

A Unified Convergence Analysis for Semi-Decentralized Learning: Sampled-to-Sampled vs. Sampled-to-All Communication

Angelo Rodio¹, Giovanni Neglia², Zheng Chen¹, Erik G. Larsson¹

¹Department of Electrical Engineering, Linköping University, Sweden

Centre Inria d’Université Côte d’Azur, France

{angelo.rodio, zheng.chen, erik.g.larsson}@liu.se, giovanni.neglia@inria.fr

Abstract

In semi-decentralized federated learning, devices primarily rely on device-to-device communication but occasionally interact with a central server. Periodically, a sampled subset of devices uploads their local models to the server, which computes an aggregate model. The server can then either (i) share this aggregate model only with the sampled clients (sampled-to-sampled, S2S) or (ii) broadcast it to all clients (sampled-to-all, S2A). Despite their practical significance, a rigorous theoretical and empirical comparison of these two strategies remains absent. We address this gap by analyzing S2S and S2A within a unified convergence framework that accounts for key system parameters: sampling rate, server aggregation frequency, and network connectivity. Our results—both analytical and experimental—reveal distinct regimes where one strategy outperforms the other, depending primarily on the degree of data heterogeneity across devices. These insights lead to concrete design guidelines for practical semi-decentralized FL deployments.

Code — <https://github.com/arodio/SemiDec>

1 Introduction

The performance of large-scale machine learning models depends critically on the volume and diversity of data; however, in many practical scenarios, training data are decentralized, generated by edge devices such as smartphones or sensors (McMahan et al. 2017; Kairouz et al. 2021). Centralizing these data is often prohibitively expensive—or even infeasible—due to network limitations and privacy constraints (Bonawitz et al. 2019; Li et al. 2020a).

Federated learning (FL) is a distributed machine learning paradigm in which multiple devices cooperate to learn a global model under the orchestration of a central server without sharing their data (McMahan et al. 2017). Device-to-server (D2S) communication is typically expensive in FL, especially when the central server is located in a wide-area network, where limited uplink bandwidth dominates both communication latency and energy consumption. The de facto optimization method, local stochastic gradient descent (SGD) (Konečný et al. 2017; McMahan et al. 2017), addresses this communication bottleneck by enabling devices

to perform multiple local updates before server aggregation. This simple trick reduces the number of D2S communications but has a well-known drawback: multiple local SGD steps on non-identically distributed (non-IID) data lead to local over-fitting (known as model drift) and hinder convergence (Karimireddy et al. 2020; Li et al. 2020b).

Fully-decentralized learning eliminates the central server and relies solely on device-to-device (D2D) communications, where devices average their local models with those of their neighbors after each SGD update (Lian et al. 2017; Koloskova et al. 2020). These exchanges are typically inexpensive, leveraging high-bandwidth local-area networks or direct short-range wireless links. The convergence of such algorithms depends on the connectivity of the underlying communication graph. Intuitively, sparse connectivity slows convergence—a phenomenon rigorously analyzed in prior work (Yuan, Ling, and Yin 2016; Neglia et al. 2020; Le Bars et al. 2023; Larsson and Michelusi 2025). More critically, convergence is impossible when the graph is disconnected, as information cannot propagate between different graph components.

Semi-decentralized learning interleaves D2D consensus rounds within components with periodic communication between a sampled subset of devices and a central server, which aggregates their models (Chen et al. 2021; Lin et al. 2021). This hybrid design leverages the hierarchical structure of modern networks: frequent, low-cost D2D exchanges foster local consensus within components, while periodic D2S rounds ensure information sharing across components and enable global convergence. Once the server aggregates the models of the sampled devices, two communication primitives have been proposed in the literature:

- (i) *Sampled-to-Sampled* (S2S): the server sends the aggregate model *only* to the sampled devices, while the remaining devices retain their current local models (Chen, Wang, and Brinton 2024);
- (ii) *Sampled-to-All* (S2A): the server broadcasts the aggregate model to *all* devices, which then replace their current models (Chen et al. 2021; Lin et al. 2021; Guo et al. 2021).

While both variants appear in prior work, their relative merits have not been thoroughly investigated. Intuitively, S2A may spread information faster because the aggregated model is immediately disseminated to all clients. However, this

comes at the cost of introducing a bias: the sampled clients exert a disproportionate influence, as their models overwrite information from unsampled ones. In this work, we address this gap through a unified theoretical analysis and extensive experimental comparison of the two strategies.

Our contributions.

- We develop a unified theoretical framework that captures (i) intra- and inter-component statistical heterogeneity, (ii) the sampling rate, (iii) the server aggregation period, and (iv) the D2D network connectivity. Our analysis reveals a fundamental trade-off. S2A introduces a broadcast-induced bias due to the shift in the global average model after each D2S aggregation but reduces disagreement error by periodically realigning all local models. Conversely, S2S avoids this bias but suffers from greater disagreement, as non-sampled models remain misaligned after aggregation.
- By comparing convergence bounds, we identify regimes in which one communication primitive outperforms the other. Specifically, S2A converges faster when both intra- and inter-component heterogeneity are low, while S2S outperforms as inter-component heterogeneity increases—particularly at low sampling rates, short server periods, or sparse network connectivity.
- Simulations on benchmark FL datasets across varying sampling rates, aggregation periods, and network topologies confirm these regimes and highlight the importance of selecting the appropriate communication primitive. These insights translate into practical guidelines for configuring semi-decentralized FL deployments.

2 Related Work

The cost of device-to-server communication in FL has been extensively investigated (Shamir, Srebro, and Zhang 2014; Alistarh et al. 2017; Horváth et al. 2022). Both classical (Stich 2018; Reddi et al. 2021) and more refined (Mishchenko et al. 2022) analyses of local SGD establish a fundamental trade-off: a moderate number of local steps reduces wall-clock time, whereas too many local updates on non-IID data induce *model drift* and hinder convergence. Advanced algorithms, e.g., relying on control variates (Karimireddy et al. 2020) and proximal corrections (Mishchenko et al. 2022), mitigate this drift at additional computational or communication cost. However, the qualitative conclusion remains: an excessive number of local SGD steps under high statistical heterogeneity slows global progress.

In fully-decentralized SGD (D-SGD), which relies solely on D2D communications, the convergence rate is governed by the spectral gap of the doubly stochastic mixing matrix W . Specifically, the iteration complexity scales inversely with $\gamma := 1 - \lambda_2(W^\top W)$, where λ_2 denotes the second-largest eigenvalue of $W^\top W$ (Nedić and Olshevsky 2016; Yuan, Ling, and Yin 2016; Koloskova et al. 2020; Le Bars et al. 2023). Convergence becomes slower as γ approaches zero, and for $\gamma = 0$ (disconnected graph), D-SGD fails to reach the global optimum, as each connected component converges to its *local* minimizer.

Hierarchical FL assumes a multi-tier tree topology (cloud-edge-device) and aggregates along the hierarchy (Wang et al. 2021); semi-decentralized FL supports *arbitrary* D2D topologies (Chen et al. 2021; Lin et al. 2021). Prior work has analyzed the S2S and S2A primitives under convex objectives (Lin et al. 2021; Chen, Wang, and Brinton 2024) and, for S2A, also under non-convex objectives (Guo et al. 2021), but assuming that at least one device per connected component is sampled in every server round. This assumption implicitly requires the server to know the component membership of each device—a requirement that is difficult to satisfy in practice due to the large number of devices, their mobility (resulting in time-varying communication graphs), and privacy constraints (as it may indirectly reveal user locations). To the best of our knowledge, a convergence analysis of the S2S primitive is still lacking for non-convex objectives, and there is no systematic theoretical or empirical comparison of S2S and S2A within a unified framework.

Our analysis tackles the following technical challenges:

- (i) The broadcast-induced bias error in S2A, defined as the change in the average model before and after a D2S communication, and the disagreement error in S2S, measuring the divergence of the local models from the global average, scale *differently* with stepsize, sampling rate, server period, and network connectivity, making their comparison non-trivial.
- (ii) The S2A update rule can be modeled as a rank-one, column-stochastic but not row-stochastic averaging operator; thus, standard spectral-gap arguments for doubly stochastic W matrices in D-SGD analyses (e.g., Koloskova et al. (2020)) are not applicable.
- (iii) Although the S2S update rule involves a symmetric, stochastic weight matrix—formally compatible with the assumptions in Koloskova et al. (2020)—their analysis fails to capture the fundamental asymmetry between D2D and D2S rounds, where inter-component statistical heterogeneity is reduced *only* through server aggregation. This distinction is crucial for the comparison of S2S and S2A and motivates our analysis.

We address these challenges by (a) characterizing bias and disagreement errors through the properties of the S2S and S2A operators, (b) introducing an orthogonal decomposition of the total disagreement into intra- and inter-component terms, and (c) capturing the distinct effects of D2D and D2S communication on intra- versus inter-component heterogeneity.

3 Problem Setting

Network model. We consider a network consisting of a central server and n devices, organized in C disjoint components (or clusters). Each component $c \in \{1, \dots, C\}$ is modeled as an undirected, connected, and time-varying graph $\mathcal{G}_c^{(t)} = (\mathcal{V}_c, \mathcal{E}_c^{(t)})$, where \mathcal{V}_c denotes the set of $n_c := |\mathcal{V}_c|$ devices in component c , and $(i, j) \in \mathcal{E}_c^{(t)}$ indicates that devices $i, j \in \mathcal{V}_c$ communicate via D2D links at round t . In addition, each device can communicate with the central server through D2S links. The overall network at round t

is modeled as $\mathcal{G}^{(t)} = (\mathcal{V}, \mathcal{E}^{(t)})$, where $\mathcal{V} := \bigcup_{c=1}^C \mathcal{V}_c$ and $\mathcal{E}^{(t)} := \bigcup_{c=1}^C \mathcal{E}_c^{(t)}$.

Learning task. We consider an FL system where the central server and the devices collaborate to learn the parameters $\mathbf{x} \in \mathbb{R}^d$ of a machine learning model, where $d \in \mathbb{N}$ is the model dimension. Each device $i \in \mathcal{V}$ has a local dataset \mathcal{D}_i of data samples $\xi \in \mathcal{D}_i$. We denote by $F_i(\mathbf{x}; \xi)$ the loss incurred by the model with parameters \mathbf{x} on data sample ξ . The goal is to solve an optimization problem of the form:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}), \quad (1)$$

where $f_i(\mathbf{x}) := \frac{1}{|\mathcal{D}_i|} \sum_{\xi \in \mathcal{D}_i} F_i(\mathbf{x}, \xi)$ is the local objective of device $i \in \mathcal{V}$.

Notation. All vectors are by default column vectors. $\mathbf{0}$ and $\mathbf{1}$ denote the all-zeros and all-ones vectors of appropriate dimension. I is the identity matrix. The global averaging projector is $\Pi := \frac{1}{n} \mathbf{1} \mathbf{1}^\top$. Given n vectors $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, we write their average as $\bar{\mathbf{x}} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \in \mathbb{R}^d$. We stack the n vectors as columns in the matrix $X := [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, such that right-multiplication by Π performs column averaging: $\bar{X} := X\Pi = [\bar{\mathbf{x}}, \dots, \bar{\mathbf{x}}] \in \mathbb{R}^{d \times n}$. We use $\|\cdot\|_2$ for both the Euclidean norm of a vector and the spectral norm of a matrix, and $\|\cdot\|_F$ for the Frobenius norm.

4 Two Communication Primitives for Semi-Decentralized FL

We study two semi-decentralized learning primitives, summarized in Algorithm 1, for solving Problem (1). The training proceeds over T communication rounds, where each round $t \in \{0, \dots, T-1\}$ consists of two or three steps:

- (i) *Local stochastic descent.* Each device $i \in \mathcal{V}$ updates its local model $\mathbf{x}_i^{(t)}$ by one local SGD step:

$$\mathbf{x}_i^{(t+1/3)} = \mathbf{x}_i^{(t)} - \eta_t \nabla F_i(\mathbf{x}_i^{(t)}, \mathcal{B}_i^{(t)}), \quad (2)$$

where η_t is the stepsize, $\mathcal{B}_i^{(t)}$ is a mini-batch sampled from the local dataset \mathcal{D}_i , and $\nabla F_i(\mathbf{x}_i^{(t)}, \mathcal{B}_i^{(t)})$ is an unbiased estimate of $\nabla F_i(\mathbf{x}_i^{(t)})$.

- (ii) *Device-to-device (D2D) mixing.* Each device $i \in \mathcal{V}$ averages its local model $\mathbf{x}_i^{(t+1/3)}$ with neighbors via mixing weight $w_{ji}^{(t)}$, where $w_{ji}^{(t)} > 0$ iff $(j, i) \in \mathcal{E}^{(t)}$:

$$\mathbf{x}_i^{(t+2/3)} = \sum_{j=1}^n w_{ji}^{(t)} \mathbf{x}_j^{(t+1/3)}. \quad (3)$$

In fully-decentralized rounds, $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t+2/3)}$.

- (iii) *Device-to-server (D2S) aggregation.* Every H rounds, the server samples a subset $\mathcal{S}^{(t)} \subseteq \mathcal{V}$ of K devices uniformly at random without replacement and averages their local models:

$$\hat{\mathbf{x}}^{(t+1)} = \frac{1}{|\mathcal{S}^{(t)}|} \sum_{i \in \mathcal{S}^{(t)}} \mathbf{x}_i^{(t+2/3)}. \quad (4)$$

The dissemination of this aggregate from the server to the devices can follow two distinct communication primitives: Sampled-to-Sampled (S2S) and Sampled-to-All (S2A).

Algorithm 1: Semi-Decentralized Federated Learning

Input: $X^{(0)} \in \mathbb{R}^{d \times n}$, rounds T , period H , stepsizes $\{\eta_t\}$, mixing matrices $W^{(t)} \sim \mathcal{W}$

```

1: for  $t = 0, \dots, T-1$  do
2:    $X^{(t+1/3)} \leftarrow X^{(t)} - \eta_t \nabla F(X^{(t)}, \mathcal{B}^{(t)})$ 
3:    $X^{(t+2/3)} \leftarrow X^{(t+1/3)} W^{(t)}$ 
4:   if  $t \equiv 0 \pmod{H}$  then
5:     sample  $\mathcal{S}^{(t)} \subseteq \mathcal{V}$ ,  $|\mathcal{S}^{(t)}| = K$ 
6:     build  $W_{\text{S2S/A}}^{(t)}$  by Eq. (5) (S2S) or Eq. (6) (S2A)
7:      $X^{(t+1)} \leftarrow X^{(t+2/3)} W_{\text{S2S/A}}^{(t)}$ 
8:   else
9:      $X^{(t+1)} \leftarrow X^{(t+2/3)}$ 
10:  return  $X^{(T)}$ 

```

Sampled-to-Sampled (S2S). The server transmits the aggregate model *only* to the sampled devices: $\mathbf{x}_i^{(t+1)} = \hat{\mathbf{x}}^{(t+1)}$, $i \in \mathcal{S}^{(t)}$; the other devices retain their local model. The evolution of the local models can be represented as the matrix multiplication $X^{(t+1)} = X^{(t+2/3)} W_{\text{S2S}}^{(t)}$, where:

$$(W_{\text{S2S}}^{(t)})_{ij} = \begin{cases} \frac{1}{K}, & i, j \in \mathcal{S}^{(t)}; \\ 1, & i = j \notin \mathcal{S}^{(t)}; \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Sampled-to-All (S2A). The server broadcasts $\mathbf{x}^{(t+1)}$ to *all* devices: $\mathbf{x}_i^{(t+1)} = \hat{\mathbf{x}}^{(t+1)}$ for all $i \in \mathcal{V}$. As above, this can be represented as $X^{(t+1)} = X^{(t+2/3)} W_{\text{S2A}}^{(t)}$, where:

$$(W_{\text{S2A}}^{(t)})_{ij} = \begin{cases} \frac{1}{K}, & i \in \mathcal{S}^{(t)}; \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

4.1 High-level Comparison of S2S and S2A

We identify the following two errors after the D2S round:

- (i) the *bias error*, which quantifies the change in the global average model induced by the D2S step, defined as $\mathbb{E}[\|\bar{X}^{(t+1)} - \bar{X}^{(t+2/3)}\|_F^2]$;
- (ii) the *disagreement error*, which quantifies the divergence of the local models from the global average model, defined as $\mathbb{E}[\|X^{(t+1)} - \bar{X}^{(t+1)}\|_F^2]$.

The two primitives, S2S and S2A, exhibit opposite error behaviors: S2S preserves the global average (zero bias) but leaves residual disagreement, whereas S2A enforces perfect consensus (zero disagreement) at the cost of a non-zero bias error.

For S2S, the matrix W_{S2S} is symmetric and doubly stochastic, satisfying $W_{\text{S2S}} \Pi = \Pi W_{\text{S2S}} = \Pi$.

Therefore, the bias error vanishes since:

$$\bar{X}^{(t+1)} = X^{(t+2/3)} W_{\text{S2S}} \Pi = X^{(t+2/3)} \Pi = \bar{X}^{(t+2/3)}. \quad (7)$$

However, non-sampled devices are not updated with the server aggregate, resulting in residual disagreement:

$$X^{(t+1)} = X^{(t+2/3)} W_{\text{S2S}} \neq X^{(t+2/3)} \Pi = \bar{X}^{(t+1)}, \quad (8)$$

with magnitude (Lemma 8, supplementary material):

$$\mathbb{E}[\|X^{(t+1)} - \bar{X}^{(t+1)}\|_F^2] = \frac{n-K}{n-1} \mathbb{E}\|X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})}\|_F^2, \quad (9)$$

where $\mathbb{E}\|X^{(t+2/3)} - \bar{X}^{(t+2/3)}\|_F^2$ denotes the disagreement inherited from the D2D step at time $t + 2/3$.

Conversely, W_{S2A} is column-stochastic but *not* row-stochastic, with $\Pi W_{S2A} = \Pi$ and $W_{S2A}\Pi = W_{S2A} \neq \Pi$. This property eliminates disagreement since:

$$X^{(t+1)} - \bar{X}^{(t+1)} = X^{(t+2/3)}(W_{S2A} - W_{S2A}\Pi) = 0, \quad (10)$$

but introduces the broadcast-induced bias:

$$\bar{X}^{(t+1)} = X^{(t+2/3)}W_{S2A}\Pi \neq X^{(t+2/3)}\Pi = \bar{X}^{(t+2/3)}, \quad (11)$$

with magnitude (Lemma 11, supplementary material):

$$\mathbb{E}[\|\bar{X}^{(t+1)} - \bar{X}^{(t+2/3)}\|_F^2] = \frac{n-K}{K(n-1)} \mathbb{E}\|X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})}\|_F^2. \quad (12)$$

Although the bias factor in Eq. (12) might appear smaller than the disagreement factor in Eq. (9), the two equations describe different error sources, which propagate under different scalings with respect to stepsize, sampling rate, server period, and network connectivity. This interplay makes the comparison between S2S and S2A non-trivial and motivates our subsequent unified convergence analysis.

5 Unified Convergence Analysis

Our framework extends the convergence theory of decentralized optimization (Koloskova et al. 2020; Le Bars et al. 2023) to semi-decentralized federated learning, and provides the first theoretical comparison of S2S and S2A.

All theoretical results assume Lipschitz continuity of the stochastic gradients (Nguyen et al. 2019).

Assumption 1 (L-smoothness). *For every $i \in \mathcal{V}$ and every $\xi \sim \mathcal{D}_i$, the stochastic loss $F_i(\cdot, \xi)$ is L-smooth; i.e., there exists $L > 0$ such that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$\|\nabla F_i(\mathbf{x}, \xi) - \nabla F_i(\mathbf{y}, \xi)\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2. \quad (13)$$

For convex results, we additionally invoke convexity of the local objectives (Bubeck 2015).

Assumption 2 (Convexity). *Each $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex:*

$$f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \quad (14)$$

To keep the analysis unified across convex and non-convex settings, we assume that the stochastic variance is uniformly bounded in \mathbf{x} (Le Bars et al. 2023), although in the convex case it would suffice to bound it only at the optimum.

Assumption 3 (Bounded stochastic variance). *For every $i \in \mathcal{V}$, there exists a constant $\bar{\sigma}^2 > 0$ such that, for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbb{E}_{\xi \sim \mathcal{D}_i} [\|\nabla F_i(\mathbf{x}, \xi) - \nabla f_i(\mathbf{x})\|_2^2] \leq \bar{\sigma}^2. \quad (15)$$

For clarity of exposition, our analysis assumes a fixed deterministic mixing matrix W . However, all results extend to dynamic D2D communication graphs (Koloskova et al. 2020), which are represented by time-varying mixing matrices (see Appendix D for details).

Assumption 4 (Mixing matrix (Koloskova et al. 2020; Le Bars et al. 2023)). *The mixing matrix W is doubly stochastic, i.e., $W \in [0, 1]^{n \times n}$, $W\mathbf{1} = \mathbf{1}$, and $\mathbf{1}^\top W = \mathbf{1}^\top$.*

The matrix W is block diagonal, reflecting the C disconnected components of the communication graph \mathcal{G} . Each diagonal block $W_c := W[\mathcal{V}_c, \mathcal{V}_c] \in \mathbb{R}^{n_c \times n_c}$ corresponds to the D2D mixing matrix of component $c \in \{1, \dots, C\}$. To decompose disagreement within and across components, we define the component projector $\Pi_C \in \mathbb{R}^{n \times n}$ as:

$$(\Pi_C)_{ij} = \begin{cases} \frac{1}{n_c}, & i, j \in \mathcal{V}_c; \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

The operators $I - \Pi_C$ and $\Pi_C - \Pi$ enable an orthogonal decomposition of the global disagreement at any time t into intra-component and inter-component terms.

Lemma 1 (Orthogonal decomposition). *For any $X \in \mathbb{R}^{d \times n}$,*

$$\|X(I - \Pi)\|_F^2 = \|X(I - \Pi_C)\|_F^2 + \|X(\Pi_C - \Pi)\|_F^2. \quad (17)$$

Only the intra-component disagreement is reduced by D2D consensus steps, while the inter-component term requires periodic D2S aggregation.

Lemma 2 (Intra-component mixing parameter). *There exists a constant $p \in (0, 1]$ such that, for all $X \in \mathbb{R}^{d \times n}$,*

$$\|X(W - \Pi_C)\|_F^2 \leq (1-p)\|X(I - \Pi_C)\|_F^2. \quad (18)$$

For a fixed W , Lemma 2 holds with $p = \frac{\sum_{c=1}^C p_c(n_c-1)}{\sum_{c=1}^C (n_c-1)}$, where $p_c = 1 - \lambda_2(W_c^\top W_c)$ (see Appendix D for details). For Metropolis-Hastings weights $w_{ij} = w_{ji} = \min\{1/(\deg(i)+1), 1/(\deg(j)+1)\}$, we have $p_c = 1$ for complete graphs, $p_c = \Theta(n_c^{-1})$ for 2D grid topologies, and $p_c = \Theta(n_c^{-2})$ for ring graphs (Xiao and Boyd 2004; Boyd et al. 2006).

A key step in our analysis is to disentangle heterogeneity within components from heterogeneity across components: this distinction is crucial for comparing S2S and S2A.

Assumption 5 (Intra- and inter-component heterogeneity). *There exist $\bar{\zeta}_{\text{intra}}^2, \bar{\zeta}_{\text{inter}}^2 > 0$ such that, for all $\mathbf{x} \in \mathbb{R}^d$:*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}_\xi \|\sum_{j=1}^n (W - \Pi_C)_{ij} \nabla F_j(\mathbf{x}, \xi)\|_2^2 \leq \bar{\zeta}_{\text{intra}}^2, \quad (19)$$

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}_\xi \|\sum_{j=1}^n (\Pi_C - \Pi)_{ij} \nabla F_j(\mathbf{x}, \xi)\|_2^2 \leq \bar{\zeta}_{\text{inter}}^2. \quad (20)$$

The constants $\bar{\zeta}_{\text{intra}}$ and $\bar{\zeta}_{\text{inter}}$ quantify intra- and inter-component noise arising from both stochastic variance and statistical heterogeneity. We treat these two sources of noise jointly: our intra-component bound (Eq. 19) generalizes the neighborhood heterogeneity of (Le Bars et al. 2023)—defined as the deviation between the W -weighted neighborhood gradients and their intra-component average—and is weaker than Assumption 4 in (Guo et al. 2021).

5.1 Main Results

We are now ready to present our main convergence results; all proofs are deferred to the supplementary material.

Theorem 1 (Sampled-to-Sampled). *Under Assumptions 1–5, there exists a constant stepsize $\eta \leq \frac{p}{8L}$ such that, for any*

target accuracy $\epsilon > 0$, Algorithm 1 (S2S) achieves

Convex: $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} (f(\bar{\mathbf{x}}^{(t)}) - f^*) \leq \epsilon$ after:

$$T \geq \mathcal{O} \left(\frac{\bar{\sigma}^2}{n\epsilon^2} + \sqrt{\frac{n-1}{K-1}} \frac{\sqrt{L}\bar{\zeta}_{\text{intra}}}{p\epsilon^{3/2}} + \frac{n-1}{K-1} \frac{\sqrt{L}H\bar{\zeta}_{\text{inter}}}{\epsilon^{3/2}} + \frac{L}{p\epsilon} \right) R_0^2, \quad (21)$$

Non-Convex: $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 \leq \epsilon$ after:

$$T \geq \mathcal{O} \left(\frac{\bar{\sigma}^2}{n\epsilon^2} + \sqrt{\frac{n-1}{K-1}} \frac{\bar{\zeta}_{\text{intra}}}{p\epsilon^{3/2}} + \frac{n-1}{K-1} \frac{H\bar{\zeta}_{\text{inter}}}{\epsilon^{3/2}} + \frac{1}{p\epsilon} \right) Lf_0, \quad (22)$$

where $R_0 := \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2$ and $f_0 := f(\mathbf{x}^{(0)}) - f^*$ denote the initial errors, and $\mathcal{O}(\cdot)$ hides the numerical constants explicitly provided in the supplementary material.

Theorem 2 (Sampled-to-All). Under Assumptions 1–5, there exists a constant stepsize $\eta \leq \frac{p}{8L}$ such that, for any target accuracy $\epsilon > 0$, Algorithm 1 (S2A) achieves

Convex: $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} (f(\bar{\mathbf{x}}^{(t)}) - f^*) \leq \epsilon$ after:

$$T \geq \mathcal{O} \left(\frac{\bar{\sigma}^2}{n\epsilon^2} + \frac{n-K}{K(n-1)} \frac{\bar{\zeta}_{\text{intra}}^2}{H p^2 \epsilon^2} + \frac{n-K}{K(n-1)} \frac{H\bar{\zeta}_{\text{inter}}^2}{\epsilon^2} + \frac{\sqrt{L}\bar{\zeta}_{\text{intra}}}{p\epsilon^{3/2}} + \frac{\sqrt{L}H\bar{\zeta}_{\text{inter}}}{\epsilon^{3/2}} + \frac{L}{p\epsilon} \right) R_0^2, \quad (23)$$

Non-Convex: $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 \leq \epsilon$ after:

$$T \geq \mathcal{O} \left(\frac{\bar{\sigma}^2}{n\epsilon^2} + \frac{n-K}{K(n-1)} \frac{\bar{\zeta}_{\text{intra}}^2}{H p^2 \epsilon^2} + \frac{n-K}{K(n-1)} \frac{H\bar{\zeta}_{\text{inter}}^2}{\epsilon^2} + \frac{\bar{\zeta}_{\text{intra}}}{p\epsilon^{3/2}} + \frac{H\bar{\zeta}_{\text{inter}}}{\epsilon^{3/2}} + \frac{1}{p\epsilon} \right) Lf_0, \quad (24)$$

where $R_0 = \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2$ and $f_0 := f(\mathbf{x}^{(0)}) - f^*$ denote the initial errors, and $\mathcal{O}(\cdot)$ hides the numerical constants explicitly provided in the supplementary material.

5.2 Discussion

We compare S2S and S2A under the convergence bounds of Theorems 1–2. Overall, S2S achieves a faster convergence than S2A: neglecting common factors, the dominant error terms scale as $\mathcal{O}(\epsilon^{-3/2})$ in Eqs. (21)–(22), as compared to $\mathcal{O}(\epsilon^{-2})$ in Eqs. (23)–(24). The slower convergence of S2A is primarily due to the broadcast-induced bias error discussed in Section 4.1. Moreover, S2A incurs an *extra quadratic* dependence on the intra- and inter-component heterogeneity terms, $\bar{\zeta}_{\text{intra}}$ and $\bar{\zeta}_{\text{inter}}$, which can dominate the bounds in Eqs. (23)–(24) under statistically diverse data distributions.

Effect of sampling rate (K/n). The number of sampled devices K affects both heterogeneity terms, $\bar{\zeta}_{\text{intra}}$ and $\bar{\zeta}_{\text{inter}}$, with different multiplicative factors for S2S and S2A. Two limiting cases are noteworthy:

- When *all* devices are sampled ($K = n$), the two update rules coincide ($W_{\text{S2A}} = W_{\text{S2S}} = \Pi$), and the two algorithms share the same convergence rate.

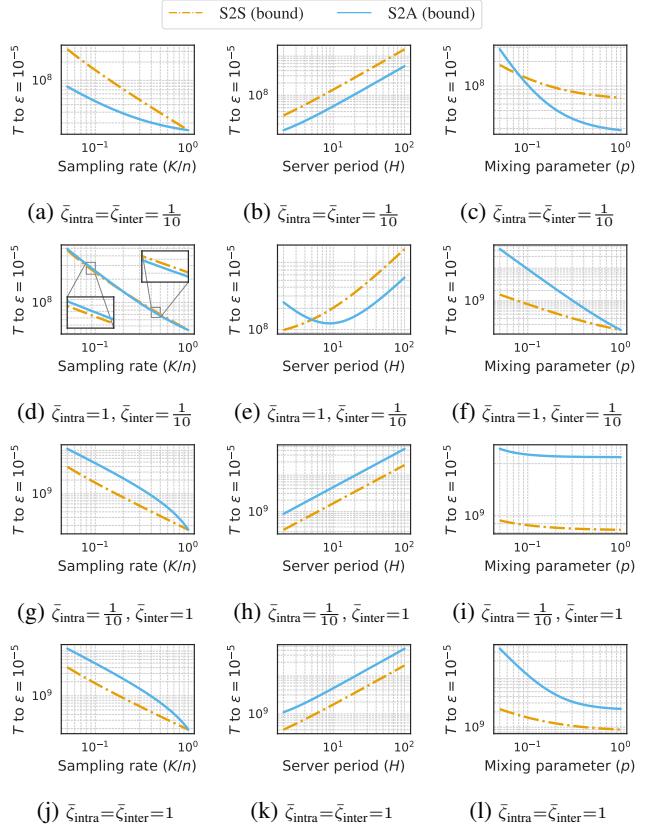


Figure 1: Convergence rates for S2S and S2A, comparing Eqs. (22)–(24) with $n=100$, $L=f_0=1$, $\bar{\sigma}=0$. Left column: Sampling rate (K/n) with $H=5$, $p=1$. Center column: Server period (H) with $K/n=0.2$, $p=1$. Right column: Mixing parameter (p) with $K/n=0.2$, $H=5$.

- When only *one* device is sampled ($K = 1$), $W_{\text{S2S}} = I$, S2S is unable to mix the sampled model across components, and the bounds in Eqs. (21)–(22) diverge. In contrast, S2A still broadcasts the (single) sampled model to all devices and thus converges, albeit at a slower rate.

Effect of server period (H). All $\bar{\zeta}_{\text{inter}}$ terms are penalized by a factor H in both bounds, reflecting the fact that only D2S rounds mitigate inter-component heterogeneity. For $H \rightarrow \infty$, both bounds diverge, as each components may reach consensus to their local optima, but no convergence to the global optimum can be guaranteed. Nonetheless, S2A grows quadratically in $\bar{\zeta}_{\text{inter}}$, whereas S2S grows linearly.

Effect of mixing parameter (p). All $\bar{\zeta}_{\text{intra}}$ terms are multiplied by the inverse of the mixing parameter p , as D2D rounds can only mitigate intra-component heterogeneity.

5.3 Theoretical Heterogeneity Regimes

To better interpret Theorems 1–2, Figure 1 shows the right-hand sides of Eqs. (22) and (24), comparing the number of rounds T required to achieve the target accuracy $\epsilon = 10^{-5}$ as a function of the sampling rate (left column), server

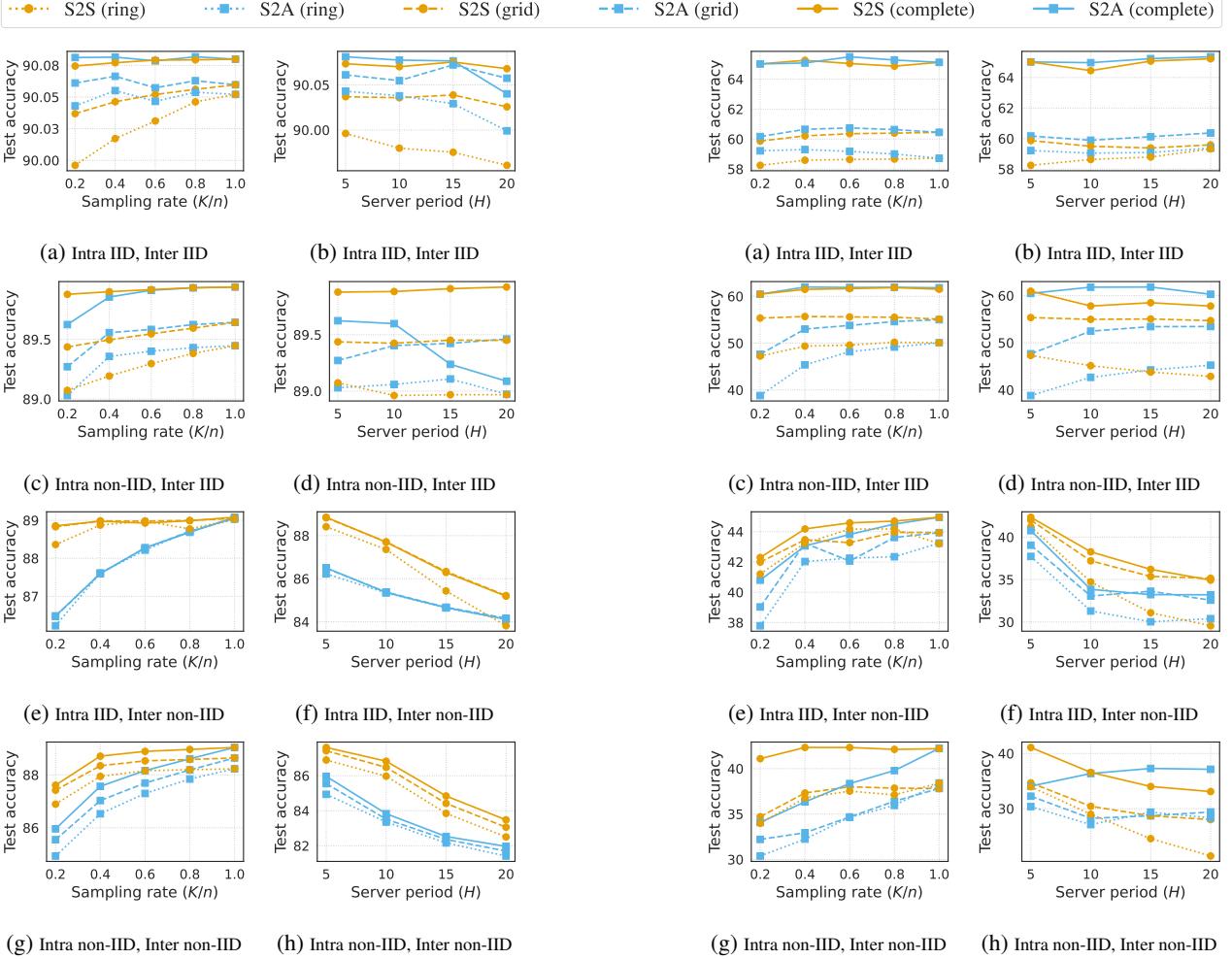


Figure 2: Test accuracy on MNIST dataset. Left column: Sampling rate (K/n) with $H = 5$. Right column: Server period (H) with $K/n = 0.2$.

period (center column), and mixing parameter (right column). We consider $n = 100$ devices, and set the parameters $L = f_0 = 1$ and $\bar{\sigma} = 0$ (as they are common to both S2S and S2A, their choice does not influence the comparison).

We identify three main qualitative regimes:

- R1.** $\zeta_{\text{intra}}, \zeta_{\text{inter}}$ are low: S2A converges faster than S2S for most sampling rates (Fig. 1(a)), server periods (Fig. 1(b)), and mixing parameters (Fig. 1(c)).
- R2.** $\zeta_{\text{inter}} \ll \zeta_{\text{intra}}$: S2S converges slightly faster for low sampling rates, low server periods, and for most mixing parameters ($K/n < 0.2$, $H < 5$, and $p < 1$); S2A converges slightly faster otherwise (Figs. 1(d,e,f)).
- R3.** ζ_{inter} is high: S2S converges faster for most values of K/n , H , and p , irrespective of ζ_{intra} (Figs. 1(g–l)).

6 Experimental Results

We simulate a semi-decentralized FL system consisting of a central server and $n = 100$ devices partitioned into

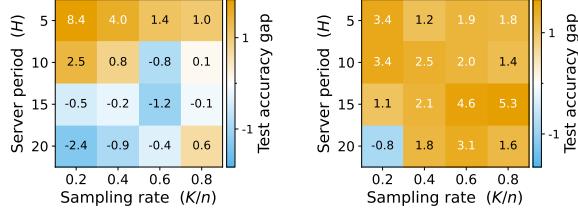
Figure 3: Test accuracy on CIFAR-10 dataset. Left column: Sampling rate (K/n) with $H = 5$. Right column: Server period (H) with $K/n = 0.2$.

$C = 2$ equal-sized components ($n_1 = n_2 = 50$). For the D2S communication network, we vary the sampling rate $K/n \in \{0.2, 0.4, 0.6, 0.8, 1\}$ and the aggregation period $H \in \{5, 10, 15, 20\}$. For the D2D communication graph, we consider three representative topologies: ring, grid, and complete graph, with Metropolis-Hastings mixing weights.

We benchmark our comparison on two image-classification tasks widely adopted in prior work on semi-decentralized FL for evaluating S2S and S2A separately: the MNIST dataset (Deng 2012) trained with a single-hidden-layer logistic classifier ($d = 7,850$ parameters), and the CIFAR-10 dataset (Krizhevsky and Hinton 2009) trained with a reference convolutional neural network ($d \approx 1.1$ million parameters) (Lin et al. 2021; Guo et al. 2021; Chen, Wang, and Brinton 2024).

We introduce intra- and inter-component heterogeneity mimicking the constants ζ_{intra} and ζ_{inter} of Assumption 5:

- **Inter-component heterogeneity.** We partition the dataset across components either through an IID split (each com-



(a) Intra non-IID, Inter IID (b) Intra IID, Inter non-IID

Figure 4: Accuracy gap on CIFAR-10 with ring topology.

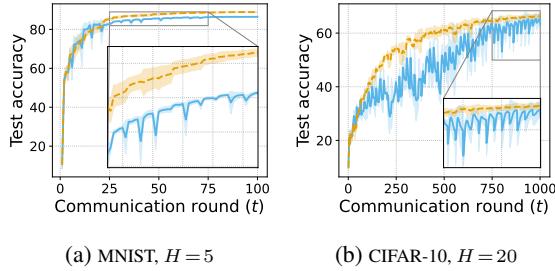


Figure 5: Test accuracy over communication rounds for intra IID, inter non-IID heterogeneity, $K/n=0.2$, ring topology.

ponent receives samples from all classes), or through a pathological non-IID split (each component receives samples from only half of the classes, with disjoint class sets) (McMahan et al. 2017).

- *Intra-component heterogeneity.* Within each component, we partition the dataset across devices either IID or non-IID, the latter through a Dirichlet distribution with concentration parameter 0.1 (Wang et al. 2019).

All models are trained with mini-batch SGD (batch size 128) for $T = 100$ rounds. For each algorithm, we tune the stepsize $\eta \in \{10^{-2.5}, 10^{-2}, 10^{-1.5}, 10^{-1}\}$. Results are averaged over five independent runs. Experiments are conducted on a Linux server with 4 Nvidia GeForce GTX 1080 Ti GPUs. The code is included in the supplementary material.

6.1 Experimental Heterogeneity Regimes

Figures 2 and 3 report the test accuracy achieved by S2S and S2A on MNIST and CIFAR-10 datasets, respectively.

Effect of sampling rate (Figs. 2–3, left column). For both S2S and S2A, accuracy improves as the sampling rate increases, with an average gain of +2 percentage points (p.p.) between $K/n = 0.2$ and $K/n = 1$. Interestingly, our experiments confirm the same qualitative heterogeneity regimes identified by our theoretical analysis:

- R1. Intra IID, Inter IID (Figs. 2–3(a)).** S2A outperforms S2S in over 80% of configurations, although the gain is modest (up to 1 p.p. on the ring for $K/n=0.2$).
- R2. Intra non-IID, Inter IID (Figs. 2–3(c)).** S2A outperforms in 40% of cases (up to +0.5 p.p. on the complete graph with high K/n), while S2S prevails in the remaining 60% (up to +8.4 p.p. on the ring at $K/n=0.2$).

- R3. Inter non-IID (Figs. 2–3(e,g)).** S2S outperforms S2A in over 90% of settings, with the largest gain at $K/n=0.2$ (+2.4 p.p. on MNIST, +7 p.p. on CIFAR-10).

Across the 96 evaluated configurations, S2S outperforms S2A in about 60% of cases, S2A in 30%, and the remaining 10% are not statistically significant (gap below standard error). Topology also plays a role: ring accounts for 45% of the largest gaps, grid for 30%, and complete graph for 25%.

Effect of server period (Figs. 2–3, right column). Accuracy decreases as the server period H increases (by an average of -2.4 p.p. from $H = 5$ to $H = 20$), highlighting the importance of frequent D2S communication. Again, our experiments confirm the theoretical regimes from Section 5.3:

- R1. Intra IID, Inter IID (Figs. 2–3(b)).** S2A consistently outperforms S2S in over 95% of cases, although the gap remains modest (below 1 p.p. at $H=5$, ring).
- R2. Intra non-IID, Inter IID (Figs. 2–3(d)).** S2S outperforms in 70% of configurations (up to +8.5 p.p. at $H = 5$, ring), whereas S2A prevails in the remaining 30% (up to +4 p.p. at $H = 10$, complete).
- R3. Inter non-IID (Figs. 2–3(g,h)).** S2S prevails in over 90% of configurations, with the largest gap at $H = 5$ (+2 p.p. on MNIST, +7 p.p. on CIFAR-10).

Across 96 comparisons, S2S outperforms in 60% of them, while S2A in 40%. Interestingly, in 80% of heterogeneity regimes, S2S shows a steeper accuracy drop with increasing H , yet it still outperforms S2A in 60% of these cases.

Intra vs. Inter Heterogeneity (Fig. 4). Figure 4 compares the accuracy gap (S2S minus S2A) on CIFAR-10 with ring topology under two opposite heterogeneity regimes. With non-IID intra and IID inter-component heterogeneity (Fig. 4(a)), S2S prevails at low sampling rates or low server periods (+8.4 p.p. at $K/n=0.2, H=5$), while S2A prevails for higher K/n or H (+1.2 p.p. at $K/n = 0.6, H = 15$). In the opposite regime, with IID intra and non-IID inter heterogeneity (Fig. 4(b)), S2S consistently outperforms S2A.

Learning curves (Fig. 5). To better understand why S2S outperforms S2A in the inter non-IID regime, Figure 5 reports representative test accuracy over communication rounds. While S2A’s broadcast step initially accelerates inter-component information exchange and achieves higher early-round accuracy, it becomes detrimental in later stages, with periodic drops in test accuracy at every D2S round.

7 Conclusion

This paper provides the first theoretical and empirical comparison of two fundamental server-to-device communication primitives for semi-decentralized federated learning: sampled-to-all (S2A) and sampled-to-sampled (S2S). Our results yield practical configuration guidelines: S2S is the better choice when (i) inter-component heterogeneity is high; or (ii) intra-component heterogeneity is high, and the server can sample only a small subset of devices while D2S communication is more frequent. Conversely, when data are nearly IID across components or when a high sampling rate and a well-connected topology mitigate intra-component noise, S2A offers the potential to accelerate convergence.

Acknowledgments

This research was supported by the Knut and Alice Wallenberg Foundation; by ELLIIT and the Swedish Research Council (VR); by the French government through the “Plan de relance” and the 3IA Côte d’Azur Investments in the Future project, managed by the National Research Agency (ANR) under reference ANR-19-P3IA-0002; by the European Network of Excellence dAIEDGE (Grant Agreement No. 101120726) and the EU HORIZON MSCA 2023 DN project FINALITY (Grant Agreement No. 101168816); and by Groupe La Poste, sponsor of the Inria Foundation, within the framework of the FedMalin Inria Challenge. Experiments presented in this paper were carried out using the Grid’5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see <https://www.grid5000.fr>).

References

- Alistarh, D.; Grubic, D.; Li, J. et al. 2017. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Bonawitz, K.; Eichner, H.; Grieskamp, W. et al. 2019. Towards Federated Learning at Scale: System Design. *Proceedings of Machine Learning and Systems*, 1: 374–388.
- Boyd, S.; Ghosh, A.; Prabhakar, B. et al. 2006. Randomized Gossip Algorithms. *IEEE Transactions on Information Theory*, 52(6): 2508–2530.
- Bubeck, S. 2015. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning*, 8(3-4): 231–357.
- Chen, E.; Wang, S.; and Brinton, C. G. 2024. Taming Subnet-Drift in D2D-Enabled Fog Learning: A Hierarchical Gradient Tracking Approach. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*, 2438–2447.
- Chen, Y.; Yuan, K.; Zhang, Y. et al. 2021. Accelerating Gossip SGD with Periodic Global Averaging. In *Proceedings of the 38th International Conference on Machine Learning*, 1791–1802. PMLR.
- Deng, L. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine*, 29(6): 141–142.
- Guo, Y.; Sun, Y.; Hu, R. et al. 2021. Hybrid Local SGD for Federated Learning with Heterogeneous Communications. In *International Conference on Learning Representations*.
- He, C.; Li, S.; So, J. et al. 2020. FedML: A Research Library and Benchmark for Federated Machine Learning. arXiv:2007.13518.
- Horvóth, S.; Ho, C.-Y.; Horvath, L. et al. 2022. Natural Compression for Distributed Deep Learning. In *Proceedings of Mathematical and Scientific Machine Learning*, 129–141. PMLR.
- Jhunjhunwala, D.; Sharma, P.; Nagarkatti, A. et al. 2022. Fedvarp: Tackling the Variance Due to Partial Client Participation in Federated Learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, 906–916. PMLR.
- Kairouz, P.; McMahan, H. B.; Avent, B. et al. 2021. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1-2): 1–210.
- Karimireddy, S. P.; Kale, S.; Mohri, M. et al. 2020. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning*, 5132–5143. PMLR.
- Koloskova, A.; Loizou, N.; Boreiri, S. et al. 2020. A Unified Theory of Decentralized SGD with Changing Topology and Local Updates. In *Proceedings of the 37th International Conference on Machine Learning*, 5381–5393. PMLR.
- Konečný, J.; McMahan, H. B.; Yu, F. X. et al. 2017. Federated Learning: Strategies for Improving Communication Efficiency. arXiv:1610.05492.
- Krizhevsky, A.; and Hinton, G. 2009. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto.
- Larsson, E. G.; and Michelusi, N. 2025. Unified Analysis of Decentralized Gradient Descent: A Contraction Mapping Framework. *IEEE Open Journal of Signal Processing*, 6: 507–529.
- Le Bars, B.; Bellet, A.; Tommasi, M. et al. 2023. Refined Convergence and Topology Learning for Decentralized SGD with Heterogeneous Data. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 1672–1702. PMLR.
- Li, T.; Sahu, A. K.; Talwalkar, A. et al. 2020a. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3): 50–60.
- Li, X.; Huang, K.; Yang, W. et al. 2020b. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.
- Lian, X.; Zhang, C.; Zhang, H. et al. 2017. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Lin, F. P.-C.; Hosseinalipour, S.; Azam, S. S. et al. 2021. Semi-Decentralized Federated Learning With Cooperative D2D Local Model Aggregations. *IEEE Journal on Selected Areas in Communications*, 39(12): 3851–3869.
- McMahan, B.; Moore, E.; Ramage, D. et al. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273–1282. PMLR.
- Mishchenko, K.; Malinovsky, G.; Stich, S. et al. 2022. Prox-Skip: Yes! Local Gradient Steps Provably Lead to Communication Acceleration! Finally! In *Proceedings of the 39th International Conference on Machine Learning*, 15750–15769. PMLR.

Nedić, A.; and Olshevsky, A. 2016. Stochastic Gradient-Push for Strongly Convex Functions on Time-Varying Directed Graphs. *IEEE Transactions on Automatic Control*, 61(12): 3936–3947.

Neglia, G.; Xu, C.; Towsley, D. et al. 2020. Decentralized Gradient Methods: Does Topology Matter? In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2348–2358. PMLR.

Nesterov, Y. 2004. *Introductory Lectures on Convex Optimization*. Applied Optimization. Boston, MA: Springer US. ISBN 978-1-4613-4691-3 978-1-4419-8853-9.

Nguyen, L. M.; Nguyen, P. H.; Richtárik, P. et al. 2019. New Convergence Aspects of Stochastic Gradient Algorithms. *Journal of Machine Learning Research*, 20(176): 1–49.

Reddi, S. J.; Charles, Z.; Zaheer, M. et al. 2021. Adaptive Federated Optimization. In *International Conference on Learning Representations*.

Shamir, O.; Srebro, N.; and Zhang, T. 2014. Communication-Efficient Distributed Optimization Using an Approximate Newton-type Method. In *Proceedings of the 31st International Conference on Machine Learning*, 1000–1008. PMLR.

Stich, S. U. 2018. Local SGD Converges Fast and Communicates Little. In *International Conference on Learning Representations*.

Wang, H.; Yurochkin, M.; Sun, Y. et al. 2019. Federated Learning with Matched Averaging. In *International Conference on Learning Representations*.

Wang, Z.; Xu, H.; Liu, J. et al. 2021. Resource-Efficient Federated Learning with Hierarchical Aggregation in Edge Computing. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 1–10.

Xiao, L.; and Boyd, S. 2004. Fast Linear Iterations for Distributed Averaging. *Systems & Control Letters*, 53(1): 65–78.

Yuan, K.; Ling, Q.; and Yin, W. 2016. On the Convergence of Decentralized Gradient Descent. *SIAM Journal on Optimization*, 26(3): 1835–1854.

APPENDIX

A Unified Convergence Analysis for Semi-Decentralized Learning: Sampled-to-All vs. Sampled-to-Sampled Communication

The appendix is organized as follows:

Appendix A	Unified Framework for the Convergence Analysis
Appendix B	Convergence Analysis of S2S
Appendix C	Convergence Analysis of S2A
Appendix D	Additional Theoretical Results
Appendix E	Additional Experimental Results

Algorithm 2.A: S2S — VECTOR NOTATION

Input: initial parameters $\mathbf{x}_i^{(0)} = \mathbf{x}^{(0)} \in \mathbb{R}^d$ for all $i \in \mathcal{V}$, communication rounds T , server aggregation period H , stepsizes $\{\eta_t\}$, mixing distribution \mathcal{W}

```

1: for  $t = 0, \dots, T - 1$  do
2:   sample mixing matrix  $W^{(t)} \sim \mathcal{W}$ 
3:   for each device  $i \in \mathcal{V}$ , in parallel do
4:     sample batch  $\mathcal{B}_i^{(t)}$  and compute  $\nabla F_i(\mathbf{x}_i^{(t)}, \mathcal{B}_i^{(t)})$ 
5:      $\mathbf{x}_i^{(t+1/3)} = \mathbf{x}_i^{(t)} - \eta_t \nabla F_i(\mathbf{x}_i^{(t)}, \mathcal{B}_i^{(t)})$ 
6:      $\mathbf{x}_i^{(t+2/3)} = \sum_{j=1}^n W_{ij}^{(t)} \mathbf{x}_j^{(t+1/3)}$ 
7:   if  $t \in \mathcal{H}$  then
8:     sample devices  $\mathcal{S}^{(t)} \subseteq \mathcal{V}$ ,  $|\mathcal{S}^{(t)}| = K$ 
9:     compute  $\hat{\mathbf{x}}^{(t+1)} = \frac{1}{K} \sum_{i \in \mathcal{S}^{(t)}} \mathbf{x}_i^{(t+2/3)}$ 
10:    send  $\hat{\mathbf{x}}^{(t+1)}$  to the sampled devices only:
        
$$\mathbf{x}_i^{(t+1)} = \begin{cases} \hat{\mathbf{x}}^{(t+1)}, & i \in \mathcal{S}^{(t)} \\ \mathbf{x}_i^{(t+2/3)}, & \text{otherwise} \end{cases}$$

11:   else
12:      $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t+2/3)}$ 
13: return  $\{\mathbf{x}_i^{(T)}\}_{i \in \mathcal{V}}$ 

```

Algorithm 2.B: S2S — MATRIX NOTATION

Input: initial parameters $X^{(0)} \in \mathbb{R}^{d \times n}$, communication rounds T , server aggregation period H , stepsizes $\{\eta_t\}$, mini-batches $\mathcal{B}^{(t)}$, mixing distribution \mathcal{W}

```

1: for  $t = 0, \dots, T - 1$  do
2:   sample mixing matrix  $W^{(t)} \sim \mathcal{W}$ 
3:    $X^{(t+1/3)} \leftarrow X^{(t)} - \eta_t \nabla F(X^{(t)}, \mathcal{B}^{(t)})$ 
4:    $X^{(t+2/3)} \leftarrow X^{(t+1/3)} W^{(t)}$ 
5:   if  $t \in \mathcal{H}$  then
6:     sample devices  $\mathcal{S}^{(t)} \subseteq \mathcal{V}$ ,  $|\mathcal{S}^{(t)}| = K$ 
7:     build  $(W_{\text{S2S}}^{(t)})_{ij} = \begin{cases} \frac{1}{K}, & i, j \in \mathcal{S}^{(t)} \\ 1, & i = j \notin \mathcal{S}^{(t)} \\ 0, & \text{otherwise} \end{cases}$ 
8:      $X^{(t+1)} \leftarrow X^{(t+2/3)} W_{\text{S2S}}^{(t)}$ 
9:   else
10:     $X^{(t+1)} \leftarrow X^{(t+2/3)}$ 
11: return  $X^{(T)}$ 

```

Algorithm 3.A: S2A — VECTOR NOTATION

Input: initial parameters $\mathbf{x}_i^{(0)} = \mathbf{x}^{(0)} \in \mathbb{R}^d$ for all $i \in \mathcal{V}$, communication rounds T , server aggregation period H , stepsizes $\{\eta_t\}$, mixing distribution \mathcal{W}

```

1: for  $t = 0, \dots, T - 1$  do
2:   sample mixing matrix  $W^{(t)} \sim \mathcal{W}$ 
3:   for each device  $i \in \mathcal{V}$ , in parallel do
4:     sample batch  $\mathcal{B}_i^{(t)}$  and compute  $\nabla F_i(\mathbf{x}_i^{(t)}, \mathcal{B}_i^{(t)})$ 
5:      $\mathbf{x}_i^{(t+1/3)} = \mathbf{x}_i^{(t)} - \eta_t \nabla F_i(\mathbf{x}_i^{(t)}, \mathcal{B}_i^{(t)})$ 
6:      $\mathbf{x}_i^{(t+2/3)} = \sum_{j=1}^n W_{ij}^{(t)} \mathbf{x}_j^{(t+1/3)}$ 
7:   if  $t \in \mathcal{H}$  then
8:     sample devices  $\mathcal{S}^{(t)} \subseteq \mathcal{V}$ ,  $|\mathcal{S}^{(t)}| = K$ 
9:     compute  $\hat{\mathbf{x}}^{(t+1)} = \frac{1}{K} \sum_{i \in \mathcal{S}^{(t)}} \mathbf{x}_i^{(t+2/3)}$ 
10:    broadcast  $\hat{\mathbf{x}}^{(t+1)}$  to all devices:
        
$$\mathbf{x}_i^{(t+1)} = \hat{\mathbf{x}}^{(t+1)} \text{ for all } i \in \mathcal{V}$$

11:   else
12:      $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t+2/3)}$ 
13: return  $\{\mathbf{x}_i^{(T)}\}_{i \in \mathcal{V}}$ 

```

Algorithm 3.B: S2A — MATRIX NOTATION

Input: initial parameters $X^{(0)} \in \mathbb{R}^{d \times n}$, communication rounds T , server aggregation period H , stepsizes $\{\eta_t\}$, mini-batches $\mathcal{B}^{(t)}$, mixing distribution \mathcal{W}

```

1: for  $t = 0, \dots, T - 1$  do
2:   sample mixing matrix  $W^{(t)} \sim \mathcal{W}$ 
3:    $X^{(t+1/3)} \leftarrow X^{(t)} - \eta_t \nabla F(X^{(t)}, \mathcal{B}^{(t)})$ 
4:    $X^{(t+2/3)} \leftarrow X^{(t+1/3)} W^{(t)}$ 
5:   if  $t \in \mathcal{H}$  then
6:     sample devices  $\mathcal{S}^{(t)} \subseteq \mathcal{V}$ ,  $|\mathcal{S}^{(t)}| = K$ 
7:     build  $(W_{\text{S2A}}^{(t)})_{ij} = \begin{cases} \frac{1}{K}, & i \in \mathcal{S}^{(t)}; \\ 0, & \text{otherwise.} \end{cases}$ 
8:      $X^{(t+1)} \leftarrow X^{(t+2/3)} W_{\text{S2A}}^{(t)}$ 
9:   else
10:     $X^{(t+1)} \leftarrow X^{(t+2/3)}$ 
11: return  $X^{(T)}$ 

```

A Unified Framework for the Convergence Analysis

We can now rewrite Algorithm 1, separately for S2S and S2A, in both vector and matrix form: Algorithms 2.A and 2.B for S2S, and Algorithms 3.A and 3.B for S2A. The following notation extends the one in the main text:

$$\begin{aligned}\mathcal{H} &:= \left\{ t \leq T \mid t \equiv 0 \pmod{H} \right\}, \\ X^{(t)} &:= \left[\mathbf{x}_1^{(t)}, \dots, \mathbf{x}_n^{(t)} \right] \in \mathbb{R}^{d \times n}, \\ \bar{X}^{(t)} &:= \left[\bar{\mathbf{x}}^{(t)}, \dots, \bar{\mathbf{x}}^{(t)} \right] \in \mathbb{R}^{d \times n}, \\ \nabla F(X^{(t)}, \xi^{(t)}) &:= \left[\nabla F_1(\mathbf{x}_1^{(t)}, \xi_1^{(t)}), \dots, \nabla F_n(\mathbf{x}_n^{(t)}, \xi_n^{(t)}) \right] \in \mathbb{R}^{d \times n}.\end{aligned}$$

Descent Lemmas

The following lemmas bound the per-iteration descent from iterate $\bar{\mathbf{x}}^{(t+1)}$ to iterate $\bar{\mathbf{x}}^{(t)}$ in terms of disagreement error ($\frac{1}{n} \mathbb{E} \|X^{(t)} - \bar{X}^{(t)}\|_F^2$) and bias error ($\mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}\|_2^2$).

Lemma 3 assumes that the local objectives are convex, while Lemma 4 is for non-convex objectives.

Lemma 3 (Descent Lemma (Convex Objectives)). *Under Assumptions 1–6, for all $t \geq 0$, the average $\bar{\mathbf{x}}^{(t)} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(t)}$ of the iterates of Algorithms 1–2 with the stepsize $\eta_t \leq \frac{1}{4L}$ satisfies:*

$$\mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|_2^2 \leq \mathbb{E} \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|_2^2 - \eta_t \mathbb{E} (f(\bar{\mathbf{x}}^{(t)}) - f^*) + \frac{\eta_t^2 \bar{\sigma}^2}{n} + \frac{3\eta_t L}{2} \frac{1}{n} \mathbb{E} \|X^{(t)} - \bar{X}^{(t)}\|_F^2 + \mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}\|_2^2,$$

where the term $\Xi^{(t)} := \frac{1}{n} \mathbb{E} \|X^{(t)} - \bar{X}^{(t)}\|_F^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\|_2^2$ is the disagreement error, and the term $\mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}\|_2^2$ is the bias error.

Proof of Lemma 3. We decompose the optimality gap into:

$$\mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|_2^2 = \mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}\|_2^2 + \mathbb{E} \|\bar{\mathbf{x}}^{(t+\frac{2}{3})} - \mathbf{x}^*\|_2^2,$$

where orthogonality follows from $\mathbb{E}_{S^{(t)}} [\bar{\mathbf{x}}^{(t+1)}] = \bar{\mathbf{x}}^{(t+\frac{2}{3})}$ for all $t \geq 0$, which holds immediately for $t \notin \mathcal{H}$, and is proved in the subsequent Lemmas 8 (ii) and 11 (iii) for $t \in \mathcal{H}$.

The bias error $\mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}\|_2^2$ is a key difference for S2S and S2A. We bound it separately in Lemmas 9 and 12.

The term $\mathbb{E} \|\bar{\mathbf{x}}^{(t+\frac{2}{3})} - \mathbf{x}^*\|_2^2$ follows the D-SGD descent lemma for convex objectives. For this reason, we refer the reader to (Koloskova et al. 2020, Lemma 8) and (Le Bars et al. 2023, Lemma 1). \square

Lemma 4 (Descent Lemma (Non-Convex Objectives)). *Under Assumptions 1 and 3–6, for all $t \geq 0$, the average $\bar{\mathbf{x}}^{(t)} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(t)}$ of the iterates of Algorithms 1–2 with the stepsize $\eta_t \leq \frac{1}{4L}$ satisfies:*

$$\mathbb{E}[f(\bar{\mathbf{x}}^{(t+1)})] \leq \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})] - \frac{\eta_t}{4} \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 + \frac{\eta_t^2 L \bar{\sigma}^2}{2n} + \eta_t L^2 \frac{1}{n} \mathbb{E} \|X^{(t)} - \bar{X}^{(t)}\|_F^2 + \frac{L}{2} \mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}\|_2^2.$$

Proof of Lemma 4. By L -smoothness of $f(\cdot)$ (Assumption 1, see (Nesterov 2004)):

$$\mathbb{E}[f(\bar{\mathbf{x}}^{(t+1)})] \leq f(\bar{\mathbf{x}}^{(t)}) + \mathbb{E} \langle \nabla f(\bar{\mathbf{x}}^{(t)}), \bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)} \rangle + \frac{L}{2} \mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\|_2^2.$$

For the term $\mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\|_2^2$, we again invoke the error decomposition:

$$\mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\|_2^2 = \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}\|_2^2 + \|\bar{\mathbf{x}}^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+1)}\|_2^2,$$

where the cross product is zero because $\mathbb{E}_{S^{(t)}} [\bar{\mathbf{x}}^{(t+1)}] = \bar{\mathbf{x}}^{(t+\frac{2}{3})}$.

Therefore:

$$\mathbb{E}[f(\bar{\mathbf{x}}^{(t+1)})] \leq \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})] + \mathbb{E} \langle \nabla f(\bar{\mathbf{x}}^{(t)}), \bar{\mathbf{x}}^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t)} \rangle + \frac{L}{2} \mathbb{E} \|\bar{\mathbf{x}}^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t)}\|_2^2 + \frac{L}{2} \mathbb{E} \|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}\|_2^2.$$

Once we isolated the bias error, the remaining terms follow the standard descent lemma for D-SGD on non-convex objectives; we therefore refer the reader to (Koloskova et al. 2020, Lemma 11) and (Le Bars et al. 2023, Lemma 2). \square

D2D Disagreement Errors

The next lemma bounds the disagreement error after the D2D round, $\mathbb{E}\|X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})}\|_F^2$, common to both S2S and S2A. We will bound the D2S disagreement error, $\mathbb{E}\|X^{(t+1)} - \bar{X}^{(t+1)}\|_F^2$, specific to each primitive, separately in Lemmas 10 and 13.

Lemma 5 (Disagreement Errors (D2D)). *For all $t \geq 0$,*

Error Decomposition. *We decompose the D2D disagreement into intra- and inter-component terms:*

$$\mathbb{E}\left\|X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})}\right\|_F^2 = \mathbb{E}\left\|X^{(t+\frac{2}{3})} - X^{(t+\frac{2}{3})}\Pi_C\right\|_F^2 + \mathbb{E}\left\|X^{(t+\frac{2}{3})}\Pi_C - \bar{X}^{(t+\frac{2}{3})}\right\|_F^2.$$

Intra-Component Disagreement. *Under Assumptions 1 and 3–6, for $\eta_t \leq \frac{p}{8L}$:*

$$\mathbb{E}\left\|X^{(t+\frac{2}{3})} - X^{(t+\frac{2}{3})}\Pi_C\right\|_F^2 \leq \left(1 - \frac{p}{4}\right)\mathbb{E}\left\|X^{(t)} - X^{(t)}\Pi_C\right\|_F^2 + \frac{6n\eta^2\bar{\zeta}_{\text{intra}}^2}{p}.$$

Inter-Component Disagreement. *Under Assumptions 1 and 3–6, for $\eta_t \leq \frac{p}{8L}$, there exists $\rho > 0$ such that:*

$$\mathbb{E}\left\|X^{(t+\frac{2}{3})}\Pi_C - \bar{X}^{(t+\frac{2}{3})}\right\|_F^2 \leq (1 + \rho)\mathbb{E}\left\|X^{(t)}\Pi_C - \bar{X}^{(t)}\right\|_F^2 + (1 + \rho^{-1})n\eta^2\bar{\zeta}_{\text{inter}}^2.$$

The D2D round alone does not contract the inter-component disagreement ($\rho > 0$).

Proof of Lemma 5. The error decomposition follows from Lemma 1 in the main paper.

For the intra-component disagreement,

$$\begin{aligned} \mathbb{E}\left\|X^{(t+\frac{2}{3})} - X^{(t+\frac{2}{3})}\Pi_C\right\|_F^2 &\stackrel{(a)}{=} \mathbb{E}\left\|X^{(t+\frac{2}{3})}(I - \Pi_C)\right\|_F^2 \\ &\stackrel{(b)}{=} \mathbb{E}\left\|X^{(t+\frac{1}{3})}W^{(t)}(I - \Pi_C)\right\|_F^2 \\ &\stackrel{(c)}{=} \mathbb{E}\left\|\left(X^{(t)} - \eta\nabla F(X^{(t)}, \xi^{(t)})\right)(W^{(t)} - \Pi_C)\right\|_F^2 \\ &\stackrel{(d)}{\leq} (1 + \alpha)\mathbb{E}\left\|X^{(t)}\left(W^{(t)} - \Pi_C\right)\right\|_F^2 + (1 + \alpha^{-1})\eta^2\mathbb{E}\left\|\nabla F(X^{(t)}, \xi^{(t)})\left(W^{(t)} - \Pi_C\right)\right\|_F^2 \\ &\stackrel{(e)}{\leq} (1 + \alpha)(1 - p)\mathbb{E}\left\|X^{(t)}(I - \Pi_C)\right\|_F^2 + (1 + \alpha^{-1})(1 - p)n\eta^2\bar{\zeta}_{\text{intra}}^2 \\ &\stackrel{(f)}{\leq} \left(1 - \frac{p}{4}\right)\mathbb{E}\left\|X^{(t)} - X^{(t)}\Pi_C\right\|_F^2 + \frac{6n\eta^2\bar{\zeta}_{\text{intra}}^2}{p}, \end{aligned}$$

where equalities (a)–(c) follow from the D2D update rule $X^{(t+\frac{2}{3})} = (X^{(t)} - \eta\nabla F(X^{(t)}, \xi^{(t)}))W^{(t)}$; inequality (d) applies $\|\mathbf{a} + \mathbf{b}\|_2^2 = (1 + \alpha)\|\mathbf{a}\|_2^2 + (1 + \alpha^{-1})\|\mathbf{b}\|_2^2$ for any $\alpha > 0$; inequality (e) uses intermediate steps detailed in (Le Bars et al. 2023, Lemma 3); and inequality (f) sets $\alpha = \frac{p}{2}$ and uses $p \in (0, 1]$, therefore $1 - p < 1$.

For the inter-component disagreement,

$$\begin{aligned} \mathbb{E}\left\|X^{(t+\frac{2}{3})}\Pi_C - \bar{X}^{(t+\frac{2}{3})}\right\|_F^2 &\stackrel{(g)}{=} \mathbb{E}\left\|X^{(t+\frac{2}{3})}(\Pi_C - \Pi)\right\|_F^2 \\ &\stackrel{(h)}{=} \mathbb{E}\left\|X^{(t+\frac{1}{3})}W^{(t)}(\Pi_C - \Pi)\right\|_F^2 \\ &\stackrel{(i)}{=} \mathbb{E}\left\|\left(X^{(t)} - \eta\nabla F(X^{(t)}, \xi^{(t)})\right)(\Pi_C - \Pi)\right\|_F^2 \\ &\stackrel{(j)}{\leq} (1 + \rho)\mathbb{E}\left\|X^{(t)}(\Pi_C - \Pi)\right\|_F^2 + (1 + \rho^{-1})\eta^2\mathbb{E}\left\|\nabla F(X^{(t)}, \xi^{(t)})(\Pi_C - \Pi)\right\|_F^2 \\ &\stackrel{(k)}{\leq} (1 + \rho)\mathbb{E}\left\|X^{(t)}\Pi_C - \bar{X}^{(t)}\right\|_F^2 + (1 + \rho^{-1})n\eta^2\bar{\zeta}_{\text{inter}}^2, \end{aligned}$$

where steps (g)–(j) replicate the arguments in (a)–(d), with $\rho > 0$, while inequality (k) is borrowed from (Le Bars et al. 2023, Lemma 3). The choice of ρ will be addressed separately for S2S and S2A in Theorems 1 and 2. \square

Alternating Disagreement Recursion

The following lemma is a central building block of our unified analysis. It applies to both S2S and S2A, and is used in the proofs of Theorems 1 and 2. The recursion parameters, however, will differ for the two primitives, producing different results.

Lemma 6 (Disagreement Recursion). *Let $\{\Xi^{(t)}\}_{t \geq 0}$ be a nonnegative sequence satisfying the recursion*

$$\Xi^{(t)} \leq \begin{cases} a_1 \Xi^{(t-1)} + b_1, & t \equiv 0 \pmod{H}, \\ a_2 \Xi^{(t-1)} + b_2, & t \not\equiv 0 \pmod{H}, \end{cases}$$

with constants $a_1, a_2, b_1, b_2 \geq 0$, and $H \geq 1$. Define $0 \leq C := a_1 a_2^{H-1} < 1$, and $D := a_1 b_2 \frac{1-a_2^{H-1}}{1-a_2} + b_1$.

Then, for any horizon $T \geq 0$,

$$\bar{\Xi}^{(T+1)} := \frac{1}{T+1} \sum_{t=0}^T \Xi^{(t)} \leq \begin{cases} \left[\frac{D}{(1-C)(1-a_2)} + \frac{b_2}{1-a_2}(H-1) \right] \left(\frac{1}{H} + \frac{1}{T+1} \right), & 0 \leq a_2 < 1, \\ \left[\frac{D}{1-C} \frac{a_2^H - 1}{a_2 - 1} + \frac{b_2(a_2^H - a_2 H + H - 1)}{(a_2 - 1)^2} \right] \left(\frac{1}{H} + \frac{1}{T+1} \right), & a_2 > 1. \end{cases}$$

Proof of Lemma 6. For any $t = mH + s$, where $m := \lfloor \frac{t}{H} \rfloor$ and $s \in \{0, \dots, H-1\}$,

$$\Xi^{(mH+s)} \leq a_2^s \Xi^{(mH)} + b_2 \sum_{k=0}^{s-1} a_2^k = a_2^s \Xi^{(mH)} + b_2 \frac{1-a_2^s}{1-a_2}.$$

First, consider the behavior at $t = mH$:

$$\begin{aligned} \Xi^{(mH)} &\leq C \Xi^{((m-1)H)} + D \\ &\leq C^m \underbrace{\Xi^{(0)}}_{=0} + D \sum_{k=0}^{m-1} C^k = \frac{D}{1-C} (1 - C^m). \end{aligned}$$

Combining:

$$\Xi^{(t)} \leq a_2^s \left[\frac{D}{1-C} (1 - C^m) \right] + b_2 \frac{1-a_2^s}{1-a_2}.$$

Second, sum over one full period ($s = 0, \dots, H-1$):

$$\begin{aligned} \sum_{s=0}^{H-1} \Xi^{(mH+s)} &\leq \frac{D}{1-C} \underbrace{\sum_{s=0}^{H-1} a_2^s}_{:=S_1} - \underbrace{\frac{C^m}{1-C} \sum_{s=0}^{H-1} a_2^s}_{\text{decays as } C^m, C < 1} + \frac{b_2}{1-a_2} \underbrace{\sum_{s=1}^{H-1} (1-a_2^s)}_{:=S_2} \\ &\leq \underbrace{\frac{D}{1-C} S_1}_{:=Q} + \frac{b_2}{1-a_2} S_2. \end{aligned}$$

Third, sum up to $T = MH + S$, with $M := \lfloor \frac{T}{H} \rfloor \leq (T + H - 1)/H$, $S \in \{0, \dots, H-1\}$.

The contribution of the M complete periods is:

$$\sum_{m=0}^{M-1} \sum_{s=0}^{H-1} \Xi^{(mH+s)} \leq MQ.$$

The contribution of the partial period S is:

$$\sum_{s=0}^{S-1} \Xi^{(MH+s)} \leq Q.$$

Combining:

$$\sum_{t=1}^T \Xi^{(t)} \leq (M+1)Q \leq \frac{T+H}{H} Q.$$

Next, divide by $T + 1$:

$$\bar{\Xi}^{(T+1)} := \frac{1}{T+1} \sum_{t=0}^T \Xi^{(t)} \leq \frac{T+H}{H(T+1)} Q \leq \left(\frac{1}{H} + \frac{1}{T+1} \right) \left[\frac{D}{1-C} S_1 + \frac{b_2}{1-a_2} S_2 \right].$$

Finally, consider the two signs of a_2 separately.

Contractive step ($0 \leq a_2 < 1$). $S_1 = \sum_{s=0}^{H-1} a_2^s = \frac{1-a_2^H}{1-a_2} \leq \frac{1}{1-a_2}$ and $S_2 \leq H - 1 - \frac{a_2}{1-a_2} \leq H - 1$:

$$\bar{\Xi}^{(T+1)} \leq \left(\frac{1}{H} + \frac{1}{T+1} \right) \left[\frac{D}{(1-C)(1-a_2)} + \frac{b_2}{1-a_2}(H-1) \right].$$

Expansive step ($a_2 > 1$). $S_1 = \sum_{s=0}^{H-1} a_2^s = \frac{a_2^H - 1}{a_2 - 1}$ and $S_2 = \sum_{s=1}^{H-1} (1 - a_2^s) = -\frac{a_2^H - a_2 H + H - 1}{a_2 - 1}$:

$$\bar{\Xi}^{(T+1)} \leq \left(\frac{1}{H} + \frac{1}{T+1} \right) \left[\frac{D}{1-C} \frac{a_2^H - 1}{a_2 - 1} + \frac{b_2(a_2^H - a_2 H + H - 1)}{(a_2 - 1)^2} \right].$$

□

Alternating Convergence Recursion

The next lemma telescopes the descent recursion over a time horizon T , alternating D2D and D2S rounds with period H .

Lemma 7 (Convergence Recursion). *Let $\mathcal{H} := \{t \leq T : t \equiv 0 \pmod{H}\}$.*

Consider a nonnegative sequence $\{r^{(t)}\}_{t \geq 0}$ satisfying the descent recursion

$$r^{(t+1)} \leq \begin{cases} r^{(t)} - \eta \Delta^{(t)} + \frac{\eta^2 \bar{\sigma}^2}{n} + a \Xi_{\text{intra}}^{(t)} + b \Xi_{\text{inter}}^{(t)} + c + d, & t \in \mathcal{H}, \\ r^{(t)} - \eta \Delta^{(t)} + \frac{\eta^2 \bar{\sigma}^2}{n} + e (\Xi_{\text{intra}}^{(t)} + \Xi_{\text{inter}}^{(t)}), & t \notin \mathcal{H}. \end{cases}$$

Then, for any horizon $T \geq 0$,

$$\frac{1}{T+1} \sum_{t=0}^T \Delta^{(t)} \leq \frac{r^{(0)}}{\eta(T+1)} + \frac{\eta \bar{\sigma}^2}{n} + \frac{e + \frac{a-e}{H}}{\eta} \bar{\Xi}_{\text{intra}}^{(T+1)} + \frac{e + \frac{b-e}{H}}{\eta} \bar{\Xi}_{\text{inter}}^{(T+1)} + \frac{c+d}{\eta H},$$

where $\bar{\Xi}_{\text{intra}}^{(T+1)} := \frac{1}{T+1} \sum_{t=0}^T \Xi_{\text{intra}}^{(t)}$, and $\bar{\Xi}_{\text{inter}}^{(T+1)} := \frac{1}{T+1} \sum_{t=0}^T \Xi_{\text{inter}}^{(t)}$.

Proof of Lemma 7. First, isolate $\Delta^{(t)}$ by rearranging the descent recursion and dividing by η :

$$\Delta^{(t)} \leq \frac{r^{(t)} - r^{(t+1)}}{\eta} + \frac{\eta \bar{\sigma}^2}{n} + \begin{cases} \frac{a}{\eta} \Xi_{\text{intra}}^{(t)} + \frac{b}{\eta} \Xi_{\text{inter}}^{(t)} + \frac{c+d}{\eta}, & t \in \mathcal{H}, \\ \frac{e}{\eta} (\Xi_{\text{intra}}^{(t)} + \Xi_{\text{inter}}^{(t)}), & t \notin \mathcal{H}. \end{cases}$$

Next, sum and average over $t = 0, \dots, T$:

$$\begin{aligned} \frac{1}{T+1} \sum_{t=0}^T \Delta^{(t)} &\leq \frac{r^{(0)}}{\eta(T+1)} + \frac{\eta \bar{\sigma}^2}{n} \\ &\quad + \frac{1}{T+1} \sum_{t=0}^T \frac{\mathbb{1}_{t \in \mathcal{H}} a + \mathbb{1}_{t \notin \mathcal{H}} e}{\eta} \Xi_{\text{intra}}^{(t)} + \frac{1}{T+1} \sum_{t=0}^T \frac{\mathbb{1}_{t \in \mathcal{H}} b + \mathbb{1}_{t \notin \mathcal{H}} e}{\eta} \Xi_{\text{inter}}^{(t)} + \frac{1}{T+1} \sum_{t=0}^T \frac{\mathbb{1}_{t \in \mathcal{H}} (c+d)}{\eta}. \end{aligned}$$

Observe that $\frac{|\mathcal{H}|}{T+1} \leq \frac{1}{H}$, because exactly one index in every block of length H lies in \mathcal{H} .

Consequently,

$$\frac{1}{T+1} \sum_{t=0}^T (\mathbb{1}_{t \in \mathcal{H}} a + \mathbb{1}_{t \notin \mathcal{H}} e) \Xi_{\text{intra}}^{(t)} \leq \left(e + \frac{a-e}{H} \right) \bar{\Xi}_{\text{intra}}^{(T+1)},$$

and an analogous bound holds for the inter-component term. Moreover, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{1}_{t \in \mathcal{H}} \leq \frac{1}{H}$. □

B Convergence Analysis of S2S

Properties of S2S

Lemma 8 (Sampled-to-Sampled). *Define:*

$$(W_{S2S}^{(t)})_{ij} = \begin{cases} \frac{1}{K}, & i, j \in \mathcal{S}^{(t)}, \\ 1, & i = j \notin \mathcal{S}^{(t)}, \\ 0, & \text{otherwise.} \end{cases}$$

The matrix $W_{S2S}^{(t)}$ satisfies the following properties:

(i) **Symmetric and stochastic.** $(W_{S2S}^{(t)})^\top = W_{S2S}^{(t)}$, and $W_{S2S}^{(t)}\mathbf{1} = \mathbf{1}$.

Consequently, $W_{S2S}^{(t)}\Pi = \Pi W_{S2S}^{(t)} = \Pi$.

(ii) **The bias error is zero.** $W_{S2S}^{(t)}$ preserves the average of the iterates between D2D and D2S rounds:

$$\bar{X}^{(t+1)} = X^{(t+2/3)}W_{S2S}^{(t)}\Pi = X^{(t+2/3)}\Pi =: \bar{X}^{(t+2/3)}.$$

(iii) **Disagreement error.** The matrix $W_{S2S}^{(t)}$ leaves residual disagreement:

$$X^{(t+1)} = X^{(t+2/3)}W_{S2S}^{(t)} \neq \bar{X}^{(t+1)} := X^{(t+2/3)}W_{S2S}^{(t)}\Pi = X^{(t+2/3)}\Pi.$$

We define the disagreement error at time $t + 1$ as:

$$n\Xi^{(t+1)} := \mathbb{E} \left\| X^{(t+1)} - \bar{X}^{(t+1)} \right\|_F^2 = \mathbb{E} \left\| X^{(t+2/3)} (W_{S2S}^{(t)} - \Pi) \right\|_F^2 \geq 0.$$

For $t \in \mathcal{H}$, the disagreement error satisfies:

$$\begin{aligned} \mathbb{E} \left[\left\| X^{(t+1)} - \bar{X}^{(t+1)} \right\|_F^2 \right] &= \mathbb{E} \left[\left\| X^{(t+2/3)} (W_{S2S}^{(t)} - \Pi) \right\|_F^2 \right] \\ &= \frac{n-K}{n-1} \left\| X^{(t+2/3)} (I - \Pi) \right\|_F^2 \\ &= \frac{n-K}{n-1} \left\| X^{(t+2/3)} - \bar{X}^{(t+2/3)} \right\|_F^2. \end{aligned}$$

- For $K = 1$, the factor $\frac{n-K}{n-1}$ equals 1, indicating no contraction.
- For $K = n$, the factor $\frac{n-K}{n-1}$ equals 0, corresponding to full contraction.

Proof of Lemma 8 (iii). The result is a consequence of the variance of sampling without replacement:

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}_{\mathcal{S}^{(t)}} \left\| \mathbf{x}_i^{(t+1)} - \bar{\mathbf{x}}^{(t+1)} \right\|_2^2 &= \sum_{i=1}^n \mathbb{E}_{\mathcal{S}^{(t)}} \left\| \mathbb{1}_{i \in \mathcal{S}^{(t)}} \left[\frac{1}{K} \sum_{j \in \mathcal{S}^{(t)}} (\mathbf{x}_j^{(t+2/3)} - \bar{\mathbf{x}}^{(t+2/3)}) \right] + \mathbb{1}_{i \notin \mathcal{S}^{(t)}} (\mathbf{x}_i^{(t+2/3)} - \bar{\mathbf{x}}^{(t+2/3)}) \right\|_2^2 \\ &= K \mathbb{E}_{\mathcal{S}^{(t)}} \underbrace{\left\| \frac{1}{K} \sum_{j \in \mathcal{S}^{(t)}} (\mathbf{x}_j^{(t+2/3)} - \bar{\mathbf{x}}^{(t+2/3)}) \right\|_2^2}_{\text{bounded in Lemma 12}} + \frac{n-K}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(t+2/3)} - \bar{\mathbf{x}}^{(t+2/3)} \right\|_2^2 \\ &= \frac{n-K}{n(n-1)} \sum_{i=1}^n \left\| \mathbf{x}_i^{(t+2/3)} - \bar{\mathbf{x}}^{(t+2/3)} \right\|_2^2 + \frac{n-K}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(t+2/3)} - \bar{\mathbf{x}}^{(t+2/3)} \right\|_2^2 \\ &= \frac{n-K}{n-1} \sum_{i=1}^n \left\| \mathbf{x}_i^{(t+2/3)} - \bar{\mathbf{x}}^{(t+2/3)} \right\|_2^2. \end{aligned}$$

□

Intermediate Lemmas

Lemma 9 (S2S: Bias Error). *For every $t \geq 0$,*

$$\bar{X}^{(t+1)} = \bar{X}^{(t+\frac{2}{3})}.$$

Proof of Lemma 9. Let $\mathcal{H} := \{t \leq T \mid t \equiv 0 \pmod{H}\}$. We have:

- For D2D rounds ($t \notin \mathcal{H}$), by definition, $X^{(t+1)} = X^{(t+\frac{2}{3})}$ and $\bar{X}^{(t+1)} = \bar{X}^{(t+\frac{2}{3})}$.
- For D2S rounds ($t \in \mathcal{H}$), by Lemma 8 (ii), $\bar{X}^{(t+1)} = \bar{X}^{(t+\frac{2}{3})}$.

□

Lemma 10 (S2S: Disagreement Error). *For every $t \geq 0$,*

$$\mathbb{E} \|X^{(t+1)} - \bar{X}^{(t+1)}\|_F^2 = \begin{cases} \mathbb{E} \|X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})}\|_F^2, & t \notin \mathcal{H}; \\ \frac{n-K}{n-1} \mathbb{E} \|X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})}\|_F^2, & t \in \mathcal{H}, \end{cases}$$

where $\mathbb{E} \|X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})}\|_F^2$ is the D2D disagreement error already bounded in Lemma 5.

Proof of Lemma 10. We have:

- For D2D rounds ($t \notin \mathcal{H}$), by definition, $X^{(t+1)} = X^{(t+\frac{2}{3})}$.
- For D2S rounds ($t \in \mathcal{H}$), we apply Lemma 8 (iii).

□

Proof of Theorem 1

Proof of Theorem 1 (Convex Objectives).

Combine Lemma 3 and Lemma 9. For every $t \geq 0$,

$$\mathbb{E} \|\bar{x}^{(t+1)} - x^*\|_2^2 \leq \mathbb{E} \|\bar{x}^{(t)} - x^*\|_2^2 - \eta \mathbb{E} (f(\bar{x}^{(t)}) - f^*) + \frac{\eta^2 \bar{\sigma}^2}{n} + \frac{3\eta L}{2} \Xi^{(t)}.$$

Apply Lemma 7 with $r^{(t)} = \mathbb{E} \|\bar{x}^{(t+1)} - x^*\|_2^2$, $\Delta^{(t)} = \mathbb{E} (f(\bar{x}^{(t)}) - f^*)$, $a = b = e = \frac{3\eta L}{2}$, and $c = d = 0$:

$$\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} (f(\bar{x}^{(t)}) - f^*) \leq \frac{\|x^{(0)} - x^*\|^2}{\eta(T+1)} + \frac{\eta \bar{\sigma}^2}{n} + \frac{3L}{2} (\bar{\Xi}_{\text{intra}}^{(T+1)} + \bar{\Xi}_{\text{inter}}^{(T+1)}).$$

For the intra-component disagreement, apply Lemmas 5 and 10:

$$\bar{\Xi}_{\text{intra}}^{(t)} \leq \begin{cases} \left(1 - \frac{p}{4}\right) \Xi_{\text{intra}}^{(t-1)} + \frac{6\eta^2 \bar{\zeta}_{\text{intra}}^2}{p}, & t \notin \mathcal{H}; \\ \frac{n-K}{n-1} \left(1 - \frac{p}{4}\right) \Xi_{\text{intra}}^{(t-1)} + \frac{n-K}{n-1} \frac{6\eta^2 \bar{\zeta}_{\text{intra}}^2}{p}, & t \in \mathcal{H}. \end{cases}$$

For the recursion, apply Lemma 6 with $a_2 = 1 - \frac{p}{4} < 1$, $b_2 = \frac{6\eta^2 \bar{\zeta}_{\text{intra}}^2}{p}$, $a_1 = (1 - \frac{K-1}{n-1})a_2$, $b_1 = \frac{n-K}{n-1}b_2$.
For simplicity, define $\gamma := \frac{p}{4}$ and $\delta := \frac{n-K}{n-1}$, such that $a_2 = 1 - \gamma$, $a_1 = \delta a_2$, and $b_1 = \delta b_2$:

$$\begin{aligned}\bar{\Xi}_{\text{intra}}^{(T+1)} &\leq \left[\frac{a_1 b_2 (1 - a_2^{H-1})}{(1 - a_1 a_2^{H-1})(1 - a_2)^2} + \frac{b_1}{(1 - a_1 a_2^{H-1})(1 - a_2)} + \frac{b_2}{1 - a_2} (H-1) \right] \left(\frac{1}{H} + \frac{1}{T+1} \right) \\ &= \left[\frac{\delta a_2 b_2 (1 - a_2^{H-1})}{(1 - \delta a_2^H)(1 - a_2)^2} + \frac{\delta b_2}{(1 - \delta a_2^H)(1 - a_2)} + \frac{b_2}{1 - a_2} (H-1) \right] \left(\frac{1}{H} + \frac{1}{T+1} \right) \\ &= \left[\underbrace{\frac{\delta(1-\gamma)b_2[1-(1-\gamma)^{H-1}]}{[1-\delta(1-\gamma)^H]\gamma^2}}_{:=T_1} + \underbrace{\frac{\delta b_2}{[1-\delta(1-\gamma)^H]\gamma}}_{:=T_2} + \frac{b_2}{\gamma}(H-1) \right] \left(\frac{1}{H} + \frac{1}{T+1} \right).\end{aligned}$$

Using that $1 - (1 - \gamma)^{H-1} \leq (H-1)\gamma$ and that $1 - \delta(1 - \gamma)^H \geq 1 - \gamma$, we have $T_1 \leq \frac{\delta b_2(H-1)}{(1-\delta)\gamma}$ and $T_2 = \frac{\delta b_2}{(1-\delta)\gamma}$:

$$\begin{aligned}\bar{\Xi}_{\text{intra}}^{(T+1)} &\leq \frac{b_2}{\gamma} \left[(H-1) + \frac{\delta H}{(1-\delta)} \right] \left(\frac{1}{H} + \frac{1}{T+1} \right) \\ &\leq \frac{b_2}{\gamma} \frac{H}{1-\delta} \left(\frac{1}{H} + \frac{1}{T+1} \right) \\ &\leq \frac{n-1}{K-1} \frac{24\eta^2 \bar{\zeta}_{\text{intra}}^2}{p^2} \left(1 + \frac{H}{T+1} \right) \\ &\leq \frac{n-1}{K-1} \frac{48\eta^2 \bar{\zeta}_{\text{intra}}^2}{p^2},\end{aligned}$$

where, in the last inequality, we simplified the bound using that, for $T \geq H-1$, $\frac{H}{T+1} \leq 1$.

For the inter-component disagreement, combine Lemmas 5 and 10, with $\rho > 0$:

$$\Xi_{\text{inter}}^{(t)} \leq \begin{cases} (1+\rho)\Xi_{\text{inter}}^{(t-1)} + (1+\rho^{-1})\eta^2 \bar{\zeta}_{\text{inter}}^2, & t \notin \mathcal{H}; \\ \frac{n-K}{n-1}(1+\rho)\Xi_{\text{inter}}^{(t-1)} + \frac{n-K}{n-1}(1+\rho^{-1})\eta^2 \bar{\zeta}_{\text{inter}}^2, & t \in \mathcal{H}. \end{cases}$$

For the recursion, apply Lemma 6 with $a_2 = (1 + \rho) > 1$, $b_2 = (1 + \rho^{-1})\eta^2 \bar{\zeta}_{\text{inter}}^2$, $a_1 = \delta a_2$, $b_1 = \delta b_2$:

$$\begin{aligned}\bar{\Xi}_{\text{inter}}^{(T+1)} &\leq \left[\frac{a_1 b_2 (a_2^{H-1} - 1)(a_2^H - 1)}{(1 - a_1 a_2^{H-1})(a_2 - 1)^2} + \frac{b_1 (a_2^H - 1)}{(1 - a_1 a_2^{H-1})(a_2 - 1)} + \frac{b_2 (a_2^H - a_2 H + H - 1)}{(a_2 - 1)^2} \right] \left(\frac{1}{H} + \frac{1}{T+1} \right) \\ &\leq \left[\frac{\delta a_2 b_2 (a_2^{H-1} - 1)(a_2^H - 1)}{(1 - \delta a_2^H)(a_2 - 1)^2} + \frac{\delta b_2 (a_2^H - 1)}{(1 - \delta a_2^H)(a_2 - 1)} + \frac{b_2 (a_2^H - a_2 H + H - 1)}{(a_2 - 1)^2} \right] \left(\frac{1}{H} + \frac{1}{T+1} \right) \\ &= \left[\underbrace{\frac{\delta(1+\rho)(1+\rho^{-1})[(1+\rho)^{H-1}-1][(1+\rho)^H-1]}{[1-\delta(1+\rho)^H]\rho^2}}_{:=T_3} + \underbrace{\frac{\delta(1+\rho^{-1})[(1+\rho)^H-1]}{[1-\delta(1+\rho)^H]\rho}}_{:=T_4} \right. \\ &\quad \left. + \underbrace{\frac{(1+\rho^{-1})[(1+\rho)^H-(1+\rho)H+H-1]}{\rho^2}}_{:=T_5} \right] \eta^2 \bar{\zeta}_{\text{inter}}^2 \left(\frac{1}{H} + \frac{1}{T+1} \right).\end{aligned}$$

We choose $\rho = \frac{1-\delta}{2H}$, such that $C = a_1 a_2^{H-1} = \frac{n-K}{n-1}(1+\rho)^H < 1$ in Lemma 6.

As a consequence, we have: $T_3 \leq \frac{54\delta H^2(H-1)}{(1-\delta)^2}$, $T_4 \leq \frac{12\delta H^2}{(1-\delta)^2}$, and $T_5 \leq \frac{4H^2(H-1)}{1-\delta}$:

$$\begin{aligned}\bar{\Xi}_{\text{inter}}^{(T+1)} &\leq 70\eta^2\bar{\zeta}_{\text{inter}}^2 \frac{\delta H^2(H-1)}{(1-\delta)^2} \left(\frac{1}{H} + \frac{1}{T+1} \right) \\ &\leq \left(\frac{n-1}{K-1} \right)^2 70\eta^2\bar{\zeta}_{\text{inter}}^2 \left(H(H-1) + \frac{H^2(H-1)}{T+1} \right) \\ &\leq \left(\frac{n-1}{K-1} \right)^2 140\eta^2\bar{\zeta}_{\text{inter}}^2 H(H-1),\end{aligned}$$

where, in the last inequality, we again simplified the bound using that, for $T \geq H-1$, $\frac{H}{T+1} \leq 1$.

Replace $\bar{\Xi}_{\text{intra}}^{(T+1)}$ and $\bar{\Xi}_{\text{inter}}^{(T+1)}$ in the bound:

$$\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left(f(\bar{\mathbf{x}}^{(t)}) - f^* \right) \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2}{\eta(T+1)} + \frac{\eta\bar{\sigma}^2}{n} + \frac{n-1}{K-1} \frac{72\eta^2 L \bar{\zeta}_{\text{intra}}^2}{p^2} + \left(\frac{n-1}{K-1} \right)^2 210\eta^2 LH(H-1)\bar{\zeta}_{\text{inter}}^2.$$

Finally, apply (Koloskova et al. 2020, Lemmas 16 and 17) with $r_t = \mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|_2^2$, $e_t = \mathbb{E}(f(\bar{\mathbf{x}}^{(t)}) - f^*)$, $a = 0$, $b = 1$, $c = \frac{\bar{\sigma}^2}{n}$, $d = \frac{8L}{p}$, and $e = \frac{n-1}{K-1} \frac{72L\bar{\zeta}_{\text{intra}}^2}{p^2} + \left(\frac{n-1}{K-1} \right)^2 210LH(H-1)\bar{\zeta}_{\text{inter}}^2$. \square

Proof of Theorem 2 (Non-Convex Objectives). Combine Lemma 4 and Lemma 9 ($\|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}\|_2^2 = 0$).

For every $t \geq 0$:

$$\mathbb{E}[f(\bar{\mathbf{x}}^{(t+1)})] \leq \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})] - \frac{\eta}{4} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 + \frac{\eta^2 L \bar{\sigma}^2}{2n} + \eta L^2 \Xi^{(t)}.$$

Apply Lemma 7 with $r^{(t)} = \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})]$, $\Delta^{(t)} = \frac{1}{4}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2$, $a = b = e = \eta L^2$, and $c = d = 0$:

$$\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 \leq \frac{4(f(\bar{\mathbf{x}}^{(0)}) - f^*)}{\eta(T+1)} + \frac{2\eta L \bar{\sigma}^2}{n} + 4L^2 \left(\bar{\Xi}_{\text{intra}}^{(T+1)} + \bar{\Xi}_{\text{inter}}^{(T+1)} \right).$$

Replace the values $\bar{\Xi}_{\text{intra}}^{(T+1)} \leq \frac{n-1}{K-1} \frac{48\eta^2 \bar{\zeta}_{\text{intra}}^2}{p^2}$, $\bar{\Xi}_{\text{intra}}^{(T+1)} \leq \left(\frac{n-1}{K-1} \right)^2 140\eta^2 H(H-1)\bar{\zeta}_{\text{inter}}^2$ found previously:

$$\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 \leq \frac{4(f(\bar{\mathbf{x}}^{(0)}) - f^*)}{\eta(T+1)} + \frac{2\eta L \bar{\sigma}^2}{n} + \frac{n-1}{K-1} \frac{192\eta^2 L^2 \bar{\zeta}_{\text{intra}}^2}{p^2} + \left(\frac{n-1}{K-1} \right)^2 560\eta^2 L^2 H(H-1)\bar{\zeta}_{\text{inter}}^2.$$

Apply (Koloskova et al. 2020, Lemmas 16 and 17) with $r_t = \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})]$, $e_t = \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2$,

$a = 0$, $b = 1$, $c = \frac{2L\bar{\sigma}^2}{n}$, $d = \frac{8L}{p}$, and $e = \frac{n-1}{K-1} \frac{192L^2 \bar{\zeta}_{\text{intra}}^2}{p^2} + \left(\frac{n-1}{K-1} \right)^2 560L^2 H(H-1)\bar{\zeta}_{\text{inter}}^2$. \square

C Convergence Analysis of S2A

Properties of S2A

Lemma 11 (Sampled-to-All). Define:

$$(W_{S2A}^{(t)})_{ij} = \begin{cases} \frac{1}{K}, & i \in \mathcal{S}^{(t)}; \\ 0, & \text{otherwise.} \end{cases}$$

The matrix $W_{S2A}^{(t)}$ satisfies the following properties:

(i) **Column-stochastic, not row-stochastic.** $\mathbf{1}^\top W_{S2A}^{(t)} = \mathbf{1}^\top$, whereas $W_{S2A}^{(t)} \mathbf{1} = \frac{K}{n} \mathbf{1}_{\{i \in \mathcal{S}\}}$.

Consequently, $W_{S2A}^{(t)} \Pi = W_{S2A}^{(t)}$, and $\Pi W_{S2A}^{(t)} = \Pi$.

(ii) **Bias error.** $W_{S2A}^{(t)}$ does not preserve the average of the iterates between D2D and D2S rounds:

$$\bar{X}^{(t+1)} = X^{(t+\frac{2}{3})} W_{S2A}^{(t)} \Pi = X^{(t+\frac{2}{3})} W_{S2A}^{(t)} \neq X^{(t+\frac{2}{3})} \Pi =: \bar{X}^{(t+\frac{2}{3})}.$$

We define the bias error as:

$$\mathbb{E} \left\| \bar{X}^{(t+1)} - \bar{X}^{(t+\frac{2}{3})} \right\|_F^2 = \mathbb{E} \left\| X^{(t+\frac{2}{3})} (W_{S2A}^{(t)} - \Pi) \right\|_F^2 \geq 0.$$

For $t \in \mathcal{H}$, the bias error satisfies:

$$\begin{aligned} \mathbb{E} \left[\left\| \bar{X}^{(t+1)} - \bar{X}^{(t+\frac{2}{3})} \right\|_F^2 \right] &= \mathbb{E} \left[\left\| X^{(t+\frac{2}{3})} (W_{S2A}^{(t)} - \Pi) \right\|_F^2 \right] \\ &= \frac{n-K}{K(n-1)} \mathbb{E} \left[\left\| X^{(t+\frac{2}{3})} (I - \Pi) \right\|_F^2 \right] \\ &= \frac{n-K}{K(n-1)} \mathbb{E} \left[\left\| X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})} \right\|_F^2 \right]. \end{aligned}$$

- For $K = 1$, the factor $\frac{n-K}{K(n-1)}$ equals 1, indicating no contraction.
- For $K = n$, the factor $\frac{n-K}{K(n-1)}$ equals 0, corresponding to full contraction.

(iii) **The global average is unbiased in expectation.** Since $\mathbb{E}_{\mathcal{S}^{(t)}}[W_{S2A}^{(t)}] = \Pi$,

$$\mathbb{E}_{\mathcal{S}^{(t)}}[\bar{X}^{(t+1)}] = \mathbb{E}_{\mathcal{S}^{(t)}}[X^{(t+\frac{2}{3})} W_{S2A}^{(t)} \Pi] = X^{(t+\frac{2}{3})} \mathbb{E}_{\mathcal{S}^{(t)}}[W_{S2A}^{(t)}] \Pi = X^{(t+\frac{2}{3})} \Pi^2 = X^{(t+\frac{2}{3})} \Pi =: \bar{X}^{(t+\frac{2}{3})}.$$

In other words, W_{S2A} propagates the average of the iterates in expectation.

(iv) **The disagreement error is zero.** For $t \in \mathcal{H}$,

$$X^{(t+1)} - \bar{X}^{(t+1)} = X^{(t+2/3)} W_{S2A}^{(t)} - X^{(t+2/3)} W_{S2A}^{(t)} \Pi = X^{(t+2/3)} W_{S2A}^{(t)} - X^{(t+2/3)} W_{S2A}^{(t)} = 0.$$

Proof of Lemma 11 (ii).

This result can be derived from the variance of sampling without replacement (Jhunjhunwala et al. 2022, Lemma 4):

$$\begin{aligned} &\mathbb{E}_{\mathcal{S}^{(t)}} \left\| \bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right\|_2^2 \\ &= \mathbb{E}_{\mathcal{S}^{(t)}} \left\| \frac{1}{K} \sum_{i \in \mathcal{S}^{(t)}} \left(\mathbf{x}_i^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right) \right\|_2^2 \\ &= \mathbb{E}_{\mathcal{S}^{(t)}} \left\| \frac{1}{K} \sum_{i=1}^n \mathbb{1}_{i \in \mathcal{S}^{(t)}} \left(\mathbf{x}_i^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right) \right\|_2^2 \\ &= \frac{1}{K^2} \mathbb{E}_{\mathcal{S}^{(t)}} \left[\sum_{i=1}^n [\mathbb{1}_{i \in \mathcal{S}^{(t)}}]^2 \left\| \mathbf{x}_i^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right\|_2^2 + \sum_{j=1}^n \sum_{\substack{i=1 \\ i \neq j}}^n \mathbb{1}_{i \in \mathcal{S}^{(t)}} \mathbb{1}_{j \in \mathcal{S}^{(t)}} \langle \mathbf{x}_i^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}, \mathbf{x}_j^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \rangle \right] \\ &= \frac{1}{K^2} \sum_{i=1}^n \frac{K}{n} \left\| \mathbf{x}_i^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right\|_2^2 + \frac{1}{K^2} \sum_{i \neq j} \frac{K}{n} \frac{K-1}{n-1} \langle \mathbf{x}_i^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})}, \mathbf{x}_j^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \rangle \\ &= \frac{1}{K^2} \sum_{i=1}^n \left\| \mathbf{x}_i^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right\|_2^2 \underbrace{\left[\frac{K}{n} - \frac{K(K-1)}{n(n-1)} \right]}_{=0} + \frac{1}{K^2} \frac{K(K-1)}{n(n-1)} \underbrace{\left\| \sum_{i=1}^n \left(\mathbf{x}_i^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right) \right\|_2^2}_{=0} \\ &= \frac{(n-K)}{K(n-1)} \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(t+\frac{2}{3})} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right\|_2^2 = \frac{(n-K)}{K(n-1)} \frac{1}{n} \left\| X^{(t)} - \bar{X}^{(t)} \right\|_F^2. \end{aligned}$$

□

Intermediate Lemmas

Lemma 12 (S2A: Bias Error). *For every $t \geq 0$,*

$$\mathbb{E} \left\| \bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right\|_2^2 = \begin{cases} \frac{n-K}{K(n-1)} \frac{1}{n} \mathbb{E} \left\| X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})} \right\|_F^2, & t \in \mathcal{H}, \\ 0, & t \notin \mathcal{H}, \end{cases}$$

where $\mathbb{E} \|X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})}\|_F^2$ is the D2D disagreement error already bounded in Lemma 5.

Proof of Lemma 12. We have:

- For D2D rounds ($t \notin \mathcal{H}$), by definition, $\bar{\mathbf{x}}^{(t+1)} = \bar{\mathbf{x}}^{(t+\frac{2}{3})}$.
- For D2S rounds ($t \in \mathcal{H}$), by Lemma 11 (ii):

$$\mathbb{E} \left\| \bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right\|_2^2 = \frac{1}{n} \mathbb{E} \left\| \bar{X}^{(t+1)} - \bar{X}^{(t+\frac{2}{3})} \right\|_F^2 = \frac{n-K}{K(n-1)} \frac{1}{n} \mathbb{E} \left\| X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})} \right\|_F^2.$$

□

Lemma 13 (S2A: Disagreement Error). *For every $t \geq 0$,*

$$\left\| X^{(t+1)} - \bar{X}^{(t+1)} \right\|_F^2 = \begin{cases} \left\| X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})} \right\|_F^2, & t \notin \mathcal{H}; \\ 0, & t \in \mathcal{H}, \end{cases}$$

where $\mathbb{E} \|X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})}\|_F^2$ is the D2D disagreement error already bounded in Lemma 5.

Proof of Lemma 13. We have:

- For D2D rounds ($t \notin \mathcal{H}$), by definition, $X^{(t+1)} = X^{(t+\frac{2}{3})}$ and $\bar{X}^{(t+1)} = \bar{X}^{(t+\frac{2}{3})}$.
- For D2S rounds ($t \in \mathcal{H}$), by Lemma 11 (iv), $X^{(t+1)} = \bar{X}^{(t+1)}$.

□

Proof of Theorem 2

Proof of Theorem 2 (Convex Objectives). Combine Lemmas 3, 5, 12, and 13:

- If $t \in \mathcal{H}$:

$$\begin{aligned} \mathbb{E} \left\| \bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^* \right\|_2^2 &\leq \mathbb{E} \left\| \bar{\mathbf{x}}^{(t)} - \mathbf{x}^* \right\|_2^2 - \eta \mathbb{E} \left(f(\bar{\mathbf{x}}^{(t)}) - f^* \right) + \frac{\eta^2 \bar{\sigma}^2}{n} \\ &\quad + \left[\frac{3\eta L}{2} + \frac{(n-K)}{K(n-1)} \left(1 - \frac{p}{4} \right) \right] \Xi_{\text{intra}}^{(t)} \\ &\quad + \left[\frac{3\eta L}{2} + \frac{(n-K)}{K(n-1)} (1 + \rho) \right] \Xi_{\text{inter}}^{(t)} \\ &\quad + \frac{(n-K)}{K(n-1)} \frac{6\eta^2 \bar{\zeta}_{\text{intra}}^2}{p} + \frac{(n-K)}{K(n-1)} (1 + \rho^{-1}) \eta^2 \bar{\zeta}_{\text{inter}}^2. \end{aligned}$$

- If $t \notin \mathcal{H}$:

$$\begin{aligned} \mathbb{E} \left\| \bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^* \right\|_2^2 &\leq \mathbb{E} \left\| \bar{\mathbf{x}}^{(t)} - \mathbf{x}^* \right\|_2^2 - \eta \mathbb{E} \left(f(\bar{\mathbf{x}}^{(t)}) - f^* \right) + \frac{\eta^2 \bar{\sigma}^2}{n} \\ &\quad + \frac{3\eta L}{2} \Xi_{\text{intra}}^{(t)} + \frac{3\eta L}{2} \eta \Xi_{\text{inter}}^{(t)}. \end{aligned}$$

Apply Lemma 7 with $r^{(t)} := \mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|_2^2$, $\Delta^{(t)} := \mathbb{E}(f(\bar{\mathbf{x}}^{(t)}) - f^*)$, $a := \frac{3\eta L}{2} + \frac{(n-K)}{K(n-1)}(1 - \frac{p}{4})$, $b := \frac{3\eta L}{2} + \frac{(n-K)}{K(n-1)}(1 + \rho)$, $c := \frac{n-K}{K(n-1)}\frac{6\eta^2\bar{\zeta}_{\text{intra}}^2}{p}$, $d := \frac{n-K}{K(n-1)}(1 + \rho^{-1})\eta^2\bar{\zeta}_{\text{inter}}^2$, and $e := \frac{3\eta L}{2}$:

$$\begin{aligned} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left(f(\bar{\mathbf{x}}^{(t)}) - f^* \right) &\leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2}{\eta(T+1)} + \frac{\eta\bar{\sigma}^2}{n} + \frac{2\eta L + \frac{n-K}{K(n-1)}\frac{1-\frac{p}{4}}{H}\bar{\Xi}_{\text{intra}}^{(T+1)}}{\eta} + \frac{2\eta L + \frac{n-K}{K(n-1)}\frac{1+\rho}{H}\bar{\Xi}_{\text{inter}}^{(T+1)}}{\eta} \\ &\quad + \frac{n-K}{K(n-1)}\frac{6\eta\bar{\zeta}_{\text{intra}}^2}{Hp} + \frac{n-K}{K(n-1)}\frac{1+\rho^{-1}}{H}\eta\bar{\zeta}_{\text{inter}}^2. \end{aligned}$$

For the intra-component disagreement, combine Lemmas 5 and 13:

$$\Xi_{\text{intra}}^{(t)} \leq \begin{cases} 0, & t \in \mathcal{H}, \\ \left(1 - \frac{p}{4}\right)\Xi_{\text{intra}}^{(t-1)} + \frac{6\eta^2\bar{\zeta}_{\text{intra}}^2}{p}, & t \notin \mathcal{H}. \end{cases}$$

For the recursion, apply Lemma 6 with $a_1 = b_1 = 0$, $a_2 = 1 - \frac{p}{4} < 1$ and $b_2 = \frac{6\eta^2\bar{\zeta}_{\text{intra}}^2}{p}$:

$$\bar{\Xi}_{\text{intra}}^{(T+1)} \leq \frac{24\eta^2\bar{\zeta}_{\text{intra}}^2}{p^2} \left(\frac{H-1}{H} + \frac{H-1}{T+1} \right) \leq \frac{48\eta^2\bar{\zeta}_{\text{intra}}^2}{p^2},$$

where, to simplify the bound, we used that for $T \geq H-1$, $\frac{1}{T+1} \leq \frac{1}{H}$, and that $\frac{H-1}{H} < 1$.

For the inter-component disagreement, combine Lemmas 5 and 13, with $\rho > 0$:

$$\Xi_{\text{inter}}^{(t)} \leq \begin{cases} 0, & t \in \mathcal{H}, \\ (1+\rho)\Xi_{\text{inter}}^{(t-1)} + (1+\rho^{-1})\eta^2\bar{\zeta}_{\text{inter}}^2, & t \notin \mathcal{H}. \end{cases}$$

For the recursion, apply Lemma 6 with $a_1 = b_1 = 0$, $a_2 = 1 + \rho > 1$ and $b_2 = (1 + \rho^{-1})\eta^2\bar{\zeta}_{\text{inter}}^2$:

$$\begin{aligned} \bar{\Xi}_{\text{inter}}^{(T+1)} &\leq \eta^2\bar{\zeta}_{\text{inter}}^2(1 + \rho^{-1}) \frac{(1+\rho)^H - (1+\rho)H + H - 1}{\rho^2} \left(\frac{1}{H} + \frac{1}{T+1} \right) \\ &\leq 4\eta^2\bar{\zeta}_{\text{inter}}^2 \left(H(H-1) + \frac{H^2(H-1)}{T+1} \right) \\ &\leq 8\eta^2\bar{\zeta}_{\text{inter}}^2 H(H-1). \end{aligned}$$

In the second inequality, we chose $\rho = \frac{2}{H}$, such that $F_H(\rho) := \frac{(1+\rho)^H - (1+\rho)H + H - 1}{\rho^2} = \sum_{k=2}^H \binom{H}{k} \rho^{k-2} \leq 2H(H-1)$, $(1+\rho^{-1}) = 1 + \frac{H}{2} \leq 2H$, $(1+\rho) = 1 + \frac{2}{H} \leq 3$, $(1+\rho^{-1})F_H(\rho) \leq 4H^2(H-1)$, and $(1+\rho)(1+\rho^{-1})F_H(\rho) \leq 12H^2(H-1)$.

In the last inequality, to simplify the bound, we used again that for $T \geq H-1$, $\frac{1}{T+1} \leq \frac{1}{H}$.

Replace $\bar{\Xi}_{\text{intra}}^{(T+1)}$ and $\bar{\Xi}_{\text{inter}}^{(T+1)}$ in the bound:

$$\begin{aligned} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left(f(\bar{\mathbf{x}}^{(t)}) - f^* \right) &\leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2}{\eta(T+1)} + \frac{\eta\bar{\sigma}^2}{n} \\ &\quad + \frac{96\eta^2L\bar{\zeta}_{\text{intra}}^2}{p^2} + \frac{n-K}{K(n-1)} \left(1 - \frac{p}{4}\right) \frac{48\eta\bar{\zeta}_{\text{intra}}^2}{p^2} \left(\frac{H-1}{H^2}\right) + \frac{n-K}{K(n-1)} \frac{6\eta\bar{\zeta}_{\text{intra}}^2}{Hp} \\ &\quad + 16\eta^2L\bar{\zeta}_{\text{inter}}^2 H(H-1) + \frac{n-K}{K(n-1)} 12\eta\bar{\zeta}_{\text{inter}}^2 \left(H-1 + \frac{H-1}{H}\right) + \frac{n-K}{K(n-1)} 2\eta\bar{\zeta}_{\text{inter}}^2. \end{aligned}$$

To simplify the bound, use that $p \leq 1$, therefore $(1 - \frac{p}{4}) < 1$, and that $\frac{H-1}{H} \leq 1$:

$$\begin{aligned} \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left(f(\bar{\mathbf{x}}^{(t)}) - f^* \right) &\leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2}{\eta(T+1)} + \frac{\eta\bar{\sigma}^2}{n} \\ &\quad + \frac{96\eta^2L\bar{\zeta}_{\text{intra}}^2}{p^2} + \frac{n-K}{K(n-1)} \frac{54\eta\bar{\zeta}_{\text{intra}}^2}{Hp^2} + 16\eta^2LH^2\bar{\zeta}_{\text{inter}}^2 + \frac{n-K}{K(n-1)} 26\eta H\bar{\zeta}_{\text{inter}}^2. \end{aligned}$$

Finally, apply (Koloskova et al. 2020, Lemmas 16 and 17) with $r_t = \mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|_2^2$, $e_t = \mathbb{E}(f(\bar{\mathbf{x}}^{(t)}) - f^*)$, $a = 0$, $b = 1$, $c = \frac{\bar{\sigma}^2}{n} + \frac{n-K}{K(n-1)} 54\frac{\bar{\zeta}_{\text{intra}}^2}{Hp^2} + \frac{n-K}{K(n-1)} 26H\bar{\zeta}_{\text{inter}}^2$, $d = \frac{8L}{p}$, and $e = \frac{96L\bar{\zeta}_{\text{intra}}^2}{p^2} + 16LH^2\bar{\zeta}_{\text{inter}}^2$. \square

Proof of Theorem 1 (Non-Convex Objectives). Combine Lemmas 4, 5, 12, and 13:

- If $t \in \mathcal{H}$:

$$\begin{aligned}
& \mathbb{E}[f(\bar{\mathbf{x}}^{(t+1)})] \leq \\
& \leq \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})] - \frac{\eta}{4} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 + \frac{\eta^2 L \bar{\sigma}^2}{2n} + \eta L^2 \frac{1}{n} \mathbb{E} \left\| X^{(t)} - \bar{X}^{(t)} \right\|_F^2 + \frac{L}{2} \mathbb{E} \left\| \bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t+\frac{2}{3})} \right\|_2^2 \\
& \leq \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})] - \frac{\eta}{4} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 + \frac{\eta^2 L \bar{\sigma}^2}{2n} + \eta L^2 \frac{1}{n} \mathbb{E} \left\| X^{(t)} - \bar{X}^{(t)} \right\|_F^2 + \frac{L}{2} \frac{(n-K)}{K(n-1)} \frac{1}{n} \mathbb{E} \left\| X^{(t+\frac{2}{3})} - \bar{X}^{(t+\frac{2}{3})} \right\|_F^2 \\
& \leq \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})] - \frac{\eta}{4} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 + \frac{\eta^2 L \bar{\sigma}^2}{2n} + \frac{L}{2} \left[2\eta L + \frac{(n-K)}{K(n-1)} \left(1 - \frac{p}{4} \right) \right] \Xi_{\text{intra}}^{(t)} \\
& \quad + \frac{L}{2} \left[2\eta L + \frac{(n-K)}{K(n-1)} (1 + \rho) \right] \Xi_{\text{inter}}^{(t)} + \frac{(n-K)}{K(n-1)} \frac{3\eta^2 L \bar{\zeta}_{\text{intra}}^2}{p} + \frac{(n-K)}{K(n-1)} (1 + \rho^{-1}) \frac{L\eta^2 \bar{\zeta}_{\text{inter}}^2}{2}.
\end{aligned}$$

- If $t \notin \mathcal{H}$:

$$\mathbb{E}[f(\bar{\mathbf{x}}^{(t+1)})] \leq \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})] - \frac{\eta}{4} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 + \frac{\eta^2 L \bar{\sigma}^2}{2n} + \eta L^2 \Xi_{\text{intra}}^{(t)} + \eta L^2 \Xi_{\text{inter}}^{(t)}.$$

Apply Lemma 7 with $r^{(t)} = \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})]$, $\Delta^{(t)} = \frac{1}{4} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2$, $a = \frac{L}{2} \left[2\eta L + \frac{(n-K)}{K(n-1)} \left(1 - \frac{p}{4} \right) \right]$, $b = \frac{L}{2} \left[2\eta L + \frac{(n-K)}{K(n-1)} (1 + \rho) \right]$, $c = \frac{(n-K)}{K(n-1)} \frac{3\eta^2 L \bar{\zeta}_{\text{intra}}^2}{p}$, $d = \frac{(n-K)}{K(n-1)} (1 + \rho^{-1}) \frac{L\eta^2 \bar{\zeta}_{\text{inter}}^2}{2}$, and $e = \eta L^2$:

$$\begin{aligned}
\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 & \leq \frac{4(f(\bar{\mathbf{x}}^{(0)}) - f^*)}{\eta(T+1)} + \frac{2\eta L \bar{\sigma}^2}{n} \\
& \quad + \frac{2L}{\eta} \left[2\eta L + \frac{(n-K)}{K(n-1)} \frac{1 - \frac{p}{4}}{H} \right] \Xi_{\text{intra}}^{(T+1)} + \frac{2L}{\eta} \left[2\eta L + \frac{(n-K)}{K(n-1)} \frac{1 + \rho}{H} \right] \Xi_{\text{inter}}^{(T+1)} \\
& \quad + \frac{(n-K)}{K(n-1)} \frac{12\eta L \bar{\zeta}_{\text{intra}}^2}{Hp} + \frac{(n-K)}{K(n-1)} \frac{1 + \rho^{-1}}{H} 2\eta L \bar{\zeta}_{\text{inter}}^2.
\end{aligned}$$

Replace the values $\Xi_{\text{intra}}^{(T+1)} \leq \frac{48\eta^2 \bar{\zeta}_{\text{intra}}^2}{p^2}$, $\Xi_{\text{inter}}^{(T+1)} \leq 8\eta^2 \bar{\zeta}_{\text{inter}}^2 H(H-1)$, $1 + \rho \leq 3$, $1 + \rho^{-1} \leq 2H$ found previously, and simplify again using $p \leq 1$, therefore $(1 - \frac{p}{4}) < 1$, and $\frac{H-1}{H} \leq 1$:

$$\begin{aligned}
\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 & \leq \frac{4(f(\bar{\mathbf{x}}^{(0)}) - f^*)}{\eta(T+1)} + \frac{2\eta L \bar{\sigma}^2}{n} \\
& \quad + \frac{192\eta^2 L^2 \bar{\zeta}_{\text{intra}}^2}{p^2} + \frac{n-K}{K(n-1)} \frac{108\eta L \bar{\zeta}_{\text{intra}}^2}{Hp^2} + 32\eta^2 L^2 H^2 \bar{\zeta}_{\text{inter}}^2 + \frac{n-K}{K(n-1)} 52\eta L H \bar{\zeta}_{\text{inter}}^2.
\end{aligned}$$

Finally, apply (Koloskova et al. 2020, Lemmas 16 and 17) with $r^{(t)} = \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})]$, $\Delta^{(t)} = \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2$, $a = 0$, $b = 1$, $c = \frac{2L\bar{\sigma}^2}{n} + \frac{n-K}{K(n-1)} \frac{108L\bar{\zeta}_{\text{intra}}^2}{Hp^2} + \frac{n-K}{K(n-1)} 52LH\bar{\zeta}_{\text{inter}}^2$, $d = \frac{8L}{p}$, and $e = \frac{192L^2\bar{\zeta}_{\text{intra}}^2}{p^2} + 32L^2H^2\bar{\zeta}_{\text{inter}}^2$. \square

D Additional Theoretical Results

Mixing Parameter for Multi-Component Communication Graphs

Lemma 14 (Mixing parameter for a block-diagonal communication matrix). *Let $W \in \mathbb{R}^{n \times n}$ be a block-diagonal matrix with C blocks, where each block $W_c \in \mathbb{R}^{n_c \times n_c}$ corresponds to the D2D mixing matrix of component c .*

For each c , define the local averaging projector $\Pi_c := \frac{1}{n_c} \mathbf{1} \mathbf{1}^\top \in \mathbb{R}^{n_c \times n_c}$ and the local mixing parameter $p_c \in (0, 1]$ such that:

$$\|W_c - \Pi_c\|_F^2 \leq (1 - p_c) \|I_{n_c} - \Pi_c\|_F^2.$$

For fixed W_c , one can take $p_c = 1 - \lambda_2(W_c^\top W_c)$ (Boyd et al. 2006).

Then, the matrix W satisfies:

$$\|W - \Pi_C\|_F^2 \leq (1 - p) \|I_n - \Pi_C\|_F^2,$$

with mixing parameter $p \geq \frac{\sum_{c=1}^C p_c(n_c - 1)}{\sum_{c=1}^C (n_c - 1)}$.

Proof. Because W and Π_C are block-diagonal, the Frobenius norm decomposes over components:

$$\|W - \Pi_C\|_F^2 = \sum_{c=1}^C \|W_c - \Pi_c\|_F^2 \leq \sum_{c=1}^C (1 - p_c) \|I_{n_c} - \Pi_c\|_F^2 = \sum_{c=1}^C (1 - p_c)(n_c - 1),$$

where the rightmost equality uses $\|I_{n_c} - \Pi_c\|_F^2 = n_c - 1$.

For the same argument,

$$\|I_n - \Pi_C\|_F^2 = \sum_{c=1}^C \|I_{n_c} - \Pi_c\|_F^2 = \sum_{c=1}^C (n_c - 1).$$

Combining:

$$p \geq 1 - \frac{\|W - \Pi_C\|_F^2}{\|I_n - \Pi_C\|_F^2} \geq 1 - \frac{\sum_{c=1}^C (1 - p_c)(n_c - 1)}{\sum_{c=1}^C (n_c - 1)} = \frac{\sum_{c=1}^C p_c(n_c - 1)}{\sum_{c=1}^C (n_c - 1)}.$$

□

We observe that $p = \frac{\sum_{c=1}^C p_c(n_c - 1)}{\sum_{c=1}^C (n_c - 1)} \geq p_{\min} := \min_{1 \leq c \leq C} p_c = 1 - \lambda_{C+1}(W^\top W)$,

where $\lambda_{C+1}(W^\top W)$ denotes the largest eigenvalue of $W^\top W$ strictly below 1 .

Extension to Random Mixing Matrices

As we mentioned in Section 5, our theoretical results extend to random mixing matrices, following the approach in (Koloskova et al. 2020; Le Bars et al. 2023). At each round t of Algorithm 1, the mixing matrix $W^{(t)}$ is sampled from a distribution $\mathcal{W}^{(t)}$, independent of the iterates $\mathbf{x}^{(t)}$, and possibly time-varying.

To analyze convergence in this setting, we modify Lemma 2 and Assumption 5 by taking expectation with respect to W . Specifically, $\mathbb{E}_W \|X(W - \Pi_C)\|_F^2 \leq (1 - p) \|X(I - \Pi_C)\|_F^2$ and $\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W, \xi} \|\sum_{j=1}^n (W - \Pi_C)_{ij} \nabla F_j(\mathbf{x}, \xi)\|_2^2 \leq \bar{\zeta}_{\text{intra}}^2$. Theorems 1 and 2 then hold under the assumption that the distributions $\mathcal{W}^{(0)}, \dots, \mathcal{W}^{(T)}$ satisfy these conditions.

The proof follows exactly the same steps as in the deterministic case, by appropriately conditioning on the realization of the random mixing matrices and on the iterates (Koloskova et al. 2020; Le Bars et al. 2023).

Iteration vs. Communication Complexity

While Section 5.3 compares S2S and S2A in terms of iteration complexity, we now analyze their total communication cost, measured as the number of messages required to reach a target accuracy $\epsilon \geq 0$.

Per-round communication cost. At each D2S round $t \in \mathcal{H}$, both primitives involve uplink (device-to-server) and downlink (server-to-device) communication. The number of messages exchanged per round is:

Primitive	uplinks	downlinks
S2S	K	K
S2A	K	n

In typical federated learning settings, where uplink cost dominates, both primitives incur a per-round communication cost of K messages, making iteration complexity a reasonable proxy for communication cost, and our theoretical comparison of S2S and S2A in Section 5.3 remains valid. However, when downlink cost is not negligible, S2S incurs a lower communication cost, saving $n - K$ downlink messages per server round, and we could keep this into account for the comparison.

Total communication cost. Given the iteration complexities $T_{\text{S2S}}(\epsilon)$ and $T_{\text{S2A}}(\epsilon)$ from Theorems 1 and 2, and defining the number of server rounds as $R := \lceil \frac{T}{H} \rceil$, the total number of messages exchanged by S2S and S2A to reach the accuracy ϵ are:

$$\Gamma_{\text{S2S}}(\epsilon) = 2KR_{\text{S2S}}(\epsilon) = \frac{2K}{H}T_{\text{S2S}}(\epsilon),$$

$$\Gamma_{\text{S2A}}(\epsilon) = (K + n)R_{\text{S2A}}(\epsilon) = \frac{K + n}{H}T_{\text{S2A}}(\epsilon).$$

The communication cost ratio is:

$$\frac{\Gamma_{\text{S2A}}(\epsilon)}{\Gamma_{\text{S2S}}(\epsilon)} = \frac{K + n}{2K} \frac{T_{\text{S2A}}(\epsilon)}{T_{\text{S2S}}(\epsilon)}.$$

Interestingly, the qualitative regimes follow those in Section 5.3:

- R1. $\bar{\zeta}_{\text{intra}}, \bar{\zeta}_{\text{inter}}$ are low:** S2A converges faster than S2S for high sampling rates ($K/n \geq 0.1$, Fig. 6(a)), most server periods (Fig. 6(e)), and higher D2D network connectivities ($p \geq 0.3$, Fig. 6(i)).
- R2. $\bar{\zeta}_{\text{inter}} \ll \bar{\zeta}_{\text{intra}}$:** S2S converges faster for most sampling rates (Fig. 6(b)), low server periods ($H \leq 10$, Fig 6(f)), and for most mixing parameters (Fig 6(j)); S2A converges faster otherwise.
- R3. $\bar{\zeta}_{\text{inter}}$ is high:** S2S converges faster for most values of K/n , H , and p , irrespective of $\bar{\zeta}_{\text{intra}}$ (Figs. 6(c,d,g,h,k,l)).

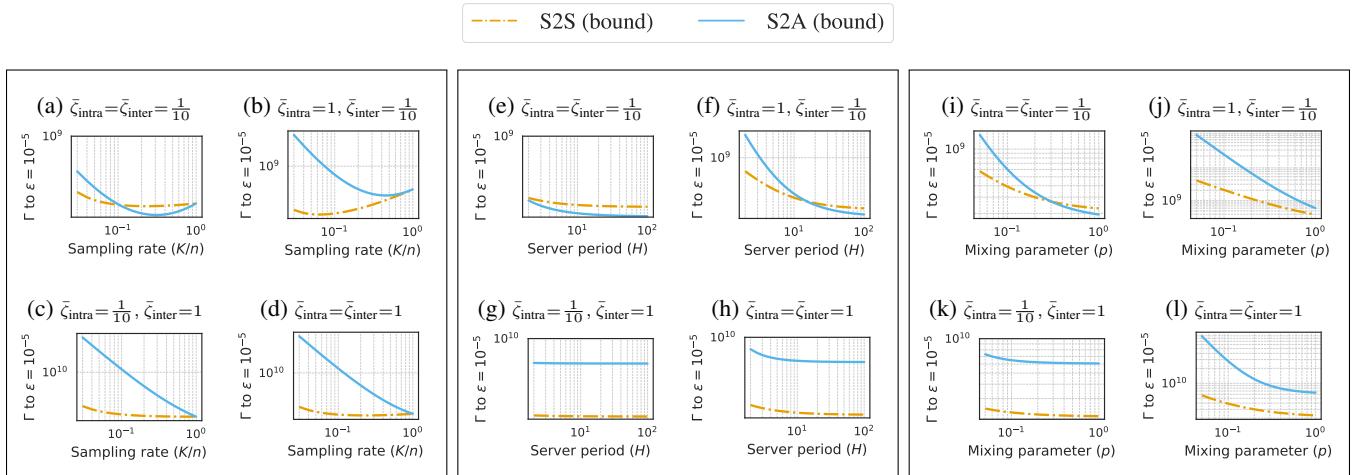


Figure 6: Communication costs $\Gamma_{\text{S2S}}(\epsilon)$ and $\Gamma_{\text{S2A}}(\epsilon)$ for $n=100$, $\epsilon = 10^{-5}$, $L=f_0=1$, $\bar{\sigma}=0$. Left panel: Sampling rate (K/n) at $H=5$, $p=1$. Center panel: Server period (H) at $K/n=0.2$, $p=1$. Right panel: Mixing parameter (p) at $K/n=0.2$, $H=5$.

E Additional Experimental Results

This appendix complements Section 6 by providing further details on the experimental setup, supporting the main results (via Tables 1–6), and presenting new experiments (Figs. 7–14), including heatmap summaries on MNIST and CIFAR-10 (Figs. 7–8), a deeper analysis of outlier behaviors (Figs. 9–10), comparison of fixed and dynamic topologies (Fig. 11), different server optimizers (Fig. 12), additional CIFAR-100 runs (Fig. 13), and empirical measurements of bias and disagreement errors (Fig. 14).

Detailed Experimental Setup. In line with prior work on semi-decentralized federated learning (Lin et al. 2021; Guo et al. 2021; Chen, Wang, and Brinton 2024), we benchmark the S2S and S2A primitives on two standard image classification datasets: MNIST (Deng 2012) and CIFAR-10 (Krizhevsky and Hinton 2009). These datasets offer a controlled and reproducible testbed for evaluating algorithmic performance in federated settings.

For MNIST, we use a single-layer linear classifier with 7,850 trainable parameters. This model is computationally lightweight and supports efficient training for our 9,600 runs on an Intel(R) Xeon(R) E-2224G CPU @ 3.50GHz.

For CIFAR-10, we adopt a standard convolutional neural network consisting of two 5×5 convolutional layers. The first layer maps the 3-channel RGB input to 32 channels, followed by ReLU activation and 2×2 max pooling. The second layer outputs 64 channels and is again followed by ReLU and max pooling. The resulting features are flattened and passed through a fully connected layer with 2048 ReLU units, followed by a final linear classifier with 10 logits.

For CIFAR-100, we consider a deeper convolutional neural network. The feature extractor consists of three convolutional blocks with 3×3 convolutions and ELU activations. The first block maps the 3-channel RGB input to 128 channels through two convolutional layers, followed by 2×2 max pooling. The second block increases the width to 256 channels via two further 3×3 convolutions, again followed by max pooling and a dropout layer with rate 0.25. The third block maps to 512 channels with two 3×3 convolutions, a final max-pooling layer, and dropout 0.25, yielding a $512 \times 2 \times 2$ feature map. The classifier flattens these features and applies a fully connected layer with 1024 ELU units and dropout 0.5, followed by a final linear layer producing 100 logits.

These model architectures are included in benchmark libraries such as FedML (He et al. 2020) and has been widely used in prior semi-decentralized FL work—e.g., for evaluating S2A (Lin et al. 2021; Guo et al. 2021) and S2S (Chen, Wang, and Brinton 2024).

Our experiments on CIFAR-10 and CIFAR-100 comprise 9,600 training runs on a Linux server with four NVIDIA GeForce GTX 1080 Ti GPUs.

Detailed Comparison of Figures 2–3. Tables 1–6 complement the visual comparison in Figures 2–3 by quantifying the test accuracy gap between S2S and S2A across the four heterogeneity regimes and key parameters: sampling rate (K/n), server period (H), and D2D topology. Tables 1–2 aggregate the average performance gap over both datasets and all topologies, reporting mean and maximum accuracy gaps, and their statistical significance. Tables 3–6 report detailed results by dataset, heterogeneity regime, and key parameter: for MNIST, Tables 3 and 5 vary sampling rate and server period, respectively; for CIFAR-10, the corresponding results are shown in Tables 4 and 6. These tables serve to support the empirical trends discussed in Section 6 and to identify regimes where either primitive yields statistically significant improvements.

Heatmap Summaries. Figures 7 and 8 extend Figure 4 by reporting accuracy gaps on both MNIST and CIFAR-10 datasets for the ring topology, across all four heterogeneity regimes. Each heatmap reports the test accuracy difference (S2S minus S2A) after $T = 100$ rounds, with sampling rate (K/n) varying across columns and server period (H) varying across rows. Positive values favor S2S. The new panels in Figures 7 (a)–(d) and 8 (a), (d) reinforce the empirical regimes discussed in Section 6. While the trends largely follow theoretical predictions, a few configurations—e.g., $(K/n, H) = (0.2, 20)$ in Figure 8 (c), and $(0.2, 15)$ and $(0.2, 20)$ in Figure 8 (d)—depart from the expected behavior. These are examined in detail below.

Analysis of Outlier Cases. Figures 9 and 10 report test accuracy over $T = 1000$ communication rounds on CIFAR-10 with $K/n = 0.2$ and ring topology, focusing on the outlier configurations identified in Figure 8 (c)–(d). Figure 9 fixes $H = 20$ and compares the two opposite heterogeneity regimes: (a) intra non-IID, inter IID; and (b) intra IID, inter non-IID. Figure 10 fixes the regime to intra non-IID, inter non-IID and reports two configurations from Figure 8 (d) with $H \in \{15, 20\}$. Specifically:

- In Figure 9 (a), S2A performs comparably to or better than S2S when inter-component heterogeneity is IID, even under non-IID intra-component distributions (Regimes R1 and R2).
- In Figure 9 (b), where inter-component heterogeneity is non-IID, S2A performs better in early rounds (explaining the accuracy gap at $T = 100$ for $(K/n, H) = (0.2, 20)$ in Figure 8 (c)), but its performance deteriorates over time, with periodic drops at each D2S round. S2S, in contrast, shows more stable convergence and outperforms S2A for $T > 100$ (Regime R3).

- Figure 10 (a) reports the case $(K/n, H) = (0.2, 15)$ from Figure 8 (d), where S2A performs best at $T = 100$ (+4.8 p.p.), but is eventually outperformed by S2S (+6 p.p. at $T = 1000$), consistent with theoretical results under high inter heterogeneity.
- Similarly, Figure 10 (b) reports the case $(K/n, H) = (0.2, 20)$ from Figure 8 (d), where S2A initially performs better than S2S (+6.9 p.p. at $T = 100$), but is eventually outperformed by S2S (+11 p.p. at $T = 1000$).

Dynamic Topologies. Figure 11 compares fixed and time-varying D2D graphs on CIFAR-10 in Regime R3 (intra IID, inter non-IID). In these experiments, we fix $K/n = 0.2$ and $H = 20$, and consider a fixed regular graph and a random regular graph with the same degree (4). Recall that in our analysis, D2D variability is fully captured by p (Lemma 2); the random regular topology yields faster intra-component mixing (we measure $p_{\text{random}} \approx 0.8 \gg p_{\text{fixed}} \approx 0.2$). We observe that moving from a fixed to a dynamically switching topology improves the final test accuracy of both S2S and S2A (up to +3.4 and +4.5 p.p., respectively), while leaving the qualitative S2S/S2A regime unchanged. Specifically, the average S2S/S2A gap increases from $+8.58 \pm 0.32$ p.p. (fixed) to $+11.52 \pm 0.42$ p.p. (random), i.e., by $+2.94 \pm 0.20$ p.p. These experiments with dynamic topologies (random regular graphs) are consistent with our theory for time-varying graphs with randomly switching links (Appendix D) and suggest that dynamic topologies benefit S2S more than S2A in Regime R3.

Alternative Server Optimizers. Figure 12 compares FedAvg and FedAvgM (with momentum $\beta = 0.9$) on CIFAR-10 with ring topology in Regime R3, for $K/n = 0.2$ and $H = 20$. We observe that the mean S2S–S2A gap is essentially unchanged: $+8.10 \pm 0.37$ p.p. for FedAvg and $+8.07 \pm 0.35$ p.p. for FedAvgM, yielding a difference of -0.03 ± 0.10 p.p. Thus, the empirical results suggest that introducing momentum does not significantly alter the relative performance between S2S and S2A. In the last 100 rounds, however, FedAvgM+S2A exhibits smaller accuracy drops (about 20% reduction) compared to FedAvg+S2A, indicating that momentum can help reduce the S2A accuracy drops under high inter-component heterogeneity, although without significantly changing its asymptotic accuracy gap to S2S.

Experiments on CIFAR-100. Figure 13 reports test accuracy on CIFAR-100 under Regimes R2 and R3 for $K/n = 0.2$, $H = 20$, and complete topology. In Regime R2 (intra non-IID, inter IID), S2A outperforms S2S by approximately 1.9 ± 0.02 p.p. on average, consistent with the conclusion that the broadcast operator (S2A) is beneficial when inter-component heterogeneity is low. In Regime R3 (intra IID, inter non-IID), the performance advantage switches in favor of S2S, which now outperforms S2A by approximately 13.6 ± 1.0 p.p. on average, confirming that the broadcast-induced bias of S2A becomes detrimental under high inter-component heterogeneity.

Bias and Disagreement Errors. To further validate our theoretical analysis and the S2S/S2A comparison in Section 4.1, Figure 14 tracks the bias and disagreement errors, as defined in §4.1(i)–(ii), over $T = 1000$ rounds on CIFAR-10 (ring topology, Regime R3, $K/n = 0.2$, $H = 20$). In detail:

- Figure 14(a) reports the test loss over communication rounds;
- Figure 14(b) reports the disagreement error at D2D rounds;
- Figures 14(c)–(d) report the disagreement and bias errors at D2S rounds;
- Figures 14(e)–(f) report the empirical disagreement and bias ratios at D2S rounds.

Consistently with our analysis, for S2A we observe pronounced spikes in the bias error after each D2S round and (numerically) zero disagreement error, whereas for S2S the bias error is zero but a non-zero disagreement remains after the D2S step. The non-zero bias of S2A correlates with its performance degradation relative to S2S in Regime R3, while the residual disagreement of S2S does not prevent it from achieving higher final accuracy in this regime. Finally, the empirical disagreement and bias ratios in Figs. 14(e)–(f) oscillate around the theoretical values predicted by Eqs. (9)–(12): for $n = 100$ and $K = 20$, we have a disagreement ratio of $\frac{n-K}{n-1} \approx 0.81$ and a bias ratio of $\frac{n-K}{K(n-1)} \approx 0.04$, reinforcing the consistency between our analysis and experimental results.

Table 1: Test accuracy gap (percentage-point difference) between S2S and S2A, aggregated from Figures 2 and 3 over 12 configurations ($K/n \in \{0.2, 0.4, 0.6, 0.8\} \times 3$ topologies) with fixed server period $H = 5$. S2S/S2A/- counts the number of configurations where each method outperforms the other (gaps below the standard error are denoted by “–”). Positive gaps favor S2S. We report the mean, standard error, maximum gap values, and p -values from a t -test over the 12 comparisons.

Dataset	Intra/Inter Regime	S2S/S2A/-	Gap (mean \pm se)	p-value	Gap (max)	(K/n, D2D topology)
MNIST	IID / IID	0/9/3	-0.01 ± 0.00	0.007	-0.05	(0.2, ring)
MNIST	non-IID / IID	6/6/0	$+0.00 \pm 0.03$	0.859	+0.25	(0.2, complete)
MNIST	IID / non-IID	12/0/0	$+0.91 \pm 0.22$	0.001	+2.37	(0.2, complete)
MNIST	non-IID / non-IID	12/0/0	$+0.86 \pm 0.18$	<0.001	+1.96	(0.2, ring)
CIFAR-10	IID / IID	1/10/1	-0.31 ± 0.08	0.002	-0.97	(0.2, ring)
CIFAR-10	non-IID / IID	9/3/0	$+1.80 \pm 0.73$	0.028	+8.43	(0.2, ring)
CIFAR-10	IID / non-IID	12/0/0	$+1.11 \pm 0.28$	0.001	+3.41	(0.2, ring)
CIFAR-10	non-IID / non-IID	12/0/0	$+2.87 \pm 0.55$	<0.001	+7.01	(0.2, complete)

Table 2: Test accuracy gap (percentage-point difference) between S2S and S2A, aggregated from Figures 2 and 3 over 12 configurations ($H \in \{5, 10, 15, 20\} \times 3$ topologies) with fixed sampling rate $K/n = 0.2$. S2S/S2A/- counts the number of configurations where each method outperforms the other. Positive gaps favor S2S. We report the mean, standard error, maximum gap values, and p -values from a t -test over the 12 comparisons.

Dataset	Intra/Inter Regime	S2S/S2A/-	Gap (mean \pm se)	p-value	Gap (max)	(H, D2D topology)
MNIST	IID / IID	1/10/1	-0.02 ± 0.01	0.006	-0.06	(10, ring)
MNIST	non-IID / IID	8/2/2	$+0.17 \pm 0.09$	0.076	+0.84	(20, complete)
MNIST	IID / non-IID	11/1/0	$+1.62 \pm 0.24$	<0.001	+2.35	(5, complete)
MNIST	non-IID / non-IID	12/0/0	$+2.01 \pm 0.18$	<0.001	+3.00	(10, complete)
CIFAR-10	IID / IID	0/10/2	-0.40 ± 0.09	<0.001	-0.97	(5, ring)
CIFAR-10	non-IID / IID	7/5/0	$+0.99 \pm 1.15$	0.411	+8.52	(5, ring)
CIFAR-10	IID / non-IID	11/1/0	$+2.44 \pm 0.42$	<0.001	+4.42	(10, complete)
CIFAR-10	non-IID / non-IID	7/5/0	-0.26 ± 1.15	0.827	+7.01	(5, complete)

Table 3: Test accuracy gap (percentage-point difference) between S2S and S2A on MNIST, reported from Figure 2 for varying sampling rates $K/n \in \{0.2, 0.4, 0.6, 0.8\}$ with fixed server period $H = 5$. Positive gaps favor S2S. Each row corresponds to a heterogeneity regime and sampling rate (K/n), and each column to a D2D topology. Each entry is annotated with the best strategy: S2S, S2A, or – (gap below standard error).

Intra/Inter Regime	Sampling rate (K/n)	Complete	Grid	Ring
IID / IID	0.2	-0.01 (S2A)	-0.02 (S2A)	-0.05 (S2A)
	0.4	–	-0.02 (S2A)	-0.04 (S2A)
	0.6	–	-0.01 (S2A)	-0.02 (S2A)
	0.8	–	-0.01 (S2A)	-0.01 (S2A)
non-IID / IID	0.2	+0.25 (S2S)	+0.17 (S2S)	+0.04 (S2S)
	0.4	+0.04 (S2S)	-0.06 (S2A)	-0.16 (S2A)
	0.6	+0.06 (S2S)	-0.04 (S2A)	-0.10 (S2A)
	0.8	+0.01 (S2S)	-0.03 (S2A)	-0.05 (S2A)
IID / non-IID	0.2	+2.37 (S2S)	+2.36 (S2S)	+2.14 (S2S)
	0.4	+1.38 (S2S)	+1.38 (S2S)	+1.26 (S2S)
	0.6	+0.65 (S2S)	+0.71 (S2S)	+0.76 (S2S)
	0.8	+0.31 (S2S)	+0.29 (S2S)	+0.06 (S2S)
non-IID / non-IID	0.2	+1.65 (S2S)	+1.87 (S2S)	+1.96 (S2S)
	0.4	+1.14 (S2S)	+1.31 (S2S)	+1.41 (S2S)
	0.6	+0.73 (S2S)	+0.84 (S2S)	+0.86 (S2S)
	0.8	+0.35 (S2S)	+0.39 (S2S)	+0.35 (S2S)

Table 4: Test accuracy gap (percentage-point difference) between S2S and S2A on CIFAR-10, reported from Figure 3 for varying sampling rates $K/n \in \{0.2, 0.4, 0.6, 0.8\}$ with fixed server period $H = 5$. Positive gaps favor S2S. Each row corresponds to a heterogeneity regime and sampling rate (K/n); each column to a D2D topology. Every entry is annotated with the best strategy: S2S, S2A, or – (gap below standard error).

Intra/Inter Regime	Sampling rate (K/n)	Complete	Grid	Ring
IID / IID	0.2	–	-0.30 (S2A)	-0.97 (S2A)
	0.4	+0.18 (S2S)	-0.44 (S2A)	-0.71 (S2A)
	0.6	-0.45 (S2A)	-0.38 (S2A)	-0.54 (S2A)
	0.8	-0.41 (S2A)	-0.24 (S2A)	-0.34 (S2A)
non-IID / IID	0.2	+0.01 (S2S)	+7.74 (S2S)	+8.43 (S2S)
	0.4	-0.50 (S2A)	+2.69 (S2S)	+3.99 (S2S)
	0.6	-0.25 (S2A)	+1.81 (S2S)	+1.42 (S2S)
	0.8	-0.08 (S2A)	+0.92 (S2S)	+1.03 (S2S)
IID / non-IID	0.2	+1.50 (S2S)	+2.96 (S2S)	+3.42 (S2S)
	0.4	+1.12 (S2S)	+0.24 (S2S)	+1.21 (S2S)
	0.6	+0.76 (S2S)	+1.21 (S2S)	+1.92 (S2S)
	0.8	+0.19 (S2S)	+0.34 (S2S)	+2.03 (S2S)
non-IID / non-IID	0.2	+7.01 (S2S)	+2.50 (S2S)	+3.60 (S2S)
	0.4	+5.99 (S2S)	+4.38 (S2S)	+4.54 (S2S)
	0.6	+3.97 (S2S)	+3.33 (S2S)	+2.88 (S2S)
	0.8	+2.33 (S2S)	+1.44 (S2S)	+1.15 (S2S)

Table 5: Test accuracy gap (percentage-point difference) between S2S and S2A on MNIST, reported from Figure 2 for varying server periods $H \in \{5, 10, 15, 20\}$ with fixed sampling rate $K/n = 0.2$. Positive gaps favor S2S. Each row corresponds to a heterogeneity regime and a server period; each column to a D2D topology. Every entry is annotated with the best strategy: S2S, S2A, or – (gap below standard error).

Intra/Inter Regime	Server period (H)	Complete	Grid	Ring
IID / IID	5	–0.01 (S2A)	–0.02 (S2A)	–0.05 (S2A)
	10	–0.01 (S2A)	–0.02 (S2A)	–0.06 (S2A)
	15	–	–0.03 (S2A)	–0.05 (S2A)
	20	+0.03 (S2S)	–0.03 (S2A)	–0.04 (S2A)
non-IID / IID	5	+0.25 (S2S)	+0.17 (S2S)	+0.04 (S2S)
	10	+0.29 (S2S)	+0.02 (S2S)	–0.10 (S2A)
	15	+0.71 (S2S)	+0.29 (S2S)	–0.14 (S2A)
	20	+0.84 (S2S)	–	–
IID / non-IID	5	+2.35 (S2S)	+2.35 (S2S)	+2.18 (S2S)
	10	+2.34 (S2S)	+2.33 (S2S)	+2.02 (S2S)
	15	+1.64 (S2S)	+1.66 (S2S)	+0.79 (S2S)
	20	+1.08 (S2S)	+1.05 (S2S)	–0.30 (S2A)
non-IID / non-IID	5	+1.65 (S2S)	+1.87 (S2S)	+1.96 (S2S)
	10	+3.00 (S2S)	+2.96 (S2S)	+2.62 (S2S)
	15	+2.31 (S2S)	+2.06 (S2S)	+1.69 (S2S)
	20	+1.51 (S2S)	+1.36 (S2S)	+1.09 (S2S)

Table 6: Test accuracy gap (percentage-point difference) between S2S and S2A on CIFAR-10, reported from Figure 3 for varying server periods $H \in \{5, 10, 15, 20\}$ with fixed sampling rate $K/n = 0.2$. Positive gaps favor S2S. Each row corresponds to a heterogeneity regime and a server period; each column to a D2D topology. Every entry is annotated with the best strategy: S2S, S2A, or – (gap below standard error).

Intra/Inter Regime	Server period (H)	Complete	Grid	Ring
IID / IID	5	–	–0.30 (S2A)	–0.97 (S2A)
	10	–0.52 (S2A)	–0.40 (S2A)	–0.41 (S2A)
	15	–0.16 (S2A)	–0.74 (S2A)	–0.30 (S2A)
	20	–0.13 (S2A)	–0.78 (S2A)	–
non-IID / IID	5	+0.46 (S2S)	+7.28 (S2S)	+8.52 (S2S)
	10	–3.99 (S2A)	+2.50 (S2S)	+2.46 (S2S)
	15	–3.33 (S2A)	+1.14 (S2S)	–0.45 (S2A)
	20	–2.50 (S2A)	+1.25 (S2S)	–2.37 (S2A)
IID / non-IID	5	+1.58 (S2S)	+2.94 (S2S)	+3.49 (S2S)
	10	+4.43 (S2S)	+4.14 (S2S)	+3.42 (S2S)
	15	+2.98 (S2S)	+1.74 (S2S)	+1.07 (S2S)
	20	+1.23 (S2S)	+2.58 (S2S)	–0.84 (S2A)
non-IID / non-IID	5	+7.01 (S2S)	+2.42 (S2S)	+3.60 (S2S)
	10	+0.21 (S2S)	+2.18 (S2S)	+1.87 (S2S)
	15	–3.27 (S2A)	–	–4.79 (S2A)
	20	–4.04 (S2A)	–1.38 (S2A)	–6.90 (S2A)

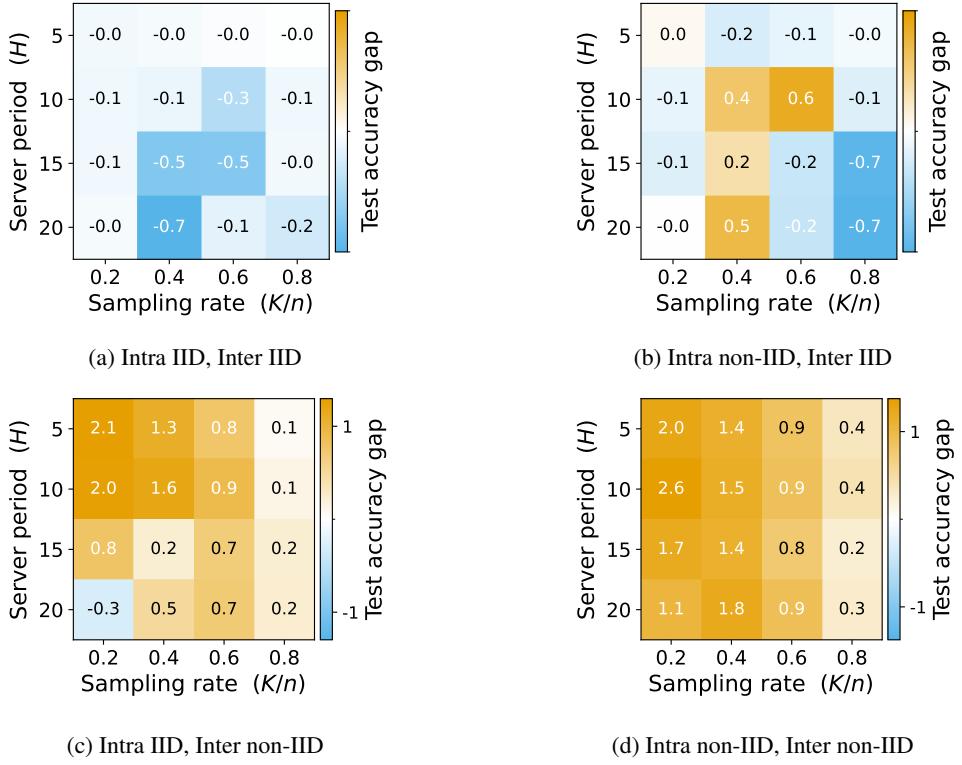


Figure 7: Accuracy gap on MNIST with ring topology.

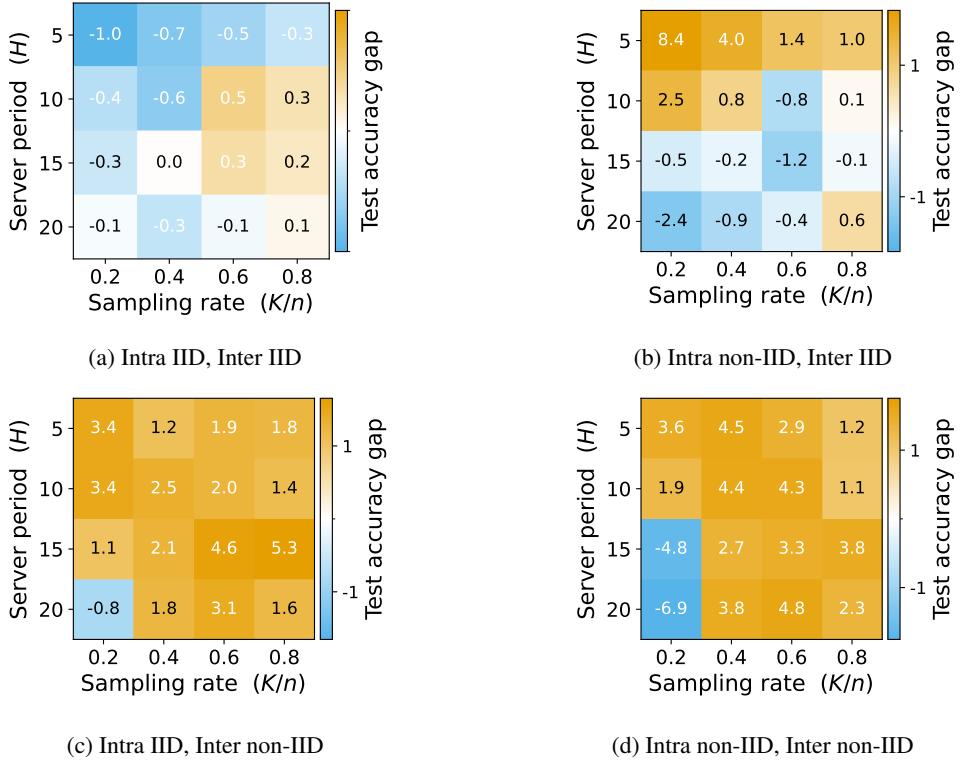


Figure 8: Accuracy gap on CIFAR-10 with ring topology.

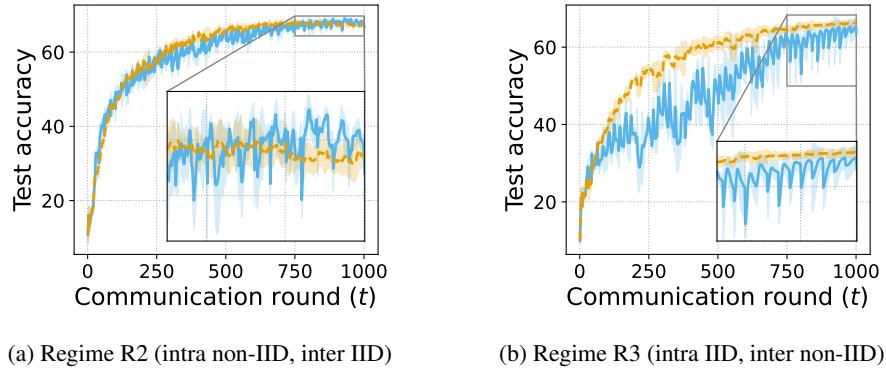


Figure 9: Test accuracy on CIFAR-10; $K/n = 0.2$, $H = 20$, ring topology.

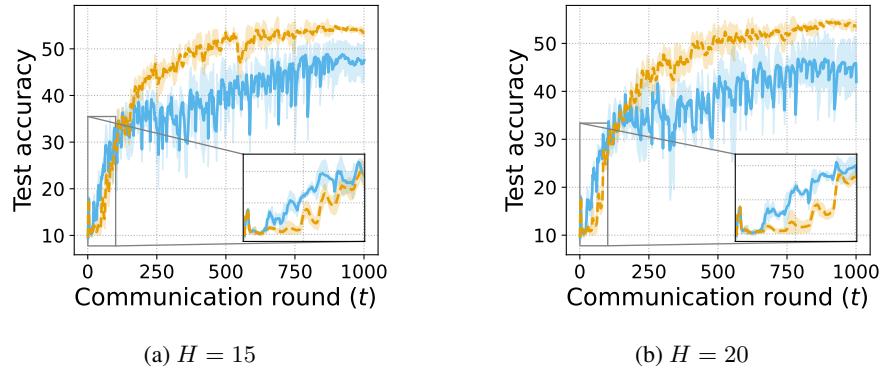


Figure 10: Test accuracy on CIFAR-10; Regime R3 (intra non-IID, inter non-IID), $K/n = 0.2$, ring topology.

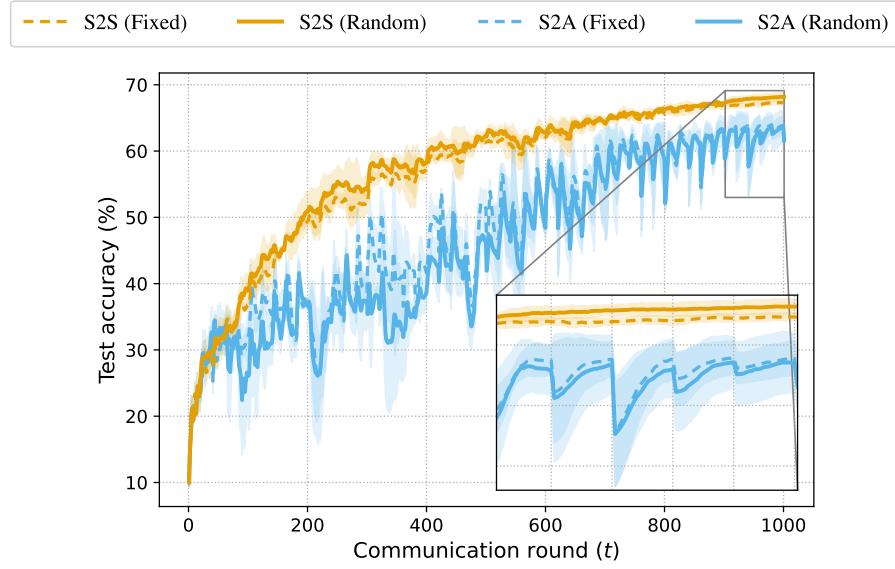


Figure 11: Test accuracy on CIFAR-10, Regime R3 (intra IID, inter non-IID), $K/n=0.2$, $H=20$, fixed/random regular graphs.

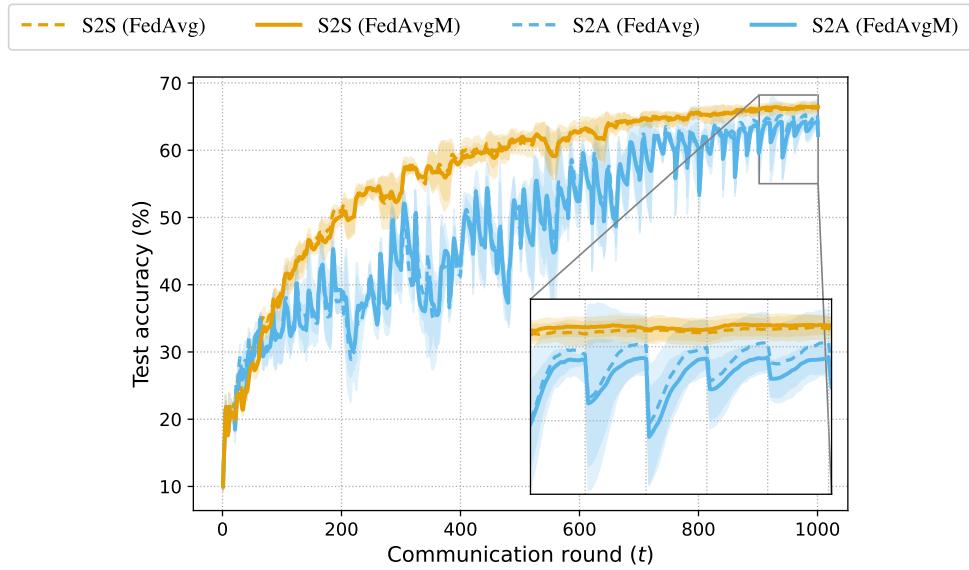
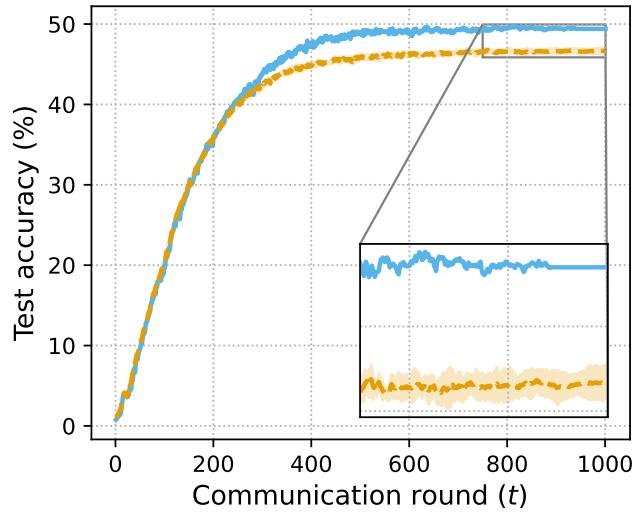
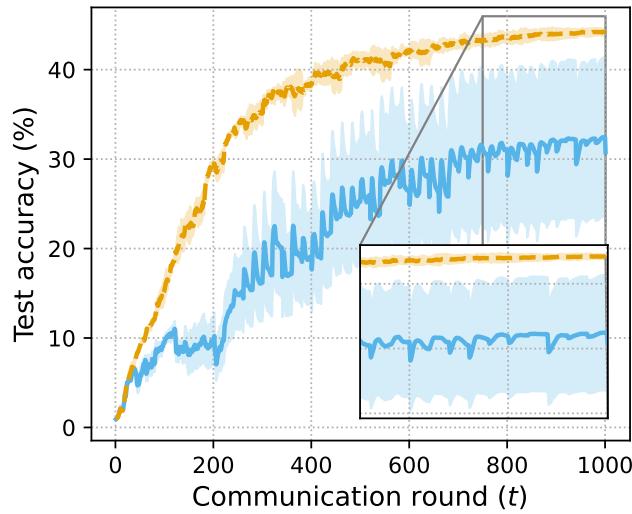


Figure 12: Test accuracy on CIFAR-10, Regime R3, $K/n=0.2$, $H=20$, ring topology, momentum $\beta = 0.9$.

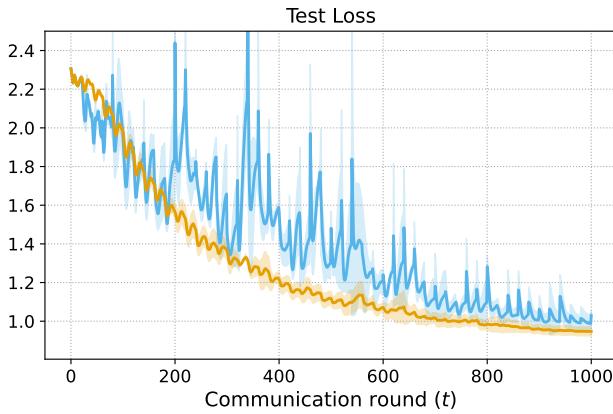


(a) Regime R2 (intra non-IID, inter IID)

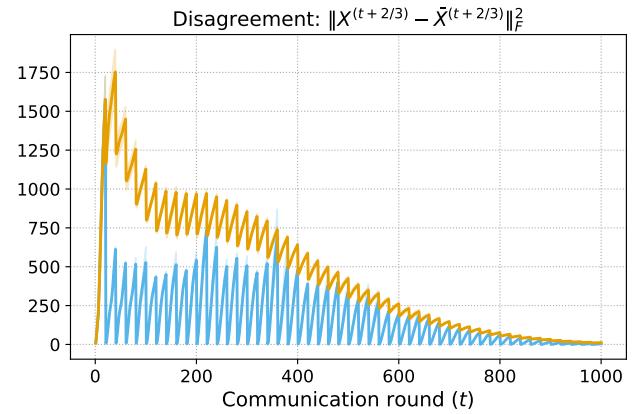


(b) Regime R3 (intra IID, inter non-IID)

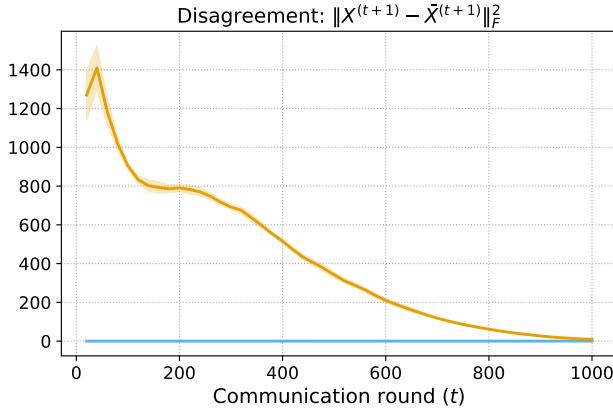
Figure 13: Test accuracy on CIFAR-100; $K/n = 0.2$, $H = 20$, complete topology.



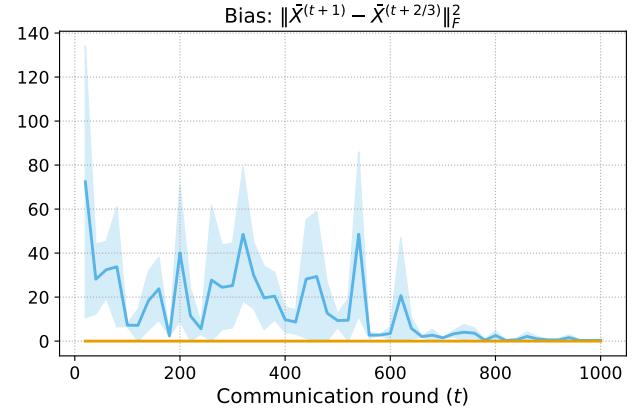
(a) Test loss over communication rounds.



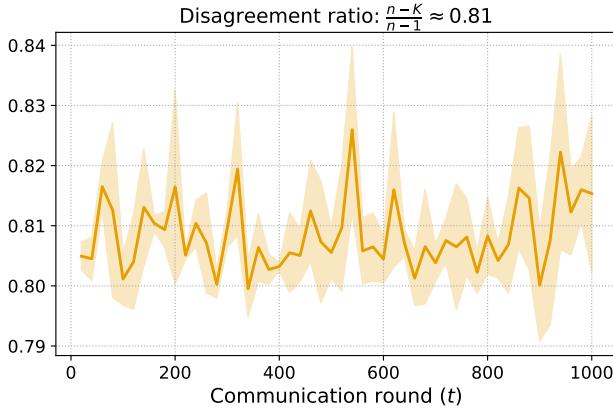
(b) Disagreement error at D2D rounds.



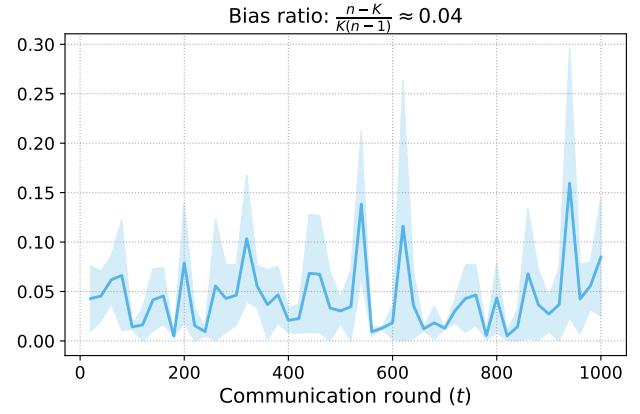
(c) Disagreement error at D2S rounds.



(d) Bias error at D2S rounds.



(e) Disagreement ratio at D2S rounds.



(f) Bias ratio at D2S rounds.

Figure 14: Bias and disagreement errors on CIFAR-10, Regime R3, $K/n = 0.2$, $H = 20$, ring topology.