SPORADIC GRADIENT TRACKING FOR DECENTRALIZED FEDERATED LEARNING OVER DIRECTED GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized Federated Learning (DFL) enables clients with local data to collaborate in a peer-to-peer manner to train a generalized model. In this paper, we propose a novel formulation that unifies two branches of work that have separately solved important challenges in DFL as special cases: (i) push-pull gradient tracking techniques for stable convergence over directed graphs and (ii) resource-aware methods for handling heterogeneous client. We propose *Sporadic Gradient Tracking* (`Spod-GT`), the first DFL algorithm that incorporates these factors over directed graphs by allowing flexible (i) client-specific gradient computation frequencies and (ii) heterogeneous and asymmetric inter-client communication frequencies. We conduct a rigorous convergence analysis of our methodology with a relaxed assumption on gradient diversity, providing $\mathcal{O}(1/K)$ consensus and stationarity guarantees for GT over directed graphs under intermittent client participation. Through numerical experiments on image classification datasets, we demonstrate the efficacy of `Spod-GT` compared to well-known GT baselines.

1 INTRODUCTION

Federated learning (FL) has been well-explored as a distributed privacy-preserving training method in the past decade (Kairouz et al., 2021). Beyond FL's conventional server-client architecture, there has been recent interest in decentralized federated learning (DFL) over peer-to-peer networks, where model aggregation occurs through a consensus algorithm executed over the client network graph (Koloskova et al., 2020)

DFL faces heterogeneity in client communication and computation capabilities and in local dataset statistics (Li et al., 2020), while the connectivity and structure (directed or undirected) of the network graph critically influence training performance (Nedić et al., 2018). This has given rise to two parallel branches of work in DFL literature. The first has focused on resource and data heterogeneity, simplifying inter-client communications to an undirected graph (Zehtabi et al., 2025). However, client communications are often asymmetric, e.g., in wireless networks where devices transmit at different power and noise levels, causing unidirectional channels (Nedić et al., 2025). The second branch of work has modeled inter-client communications via a directed graph, with focusing on improving the convergence rate of DFL via gradient tracking (GT) (Nguyen et al., 2024), but has largely overlooked resource heterogeneity.

We propose a novel framework to unify these two branches of work by answering the question:

How can we develop a consensus-based GT algorithm for DFL over directed graphs, that is aware of heterogeneities in computations and communications of the clients, while maintaining convergence guarantees under heterogeneous data distributions of clients?

Contributions. Our main contributions are as follows:

- We develop *Sporadic Gradient Tracking* (`Spod-GT`) in Sec. 3, the first computation- and communication-aware algorithm for consensus-based DFL over directed graphs. Each client participates in (i) gradient computation based on its processing resources and (ii) inter-client communication based on its communication resources (Fig. 1), so that clients with lower available re-

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.

Table 1: Comparison of our work with representative related papers.

Paper	Properties of Framework					Theoretical Assumptions	
	Directed Graphs	Non-doubly Stochastic	Gradient Tracking	Sporadic Comp.	Sporadic Comm.	Relaxed Gradient Diversity	Non-Convex Loss
(Tang et al., 2018), (Koloskova et al., 2021)			✓				✓
(Koloskova et al., 2020)					✓	✓	✓
(Pu et al., 2020)	✓	✓	✓		✓		
(Xin et al., 2021)	✓		✓				✓
(Nguyen et al., 2024), (Nedić et al., 2025)	✓	✓	✓				
(Liu et al., 2025)			✓			✓	✓
(Zehtabi et al., 2025)				✓	✓	✓	✓
(Liang et al., 2025), (You & Pu, 2026)	✓	✓	✓				✓
Ours [a]	✓	✓	✓	✓	✓	✓	✓

[a] We discuss how our theoretical analysis recover existing results in Secs. 4.3 and 4.4.

sources (stragglers) participate less frequency than resource-abundant clients, balancing the training load in a client-specific way.

- We provide a rigorous convergence analysis of GT-enhanced DFL for non-convex losses over directed graphs in Sec. 4, under a relaxed gradient diversity assumption capturing the data heterogeneity among clients. Unique to our setup, we factor in sporadicity in both (i) gradient calculations and (ii) inter-client communications, and illustrate how our analysis reverts back to conventional GT methods as special cases. We overcome novel challenges in jointly analyzing sporadic GT updates and sporadic communications over directed graphs, introduced as Theoretical Challenges 1 and 2 in Sec. 4.
- Our theoretical results elucidate novel DFL dynamics over directed graphs under sporadic participation. Our Proposition 4.8 is the first to show a clear dependence of (i) minimum gradient computation probability on gradient diversity and (ii) minimum inter-client communication probability on spectral characteristics of the underlying directed graph.
- In Sec. 5, we show that `Spod-GT` consistently achieves higher accuracies in a shorter time than state-of-the-art GT methods for directed graphs, which are not aware of the underlying resource heterogeneity.

See Appendix B.1 for a summary of notation.

2 Related Work

Table 1 summarizes key contributions of our work relative to closely related literature on GT algorithms for DFL. Our paper is the first to consider sporadic GT steps and aggregations simultaneously, capturing heterogeneous resources in fully-decentralized directed networks. In the following, we discuss related work along key considerations of GT and sporadicity of communications in DFL.

Decentralized Federated Learning and Gradient Tracking. In DFL, clients collaboratively train a global model without a centralized parameter server by communicating directly over a network and performing iterative consensus operations. Early decentralized methods studied distributed stochastic gradient descent (SGD) for convex objectives (Nedić & Ozdaglar, 2009), and D-PSGD (Lian et al., 2017) extended decentralized SGD to smooth non-convex objectives. However, decentralized SGD converges more slowly than its centralized counterpart due to the disagreement introduced by local updates and consensus operations (Nedić et al., 2018). Gradient tracking (GT) addresses this issue by introducing an auxiliary variable at each client that estimates the global gradient, achieving convergence rates closer to centralized optimization. GT was first explored in EXTRA (Shi et al., 2015) and further developed in NEXT (Di Lorenzo & Scutari, 2016), DIGing (Nedić et al., 2017), and related approaches.

GT over Undirected Graphs. Most early GT methods assume an undirected graph with symmetric links. EXTRA (Shi et al., 2015) introduced gradient correction for convex objectives, NEXT (Di Lorenzo & Scutari, 2016) extended GT to non-convex objectives with bounded gradients, and D^2 (Tang et al., 2018) studied general smooth non-convex objectives. These approaches rely on undirected graphs or doubly-stochastic mixing matrices, which implicitly assume balanced information exchange and may not hold when links are asymmetric or clients have heterogeneous communication capabilities. GT-DSGD (Xin et al., 2021) studied directed graphs while maintaining a doubly-stochastic mixing matrix, resulting in an analysis similar to the undirected setting. In contrast, we consider general directed graphs that do not necessarily admit doubly-stochastic matrices (Gharesifard & Cortés, 2010).

GT over Directed Graphs. Early methods over directed graphs, such as subgradient-push (Nedić & Olshevsky, 2014), combined push-sum consensus with SGD for convex objectives. However, SGD converges slowly, and push-sum consensus results in nonlinear update dynamics that complicate analysis (Xin et al., 2020). Two major directions have since emerged: incorporating GT into push-sum-based frameworks, e.g., Push-DIGing (Nedić et al., 2017) and FROST (Xin et al., 2019), and avoiding nonlinear push-sum updates via row-stochastic and column-stochastic mixing matrix pairs, e.g., AB (Xin & Khan, 2018), Push-Pull (Pu et al., 2020), and AB/Push-Pull (Nedić et al., 2025). Recent works have extended directed GT to stochastic gradients using multi-step descent frameworks (Liang et al., 2025; You & Pu, 2026). However, these methods assume homogeneous computation and communication resources among clients, and do not consider sporadic computations or aggregations caused by resource heterogeneity. We focus on the exact-gradient setting in the main text and provide stochastic-gradient extensions in Appendix A.

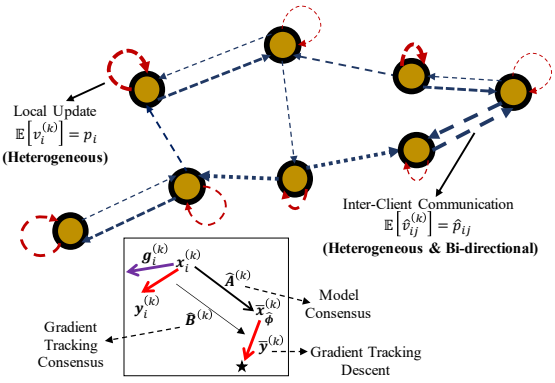


Figure 1: Illustration of GT-enhanced DFL over directed graphs with clients having heterogeneous computation and communication capabilities. Unlike conventional GT, where local updates and inter-client communications occur at every iteration, `Spod-GT` makes both sporadic (dashed lines, thickness representing relative frequency).

Sporadic Communications. One approach to reduce communications is randomized gossip (Boyd et al., 2006), applied to SGD over undirected graphs in DSGD (Koloskova et al., 2020) and DSpodFL (Zehtabi et al., 2025), to GT over undirected graphs in GSGT (Pu & Nedić, 2021), to SGD with push-sum over directed graphs in SGP (Assran et al., 2019), and to GT over directed graphs in G-Push-Pull (Pu et al., 2020). While DSGD and DSpodFL work with the expected contraction of mixing matrices over undirected graphs, works on directed graphs either assume strong connectivity of every instantaneous graph or within a deterministic sliding window (Nguyen et al., 2024), or allow clients to communicate through only one edge at each iteration (Pu et al., 2020). We relax this by only relying on the expected graph being strongly connected. Another approach is to take multiple local gradient steps between aggregations, as in DFedAvg (Sun et al., 2022) with SGD and K-GT (Liu et al., 2025) with GT, which however enforce a fixed number of local steps for all clients, overlooking their computational heterogeneity. Our framework instead allows for an adaptive number of local steps for each client.

3 Methodology

3.1 Problem Formulation

We consider a DFL system with m clients $\mathcal{M} := \{1, \dots, m\}$ and a series of training iterations $k = 0, ..., K-1$ (see Appendix B.1 for a summary of notation). Clients are connected through a *directed graph* $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, where $(i, j) \in \mathcal{E}$ if there exists a physical outgoing link from client i to j. Each client $i \in \mathcal{M}$ has a local dataset $\mathcal{D}_i$ of size $|\mathcal{D}_i| = D_i$, and aims to train a model parameterized by $\boldsymbol{x} \in \mathbb{R}^n$ by collaborating with its neighbors without exchanging data. Let $\mathcal{L}(\boldsymbol{x}, \xi)$ denote the loss of model $\boldsymbol{x}$ evaluated at data sample ξ. We define the local and the global loss

functions as $F_i(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathbb{P}(\mathcal{D}_i)}[\mathcal{L}(\boldsymbol{x},\xi)]$ and $F(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m F_i(\boldsymbol{x})$, respectively. The goal of DFL is for clients to learn the optimal global model $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} F(\boldsymbol{x})$.

3.2 Sporadic Gradient Tracking

To minimize the global loss function, we employ *gradient tracking* (GT), which has been shown to have good convergence properties for directed graphs (Xin & Khan, 2018; Pu et al., 2020). In GT methods, each client has an additional set of parameters $\boldsymbol{y}$ that tracks an estimate of the global gradient at each iteration of training. Although existing work on GT has assumed that each client computes gradients (or multiple local gradient steps) at each communication round, we focus on a relaxed and generalized extension of GT where the clients may opt out of computing gradients at some iterations. The iterative updates of our method, which we call `Spod-GT`, can be formally written in two steps as follows:

$$\boldsymbol{x}_i^{(k+\frac{1}{2})} = \boldsymbol{x}_i^{(k)} + \sum_{j=1}^m a_{ij}(\boldsymbol{x}_j^{(k)} - \boldsymbol{x}_i^{(k)})\hat{v}_{ij}^{(k)}, \quad \boldsymbol{y}_i^{(k+\frac{1}{2})} = \boldsymbol{y}_i^{(k)} + \sum_{j=1}^m (b_{ij}\boldsymbol{y}_j^{(k)}\hat{v}_{ij}^{(k)} - b_{ji}\boldsymbol{y}_i^{(k)}\hat{v}_{ji}^{(k)}), \quad (1)$$

which are the aggregation steps, and

$$\boldsymbol{x}_i^{(k+1)} = \boldsymbol{x}_i^{(k+\frac{1}{2})} - \eta_i^{(k)}\boldsymbol{y}_i^{(k+\frac{1}{2})}, \quad \boldsymbol{y}_i^{(k+1)} = \boldsymbol{y}_i^{(k+\frac{1}{2})} + \nabla F_i(\boldsymbol{x}_i^{(k+1)})v_i^{(k+1)} - \nabla F_i(\boldsymbol{x}_i^{(k)})v_i^{(k)}, \quad (2)$$

which are the update steps. Here, $\boldsymbol{x}_i^{(k)}$ and $\boldsymbol{y}_i^{(k)}$ are the model parameters and the global gradient estimate of the client $i \in \mathcal{M}$ in the communication round $k \geq 0$, respectively, and we initialize $\boldsymbol{y}_i^{(0)} = \nabla F_i(\boldsymbol{x}_i^{(0)})v_i^{(0)}$, where $\eta_i^{(k)}$ is the learning rate of the client i in the round k.

Sporadicity. Unique to our setup, $v_i^{(k)} \in \{0,1\}$ and $\hat{v}_{ij}^{(k)} \in \{0,1\}$ in Eqs. 1 and 2 are indicator random variables that capture whether or not a client participates in the computation of gradients and inter-client communication in a given round, respectively, and are a relaxation of conventional GT methods (Koloskova et al., 2021; Nedić et al., 2025; You & Pu, 2026) where they always consider $v_i^{(k)} = \hat{v}_{ij}^{(k)} = 1$. Our framework also generalizes algorithms such as K-GT (Liu et al., 2025) where the number of local GT steps K is the same for all clients and is fixed throughout the training, since the combination of indicator variables $v_i^{(k)}$ and $\hat{v}_{ij}^{(k)}$ can give rise to heterogeneous and dynamic number of local GT steps (see Appendix B.2).

Lastly, a_{ij} and b_{ij} in Eq. 1 are the mixing weights for the model parameters and the tracked gradients that are assigned to the unidirectional link $(i,j) \in \mathcal{E}$, respectively. While a_{ij} are chosen by the client receiving the model parameters (client i), b_{ij} are determined by the client transmitting the tracked gradients (client j). To simplify the notation in Eq. 1, let $\hat{a}_{ij}^{(k)} = a_{ij}\hat{v}_{ij}^{(k)}$ and $\hat{b}_{ij}^{(k)} = b_{ij}\hat{v}_{ij}^{(k)}$ for $i \neq j$, and $\hat{a}_{ii}^{(k)} = 1 - \sum_{\substack{\ell=1\\\ell\neq i}}^m a_{i\ell}\hat{v}_{i\ell}^{(k)}$ and $\hat{b}_{ii}^{(k)} = 1 - \sum_{\substack{\ell=1\\\ell\neq i}}^m b_{\ell i}\hat{v}_{\ell i}^{(k)}$ for $i = j$, using which we get

$$\boldsymbol{x}_i^{(k+1)} = \sum_{j=1}^m \hat{a}_{ij}^{(k)}\boldsymbol{x}_j^{(k)} - \eta_i^{(k)}\sum_{j=1}^m \hat{b}_{ij}^{(k)}\boldsymbol{y}_j^{(k)}, \; \boldsymbol{y}_i^{(k+1)} = \sum_{j=1}^m \hat{b}_{ij}^{(k)}\boldsymbol{y}_j^{(k)} + \nabla F_i(\boldsymbol{x}_i^{(k+1)})v_i^{(k+1)} - \nabla F_i(\boldsymbol{x}_i^{(k)})v_i^{(k)}, \quad (3)$$

where $[\hat{a}_{ij}^{(k)}]_{1\leq i,j\leq m}$ and $[\hat{b}_{ij}^{(k)}]_{1\leq i,j\leq m}$ represent sporadic mixing weights. The sporadic update rule $\nabla F_i(\boldsymbol{x}_i^{(k+1)})v_i^{(k+1)} - \nabla F_i(\boldsymbol{x}_i^{(k)})v_i^{(k)}$ in Eq. 3 is a generalization of existing GT methods of the form $\nabla F_i(\boldsymbol{x}_i^{(k+1)}) - \nabla F_i(\boldsymbol{x}_i^{(k)})$ (Koloskova et al., 2021; Nedić et al., 2025), which implies that updates to GT parameters $y_i^{(k)}$ may not occur in consecutive iterations. This core change introduces challenges in the analysis of our approach, and one of our main contributions is to ensure convergence guarantees under intermittent computations. A pseudo-code of our algorithm summarizing the steps above is presented in Appendix B.3.

Insights for Sporadicity. Resource availability is often heterogeneous and asymmetric in DFL, i.e., there are variations in (i) client computation capabilities and (ii) communication link throughputs. `Spod-GT` overcomes these limitations in Eqs. 1 and 2 by relaxing (i) the client gradient computation frequencies to $\mathbb{E}[v_i^{(k)}] = p_i \in (0,1]$ proportional to processing capabilities, and (ii) the edge utilization frequency $\mathbb{E}[\hat{v}_{ij}^{(k)}] = \hat{p}_{ij} \in (0,1]$ proportional to a link's throughput. This is in contrast to conventional DFL methods that enforce a restrictive $p_i = \hat{p}_{ij} = 1$, demanding the same work load from all clients irrespective of their heterogeneous and asymmetric capabilities.

3.3 Matrix Representation

We collect the model parameters $\boldsymbol{x}_i^{(k)}$, the gradient tracking estimates $\boldsymbol{y}_i^{(k)}$ and the local gradients $\nabla F_i(\boldsymbol{x}_i^{(k)})$ of all clients to the matrices $\boldsymbol{X}^{(k)} = [\boldsymbol{x}_1^{(k)} \cdots \boldsymbol{x}_m^{(k)}]^T \in \mathbb{R}^{m\times n}$, $\boldsymbol{Y}^{(k)} = [\boldsymbol{y}_1^{(k)} \cdots \boldsymbol{y}_m^{(k)}]^T \in \mathbb{R}^{m\times n}$ and $\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)}) = [\nabla F_1(\boldsymbol{x}_1^{(k)}) \cdots \nabla F_m(\boldsymbol{x}_m^{(k)})]^T \in \mathbb{R}^{m\times n}$, respectively. We also collect all mixing weights into matrices, i.e., $\boldsymbol{A} = [a_{ij}]_{1\le i,j\le m} \in \mathbb{R}^{m\times m}$, $\boldsymbol{B} = [b_{ij}]_{1\le i,j\le m} \in \mathbb{R}^{m\times m}$, $\hat{\boldsymbol{A}}^{(k)} = [\hat{a}_{ij}^{(k)}]_{1\le i,j\le m} \in \mathbb{R}^{m\times m}$ and $\hat{\boldsymbol{B}}^{(k)} = [\hat{b}_{ij}^{(k)}]_{1\le i,j\le m} \in \mathbb{R}^{m\times m}$. For the learning rate $\eta_i^{(k)}$ and the indicator variable $v_i^{(k)}$ of each client, which are scalars, we define the diagonal matrices $\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)} = \mathrm{diag}(\boldsymbol{\eta}^{(k)}) \in \mathbb{R}^{m\times m}$ and $\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)} = \mathrm{diag}(\boldsymbol{v}^{(k)}) \in \mathbb{R}^{m\times m}$, respectively. Using this matrix notation, we can collect the iterative update rules for each client in Eq. 3 as update rules for the whole decentralized network as

$$\boldsymbol{X}^{(k+1)} = \hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)},\ \boldsymbol{Y}^{(k+1)} = \hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} + \boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\nabla \boldsymbol{F}(\boldsymbol{X}^{(k+1)}) - \boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)}). \tag{4}$$

In conventional GT methods without sporadicity, Eq. 4 is taken to be $\boldsymbol{X}^{(k+1)} = \boldsymbol{A}\boldsymbol{X}^{(k)} - \eta\boldsymbol{Y}^{(k)}$ and $\boldsymbol{Y}^{(k+1)} = \boldsymbol{B}\boldsymbol{Y}^{(k)} + \nabla \boldsymbol{F}(\boldsymbol{X}^{(k+1)}) - \nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})$, which is a simplified case of our update rules. Our main contribution in this work is to provide convergence guarantees despite the use of sporadic aggregation matrices $\hat{\boldsymbol{A}}^{(k)}$ and $\hat{\boldsymbol{B}}^{(k)}$, and sporadic GT updates $\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\nabla \boldsymbol{F}(\boldsymbol{X}^{(k+1)}) - \boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})$.

4 Convergence Analysis

We lay out our assumptions in Sec. 4.1 and definitions in Sec. 4.2, then discuss our key propositions in Sec. 4.3 and main theorem in Sec. 4.4. We focus on the exact gradient (i.e., full-batch) setting to more clearly highlight the implications of sporadicity, and provide our analysis for stochastic gradients in Appendix A.

4.1 Assumptions and Immediate Implications

Assumption 4.1 (Lipschitz-Continuous Gradients)**.** Each local loss function $F_i : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient $\nabla F_i : \mathbb{R}^n \to \mathbb{R}^n$ is L_i-Lipschitz continuous, i.e., $\|\nabla F_i(\boldsymbol{x}) - \nabla F_i(\boldsymbol{x}')\| \le L_i\|\boldsymbol{x} - \boldsymbol{x}'\|$, with $\{\boldsymbol{x}, \boldsymbol{x}'\} \subseteq \mathbb{R}^n$ and $L_i > 0$ for all $i \in \mathcal{M}$.

Since the global loss function is defined as $F(\boldsymbol{x}) = (1/m)\sum_{i=1}^m F_i(\boldsymbol{x})$, it is also continuously differentiable as a convex combination of continuously differentiable functions. Furthermore, its gradient function is $\bar{L}$-Lipschitz with $\bar{L} = (1/m)\sum_{i=1}^m L_i$, and we have that $F(\boldsymbol{x}) \le F(\boldsymbol{x}') + \langle \nabla F(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}'\rangle + \frac{1}{2}\bar{L}\|\boldsymbol{x} - \boldsymbol{x}'\|^2$, for all $\{\boldsymbol{x}, \boldsymbol{x}'\} \subseteq \mathbb{R}^n$ (Bottou et al., 2018).

Assumption 4.2 (Gradient Diversity)**.** Each local loss function $F_i : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and the distance from its gradient $\nabla F_i : \mathbb{R}^n \to \mathbb{R}^n$ to the global loss gradient function $\nabla F = (1/m)\sum_{i=1}^m \nabla F_i : \mathbb{R}^n \to \mathbb{R}^n$ is bounded as $\|\nabla F_i(\boldsymbol{x})\|^2 \le \delta_{0,i}^2 + \delta_{1,i}^2\|\nabla F(\boldsymbol{x})\|^2$, with $\boldsymbol{x} \in \mathbb{R}^n$ and $\delta_{0,i}, \delta_{1,i} > 0$.

Although earlier papers worked with the restrictive bounded gradient assumption, i.e., $\delta_{1,i}^2 = 0$, Assumption 4.2 has been used more recently as a relaxed condition on gradient diversity (Koloskova et al., 2020; Li & Lyu, 2023; Hu et al., 2023) (see Appendices B.4 and B.5 for further discussion).

Assumption 4.3 (Bernoulli Indicator Random Variables)**.** The indicator random variables $v_i^{(k)}$ and $\hat{v}_{ij}^{(k)}$ that capture whether a client $i \in \mathcal{M}$ computes local gradient or participates in inter-client communication with its neighbor j at iteration $k \ge 0$, respectively, are statistically independent Bernoulli random variables with non-zero probabilities. In other words, we have $\mathbb{E}[v_i^{(k)}] = p_i$ and $\mathbb{E}[\hat{v}_{ij}^{(k)}] = \hat{p}_{ij}$, where $0 < p_i, \hat{p}_{ij} \le 1$.

Let us define the expected mixing matrices with $\hat{a}_{ij} = a_{ij}p_{ij}$ and $\hat{b}_{ij} = b_{ij}p_{ij}$ if $i \ne j$, and $\hat{a}_{ii} = 1 - \sum_{j=1}^m a_{ij}p_{ij}$ and $\hat{b}_{jj} = 1 - \sum_{i=1}^m b_{ij}p_{ij}$ if $i = j$ as

$$\hat{\boldsymbol{A}} = \mathbb{E}\left[\hat{\boldsymbol{A}}^{(k)}\right], \qquad \hat{\boldsymbol{B}} = \mathbb{E}\left[\hat{\boldsymbol{B}}^{(k)}\right]. \tag{5}$$

Assumption 4.4 (Strongly Connected Directed Graph)**.** The directed graph $\mathcal{G}$ is strongly connected, i.e., for any pair of clients $\{i, j\} \subseteq \mathcal{M}$, there exist a path from i to j using the edges in $\mathcal{E}$. Also, we assume that each client i has a self-loop.

We design mixing matrices $\boldsymbol{A}$, $\boldsymbol{B}$ defined in Sec. 3.3 compatible with the graph $\mathcal{G}$, with $\boldsymbol{A}$ row-stochastic and $\boldsymbol{B}$ column-stochastic. In other words, $a_{ij} > 0$ if $j \in \mathcal{N}_i^{\text{in}} \cup \{i\}$ and $a_{ij} = 0$ otherwise, and $b_{ji} > 0$ if $j \in \mathcal{N}_i^{\text{out}} \cup \{i\}$ and $b_{ji} = 0$ otherwise. Note that $\mathcal{N}_i^{\text{in}}$ and $\mathcal{N}_i^{\text{out}}$ denote the in-neighbors and out-neighbors of a client, respectively. If we design $\boldsymbol{A}$ as row-stochastic and $\boldsymbol{B}$ as column-stochastic, the sporadic mixing matrices $\hat{\boldsymbol{A}}^{(k)}$, $\hat{\boldsymbol{B}}^{(k)}$ defined in Eq. 4 and expected mixing matrices $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$ defined in Eq. 5 will also be row-stochastic and column-stochastic, respectively. In matrix notation, we have $\hat{\boldsymbol{A}}^{(k)}\mathbf{1} = \mathbf{1}$, $\hat{\boldsymbol{A}}\mathbf{1} = \mathbf{1}$, $\mathbf{1}^T\hat{\boldsymbol{B}}^{(k)} = \mathbf{1}^T$ and $\mathbf{1}^T\hat{\boldsymbol{B}} = \mathbf{1}^T$, implying that the vector $\mathbf{1}$ is the right-eigenvector of $\hat{\boldsymbol{A}}^{(k)}$, $\hat{\boldsymbol{A}}$ and the left-eigenvector of $\hat{\boldsymbol{B}}^{(k)}$, $\hat{\boldsymbol{B}}$ corresponding to the eigenvalue 1.

Moreover, the elements of the expected mixing matrices $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ are nonnegative scaled versions of the elements in $\boldsymbol{A}$ and $\boldsymbol{B}$, which correspond to a strongly connected graph. Hence, they also correspond to a strongly connected graph and are therefore irreducible. In addition, the presence of self-loops implies that they are primitive. Hence, by the Perron-Frobenius Theorem, the eigenvalue 1 is simple and is the spectral radius for matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ (Pu et al., 2020). This implies that there exists a unique positive left-eigenvector $\hat{\boldsymbol{\phi}}$ for $\hat{\boldsymbol{A}}$ and a unique positive right-eigenvector $\hat{\boldsymbol{\pi}}$ for $\hat{\boldsymbol{B}}$, i.e.,

$$\hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}} = \hat{\boldsymbol{\phi}}^T,\ \hat{\boldsymbol{B}}\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}, \qquad \hat{\boldsymbol{\phi}}^T\mathbf{1} = 1,\ \mathbf{1}^T\hat{\boldsymbol{\pi}} = 1, \tag{6}$$

where we choose the eigenvectors $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\pi}}$ to be stochastic.

Theoretical Challenge 1: Loss of Instantaneous Spectral Properties. In contrast to the expected matrices $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$ defined in Eq. 5, the per-round matrices $\hat{\boldsymbol{A}}^{(k)}$, $\hat{\boldsymbol{B}}^{(k)}$, which fluctuate due to *sporadic communications* (random link dropouts), do not satisfy per-iteration strong-connectivity. Consequently, we cannot apply the standard Perron-Frobenius Theorem to deduce positive eigenvectors on a per-round basis, i.e., the eigenvectors $\hat{\boldsymbol{\phi}}^{(k)}$ and $\hat{\boldsymbol{\pi}}^{(k)}$ defined in $(\hat{\boldsymbol{\phi}}^{(k)})^T\hat{\boldsymbol{A}}^{(k)} = (\hat{\boldsymbol{\phi}}^{(k)})^T$ and $\hat{\boldsymbol{B}}^{(k)}\hat{\boldsymbol{\pi}}^{(k)} = \hat{\boldsymbol{\pi}}^{(k)}$ may contain zero entries (see (Touri & Nedić, 2013) for an in-depth discussion on dynamic random Markov chains). Moreover, for undirected graphs (Zehtabi et al., 2025) and digraphs admitting a doubly-stochastic matrix (Xin et al., 2021), the all-ones vector $\mathbf{1}$ remains both a left and right eigenvector of $\hat{\boldsymbol{A}}^{(k)}$, $\hat{\boldsymbol{B}}^{(k)}$ despite communication dropouts, making their analysis significantly easier than our setup. In our framework, random link dropouts alter row- and column-stochasticity independently over directed topologies. A key contribution of our work is overcoming these communication-induced matrix instabilities by establishing a contraction property for non-doubly stochastic aggregation matrices $\hat{\boldsymbol{A}}^{(k)}$, $\hat{\boldsymbol{B}}^{(k)}$ in Lemma C.4, which operates in tandem with the state tracking analysis introduced as Theoretical Challenge 2 in Section 4.2.

4.2 Definitions

We use the properties of the left and right eigenvectors in Eq. 6 to give the following definition.

Definition 4.5. For the purpose of our theoretical analysis, we will consider the $\hat{\boldsymbol{\phi}}$-weighted average of the model parameters $\boldsymbol{x}_i^{(k)}$, the regular average of GT parameters $\boldsymbol{y}_i^{(k)}$ and the weighted average of local gradients $\nabla F_i(\boldsymbol{x}_i^{(k)})$ given as $\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)} = \sum_{i=1}^m \hat{\phi}_i \boldsymbol{x}_i^{(k)} = \hat{\boldsymbol{\phi}}^T\boldsymbol{X}^{(k)}$, $\bar{\boldsymbol{y}}^{(k)} = \frac{1}{m}\sum_{i=1}^m \boldsymbol{y}_i^{(k)} = \frac{1}{m}\mathbf{1}^T\boldsymbol{Y}^{(k)}$ and $\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} = \frac{1}{m}\sum_{i=1}^m \nabla F_i(\boldsymbol{x}_i^{(k)})v_i^{(k)} = \frac{1}{m}\mathbf{1}^T\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})$.

Applying the average vector notation of Definition 4.5 to the matrix updates of Eq. 4, we get

$$\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} = \hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \hat{\boldsymbol{\phi}}^T\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}, \quad \bar{\boldsymbol{y}}^{(k+1)} = \bar{\boldsymbol{y}}^{(k)} + \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k+1)})}_{\boldsymbol{v}^{(k+1)}} - \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}}. \tag{7}$$

Recursively expanding $\bar{\boldsymbol{y}}^{(k+1)}$ with the initialization of Eq. 2 as $\bar{\boldsymbol{y}}^{(0)} = \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(0)})}_{\boldsymbol{v}^{(0)}}$, we get

$$\bar{\boldsymbol{y}}^{(k)} = \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}}. \tag{8}$$

Theoretical Challenge 2: Masked State and Gradient Tracking. In conventional GT methods with no sporadicity, Eqs. 7 and 8 simplify to $\bar{\boldsymbol{x}}_{\boldsymbol{\phi}}^{(k+1)} = \bar{\boldsymbol{x}}_{\boldsymbol{\phi}}^{(k)} - \eta\bar{\boldsymbol{y}}^{(k)}$ and $\bar{\boldsymbol{y}}^{(k)} = \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}$. This ensures that (i) the average GT vector $\bar{\boldsymbol{y}}^{(k)}$ precisely tracks the true average gradient $\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}$, and (ii) the weighted average model $\bar{\boldsymbol{x}}_{\boldsymbol{\phi}}^{(k+1)}$ updates directly from the previous weighted average model $\bar{\boldsymbol{x}}_{\boldsymbol{\phi}}^{(k)}$. In our sporadic formulation (Eqs. 7, 8), joint randomness fundamentally distorts these state dynamics: (i) *due to sporadic computations* (driven by indicators $v_i^{(k)}$), the average GT vector $\bar{\boldsymbol{y}}^{(k)}$

tracks the dynamically-changing *masked* average gradient $\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}}$; and (ii) *due to sporadic communications* (driven by indicators $\hat{v}_{ij}^{(k)}$), the weighted average model parameters $\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}$ are updated from a *masked* weighted average of the previous model $\hat{\boldsymbol{\phi}}^T \hat{\boldsymbol{A}}^{(k)} \boldsymbol{X}^{(k)}$. Combining this masked tracking analysis with the expected matrix contraction results (discussed in Theoretical Challenge 1 in Section 4.1) forms the core of our proof techniques in Appendices C and D.

Definition 4.6 (Consensus Error)**.** We collect the consensus error for model and GT parameters $\boldsymbol{x}_i^{(k)}$ and $\boldsymbol{y}_i^{(k)}$ in a vector $\boldsymbol{\varsigma}^{(k)} \in \mathbb{R}^{2\times 1}$ as $\boldsymbol{\varsigma}^{(k)} = \left[\mathbb{E}[\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\|_{\hat{\boldsymbol{\phi}}}^2], \mathbb{E}[\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\|_{\hat{\boldsymbol{\pi}}}^2]\right]^T$.

Lemmas D.1 and D.2 upper bound $\varsigma_0^{(k)}$ and $\varsigma_1^{(k)}$, respectively, which can compactly be written as

$$\boldsymbol{\varsigma}^{(k+1)} \leq \boldsymbol{\Psi}^{(k)}\boldsymbol{\varsigma}^{(k)} + \boldsymbol{\gamma}^{(k)}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right] + \boldsymbol{\omega}^{(k)}, \tag{9}$$

where the values for $\boldsymbol{\Psi}^{(k)} = [\psi_{ij}^{(k)}]_{1\leq i,j\leq m} \in \mathbb{R}^{2\times 2}$, $\boldsymbol{\gamma}^{(k)} = [\gamma_1^{(k)}, \gamma_2^{(k)}]^T \in \mathbb{R}^{2\times 1}$ and $\boldsymbol{\omega}^{(k)} = [\omega_1^{(k)}, \omega_2^{(k)}]^T \in \mathbb{R}^{2\times 1}$ are indexed in Table 3 in Appendix B.6. In Appendices C and D we form tight bounds for Eq. 9, such that it would decrease over time as discussed next in Sec. 4.3.

4.3 Propositions

In our first proposition, we lay out the sufficient constraints on the learning rates $\eta_i^{(k)}$ to ensure that the consensus error $\boldsymbol{\varsigma}^{(k)}$ in Eq. 9 decreases over time.

Proposition 4.7 (Spectral Radius of Transition Matrix)**.** *Let Assumptions 4.1-4.4 hold, and learning rates satisfy* $\eta_i^{(k)} < \frac{1/8}{\sqrt{m}\kappa_3}\min\left\{\frac{1-\tilde{\rho}_B}{\sqrt{1+\tilde{\rho}_B}}\frac{\sqrt{\min(\hat{\pi}_i)}(1-\tilde{\rho}_A)/2}{\sqrt{\frac{2\kappa_4}{m}+16\kappa_5\hat{\rho}_{0,A}+\frac{1-\tilde{\rho}_B}{8\tilde{\rho}_B}\min(\hat{\pi}_i)\kappa_1\hat{\rho}_{0,B}}}, \sqrt{\frac{1-\tilde{\rho}_A}{m\kappa_5}}, \frac{1}{2}\frac{1-\tilde{\rho}_B}{\sqrt{1+\tilde{\rho}_B}}\sqrt{\frac{\min(\hat{\pi}_i)}{2\kappa_5}}\right\}$. *Then the spectral radius of* $\boldsymbol{\Psi}^{(k)}$ *defined in Eq. 9 is strictly less than one, i.e.,* $\rho(\boldsymbol{\Psi}^{(k)}) < 1$. *The values of* κ_3-κ_5, $\hat{\rho}_{0,A}$, $\tilde{\rho}_A$ *and* $\tilde{\rho}_B$ *are given in Table 2 in Appendix B.6.*

Proposition 4.7 establishes a decreasing consensus error $\boldsymbol{\varsigma}^{(k)}$ in Eq. 9, capturing all parameters related to the non-convex loss and the directed graph in a constraint over the learning rate. We next analyze the global loss decrement using Proposition 4.7.

Proposition 4.8 (Loss Descent Constraints)**.** *Let Assumptions 4.1-4.4 hold, and each client* $i \in \mathcal{M}$ *use a constant learning rate* $\eta_i^{(k)} = \eta_i^{(0)}$ *that satisfies the conditions outlined in Proposition 4.7 to have* $\rho(\boldsymbol{\Psi}^{(k)}) = \rho_\Psi < 1$ *for all* $k \geq 0$. *Let the gradient calculation probabilities, communication probabilities, and learning rate heterogeneities satisfy* $p_i \in (\max\{0, 1 - \frac{1}{9\Gamma_1\delta_{1,i}^2}\}, 1]$, $\hat{p}_{ij} \in [\max\{\hat{r}_A, \hat{r}_B, 1-\hat{r}'_{ij}\}, 1]$ *and* $\frac{\bar{\eta}^{(0)}}{\eta_i^{(0)}} \geq \frac{1}{\Gamma_2}$ *for each client* $i \in \mathcal{M}$, *respectively, where* $\bar{\eta}^{(0)} = \frac{1}{m}\sum_{i=1}^m \eta_i^{(0)}$ *and the values for* $\Gamma_1 > 1$, $0 < \hat{r}_A, \hat{r}_B < 1$ *and* $\hat{r}'_{ij} > 0$ *are given in Table 2 in Appendix B.6, and* $\Gamma_2 > 1$ *is arbitrary. Also, let the learning rates further satisfy* $\eta_i^{(0)} < \min\left\{\frac{\min(\hat{\phi}_i\hat{\pi}_i)}{6\Gamma_1\Gamma_2\kappa_3\kappa_2 L}, \left(\frac{\kappa_9}{3\kappa_{10}}\right)^{\frac{1}{3}}, \sqrt{\frac{\kappa_9}{3\kappa_{11}}}, \frac{\kappa_9}{3\kappa_{12}}\right\}$, *where* κ_2 *and* κ_6-κ_{12} *are given in Table 2 in Appendix B.6. Then, the loss decrements at each iteration are lower bounded as* $q\max(\eta_i^{(0)})\mathbb{E}[\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\|^2]$, *where* $q = \frac{1}{8}m^2(1-\frac{1}{\Gamma_1})\frac{1}{\Gamma_2}\min(\hat{\phi}_i\hat{\pi}_i)$.

Note how a small gradient diversity, i.e., $\delta_{1,i}^2 < 1/(9\Gamma_1)$, imposes no constraint on the allowed computation probability $p_i \in (0, 1]$, and how coordinated learning rates $\eta_i^{(0)} = \bar{\eta}^{(0)}$ automatically satisfies $\bar{\eta}^{(0)}/\eta_i^{(0)} \geq 1/\Gamma_2$ since $\Gamma_2 > 1$.

Understanding the Learning Rate Constraints. In the less complex non-sporadic centralized/FL training with $p_i = \hat{p}_{ij} = 1$ and $\hat{\phi}_i = \hat{\pi}_i = \frac{1}{m}$, we get $\hat{\rho}_{0,A} = \hat{\rho}_{0,B} = \tilde{\rho}_A = \tilde{\rho}_B = \rho_\Psi = 0$ and all constants in Table 2 (Appendix B.6) reduce to simple functions of m and L_i.

Thus, the constraints in Proposition 4.7 and 4.8 simplify to $\eta_i^{(k)} < \frac{1/32}{\max(L_i)}$ and $\eta_i^{(0)} < \frac{1/18}{\max(L_i)}$, recovering the standard inverse dependence of the learning rate on smoothness (Bottou et al., 2018)..

4.4 Main Theorem

Theorem 4.9 (Convergence for Non-Convex Loss)**.** *Let Assumptions 4.1-4.4 hold, and each client $i \in \mathcal{M}$ use a constant learning rate $\eta_i^{(k)} = \eta_i^{(0)}$ that satisfies the constraints in Propositions 4.7 and 4.8, and let $\eta_{\max} = \max(\eta_i^{(0)})$. If the gradient calculation probabilities p_i, communication probabilities $\hat{p}_{ij}$, and learning rate heterogeneities $\bar{\eta}^{(0)}/\eta_i^{(0)}$ also satisfy the constraints of Proposition 4.8, the convergence of our approach to a stationary point can be characterized as*

$$\frac{1}{k+1}\sum_{r=0}^{k}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\right\|^2\right] \leq \frac{F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) - F^\star}{\mathcal{O}\left(\eta_{\max}(k+1)\right)} + \mathcal{O}(1+\eta_{\max})\frac{\varsigma_1}{k+1} \tag{10.1}$$

$$+\left(\mathcal{O}\left(\frac{\max(\hat{p}_{ij}(1-\hat{p}_{ij}))}{\eta_{\max}}\right) + \mathcal{O}(1+\eta_{\max})\right)\frac{\varsigma_0}{k+1} \tag{10.2}$$

$$+\Bigg(\mathcal{O}\left((1+\eta_{\max}+\eta_{\max}^2)\max(\hat{p}_{ij}(1-\hat{p}_{ij}))\right) + \mathcal{O}(\overline{(1-p)\delta_0^2}) \tag{10.3}$$

$$+\mathcal{O}\left(\eta_{\max}+\eta_{\max}^2+\eta_{\max}^3\right)\Bigg)\overline{(1-p)\delta_0^2}, \tag{10.4}$$

with a convergence rate of $\mathcal{O}(1/k)$ up to a neighborhood of the stationary points, whose gap we can control using the learning rates $\eta_i^{(0)}$ and the computation and communication frequencies p_i and $\hat{p}_{ij}$. Note that $\overline{(1-p)\delta_0^2} = \frac{1}{m}\sum_{i=1}^{m}(1-p_i)\delta_{0,i}^2$.

Theorem 4.9 states that `Spod-GT` achieves a stationary point of the non-convex loss, up to an asymptotic neighborhood controlled by the learning rate $\eta_{\max}$, and computation and communication probabilities p_i and $\hat{p}_{ij}$. Eq. 10 shows how the errors for optimality $F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) - F^\star$, model consensus ς_0 and GT consensus ς_1 decrease over time.

Recovering Existing Results. If clients perform GT in every iteration, i.e., $p_i = 1$, Eqs. 10.3 and 10.4 vanish since $\overline{(1-p)\delta_0^2} = 0$, and Eqs. 10.1 and 10.2 give a convergence rate of $\mathcal{O}(1/k)$ to the exact stationary points. Also, since $\hat{p}_{ij}(1-\hat{p}_{ij}) \leq 0.25$ always holds, communication sporadicity $\hat{p}_{ij}$ neither affects the convergence rate (governed by $\eta_{\max}$) nor the stationarity gap (governed by p_i), in line with the literature on gossip-like algorithms (Pu et al., 2020; Koloskova et al., 2020; Pu & Nedić, 2021). Our bounds further capture the initial model and GT consensus errors ς_0 and ς_1 in Eqs. 10.1 and 10.2, which vanish if all clients are identically initialized. Finally, in the non-sporadic setting ($p_i = 1$) with identical initializations ($\varsigma_0 = \varsigma_1 = 0$), Eq. 10 reduces to the convergence result of centralized/federated training with $\frac{1}{k+1}\sum_{r=0}^{k}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\|^2] \leq \frac{F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) - F^\star}{\mathcal{O}(\eta_{\max}(k+1))}$ (Bottou et al., 2018).

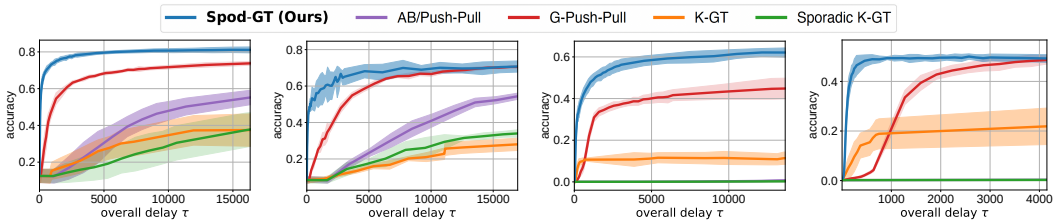
Corollary 4.10 (Arbitrarily Small Bound)**.** *Let $K \geq 0$ denote the total number of training iterations. If the gradient calculation probabilities are chosen to increase as $p_i = 1 - \mathcal{O}(1/(K+1))$ by each client $i \in \mathcal{M}$, then Theorem 4.9 implies that $\frac{1}{K+1}\sum_{r=0}^{K}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\|^2] \leq \frac{1}{\mathcal{O}(K+1)}((F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) - F^\star) + (\varsigma_0+\varsigma_1) + \overline{\delta_0^2})$, with a convergence rate of $\mathcal{O}(1/K)$. Therefore, by increasing K, we can decrease the stationarity gap to an arbitrarily small value, i.e., $\lim_{K\to\infty}\frac{1}{K+1}\sum_{r=0}^{K}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\|^2] = 0$.*

In Corollary 4.10, computation probabilities p_i asymptotically approach 1 to achieve an exact zero stationarity gap. However, the main goal in practical DFL deployments is achieving a higher target accuracy in a shorter time rather than exact convergence. Thus, the main motivation of `Spod-GT` is to continuously save resources by choosing a lower p_i for straggler clients as we show in Sec. 5.

5 Experiments

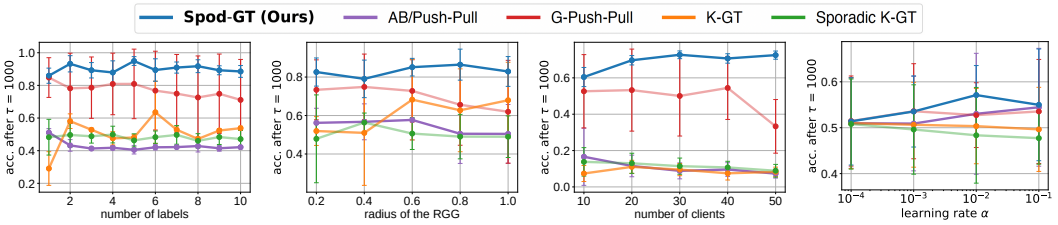
Models and Datasets. We consider image classification tasks using the Fashion-MNIST (FMNIST) (Xiao et al., 2017) and CIFAR-10 (Krizhevsky, 2009) datasets, training a Support Vector Machine (SVM) (Cortes & Vapnik, 1995) for FMNIST and ResNet-18 (He et al., 2016) for CIFAR-10.

Baselines. We compare `Spod-GT` with four baselines that use GT in directed graphs: (a) AB/Push-Pull (Nedić et al., 2025), where GT and local directed aggregations occur at every iteration; (b)



(a) FMNIST, IID. (b) FMNIST, Non-IID. (c) CIFAR-10, IID. (d) CIFAR-10, Non-IID.

Figure 2: Accuracy vs. latency plots. `Spod-GT` achieves the target accuracy much faster, emphasizing the benefit of sporadicity in DFL over directed graphs for GT iterations and aggregations simultaneously.



(a) Varying number of labels per client. (b) Varying the radius of random geometric graph. (c) Varying the number of clients m in the network. (d) Varying the learning rate η for all clients.

Figure 3: Effects of system parameters on FMNIST. The overall results confirm the advantage of `Spod-GT`.

the G-Push-Pull algorithm (Pu et al., 2020), which is a special case of our `Spod-GT` framework with sporadic aggregations but conducting GT at every iteration; (c) K-GT (Liu et al., 2025) with K local GT steps between aggregations; and (d) Sporadic K-GT, which is another special case of our `Spod-GT` framework with aggregations at every iteration but performing sporadic GT. See Appendix B.7 for additional settings of our experiments.

Accuracy vs. Total Delay Result. Fig. 2 illustrates the accuracy achieved per total delay for our `Spod-GT` algorithm and other GT baselines. We observe that when the clients in a DFL setup adopt our `Spod-GT` training algorithm, both for the Fashion-MNIST and the CIFAR-10 datasets, they are able to consistently achieve higher accuracies faster compared to their non-sporadic counterparts. For example, if we compare all algorithms after $\bar{\tau}_{\text{total}}^{(k)} = 5000$ in Figs. 2a-2d, we could see a 7-40% accuracy improvement for our `Spod-GT` algorithm compared to the second best baseline.

Ablation Studies. In Fig. 3, we conduct ablation studies on key system parameters for FMNIST. In Fig. 3a, we vary the number of labels per client from 1 (extreme non-IID) to 10 (IID), and `Spod-GT` consistently outperforms all baselines at all levels of data heterogeneity, confirming the benefit of sporadicity in both communications and computations. `Spod-GT` also performs the best for all radii of the random geometric graph, where a higher radius indicates denser connectivity (Fig. 3b), and maintains its superiority for larger networks of $m = 10$ to 50 clients (Fig. 3c). Finally, varying the learning rate η from 10^{-4} to 10^{-1} in Fig. 3d shows that $\eta = 0.01$ is the optimal choice, and that `Spod-GT` outperforms all baselines for any η.

Other Results. Appendix J provides additional results, including a ConvNeXt model (Liu et al., 2022) trained on Tiny-ImageNet (Le et al., 2015) and an ablation study on CIFAR-10.

6 Conclusion

We developed *Sporadic Gradient Tracking* (`Spod-GT`), a resource-efficient GT-based DFL algorithm designed for non-doubly stochastic directed graphs. By adjusting each client's participation in both (i) gradient computations and (ii) peer-to-peer communications according to its processing and transmission resource availabilities, respectively, our methodology achieves significant improvements in latency saving over non-heterogeneity-aware algorithms. We carried out a thorough analysis of our methodology, proving consensus and stationarity guarantees under intermittent client participation in DFL. We validate the efficacy of our approach through numerical experiments.

AI USE STATEMENT

In this work, we have **not** used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis and interpret results. Furthermore, assisting with translation and generating synthetic data sets are not applicable to this work.

REFERENCES

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pp. 344–353. PMLR, 2019.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

Bahman Gharesifard and Jorge Cortés. When does a digraph admit a doubly stochastic adjacency matrix? In *Proceedings of the 2010 American Control Conference*, pp. 2440–2445. IEEE, 2010.

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

Chung-Hsuan Hu, Zheng Chen, and Erik G Larsson. Scheduling and aggregation design for asynchronous federated learning over wireless networks. *IEEE Journal on Selected Areas in Communications*, 41(4):874–886, 2023.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International conference on machine learning*, pp. 5381–5393. PMLR, 2020.

Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34: 11422–11435, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Ya Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

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

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

Liyuan Liang, Gan Luo, and Kun Yuan. On the linear speedup of the push-pull method for decentralized optimization over digraphs. *arXiv preprint arXiv:2506.18075*, 2025.

Yue Liu, Tao Lin, Anastasia Koloskova, and Sebastian U Stich. Decentralized gradient tracking with local steps. *Optimization Methods and Software*, 40(5):1153–1180, 2025.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Grigory Malinovsky, Konstantin Mishchenko, and Peter Richtárik. Server-side stepsizes and sampling without replacement provably help in federated optimization. In *Proceedings of the 4th International Workshop on Distributed Machine Learning*, pp. 85–104, 2023.

Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on automatic control*, 54(1):48–61, 2009.

Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

Angelia Nedić, Duong Thuy Anh Nguyen, and Duong Tung Nguyen. Ab/push-pull method for distributed optimization in time-varying directed networks. *Optimization Methods and Software*, 40(5):1044–1071, 2025.

Duong Thuy Anh Nguyen, Su Wang, Duong Tung Nguyen, Angelia Nedić, and H Vincent Poor. Decentralized federated learning with gradient tracking over time-varying directed networks. *arXiv preprint arXiv:2409.17189*, 2024.

Duong Thuy Anh Nguyen, Duong Tung Nguyen, and Angelia Nedić. Distributed nash equilibrium seeking over time-varying directed communication networks. *IEEE Transactions on Control of Network Systems*, 2025.

Mathew Penrose. *Random geometric graphs*, volume 5. OUP Oxford, 2003.

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021.

Shi Pu, Wei Shi, Jinming Xu, and Angelia Nedić. Push–pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, 66(1):1–16, 2020.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Tao Sun, Dongsheng Li, and Bao Wang. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4289–4301, 2022.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pp. 4848–4856. PMLR, 2018.

Behrouz Touri and Angelia Nedić. Product of random stochastic matrices. *IEEE Transactions on Automatic Control*, 59(2):437–448, 2013.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Ran Xin and Usman A Khan. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2(3):315–320, 2018.

Ran Xin, Chenguang Xi, and Usman A Khan. Frost—fast row-stochastic optimization with uncoordinated step-sizes. *EURASIP Journal on Advances in Signal Processing*, 2019(1):1, 2019.

Ran Xin, Shi Pu, Angelia Nedić, and Usman A Khan. A general framework for decentralized optimization with first-order methods. *Proceedings of the IEEE*, 108(11):1869–1889, 2020.

Ran Xin, Usman A Khan, and Soummya Kar. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 69:1842–1858, 2021.

Runze You and Shi Pu. Stochastic push-pull for decentralized nonconvex optimization. *IEEE Transactions on Signal Processing*, 2026.

Shahryar Zehtabi, Dong-Jun Han, Rohit Parasnis, Seyyedali Hosseinalipour, and Christopher G Brinton. Decentralized sporadic federated learning: A unified algorithmic framework with convergence guarantees. In *ICLR*, 2025.

APPENDIX TABLE OF CONTENTS

A Stochastic Gradient Implications **14**

A.1 Convergence Analysis 14

B Supplementary Technical Details **16**

B.1 Notation 16

B.2 Heterogeneous and Dynamic GT Steps 16

B.3 Algorithm 16

B.4 Necessary Assumption 4.2 16

B.5 Relaxed Assumption Examples 17

B.6 Summary of Scalars 18

B.7 Experiment Settings 18

C Intermediary Lemmas **19**

D Main Lemmas **22**

E Corollaries to Proposition 4.7 **24**

F Proofs of Intermediary Lemmas **25**

F.1 Proof of Lemma C.1 25

F.2 Proof of Lemma C.2 26

F.3 Proof of Lemma C.4 28

F.4 Proof of Lemma C.5 32

G Proofs of Main Lemmas **34**

G.1 Proof of Lemma D.1 34

G.2 Proof of Lemma D.2 35

G.3 Proof of Lemma D.3 36

H Proofs of Propositions **38**

H.1 Proof of Proposition 4.7 38

H.2 Proof of Proposition 4.8 39

I Proof of Theorem 4.9 **41**

J Additional Experiments **42**

J.1 Ablation Study with CIFAR-10 42

J.2 ConvNeXt Model on the Tiny-ImageNet Dataset 42

A Stochastic Gradient Implications

In this section, we present the changes in our algorithm and theoretical analysis when using local stochastic gradients $\boldsymbol{g}_i^{(k)}$ as

$$\boldsymbol{g}_i^{(k)} = \frac{1}{B_i} \sum_{\xi \in \mathcal{B}_i^{(k)}} \nabla \mathcal{L}(\boldsymbol{x}_i^{(k)}, \xi), \tag{11}$$

where $\mathcal{B}_i^{(k)} \subseteq \mathcal{D}_i$ is the mini-batch with size $B_i = |\mathcal{B}_i^{(k)}|$. While the aggregation step in Eq. 1 and the update rule for model parameters $\boldsymbol{x}$ in Eq. 2 stay the same, the update rule for GT parameters $\boldsymbol{y}$ in Eq. 2 change to

$$\boldsymbol{y}_i^{(k+1)} = \boldsymbol{y}_i^{(k+\frac{1}{2})} + \boldsymbol{g}_i^{(k+1)} v_i^{(k+1)} - \boldsymbol{g}_i^{(k)} v_i^{(k)}, \tag{12}$$

where we initialize $\boldsymbol{y}_i^{(0)} = \boldsymbol{g}_i^{(0)} v_i^{(0)}$. Similar to the simplification we made in Eq. 3, Eq. 12 can be combined with Eq. 1 to have

$$\boldsymbol{y}_i^{(k+1)} = \sum_{j=1}^{m} \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)} + \boldsymbol{g}_i^{(k+1)} v_i^{(k+1)} - \boldsymbol{g}_i^{(k)} v_i^{(k)}. \tag{13}$$

Collecting the local stochastic gradients $\boldsymbol{g}_i^{(k)}$ of all clients to the matrix $\boldsymbol{G}^{(k)} = [\boldsymbol{g}_1^{(k)} \cdots \boldsymbol{g}_m^{(k)}]^T \in \mathbb{R}^{m \times n}$, the update rule for GT parameters $\boldsymbol{y}$ in Eq. 4 gets modified to the following when using Eq. 13

$$\boldsymbol{Y}^{(k+1)} = \hat{\boldsymbol{B}}^{(k)} \boldsymbol{Y}^{(k)} + \boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)} \boldsymbol{G}^{(k+1)} - \boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)} \boldsymbol{G}^{(k)}. \tag{14}$$

Following Definition 4.5, we define the weighted average of local stochastic gradients $\boldsymbol{g}_i^{(k)}$ as $\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)} = \frac{1}{m} \sum_{i=1}^{m} g_i^{(k)} v_i^{(k)} = \frac{1}{m} \mathbf{1}^T \boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)} \boldsymbol{G}^{(k)}$. Applying this average vector notation to the matrix form of updates given in Eq. 14, we get the following equation which is similar to Eq. 7:

$$\bar{\boldsymbol{y}}^{(k+1)} = \bar{\boldsymbol{y}}^{(k)} + \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k+1)} - \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)}. \tag{15}$$

Recursively expanding $\bar{\boldsymbol{y}}^{(k+1)}$ with the initialization discussed for Eq. 12 as $\bar{\boldsymbol{y}}^{(0)} = \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(0)}$, we will have

$$\bar{\boldsymbol{y}}^{(k)} = \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)}, \tag{16}$$

for all $k \geq 0$.

A.1 Convergence Analysis

Assumption A.1 (Stochastic Gradient Estimation)**.** The stochastic gradient $\nabla \mathcal{L} : \mathbb{R}^n \times \mathbb{R}^{|\xi|} \to \mathbb{R}^n$, where ξ is a data sample, is an unbiased estimator, i.e., for each client $i \in \mathcal{M}$ with dataset $\mathcal{D}_i$, we have $\mathbb{E}_{\xi \sim \mathcal{D}_i} [\nabla \mathcal{L}(\boldsymbol{x}, \xi)] = \nabla F_i(\boldsymbol{x})$, with $\boldsymbol{x} \in \mathbb{R}^n$. We further assume that the variance of the stochastic gradient $\nabla \mathcal{L}$ is bounded as $\mathbb{E}_{\xi \sim \mathcal{D}_i}[\|\nabla \mathcal{L}(\boldsymbol{x}, \xi) - \nabla F_i(\boldsymbol{x})\|^2] \leq \sigma_{0,i}^2 + \sigma_{1,i}^2 \|\nabla F(\boldsymbol{x})\|^2$, with $\sigma_{0,i}, \sigma_{1,i} > 0$ where $\nabla F(\boldsymbol{x})$ is the global gradient.

While most works in the literature consider a uniform bound on stochastic gradient estimation, i.e., $\sigma_{1,i}^2 = 0$, our Assumption A.1 relaxes it to the general case of $\sigma_{1,i}^2 > 0$ (Koloskova et al., 2020). When using a mini-batch $\mathcal{B}_i$ to estimate the gradient as in Eq. 11, the variance coefficients are decreased as $\sigma_{0,i}^2 / B_i$ and $\sigma_{1,i}^2 / B_i$ for sampling with replacement and $(1/B_i - 1/D_i)\sigma_{0,i}^2$ and $(1/B_i - 1/D_i)\sigma_{1,i}^2$ for sampling without replacement (Malinovsky et al., 2023). Thus, when using a full batch $\mathcal{B}_i = \mathcal{D}_i$ as done in our main text, the gradient estimation becomes exact, i.e., $\boldsymbol{g}_i^{(k)} = \nabla F_i(\boldsymbol{x}_i^{(k)})$ with the effective variance coefficients being $\sigma_{0,i}^2 = \sigma_{1,i}^2 = 0$. Without loss of generality, we will focus on a single sample SGD in subsequent analysis, adding discussions on the implications of mini-batch SGD when appropriate.

Corollary A.2 (Loss Descent with Stochastic Gradient)**.** *Let Assumptions 4.1-4.4 hold, and each client $i \in \mathcal{M}$ use stochastic gradients $\boldsymbol{g}_i^{(k)}$. Then, all conditions in Propositions 4.7 and 4.8 remain the same with the inclusion of all stochastic gradient variance terms $\sigma_{0,i}^2$ and $\sigma_{1,i}^2$ for the constant scalars in Tables 2 and 3 of B.6.*

Proof. See Appendices H.1 and H.2. □

Theorem A.3 (Convergence with Stochastic Gradients)**.** *Let Assumptions 4.1-4.4 hold, and each client $i \in \mathcal{M}$ use a constant learning rate $\eta_i^{(k)} = \eta_i^{(0)}$ that satisfies the constraints in Propositions 4.7 and 4.8, and let $\eta_{\max} = \max(\eta_i^{(0)})$. If the gradient calculation probabilities p_i, communication probabilities $\hat{p}_{ij}$, and learning rate heterogeneities $\bar{\eta}^{(0)}/\eta_i^{(0)}$ also satisfy the constraints of Proposition 4.8, the stationary point of our approach can be characterized in Eq. 17 as*

$$
\begin{aligned}
\frac{1}{k+1}\sum_{r=0}^{k}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\right\|^2\right] &\leq \frac{F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) - F^\star}{\mathcal{O}\left(\eta_{\max}(k+1)\right)} + \mathcal{O}(1+\eta_{\max})\frac{\varsigma_1}{k+1} \\
&+ \left(\mathcal{O}\left(\frac{\max(\hat{p}_{ij}(1-\hat{p}_{ij}))}{\eta_{\max}}\right) + \mathcal{O}(1+\eta_{\max})\right)\frac{\varsigma_0}{k+1} \\
&+ \Bigg(\mathcal{O}(\overline{(1-p)\delta_0^2}) + \mathcal{O}\left(\max(\hat{p}_{ij}(1-\hat{p}_{ij}))\left(1+\eta_{\max}+\eta_{\max}^2\right)\right) \\
&\qquad + \mathcal{O}\left(\eta_{\max}+\eta_{\max}^2+\eta_{\max}^3\right)\Bigg)\left(\overline{\sigma_0^2 p} + 3\overline{(1-p)\delta_0^2}\right) \\
&+ \mathcal{O}(1+\eta_{\max})\overline{\sigma_0^2 p},
\end{aligned}
\tag{17}
$$

with a convergence rate of $\mathcal{O}(1/k)$ up to a neighborhood of the stationary points, whose gap we can control using the learning rates $\eta_i^{(0)}$, computation and communication frequencies p_i and $\hat{p}_{ij}$, and the batch size B_i.

Proof. See Appendix I. □

Theorem A.3 states that `Spod-GT` achieves a stationary point of the non-convex loss, up to an asymptotic neighborhood controlled by the learning rate $\eta_{\max}$, and computation and communication probabilities p_i and $\hat{p}_{ij}$, and the batch size B_i. Eq. 17 shows how the errors for optimality $F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) - F^\star$, model consensus ς_0 and GT consensus ς_1 decrease over time.

Note that even in the simplified setting of $p_i = \hat{p}_{ij} = 1$, i.e., conventional DFL where all clients participate equally, the stationarity gap of Eq. 17 depends on $\mathcal{O}(1+\eta_{\max})\overline{\sigma_0^2}$, which could not be made zero even with a diminishing learning rate $\eta_{\max}$. This observation is also discussed in Remark 5 of Nguyen et al. (2024) following their main Theorem 1, which also analyzes the convergence of stochastic optimization in the simplified case of uniform bound on the stochastic gradient variance, i.e., with $\sigma_{1,i}^2 = 0$. The authors of Nguyen et al. (2024) mention that a decreasing gradient noise variance $\sigma_{0,i}^2$ is required for exact convergence. We note that using our formulation, this bound could be controlled with the batch size parameter B_i as $\mathcal{O}(\overline{\sigma_0^2(1-B/D)/B})$ for sampling without replacement, and $\mathcal{O}(\overline{\sigma_0^2/B})$ for sampling with replacement. This observation was not made in papers working with full-batch gradients $B_i = D_i$ such as Xin & Khan (2018); Pu et al. (2020); Nedić et al. (2025), failing to reveal any insights on the implications of stochastic approximation on convergence.

However, as shown in recent works of Liang et al. (2025); You & Pu (2026) which improve the convergence bounds of Push–Pull–type algorithms with stochastic gradients, the dependence of the bounds on the non-vanishing term $\overline{\sigma_0^2}$ is an artifact of the analysis, and not a weakness of the GT algorithm. The works in Liang et al. (2025); You & Pu (2026) introduce a new analytical framework by leveraging a multi-step descent lemma derived from smoothness, as stated in Assumption 4.1. In particular, Liang et al. (2025) elaborates that the iterative updates of Push-Pull GT over directed graphs is inherently not aligned with centralized SGD (unlike SGD over undirected graphs (Koloskova et al., 2020; Zehtabi et al., 2025), GT over undirected graph (Pu & Nedić, 2021; Koloskova et al., 2021) etc.), and thus a multi-step descent lemma is required to omit the effect of $\overline{\sigma_0^2}$ from the bounds.

Extending our sporadic approach to the multi-step descent framework of Liang et al. (2025); You & Pu (2026) would be an interesting future direction.

B SUPPLEMENTARY TECHNICAL DETAILS

B.1 NOTATION

We use boldface lowercase $\boldsymbol{z} \in \mathbb{R}^{n\times 1}$ for vectors and boldface uppercase $\boldsymbol{Z} \in \mathbb{R}^{m\times n}$ for matrices. Subscripts and superscripts are used to denote the client and the iteration, respectively, i.e., $\boldsymbol{z}_i^{(k)}$ belongs to client i at iteration k. We use the Frobenius norm and inner product for all matrices, i.e., $\|\boldsymbol{Z}\|^2 = \|\boldsymbol{Z}\|_F^2 = \sum_{i=1}^m \|\boldsymbol{Z}_i\|^2$ and $\langle \boldsymbol{Z}, \boldsymbol{W}\rangle = \langle \boldsymbol{Z}, \boldsymbol{W}\rangle_F = \sum_{i=1}^m \langle \boldsymbol{z}_i, \boldsymbol{w}_i\rangle$, where $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_m]^T$ and $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_m]^T$, and we drop the subscript F to avoid repetition. Given a non-negative vector $\boldsymbol{\phi} \in \mathbb{R}^{n\times 1}$, we define the row-based matrix norm and the inner product imposed by a vector as $\|\boldsymbol{Z}\|_{\boldsymbol{\phi}}^2 = \sum_{i=1}^m \phi_i \|\boldsymbol{z}_i\|^2$ and $\langle \boldsymbol{Z}, \boldsymbol{W}\rangle_{\boldsymbol{\phi}} = \sum_{i=1}^m \phi_i \langle \boldsymbol{z}_i, \boldsymbol{w}_i\rangle$, respectively, which is a generalization of the Frobenius norm and inner products if $\boldsymbol{\phi} = \mathbf{1}_m$. For a matrix $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_m]^T$, we define its row-based average as $\bar{\boldsymbol{z}} = \frac{1}{m}\sum_{i=1}^m \boldsymbol{z}_i$, and its weighted average imposed by a vector $\boldsymbol{\phi}$ as $\bar{\boldsymbol{z}}_{\boldsymbol{\phi}} = \sum_{i=1}^m \phi_i \boldsymbol{z}_i$. Also, given N vectors $\boldsymbol{z}_1, ..., \boldsymbol{z}_N$, we define their element-wise average as $\overline{z_1 ... z_N} = \frac{1}{m}\sum_{i=1}^m z_{1,i} ... z_{N,i}$. We use the notation $\boldsymbol{\Lambda}_{\boldsymbol{z}} = \mathrm{diag}(\boldsymbol{z}) \in \mathbb{R}^{n\times n}$ to indicate a diagonal matrix whose diagonal entries are the elements of the vector $\boldsymbol{z}$.

Let $\mathcal{F}_\sigma^{(k)}$ be the σ-algebra generated by the set of random variables $\{\mathcal{B}_i^{(r)}\}_{\substack{0\le r\le k-1\\ i\in\mathcal{M}}} \cup \{v_i^{(r)}\}_{\substack{0\le r\le k-1\\ i\in\mathcal{M}}} \cup \{\hat{v}_{ij}^{(r)}\}_{\substack{0\le r\le k-1\\ i,j\in\mathcal{M}}}$. If $z^{(k)}$ is an event in the filtered probability space with filtration $\mathcal{F}_\sigma^{(k)}$, we use the notation $\mathbb{E}[z^{(k)}]$ for brevity to denote the conditional expectation $\mathbb{E}[z^{(k)} \mid \mathcal{F}_\sigma^{(k)}]$. Let $D(\mathcal{G})$ and $K(\mathcal{G})$ denote the graph diameter and maximal edge utility in the digraph $\mathcal{G}$, respectively (see Nedić et al. (2025) for more details).

In our notation, model parameters, tracked gradients, local gradients and local stochastic gradients $\boldsymbol{x}_i^{(k)}, \boldsymbol{y}_i^{(k)}, \nabla F_i(\boldsymbol{x}_i^{(k)}), \boldsymbol{g}_i^{(k)} \in \mathbb{R}^{n\times 1}$ are column vectors, but their corresponding (weighted) averages $\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}, \bar{\boldsymbol{y}}^{(k)}, \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)} \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} \in \mathbb{R}^{1\times n}$ are row vectors.

B.2 HETEROGENEOUS AND DYNAMIC GT STEPS

Consider a setup of two clients with the following values for the indicator variables for the first 5 iterations: $v_1^{(0:4)} = [11110]$, $v_2^{(0:4)} = [11100\cdots]$, and $\hat{v}_{12}^{(0:4)} = \hat{v}_{21}^{(0:4)} = [00001\cdots]$. In this setup, which corresponds to a communication round spanning 5 iterations, the clients take a heterogeneous number of local steps 4 and 3.

Now, assume that these indicator variables change next as $v_1^{(5:14)} = [1111111000]$, $v_2^{(5:14)} = [1100000000\cdots]$, and $\hat{v}_{12}^{(5:14)} = \hat{v}_{21}^{(5:14)} = [0000000001\cdots]$. Hence, the next communication round spans 10 iterations, and each client takes a heterogeneous number of local steps 7 and 2. This shows that not only could the number of local steps be heterogeneous, but also that they could be dynamically changing throughout the training.

B.3 ALGORITHM

We present the complete steps the `Spod-GT` algorithm in Alg. 1, explicitly demonstrating what each client does when it decides to participate or not to participate in DFL training at a given iteration.

B.4 NECESSARY ASSUMPTION 4.2

When working with sporadic methods where clients may opt out of computing gradient at some iterations, an Assumption over the local gradient norms is necessary to prevent the training process from being skewed toward certain clients participating with a higher frequency. This is in contrast to conventional non-sporadic methods where all clients participate at every iteration. In fact, Lemma C.1-(iii) in our paper captures this phenomenon by setting all $p_i = 1$. In that case, $(1/m)\sum_{i=1}^m \nabla F_i(\bar{x}_{\hat{\phi}}^{(k)}) v_i^{(k)} = \nabla F(\bar{x}_{\hat{\phi}}^{(k)})$, since $v_i^{(k)} = 1$ for all clients at all iterations, and an Assumption on individual norms is no longer needed. One can check that all our subsequent Lemmas in Appendices C and D use Assumption 4.2 through Lemma C.1-(iii), so they all include the terms

Algorithm 1: Sporadic Gradient Tracking (`Spod-GT`)

Input: K, $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, $\{\mathcal{D}_i\}_{i\in\mathcal{M}}$, $\{B_i\}_{i\in\mathcal{M}}$, $\{p_i\}_{i\in\mathcal{M}}$, $\{\hat{p}_{ij}\}_{(i,j)\in\mathcal{E}}$, $\{\eta_i^{(k)}\}_{\substack{0\le k\le K-1\\ i\in\mathcal{M}}}$;

Output: $\{\boldsymbol{x}_i^{(K)}\}_{i\in\mathcal{M}}$;

Initialize: $k \leftarrow 0$, $\{\boldsymbol{x}_i^{(0)} \in \mathbb{R}^n\}_{i\in\mathcal{M}}$, $\{\boldsymbol{y}_i^{(0)} \leftarrow \boldsymbol{g}_i^{(0)} = \boldsymbol{0}_n\}_{i\in\mathcal{M}}$, $\{v_i^{(0)} \in \{0,1\}\}_{i\in\mathcal{M}}$, $\{\hat{v}_{ij}^{(0)} \in \{0,1\}\}_{(i,j)\in\mathcal{E}}$;

foreach $i \in \mathcal{M}$ **do**

  if $v_i^{(0)} = 1$ **then** $\boldsymbol{y}_i^{(0)} \leftarrow \boldsymbol{g}_i^{(0)} = \frac{1}{B_i}\sum_{\xi_{ij}^{(0)}\in\mathcal{B}_i^{(0)}} \nabla\mathcal{L}(\boldsymbol{x}_i^{(0)}, \xi_{ij}^{(0)})$ where $\mathcal{B}_i^{(0)} \sim \mathcal{D}_i$;

  foreach $j \in \mathcal{N}_i^{\text{in}}$ **do**: $a_{ij} \leftarrow 1/(1 + |\mathcal{N}_i^{\text{in}}|)$;

  foreach $j \in \mathcal{N}_i^{\text{out}}$ **do**: $b_{ji} \leftarrow 1/(1 + |\mathcal{N}_i^{\text{out}}|)$;

while $k \le K - 1$ **do**

  foreach $i \in \mathcal{M}$ **do**

    $\boldsymbol{x}_i^{(k+\frac{1}{2})} \leftarrow \boldsymbol{x}_i^{(k)}$, $\boldsymbol{y}_i^{(k+\frac{1}{2})} \leftarrow \boldsymbol{y}_i^{(k)}$;

    foreach $j \in \mathcal{N}_i^{\text{in}}$ **do**: **if** $\hat{v}_{ij}^{(k)} = 1$ **then** $\boldsymbol{x}_i^{(k+\frac{1}{2})} \leftarrow \boldsymbol{x}_i^{(k+\frac{1}{2})} + a_{ij}\left(\boldsymbol{x}_j^{(k)} - \boldsymbol{x}_i^{(k)}\right)$, $\boldsymbol{y}_i^{(k+\frac{1}{2})} \leftarrow \boldsymbol{y}_i^{(k+\frac{1}{2})} + b_{ij}\boldsymbol{y}_j^{(k)}$;

    foreach $j \in \mathcal{N}_i^{\text{out}}$ **do**: **if** $\hat{v}_{ji}^{(k)} = 1$ **then** $\boldsymbol{y}_i^{(k+\frac{1}{2})} \leftarrow \boldsymbol{y}_i^{(k+\frac{1}{2})} - b_{ji}\boldsymbol{y}_i^{(k)}$;

  foreach $i \in \mathcal{M}$ **do**: $\boldsymbol{x}_i^{(k+1)} \leftarrow \boldsymbol{x}_i^{(k+\frac{1}{2})} - \eta_i^{(k)}\boldsymbol{y}_i^{(k+\frac{1}{2})}$, $\boldsymbol{y}_i^{(k+1)} \leftarrow \boldsymbol{y}_i^{(k+\frac{1}{2})} - \boldsymbol{g}_i^{(k)}$;

  foreach $i \in \mathcal{M}$ **do**

    if $U_i^{(k+1)} \le p_i$ *where* $U_i^{(k+1)} \sim \text{Uniform}(0,1]$ **then**

      $\boldsymbol{g}_i^{(k+1)} \leftarrow \frac{1}{B_i}\sum_{\xi_{ij}^{(k+1)}\in\mathcal{B}_i^{(k+1)}} \nabla\mathcal{L}(\boldsymbol{x}_i^{(k+1)}, \xi_{ij}^{(k+1)})$ where $\mathcal{B}_i^{(k+1)} \sim \mathcal{D}_i$;

      $v_i^{(k+1)} \leftarrow 1$, $\boldsymbol{y}_i^{(k+1)} \leftarrow \boldsymbol{y}_i^{(k+1)} + \boldsymbol{g}_i^{(k+1)}$;

    else $\boldsymbol{g}_i^{(k+1)} \leftarrow \boldsymbol{0}_n$, $v_i^{(k+1)} \leftarrow 0$;

    foreach $j \in \mathcal{N}_i^{out}$ **do**

      if $\hat{U}_{ij}^{(k+1)} \le \hat{p}_{ij}$ where $\hat{U}_{ij}^{(k+1)} \sim \text{Uniform}(0,1]$ **then** $\hat{v}_{ij}^{(k+1)} \leftarrow 1$;

      else $\hat{v}_{ij}^{(k+1)} \leftarrow 0$;

  $k \leftarrow k + 1$;

$(1-p_i)\delta_{0,i}^2$ and $(1-p_i)\delta_{1,i}^2$ in their upper bounds, i.e., terms which would be zero in conventional non-sporadic methods with $p_i = 1$ Xin & Khan (2018); Pu et al. (2020); Koloskova et al. (2021); Pu & Nedić (2021); Nedić et al. (2025); Liang et al. (2025); You & Pu (2026).

B.5 Relaxed Assumption Examples

Assumptions 4.2 and A.1 are relaxed since they do not enforce a uniform bound on gradient diversity (Koloskova et al., 2020) and gradient variance (Bottou et al., 2018) via the inclusion of non-zero terms $\delta_{1,i}^2$ and $\sigma_{1,i}^2$. Consider the following two examples where a uniform bound on these quantities would fail.

Example 1. Consider local losses $F_i(\boldsymbol{x}) = a_i F_0(\boldsymbol{x})$ with $a_i > 0$, where $F_0(\boldsymbol{x})$ is a common loss among all clients. Then, the local and global gradients will be $\nabla F_i(\boldsymbol{x}) = a_i \nabla F_0(x)$ and $\nabla F(\boldsymbol{x}) = \bar{a}\nabla F_0(x)$, where $\bar{a} = (1/m)\sum_{i=1}^m a_i$. Thus, $\|\nabla F_i(\boldsymbol{x})\|^2 \le (a_i^2/\bar{a}^2)\|\nabla F(\boldsymbol{x})\|^2$, with $\delta_{0,i} = 0$ and $\delta_{1,i} = a_i/\bar{a}$. We see how assuming a uniform bound with only $\delta_{0,i}^2$, i.e., $\delta_{1,i}^2 = 0$, can be restrictive.

Example 2. Consider for a data point $\xi \in \mathcal{D}_i$, we have $\nabla\mathcal{L}(\boldsymbol{x}, \xi) = a(\xi)\nabla F_i(\boldsymbol{x}) + b(\xi)$, where $\mathbb{E}[a(\xi)] = 1$, $\mathbb{E}[b(\xi)] = 0$ and $\mathbb{E}[a(\xi)b(\xi)] = 0$. In other words, the stochastic gradient $\nabla\mathcal{L}(\boldsymbol{x}, \xi)$ is an unbiased estimator of the gradient $F_i(\boldsymbol{x})$, with the multiplicative noise coefficient $a(\xi)$ and the additive noise coefficient $b(\xi)$ not uncorrelated. The variance of the stochastic gradient would then be $\mathbb{E}[\|\nabla\mathcal{L}(\boldsymbol{x}, \xi) - \nabla F_i(\boldsymbol{x})\|^2] = \mathbb{E}[(a(\xi) - 1)^2]\|\nabla F_i(\boldsymbol{x})\|^2 + \mathbb{E}[b(\xi)^2]$, with $\sigma_{0,i}^2 = \mathbb{E}[b(\xi)^2]$ and

$\sigma_{1,i}^2 = \mathbb{E}[(a(\xi)-1)^2]$. Noting that unless $a(\xi) = 1$ for all data $\xi \in \mathcal{D}_i$, then assuming a uniform bound with only $\sigma_{0,i}^2$, i.e., $\sigma_{1,i}^2 = 0$, is restrictive.

B.6 Summary of Scalars

We define several constant scalars throughout our theoretical analysis that depend on the parameters of our Assumptions given in Sec. 4.1. Table 2 outlines the auxiliary constant scalars used in our Intermediary Lemmas and Propositions that are presented in Appendix C and Sec. 4.3, respectively. Moreover, Table 3 includes the constant scalars used in our Main Lemmas that are discussed in Appendix. D, which are in turn are central to the proofs of our Propositions.

Table 2: Auxiliary constant scalars defined throughout our intermediary Lemmas (Appendix. C) and Propositions (Sec. 4.3).

Symbol	Value
κ_1	$\max(\frac{p_i}{\hat{\phi}_i}(\frac{2}{m}\sigma_{1,i}^2\bar{L}^2 + 3L_i^2))$
κ_2	$\frac{2}{m}\overline{\sigma_1^2 p} + 3\overline{(1-p)\delta_1^2} + 3$
κ_3	$\max(\hat{\phi}_i)(\frac{\max(\hat{\pi}_i)^2}{\min(\hat{\pi}_i)} + 2m\max(\hat{\pi}_j\overline{b_{:j}^2\hat{p}_{:j}(1-\hat{p}_{:j})}))$
κ_4	$\max\left(\frac{p_i}{\hat{\phi}_i}\left(3\sigma_{1,i}^2\bar{L}^2 + 4L_i^2\right)\right)$
κ_5	$\overline{\sigma_1^2 p}\bar{L}^2 + \overline{pL^2}$
$\hat{\rho}_{0,A}$	$4(m-1)\frac{\max(\hat{\phi}_i)}{\min(\hat{\phi}_i)}\max(a_{ij}^2\hat{p}_{ij}(1-\hat{p}_{ij}))$
$\hat{\rho}_{0,B}$	$2(m-1)\frac{\max(\hat{\pi}_i)}{\min(\hat{\pi}_i)}\max(b_{ij}^2\hat{p}_{ij}(1-\hat{p}_{ij}))$
$\tilde{\rho}_A$	$\hat{\rho}_A + \hat{\rho}_{0,A}$
$\tilde{\rho}_B$	$\hat{\rho}_B + \hat{\rho}_{0,B}$
κ_6	$\frac{2}{1-\tilde{\rho}_A}$
κ_7	$2(m + \min(\hat{\pi}_i)^{-1})\frac{1+\tilde{\rho}_B}{1-\tilde{\rho}_B}$
κ_8	$\frac{m}{2}\frac{1+\tilde{\rho}_B}{\tilde{\rho}_B}$
κ_9	$\frac{1-\rho_\Psi}{\Gamma_2\Gamma_3}\left(1-\frac{1}{\Gamma_1}\right)\left(1-\frac{1}{\Gamma_4}\right)\min(\hat{\phi}_i\hat{\pi}_i)$
κ_{10}	$2\left(m\kappa_1\kappa_6 + \kappa_7\kappa_4\kappa_6 + 16m\kappa_7\kappa_5\right)\kappa_3^2\kappa_2\bar{L}$
κ_{11}	$3\left(\left(\max\left(\frac{p_i}{\hat{\phi}_i}L_i^2\right) + \hat{\rho}_{0,B}\kappa_1\right)\kappa_6 + \left(\frac{\kappa_4\kappa_6}{m} + 16\kappa_5\right)\tilde{\rho}_B\kappa_7\right)\max(\hat{\phi}_i)\kappa_3\kappa_2$
κ_{12}	$\frac{2}{m}\left(2\kappa_7(3\overline{\sigma_1^2 p} + 4\overline{p(1-p)\delta_1^2}) + m\kappa_2\kappa_6\hat{\rho}_{0,A} + \kappa_8\kappa_2\hat{\rho}_{0,B}\right)\kappa_3\bar{L}$
$\hat{\tau}_A$	$8m^2(m-1)D(\mathcal{G})K(\mathcal{G})\frac{\max(a_{ij}^+)^2}{\min(a_{ij}^+)^{2(m+1)}}$
$\hat{\tau}_B$	$4m^3(m-1)D(\mathcal{G})K(\mathcal{G})\frac{\max(b_{ij}^+)^2}{\min(b_{ij}^+)^{3m+2}}$
$\hat{r}_A$	$f_A(\hat{r}_A) = 0; f_A(x) = x^{2(m+1)} + \hat{\tau}_A x - \hat{\tau}_A$
$\hat{r}_B$	$f_B(\hat{r}_B) = 0; f_B(x) = x^{3m+2} + \hat{\tau}_B x - \hat{\tau}_B$
$\hat{r}'_{ij}$	$\frac{1}{18\Gamma_1\Gamma_2\kappa_2 b_{ij}^2}\frac{m}{m-1}\frac{\min(\hat{\phi}_i)}{\max(\hat{\phi}_i)}\frac{\min(\hat{\pi}_i)^2}{\max(\hat{\pi}_i)}$
Γ_1	$1 + \frac{2\Gamma_4\Gamma_3\tilde{\rho}_B}{m(1-\rho_\Psi)}\max\left\{\frac{6\Gamma_2\kappa_7}{m}\frac{\max(\hat{\phi}_i)}{\min(\hat{\phi}_i\hat{\pi}_i)}(3\sigma_{1,i}^2 + 4\delta_{1,i}^2), \frac{\kappa_8}{3}\right\}$
$\Gamma_2, \Gamma_3, \Gamma_4$	$\Gamma_2, \Gamma_3, \Gamma_4 > 1$

B.7 Experiment Settings

By default, we consider $m = 10$ clients connected via a random geometric graph (RGG) (Penrose, 2003). We design our directed network graph $\mathcal{G}$ by first generating an undirected RGG, then making all edges bidirectional with heterogeneous weights for each link. We adopt a constant learning rate $\eta = 0.01$ and a batch size of $B = 16$ for all clients. The probability of gradient computation p_i for each client and the probability of transmission $\hat{p}_{ij}$ through a link $(i,j) \in \mathcal{E}$ on the directed graph are randomly chosen according to the Beta distribution $\mathrm{Beta}(0.5, 0.5)$, yielding an inverted bell-shaped distribution, corresponding to significant resource heterogeneity. We consider two dif-

Table 3: Core constant scalars central to our main Lemmas (Appendix D).

Symbol	Value
$\psi_{11}^{(k)}$	$\frac{1+\tilde{\rho}_A}{2} + m\kappa_1\kappa_6\kappa_3 \max(\eta_i^{(k)})^2$
$\psi_{12}^{(k)}$	$\kappa_6\kappa_3 \max(\eta_i^{(k)})^2$
$\gamma_1^{(k)}$	$m^2\kappa_2\kappa_6\kappa_3 \max(\eta_i^{(k)})^2$
$\omega_1^{(k)}$	$m^2\kappa_6\kappa_3 \max(\eta_i^{(k)})^2(\frac{1}{m}\overline{\sigma_0^2 p} + 3\overline{(1-p)\delta_0^2})$
$\psi_{21}^{(k)}$	$\kappa_7\kappa_4(1+\psi_{11}^{(k)}) + 16m\kappa_7\kappa_5\hat{\rho}_{0,A} + \kappa_8\kappa_1\hat{\rho}_{0,B} + 16m^2\kappa_7\kappa_5\kappa_3\kappa_1 \max(\eta_i^{(k)})^2$
$\psi_{22}^{(k)}$	$\frac{1+\tilde{\rho}_B}{2} + \kappa_7\kappa_4\psi_{12}^{(k)} + 16m\kappa_7\kappa_5\kappa_3 \max(\eta_i^{(k)})^2$
$\gamma_2^{(k)}$	$\kappa_7\kappa_4\gamma_1^{(k)} + 2m\kappa_7(3\overline{\sigma_1^2 p} + \overline{4p(1-p)\delta_1^2}) + m\kappa_8\kappa_2\hat{\rho}_{0,B} + 16m^3\kappa_7\kappa_5\kappa_3\kappa_2 \max(\eta_i^{(k)})^2$
$\omega_2^{(k)}$	$\kappa_7\kappa_4\omega_1^{(k)} + 2m\kappa_7(\overline{\sigma_0^2 p} + \overline{4p(1-p)\delta_0^2}) + m\kappa_8\hat{\rho}_{0,B}(\overline{\sigma_0^2 p} + 3\overline{(1-p)\delta_0^2})$ $+ 16m^3(\frac{1}{m}\overline{\sigma_0^2 p} + 3\overline{(1-p)\delta_0^2})\kappa_7\kappa_5\kappa_3 \max(\eta_i^{(k)})^2$
$\psi_{01}^{(k)}$	$\frac{1}{2}(3m\overline{\hat{\pi}\eta^{(k)}} \max(p_i L_i^2) + \bar{L}\hat{\rho}_{0,A} + (3\max(\hat{\phi}_i\eta_i^{(k)})\hat{\rho}_{0,B} + 2m\bar{L}\kappa_3 \max(\eta_i^{(k)})^2)\kappa_1)$
$\psi_{02}^{(k)}$	$\frac{1}{m}(\frac{3}{2}\max(\hat{\phi}_i\eta_i^{(k)})(\hat{\rho}_B + \hat{\rho}_{0,B}) + m\bar{L}\kappa_3 \max(\eta_i^{(k)})^2)$
$\gamma_0^{(k)}$	$\frac{1}{2}m(m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}} - 3m\hat{\phi}\hat{\pi}\eta^{(k)}(1-p)\delta_1^2 - (3\max(\hat{\phi}_i\eta_i^{(k)})\hat{\rho}_{0,B} + 2m\bar{L}\kappa_3 \max(\eta_i^{(k)})^2)\kappa_2)$
$\omega_0^{(k)}$	$m(\frac{3}{2}m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}(1-p)\delta_0^2} + \frac{3}{2}\max(\hat{\phi}_i\eta_i^{(k)})\hat{\rho}_{0,B} + m\bar{L}\kappa_3 \max(\eta_i^{(k)})^2)(\overline{\frac{\sigma_0^2 p}{m}} + 3\overline{(1-p)\delta_0^2})$

ferent data distribution scenarios: (i) IID, where each client receives samples from all 10 classes in the dataset, and (ii) non-IID, where each client receives samples from just 1 class. We measure the test accuracy of each scheme achieved over the average total delay incurred up to iteration k. Specifically, $\bar{\tau}_{\text{total}}^{(k)} = \bar{\tau}_{\text{trans, in}}^{(k)} + \bar{\tau}_{\text{proc}}^{(k)} + \bar{\tau}_{\text{trans, out}}^{(k)}$, in which $\bar{\tau}_{\text{trans, in}}^{(k)} = \frac{1}{m}\sum_{i=1}^m (1/|\mathcal{N}_i^{\text{in}}|)\sum_{j=1}^m (\hat{v}_{ij}^{(k)}/\hat{p}_{ij})$, $\bar{\tau}_{\text{trans, out}}^{(k)} = \frac{1}{m}\sum_{i=1}^m (1/|\mathcal{N}_i^{\text{out}}|)\sum_{j=1}^m (\hat{v}_{ji}^{(k)}/\hat{p}_{ji})$ and $\bar{\tau}_{\text{proc}}^{(k)} = \frac{1}{m}\sum_{i=1}^m (v_i^{(k)}/p_i)$ are the transmission delays per-client incurred across the links of in-neighbors and out-neighbors, and the processing delays incurred across clients in iteration k, respectively. For a fair comparison, we determine the aggregation interval K for the K-GT algorithm based on these p_i, i.e., $K = \lceil \frac{1}{m}\sum_{i=1}^m (1/p_i) \rceil$. We conduct the experiments based on a cluster of three NVIDIA A100 GPUs with 40GB memory. We run the experiments five times in each setup and present the mean and one-sigma standard deviation error bars.

C Intermediary Lemmas

In Definition 4.5 we see that while we took the regular average of $\boldsymbol{Y}^{(k)}$, we needed to take the $\hat{\boldsymbol{\phi}}$-weighted average of $\boldsymbol{X}^{(k)}$ to arrive at the average iterative updates in Eq. 7. To analyze the convergence of our methodology under these expressions, we need to obtain an upper bound on $\|\sum_{j=1}^m b_{ij}\boldsymbol{y}_j^{(k)} - \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\|^2$, which is the difference between the aggregated gradient tracking at client $i \in \mathcal{M}$ and the global gradient in any iteration $k \geq 0$. In order to obtain this bound using Assumptions 4.1-4.4, we need to break it down to multiple steps, i.e.,

$$\sum_{j=1}^m b_{ij}\boldsymbol{y}_j^{(k)} \overset{(a)}{\to} \bar{\boldsymbol{y}}^{(k)} \overset{(b)}{\to} \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)} \overset{(c)}{\to} \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} \overset{(d)}{\to} \overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} \overset{(e)}{\to} \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}). \tag{18}$$

We will postpone discussing (a) to Lemma C.4, and the equality for (b) was obtained in Eq. 8.

Lemma C.1 (Sporadic Gradient Tracking Discrepancy)**.** *Let $\hat{\boldsymbol{\phi}}$ be a positive stochastic vector. The following bounds hold for the steps (c)-(e) of Eq. 18:*

(i) Let Assumptions 4.1, A.1 and 4.3 hold. We have

$$\mathbb{E}\left[\left\|\bar{\boldsymbol{g}}_v^{(k)} - \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}}\right\|^2\right] \leq \frac{1}{m}\overline{\sigma_0^2 p} + \frac{2\bar{L}^2}{m^2}\max\left(\frac{p_i}{\hat{\phi}_i}\sigma_{1,i}^2\right)\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] + \frac{2}{m}\overline{\sigma_1^2 p}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right]. \tag{19}$$

(ii) Let Assumptions 4.1 and 4.3 hold. We have

$$\mathbb{E}\left[\left\|\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} - \overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}}\right\|^2\right] \leq \frac{1}{m}\max\left(\frac{p_i}{\hat{\phi}_i}L_i^2\right)\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right]. \tag{20}$$

(iii) Let Assumptions 4.2 and 4.3 hold. We have

$$\mathbb{E}\left[\left\|\overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} - \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right] \leq \overline{(1-p)\delta_0^2} + \overline{(1-p)\delta_1^2}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right]. \tag{21}$$

Proof. See Appendix F.1 □

We observe that when clients conduct GT in every training iteration, i.e., $p_i = 1$ for all $i \in \mathcal{M}$, the upper bound in Lemma C.1-(iii) becomes zero. On the other hand, if the gradient estimation variance is uniformly bounded, i.e., $\sigma_{1,i}^2 = 0$, we get a uniform bound for Lemma C.1-(i). Also, when clients use a full batch, i.e., $B_i = D_i$, the bound in Lemma C.1-(i) becomes zero.

Next, we obtain several key upper bounds involving the gradient tracking parameters $\boldsymbol{y}_i^{(k)}$.

Lemma C.2 (Descent Effect on Model Consensus)**.** *Let $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\pi}}$ be positive stochastic vectors. Then, the following bounds hold:*

(i) Let Assumptions 4.1-4.3 hold. We will then have according to Lemma C.1 that

$$\mathbb{E}\left[\|\bar{\boldsymbol{y}}^{(k)}\|^2\right] \leq \frac{1}{m}\overline{\sigma_0^2 p} + 3\overline{(1-p)\delta_0^2} + \frac{\kappa_1}{m}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] + \kappa_2\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right], \tag{22}$$

where $\kappa_1 = \max(\frac{p_i}{\hat{\phi}_i}(\frac{2}{m}\sigma_{1,i}^2\bar{L}^2 + 3L_i^2))$ and $\kappa_2 = \frac{2}{m}\overline{\sigma_1^2 p} + 3\overline{(1-p)\delta_1^2} + 3$.

(ii) The $\hat{\boldsymbol{\pi}}^{-1}$-weighted norm of gradient tracking parameters $\boldsymbol{y}_i^{(k)}$ can be upper bounded as

$$\left\|\boldsymbol{Y}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2 = \left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^2 + m^2\left\|\bar{\boldsymbol{y}}^{(k)}\right\|^2. \tag{23}$$

(iii) Let Assumption 4.3 hold. The descent step deviation from its $\hat{\boldsymbol{\phi}}$-weighted average is bounded as

$$\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - \mathbf{1}\hat{\boldsymbol{\phi}}^T\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] \leq \kappa_3\max(\eta_i^{(k)})^2\mathbb{E}\left[\left\|\boldsymbol{Y}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2\right], \tag{24}$$

where $\kappa_3 = \max(\hat{\phi}_i)(\frac{\max(\hat{\pi}_i)^2}{\min(\hat{\pi}_i)} + 2m\max(\hat{\pi}_j\overline{\hat{b}_{:j}^2\hat{p}_{:j}(1-\hat{p}_{:j})}))$.

Proof. See Appendix F.2 □

In Eq. 4 where we obtained the matrix update equations for `Spod-GT`, we see that the aggregations are governed by the row-stochastic matrix $\hat{\boldsymbol{A}}^{(k)}$ and column-stochastic matrix $\hat{\boldsymbol{B}}^{(k)}$. Furthermore, in Eq. 7, we obtained the weighted average of model parameters $\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}$ and the average of GT parameters $\hat{\boldsymbol{y}}^{(k)}$. In the following Lemma, we first characterize error upper bounds on the expected mixing matrices $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, which were defined in Eq. 5.

Lemma C.3 (Expected Matrix Contraction). *Let Assumption 4.4 hold. Let the row-stochastic $\hat{\boldsymbol{A}}$ and column-stochastic $\hat{\boldsymbol{B}}$ expected mixing matrices (Eq. 5) be designed to be irreducible and primitive such that $\hat{\boldsymbol{\phi}}$ is the left-eigenvector of $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{\pi}}$ is the right-eigenvector of $\hat{\boldsymbol{B}}$ (Eq. 6). Then, the following bounds will hold for the vector-induced matrix norms:*

(i) $\left\|\hat{\boldsymbol{A}}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}^{(k)}_{\hat{\boldsymbol{\phi}}}\right\|^2_{\hat{\boldsymbol{\phi}}} \leq \hat{\rho}_A \left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}^{(k)}_{\hat{\boldsymbol{\phi}}}\right\|^2_{\hat{\boldsymbol{\phi}}}$,

(ii) $\left\|\boldsymbol{\Lambda}^{-1}_{\hat{\boldsymbol{\pi}}}\hat{\boldsymbol{B}}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|^2_{\hat{\boldsymbol{\pi}}} \leq \hat{\rho}_B \left\|\boldsymbol{\Lambda}^{-1}_{\hat{\boldsymbol{\pi}}}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|^2_{\hat{\boldsymbol{\pi}}}$.

where $\boldsymbol{X}^{(k)}, \boldsymbol{Y}^{(k)} \in \mathbb{R}^{m\times n}$ are a collection of m arbitrary vectors and $0 \leq \hat{\rho}_A, \hat{\rho}_B < 1$. In particular, we have $\hat{\rho}_A = 1 - \frac{\min(\hat{\phi}_i)\min(a^+_{ij}\hat{p}_{ij})^2}{\max(\hat{\phi}_i)^2 D(\mathcal{G})K(\mathcal{G})}$ and $\hat{\rho}_B = 1 - \frac{\min(\hat{\pi}_i)^2\min(b^+_{ij}\hat{p}_{ij})^2}{\max(\hat{\pi}_i)^3 D(\mathcal{G})K(\mathcal{G})}$, in which $D(\mathcal{G})$ is the diameter of graph $\mathcal{G}$ and $K(\mathcal{G})$ is its maximal edge utility.

Proof. See Lemmas 4.2 and 4.5 in Nedić et al. (2025) applied to expected mixing matrices in Eq. 5, noting that $\hat{a}_{ij} = a_{ij}\hat{p}_{ij}$ and $\hat{b}_{ij} = b_{ij}\hat{p}_{ij}$. □

In the simpler case where the directed graph $\mathcal{G}$ admits a doubly-stochastic mixing matrix Gharesifard & Cortés (2010); Xin et al. (2021), we would have $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\pi}} = \frac{1}{m}\mathbf{1}$, and $\hat{\rho}_A$ and $\hat{\rho}_B$ in Lemma C.3 would correspond to the spectral radius of the matrices $\hat{\boldsymbol{A}} - \frac{1}{m}\mathbf{1}\mathbf{1}^T$ and $\hat{\boldsymbol{B}} - \frac{1}{m}\mathbf{1}\mathbf{1}^T$, respectively. Furthermore, if $\mathcal{G}$ is a fully-connected graph, which results in $D(\mathcal{G}) = K(\mathcal{G}) = 1$, and we assign the mixing weights to $a_{ij} = b_{ij} = 1/m$ for all $(i,j) \in \mathcal{E}$, the spectral radii simplify to $\hat{\rho}_A = \hat{\rho}_B = 1 - \min(\hat{p}_{ij})^2$. Notice how assigning $\hat{p}_{ij} = 1$ for all the edges further gives us $\hat{\rho}_A = \hat{\rho}_B = 0$.

Having established a contraction property for the expected mixing matrices in Lemma C.3, we now extend it to sporadic mixing matrices $\hat{\boldsymbol{A}}^{(k)}$ and $\hat{\boldsymbol{B}}^{(k)}$.

Lemma C.4 (Sporadic Matrix Contraction). *Let Assumptions 4.3 and 4.4 hold. Let the row-stochastic $\hat{\boldsymbol{A}}^{(k)}$ and column-stochastic $\hat{\boldsymbol{B}}^{(k)}$ sporadic mixing matrices (Eq. 4) be designed such that their respective expected values $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ (Eq. 5) are irreducible and primitive. Then, Lemma C.3 implies the following bounds for the vector-induced sporadic matrix norms:*

(i) If the communication probabilities satisfy $\min(\hat{p}_{ij}) \in [\hat{r}_A, 1]$ where $\hat{r}_A \in (0,1)$ is the solution of the equation $f_A(x) = x^{2(m+1)} + \hat{\tau}_A x - \hat{\tau}_A$ with $\hat{\tau}_A = 8m^2(m-1)D(\mathcal{G})K(\mathcal{G})\frac{\max(a^+_{ij})^2}{\min(a^+_{ij})^{2(m+1)}}$, then

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}^{(k)}_{\hat{\boldsymbol{\phi}}}\right\|^2_{\hat{\boldsymbol{\phi}}}\right] \leq (\hat{\rho}_A + \hat{\rho}_{0,A})\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}^{(k)}_{\hat{\boldsymbol{\phi}}}\right\|^2_{\hat{\boldsymbol{\phi}}}\right], \tag{25}$$

where $\hat{\rho}_{0,A} = 4(m-1)\frac{\max(\hat{\phi}_i)}{\min(\hat{\phi}_i)}\max(a^2_{ij}\hat{p}_{ij}(1-\hat{p}_{ij})) \leq \frac{1-\hat{\rho}_A}{2}$.

(ii) If the communication probabilities satisfy $\min(\hat{p}_{ij}) \in [\hat{r}_B, 1]$ where $\hat{r}_B \in (0,1)$ is the solution of the equation $f_B(x) = x^{3m+2} + \hat{\tau}_B x - \hat{\tau}_B$ with $\hat{\tau}_B = 4m^3(m-1)D(\mathcal{G})K(\mathcal{G})\frac{\max(b^+_{ij})^2}{\min(b^+_{ij})^{3m+2}}$, then

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}^{-1}_{\hat{\boldsymbol{\pi}}}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|^2_{\hat{\boldsymbol{\pi}}}\right] \leq (\hat{\rho}_B + \hat{\rho}_{0,B})\mathbb{E}\left[\left\|\boldsymbol{\Lambda}^{-1}_{\hat{\boldsymbol{\pi}}}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|^2_{\hat{\boldsymbol{\pi}}}\right] \\ + m^2\hat{\rho}_{0,B}\mathbb{E}\left[\|\bar{\boldsymbol{y}}^{(k)}\|^2\right],\end{aligned} \tag{26}$$

where $\hat{\rho}_{0,B} = 2(m-1)\frac{\max(\hat{\pi}_i)}{\min(\hat{\pi}_i)}\max(b^2_{ij}\hat{p}_{ij}(1-\hat{p}_{ij})) \leq \frac{1-\hat{\rho}_B}{2}$.

Here, $\boldsymbol{X}^{(k)}, \boldsymbol{Y}^{(k)} \in \mathbb{R}^{m\times n}$ are a collection of m arbitrary n-dimensional vectors.

Proof. See Appendix F.3. □

Lemma C.4 is one of the main contributions of our paper, as we show that the row-stochastic $\hat{\boldsymbol{A}}^{(k)}$ and column-stochastic $\hat{\boldsymbol{B}}^{(k)}$ transition matrices have a contraction property, despite them being sporadic. In other words, $\hat{\boldsymbol{A}}^{(k)}$ and $\hat{\boldsymbol{B}}^{(k)}$ contract in expectation despite not corresponding to a strongly-connected digraph. Moreover, we observe that when clients participate in peer-to-peer communications in every training iteration, i.e., $\hat{p}_{ij} = 1$ for all $(i, j) \in \mathcal{E}$, we will have both $\hat{\rho}_{0,A} = \hat{\rho}_{0,B} = 0$, and Lemma C.4 reduces back to Lemma C.3.

We see in Eqs. 3, and 4, and also in stochastic versions of them in Eqs. 13 and 14, that each client $i \in \mathcal{M}$ updates its gradient tracking parameters $\boldsymbol{y}_i^{(k+1)}$ by subtracting the old local gradient from it and adding the new local gradient, i.e., $\nabla F_i(\boldsymbol{x}_i^{(k+1)})v_i^{(k+1)} - \nabla F_i(\boldsymbol{x}_i^{(k)})v_i^{(k)}$, or $\boldsymbol{g}_i^{(k+1)}v_i^{(k+1)} - \boldsymbol{g}_i^{(k)}v_i^{(k)}$ for the stochastic version. Therefore, we need to characterize how much this update affects the gradient tracking consensus.

Lemma C.5 (Update Effect on Gradient Tracking Consensus)**.** *Let $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\pi}}$ be positive stochastic vectors. Then, the following bounds hold:*

(i) Let Assumptions 4.3 and 4.4 hold. Then, Lemma C.4-(i) can be used to obtain an upper bound on the sporadic consensus variance as

$$\mathbb{E}\left[\left\|\left(\hat{\boldsymbol{A}}^{(k)} - \hat{\boldsymbol{A}}\right)\boldsymbol{X}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] \leq \hat{\rho}_{0,A}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right]. \tag{27}$$

(ii) Let Assumptions 4.3 and 4.4 hold. Then, Lemmas C.2-(iii) and C.5-(i)can be used to obtain an upper bound on the one-step average model update as

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2\right] \leq 2\hat{\rho}_{0,A}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] \\ + 2\kappa_3 \max(\eta_i^{(k)})^2\mathbb{E}\left[\left\|\boldsymbol{Y}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2\right].\end{aligned} \tag{28}$$

(iii) Let Assumptions 4.1-4.3 hold. Then, Lemma C.1-(i) can be used to obtain an upper bound on the sporadic stochastic gradient diversity can be obtained

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)} - \boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\right\|^2\right] \leq 3m\overline{\sigma_0^2 p} + 16m\overline{p(1-p)\delta_0^2} \\ &+ \kappa_4\left(\mathbb{E}\left[\left\|\boldsymbol{X}^{(k+1)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] + \mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right]\right) \\ &+ 4m\left(3\overline{\sigma_1^2 p} + 4\overline{p(1-p)\delta_1^2}\right)\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right] + 8m\kappa_5\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2\right],\end{aligned} \tag{29}$$

where $\kappa_4 = \max\left(\frac{p_i}{\hat{\phi}_i}\left(3\sigma_{1,i}^2\bar{L}^2 + 4L_i^2\right)\right)$ and $\kappa_5 = \overline{\sigma_1^2 p}\bar{L}^2 + \overline{pL^2}$.

Proof. See Appendix F.4. □

We observe that when clients participate in peer-to-peer communications in every training iteration, i.e., $\hat{p}_{ij} = 1$ for all $(i, j) \in \mathcal{E}$ resulting in $\hat{\rho}_{0,A} = 0$, the upper bound of Lemma C.5-(i) becomes zero, and the upper bound of Lemma C.5-(ii) only involves the term $2\kappa_3 \max(\eta_i^{(k)})^2\mathbb{E}[\|\boldsymbol{Y}^{(k)}\|_{\hat{\boldsymbol{\pi}}^{-1}}^2]$. On the other hand, if the gradient estimation variance is uniformly bounded, i.e., $\sigma_{1,i}^2 = 0$, the bound of Lemma C.5-(iii) will no longer depend on $\mathbb{E}[\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\|^2]$.

D Main Lemmas

Our first main lemma is a consensus upper bound of model parameters $\boldsymbol{x}_i^{(k)}$ for each client compared to the weighted average of all model parameters in the directed network graph $\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}$.

Lemma D.1 (Model Consensus)**.** *Let Assumptions 4.1-4.4 hold. Then, Lemmas C.2, C.4-(i) and C.5-(i) imply that*

$$\mathbb{E}\left[\left\|\boldsymbol{X}^{(k+1)}-\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}\right\|_{\hat{\boldsymbol{\phi}}}^{2}\right]\leq\psi_{11}^{(k)}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)}-\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^{2}\right]+\psi_{12}^{(k)}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^{2}\right]+\gamma_{1}^{(k)}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^{2}\right]+\omega_{1}^{(k)},\tag{30}$$

where $\psi_{11}^{(k)}=\frac{1+\tilde{\rho}_A}{2}+m\kappa_1\kappa_6\kappa_3\max(\eta_i^{(k)})^2$, $\psi_{12}^{(k)}=\kappa_6\kappa_3\max(\eta_i^{(k)})^2$, $\gamma_1^{(k)}=m^2\kappa_2\kappa_6\kappa_3\max(\eta_i^{(k)})^2$ *and* $\omega_1^{(k)}=m^2\kappa_6\kappa_3\max(\eta_i^{(k)})^2(\frac{1}{m}\overline{\sigma_0^2p}+3\overline{(1-p)\delta_0^2})$, *in which* $\kappa_6=\frac{2}{1-\tilde{\rho}_A}$ *and* $\tilde{\rho}_A=\hat{\rho}_A+\hat{\rho}_{0,A}$.

Proof. See Appendix G.1. □

We observe that when clients conduct GT in every training iteration, i.e., $p_i=1$ for all $i\in\mathcal{M}$, and use a full batch, i.e., $B_i=D_i$, we will have $\omega_1^{(k)}=0$.

Our next main lemma is a consensus upper bound of scaled GT parameters $\frac{1}{\hat{\pi}_i}\boldsymbol{y}_i^{(k)}$ for each client compared to the sum of all GT parameters in the directed network graph $m\bar{\boldsymbol{y}}^{(k)}$.

Lemma D.2 (Gradient Tracking Consensus)**.** *Let Assumptions 4.1-4.4 hold. Then, Lemmas C.2, C.4-(ii), C.5 and D.1 imply that*

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k+1)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k+1)}\right\|_{\hat{\boldsymbol{\pi}}}^{2}\right]&\leq\psi_{21}^{(k)}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)}-\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^{2}\right]\\&+\psi_{22}^{(k)}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^{2}\right]+\gamma_{2}^{(k)}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^{2}\right]+\omega_{2}^{(k)},\end{aligned}\tag{31}$$

where $\psi_{21}^{(k)}=\kappa_7\kappa_4(1+\psi_{11}^{(k)})+16m\kappa_7\kappa_5\hat{\rho}_{0,A}+\kappa_8\kappa_1\hat{\rho}_{0,B}+16m^2\kappa_7\kappa_5\kappa_3\kappa_1\max(\eta_i^{(k)})^2$, $\psi_{22}^{(k)}=\frac{1+\tilde{\rho}_B}{2}+\kappa_7\kappa_4\psi_{12}^{(k)}+16m\kappa_7\kappa_5\kappa_3\max(\eta_i^{(k)})^2$, $\gamma_2^{(k)}=\kappa_7\kappa_4\gamma_1^{(k)}+2m\kappa_7(3\overline{\sigma_1^2p}+4\overline{p(1-p)\delta_1^2})+m\kappa_8\kappa_2\hat{\rho}_{0,B}+16m^3\kappa_7\kappa_5\kappa_3\kappa_2\max(\eta_i^{(k)})^2$ *and* $\omega_2^{(k)}=\kappa_7\kappa_4\omega_1^{(k)}+2m\kappa_7(\overline{\sigma_0^2p}+4\overline{p(1-p)\delta_0^2})+m\kappa_8\hat{\rho}_{0,B}(\overline{\sigma_0^2p}+3\overline{(1-p)\delta_0^2})+8m^3(\frac{1}{m}\overline{\sigma_0^2p}+6\overline{(1-p)\delta_0^2})\kappa_7\kappa_5\kappa_3\max(\eta_i^{(k)})^2$, *in which* $\kappa_7=2(m+\min(\hat{\pi}_i)^{-1})\frac{1+\tilde{\rho}_B}{1-\tilde{\rho}_B}$, $\kappa_8=\frac{m}{2}\frac{1+\tilde{\rho}_B}{\tilde{\rho}_B}$ *and* $\tilde{\rho}_B=\hat{\rho}_B+\hat{\rho}_{0,B}$.

Proof. See Appendix G.2. □

We observe that if clients perform GT in every training iteration, i.e., $p_i=1$ for all $i\in\mathcal{M}$, and the gradient estimation variance is uniformly bounded, i.e., $\sigma_{1,i}^2=0$, then $\gamma_2^{(k)}\propto\max(\eta_i^{(k)})^2$. On the other hand, when clients perform GT in every training iteration, i.e., $p_i=1$ for all $i\in\mathcal{M}$, and clients use a full batch, i.e., $B_i=D_i$, we will have $\omega_2^{(k)}=0$.

Recursively expanding Eq. 9, we obtain the following explicit inequality for the error vector $\boldsymbol{\varsigma}^{(k)}$

$$\boldsymbol{\varsigma}^{(k+1)}\leq\boldsymbol{\Psi}^{(k:0)}\boldsymbol{\varsigma}^{(0)}+\sum_{r=0}^{k}\boldsymbol{\Psi}^{(k:r+1)}\boldsymbol{\gamma}^{(r)}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\right\|^{2}\right]+\sum_{r=0}^{k}\boldsymbol{\Psi}^{(k:r+1)}\boldsymbol{\omega}^{(r)},\tag{32}$$

in which $\boldsymbol{\Psi}^{(k:r)}=\boldsymbol{\Psi}^{(k)}\boldsymbol{\Psi}^{(k-1)}...\boldsymbol{\Psi}^{(r)}$ for $k>r$, $\boldsymbol{\Psi}^{(k:k)}=\boldsymbol{\Psi}^{(k)}$ for $k=r$ and $\boldsymbol{\Psi}^{(k:r)}=\boldsymbol{I}$ for $k<r$.

Finally, we provide an upper bound on the global loss value at each training iteration.

Lemma D.3 (Loss Descent)**.** *Let Assumptions 4.1-4.4 hold. Then, Lemmas C.1, C.2-(i), C.4-(ii) and C.5-(ii) imply that*

$$\begin{aligned}\mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\right]&\leq\mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right]-\gamma_0^{(k)}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^{2}\right]+\psi_{01}^{(k)}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)}-\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^{2}\right]\\&+\psi_{02}^{(k)}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^{2}\right]+\omega_0^{(k)},\end{aligned}\tag{33}$$

in which we have $\gamma_0^{(k)} = \frac{1}{2}m(m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}} - 3m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}(1-p)\delta_1^2} - (3\max(\hat{\phi}_i\eta_i^{(k)})\hat{\rho}_{0,B} + 2m\bar{L}\kappa_3\max(\eta_i^{(k)})^2)\kappa_2)$, $\psi_{01}^{(k)} = (\frac{3}{2}m\overline{\hat{\pi}\eta^{(k)}}\max(p_iL_i^2) + \bar{L}\hat{\rho}_{0,A} + (\frac{3}{2}\max(\hat{\phi}_i\eta_i^{(k)})\hat{\rho}_{0,B} + m\bar{L}\kappa_3\max(\eta_i^{(k)})^2)\kappa_1)$, $\psi_{02}^{(k)} = \frac{1}{m}(\frac{3}{2}\max(\hat{\phi}_i\eta_i^{(k)})\tilde{\rho}_B + m\bar{L}\kappa_3\max(\eta_i^{(k)})^2)$ *and* $\omega_0^{(k)} = m(\frac{3}{2}m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}(1-p)\delta_0^2} + (\frac{3}{2}\max(\hat{\phi}_i\eta_i^{(k)})\hat{\rho}_{0,B} + m\bar{L}\kappa_3\max(\eta_i^{(k)})^2)(\frac{1}{m}\overline{\sigma_0^2p} + 3\overline{(1-p)\delta_0^2}))$.

Proof. See Appendix G.3. □

We observe that when clients conduct GT in every training iteration, i.e., $p_i = 1$ for all $i \in \mathcal{M}$, and they participate in peer-to-peer communications in every training iteration, i.e., $\hat{p}_{ij} = 1$ for all $(i,j) \in \mathcal{E}$ resulting in $\hat{\rho}_{0,B} = 0$, and also the gradient estimation variance is uniformly bounded, i.e., $\sigma_{1,i}^2 = 0$, then $\gamma_0^{(k)} = \frac{1}{2}m^2\overline{\hat{\phi}\hat{\pi}\eta^{(k)}}$. Furthermore, if the underlying graph admits doubly stochastic matrices $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ to get $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\pi}} = \frac{1}{m}\mathbf{1}$, this simplifies to $\gamma_0^{(k)} = \frac{1}{2}\overline{\eta^{(k)}}$. On the other hand, if clients conduct GT in every training iteration, i.e., $p_i = 1$ for all $i \in \mathcal{M}$, and they use a full batch, i.e., $B_i = D_i$, we will have $\omega_0^{(k)} = 0$.

Applying Definition 4.6 further to Lemma D.3, we get

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\right] \leq \mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right] - \gamma_0^{(k)}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right] + (\boldsymbol{\psi}_0^{(k)})^T\boldsymbol{\varsigma}^{(k)} + \omega_0^{(k)}, \tag{34}$$

and recursively expanding Eq. 34, we obtain the following explicit inequality for the expected loss

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\right] \leq F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) - \sum_{r=0}^{k}\gamma_0^{(r)}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\right\|^2\right] + \sum_{r=0}^{k}(\boldsymbol{\psi}_0^{(r)})^T\boldsymbol{\varsigma}^{(r)} + \sum_{r=0}^{k}\omega_0^{(r)}. \tag{35}$$

Finally, substituting Eq. 32 into Eq. 35 gives us

$$\begin{aligned}\boxed{\begin{aligned}&\mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\right] \leq F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) - \gamma_0^{(k)}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right] \\ &- \sum_{r=0}^{k-1}\left(\gamma_0^{(r)} - \left(\sum_{s=r+1}^{k}(\boldsymbol{\psi}_0^{(s)})^T\boldsymbol{\Psi}^{(s-2:r)}\right)\boldsymbol{\gamma}^{(r)}\right)\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\right\|^2\right] \\ &+ \left(\sum_{r=0}^{k}(\boldsymbol{\psi}_0^{(r)})^T\boldsymbol{\Psi}^{(r-1:0)}\right)\boldsymbol{\varsigma}^{(0)} + \sum_{r=0}^{k-1}\left(\sum_{s=r+1}^{k}(\boldsymbol{\psi}_0^{(s)})^T\boldsymbol{\Psi}^{(s-2:r)}\right)\boldsymbol{\omega}^{(r)} + \sum_{r=0}^{k}\omega_0^{(r)}.\end{aligned}}\end{aligned} \tag{36}$$

Eq. 36 is a key equation that is used in both Propositions in Sec. 4.3.

E COROLLARIES TO PROPOSITION 4.7

We proceed with two corollaries to Proposition 4.7 when further structure is placed on the learning rates.

Corollary E.1 (Non-Increasing Learning Rates)**.** *Let Assumptions 4.1-4.4 hold. Let each client $i \in \mathcal{M}$ use a non-increasing learning rate $\eta_i^{(k+1)} \leq \eta_i^{(k)}$. Then, it is sufficient for the constraint of Proposition 4.7 to hold only for the initial values $\eta_i^{(0)}$ to ensure $\rho(\boldsymbol{\Psi}^{(k)}) < 1$ for all $k \geq 0$.*

Corollary E.2 (Constant Learning Rates)**.** *Let Assumptions 4.1-4.4 hold. Let each client $i \in \mathcal{M}$ use a constant learning rate, i.e., $\eta_i^{(k)} = \eta_i^{(0)}$. Then, the values defined in Lemmas D.1-D.3 will all become constants, i.e., $\boldsymbol{\Psi}^{(k)} = \boldsymbol{\Psi}^{(0)}$, $\boldsymbol{\gamma}^{(k)} = \boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\omega}^{(k)} = \boldsymbol{\omega}^{(0)}$, $\boldsymbol{\psi}_0^{(k)} = \boldsymbol{\psi}_0^{(0)}$, $\gamma_0^{(k)} = \gamma_0^{(0)}$ and $\omega_0^{(k)} = \omega_0^{(0)}$, for all $k \geq 0$. Furthermore, if constant learning rates satisfy the conditions outlined in Corollary E.1, then there exists a scalar $0 < \rho_\Psi < 1$ such that $\rho(\boldsymbol{\Psi}^{(k)}) = \rho_\Psi$ for all $k \geq 0$.*

F PROOFS OF INTERMEDIARY LEMMAS

F.1 PROOF OF LEMMA C.1

(i) We have

$$\begin{aligned}\left\|\bar{\boldsymbol{g}}_v^{(k)} - \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}}\right\|^2 &= \left\|\frac{1}{m}\sum_{i=1}^m (\boldsymbol{g}_i^{(k)} - \nabla F_i(\boldsymbol{x}_i^{(k)}))v_i^{(k)}\right\|^2 \\ &= \frac{1}{m^2}\sum_{i=1}^m \left\|\boldsymbol{g}_i^{(k)} - \nabla F_i(\boldsymbol{x}_i^{(k)})\right\|^2 v_i^{(k)} \\ &\quad + \frac{1}{m^2}\sum_{i=1}^m\sum_{\substack{j=1\\ j\neq i}}^m \left\langle \boldsymbol{g}_i^{(k)} - \nabla F_i(\boldsymbol{x}_i^{(k)}), \boldsymbol{g}_j^{(k)} - \nabla F_j(\boldsymbol{x}_j^{(k)})\right\rangle v_i^{(k)} v_j^{(k)}.\end{aligned} \tag{37}$$

Taking the expected value of this expression and using Assumptions A.1 and 4.3, we obtain the following:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{m^2}\sum_{i=1}^m \left\|\boldsymbol{g}_i^{(k)} - \nabla F_i(\boldsymbol{x}_i^{(k)})\right\|^2 v_i^{(k)}\right] &= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}\left[\left\|\boldsymbol{g}_i^{(k)} - \nabla F_i(\boldsymbol{x}_i^{(k)})\right\|^2\right]\mathbb{E}[v_i^{(k)}] \\ &\leq \frac{1}{m^2}\sum_{i=1}^m p_i\left(\sigma_{0,i}^2 + \sigma_{1,i}^2\mathbb{E}\left[\left\|\nabla F(\boldsymbol{x}_i^{(k)})\right\|^2\right]\right) \\ &= \frac{1}{m}\overline{\sigma_0^2 p} + \frac{1}{m^2}\sum_{i=1}^m \sigma_{1,i}^2 p_i \mathbb{E}\left[\left\|\nabla F(\boldsymbol{x}_i^{(k)})\right\|^2\right].\end{aligned} \tag{38}$$

We finally use the Lipschitz continuity of the gradients of the global loss function F which is implied by Assumption 4.1 to write

$$\begin{aligned}\frac{1}{m^2}\sum_{i=1}^m \sigma_{1,i}^2 p_i \left\|\nabla F(\boldsymbol{x}_i^{(k)})\right\|^2 &= \frac{1}{m^2}\sum_{i=1}^m \sigma_{1,i}^2 p_i \left\|\nabla F(\boldsymbol{x}_i^{(k)}) \mp \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 \\ &\leq \frac{2}{m^2}\sum_{i=1}^m \sigma_{1,i}^2 p_i \left(\left\|\nabla F(\boldsymbol{x}_i^{(k)}) - \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 + \left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right) \\ &\leq \frac{2}{m^2}\sum_{i=1}^m \sigma_{1,i}^2 p_i \bar{L}^2 \left\|\boldsymbol{x}_i^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2 + \frac{2}{m}\overline{\sigma_1^2 p}\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 \\ &= \frac{2\bar{L}^2}{m^2}\sum_{i=1}^m \sigma_{1,i}^2 p_i \frac{\hat{\phi}_i}{\hat{\phi}_i}\left\|\boldsymbol{x}_i^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2 + \frac{2}{m}\overline{\sigma_1^2 p}\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 \\ &\leq \frac{2\bar{L}^2}{m^2}\max\left(\frac{\sigma_{1,i}^2 p_i}{\hat{\phi}_i}\right)\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2 + \frac{2}{m}\overline{\sigma_1^2 p}\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2.\end{aligned} \tag{39}$$

(ii) According to Jensen's inequality, we have

$$\begin{aligned}\left\|\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} - \overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}}\right\|^2 &= \left\|\frac{1}{m}\sum_{i=1}^m \left(\nabla F_i(\boldsymbol{x}_i^{(k)}) - \nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right) v_i^{(k)}\right\|^2 \\ &\leq \frac{1}{m}\sum_{i=1}^m \left\|\nabla F_i(\boldsymbol{x}_i^{(k)}) - \nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 v_i^{(k)} \\ &\leq \frac{1}{m}\sum_{i=1}^m L_i^2 \left\|\boldsymbol{x}_i^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2 v_i^{(k)},\end{aligned} \tag{40}$$

where we used Assumption 4.1 in the end. Taking the expected value of this expression and using Assumption 4.3, we get

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} L_i^2 \left\|\boldsymbol{x}_i^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2 v_i^{(k)}\right] = \frac{1}{m}\sum_{i=1}^{m} L_i^2 \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2\right] \mathbb{E}[v_i^{(k)}] \\
&= \frac{1}{m}\sum_{i=1}^{m} L_i^2 p_i \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2\right] = \frac{1}{m}\sum_{i=1}^{m} L_i^2 p_i \frac{\hat{\phi}_i}{\hat{\phi}_i} \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2\right] \\
&\leq \frac{1}{m}\max\left(\frac{L_i^2 p_i}{\hat{\phi}_i}\right)\sum_{i=1}^{m}\hat{\phi}_i \mathbb{E}\left[\left\|\boldsymbol{x}_i^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2\right] = \frac{1}{m}\max\left(\frac{L_i^2 p_i}{\hat{\phi}_i}\right)\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right].
\end{aligned} \tag{41}$$

(iii) According to Jensen's inequality, we have

$$\begin{aligned}
&\left\|\overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} - \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 = \left\|\frac{1}{m}\sum_{i=1}^{m}\left(\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) v_i^{(k)} - \nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right)\right\|^2 \\
&= \left\|\frac{1}{m}\sum_{i=1}^{m}\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})(v_i^{(k)} - 1)\right\|^2 \leq \frac{1}{m}\sum_{i=1}^{m}\left\|\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 (1 - v_i^{(k)}) \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\left(\delta_{0,i}^2 + \delta_{1,i}^2\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right)(1 - v_i^{(k)}).
\end{aligned} \tag{42}$$

Taking the expected value of this expression and using Assumptions 4.2 and 4.3, we obtain the following.

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\delta_{0,i}^2 + \delta_{1,i}^2\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right)(1 - v_i^{(k)})\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\delta_{0,i}^2 + \delta_{1,i}^2\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right]\right)(1 - \mathbb{E}[v_i^{(k)}]) \\
&= \frac{1}{m}\sum_{i=1}^{m}(1 - p_i)\left(\delta_{0,i}^2 + \delta_{1,i}^2\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right]\right) = \overline{(1-p)\delta_0^2} + \overline{(1-p)\delta_1^2}\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2.
\end{aligned} \tag{43}$$

F.2 Proof of Lemma C.2

(i) We derive this upper bound using steps (b)-(e) of Eq. 18. We have

$$\begin{aligned}
&\left\|\bar{\boldsymbol{y}}^{(k)}\right\|^2 = \left\|\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)}\right\|^2 = \left\|\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)} \mp \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 \\
&= \left\|\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)} - \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}}\right\|^2 + \left\|\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 \\
&\quad + 2\left\langle \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)} - \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}}, \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\rangle,
\end{aligned} \tag{44}$$

where we separated the inner product term because it will be equal to 0 when taking its expected value, since $\boldsymbol{g}_i^{(k)}$ is an unbiased estimate of $\nabla F_i(\boldsymbol{x}_i^{(k)})$ according to the Assumption A.1. We further expand the bound as follows.

$$\begin{aligned}
&\left\|\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 \\
&\leq 3\Bigg(\left\|\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} - \overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}}\right\|^2 + \left\|\overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} - \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2 \\
&\qquad + \left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\Bigg),
\end{aligned} \tag{45}$$

in which we used Young's inequality. We add Eq. 45 back to Eq. 44, which results in an upper bound involving the three expressions given in Lemma C.1. Taking the expected value of the result concludes the proof. While the upper bounds in Lemmas C.1-(i) and C.1-(ii) each involve a $\max(\cdot)$ term, their proofs in Appendix F.1 show that we can first combine their coefficients and then take their maximum.

(ii) See Proposition 4.6 in Nedić et al. (2025).

(iii) According to Corollary 1 of Lemma 1(b) in Nguyen et al. (2025), we have $\sum_{i=1}^m \gamma_i \|u_i - \sum_{\ell=1}^m \gamma_\ell u_\ell\|^2 \leq \sum_{i=1}^m \gamma_i \|u_i\|^2$, which holds for any collection of vectors $\{u_i\}_{i=1}^m$ and scalars $\{\gamma_i\}_{i=1}^m$ such that $\sum_{i=1}^m \gamma_i = 1$. Thus, we have

$$
\begin{aligned}
&\left\| \boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)} \hat{\boldsymbol{B}}^{(k)} \boldsymbol{Y}^{(k)} - \mathbf{1} \hat{\boldsymbol{\phi}}^T \boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)} \hat{\boldsymbol{B}}^{(k)} \boldsymbol{Y}^{(k)} \right\|_{\hat{\boldsymbol{\phi}}}^2 \leq \left\| \boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)} \hat{\boldsymbol{B}}^{(k)} \boldsymbol{Y}^{(k)} \right\|_{\hat{\boldsymbol{\phi}}}^2 = \sum_{i=1}^m \hat{\phi}_i (\eta_i^{(k)})^2 \left\| \sum_{j=1}^m \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)} \right\|^2 \\
&= \sum_{i=1}^m \hat{\phi}_i (\eta_i^{(k)})^2 \left\| \left(\sum_{\ell=1}^m \hat{b}_{i\ell}^{(k)} \right) \frac{\sum_{j=1}^m \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)}}{\sum_{\ell=1}^m \hat{b}_{i\ell}^{(k)}} \right\|^2 \leq \sum_{i=1}^m \hat{\phi}_i (\eta_i^{(k)})^2 \left(\sum_{\ell=1}^m \hat{b}_{i\ell}^{(k)} \right)^2 \frac{\sum_{j=1}^m \hat{b}_{ij}^{(k)} \left\| \boldsymbol{y}_j^{(k)} \right\|^2}{\sum_{\ell=1}^m \hat{b}_{i\ell}^{(k)}} \\
&= \sum_{i=1}^m \hat{\phi}_i (\eta_i^{(k)})^2 \left(\sum_{\ell=1}^m \hat{b}_{i\ell}^{(k)} \right) \sum_{j=1}^m \hat{b}_{ij}^{(k)} \left\| \boldsymbol{y}_j^{(k)} \right\|^2 \\
&\leq \sum_{i=1}^m \hat{\phi}_i (\eta_i^{(k)})^2 \Bigg(\Bigg(\sum_{\substack{\ell=1 \\ \ell \neq i}}^m \hat{b}_{ii}^{(k)} \hat{b}_{i\ell}^{(k)} + (\hat{b}_{ii}^{(k)})^2 \Bigg) \left\| \boldsymbol{y}_i^{(k)} \right\|^2 \\
&\qquad + \sum_{\substack{j=1 \\ j \neq i}}^m \Bigg(\sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^m \hat{b}_{ij}^{(k)} \hat{b}_{i\ell}^{(k)} + \hat{b}_{ij}^{(k)} \hat{b}_{ii}^{(k)} + (\hat{b}_{ij}^{(k)})^2 \Bigg) \left\| \boldsymbol{y}_j^{(k)} \right\|^2 \Bigg),
\end{aligned}
\tag{46}
$$

where we used Jensen's inequality noting that $\hat{\boldsymbol{B}}^{(k)}$ is a column-stochastic matrix. Taking the expected value of this expression and employing Assumption 4.3, we get

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)} \hat{\boldsymbol{B}}^{(k)} \boldsymbol{Y}^{(k)} \right\|_{\hat{\boldsymbol{\phi}}}^2 \right] \\
&\leq \sum_{i=1}^m \hat{\phi}_i (\eta_i^{(k)})^2 \Bigg[\Bigg(\sum_{\substack{\ell=1 \\ \ell \neq i}}^m \Bigg(1 - \sum_{\substack{\ell=1 \\ \ell \neq i}} b_{\ell i} \hat{p}_{\ell i} \Bigg) b_{i\ell} \hat{p}_{i\ell} \\
&\qquad + \Bigg(1 - 2 \sum_{\substack{\ell=1 \\ \ell \neq i}}^m b_{\ell i} \hat{p}_{\ell i} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^m \sum_{\substack{\ell'=1 \\ \ell' \neq \ell}}^m b_{\ell i} b_{\ell' i} \hat{p}_{\ell i} \hat{p}_{\ell' i} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^m b_{\ell i}^2 \hat{p}_{\ell i} \Bigg) \Bigg) \mathbb{E}\left[\left\| \boldsymbol{y}_i^{(k)} \right\|^2 \right] \\
&\qquad + \sum_{\substack{j=1 \\ j \neq i}}^m \Bigg(\sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^m b_{ij} b_{i\ell} \hat{p}_{ij} \hat{p}_{i\ell} + b_{ij} \hat{p}_{ij} \Bigg(1 - \sum_{\substack{\ell=1 \\ \ell \neq i}}^m b_{\ell i} \hat{p}_{\ell i} \Bigg) + b_{ij}^2 \hat{p}_{ij} \Bigg) \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \Bigg] \\
&= \sum_{i=1}^m \hat{\phi}_i (\eta_i^{(k)})^2 \left(\sum_{\ell=1}^m \hat{b}_{i\ell} \right) \sum_{j=1}^m \hat{b}_{ij} \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \\
&\qquad + \sum_{i=1}^m \hat{\phi}_i (\eta_i^{(k)})^2 \Bigg[\Bigg(\sum_{\substack{\ell=1 \\ \ell \neq i}}^m b_{\ell i}^2 \hat{p}_{\ell i} (1 - \hat{p}_{\ell i}) \Bigg) \mathbb{E}\left[\left\| \boldsymbol{y}_i^{(k)} \right\|^2 \right] + \sum_{\substack{j=1 \\ j \neq i}}^m b_{ij}^2 \hat{p}_{ij} (1 - \hat{p}_{ij}) \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \Bigg]
\end{aligned}
\tag{47}
$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{m} \hat{\phi}_i (\eta_i^{(k)})^2 \left(\sum_{\ell=1}^{m} \frac{\hat{\pi}_\ell}{\hat{\pi}_\ell} \hat{b}_{i\ell} \right) \sum_{j=1}^{m} \hat{b}_{ij} \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \\
&\quad + 2 \max(\hat{\phi}_i (\eta_i^{(k)})^2) \sum_{j=1}^{m} \left(\sum_{\substack{\ell=1 \\ \ell \neq j}}^{m} b_{\ell j}^2 \hat{p}_{\ell j} (1 - \hat{p}_{\ell j}) \right) \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \\
&\leq \frac{1}{\min(\hat{\pi}_i)} \sum_{i=1}^{m} \hat{\phi}_i (\eta_i^{(k)})^2 \hat{\pi}_i \sum_{j=1}^{m} \hat{b}_{ij} \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \\
&\qquad\qquad + 2m \max(\hat{\phi}_i (\eta_i^{(k)})^2) \sum_{j=1}^{m} \frac{\hat{\pi}_j}{\hat{\pi}_j} \overline{b_{:j}^2 \hat{p}_{:j} (1 - \hat{p}_{:j})} \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \\
&\leq \frac{\max(\hat{\phi}_i (\eta_i^{(k)})^2 \hat{\pi}_i)}{\min(\hat{\pi}_i)} \sum_{j=1}^{m} \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \sum_{i=1}^{m} \hat{b}_{ij} \\
&\qquad\qquad + 2m \max(\hat{\phi}_i (\eta_i^{(k)})^2) \max(\hat{\pi}_j \overline{b_{:j}^2 \hat{p}_{:j} (1 - \hat{p}_{:j})}) \sum_{j=1}^{m} \frac{1}{\hat{\pi}_j} \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \\
&\leq \frac{\max(\hat{\phi}_i (\eta_i^{(k)})^2 \hat{\pi}_i)}{\min(\hat{\pi}_i)} \sum_{j=1}^{m} \mathbb{E}\left[\left\| \boldsymbol{y}_j^{(k)} \right\|^2 \right] \\
&\qquad\qquad + 2m \max(\hat{\phi}_i (\eta_i^{(k)})^2) \max(\hat{\pi}_j \overline{b_{:j}^2 \hat{p}_{:j} (1 - \hat{p}_{:j})}) \mathbb{E}\left[\left\| \boldsymbol{Y}^{(k)} \right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2 \right] \\
&\leq \max(\hat{\phi}_i (\eta_i^{(k)})^2) \left(\frac{\max(\hat{\pi}_i)^2}{\min(\hat{\pi}_i)} + 2m \max(\hat{\pi}_j \overline{b_{:j}^2 \hat{p}_{:j} (1 - \hat{p}_{:j})}) \right) \mathbb{E}\left[\left\| \boldsymbol{Y}^{(k)} \right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2 \right],
\end{aligned}
$$

where we used the fact that $\hat{\boldsymbol{B}} \hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}$ and $\mathbf{1}^T \hat{\boldsymbol{B}} = \mathbf{1}^T$ to simplify the first term.

F.3 Proof of Lemma C.4

(i) We start by expanding the norm as

$$
\begin{aligned}
&\left\| \hat{\boldsymbol{A}}^{(k)} \boldsymbol{X}^{(k)} - \mathbf{1} \hat{\boldsymbol{\phi}}^T \boldsymbol{X}^{(k)} \right\|_{\hat{\boldsymbol{\phi}}}^2 = \sum_{i=1}^{m} \hat{\phi}_i \left\| (\hat{\boldsymbol{a}}_i^{(k)})^T \boldsymbol{X}^{(k)} - \hat{\boldsymbol{\phi}}^T \boldsymbol{X}^{(k)} \right\|^2 \\
&= \sum_{i=1}^{m} \left\| \sqrt{\hat{\phi}_i} \left((\hat{\boldsymbol{a}}_i^{(k)})^T \boldsymbol{X}^{(k)} - \hat{\boldsymbol{\phi}}^T \boldsymbol{X}^{(k)} \right) \right\|^2 = \left\| \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}}^{1/2} \left(\hat{\boldsymbol{A}}^{(k)} \boldsymbol{X}^{(k)} - \mathbf{1} \hat{\boldsymbol{\phi}}^T \boldsymbol{X}^{(k)} \right) \right\|^2 \\
&= \left\| \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}}^{1/2} \left(\hat{\boldsymbol{A}}^{(k)} \boldsymbol{X}^{(k)} - \hat{\boldsymbol{A}}^{(k)} \mathbf{1} \hat{\boldsymbol{\phi}}^T \boldsymbol{X}^{(k)} \right) \right\|^2 = \left\| \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}}^{1/2} \hat{\boldsymbol{A}}^{(k)} \left(\boldsymbol{I} - \mathbf{1} \hat{\boldsymbol{\phi}}^T \right) \boldsymbol{X}^{(k)} \right\|^2 \\
&= \sum_{d=1}^{n} \left\| \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}}^{1/2} \hat{\boldsymbol{A}}^{(k)} \left(\boldsymbol{I} - \mathbf{1} \hat{\boldsymbol{\phi}}^T \right) [\boldsymbol{X}^{(k)}]_{[:,d]} \right\|^2 \\
&= \sum_{d=1}^{n} [\boldsymbol{X}^{(k)}]_{[:,d]}^T \left(\boldsymbol{I} - \hat{\boldsymbol{\phi}} \mathbf{1}^T \right) (\hat{\boldsymbol{A}}^{(k)})^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}} \hat{\boldsymbol{A}}^{(k)} \left(\boldsymbol{I} - \mathbf{1} \hat{\boldsymbol{\phi}}^T \right) [\boldsymbol{X}^{(k)}]_{[:,d]}.
\end{aligned} \tag{48}
$$

Noting that when taking the expected value of the above expression we can push the expectation operation inside to $\mathbb{E}[(\hat{\boldsymbol{A}}^{(k)})^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}} \hat{\boldsymbol{A}}^{(k)}]$, we first determine the values of this product. We have

$$
\begin{aligned}
(\hat{\boldsymbol{A}}^{(k)})^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}} \hat{\boldsymbol{A}}^{(k)} &= \left[\sum_{\ell=1}^{m} \hat{\phi}_\ell \hat{a}_{\ell i}^{(k)} \hat{a}_{\ell j}^{(k)} \right]_{1 \le i,j \le m} \\
&= \begin{cases} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \hat{\phi}_\ell a_{\ell i}^2 \hat{v}_{\ell i}^{(k)} + \hat{\phi}_i \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} a_{i\ell} \hat{v}_{i\ell}^{(k)} \right)^2 & i = j, \\ \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{m} \hat{\phi}_\ell a_{\ell i} a_{\ell j} \hat{v}_{\ell i}^{(k)} \hat{v}_{\ell j}^{(k)} + \hat{\phi}_i \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} a_{i\ell} \hat{v}_{i\ell}^{(k)} \right) a_{ij} \hat{v}_{ij}^{(k)} + \hat{\phi}_j a_{ji} \hat{v}_{ji}^{(k)} \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq j}}^{m} a_{j\ell} \hat{v}_{j\ell}^{(k)} \right) & i \neq j \end{cases} \\
&= \begin{cases} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \hat{\phi}_\ell a_{\ell i}^2 \hat{v}_{\ell i}^{(k)} + \hat{\phi}_i \left(1 - 2 \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} a_{i\ell} \hat{v}_{i\ell}^{(k)} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \sum_{\substack{\ell'=1 \\ \ell' \neq i \\ \ell' \neq \ell}}^{m} a_{i\ell} a_{i\ell'} \hat{v}_{i\ell}^{(k)} \hat{v}_{i\ell'}^{(k)} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} a_{i\ell}^2 \hat{v}_{i\ell}^{(k)} \right) & i = j, \\ \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{m} \hat{\phi}_\ell a_{\ell i} a_{\ell j} \hat{v}_{\ell i}^{(k)} \hat{v}_{\ell j}^{(k)} + \hat{\phi}_i \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{m} a_{i\ell} \hat{v}_{i\ell}^{(k)} \right) a_{ij} \hat{v}_{ij}^{(k)} - \hat{\phi}_i a_{ij}^2 \hat{v}_{ij}^{(k)} & \\ \qquad + \hat{\phi}_j a_{ji} \hat{v}_{ji}^{(k)} \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq j \\ \ell \neq i}}^{m} a_{j\ell} \hat{v}_{j\ell}^{(k)} \right) - \hat{\phi}_j a_{ji}^2 \hat{v}_{ji}^{(k)} & i \neq j, \end{cases}
\end{aligned} \tag{49}
$$

where we considered the terms that are statistically dependent on each other together. Now, taking the expected value of this product and using Assumption 4.3, we get

$$
\begin{aligned}
&\mathbb{E}\left[(\hat{\boldsymbol{A}}^{(k)})^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}} \hat{\boldsymbol{A}}^{(k)} \right] \\
&= \begin{cases} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \hat{\phi}_\ell a_{\ell i}^2 \hat{p}_{\ell i} + \hat{\phi}_i \left(1 - 2 \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} a_{i\ell} \hat{p}_{i\ell} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \sum_{\substack{\ell'=1 \\ \ell' \neq i \\ \ell' \neq \ell}}^{m} a_{i\ell} a_{i\ell'} \hat{p}_{i\ell} \hat{p}_{i\ell'} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} a_{i\ell}^2 \hat{p}_{i\ell} \right) & i = j, \\ \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{m} \hat{\phi}_\ell a_{\ell i} a_{\ell j} \hat{p}_{\ell i} \hat{p}_{\ell j} + \hat{\phi}_i \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{m} a_{i\ell} \hat{p}_{i\ell} \right) a_{ij} \hat{p}_{ij} - \hat{\phi}_i a_{ij}^2 \hat{p}_{ij} & \\ \qquad + \hat{\phi}_j a_{ji} \hat{p}_{ji} \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq j \\ \ell \neq i}}^{m} a_{j\ell} \hat{p}_{j\ell} \right) - \hat{\phi}_j a_{ji}^2 \hat{p}_{ji} & i \neq j \end{cases} \\
&= \hat{\boldsymbol{A}}^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}} \hat{\boldsymbol{A}} + \begin{cases} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \hat{\phi}_\ell a_{\ell i}^2 \hat{p}_{\ell i} (1 - \hat{p}_{\ell i}) + \hat{\phi}_i \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} a_{i\ell}^2 \hat{p}_{i\ell} (1 - \hat{p}_{i\ell}) & i = j, \\ -\hat{\phi}_i a_{ij}^2 \hat{p}_{ij} (1 - \hat{p}_{ij}) - \hat{\phi}_j a_{ji}^2 \hat{p}_{ji} (1 - \hat{p}_{ji}) & i \neq j \end{cases} \\
&= \hat{\boldsymbol{A}}^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}} \hat{\boldsymbol{A}} + \begin{cases} (m-1) \hat{\phi}_i \left(\overline{\hat{\phi} \hat{\phi}_i^{-1} a_{:i}^2 \hat{p}_{:i} (1 - \hat{p}_{:i})} + \overline{a_{i:}^2 \hat{p}_{i:} (1 - \hat{p}_{i:})} \right) & i = j, \\ -\sqrt{\hat{\phi}_i \hat{\phi}_j} \left(\sqrt{\frac{\hat{\phi}_i}{\hat{\phi}_j}} a_{ij}^2 \hat{p}_{ij} (1 - \hat{p}_{ij}) + \sqrt{\frac{\hat{\phi}_j}{\hat{\phi}_i}} a_{ji}^2 \hat{p}_{ji} (1 - \hat{p}_{ji}) \right) & i \neq j \end{cases} \\
&= \hat{\boldsymbol{A}}^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}} \hat{\boldsymbol{A}} + \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}}^{1/2} \hat{\boldsymbol{A}}_0 \boldsymbol{\Lambda}_{\hat{\boldsymbol{\phi}}}^{1/2}.
\end{aligned} \tag{50}
$$

The matrix $\hat{\boldsymbol{A}}_0$ will be an all-zeroes matrix if $\hat{p}_{ij} = 1$ for all links $(i,j) \in \mathcal{E}$. Plugging the result in Eq. 50 back into Eq. 48, we get

$$
\mathbb{E}\left[\left\| \hat{\boldsymbol{A}}^{(k)} \boldsymbol{X}^{(k)} - \mathbf{1} \hat{\boldsymbol{\phi}}^T \boldsymbol{X}^{(k)} \right\|_{\hat{\boldsymbol{\phi}}}^2 \right] \le \left(\hat{\rho}_A + \rho(\hat{\boldsymbol{A}}_0) \right) \mathbb{E}\left[\left\| \boldsymbol{X}^{(k)} - \mathbf{1} \hat{\boldsymbol{\phi}}^T \boldsymbol{X}^{(k)} \right\|_{\hat{\boldsymbol{\phi}}}^2 \right], \tag{51}
$$

in which $\rho(\hat{\boldsymbol{A}}_0) \leq (m-1)\max((\hat{\phi}\hat{\phi}_i^{-1} + \sqrt{\hat{\phi}\hat{\phi}_i^{-1}})a_{:i}^2\hat{p}_{:i}(1-\hat{p}_{:i}) + (1+\sqrt{\hat{\phi}_i\hat{\phi}^{-1}})a_{i:}^2\hat{p}_{i:}(1-\hat{p}_{i:}))$. To ensure a contractive property for the upper bound in Eq. 51, we need to ensure that $\rho(\hat{\boldsymbol{A}}_0) < \frac{1-\hat{\rho}_A}{2}$. Noting that $\hat{\rho}_A = 1 - \frac{\min(\hat{\phi}_i)\min(a_{ij}^+\hat{p}_{ij})^2}{\max(\hat{\phi}_i)^2 D(\mathcal{G})K(\mathcal{G})}$ from Lemma C.3-(i), and that we have the bounds $\frac{\min(a_{ij}^+\hat{p}_{ij})^m}{m} \leq \min(\hat{\phi}_i) \leq \frac{1}{m} \leq \max(\hat{\phi}_i) < 1$ for the eigenvector entries (see Nguyen et al. (2025)), we need to solve for

$$\begin{aligned} \max(\hat{p}_{ij}(1-\hat{p}_{ij})) &\leq \frac{\frac{\min(\hat{\phi}_i)\min(a_{ij}^+\hat{p}_{ij})^2}{\max(\hat{\phi}_i)^2 D(\mathcal{G})K(\mathcal{G})}}{8(m-1)\max(a_{ij}^+)^2}\frac{\min(\hat{\phi}_i)}{\max(\hat{\phi}_i)} \\ \Rightarrow \min(\hat{p}_{ij})^{2(m+1)} &\geq \frac{8m^2(m-1)\max(a_{ij}^+)^2 D(\mathcal{G})K(\mathcal{G})}{\min(a_{ij}^+)^{2(m+1)}}(1-\min(\hat{p}_{ij})). \end{aligned} \tag{52}$$

Solving for a range of unknown values $x = \hat{p}_{ij}$ and $\hat{\tau}_A = \frac{8m^2(m-1)\max(a_{ij}^+)^2 D(\mathcal{G})K(\mathcal{G})}{\min(a_{ij}^+)^{2(m+1)}}$, we write the equation $f_A(x) = x^{2(m+1)} + \hat{\tau}_A x - \hat{\tau}_A$. Noting that $f_A(0) = -\hat{\tau}_A < 0$ and $f_A(1) = 1 > 0$, there exists a root $\hat{r}_A \in (0,1)$ such that $f_A(\hat{r}_A) = 0$. Since $f_A(x)$ is an increasing function in x for $x > 0$, which is the interval we are interested in analyzing, the solution to $f_A(x) \geq 0$ will be $x \in [\hat{r}_A, 1]$.

(ii) We start by expanding the norm as

$$\begin{aligned} &\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^2 = \left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{1/2}\left(\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - \mathbf{1}\mathbf{1}^T\boldsymbol{Y}^{(k)}\right)\right\|^2 \\ &= \left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1/2}\left(\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - \hat{\boldsymbol{\pi}}\mathbf{1}^T\boldsymbol{Y}^{(k)}\right)\right\|^2 = \left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1/2}\left(\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - \hat{\boldsymbol{\pi}}\mathbf{1}^T\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}\right)\right\|^2 \\ &= \left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1/2}\left(\boldsymbol{I} - \hat{\boldsymbol{\pi}}\mathbf{1}^T\right)\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}\right\|^2 = \sum_{d=1}^{n}\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1/2}\left(\boldsymbol{I} - \hat{\boldsymbol{\pi}}\mathbf{1}^T\right)\hat{\boldsymbol{B}}^{(k)}[\boldsymbol{Y}^{(k)}]_{[:,d]}\right\|^2 \\ &= \sum_{d=1}^{n}[\boldsymbol{Y}^{(k)}]_{[:,d]}^T(\hat{\boldsymbol{B}}^{(k)})^T\left(\boldsymbol{I} - \mathbf{1}\hat{\boldsymbol{\pi}}^T\right)\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\left(\boldsymbol{I} - \hat{\boldsymbol{\pi}}\mathbf{1}^T\right)\hat{\boldsymbol{B}}^{(k)}[\boldsymbol{Y}^{(k)}]_{[:,d]} \\ &= \sum_{d=1}^{n}[\boldsymbol{Y}^{(k)}]_{[:,d]}^T(\hat{\boldsymbol{B}}^{(k)})^T\left(\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} - \mathbf{1}\mathbf{1}^T\right)\hat{\boldsymbol{B}}^{(k)}[\boldsymbol{Y}^{(k)}]_{[:,d]} \\ &= \sum_{d=1}^{n}[\boldsymbol{Y}^{(k)}]_{[:,d]}^T\left((\hat{\boldsymbol{B}}^{(k)})^T\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\hat{\boldsymbol{B}}^{(k)} - \mathbf{1}\mathbf{1}^T\right)[\boldsymbol{Y}^{(k)}]_{[:,d]}. \end{aligned} \tag{53}$$

Noting that when taking the expected value of the above expression we can push the expectation operation inside to $\mathbb{E}[(\hat{\boldsymbol{B}}^{(k)})^T\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\hat{\boldsymbol{B}}^{(k)} - \mathbf{1}\mathbf{1}^T]$, we first determine the values of this product. We

have

$$
\begin{aligned}
(\hat{\boldsymbol{B}}^{(k)})^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} \hat{\boldsymbol{B}}^{(k)} &= \left[\sum_{\ell=1}^{m} \hat{\pi}_\ell^{-1} \hat{b}_{\ell i}^{(k)} \hat{b}_{\ell j}^{(k)} \right]_{1 \le i,j \le m} \\
&= \begin{cases}
\sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \hat{\pi}_\ell^{-1} b_{\ell i}^2 \hat{v}_{\ell i}^{(k)} + \hat{\pi}_i^{-1} \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} b_{\ell i} \hat{v}_{\ell i}^{(k)} \right)^2 & i = j, \\
\sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{m} \hat{\pi}_\ell^{-1} b_{\ell i} b_{\ell j} \hat{v}_{\ell i}^{(k)} \hat{v}_{\ell j}^{(k)} + \hat{\pi}_i^{-1} \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} b_{\ell i} \hat{v}_{\ell i}^{(k)} \right) b_{ij} \hat{v}_{ij}^{(k)} & \\
\qquad + \hat{\pi}_j^{-1} b_{ji} \hat{v}_{ji}^{(k)} \left(1 - \sum_{\substack{\ell'=1 \\ \ell' \neq j}}^{m} b_{\ell' j} \hat{v}_{\ell' j}^{(k)} \right) & i \neq j
\end{cases} \\
&= \begin{cases}
\sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \hat{\pi}_\ell^{-1} b_{\ell i}^2 \hat{v}_{\ell i}^{(k)} + \hat{\pi}_i^{-1} \left(1 - 2 \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} b_{\ell i} \hat{v}_{\ell i}^{(k)} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \sum_{\substack{\ell'=1 \\ \ell' \neq i \\ \ell' \neq \ell}}^{m} b_{\ell i} b_{\ell' i} \hat{v}_{\ell i}^{(k)} \hat{v}_{\ell' i}^{(k)} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} b_{\ell i}^2 \hat{v}_{\ell i}^{(k)} \right) & i = j, \\
\sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{m} \hat{\pi}_\ell^{-1} b_{\ell i} b_{\ell j} \hat{v}_{\ell i}^{(k)} \hat{v}_{\ell j}^{(k)} + \hat{\pi}_i^{-1} \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} b_{\ell i} \hat{v}_{\ell i}^{(k)} \right) b_{ij} \hat{v}_{ij}^{(k)} & \\
\qquad + \hat{\pi}_j^{-1} b_{ji} \hat{v}_{ji}^{(k)} \left(1 - \sum_{\substack{\ell'=1 \\ \ell' \neq j}}^{m} b_{\ell' j} \hat{v}_{\ell' j}^{(k)} \right) & i \neq j,
\end{cases}
\end{aligned}
\tag{54}
$$

where we considered the terms that are statistically dependent on each other together. Now, taking the expected value of this product and using Assumption 4.3, we get

$$
\begin{aligned}
&\mathbb{E}\left[(\hat{\boldsymbol{B}}^{(k)})^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} \hat{\boldsymbol{B}}^{(k)} \right] \\
&= \begin{cases}
\sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \hat{\pi}_\ell^{-1} b_{\ell i}^2 \hat{p}_{\ell i} + \hat{\pi}_i^{-1} \left(1 - 2 \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} b_{\ell i} \hat{p}_{\ell i} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} \sum_{\substack{\ell'=1 \\ \ell' \neq i \\ \ell' \neq \ell}}^{m} b_{\ell i} b_{\ell' i} \hat{p}_{\ell i} \hat{p}_{\ell' i} + \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} b_{\ell i}^2 \hat{p}_{\ell i} \right) & i = j, \\
\sum_{\substack{\ell=1 \\ \ell \neq i \\ \ell \neq j}}^{m} \hat{\pi}_\ell^{-1} b_{\ell i} b_{\ell j} \hat{p}_{\ell i} \hat{p}_{\ell j} + \hat{\pi}_i^{-1} \left(1 - \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} b_{\ell i} \hat{p}_{\ell i} \right) b_{ij} \hat{p}_{ij} & \\
\qquad + \hat{\pi}_j^{-1} b_{ji} \hat{p}_{ji} \left(1 - \sum_{\substack{\ell'=1 \\ \ell' \neq j}}^{m} b_{\ell' j} \hat{p}_{\ell' j} \right) & i \neq j
\end{cases} \\
&= \hat{\boldsymbol{B}}^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} \hat{\boldsymbol{B}} + \begin{cases} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{m} (\hat{\pi}_\ell^{-1} + \hat{\pi}_i^{-1}) b_{\ell i}^2 \hat{p}_{\ell i} (1 - \hat{p}_{\ell i}) & i = j, \\ 0 & i \neq j \end{cases} \\
&= \hat{\boldsymbol{B}}^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} \hat{\boldsymbol{B}} + \begin{cases} \frac{m-1}{\hat{\pi}_i} \overline{(\hat{\pi}_i \hat{\pi}^{-1} + 1) b_{:i}^2 \hat{p}_{:i} (1 - \hat{p}_{:i})} & i = j, \\ 0 & i \neq j \end{cases} = \hat{\boldsymbol{B}}^T \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} \hat{\boldsymbol{B}} + \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1/2} \boldsymbol{\Lambda}_{\hat{\boldsymbol{b}}_0} \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1/2}.
\end{aligned}
\tag{55}
$$

The diagonal matrix $\boldsymbol{\Lambda}_{\hat{\boldsymbol{b}}_0}$ will be an all-zeroes matrix if $\hat{p}_{ij} = 1$ for all links $(i,j) \in \mathcal{E}$. Plugging the result in Eq. 55 back into Eq. 53, we get

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} \hat{\boldsymbol{B}}^{(k)} \boldsymbol{Y}^{(k)} - m \mathbf{1} \bar{\boldsymbol{y}}^{(k)} \right\|_{\hat{\boldsymbol{\pi}}}^2 \right] \le \hat{\rho}_B \mathbb{E}\left[\left\| \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} \boldsymbol{Y}^{(k)} - m \mathbf{1} \bar{\boldsymbol{y}}^{(k)} \right\|_{\hat{\boldsymbol{\pi}}}^2 \right] + \rho(\boldsymbol{\Lambda}_{\hat{\boldsymbol{b}}_0}) \mathbb{E}\left[\left\| \boldsymbol{Y}^{(k)} \right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2 \right] \\
&\le \left(\hat{\rho}_B + \rho(\boldsymbol{\Lambda}_{\hat{\boldsymbol{b}}_0}) \right) \mathbb{E}\left[\left\| \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} \boldsymbol{Y}^{(k)} - m \mathbf{1} \bar{\boldsymbol{y}}^{(k)} \right\|_{\hat{\boldsymbol{\pi}}}^2 \right] + m^2 \rho(\boldsymbol{\Lambda}_{\hat{\boldsymbol{b}}_0}) \mathbb{E}\left[\left\| \bar{\boldsymbol{y}}^{(k)} \right\|^2 \right],
\end{aligned}
\tag{56}
$$

in which $\rho(\mathbf{\Lambda}_{\hat{\boldsymbol{b}}_0}) \leq (m-1)\max(\overline{(\hat{\pi}_i\hat{\pi}^{-1}+1)b_{:i}^2\hat{p}_{:i}(1-\hat{p}_{:i})})$. To ensure a contractive property for the upper bound in Eq. 56, we need to ensure that $\rho(\mathbf{\Lambda}_{\hat{\boldsymbol{b}}_0}) \leq \frac{1-\hat{\rho}_B}{2}$. Noting that $\hat{\rho}_B = 1 - \frac{\min(\hat{\pi}_i)^2\min((b_{ij}\hat{p}_{ij})^+)^2}{\max(\hat{\pi}_i)^3 D(\mathcal{G})K(\mathcal{G})}$ from Lemma C.3-(ii), and that we have the bounds $\frac{\min(b_{ij}^+\hat{p}_{ij})^m}{m} \leq \min(\hat{\pi}_i) \leq \frac{1}{m} \leq \max(\hat{\pi}_i) < 1$ for the eigenvector entries (see Nguyen et al. (2025)), we need to solve for

$$\begin{aligned} \max(\hat{p}_{ij}(1-\hat{p}_{ij})) &\leq \frac{\frac{\min(\hat{\pi}_i)^2\min((b_{ij}\hat{p}_{ij})^+)^2}{\max(\hat{\pi}_i)^3 D(\mathcal{G})K(\mathcal{G})}}{4(m-1)\max(b_{ij}^+)^2}\frac{\min(\hat{\pi}_i)}{\max(\hat{\pi}_i)} \\ \Rightarrow \min(\hat{p}_{ij})^{3m+2} &\geq \frac{4m^3(m-1)\max(b_{ij}^+)^2 D(\mathcal{G})K(\mathcal{G})}{\min(b_{ij}^+)^{3m+2}}(1-\min(\hat{p}_{ij})). \end{aligned} \tag{57}$$

Solving for a range of unknown values $x = \hat{p}_{ij}$ and $\hat{\tau}_B = \frac{4m^3(m-1)\max(b_{ij}^+)^2 D(\mathcal{G})K(\mathcal{G})}{\min(b_{ij}^+)^{3m+2}}$, we write the equation $f_B(x) = x^{3m+2} + \hat{\tau}_B x - \hat{\tau}_B$. Noting that $f_B(0) = -\hat{\tau}_B < 0$ and $f_B(1) = 1 > 0$, there exists a root $\hat{r}_B \in (0,1)$ such that $f_B(\hat{r}_B) = 0$. Since $f(x)$ is an increasing function in x for $x > 0$, which is the interval we are interested in analyzing, the solution to $f_B(x) \geq 0$ will be $x \in [\hat{r}_B, 1]$.

F.4 Proof of Lemma C.5

(i) We have

$$\begin{aligned} &\left\|\left(\hat{\boldsymbol{A}}^{(k)} - \hat{\boldsymbol{A}}\right)\boldsymbol{X}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2 = \left\|\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} \mp \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)} - \hat{\boldsymbol{A}}\boldsymbol{X}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2 \\ &= \left\|\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2 + \left\|\hat{\boldsymbol{A}}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2 - 2\left\langle \hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}, \hat{\boldsymbol{A}}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\rangle_{\hat{\boldsymbol{\phi}}}. \end{aligned} \tag{58}$$

Next, we take the expected value of the result and employ Eq. 5 to write

$$\begin{aligned} \mathbb{E}\left[\left\|\left(\hat{\boldsymbol{A}}^{(k)} - \hat{\boldsymbol{A}}\right)\boldsymbol{X}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] &= \mathbb{E}\left[\left\|\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] - \mathbb{E}\left[\left\|\hat{\boldsymbol{A}}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] \\ &\leq \hat{\rho}_{0,A}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right], \end{aligned} \tag{59}$$

where we applied Lemma C.4-(i) at the end.

(ii) We use Eq. 7 to write

$$\begin{aligned} \left\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2 &= \left\|\hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \hat{\boldsymbol{\phi}}^T\mathbf{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2 \\ &\leq 2\left\|\hat{\boldsymbol{\phi}}^T\left(\hat{\boldsymbol{A}}^{(k)} - \hat{\boldsymbol{A}}\right)\boldsymbol{X}^{(k)}\right\|^2 + 2\left\|\hat{\boldsymbol{\phi}}^T\mathbf{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}\right\|^2 \\ &= 2\left\|\sum_{i=1}^m \hat{\phi}_i\sum_{j=1}^m(\hat{a}_{ij}^{(k)} - \hat{a}_{ij})\boldsymbol{x}_j^{(k)}\right\|^2 + 2\left\|\sum_{i=1}^m \hat{\phi}_i\eta_i^{(k)}\sum_{j=1}^m \hat{b}_{ij}^{(k)}\boldsymbol{y}_j^{(k)}\right\|^2 \\ &\leq 2\sum_{i=1}^m \hat{\phi}_i\left\|\sum_{j=1}^m(\hat{a}_{ij}^{(k)} - \hat{a}_{ij})\boldsymbol{x}_j^{(k)}\right\|^2 + 2\sum_{i=1}^m \hat{\phi}_i\left\|\eta_i^{(k)}\sum_{j=1}^m \hat{b}_{ij}^{(k)}\boldsymbol{y}_j^{(k)}\right\|^2 \\ &= 2\left\|\left(\hat{\boldsymbol{A}}^{(k)} - \hat{\boldsymbol{A}}\right)\boldsymbol{X}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2 + 2\left\|\mathbf{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2, \end{aligned} \tag{60}$$

where we used Young's inequality and Jensen's inequality, respectively. Next, we take the expected value of the result and employ Lemmas C.2-(iii) (it's implication in Eq. 46 in its proof in Appendix F.2) and C.5-(i) to conclude the proof.

(iii) We have

$$
\begin{aligned}
&\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)}-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\right\|^2\\
&=\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)}\mp\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k+1)})\mp\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k)})-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\right\|^2\\
&\leq\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\left(\boldsymbol{G}^{(k+1)}-\nabla\boldsymbol{F}(\boldsymbol{X}^{(k+1)})\right)\right\|^2+2\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\left(\nabla\boldsymbol{F}(\boldsymbol{X}^{(k)})-\boldsymbol{G}^{(k)}\right)\right\|^2\\
&\qquad+2\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k+1)})-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k)})\right\|^2\\
&\qquad+2\Big\langle\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\left(\boldsymbol{G}^{(k+1)}-\nabla\boldsymbol{F}(\boldsymbol{X}^{(k+1)})\right),\\
&\qquad\qquad\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\left(\nabla\boldsymbol{F}(\boldsymbol{X}^{(k)})-\boldsymbol{G}^{(k)}\right)+\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k+1)})-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k)})\Big\rangle.
\end{aligned}
\tag{61}
$$

where we separated the inner product term because it will be equal to 0 when taking its expected value, since $\boldsymbol{g}_i^{(k)}$ is an unbiased estimate of $\nabla F_i(\boldsymbol{x}_i^{(k)})$ according to the Assumption A.1. We further expand the bound as

$$
\begin{aligned}
&\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k+1)})-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k)})\right\|^2\\
&=\Big\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k+1)})\mp\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\nabla\boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\mp\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\nabla\boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\\
&\qquad\qquad\mp\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\nabla\boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\nabla\boldsymbol{F}(\boldsymbol{X}^{(k)})\Big\|^2\\
&\leq 4\Bigg(\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\left(\nabla\boldsymbol{F}(\boldsymbol{X}^{(k+1)})-\nabla\boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\right)\right\|^2+\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\left(\nabla\boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})-\nabla\boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right)\right\|^2\\
&\qquad+\left\|\left(\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\right)\nabla\boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2+\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\left(\nabla\boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})-\nabla\boldsymbol{F}(\boldsymbol{X}^{(k)})\right)\right\|^2\Bigg)\\
&=4\sum_{i=1}^{m}\Bigg(\left\|\nabla F_i(\boldsymbol{x}_i^{(k+1)})-\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\right\|^2v_i^{(k+1)}+\left\|\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})-\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2v_i^{(k+1)}\\
&\qquad+\left\|\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2(v_i^{(k+1)}+v_i^{(k)}-2v_i^{(k+1)}v_i^{(k)})+\left\|\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})-\nabla F_i(\boldsymbol{x}_i^{(k)})\right\|^2v_i^{(k)}\Bigg),\\
&\leq 4\sum_{i=1}^{m}\Bigg(L_i^2\left\|\boldsymbol{x}_i^{(k+1)}-\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}\right\|^2v_i^{(k+1)}+L_i^2\left\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}-\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2v_i^{(k+1)}\\
&\qquad+\left\|\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2(v_i^{(k+1)}+v_i^{(k)}-2v_i^{(k+1)}v_i^{(k)})+L_i^2\left\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}-\boldsymbol{x}_i^{(k)}\right\|^2v_i^{(k)}\Bigg),
\end{aligned}
\tag{62}
$$

in which we first used Young's inequality and then invoked Assumptions 4.1 and 4.2. Plugging Eq. 62 back into Eq. 61 and taking the expected value of the result, we get

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)}-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\right\|^2\right]\\
&\leq\sum_{i=1}^{m}p_i\Bigg(\mathbb{E}\left[\left\|\boldsymbol{g}_i^{(k+1)}-\nabla F_i(\boldsymbol{x}_i^{(k+1)})\right\|^2\right]+2\mathbb{E}\left[\left\|\nabla F_i(\boldsymbol{x}_i^{(k)})-\boldsymbol{g}_i^{(k)}\right\|^2\right]\\
&\qquad+8\Bigg(L_i^2\mathbb{E}\left[\left\|\boldsymbol{x}_i^{(k+1)}-\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}\right\|^2\right]+L_i^2\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}-\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2\right]\\
&\qquad\qquad+2(1-p_i)\left\|\nabla F_i(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2+L_i^2\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}-\boldsymbol{x}_i^{(k)}\right\|^2\right]\Bigg)\Bigg)
\end{aligned}
\tag{63}
$$

$$
\begin{aligned}
&\leq 3m\overline{\sigma_0^2 p} + 3\bar{L}^2 \max\left(\frac{\sigma_{1,i}^2 p_i}{\hat{\phi}_i}\right)\left(\mathbb{E}\left[\left\|\boldsymbol{X}^{(k+1)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)}\right\|_{\hat{\phi}}^2\right] + \mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\|_{\hat{\phi}}^2\right]\right) \\
&\qquad + 3m\overline{\sigma_1^2 p}\left(\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)})\right\|^2\right] + \mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)})\right\|^2\right]\right) \\
&\qquad + 16m\overline{p(1-p)\delta_0^2} + 16m\overline{p(1-p)\delta_1^2}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)})\right\|^2\right] \\
&\qquad + 8\max(p_i L_i^2)\left(\mathbb{E}\left[\left\|\boldsymbol{X}^{(k+1)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)}\right\|^2\right] + \mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\|^2\right]\right) \\
&\qquad + 8m\overline{pL^2}\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\|^2\right], \\
&\leq 3m\overline{\sigma_0^2 p} + 16m\overline{p(1-p)\delta_0^2} \\
&\qquad + \max\left(\frac{p_i}{\hat{\phi}_i}\left(3\sigma_{1,i}^2\bar{L}^2 + 4L_i^2\right)\right)\left(\mathbb{E}\left[\left\|\boldsymbol{X}^{(k+1)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)}\right\|_{\hat{\phi}}^2\right] + \mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\|_{\hat{\phi}}^2\right]\right) \\
&\qquad + 4m\overline{\sigma_1^2 p}\left(\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)})\right\|^2\right] + \mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)})\right\|^2\right]\right) \\
&\qquad + 16m\overline{p(1-p)\delta_1^2}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)})\right\|^2\right] + 8m\overline{pL^2}\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\|^2\right],
\end{aligned}
$$

where we employed Assumption 4.3 and A.1. Noting that $\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)})\|^2 \leq 2\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)}) - \nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)})\|^2 + 2\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)})\|^2 \leq 2\bar{L}^2\|\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\|^2 + 2\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)})\|^2$ concludes the proof.

G Proofs of Main Lemmas

G.1 Proof of Lemma D.1

According to Young's inequality with $\varepsilon > 0$ to be tuned later, we have

$$
\begin{aligned}
&\left\|\boldsymbol{X}^{(k+1)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k+1)}\right\|_{\hat{\phi}}^2 \\
&= \left\|\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - \mathbf{1}(\hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \hat{\boldsymbol{\phi}}^T\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)})\right\|_{\hat{\phi}}^2 \\
&= \Bigg\|\left(\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right) - \mathbf{1}\left(\hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right) \\
&\qquad\qquad - \left(\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - \mathbf{1}\hat{\boldsymbol{\phi}}^T\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}\right)\Bigg\|_{\hat{\phi}}^2 \qquad (64) \\
&\leq (1+\varepsilon)\left\|\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\|_{\hat{\phi}}^2 + 2\left\|\hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\|^2 \\
&\qquad + \left(2+\varepsilon^{-1}\right)\left\|\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - \mathbf{1}\hat{\boldsymbol{\phi}}^T\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}\right\|_{\hat{\phi}}^2 \\
&\qquad - 2\left\langle\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}, \mathbf{1}\left(\hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right)\right\rangle_{\hat{\phi}}.
\end{aligned}
$$

We can simplify the inner-product term as

$$
\begin{aligned}
&2\left\langle\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}, \mathbf{1}\left(\hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right)\right\rangle_{\hat{\phi}} \\
&= 2\sum_{i=1}^m \hat{\phi}_i\left\langle(\hat{\boldsymbol{a}}_i^{(k)})^T\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}, \hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\rangle \qquad (65) \\
&= 2\left\langle\sum_{i=1}^m \hat{\phi}_i(\hat{\boldsymbol{a}}_i^{(k)})^T\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}, \hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\rangle = 2\left\|\hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\phi}}^{(k)}\right\|^2.
\end{aligned}
$$

Plugging Eq. 65 back in Eq. 64 and taking the expected value of the result, we get

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k+1)}-\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] \leq & (1+\varepsilon)(\hat{\rho}_A+\hat{\rho}_{0,A})\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)}-\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] \\ & +(2+\varepsilon^{-1})\kappa_3\max(\eta_i^{(k)})^2\mathbb{E}\left[\left\|\boldsymbol{Y}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2\right],\end{aligned} \tag{66}$$

where we used Lemmas C.4-(i) and C.2-(iii). Next, employing Lemma C.2-(ii), we get

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k+1)}-\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] \leq & \left(\hat{\rho}_A+\hat{\rho}_{0,A}+\varepsilon(\hat{\rho}_A+\hat{\rho}_{0,A})\right)\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)}-\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] \\ & +(2+\varepsilon^{-1})\kappa_3\max(\eta_i^{(k)})^2\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2\right] \\ & +(2+\varepsilon^{-1})\kappa_3 m^2\max(\eta_i^{(k)})^2\mathbb{E}\left[\left\|\bar{\boldsymbol{y}}^{(k)}\right\|^2\right].\end{aligned} \tag{67}$$

Using Lemma C.2-(i) on the last expression and letting $\varepsilon=\frac{1-(\hat{\rho}_A+\hat{\rho}_{0,A})}{2(\hat{\rho}_A+\hat{\rho}_{0,A})}$ conclude the proof.

G.2 Proof of Lemma D.2

According to Young's inequality, we have

$$\begin{aligned}&\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k+1)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k+1)}\right\|_{\hat{\boldsymbol{\pi}}}^2 \\ &=\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\left(\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}+\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)}-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\right)-m\mathbf{1}\left(\bar{\boldsymbol{y}}^{(k)}+\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k+1)}-\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)}\right)\right\|_{\hat{\boldsymbol{\pi}}}^2 \\ &=\Bigg\|\left(\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right)+\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\left(\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)}-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\right) \\ &\qquad\qquad -m\mathbf{1}\left(\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k+1)}-\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)}\right)\Bigg\|_{\hat{\boldsymbol{\pi}}}^2 \\ &\leq\left(1+2\frac{1-(\hat{\rho}_B+\hat{\rho}_{0,B})}{4(\hat{\rho}_B+\hat{\rho}_{0,B})}\right)\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^2 \\ &\quad+\left(2+\frac{4(\hat{\rho}_B+\hat{\rho}_{0,B})}{1-(\hat{\rho}_B+\hat{\rho}_{0,B})}\right)\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\left(\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)}-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\right)\right\|_{\hat{\boldsymbol{\pi}}}^2 \\ &\quad+\left(2+\frac{4(\hat{\rho}_B+\hat{\rho}_{0,B})}{1-(\hat{\rho}_B+\hat{\rho}_{0,B})}\right)\left\|m\mathbf{1}\left(\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k+1)}-\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)}\right)\right\|_{\hat{\boldsymbol{\pi}}}^2 \\ &\leq\frac{1+\hat{\rho}_B+\hat{\rho}_{0,B}}{2(\hat{\rho}_B+\hat{\rho}_{0,B})}\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^2 \\ &\quad+\frac{2(1+\hat{\rho}_B+\hat{\rho}_{0,B})}{1-(\hat{\rho}_B+\hat{\rho}_{0,B})}\left(\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)}-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2+m^2\left\|\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k+1)}-\bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)}\right\|^2\right) \\ &\leq\frac{1+\hat{\rho}_B+\hat{\rho}_{0,B}}{2(\hat{\rho}_B+\hat{\rho}_{0,B})}\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^2 \\ &\quad+\frac{2(1+\hat{\rho}_B+\hat{\rho}_{0,B})}{1-(\hat{\rho}_B+\hat{\rho}_{0,B})}\left(\frac{1}{\min(\hat{\pi}_i)}+m\right)\left\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)}-\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\right\|^2.\end{aligned} \tag{68}$$

We can take the expected value of the above expression and use Lemmas C.4-(ii) to upper bound $\mathbb{E}[\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\|_{\hat{\boldsymbol{\pi}}}^2]$ and C.5-(iii) to upper bound $\|\boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k+1)}\boldsymbol{G}^{(k+1)} - \boldsymbol{\Lambda}_{\boldsymbol{v}}^{(k)}\boldsymbol{G}^{(k)}\|$ to get

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k+1)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k+1)}\right\|_{\hat{\boldsymbol{\pi}}}^2\right] &\leq \frac{1+\hat{\rho}_B+\hat{\rho}_{0,B}}{2}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^2\right] \\
&+ \frac{(1+\hat{\rho}_B+\hat{\rho}_{0,B})\hat{\rho}_{0,B}}{2(\hat{\rho}_B+\hat{\rho}_{0,B})}m^2\mathbb{E}\left[\|\bar{\boldsymbol{y}}^{(k)}\|^2\right] \\
&+ \frac{2(m+\min(\hat{\pi}_i)^{-1})(1+\hat{\rho}_B+\hat{\rho}_{0,B})}{1-(\hat{\rho}_B+\hat{\rho}_{0,B})}\Bigg[2m\overline{\sigma_0^2 p} + 8m\overline{p(1-p)\delta_0^2} \\
&\quad + \kappa_4\left(\mathbb{E}\left[\left\|\boldsymbol{X}^{(k+1)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] + \mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right]\right) \\
&\quad + 2m\left(3\overline{\sigma_1^2 p} + 4\overline{p(1-p)\delta_1^2}\right)\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right] + 8m\kappa_5\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2\right]\Bigg].
\end{aligned} \tag{69}$$

Next, we use Lemma C.5-(ii) to upper bound $\mathbb{E}[\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\|^2]$ and Lemma D.1 to upper bound $\mathbb{E}[\|\boldsymbol{X}^{(k+1)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}\|_{\hat{\boldsymbol{\phi}}}^2]$ as

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k+1)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k+1)}\right\|_{\hat{\boldsymbol{\pi}}}^2\right] &\leq \frac{1+\hat{\rho}_B+\hat{\rho}_{0,B}}{2}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^2\right] \\
&+ \frac{(1+\hat{\rho}_B+\hat{\rho}_{0,B})\hat{\rho}_{0,B}}{2(\hat{\rho}_B+\hat{\rho}_{0,B})}m^2\mathbb{E}\left[\|\bar{\boldsymbol{y}}^{(k)}\|^2\right] \\
&+ \frac{2(m+\min(\hat{\pi}_i)^{-1})(1+\hat{\rho}_B+\hat{\rho}_{0,B})}{1-(\hat{\rho}_B+\hat{\rho}_{0,B})}\Bigg[2m\overline{\sigma_0^2 p} + 8m\overline{p(1-p)\delta_0^2} + \kappa_4\omega_1^{(k)} \\
&\quad + \left(\kappa_4(1+\psi_{11}^{(k)}) + 16m\kappa_5\hat{\rho}_{0,A}\right)\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right] \\
&\quad + \kappa_4\psi_{12}^{(k)}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^2\right] + 16m\kappa_5\kappa_3\max(\eta_i^{(k)})^2\mathbb{E}\left[\left\|\boldsymbol{Y}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}^{-1}}^2\right] \\
&\quad + \left(\kappa_4\gamma_1^{(k)} + 6m\overline{\sigma_1^2 p} + 8m\overline{p(1-p)\delta_1^2}\right)\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right]\Bigg].
\end{aligned} \tag{70}$$

Finally, since Lemma C.2-(ii) gives us $\|\boldsymbol{Y}^{(k)}\|_{\hat{\boldsymbol{\pi}}^{-1}}^2 \leq \|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\|_{\hat{\boldsymbol{\pi}}}^2 + m^2\|\bar{\boldsymbol{y}}^{(k)}\|^2$, we can Lemma C.2-(i) to upper bound $\mathbb{E}[\|\bar{\boldsymbol{y}}^{(k)}\|^2]$ and conclude the proof.

G.3 Proof of Lemma D.3

Let $\boldsymbol{x} = \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}$ and $\boldsymbol{x}' = \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}$ in $F(\boldsymbol{x}) \leq F(\boldsymbol{x}') + \langle\nabla F(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}'\rangle + \frac{1}{2}\bar{L}\|\boldsymbol{x} - \boldsymbol{x}'\|^2$, which is a consequence of Assumption 4.1. We get

$$F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)}) \leq F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) + \left\langle\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\rangle + \frac{1}{2}\bar{L}\left\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|^2. \tag{71}$$

First, we use Eq. 7 (see Definition 4.5 for the notation) to write

$$\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)} = \hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)} - \hat{\boldsymbol{\phi}}^T\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}. \tag{72}$$

Therefore, we need to obtain upper bounds for (i) $\langle\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), -\hat{\boldsymbol{\phi}}^T\boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)}\hat{\boldsymbol{B}}^{(k)}\boldsymbol{Y}^{(k)}\rangle$, (ii) $\langle\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), \hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{A}}^{(k)}\boldsymbol{X}^{(k)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\rangle$ and (iii) $\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\|^2$. We already obtained an upper bound for (iii) in Lemma C.5-(ii), and the expected value of (ii) will be zero according to Assumption 4.3

(See Eq. 5). We proceed with a bound for (i) as

$$
\begin{aligned}
&\left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), -\hat{\boldsymbol{\phi}}^T \boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)} \hat{\boldsymbol{B}}^{(k)} \boldsymbol{Y}^{(k)} \right\rangle = \sum_{i=1}^{m} \hat{\phi}_i \eta_i^{(k)} \left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), -\sum_{j=1}^{m} \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)} \right\rangle \\
&= \sum_{i=1}^{m} \hat{\phi}_i \eta_i^{(k)} \left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), \mp m \hat{\pi}_i \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) - \sum_{j=1}^{m} \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)} \right\rangle \\
&= \sum_{i=1}^{m} \hat{\phi}_i \hat{\pi}_i \eta_i^{(k)} \left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), \mp m \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) - \frac{1}{\hat{\pi}_i} \sum_{j=1}^{m} \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)} \right\rangle \\
&= \sum_{i=1}^{m} \hat{\phi}_i \hat{\pi}_i \eta_i^{(k)} \left(-m \left\| \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) \right\|^2 + \left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), m \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) - \frac{1}{\hat{\pi}_i} \sum_{j=1}^{m} \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)} \right\rangle \right).
\end{aligned}
\tag{73}
$$

Focusing on the inner-product term, we can expand it as

$$
\begin{aligned}
&\left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), m \left(\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) \mp \overline{\nabla \boldsymbol{F}(\mathbf{1} \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} \mp \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)} \mp \bar{\boldsymbol{y}}^{(k)} \right) \right. \\
&\qquad \left. - \frac{1}{\hat{\pi}_i} \sum_{j=1}^{m} \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)} \right\rangle \\
&= \left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), m \left(\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) \mp \overline{\nabla \boldsymbol{F}(\mathbf{1} \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} - \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} + \bar{\boldsymbol{y}}^{(k)} \right) - \frac{1}{\hat{\pi}_i} \sum_{j=1}^{m} \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)} \right\rangle \\
&\qquad + \left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), m \left(\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} - \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)} \right) \right\rangle \\
&\leq \frac{m}{2} \left\| \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) \right\|^2 + \frac{3}{2m} \left(m^2 \left\| \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) - \overline{\nabla \boldsymbol{F}(\mathbf{1} \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} \right\|^2 \right. \\
&\qquad \left. + m^2 \left\| \overline{\nabla \boldsymbol{F}(\mathbf{1} \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} - \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} \right\|^2 + \left\| m \bar{\boldsymbol{y}}^{(k)} - \frac{1}{\hat{\pi}_i} \sum_{j=1}^{m} \hat{b}_{ij}^{(k)} \boldsymbol{y}_j^{(k)} \right\|^2 \right) \\
&\qquad + \left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), m \left(\overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} - \bar{\boldsymbol{g}}_{\boldsymbol{v}}^{(k)} \right) \right\rangle,
\end{aligned}
\tag{74}
$$

where we separated the inner product term from the others because it will be equal to 0 when taking its expected value, since $\boldsymbol{g}_i^{(k)}$ is an unbiased estimate of $\nabla F_i(\boldsymbol{x}_i^{(k)})$ according to the Assumption A.1. Plugging Eq. 74 back in Eq. 73, and taking its expected value, we get

$$
\begin{aligned}
&\mathbb{E}\left[\left\langle \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}), -\hat{\boldsymbol{\phi}}^T \boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{(k)} \hat{\boldsymbol{B}}^{(k)} \boldsymbol{Y}^{(k)} \right\rangle \right] \leq -\frac{m^2}{2} \overline{\hat{\phi} \hat{\pi} \eta^{(k)}} \mathbb{E}\left[\left\| \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) \right\|^2 \right] \\
&\quad + \frac{3m^2}{2} \overline{\hat{\phi} \hat{\pi} \eta^{(k)}} \left(\mathbb{E}\left[\left\| \nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) - \overline{\nabla \boldsymbol{F}(\mathbf{1} \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} \right\|^2 \right] \right. \\
&\qquad \left. + \mathbb{E}\left[\left\| \overline{\nabla \boldsymbol{F}(\mathbf{1} \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} - \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}} \right\|^2 \right] \right) \\
&\quad + \frac{3 \max(\hat{\phi}_i \eta_i^{(k)})}{2m} (\hat{\rho}_B + \hat{\rho}_{0,B}) \mathbb{E}\left[\left\| \boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1} \boldsymbol{Y}^{(k)} - m \mathbf{1} \bar{\boldsymbol{y}}^{(k)} \right\|_{\hat{\boldsymbol{\pi}}}^2 \right] \\
&\quad + \frac{3m \max(\hat{\phi}_i \eta_i^{(k)})}{2} \hat{\rho}_{0,B} \mathbb{E}\left[\| \bar{\boldsymbol{y}}^{(k)} \|^2 \right],
\end{aligned}
\tag{75}
$$

where we invoked Lemma C.4-(ii). We then substitute Eq. 75 back into Eq. 71 and take the expected value of the result to get

$$\begin{aligned}
\mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\right] &\leq \mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right] - \frac{m^2}{2}\overline{\hat{\phi}\hat{\pi}\eta^{(k)}}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right] \\
&+\frac{3m^2}{2}\overline{\hat{\phi}\hat{\pi}\eta^{(k)}}\Bigg(\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}) - \overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}}\right\|^2\right] \\
&\qquad + \mathbb{E}\left[\left\|\overline{\nabla \boldsymbol{F}(\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})}_{\boldsymbol{v}^{(k)}} - \overline{\nabla \boldsymbol{F}(\boldsymbol{X}^{(k)})}_{\boldsymbol{v}^{(k)}}\right\|^2\right]\Bigg) \\
&+\left(\frac{3\max(\hat{\phi}_i)}{2m}(\hat{\rho}_B + \hat{\rho}_{0,B}) + \bar{L}\kappa_3\max(\eta_i^{(k)})\right)\max(\eta_i^{(k)})\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(k)} - m\mathbf{1}\bar{\boldsymbol{y}}^{(k)}\right\|_{\hat{\boldsymbol{\pi}}}^2\right] \\
&+m\left(\frac{3}{2}\max(\hat{\phi}_i)\hat{\rho}_{0,B} + m\bar{L}\kappa_3\max(\eta_i^{(k)})\right)\max(\eta_i^{(k)})\mathbb{E}\left[\|\bar{\boldsymbol{y}}^{(k)}\|^2\right] \\
&+\bar{L}\hat{\rho}_{0,A}\mathbb{E}\left[\left\|\boldsymbol{X}^{(k)} - \mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right],
\end{aligned} \tag{76}$$

where we invoked Lemma C.5-(ii) as discussed in the beginning of the proof for $\|\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)} - \bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)}\|^2$. Finally applying Lemmas C.1-(ii), C.1-(iii) and C.2-(i) completes the proof.

H Proofs of Propositions

H.1 Proof of Proposition 4.7

We have that $\rho(\boldsymbol{\Psi}^{(k)}) = \max(|\lambda_1|, |\lambda_2|)$, where λ_1 and λ_2 are the eigenvalues of the matrix $\boldsymbol{\Psi}^{(k)}$. Since $\mathrm{tr}(\boldsymbol{\Psi}^{(k)}) = \lambda_1 + \lambda_2$ and $\det(\boldsymbol{\Psi}^{(k)}) = \lambda_1\lambda_2$, we have that $\max(|\lambda_1|, |\lambda_2|) = \max(\frac{1}{2}|\,\mathrm{tr}(\boldsymbol{\Psi}^{(k)}) \pm (\mathrm{tr}(\boldsymbol{\Psi}^{(k)})^2 - 4\det(\boldsymbol{\Psi}^{(k)}))^{1/2}|)$. Therefore, by noting that $\mathrm{tr}(\boldsymbol{\Psi}^{(k)}) = \psi_{11}^{(k)} + \psi_{22}^{(k)} > 0$ due to both $\psi_{11}^{(k)} > 0$ and $\psi_{22}^{(k)} > 0$, we will have that $\rho(\boldsymbol{\Psi}^{(k)}) = \frac{1}{2}(\mathrm{tr}(\boldsymbol{\Psi}^{(k)}) + (\mathrm{tr}(\boldsymbol{\Psi}^{(k)})^2 - 4\det(\boldsymbol{\Psi}^{(k)}))^{1/2})$. Solving this equation, we get $\rho(\boldsymbol{\Psi}^{(k)}) = \frac{1}{2}(\psi_{11}^{(k)} + \psi_{22}^{(k)} + ((\psi_{11}^{(k)} - \psi_{22}^{(k)})^2 + 4\psi_{12}^{(k)}\psi_{21}^{(k)})^{1/2})$, and to have $\rho(\boldsymbol{\Psi}^{(k)}) < 1$, the following must hold.

$$\begin{aligned}
&(\psi_{11}^{(k)} - \psi_{22}^{(k)})^2 + 4\psi_{12}^{(k)}\psi_{21}^{(k)} < 4 + (\psi_{11}^{(k)})^2 + (\psi_{22}^{(k)})^2 - 4\psi_{11}^{(k)} - 4\psi_{22}^{(k)} + 2\psi_{11}^{(k)}\psi_{22}^{(k)} \\
&\Rightarrow \psi_{12}^{(k)}\psi_{21}^{(k)} < 1 - \psi_{11}^{(k)} - \psi_{22}^{(k)} + \psi_{11}^{(k)}\psi_{22}^{(k)} = (1 - \psi_{11}^{(k)})(1 - \psi_{22}^{(k)})
\end{aligned} \tag{77}$$

Since we have both $\psi_{12}^{(k)} > 0$ and $\psi_{21}^{(k)} > 0$, we need to ensure that both $\psi_{11}^{(k)} < 1$ and $\psi_{22}^{(k)} < 1$ to be able to solve for Eq. 77. We will enforce these two constraints by solving for (i) $\psi_{11}^{(k)} < \frac{3+\tilde{\rho}_A}{4}$ and (ii) $\psi_{22}^{(k)} < \frac{7+\tilde{\rho}_B}{8}$. Furthermore, noting that $\psi_{21}^{(k)}$ and $\psi_{22}^{(k)}$ are comprised of the terms $\kappa_7\kappa_4(1 + \psi_{11}^{(k)}) + 16m\kappa_7\kappa_5\hat{\rho}_{0,A} + \kappa_8\kappa_1\hat{\rho}_{0,B}$ and $\kappa_7\kappa_4\psi_{12}^{(k)}$ according to Lemma D.2, respectively, we will also enforce $\psi_{21}^{(k)} < 2\kappa_7\kappa_4 + 16m\kappa_7\kappa_5\hat{\rho}_{0,A} + \kappa_8\kappa_1\hat{\rho}_{0,B}$ and $\psi_{12}^{(k)} < \frac{1-\tilde{\rho}_B}{4\kappa_7\kappa_4}$. However, plugging the constraints (i) and (ii) into Eq. 77 also requires us to satisfy $\psi_{12}^{(k)}\psi_{21}^{(k)} < \frac{1-\tilde{\rho}_A}{4}\frac{1-\tilde{\rho}_B}{8}$. Due to more flexibility satisfying stricter conditions over $\psi_{12}^{(k)}$, we choose the last two constraints as (iii) $\psi_{12}^{(k)} < \frac{1-\tilde{\rho}_B}{8(2\kappa_7\kappa_4 + 16m\kappa_7\kappa_5\hat{\rho}_{0,A} + \kappa_8\kappa_1\hat{\rho}_{0,B})}\frac{1-\tilde{\rho}_A}{4}$ and $\psi_{21}^{(k)} < 2\kappa_7\kappa_4 + 16m\kappa_7\kappa_5\hat{\rho}_{0,A} + \kappa_8\kappa_1\hat{\rho}_{0,B}$. Using the values $\psi_{ij}^{(k)}$ given in Lemmas D.1 and D.2, we have

- $\psi_{11}^{(k)} < \frac{3+\tilde{\rho}_A}{4}$:

$$\max(\eta_i^{(k)}) < \frac{1 - \tilde{\rho}_A}{2\sqrt{2m\kappa_1\kappa_3}}. \tag{78}$$

- $\psi_{12}^{(k)} < \frac{1-\tilde{\rho}_B}{8(2\kappa_7\kappa_4 + 16m\kappa_7\kappa_5\hat{\rho}_{0,A} + \kappa_8\kappa_1\hat{\rho}_{0,B})}\frac{1-\tilde{\rho}_A}{4}$:

$$\max(\eta_i^{(k)}) < \frac{1 - \tilde{\rho}_B}{\sqrt{1 + \tilde{\rho}_B}}\frac{\sqrt{\min(\hat{\pi}_i)/(m\kappa_3)}(1 - \tilde{\rho}_A)/16}{\sqrt{\frac{2\kappa_4}{m} + 16\kappa_5\hat{\rho}_{0,A} + \frac{1-\tilde{\rho}_B}{8\tilde{\rho}_B}\min(\hat{\pi}_i)\kappa_1\hat{\rho}_{0,B}}}, \tag{79}$$

where we used the fact that $m + \min(\hat{\pi}_i)^{-1} \leq 2\min(\hat{\pi}_i)^{-1}$.

- $\psi_{21}^{(k)} < 2\kappa_7\kappa_4 + 16m\kappa_7\kappa_5\hat{\rho}_{0,A} + \kappa_8\kappa_1\hat{\rho}_{0,B}$:

$$\max(\eta_i^{(k)}) < \frac{1}{8m}\sqrt{\frac{1-\tilde{\rho}_A}{\kappa_5\kappa_3}}, \tag{80}$$

where we used the fact that $\psi_{11}^{(k)} < \frac{3+\tilde{\rho}_A}{4}$ and $\kappa_1 \leq \kappa_4$.

- $\psi_{22}^{(k)} < \frac{7+\tilde{\rho}_B}{8}$:

$$\max(\eta_i^{(k)}) < \frac{1}{16}\frac{1-\tilde{\rho}_B}{\sqrt{1+\tilde{\rho}_B}}\sqrt{\frac{\min(\hat{\pi}_i)}{2m\kappa_5\kappa_3}} \tag{81}$$

where we used the fact that $m + \min(\hat{\pi}_i)^{-1} \leq 2\min(\hat{\pi}_i)^{-1}$, and used the looser upper bound $\psi_{12}^{(k)} < \frac{1-\tilde{\rho}_B}{4\kappa_7\kappa_4}$ here to obtain cleaner results, although a stricter upper bound $\psi_{12}^{(k)} < \frac{1-\tilde{\rho}_B}{8(2\kappa_7\kappa_4+16m\kappa_7\kappa_5\hat{\rho}_{0,A}+\kappa_8\kappa_1\hat{\rho}_{0,B})}\frac{1-\tilde{\rho}_A}{4}$ was guaranteed.

We obtained four constraints on the maximum learning rate $\max(\eta_i^{(k)})$ that need to be satisfied to have $\rho(\boldsymbol{\Psi}^{(k)}) < 1$. Noting that the learning rates satisfying Eq. 79 are a subset of Eq. 78 (due to $\kappa_1 \leq \kappa_4$) concludes the proof.

H.2 Proof of Proposition 4.8

We must ensure that the coefficients for $\mathbb{E}[\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\|^2]$ in Eq. 36 are negative for all $r = 0, ..., k$, i.e., $\gamma_0^{(k)} > 0$ and $\gamma_0^{(r)} - \left(\sum_{s=r+1}^{k} (\boldsymbol{\psi}_0^{(s)})^T \boldsymbol{\Psi}^{(s-2:r)}\right)\boldsymbol{\gamma}^{(r)} > 0$ for $r = 0, ..., k-1$. Using the definitions of $\gamma_0^{(k)}$ and $\boldsymbol{\psi}_0^{(k)}$ from Lemma D.3 and $\boldsymbol{\gamma}^{(k)}$ from Lemmas D.1 and D.2, we have

- $\gamma_0^{(k)} > 0$:
We need to obtain the conditions under which we have

$$m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}} - 3m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}(1-p)\delta_1^2} - (3\max(\hat{\phi}_i\eta_i^{(k)})\hat{\rho}_{0,B} + 2m\bar{L}\kappa_3\max(\eta_i^{(k)})^2)\kappa_2 > 0. \tag{82}$$

Let us enforce $\frac{m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}}}{\Gamma_1} - 3m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}(1-p)\delta_1^2} - (3\max(\hat{\phi}_i\eta_i^{(k)})\hat{\rho}_{0,B} + 2m\bar{L}\kappa_3\max(\eta_i^{(k)})^2)\kappa_2 \geq 0$ for a scalar $\Gamma_1 > 1$. which holds if (i) $3(1-p_i)\delta_{1,i}^2 \leq \frac{1}{3\Gamma_1}$, (ii) $3\max(\hat{\phi}_i\eta_i^{(k)})\hat{\rho}_{0,B}\kappa_2 \leq \frac{m\overline{\hat{\phi}\hat{\pi}\eta^{(k)}}}{3\Gamma_1}$ and (iii) $2\bar{L}\kappa_3\max(\eta_i^{(k)})^2\kappa_2 \leq \frac{\overline{\hat{\phi}\hat{\pi}\eta^{(k)}}}{3\Gamma_1}$ are satisfied. Noting that (ii) and (iii) have dependencies on both $\max(\eta_i^{(k)})$ and $\bar{\eta}^{(k)}$, we first enforce the following using an arbitrary scalar $\Gamma_2 > 1$

$$\frac{\bar{\eta}^{(k)}}{\max(\eta_i^{(k)})} \geq \frac{1}{\Gamma_2}, \tag{83}$$

Using this, we need to ensure

$$\begin{aligned}
&\text{(i) } p_i \geq 1 - \frac{1}{9\Gamma_1\delta_{1,i}^2},\\
&\text{(ii) } \hat{p}_{ij}(1-\hat{p}_{ij}) \leq \frac{1}{18\Gamma_1\Gamma_2\kappa_2 b_{ij}^2}\frac{m}{m-1}\frac{\min(\hat{\phi}_i)}{\max(\hat{\phi}_i)}\frac{\min(\hat{\pi}_i)^2}{\max(\hat{\pi}_i)},\\
&\text{(iii) } \eta_i^{(k)} \leq \frac{\min(\hat{\phi}_i\hat{\pi}_i)}{6\Gamma_1\Gamma_2\kappa_3\kappa_2\bar{L}},
\end{aligned} \tag{84}$$

where we used the fact that $\hat{\rho}_{0,B} = 2(m-1)\frac{\max(\hat{\pi}_i)}{\min(\hat{\pi}_i)}\max(b_{ij}^2\hat{p}_{ij}(1-\hat{p}_{ij}))$ from Lemma C.4-(ii). Solving for (ii), we get $\hat{p}_{ij} \geq 1 - \hat{r}'_{ij}$ with $\hat{r}'_{ij} = \frac{1}{18\Gamma_1\Gamma_2\kappa_2 b_{ij}^2}\frac{m}{m-1}\frac{\min(\hat{\phi}_i)}{\max(\hat{\phi}_i)}\frac{\min(\hat{\pi}_i)^2}{\max(\hat{\pi}_i)}$, which also implies that $\hat{\rho}_{0,B} \leq \frac{\max(\hat{p}_{ij})}{9\Gamma_1\Gamma_2\kappa_2}\frac{m\overline{\hat{\phi}\hat{\pi}}}{\max(\hat{\phi}_i)}$. All of the constraints (i)-(iii) combined result in the lower bound for $\gamma_0^{(k)} > \frac{1}{2}m^2(1-\frac{1}{\Gamma_1})\overline{\hat{\phi}\hat{\pi}\eta^{(k)}}$.

- $\gamma_0^{(r)} - (\sum_{s=r+1}^{k} (\boldsymbol{\psi}_0^{(s)})^T \boldsymbol{\Psi}^{(s-2:r)}) \boldsymbol{\gamma}^{(r)} > 0$:

  To simplify this constraint, we further strengthen it to $\gamma_0^{(r)} - \sum_{s=r+1}^{k} (\boldsymbol{\psi}_0^{(s)})^T (\prod_{q=r}^{s-2} \rho(\boldsymbol{\Psi}^{(q)})) \boldsymbol{\gamma}^{(r)} > 0$. Since the values of $\boldsymbol{\psi}_0^{(k)}$ are proportional to the non-increasing learning rate $\max(\eta_i^{(k)})$ according to the Lemma D.3, the values of $\boldsymbol{\psi}_0^{(k)}$ will also be non-increasing, i.e., $\boldsymbol{\psi}_0^{(k+1)} \leq \boldsymbol{\psi}_0^{(k)}$. We can use this fact to ensure that the constraint is valid for all $s = r+1, ..., k$ by strengthening it to $\gamma_0^{(r)} - (\boldsymbol{\psi}_0^{(r+1)})^T (\sum_{s=r+1}^{k} (\prod_{q=r}^{s-2} \rho(\boldsymbol{\Psi}^{(q)}))) \boldsymbol{\gamma}^{(r)} > 0$, or $\gamma_0^{(r)} - (\boldsymbol{\psi}_0^{(r)})^T (\sum_{s=r+1}^{k} (\prod_{q=r}^{s-2} \rho(\boldsymbol{\Psi}^{(q)}))) \boldsymbol{\gamma}^{(r)} > 0$ to make the superscripts the same. Since we are using a constant learning rate $\eta_i^{(k)} = \eta_i^{(0)}$, we will have $\rho(\boldsymbol{\Psi}^{(k)}) = \rho_\Psi$ for all $k \geq 0$, and thus the constraint can be simplified to $\gamma_0^{(r)} - (\boldsymbol{\psi}_0^{(r)})^T (\sum_{s=r+1}^{k} \rho_\Psi^{s-1-r}) \boldsymbol{\gamma}^{(r)} > 0$. We strengthen this bound further by letting the sum run to infinity, and by employing the fact that $0 < \rho_\Psi < 1$, we get the final constraint to be

$$\gamma_0^{(r)} - \frac{(\boldsymbol{\psi}_0^{(r)})^T \boldsymbol{\gamma}^{(r)}}{1 - \rho_\Psi} > 0. \tag{85}$$

Let us enforce a stronger constraint $\frac{(\boldsymbol{\psi}_0^{(r)})^T \boldsymbol{\gamma}^{(r)}}{1-\rho_\Psi} < \frac{\gamma_0^{(r)}}{\Gamma_3}$ with $\Gamma_3 > 1$. Plugging in the values of these parameters from Lemmas D.1-D.3 and doing some algebra, we get

$$\begin{aligned}
&2\left(m\kappa_6\kappa_1 + \kappa_7\kappa_6\kappa_4 + 16m\kappa_7\kappa_5\right)\kappa_3^2\kappa_2\bar{L}\max(\eta_i^{(r)})^4 \\
&+3\Bigg(\kappa_6\left(\kappa_1\hat{\rho}_{0,B} + \max\left(\frac{p_i}{\hat{\phi}_i}L_i^2\right)\right) \\
&\qquad + \kappa_7\tilde{\rho}_B\left(16\kappa_5 + \frac{\kappa_6\kappa_4}{m}\right)\Bigg)\max(\hat{\phi}_i)\kappa_3\kappa_2\max(\eta_i^{(r)})^3 \\
&+\frac{2}{m}\left(2\kappa_7(3\overline{\sigma_1^2 p} + 4\overline{p(1-p)\delta_1^2}) + \kappa_8\kappa_2\hat{\rho}_{0,B} + m\kappa_6\kappa_2\hat{\rho}_{0,A}\right)\kappa_3\bar{L}\max(\eta_i^{(r)})^2 \\
&< \frac{1-\rho_\Psi}{\Gamma_3}\left(1 - \frac{1}{\Gamma_1}\right)\min(\hat{\phi}_i\hat{\pi}_i)\bar{\eta}^{(r)} \\
&\qquad - \frac{3}{m^2}\max(\hat{\phi}_i)\tilde{\rho}_B\left(\kappa_8\kappa_2\hat{\rho}_{0,B} + 2\kappa_7(3\overline{\sigma_1^2 p} + 4\overline{p(1-p)\delta_1^2})\right)\max(\eta_i^{(r)}),
\end{aligned} \tag{86}$$

where we used $\gamma_0^{(r)} > \frac{1}{2}m^2(1 - \frac{1}{\Gamma_1})\overline{\hat{\phi}\hat{\pi}\eta^{(r)}}$ as a result of the first part of this proof. To be able to satisfy this, we first need to ensure that the right-hand side of the inequality is positive. We have

$$\frac{\bar{\eta}^{(r)}}{\max(\eta_i^{(r)})} > \frac{3\Gamma_3}{1 - \frac{1}{\Gamma_1}} \frac{\tilde{\rho}_B\left(2\kappa_7(3\overline{\sigma_1^2 p} + 4\overline{p(1-p)\delta_1^2}) + \kappa_8\kappa_2\hat{\rho}_{0,B}\right)}{m^2(1-\rho_\Psi)} \frac{\max(\hat{\phi}_i)}{\min(\hat{\phi}_i\hat{\pi}_i)}. \tag{87}$$

Noting that we had $\frac{\bar{\eta}^{(r)}}{\max(\eta_i^{(r)})} \geq \frac{1}{\Gamma_2}$ from Eq. 83, let us enforce $\frac{3\Gamma_3}{1-\frac{1}{\Gamma_1}} \frac{\tilde{\rho}_B\left(2\kappa_7(3\overline{\sigma_1^2 p} + 4\overline{p(1-p)\delta_1^2}) + \kappa_8\kappa_2\hat{\rho}_{0,B}\right)}{m^2(1-\rho_\Psi)} \frac{\max(\hat{\phi}_i)}{\min(\hat{\phi}_i\hat{\pi}_i)} < \frac{1}{\Gamma_2\Gamma_4}$ for arbitrary $\Gamma_4 > 1$. Using $\hat{\rho}_{0,B} \leq \frac{\max(\hat{p}_{ij})}{9\Gamma_1\Gamma_2\kappa_2} \frac{m\overline{\hat{\phi}\hat{\pi}}}{\max(\hat{\phi}_i)}$ from Eq. 84, we have

$$\begin{aligned}
&2\kappa_7 p_i(3\sigma_{1,i}^2 + 4(1-p_i)\delta_{1,i}^2) + \kappa_8\frac{\max(\hat{p}_{ij})}{9\Gamma_1\Gamma_2}\frac{m\overline{\hat{\phi}\hat{\pi}}}{\max(\hat{\phi}_i)} \\
&\qquad\qquad \leq \frac{m^2(1-\rho_\Psi)}{3\Gamma_4\Gamma_3\Gamma_2\tilde{\rho}_B}\left(1 - \frac{1}{\Gamma_1}\right)\frac{\min(\hat{\phi}_i\hat{\pi}_i)}{\max(\hat{\phi}_i)} \\
\Rightarrow p_i \leq \frac{m^2(1-\rho_\Psi)}{12\Gamma_4\Gamma_3\Gamma_2\tilde{\rho}_B\kappa_7}&\frac{1 - 1/\Gamma_1}{3\sigma_{1,i}^2 + 4\delta_{1,i}^2}\frac{\min(\hat{\phi}_i\hat{\pi}_i)}{\max(\hat{\phi}_i)}, \qquad \hat{p}_{ij} \leq \frac{3m(1-\rho_\Psi)}{2\Gamma_4\Gamma_3\tilde{\rho}_B\kappa_8}\left(\Gamma_1 - 1\right).
\end{aligned} \tag{88}$$

So far, we have obtained the allowed range $p_i \in (1 - \frac{1}{9\Gamma_1\delta_{1,i}^2}, \frac{m^2(1-\rho_\Psi)}{12\Gamma_4\Gamma_3\Gamma_2\tilde{\rho}_B\kappa_7}\frac{1-1/\Gamma_1}{3\sigma_{1,i}^2+4\delta_{1,i}^2}\frac{\min(\hat{\phi}_i\hat{\pi}_i)}{\max(\hat{\phi}_i)}]$ for gradient probabilities and $\hat{p}_{ij} \in (1 - \hat{r}'_{ij}, \frac{3m(1-\rho_\Psi)}{2\Gamma_4\Gamma_3\tilde{\rho}_B\kappa_8}(\Gamma_1 - 1)]$ for communication probabilities. However, by tuning Γ_1, we can make the maximum values of these ranges to be one as follows

$$\Gamma_1 = 1 + \frac{2\Gamma_4\Gamma_3\tilde{\rho}_B}{m(1-\rho_\Psi)}\max\left\{\frac{6\Gamma_2\kappa_7}{m}\frac{\max(\hat{\phi}_i)}{\min(\hat{\phi}_i\hat{\pi}_i)}(3\sigma_{1,i}^2 + 4\delta_{1,i}^2), \frac{\kappa_8}{3}\right\}. \tag{89}$$

Finally, using the lower bounds obtained in Eqs. 83 and 88 in Eq. 86, we get

$$\begin{aligned}
&2\left(m\kappa_1\kappa_6 + \kappa_7\kappa_4\kappa_6 + 16m\kappa_7\kappa_5\right)\kappa_3^2\kappa_2\bar{L}\max(\eta_i^{(r)})^4 \\
&+3\Bigg(\left(\max\left(\frac{p_i}{\hat{\phi}_i}L_i^2\right) + \hat{\rho}_{0,B}\kappa_1\right)\kappa_6 \\
&\qquad + \left(\frac{\kappa_4\kappa_6}{m} + 16\kappa_5\right)\tilde{\rho}_B\kappa_7\Bigg)\max(\hat{\phi}_i)\kappa_3\kappa_2\max(\eta_i^{(r)})^3 \\
&+\frac{2}{m}\left(2\kappa_7(3\overline{\sigma_1^2 p} + 4\overline{p(1-p)\delta_1^2}) + m\kappa_2\kappa_6\hat{\rho}_{0,A} + \kappa_8\kappa_2\hat{\rho}_{0,B}\right)\kappa_3\bar{L}\max(\eta_i^{(r)})^2 \\
&< \frac{1-\rho_\Psi}{\Gamma_2\Gamma_3}\left(1 - \frac{1}{\Gamma_1}\right)\left(1 - \frac{1}{\Gamma_4}\right)\min(\hat{\phi}_i\hat{\pi}_i)\bar{\eta}^{(r)}.
\end{aligned} \tag{90}$$

Setting each one of the three terms in the left-hand side less than one third of the right-hand side, setting the gradient estimation variance $\sigma_{0,i}^2 = \sigma_{1,i}^2 = 0$ for the exact-gradient case and choosing $\Gamma_3 = \Gamma_4 = 2$ for simplicity concludes the proof.

I Proof of Theorem 4.9

We start from Eq. 36 and use Corollary E.2 to write

$$\begin{aligned}
\mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\right] &\leq F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) - \gamma_0^{(0)}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k)})\right\|^2\right] \\
&- \sum_{r=0}^{k-1}\left(\gamma_0^{(0)} - (\boldsymbol{\psi}_0^{(0)})^T\left(\sum_{s=r+1}^{k}\boldsymbol{\Psi}^{s-1-r}\right)\boldsymbol{\gamma}^{(0)}\right)\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\right\|^2\right] \\
&+ (\boldsymbol{\psi}_0^{(0)})^T\left(\sum_{r=0}^{k}\boldsymbol{\Psi}^r\right)\boldsymbol{\varsigma}^{(0)} + (\boldsymbol{\psi}_0^{(0)})^T\sum_{r=0}^{k-1}\left(\sum_{s=r+1}^{k}\boldsymbol{\Psi}^{s-1-r}\right)\boldsymbol{\omega}^{(0)} + \omega_0^{(0)}(k+1).
\end{aligned} \tag{91}$$

In the proof of Proposition 4.8 in Appendix H.2, we showed that when using a constant learning rate $\eta_i^{(0)}$ and $0 < \rho_\Psi < 1$, we get $\sum_{s=r+1}^{k}\boldsymbol{\Psi}^{s-1-r} \leq \frac{1}{1-\rho_\Psi}$. Using this result, and noting that we also similarly have $\sum_{r=0}^{k}\boldsymbol{\Psi}^r \leq \frac{1}{1-\rho_\Psi}$, we get

$$\boxed{\begin{aligned}
\mathbb{E}\left[F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(k+1)})\right] \leq F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}) &- \left(\gamma_0^{(0)} - \frac{(\boldsymbol{\psi}_0^{(0)})^T\boldsymbol{\gamma}^{(0)}}{1-\rho_\Psi}\right)\sum_{r=0}^{k}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\right\|^2\right] \\
&+ \frac{(\boldsymbol{\psi}_0^{(0)})^T\boldsymbol{\varsigma}^{(0)}}{1-\rho_\Psi} + \frac{(\boldsymbol{\psi}_0^{(0)})^T\boldsymbol{\omega}^{(0)}}{1-\rho_\Psi}(k+1) + \omega_0^{(0)}(k+1).
\end{aligned}} \tag{92}$$

The equation above is a compact reformulation of Eq. 36. Rearranging the above expression and plugging in the values of $\boldsymbol{\omega}^{(0)}$, $\boldsymbol{\psi}_0^{(0)}$ and $\omega_0^{(0)}$ from Lemmas D.1-D.3, we get

$$
\begin{aligned}
\frac{1}{k+1}\sum_{r=0}^{k}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(r)})\right\|^2\right] &\leq \frac{F(\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)})-F^\star}{q\max(\eta_i^{(0)})(k+1)} \\
&+\Bigg(\frac{\bar{L}\hat{\rho}_{0,A}}{\max(\eta_i^{(0)})}+3\left(\max(p_iL_i^2)+\max(\hat{\phi}_i)\hat{\rho}_{0,B}\kappa_1\right) \\
&\qquad +m\bar{L}\kappa_3\kappa_1\max(\eta_i^{(0)})\Bigg)\frac{\mathbb{E}\left[\left\|\boldsymbol{X}^{(0)}-\mathbf{1}\bar{\boldsymbol{x}}_{\hat{\boldsymbol{\phi}}}^{(0)}\right\|_{\hat{\boldsymbol{\phi}}}^2\right]}{2q(1-\rho_\Psi)(k+1)} \\
&+\left(3\max(\hat{\phi}_i)\tilde{\rho}_B+m\bar{L}\kappa_3\max(\eta_i^{(0)})\right)\frac{\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\hat{\boldsymbol{\pi}}}^{-1}\boldsymbol{Y}^{(0)}-m\mathbf{1}\bar{\boldsymbol{y}}^{(0)}\right\|_{\hat{\boldsymbol{\pi}}}^2\right]}{2mq(1-\rho_\Psi)(k+1)} \\
&+\frac{m^2\kappa_6\kappa_3}{2q(1-\rho_\Psi)}\Bigg[\bar{L}\hat{\rho}_{0,A}+3\left(\max(p_iL_i^2)+\max(\hat{\phi}_i)\hat{\rho}_{0,B}\kappa_1\right)\max(\eta_i^{(0)}) \\
&\qquad +m\bar{L}\kappa_3\kappa_1\max(\eta_i^{(0)})^2\Bigg]\left(\overline{\sigma_0^2p}+3\overline{(1-p)\delta_0^2}\right)\max(\eta_i^{(0)}) \\
&+\frac{\kappa_7}{q(1-\rho_\Psi)}\left(3\max(\hat{\phi}_i)\tilde{\rho}_B+m\bar{L}\kappa_3\max(\eta_i^{(0)})\right)\Bigg[\overline{\sigma_0^2p}+4\overline{p(1-p)\delta_0^2} \\
&\qquad +m\kappa_3\left(\kappa_6\frac{\kappa_4}{2}+8m\kappa_5\right)(\overline{\sigma_0^2p}+3\overline{(1-p)\delta_0^2})\max(\eta_i^{(0)})^2\Bigg] \\
&+\frac{\kappa_8\hat{\rho}_{0,B}}{2q(1-\rho_\Psi)}\left(3\max(\hat{\phi}_i)\tilde{\rho}_B+m\bar{L}\kappa_3\max(\eta_i^{(0)})\right)(\overline{\sigma_0^2p}+3\overline{(1-p)\delta_0^2}) \\
&+\frac{m}{2q}\left(3m\overline{\hat{\phi}\hat{\pi}(1-p)\delta_0^2}+3\max(\hat{\phi}_i)\hat{\rho}_{0,B}+m\bar{L}\kappa_3\max(\eta_i^{(0)})\right)\left(\overline{\sigma_0^2p}+3\overline{(1-p)\delta_0^2}\right).
\end{aligned}
\tag{93}
$$

where q was defined in Proposition 4.8. Setting the gradient estimation variance $\sigma_{0,i}^2=\sigma_{1,i}^2=0$ for the exact-gradient case concludes the proof.

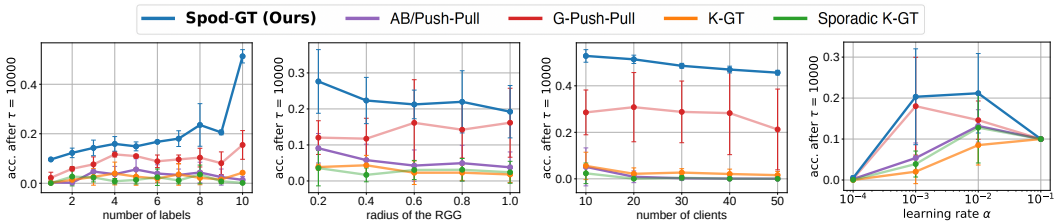
J Additional Experiments

J.1 Ablation Study with CIFAR-10

We carry out an additional ablation study of system parameters in our paper on the CIFAR-10 dataset, which is given in Fig. 4. In Fig. 4a, we vary the number of labels each client holds, with 1 corresponding to the extreme non-IID case, and 10 corresponding to the completely IID case. In Fig. 4b, we vary the radius of the RGG, with higher radius indicating a denser connectivity in the network. In Fig. 4c, we vary the number of clients m from 10 to 50 to validate that our methodology maintains its superiority for larger network sizes. Finally, in Fig. 4d, we do a hyperparameter analysis of the learning rate η from 10^{-4} to 10^{-1}, which shows that the choice of $\eta=0.01$ is the optimal choice for both datasets. In all of these studies, we observe that our `Spod-GT` algorithm is capable of consistently outperforming all baselines at different levels of data heterogeneity, confirming the benefit of integrating the notion of sporadicity in both communications and computations.

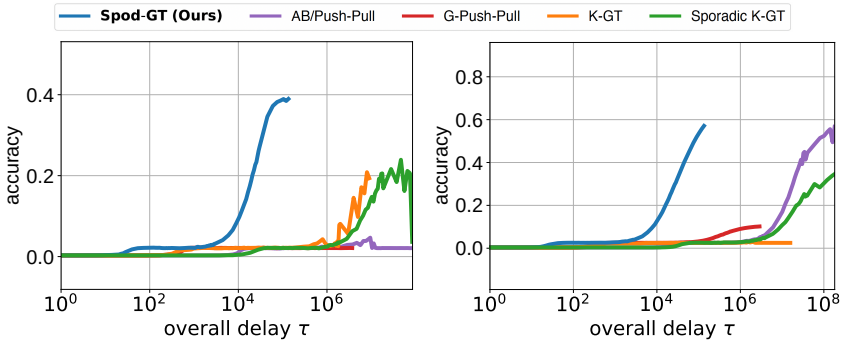
J.2 ConvNeXt Model on the Tiny-ImageNet Dataset

In this section, we conduct additional numerical experiments on the Tiny-ImageNet dataset. Fig. 5 illustrates the accuracy achieved per total delay for our `Spod-GT` algorithm and other GT baselines, noting that the x-axis is given in log-scale. We observe that when the clients in a DFL setup adopt our `Spod-GT` algorithm to train the ConvNeXt model on the Tiny-ImageNet dataset, they are able to consistently achieve higher accuracies faster compared to their non-sporadic counterparts. For example, if we compare all algorithms in Fig. 5b in terms of how long it takes them to achieve an accuracy of 40%, we could see a $100x$ speedup for our `Spod-GT` algorithm compared to the second best baseline.



(a) Varying number of labels per client. (b) Varying the radius of random geometric graph. (c) Varying the number of clients m in the network. (d) Varying the learning rate η for all clients.

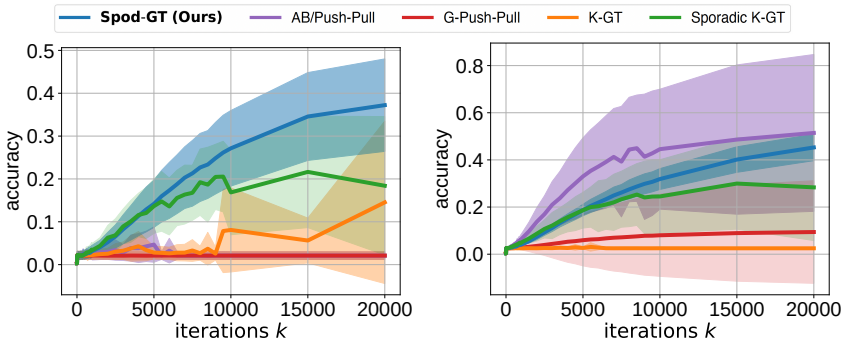
Figure 4: Effects of system parameters on CIFAR-10. The overall results confirm the advantage of `Spod-GT`.



(a) IID. (b) Non-IID.

Figure 5: Accuracy vs. latency plots for the ConvNeXt model on the Tiny-ImageNet dataset. `Spod-GT` achieves the target accuracy much faster with less delay, emphasizing the benefit of sporadicity in DFL over directed graphs for GT iterations and aggregations simultaneously.

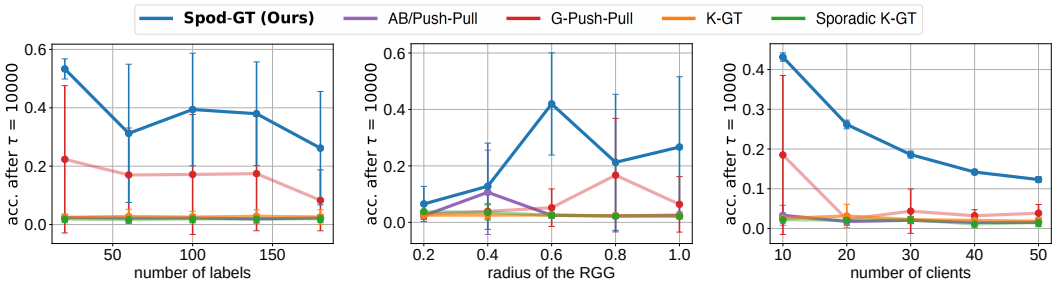
In Fig. 6, we change the x-axis of the same accuracy plots in Fig. 5 on the Tiny-ImageNet dataset to iterations. This helps us gain more insight into the training of `Spod-GT` and other baselines in terms of accuracy per pure iteration, which fails to capture the total time it takes to finish each iteration due to the clients having heterogeneous resources. We observe that different algorithms achieve various convergence rates even when resource heterogeneities are not accounted for, i.e., sporadicity in gradient tracking and aggregation operations affect the convergence speed of DFL algorithms on directed network graphs. However, our `Spod-GT` is not designed to outperform other DFL algorithms in terms of accuracy per pure iterations, although this remarkably occurs in Fig. 6a, and it is designed to outperform other baselines when compared in terms of overall actual delay as shown in Fig. 5.



(a) IID. (b) Non-IID.

Figure 6: Accuracy vs. iteration plots for the ConvNeXt model on the Tiny-ImageNet dataset.

We further carry out an ablation study of system parameters in our paper on the Tiny-ImageNet dataset, which is given in Fig. 7. In Fig. 7a, we vary the number of labels each client holds, with 20 corresponding to the extreme non-IID case, and 200 corresponding to the completely IID case. In Fig. 7b, we vary the radius of the RGG, with higher radius indicating a denser connectivity in the network. In Fig. 7c, we vary the number of clients m from 10 to 50 to validate that our methodology maintains its superiority for larger network sizes. In all of these studies, we observe that our `Spod-GT` algorithm is capable of consistently outperforming all baselines at different levels of data heterogeneity, confirming the benefit of integrating the notion of sporadicity in both communications and computations even when using larger models and datasets such as ConvNeXT on Tiny-ImageNet.



(a) Varying number of labels per client.

(b) Varying the radius of random geometric graph.

(c) Varying the number of clients m in the network.

Figure 7: Effects of system parameters on Tiny-ImageNet. The overall results confirm the advantage of `Spod-GT`.