}\left(\omega\odot x_t^{(k)};\xi\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')};\xi\right)\right\|^2\right]\\
&=\sum_{k=1}^{K}\mathbb{E}\Bigg[\left\|\nabla g^{(k)}\left(\omega\odot x_t^{(k)};\xi\right)-\nabla g^{(k)}\left(\omega\odot x_t^{(k)}\right)\right.\\
&\quad+\nabla g^{(k)}\left(\omega\odot x_t^{(k)}\right)-\nabla g^{(k)}\left(\omega\odot \bar{x}_t\right)+\nabla g^{(k)}\left(\omega\odot \bar{x}_t\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot \bar{x}_t\right)\\
&\quad+\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot \bar{x}_t\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')}\right)\\
&\quad\left.+\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')};\xi\right)\right\|^2\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant 5\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}\left(\omega\odot x_t^{(k)};\xi\right)-\nabla g^{(k)}\left(\omega\odot x_t^{(k)}\right)\right\|^2\right]\\
&\quad+5\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}\left(\omega\odot x_t^{(k)}\right)-\nabla g^{(k)}\left(\omega\odot \bar{x}_t\right)\right\|^2\right]\\
&\quad+5\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}\left(\omega\odot \bar{x}_t\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot \bar{x}_t\right)\right\|^2\right]\\
&\quad+5\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot \bar{x}_t\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')}\right)\right\|^2\right]\\
&\quad+5\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')};\xi\right)\right\|^2\right]\\
&\leqslant 5K\sigma_g^2(\omega)+5L_g^2\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_t^{(k)}-\bar{x}_t\right\|^2\right]+5\gamma_g^2(\omega)+5\frac{1}{K}\sum_{k=1}^{K}\sum_{k'=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k')}\left(\omega\odot \bar{x}_t\right)-\nabla g^{(k')}\left(\omega\odot x_t^{(k')}\right)\right\|^2\right]\\
&\quad+5\frac{1}{K}\sum_{k=1}^{K}\sum_{k'=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k')}\left(\omega\odot x_t^{(k')}\right)-\nabla g^{(k')}\left(\omega\odot x_t^{(k')};\xi\right)\right\|^2\right]\\
&\leqslant 5K\sigma_g^2(\omega)+5L_g^2\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_t^{(k)}-\bar{x}_t\right\|^2\right]+5\gamma_g^2(\omega)+5\frac{L_g^2}{K}\sum_{k=1}^{K}\sum_{k'=1}^{K}\mathbb{E}\left[\left\|\bar{x}_t-x_t^{(k')}\right\|^2\right]+5K\sigma_g^2(\omega)\\
&=10K\sigma_g^2(\omega)+10L_g^2\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_t^{(k)}-\bar{x}_t\right\|^2\right]+5\gamma_g^2(\omega). \qquad (25)
\end{aligned}
$$

Here, the first equality adds and subtracts $\nabla g^{(k)}(\omega\odot x_t^{(k)})$, $\nabla g^{(k)}(\omega\odot\bar{x}_t)$, $\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot\bar{x}_t)$ and $\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\omega\odot x_t^{(k')})$. The first inequality uses $\|\sum_{i=1}^{5}a_i\|^2\leqslant 5\sum_{i=1}^{5}\|a_i\|^2$. The second inequality uses Assumption 4.3 on the first term. On the second term it uses Assumption 4.1 together with Jensen's inequality. On the third term it uses Lemma D.1 at the point $\bar{x}_t$, and on the fourth and fifth terms it uses Jensen's inequality for the average over k'. The third inequality applies the same Lipschitz property to the fourth term and Assumption 4.3 to the fifth term. The last equality uses $\frac{1}{K}\sum_{k=1}^{K}\sum_{k'=1}^{K}\|\bar{x}_t-x_t^{(k')}\|^2=\sum_{k=1}^{K}\|x_t^{(k)}-\bar{x}_t\|^2$.

2. We now estimate the second term on the right side of inequality (24). For any k and any y, we have

$$
\left\|\nabla f^{(k)}(y)\right\|^2=\left\|\mathbb{E}_\zeta\left[\nabla f^{(k)}(y;\zeta)\right]\right\|^2\leqslant\mathbb{E}_\zeta\left[\left\|\nabla f^{(k)}(y;\zeta)\right\|^2\right]\leqslant\mathbb{E}_\zeta\left[\sup_{y'}\left\|\nabla f^{(k)}(y';\zeta)\right\|^2\right]\leqslant C_f^2. \qquad (26)
$$

Hence,

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k)}\left(u_{t+1}^{(k)},s_{t+1}^{(k)}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}\left(u_{t+1}^{(k')},s_{t+1}^{(k')}\right)\right\|^2\right]\\
&=\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k)}\left(u_{t+1}^{(k)},s_{t+1}^{(k)}\right)\right\|^2\right]-K\,\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}\left(u_{t+1}^{(k')},s_{t+1}^{(k')}\right)\right\|^2\right]\\
&\leqslant\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k)}\left(u_{t+1}^{(k)},s_{t+1}^{(k)}\right)\right\|^2\right]\\
&\leqslant 3\gamma_f^2+30L_f^2\gamma_g^2(\omega)+60KL_f^2\sigma_g^2(\omega)+6L_f^2(10L_g^2+1)\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_{t+1}^{(k)}-\bar{x}_{t+1}\right\|^2\right]. \qquad (27)
\end{aligned}
$$

Substitute (25) and (27) into (24), we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|z_t^{(k)}-\bar{z}_t\right\|^2\right]\\
&\leqslant 14(C_f^2+\sigma_f^2)\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla g^{(k)}\left(\omega\odot x_t^{(k)};\xi\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}\left(\omega\odot x_t^{(k')};\xi\right)\right\|^2\right]\\
&\quad+(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left\{\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f^{(k)}\left(u_{t+1}^{(k)},s_{t+1}^{(k)}\right)-\frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}\left(u_{t+1}^{(k')},s_{t+1}^{(k')}\right)\right\|^2\right]\right\}\\
&\quad+14K\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]\\
&\leqslant 14(C_f^2+\sigma_f^2)\left\{10K\sigma_g^2(\omega)+10L_g^2\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_t^{(k)}-\bar{x}_t\right\|^2\right]+5\gamma_g^2(\omega)\right\}\\
&\quad+(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left\{3\gamma_f^2+30L_f^2\gamma_g^2(\omega)+60KL_f^2\sigma_g^2(\omega)+6L_f^2(10L_g^2+1)\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_{t+1}^{(k)}-\bar{x}_{t+1}\right\|^2\right]\right\}\\
&\quad+14K\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]\\
&=14(C_f^2+\sigma_f^2)\left(10K\sigma_g^2(\omega)+5\gamma_g^2(\omega)\right)+(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left(3\gamma_f^2+30L_f^2\gamma_g^2(\omega)+60KL_f^2\sigma_g^2(\omega)\right)\\
&\quad+14K\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]+140L_g^2(C_f^2+\sigma_f^2)\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_t^{(k)}-\bar{x}_t\right\|^2\right]\\
&\quad+6L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_{t+1}^{(k)}-\bar{x}_{t+1}\right\|^2\right]. \qquad (28)
\end{aligned}
$$

Thus we complete our proof. □

D.2 Formal proof of Proposition 4.5

Proof. Recall that $\bar{x}_t := \frac{1}{K}\sum_{k=1}^{K}x_t^{(k)}$ and $\bar{z}_t := \frac{1}{K}\sum_{k=1}^{K}z_t^{(k)}$ for every t, and $D_t := \sum_{k=1}^{K}\mathbb{E}\big[\|x_t^{(k)}-\bar{x}_t\|^2\big]$. In Algorithm 1, the averaging step replaces every $x_{t+1}^{(k)}$ by $\bar{x}_{t+1}$, and hence $\bar{x}_{t+1}=\bar{x}_t-\eta\bar{z}_t$ for all t. Moreover, synchronization happens at every multiple of p, and $x_0^{(k)}=x_0$ for all k, so

$$
x_{sp}^{(k)}=\bar{x}_{sp}\quad\text{for all } k \text{ and all integers } s\geqslant 0,\qquad \text{i.e.,}\quad D_{sp}=0. \qquad (29)
$$

Denoting $s_t=\left\lfloor\frac{t}{p}\right\rfloor$, no synchronization occurs at the rounds $s_tp+1,\dots,t$. For the consensus error $\left\|\bar{x}_t-x_t^{(k)}\right\|^2$, one can get

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|\bar{x}_t-x_t^{(k)}\right\|^2\right]\\
&=\sum_{k=1}^{K}\mathbb{E}\left[\left\|\bar{x}_{s_tp}-\eta\sum_{t'=s_tp}^{t-1}\bar{z}_{t'}-\left(x_{s_tp}^{(k)}-\eta\sum_{t'=s_tp}^{t-1}z_{t'}^{(k)}\right)\right\|^2\right]\\
&=\eta^2\sum_{k=1}^{K}\mathbb{E}\left[\left\|\sum_{t'=s_tp}^{t-1}\left(z_{t'}^{(k)}-\bar{z}_{t'}\right)\right\|^2\right]\\
&\leqslant p\eta^2\sum_{k=1}^{K}\sum_{t'=s_tp}^{t-1}\mathbb{E}\left[\left\|z_{t'}^{(k)}-\bar{z}_{t'}\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant 14Kp^2\eta^2(C_f^2+\sigma_f^2)\left(10K\sigma_g^2(\omega)+5\gamma_g^2(\omega)\right)+Kp^2\eta^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left(3\gamma_f^2+30L_f^2\gamma_g^2(\omega)+60KL_f^2\sigma_g^2(\omega)\right)\\
&\quad+14K^2p^2\eta^2\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]+140p\eta^2L_g^2(C_f^2+\sigma_f^2)\sum_{k=1}^{K}\sum_{t'=s_tp}^{t-1}\mathbb{E}\left[\left\|x_{t'}^{(k)}-\bar{x}_{t'}\right\|^2\right]\\
&\quad+6p\eta^2L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)\sum_{k=1}^{K}\sum_{t'=s_tp}^{t-1}\mathbb{E}\left[\left\|x_{t'+1}^{(k)}-\bar{x}_{t'+1}\right\|^2\right]\\
&=14Kp^2\eta^2(C_f^2+\sigma_f^2)\left(10K\sigma_g^2(\omega)+5\gamma_g^2(\omega)\right)+Kp^2\eta^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left(3\gamma_f^2+30L_f^2\gamma_g^2(\omega)+60KL_f^2\sigma_g^2(\omega)\right)\\
&\quad+14K^2p^2\eta^2\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]+140p\eta^2L_g^2(C_f^2+\sigma_f^2)\sum_{k=1}^{K}\sum_{t'=s_tp}^{t-1}\mathbb{E}\left[\left\|x_{t'}^{(k)}-\bar{x}_{t'}\right\|^2\right]\\
&\quad+6p\eta^2L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)\sum_{k=1}^{K}\sum_{t'=s_tp}^{t-1}\mathbb{E}\left[\left\|x_{t'}^{(k)}-\bar{x}_{t'}\right\|^2\right]\\
&\quad+6p\eta^2L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_{t}^{(k)}-\bar{x}_{t}\right\|^2\right]. \qquad (30)
\end{aligned}
$$

Here, the first equality unrolls the local updates $x_{t'+1}^{(k)}=x_{t'}^{(k)}-\eta z_{t'}^{(k)}$ and the averaged recursion $\bar{x}_{t'+1}=\bar{x}_{t'}-\eta\bar{z}_{t'}$ from round s_tp to round t. The second equality uses (29), i.e., $\bar{x}_{s_tp}=x_{s_tp}^{(k)}$. The first inequality is based on Jensen's inequality, $\|\sum_{i=1}^{n}a_i\|^2\leqslant n\sum_{i=1}^{n}\|a_i\|^2$, with $n=t-s_tp\leqslant p$. The second inequality substitutes (22) of Lemma D.3 into each summand t'. The three constant terms of (22) are summed over at most p values of t', which gives the factor $p^2\eta^2$, and are then multiplied by $K\geqslant 1$ to keep a uniform K-scaling across the terms. The last equality shifts the index in the last sum, $\sum_{t'=s_tp}^{t-1}D_{t'+1}=\sum_{t'=s_tp}^{t-1}D_{t'}+D_t-D_{s_tp}$, and uses $D_{s_tp}=0$ from (29).

By summing t from 0 to $T-1$, one can get

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\bar{x}_t-x_t^{(k)}\right\|^2\right]\\
&\leqslant 14TKp^2\eta^2(C_f^2+\sigma_f^2)\left(10K\sigma_g^2(\omega)+5\gamma_g^2(\omega)\right)\\
&\quad+TKp^2\eta^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left(3\gamma_f^2+30L_f^2\gamma_g^2(\omega)+60KL_f^2\sigma_g^2(\omega)\right)\\
&\quad+14TK^2p^2\eta^2\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]+140p^2\eta^2L_g^2(C_f^2+\sigma_f^2)\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_t^{(k)}-\bar{x}_t\right\|^2\right]\\
&\quad+6p^2\eta^2L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_t^{(k)}-\bar{x}_t\right\|^2\right]\\
&\quad+6p\eta^2L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\left[\left\|x_t^{(k)}-\bar{x}_t\right\|^2\right]. \qquad (31)
\end{aligned}
$$

Here, the T-factors on the constants come from summing T bounds. For the two consensus-error terms that carry the inner sum $\sum_{t'=s_tp}^{t-1}$, we switch the order of summation. That is, a fixed round t' appears in this inner sum only for $t\in\{t'+1,\dots,(s_{t'}+1)p-1\}$. Hence $\sum_{t=0}^{T-1}\sum_{t'=s_tp}^{t-1}D_{t'}\leqslant p\sum_{t=0}^{T-1}D_t$, which converts their pre-factor $p\eta^2$ into $p^2\eta^2$. The last term of (30) has no inner sum, so it keeps the pre-factor $p\eta^2$.

By setting

$$
\eta^2\leqslant c_1(\omega)=\frac{1}{280p^2L_g^2(C_f^2+\sigma_f^2)+12(p+p^2)L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)}, \qquad (32)
$$

one can get

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\bar{x}_t - x_t^{(k)}\right\|^2\right] \\
&\leqslant 2\Big\{14TKp^2\eta^2(C_f^2+\sigma_f^2)\left(10K\sigma_g^2(\omega)+5\gamma_g^2(\omega)\right)+TKp^2\eta^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)\left(3\gamma_f^2+30L_f^2\gamma_g^2(\omega)+60KL_f^2\sigma_g^2\right) \\
&\quad + 14TK^2p^2\eta^2\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]\Big\} \\
&= 140TKp^2\eta^2(C_f^2+\sigma_f^2)\left(2K\sigma_g^2(\omega)+\gamma_g^2(\omega)\right)+28TK^2p^2\eta^2\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right] \\
&\quad + 42TKp^2\eta^2(2(C_g^2(\omega)+\sigma_g^2(\omega))+1)\left(\gamma_f^2+10L_f^2\gamma_g^2(\omega)+20KL_f^2\sigma_g^2(\omega)\right).
\end{aligned}
\tag{33}
$$

Here, the inequality moves the three consensus-error terms of (31) to the left-hand side. Their total coefficient is $a := \eta^2\left[140p^2L_g^2(C_f^2+\sigma_f^2)+6(p^2+p)L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)\right]$, and (32) gives $a \leqslant \frac{1}{2}$. Since the left-hand side is finite, $(1-a)\sum_{t=0}^{T-1} D_t$ is bounded by the constant terms, and $\frac{1}{1-a} \leqslant 2$.

Thus we finish our proof. □

E PROOF OF THEOREM 4.6

Proof. Due to the L_F-smoothness of $F(x)$, one can get

$$\begin{aligned} & F\left(\bar{x}_{t+1}\right) \\ \leqslant & F\left(\bar{x}_t\right)+\left\langle \nabla F\left(\bar{x}_t\right), \bar{x}_{t+1}-\bar{x}_t\right\rangle+\frac{L_F}{2}\left\|\bar{x}_{t+1}-\bar{x}_t\right\|^2 \\ = & F\left(\bar{x}_t\right)-\eta\left\langle \nabla F\left(\bar{x}_t\right), \frac{1}{K} \sum_{k=1}^K z_t^{(k)}\right\rangle+\frac{\eta^2 L_F}{2}\left\|\frac{1}{K} \sum_{k=1}^K z_t^{(k)}\right\|^2 . \end{aligned} \tag{34}$$

Here, the inequality is based on the L_F-smoothness of F (Lemma 4.4), and the equality substitutes $\bar{x}_{t+1}-\bar{x}_t=-\frac{\eta}{K} \sum_{k=1}^K z_t^{(k)}$.

Taking expectations, one can get

$$\begin{aligned} & \mathbb{E}\left[F\left(\bar{x}_{t+1}\right)\right] \\ \leqslant & \mathbb{E}\left[F\left(\bar{x}_t\right)\right]-\eta \mathbb{E}\left\langle \nabla F\left(\bar{x}_t\right), \frac{1}{K} \sum_{k=1}^K \nabla F^{(k)}(x_t^{(k)})\right\rangle+\frac{\eta^2 L_F}{2} \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^K z_t^{(k)}\right\|^2\right] \\ = & \mathbb{E}\left[F\left(\bar{x}_t\right)\right]-\frac{\eta}{2} \mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right]-\frac{\eta}{2} \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^K \nabla F^{(k)}(x_t^{(k)})\right\|^2\right] \\ & +\frac{\eta}{2} \mathbb{E}\left[\left\|\nabla F(\bar{x}_t)-\frac{1}{K} \sum_{k=1}^K \nabla F^{(k)}(x_t^{(k)})\right\|^2\right]+\frac{\eta^2 L_F}{2} \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^K z_t^{(k)}\right\|^2\right] \\ = & \mathbb{E}\left[F\left(\bar{x}_t\right)\right]-\frac{\eta}{2} \mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right]+\frac{\eta}{2}\left(\eta L_F-1\right) \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^K \nabla F^{(k)}(x_t^{(k)})\right\|^2\right] \\ & +\frac{\eta}{2} \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^K\left(\nabla F^{(k)}(\bar{x}_t)-\nabla F^{(k)}(x_t^{(k)})\right)\right\|^2\right]+\frac{\eta^2 L_F}{2} \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^K\left(z_t^{(k)}-\nabla F^{(k)}(x_t^{(k)})\right)\right\|^2\right] \\ \leqslant & \mathbb{E}\left[F\left(\bar{x}_t\right)\right]-\frac{\eta}{2} \mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right]+\frac{\eta}{2}\left(\eta L_F-1\right) \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^K \nabla F^{(k)}(x_t^{(k)})\right\|^2\right] \\ & +\frac{\eta}{2} \frac{1}{K} \sum_{k=1}^K \mathbb{E}\left[\left\|\nabla F^{(k)}(\bar{x}_t)-\nabla F^{(k)}(x_t^{(k)})\right\|^2\right]+\frac{\eta^2 L_F}{2} \frac{1}{K} \sum_{k=1}^K \mathbb{E}\left[\left\|z_t^{(k)}-\nabla F^{(k)}(x_t^{(k)})\right\|^2\right] \\ \leqslant & \mathbb{E}\left[F\left(\bar{x}_t\right)\right]-\frac{\eta}{2} \mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right]+\frac{\eta}{2}\left(\eta L_F-1\right) \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^K \nabla F^{(k)}(x_t^{(k)})\right\|^2\right] \\ & +\frac{\eta}{2} \frac{L_F^2}{K} \sum_{k=1}^K \mathbb{E}\left[\left\|\bar{x}_t-x_t^{(k)}\right\|^2\right]+\frac{\eta^2 L_F \sigma_F^2(\omega)}{2} . \end{aligned} \tag{35}$$

Here, the first inequality takes expectations in (34) and uses the tower property. The second inequality is based on Jensen's inequality, $\|\frac{1}{K} \sum_{k=1}^K a_k\|^2 \leqslant \frac{1}{K} \sum_{k=1}^K\|a_k\|^2$, applied to the fourth and fifth terms. The last inequality uses the L_F-smoothness of each $F^{(k)}$ (Lemma 4.4) on the fourth term, and the tower property on the fifth term.

Let $\eta \leqslant \frac{1}{L_F}$. Then $\eta L_F-1 \leqslant 0$, and one can get

$$\begin{aligned} & \mathbb{E}\left[F\left(\bar{x}_{t+1}\right)\right] \\ \leqslant & \mathbb{E}\left[F\left(\bar{x}_t\right)\right]-\frac{\eta}{2} \mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right]+\frac{\eta}{2} \frac{L_F^2}{K} \sum_{k=1}^K \mathbb{E}\left[\left\|\bar{x}_t-x_t^{(k)}\right\|^2\right]+\frac{\eta^2 L_F \sigma_F^2(\omega)}{2} . \end{aligned} \tag{36}$$

Here, the inequality drops the third term of (35), which is non-positive.

By summing t from 0 to $T-1$, one can get

$$\begin{aligned}&\mathbb{E}\left[F\left(\bar{x}_T\right)\right]\\&\leqslant F\left(\bar{x}_0\right)-\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{x}_t)\right\|^2\right]+\frac{\eta}{2}\frac{L_F^2}{K}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\bar{x}_t-x_t^{(k)}\right\|^2\right]+\frac{\eta^2 L_F\sigma_F^2(\omega)T}{2}.\end{aligned}\tag{37}$$

Here, the inequality telescopes $\mathbb{E}[F(\bar{x}_{t+1})]-\mathbb{E}[F(\bar{x}_t)]$ over t.

Hence,

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\bar{x}_t)\right\|^2\right]\\&\leqslant\frac{2\left(F\left(\bar{x}_0\right)-\mathbb{E}\left[F\left(\bar{x}_T\right)\right]\right)}{\eta T}+\frac{L_F^2}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\bar{x}_t-x_t^{(k)}\right\|^2\right]+\eta L_F\sigma_F^2(\omega)\\&\leqslant\frac{2\left(F\left(\bar{x}_0\right)-F^*\right)}{\eta T}+\frac{L_F^2}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\bar{x}_t-x_t^{(k)}\right\|^2\right]+\eta L_F\sigma_F^2(\omega).\end{aligned}\tag{38}$$

Here, the first inequality rearranges (37) and multiplies both sides by $\frac{2}{\eta T}$. The second inequality uses $\mathbb{E}[F(\bar{x}_T)]\geqslant F^*$.

In addition, by denoting

$$\kappa\triangleq\min\left\{\frac{1}{\sqrt{280p^2L_g^2(C_f^2+\sigma_f^2)+12(p+p^2)L_f^2(14(C_g^2(\omega)+\sigma_g^2(\omega))+7)(10L_g^2+1)}},\frac{1}{L_F}\right\},\tag{39}$$

if $\eta\leqslant\kappa$, we can get

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F\left(\bar{x}_t\right)\right\|^2\right]\\&\leqslant\frac{2(F(\bar{x}_0)-F^*)}{\eta T}+\eta L_F\sigma_F^2(\omega)+\frac{L_F^2}{KT}\Big\{140TKp^2\eta^2(C_f^2+\sigma_f^2)\left(2K\sigma_g^2(\omega)+\gamma_g^2(\omega)\right)\\&\quad+28TK^2p^2\eta^2\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]\\&\quad+42TKp^2\eta^2\left(2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right)\left(\gamma_f^2+10L_f^2\gamma_g^2(\omega)+20KL_f^2\sigma_g^2(\omega)\right)\Big\}\\&=\frac{2(F(\bar{x}_0)-F^*)}{\eta T}+\eta L_F\sigma_F^2(\omega)+140L_F^2p^2\eta^2(C_f^2+\sigma_f^2)\left(2K\sigma_g^2(\omega)+\gamma_g^2(\omega)\right)\\&\quad+28KL_F^2p^2\eta^2\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]\\&\quad+42L_F^2p^2\eta^2\left(2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right)\left(\gamma_f^2+10L_f^2\gamma_g^2(\omega)+20KL_f^2\sigma_g^2(\omega)\right).\end{aligned}\tag{40}$$

Here, the inequality applies (38) since $\eta\leqslant\kappa\leqslant\frac{1}{L_F}$, and substitutes Proposition 4.5, which is valid since $\eta^2\leqslant\kappa^2\leqslant c_1(\omega)$ (see (32)). Then this proves (12).

Let $\eta = \frac{\kappa}{\sqrt{T}}$, which satisfies $\eta \leqslant \kappa$ for all $T \geqslant 1$. Then we can get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F\left(\bar{x}_t\right)\right\|^2\right] \\
&\leqslant \frac{2(F(\bar{x}_0)-F^*)}{\kappa\sqrt{T}} + \frac{\kappa L_F \sigma_F^2(\omega)}{\sqrt{T}} + \frac{140L_F^2p^2\kappa^2(C_f^2+\sigma_f^2)\left(2K\sigma_g^2(\omega)+\gamma_g^2(\omega)\right)}{T} \\
&\quad + \frac{28KL_F^2p^2\kappa^2\sigma_f^2\left[2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right]}{T} \\
&\quad + \frac{42L_F^2p^2\kappa^2\left(2(C_g^2(\omega)+\sigma_g^2(\omega))+1\right)\left(\gamma_f^2+10L_f^2\gamma_g^2(\omega)+20KL_f^2\sigma_g^2(\omega)\right)}{T}.
\end{aligned} \tag{41}$$

Here, the inequality substitutes $\eta = \kappa/\sqrt{T}$ into (40). Since $F(\bar{x}_0) - F^*$, κ, and all other constants do not depend on T, the first two terms are $O(1/\sqrt{T})$ and the last three are $O(1/T)$, i.e., $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F\left(\bar{x}_t\right)\right\|^2\right]$ has the convergence rate $O(\frac{1}{\sqrt{T}})$.

Thus we complete our proof. □

F PROOF OF THEOREM 4.7

The proof of Theorem 4.6 in Appendix E relies on two facts, i.e., the averaging step preserves $\bar{x}_t$, and each local stochastic gradient is evaluated at a point that is known at the start of the iteration. **Neither holds for SFL-V2.** The server-side model is handed off from client to client, so the point $y_t^{(k)}$ used by client k depends on the minibatches of clients $1, \dots, k-1$ drawn in the same iteration. Moreover, line 11 of Algorithm 2 overwrites the server-side blocks, so it changes $\bar{x}_t$. Hence we need to work with the synchronized model $\hat{x}_t = \omega \odot \bar{x}_t + (\mathbf{1} - \omega) \odot x_t^{(K)}$ and a filtration that follows the client order. Here is a roadmap.

- *Preliminaries.* We define the block scaling vector $\mathbf{d}_K = \omega + K(\mathbf{1} - \omega)$ and the filtration $\{\mathcal{G}_t^{(k)}\}$. We show $z_t^{(k)}$ is conditionally unbiased and bound its conditional second moments.
- *Step 1.* We show that $\hat{x}_t$ follows the exact recursion $\hat{x}_{t+1} = \hat{x}_t - \eta\, \mathbf{d}_K \odot \bar{z}_t$ in Eq. 46, and that the synchronization step does not change $\hat{x}_t$. This is where the sequential handoff enters, i.e., the client-side block moves by the averaged gradient, while the server-side block moves by the sum of the K sequential gradients.
- *Step 2.* We apply the L_F-smoothness of F to this recursion in Eq. 47.
- *Step 3.* We lower bound the inner-product term in Eq. 48. This yields the descent term $\frac{1}{2}\mathbb{E}\|\nabla F(\hat{x}_t)\|^2$ minus the coupling error $\frac{L_F^2}{2}\sum_k \mathbb{E}\|y_t^{(k)} - \hat{x}_t\|^2$.
- *Step 4.* We bound the quadratic term by $(C_g^2(\omega) + K^2)C_f^2$ in Eq. 49. The factor K^2 comes from the server-side block, which is scaled by K.
- *Step 5.* We bound the coupling error in Eq. 50-Eq. 54.
- *Steps 6 and 7.* We combine the above bounds into a one-iteration inequality in Eq. 55, telescope it over t in Eq. 56.

Proof. Preliminaries. We follow the indexing of Eq. 7-Eq. 8 and Algorithm 2. In iteration t, client k draws a minibatch $\xi_t^{(k)}$, forms $y_t^{(k)} = \omega \odot x_t^{(k)} + (\mathbf{1} - \omega) \odot x_{t+1}^{(k-1)}$ with $x_{t+1}^{(0)} := x_t^{(K)}$, computes $z_t^{(k)} := \nabla_x F^{(k)}(y_t^{(k)}; \xi_t^{(k)}, \zeta_t^{(k)})$, and sets $x_{t+1}^{(k)} = y_t^{(k)} - \eta z_t^{(k)}$. By the masking, we have

$$\omega \odot x_{t+1}^{(k)} = \omega \odot x_t^{(k)} - \eta\, \omega \odot z_t^{(k)}, (\mathbf{1} - \omega) \odot x_{t+1}^{(k)} = (\mathbf{1} - \omega) \odot x_{t+1}^{(k-1)} - \eta\, (\mathbf{1} - \omega) \odot z_t^{(k)}. \tag{42}$$

We write $\bar{z}_t := \frac{1}{K}\sum_{k=1}^K z_t^{(k)}$ and define the block scaling vector

$$\mathbf{d}_K := \omega + K(\mathbf{1} - \omega) \in \mathbb{R}^n, \qquad \mathbf{d}_K^{1/2} := \omega + \sqrt{K}(\mathbf{1} - \omega), \tag{43}$$

so that $\mathbf{d}_K \odot a$ keeps the client-side block of a and multiplies its server-side block by K. Since $\omega \in \{0,1\}^n$, the two blocks are orthogonal, i.e., $\|a\|^2 = \|\omega \odot a\|^2 + \|(\mathbf{1} - \omega) \odot a\|^2$ for any $a \in \mathbb{R}^n$. As a consequence, $\|a\|^2 \leqslant \|\mathbf{d}_K^{1/2} \odot a\|^2 \leqslant K\|a\|^2$ and $\|\mathbf{d}_K \odot a\|^2 = \|\omega \odot a\|^2 + K^2\|(\mathbf{1} - \omega) \odot a\|^2$.

We can derive that

$$\begin{aligned} &\mathbb{E}_t^{(k)}\big[z_t^{(k)}\big] = \nabla F^{(k)}\big(y_t^{(k)}\big), \qquad \mathbb{E}_t^{(k)}\Big[\big\|\mathbf{d}_K \odot z_t^{(k)}\big\|^2\Big] \leqslant \big(C_g^2(\omega) + K^2\big)C_f^2, \\ &\mathbb{E}_t^{(k)}\Big[\big\|\omega \odot z_t^{(k)}\big\|^2\Big] \leqslant C_g^2(\omega)C_f^2, \qquad \mathbb{E}_t^{(k)}\Big[\big\|(\mathbf{1} - \omega) \odot z_t^{(k)}\big\|^2\Big] \leqslant C_f^2. \end{aligned} \tag{44}$$

Then

$$\begin{aligned} \big\|\mathbf{d}_K \odot z_t^{(k)}\big\|^2 &= \big\|\omega \odot z_t^{(k)}\big\|^2 + K^2\big\|(\mathbf{1} - \omega) \odot z_t^{(k)}\big\|^2 \\ &= \Big\|\omega \odot \big[\nabla g^{(k)}(\omega \odot y_t^{(k)}; \xi_t^{(k)})\big]^\top \nabla_g f^{(k)}\big(u, s; \zeta_t^{(k)}\big)\Big\|^2 \\ &\quad + K^2\Big\|(\mathbf{1} - \omega) \odot \nabla_{x_s} f^{(k)}\big(u, s; \zeta_t^{(k)}\big)\Big\|^2 \\ &\leqslant C_g^2(\omega)\Big\|\nabla_g f^{(k)}\big(u, s; \zeta_t^{(k)}\big)\Big\|^2 + K^2\Big\|\nabla_{x_s} f^{(k)}\big(u, s; \zeta_t^{(k)}\big)\Big\|^2 \end{aligned}$$

$$\leqslant \left(C_g^2(\omega) + K^2\right) \left\| \nabla f^{(k)}\left(u, s; \zeta_t^{(k)}\right) \right\|^2$$
$$\leqslant \left(C_g^2(\omega) + K^2\right) \sup_y \left\| \nabla f^{(k)}\left(y; \zeta_t^{(k)}\right) \right\|^2. \tag{45}$$

Here, the first inequality uses $\|\omega \odot a\| \leqslant \|a\|$, $\|A^\top v\| \leqslant \|A\| \, \|v\|$, and $\|\nabla g^{(k)}(\omega \odot y; \xi)\| \leqslant C_g(\omega)$ from Assumption 4.2 for the first term, and $\|(\mathbf{1} - \omega) \odot a\| \leqslant \|a\|$ for the second term. The second inequality uses $\|\nabla_g f^{(k)}\|^2 + \|\nabla_{x_s} f^{(k)}\|^2 = \|\nabla f^{(k)}\|^2$, since $\nabla f^{(k)}$ is the joint gradient with respect to (u, x_s). The last inequality takes the supremum over the argument.

Step 1. Before the synchronization step, we have

$$\begin{aligned}
\hat{x}_{t+1} &= \omega \odot \bar{x}_{t+1} + (\mathbf{1} - \omega) \odot x_{t+1}^{(K)} \\
&= \frac{1}{K} \sum_{k=1}^{K} \left(\omega \odot x_t^{(k)} - \eta \, \omega \odot z_t^{(k)} \right) + (\mathbf{1} - \omega) \odot x_{t+1}^{(K-1)} - \eta \, (\mathbf{1} - \omega) \odot z_t^{(K)} \\
&= \omega \odot \bar{x}_t - \eta \, \omega \odot \bar{z}_t + (\mathbf{1} - \omega) \odot x_{t+1}^{(0)} - \eta \sum_{k=1}^{K} (\mathbf{1} - \omega) \odot z_t^{(k)} \\
&= \omega \odot \bar{x}_t + (\mathbf{1} - \omega) \odot x_t^{(K)} - \eta \Big(\omega \odot \bar{z}_t + K(\mathbf{1} - \omega) \odot \bar{z}_t \Big) \\
&= \hat{x}_t - \eta \, \mathbf{d}_K \odot \bar{z}_t.
\end{aligned} \tag{46}$$

Here, the third equality unrolls the server-side update from $k = K$ down to $k = 1$. The fourth equality uses the convention $x_{t+1}^{(0)} = x_t^{(K)}$ and $\sum_{k=1}^{K} z_t^{(k)} = K \bar{z}_t$. The last equality is by definition. Note that the synchronization step does change $\bar{x}_{t+1}$, since it overwrites the server-side blocks by that of $x_{t+1}^{(K)}$.

Step 2. Due to the L_F-smoothness of F, one can get

$$\begin{aligned}
F\left(\hat{x}_{t+1}\right) &\leqslant F\left(\hat{x}_t\right) + \left\langle \nabla F\left(\hat{x}_t\right), \hat{x}_{t+1} - \hat{x}_t \right\rangle + \frac{L_F}{2} \left\| \hat{x}_{t+1} - \hat{x}_t \right\|^2 \\
&= F\left(\hat{x}_t\right) - \eta \left\langle \nabla F\left(\hat{x}_t\right), \mathbf{d}_K \odot \bar{z}_t \right\rangle + \frac{\eta^2 L_F}{2} \left\| \mathbf{d}_K \odot \bar{z}_t \right\|^2.
\end{aligned} \tag{47}$$

Step 3. Let $\bar{g}_t := \frac{1}{K} \sum_{k=1}^{K} \nabla F^{(k)}(y_t^{(k)})$. One can get

$$\begin{aligned}
&\mathbb{E} \left\langle \nabla F(\hat{x}_t), \mathbf{d}_K \odot \bar{z}_t \right\rangle \\
&= \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left\langle \mathbf{d}_K \odot \nabla F(\hat{x}_t), \mathbb{E}_t^{(k)}\left[z_t^{(k)}\right] \right\rangle \\
&= \mathbb{E} \left\langle \mathbf{d}_K \odot \nabla F(\hat{x}_t), \bar{g}_t \right\rangle \\
&= \mathbb{E} \left[\left\| \mathbf{d}_K^{1/2} \odot \nabla F(\hat{x}_t) \right\|^2 \right] + \mathbb{E} \left\langle \mathbf{d}_K^{1/2} \odot \nabla F(\hat{x}_t), \mathbf{d}_K^{1/2} \odot \left(\bar{g}_t - \nabla F(\hat{x}_t) \right) \right\rangle \\
&\geqslant \frac{1}{2} \mathbb{E} \left[\left\| \mathbf{d}_K^{1/2} \odot \nabla F(\hat{x}_t) \right\|^2 \right] - \frac{1}{2} \mathbb{E} \left[\left\| \mathbf{d}_K^{1/2} \odot \left(\bar{g}_t - \nabla F(\hat{x}_t) \right) \right\|^2 \right] \\
&\geqslant \frac{1}{2} \mathbb{E} \left[\left\| \nabla F(\hat{x}_t) \right\|^2 \right] - \frac{K}{2} \mathbb{E} \left[\left\| \frac{1}{K} \sum_{k=1}^{K} \left(\nabla F^{(k)}\left(y_t^{(k)}\right) - \nabla F^{(k)}\left(\hat{x}_t\right) \right) \right\|^2 \right] \\
&\geqslant \frac{1}{2} \mathbb{E} \left[\left\| \nabla F(\hat{x}_t) \right\|^2 \right] - \frac{1}{2} \sum_{k=1}^{K} \mathbb{E} \left[\left\| \nabla F^{(k)}\left(y_t^{(k)}\right) - \nabla F^{(k)}\left(\hat{x}_t\right) \right\|^2 \right] \\
&\geqslant \frac{1}{2} \mathbb{E} \left[\left\| \nabla F(\hat{x}_t) \right\|^2 \right] - \frac{L_F^2}{2} \sum_{k=1}^{K} \mathbb{E} \left[\left\| y_t^{(k)} - \hat{x}_t \right\|^2 \right].
\end{aligned} \tag{48}$$

Here, the first equality uses $\langle a, \mathbf{d}_K \odot b \rangle = \langle \mathbf{d}_K \odot a, b \rangle$ and the tower property. The second equality substitutes $\mathbb{E}_t^{(k)}[z_t^{(k)}] = \nabla F^{(k)}(y_t^{(k)})$ from Eq. 44 and the definition of $\bar{g}_t$. The first

inequality uses Young's inequality, $\langle a, b\rangle \geqslant -\frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2$. The second inequality uses $\|a\|^2 \leqslant \|\mathbf{d}_K^{1/2} \odot a\|^2 \leqslant K\|a\|^2$ and the definition of F. The third inequality is based on Jensen's inequality, $\|\frac{1}{K}\sum_{k=1}^K a_k\|^2 \leqslant \frac{1}{K}\sum_{k=1}^K \|a_k\|^2$. The last inequality uses Lemma 4.4.

Step 4. One can get

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{d}_K \odot \bar{z}_t\right\|^2\right] &\leqslant \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\left\|\mathbf{d}_K \odot z_t^{(k)}\right\|^2\right] \\ &= \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\mathbb{E}_t^{(k)}\left\|\mathbf{d}_K \odot z_t^{(k)}\right\|^2\right] \\ &\leqslant \left(C_g^2(\omega) + K^2\right)C_f^2. \end{aligned} \tag{49}$$

Here, the first inequality is based on Jensen's inequality, and the second inequality uses Eq. 44.

Step 5. For each k, one can get

$$\begin{aligned} y_t^{(k)} - \hat{x}_t &= \omega \odot x_t^{(k)} + (\mathbf{1}-\omega) \odot x_{t+1}^{(k-1)} - \omega \odot \bar{x}_t - (\mathbf{1}-\omega) \odot x_t^{(K)} \\ &= \omega \odot \left(x_t^{(k)} - \bar{x}_t\right) + (\mathbf{1}-\omega) \odot \left(x_{t+1}^{(k-1)} - x_{t+1}^{(0)}\right) \\ &= \omega \odot \left(x_t^{(k)} - \bar{x}_t\right) - \eta \sum_{i=1}^{k-1} (\mathbf{1}-\omega) \odot z_t^{(i)}. \end{aligned} \tag{50}$$

By orthogonality of the two blocks,

$$\sum_{k=1}^K \mathbb{E}\left[\left\|y_t^{(k)} - \hat{x}_t\right\|^2\right] = \sum_{k=1}^K \mathbb{E}\left[\left\|\omega \odot \left(x_t^{(k)} - \bar{x}_t\right)\right\|^2\right] + \eta^2 \sum_{k=1}^K \mathbb{E}\left[\left\|\sum_{i=1}^{k-1}(\mathbf{1}-\omega) \odot z_t^{(i)}\right\|^2\right]. \tag{51}$$

For the server-side lag, one can get

$$\begin{aligned} \eta^2 \sum_{k=1}^K \mathbb{E}\left[\left\|\sum_{i=1}^{k-1}(\mathbf{1}-\omega) \odot z_t^{(i)}\right\|^2\right] &\leqslant \eta^2 \sum_{k=1}^K (k-1) \sum_{i=1}^{k-1} \mathbb{E}\left[\left\|(\mathbf{1}-\omega) \odot z_t^{(i)}\right\|^2\right] \\ &\leqslant \eta^2 C_f^2 \sum_{k=1}^K (k-1)^2 \\ &= \eta^2 C_f^2 \frac{(K-1)K(2K-1)}{6} \\ &\leqslant \frac{\eta^2 C_f^2 K^3}{3}. \end{aligned} \tag{52}$$

Here, the first inequality uses $\|\sum_{i=1}^m a_i\|^2 \leqslant m\sum_{i=1}^m \|a_i\|^2$. The second inequality uses the tower property and Eq. 44, and the last inequality uses $(K-1)(2K-1) \leqslant 2K^2$.

Related to the client drift, let $t_0 := p\lfloor t/p \rfloor$ be the most recent synchronization index. Hence, $\omega \odot (x_t^{(k)} - \bar{x}_t) = -\eta \sum_{\tau=t_0}^{t-1} \omega \odot (z_\tau^{(k)} - \bar{z}_\tau)$, and one can get

$$\begin{aligned} \sum_{k=1}^K \mathbb{E}\left[\left\|\omega \odot \left(x_t^{(k)} - \bar{x}_t\right)\right\|^2\right] &= \eta^2 \sum_{k=1}^K \mathbb{E}\left[\left\|\sum_{\tau=t_0}^{t-1} \omega \odot \left(z_\tau^{(k)} - \bar{z}_\tau\right)\right\|^2\right] \\ &\leqslant \eta^2 (t-t_0) \sum_{\tau=t_0}^{t-1} \sum_{k=1}^K \mathbb{E}\left[\left\|\omega \odot \left(z_\tau^{(k)} - \bar{z}_\tau\right)\right\|^2\right] \\ &\leqslant \eta^2 (t-t_0) \sum_{\tau=t_0}^{t-1} \sum_{k=1}^K \mathbb{E}\left[\left\|\omega \odot z_\tau^{(k)}\right\|^2\right] \end{aligned}$$

$$\leqslant \eta^2 (t-t_0)^2 K\, C_g^2(\omega) C_f^2$$
$$\leqslant \eta^2 p^2 K\, C_g^2(\omega) C_f^2. \tag{53}$$

Here, the first inequality uses $\| \sum_{i=1}^m a_i \|^2 \leqslant m \sum_{i=1}^m \|a_i\|^2$ with $m = t - t_0$. The second inequality uses $\sum_{k=1}^K \|a_k - \bar{a}\|^2 = \sum_{k=1}^K \|a_k\|^2 - K\|\bar{a}\|^2 \leqslant \sum_{k=1}^K \|a_k\|^2$ with $a_k = \omega \odot z_\tau^{(k)}$. The third inequality uses the tower property and Eq. 44. The last inequality uses $t - t_0 \leqslant p - 1 < p$.

Substituting Eq. 52 and Eq. 53 into Eq. 51, we have

$$\sum_{k=1}^K \mathbb{E}\left[\left\| y_t^{(k)} - \hat{x}_t \right\|^2 \right] \leqslant \eta^2 C_f^2 \Big(p^2 K\, C_g^2(\omega) + \frac{K^3}{3} \Big). \tag{54}$$

Step 6. Taking expectations in Eq. 47, one can get

$$\begin{aligned}
&\mathbb{E}\left[F\left(\hat{x}_{t+1}\right)\right] \\
&\leqslant \mathbb{E}\left[F\left(\hat{x}_t\right)\right] - \eta\, \mathbb{E}\left\langle \nabla F\left(\hat{x}_t\right), \mathbf{d}_K \odot \bar{z}_t \right\rangle + \frac{\eta^2 L_F}{2} \mathbb{E}\left[\left\| \mathbf{d}_K \odot \bar{z}_t \right\|^2 \right] \\
&\leqslant \mathbb{E}\left[F\left(\hat{x}_t\right)\right] - \frac{\eta}{2} \mathbb{E}\left[\left\| \nabla F(\hat{x}_t) \right\|^2 \right] + \frac{\eta L_F^2}{2} \sum_{k=1}^K \mathbb{E}\left[\left\| y_t^{(k)} - \hat{x}_t \right\|^2 \right] + \frac{\eta^2 L_F}{2} \big(C_g^2(\omega) + K^2 \big) C_f^2 \\
&\leqslant \mathbb{E}\left[F\left(\hat{x}_t\right)\right] - \frac{\eta}{2} \mathbb{E}\left[\left\| \nabla F(\hat{x}_t) \right\|^2 \right] + \frac{\eta^2 L_F}{2} C_f^2 \big(C_g^2(\omega) + K^2 \big) \\
&\quad + \frac{\eta^3 L_F^2}{2} C_f^2 \Big(p^2 K\, C_g^2(\omega) + \frac{K^3}{3} \Big).
\end{aligned} \tag{55}$$

Here, the first inequality takes expectations in Eq. 47. The second inequality substitutes Eq. 48 into the second term and Eq. 49 into the third term. The last inequality substitutes Eq. 54.

Step 7. By summing t from 0 to $T-1$, one can get

$$\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \nabla F(\hat{x}_t) \right\|^2 \right] \\
&\leqslant \frac{2\big(F(x_0) - \mathbb{E}[F(\hat{x}_T)]\big)}{\eta T} + \eta L_F C_f^2 \big(C_g^2(\omega) + K^2 \big) + \eta^2 L_F^2 C_f^2 \Big(p^2 K\, C_g^2(\omega) + \frac{K^3}{3} \Big) \\
&\leqslant \frac{2\big(F(x_0) - F^*\big)}{\eta T} + \eta L_F C_f^2 \big(C_g^2(\omega) + K^2 \big) + \eta^2 L_F^2 C_f^2 \Big(p^2 K\, C_g^2(\omega) + \frac{K^3}{3} \Big).
\end{aligned} \tag{56}$$

Here, the first inequality telescopes $\mathbb{E}[F(\hat{x}_{t+1})] - \mathbb{E}[F(\hat{x}_t)]$ over t, uses that $\hat{x}_0 = x_0$ is deterministic, rearranges, and multiplies both sides by $\frac{2}{\eta T}$. The second inequality uses $\mathbb{E}[F(\hat{x}_T)] \geqslant F^*$. This proves Eq. 13.

Let $\eta = \frac{\kappa}{\sqrt{T}}$ for a constant $\kappa > 0$. Then we can get

$$\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \nabla F(\hat{x}_t) \right\|^2 \right] \\
&\leqslant \frac{2\big(F(x_0) - F^*\big)}{\kappa \sqrt{T}} + \frac{\kappa L_F C_f^2 \big(C_g^2(\omega) + K^2 \big)}{\sqrt{T}} + \frac{\kappa^2 L_F^2 C_f^2 \big(p^2 K\, C_g^2(\omega) + K^3/3 \big)}{T}.
\end{aligned} \tag{57}$$

Hence, $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(\hat{x}_t)\|^2]$ has the convergence rate $O(\frac{1}{\sqrt{T}})$, and the leading constant scales with K^2. If η is kept constant instead, the right-hand side of Eq. 56 tends to $\eta L_F C_f^2 (C_g^2(\omega) + K^2) + \eta^2 L_F^2 C_f^2 (p^2 K C_g^2(\omega) + K^3/3)$ as $T \to \infty$, which is the residual discussed after Theorem 4.7.

Thus we complete our proof. □

G Detailed Comparison of SL/SFL Convergence Results

We provide a theorem-level comparison of representative nonconvex convergence results. In particular, we use the average-iterate stationarity criterion

$$\mathcal{G}_T := \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla F(\bar{x}_t)\|^2\big]. \tag{58}$$

Equivalently, $\mathcal{G}_T = \mathbb{E}\|\nabla F(\widehat{x})\|^2$ when $\widehat{x} = \bar{x}_R$ and R is sampled uniformly from $\{0,\dots,T-1\}$ independently of training. Rows marked by † instead state a best-iterate guarantee,

$$\mathcal{M}_T := \min_{0\le t<T}\mathbb{E}\big[\|\nabla F(x_t)\|^2\big]. \tag{59}$$

Note that although $\mathcal{M}_T \le \mathcal{G}_T$, a bound on $\mathcal{M}_T$ alone does not establish the same bound on $\mathcal{G}_T$.

Notation.

- T: the iteration unit, in our analysis it corresponds to number of rounds
- K: number of clients
- p: synchronization period
- L: the relevant smoothness constant in prior work, L_F: smoothness of our global objective
- $\Delta_F = F(x_0) - F_\star$: initial objective gap, where $F_\star$ is a lower bound on F
- σ^2: stochastic-gradient variance ζ^2: the relevant additive heterogeneity bound
- β: multiplicative heterogeneity parameter Li and Lyu (2023)
- ω: split mask representing the cut layer.

The iteration units are not identical among existing studies. For example, our T local update iterations correspond to approximately T/p model-averaging events, in addition to split-layer communication. For vanilla SL, one round includes a traversal of the participating clients. We suppress numerical constants and fixed method-specific quantities unless explicitly displayed. All bounds retain the assumptions and admissibility conditions of their respective source theorems.

For Han et al. (2024), with client weights $a_k \ge 0$ and $\sum_{k=1}^K a_k = 1$, define

$$V_1 = \sum_{k=1}^K a_k^2(\sigma_k^2+\zeta^2), \qquad V_2 = \sum_{k=1}^K (a_k^2+1)(\sigma_k^2+\zeta^2), \qquad A_1 = \sum_{k=1}^K a_k^2. \tag{60}$$

Their stepsize-weighted bounds are divided by a constant-within-run stepsize before optimization.

For Li and Lyu (2023); Liang et al. (2025), the heterogeneity condition is

$$\frac{1}{K}\sum_{k=1}^K \|\nabla F_k(x) - \nabla F(x)\|^2 \le \beta^2\|\nabla F(x)\|^2 + \zeta^2. \tag{61}$$

For Lin et al. (2025), define

$$S_\sigma = \sum_{\ell=1}^{L_{\text{layers}}}\sigma_\ell^2, \qquad D_{\text{H}} = \sum_{m=1}^{M-1}\mathbf{1}_{\{I_m>1\}} I_m^2 \sum_{\ell\in\mathcal{L}_m} G_\ell^2, \tag{62}$$

where I_m is the aggregation interval of tier m, $\mathcal{L}_m$ is its layer set, and G_ℓ^2 bounds the layerwise stochastic-gradient second moment. Thus, its synchronization-dependent coefficient is retained rather than hidden in big-O notation.

For Nair et al. (2025), $\mathcal{E}_T = \sum_{t=1}^T \varepsilon_t$ is the cumulative auxiliary-gradient estimation error. A vanishing rate requires control of $\mathcal{E}_T/T$, and a pure $O(T^{-1/2})$ rate requires this average error to be $O(T^{-1/2})$ or smaller.

Table 2: A unified theorem-level comparison of nonconvex convergence results. Unmarked rows bound $\mathcal{G}_T$, while rows marked by † bound $\mathcal{M}_T$.

Work	Normalized nonconvex convergence bound
Li and Lyu (2023): vanilla SL†	$O\left(\sqrt{\frac{L\Delta_F\sigma^2}{KpT}}+\left(\frac{L^2\Delta_F^2\sigma^2}{KpT^2}\right)^{1/3}+\left(\frac{L^2\Delta_F^2\zeta^2}{KT^2}\right)^{1/3}+\frac{L\Delta_F(1+\beta)}{T}\right)$
Li and Lyu (2025): vanilla SL†	$O\left(\sqrt{\frac{L\Delta_F\sigma^2}{KpT}}+\left(\frac{L^2\Delta_F^2\sigma^2}{KpT^2}\right)^{1/3}+\left(\frac{L^2\Delta_F^2\zeta^2}{KT^2}\right)^{1/3}+\frac{L\Delta_F(1+\beta^2/K)}{T}\right)$
Han et al. (2024): SFL-V1	$O\left(\sqrt{\frac{L\Delta_F KV_1}{T}}+\frac{L\Delta_F(1+KA_1)}{T}\right)$
Han et al. (2024): SFL-V2	$O\left(\sqrt{\frac{L\Delta_F KV_2}{T}}+\frac{L\Delta_F(1+K^2)}{T}\right)$
Lin et al. (2025): hierarchical SFL	$O\left(\sqrt{\frac{L\Delta_F S_\sigma}{KT}}+\left(\frac{L^2\Delta_F^2 D_{\mathrm{H}}}{T^2}\right)^{1/3}+\frac{L\Delta_F}{T}\right)$
Nair et al. (2025): FSL-SAGE†	$O\left(\frac{1}{\sqrt{T}}+\frac{\mathcal{E}_T}{T}\right)$
Liang et al. (2025): MU-SplitFed	$O\left(\frac{\sqrt{d}\,[L\Delta_F+\zeta^2+\sigma_s^2]}{\sqrt{\tau TK}}+\frac{\zeta^2+\sigma_c^2}{\sqrt{TK}}+\frac{\sqrt{d}\,[\zeta^2+\sigma_s^2+1]}{\tau T}\right)$
Kou et al. (2026): hybrid-order SFL†	$O\left(\sqrt{\frac{\kappa}{KT}}+\frac{1}{T}\right)+C_{\mathrm{drift}}$
This work: SFL-V1	$\mathcal{O}\left(\sqrt{\frac{\Delta_F B_\omega}{T}}+\left(\frac{\Delta_F^2 Kp^2 C_\omega}{T^2}\right)^{1/3}+\frac{\Delta_F}{\eta_{\max,1}T}\right)$
This work: SFL-V2	$\mathcal{O}\left(\sqrt{\frac{\Delta_F D_\omega(K)}{T}}+\left(\frac{\Delta_F^2 E_\omega(K,p)}{T^2}\right)^{1/3}+\frac{\Delta_F}{\eta_{\max,2}T}\right)$

For Liang et al. (2025), $d = d_c + d_s$ is the total parameter dimension and τ counts server-side local updates.

For Kou et al. (2026), κ is the effective-rank bound. It also retains the representation-drift contribution C_{drift}, which the paper claims vanishes in the IID setting.

For our SFL-V1 result, define

$$S_\omega := 2\big(C_g^2(\omega)+\sigma_g^2(\omega)\big)+1, \tag{63}$$

$$B_\omega := L_F\sigma_F^2(\omega), \tag{64}$$

$$C_\omega := L_F^2\Big\{140(C_f^2+\sigma_f^2)\big[12C_g^2(\omega)+8\sigma_g^2(\omega)\big]+28\sigma_f^2 S_\omega + 42S_\omega\big[12C_f^2+6\sigma_f^2+120L_f^2C_g^2(\omega)+80L_f^2\sigma_g^2(\omega)\big]\Big\}. \tag{65}$$

For our SFL-V2 result, define

$$D_\omega(K) := L_F C_f^2\big[C_g^2(\omega)+K^2\big], \tag{66}$$

$$E_\omega(K,p) := L_F^2 C_f^2\left[p^2KC_g^2(\omega)+\frac{K^3}{3}\right]. \tag{67}$$

The quantities $\eta_{\max,1}$ and $\eta_{\max,2}$ are positive stepsize caps.

Key observations. **a. Comparable leading-order rates with different structural dependencies.** After accounting for differences in the stationarity criteria and tuning the stepsizes, the vanishing terms in canonical first-order SFL guarantees generally have a leading $O(1/\sqrt{T})$ dependence for fixed structural parameters. We therefore do not claim a faster asymptotic rate. Instead, our analysis exposes SFL-specific dependencies on the cut layer, synchronization period, client number, and update structure that are hidden by the leading-order notation.

b. Cut-layer-aware synchronization analysis for SFL-V1. Our SFL-V1 bound separates the split-dependent first-order coefficient B_ω from the synchronization-dependent contribution $p^2 C_\omega$. Here, B_ω contains both gradient-moment and stochastic-variance quantities, whereas C_ω captures the split-dependent factors that amplify disagreement between periodic averaging events. Holding the remaining coefficients and stepsize conditions fixed, a split with a smaller C_ω incurs a smaller penalty as the synchronization period p increases. This result indicates that synchronization frequency and cut-layer selection should be considered jointly.

c. Variant-specific proof mechanisms. The SFL-V1 analysis controls the disagreement accumulated between local iterates and their average during periodic synchronization. SFL-V2 instead requires a pairwise recursion that tracks how disagreement propagates through the server state passed sequentially across clients within each iteration. The quantities $e_1(\omega)$ and $e_2(\omega)$ characterize, respectively, the contraction and forcing terms in this sequential recursion. These proof mechanisms reflect the different update dynamics of SFL-V1 and SFL-V2.

d. Explicit characterization of the client-number dependence. The SFL-V2 analysis identifies where the K^2 dependence enters the current upper bound: gradients and disagreement can accumulate along the sequential client chain. Consequently, the bound contains both K-dependent vanishing terms and an additive $O(A_\omega K^2)$ residual. This residual is a feature of the present upper bound and should not be interpreted as a lower bound establishing an unavoidable error floor. Moreover, its dependence on K is understood while holding the other problem-dependent coefficients fixed.

H Additional Experiments

H.1 Hyper-parameters

For the regression task, we use ADAM with learning rate 0.001 and batch size 128. Each communication round consists of 5 local epochs.

For the classification task, we use SGD with learning rate 0.01 and batch size 128. Each communication round consists of 5 local epochs.

Note that our hyperparameters mainly follow prior SFL studies, e.g., Han et al. (2024), to keep setup consistent across methods rather than extensively tuned.

We run all experiments on a single A100 GPU.

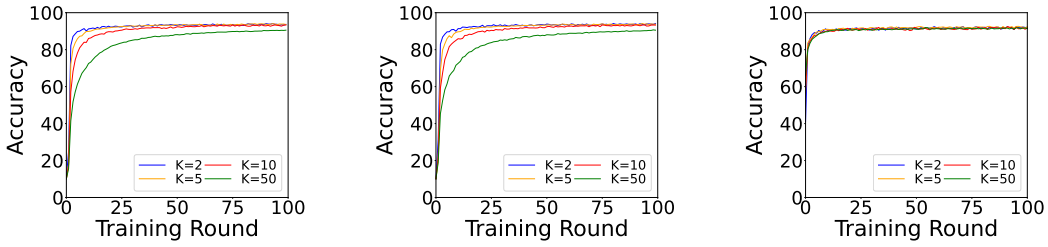
H.2 Experiments on CIFAR-10 and FEMNIST

We conduct additional experiments on CIFAR-10. Fig. 5 reports the CIFAR-10 accuracy versus K under both IID and non-IID partitions. The same trend as in Fig. 2 continues to hold. Increasing K generally worsens performance across methods. Under IID data, the degradation is relatively mild, where each method still converges to a high-accuracy regime, but larger K can slow early progress and slightly lower the final plateau. Under non-IID data, the effect is substantially amplified. As K increases, the final accuracy drops more noticeably and convergence becomes less stable, reflecting the stronger optimization difficulty induced by heterogeneous client updates.

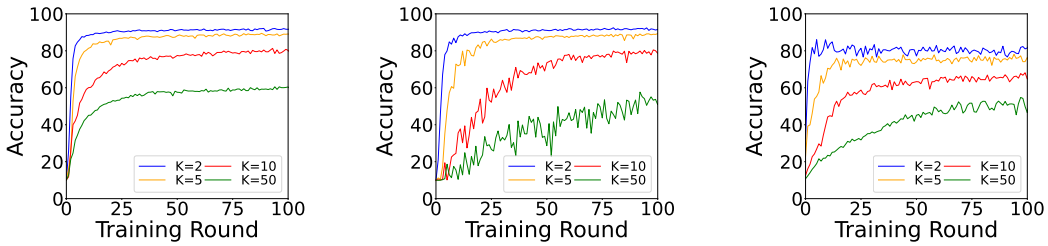
We further conduct experiments on FEMNIST. We consider SFL-V1 and SFL-V2, and use a large client number $K = 100$. The model used here is ResNet-18. The results are reported in Table 3.

Table 3: FEMNIST accuracy under different cut layers.

Cut layer	**SFL-V1**	**SFL-V2**
$L_c = 1$	69.2	71.3
$L_c = 2$	73.8	72.6
$L_c = 3$	76.6	77.4
$L_c = 4$	78.5	78.4



(a) SFL-V1 (IID) (b) SFL-V2 (IID) (c) SL (IID)



(d) SFL-V1 (non-IID) (e) SFL-V2 (non-IID) (f) SL (non-IID)

Figure 5: CIFAR-10 **accuracy** versus K under both IID and non-IID scenarios.

From the table, we observe that the results are consistent with the main paper, where performance generally improves as the cut layer moves deeper.

H.3 Experiments using ResNet-110.

We further conducted additional experiments with a deeper model, ResNet-110, under the non-IID CIFAR-100 setting with $K = 10$ clients. We considered three cut-layer choices $L_c \in \{s, m, d\}$, where s denotes a shallow cut at the 12-th block, m denotes a middle cut at the 17-th block, and d denotes a deep cut at the 26-th block. The results are summarized below.

Table 4: CIFAR-100 accuracy vs. L_c using ResNet-110.

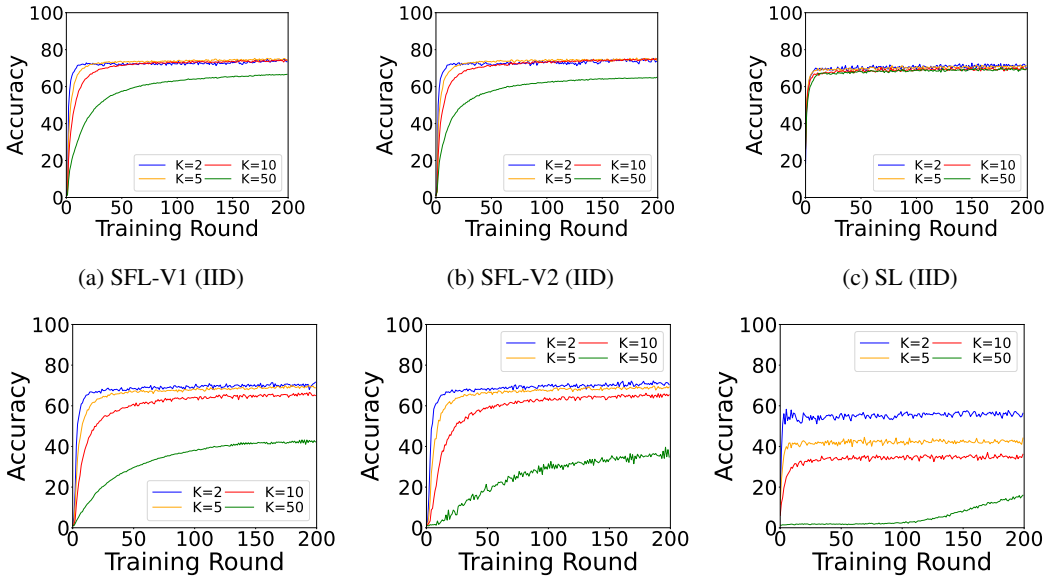
Cut layer	**SFL-V1**	**SFL-V2**
$L_c = s$	61.4	59.3
$L_c = m$	64.2	63.5
$L_c = d$	67.9	66.8

The same trend as in the main paper continues to hold, i.e., deeper cuts lead to better performance for both SFL-V1 and SFL-V2.

H.4 Experiments using different data heterogeneity α

We study the impact of data heterogeneity here. For both figures, we contrast the non-IID setting with $\alpha = 0.1$, as used in the main paper, and the IID setting, which corresponds to $\alpha = \infty$.

Impact of the number of clients K. Fig. 6 shows the CIFAR-100 accuracy versus the number of clients K under $L_c = 4$. We can see that increasing K generally worsens the performance, and the degradation is more significant under the non-IID setting. In the IID case, all methods are relatively stable as K increases, while under non-IID data, both SFL variants experience accuracy decrease when K becomes large. This trend is consistent with our theory, since a larger number of clients can amplify client drift and split-induced stochasticity. Moreover, SFL-V1 and SFL-V2 consistently outperform vanilla SL under non-IID data, especially when K is large, which validates the advantage of FL-style synchronization in mitigating heterogeneous client effects.



(a) SFL-V1 (IID) (b) SFL-V2 (IID) (c) SL (IID)

(d) SFL-V1 (non-IID) (e) SFL-V2 (non-IID) (f) SL (non-IID)

Figure 6: CIFAR-100 **accuracy** (top-1) versus K under $L_c = 4$.

Impact of the cut layer L_c. Fig. 7 shows the CIFAR-100 accuracy versus the cut layer L_c. We can see that later splits, i.e., larger L_c, generally lead to better performance, and this trend is more significant under non-IID data. In the IID setting, the performance gaps across different cut layers are relatively small, while under non-IID data, the choice of cut layer has a stronger impact, especially for SFL-V2. This observation is consistent with our theory, i.e., changing the cut layer changes the split mask ω, the client-side mapping, and the corresponding split-dependent constants in the convergence bounds. A later cut can produce more stable smashed representations and backpropagated gradients, leading to better optimization behavior.

Overall, these additional experiments are consistent with the main paper's conclusion.

H.5 Cut Layer Variability vs. Performance

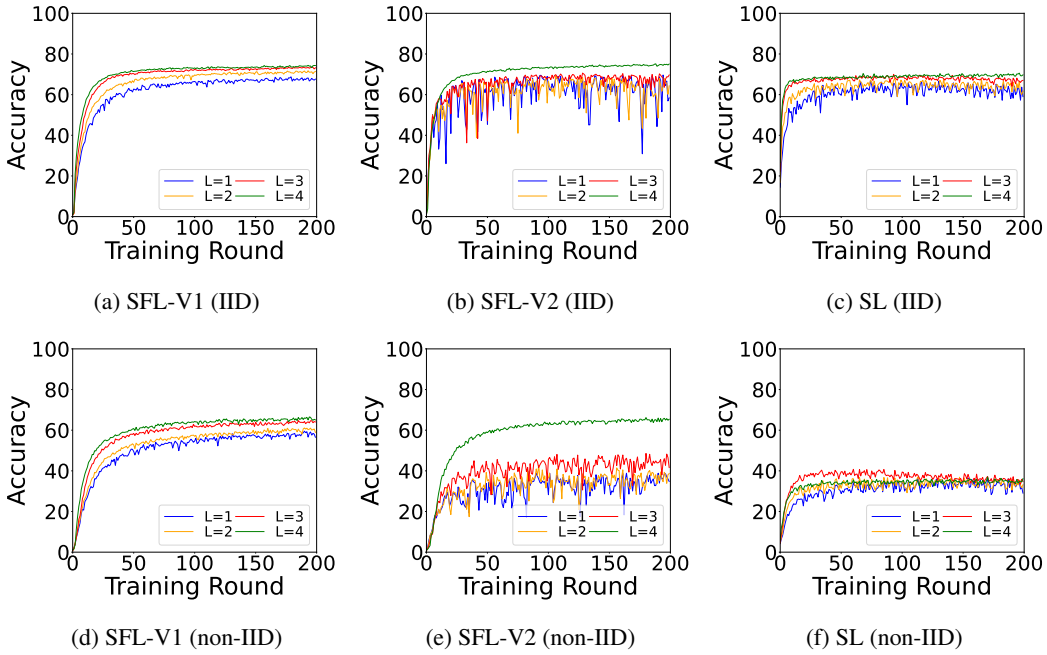
Our theory identifies the split-dependent quantities through which the cut layer affects convergence. More specifically, the cut layer determines the client-side mapping and the gradients propagated through the split interface. *A practical proxy could be the cross-client variance of the cut-layer gradients.* For a candidate layer L_c, it can be estimated during a short pilot-training phase as

$$V(L_c) = \frac{1}{K}\sum_{k=1}^{K}\left\|\nabla_u F_{L_c}^{(k)} - \frac{1}{K}\sum_{j=1}^{K}\nabla_u F_{L_c}^{(j)}\right\|^2 .$$

Here, $\nabla_u F_{L_c}^{(k)}$ denotes the gradient with respect to the cut-layer representation for client k. A smaller $V(L_c)$ means that clients produce more consistent split gradients at that cut layer, corresponding to lower split-induced variability in our analysis.

To validate this interpretation, we measured the average cross-client gradient variance for different cut layers on CIFAR-10:

Cut layer	SFL-V1	SFL-V2
1	33.8	33.2
2	15.2	15.6
3	9.8	9.7
4	7.1	7.0



(a) SFL-V1 (IID) (b) SFL-V2 (IID) (c) SL (IID)

(d) SFL-V1 (non-IID) (e) SFL-V2 (non-IID) (f) SL (non-IID)

Figure 7: CIFAR-100 **accuracy** (top-1) versus L_c.

The table shows that deeper cut layers yield lower variability for both variants, which is consistent with our theory (e.g., Theorem 4.6) and experimental observation. In practice, one can evaluate candidate layers for a few pilot iterations, compute $V(L_c)$, and balance variability against practical considerations such as communication cost and client computation.