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

Angelo Rodio[1], **Giovanni Neglia**[2], **Zheng Chen**[1], **Erik G. Larsson**[1]

[1]Department of Electrical Engineering, Linköping University, Sweden
[2]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
Extended version — https://arxiv.org/abs/2511.11560

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 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 widely studied (Shamir, Srebro, and Zhang 2014; Alistarh et al. 2017; Horvóth et al. 2022). Both classical (Stich 2018; Reddi et al. 2021) and 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 many local updates on non-IID data induce *model drift* and hinder convergence. More advanced methods, e.g., using control variates (Karimireddy et al. 2020) or proximal corrections (Mishchenko et al. 2022), mitigate this drift at additional computational or communication cost, but the main conclusion remains: too many local SGD steps under high statistical heterogeneity slow convergence.

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 $\gamma \coloneqq 1 - \lambda_2(W^\top W)$, 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 reveals 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 $n_c \coloneqq |\mathcal{V}_c|$ 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 $\mathcal{V} \coloneqq \bigcup_{c=1}^{C} \mathcal{V}_c$ and $\mathcal{E}^{(t)} \coloneqq \bigcup_{c=1}^{C} \mathcal{E}_c^{(t)}$.

Learning task. We consider an FL system where the central server and the devices collaborate to learn the parameters $\boldsymbol{x} \in \mathbb{R}^d$ 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 $F_i(\boldsymbol{x}; \xi)$ the loss incurred by the model with parameters $\boldsymbol{x}$ on data sample ξ. The goal is to solve an optimization problem of the form:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{x}), \tag{1}$$

where $f_i(\boldsymbol{x}) \coloneqq \frac{1}{|\mathcal{D}_i|} \sum_{\xi \in \mathcal{D}_i} F_i(\boldsymbol{x}, \xi)$ 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 $\Pi \coloneqq \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Given n vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, we write their average as $\bar{\boldsymbol{x}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \in \mathbb{R}^d$. We stack the n vectors as columns in the matrix $X \coloneqq [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$, such that right-multiplication by Π performs column averaging: $\bar{X} \coloneqq X\Pi = [\bar{\boldsymbol{x}}, \ldots, \bar{\boldsymbol{x}}] \in \mathbb{R}^{d \times n}$. 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 $t \in \{0, \ldots, T-1\}$ consists of two or three steps:

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

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

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

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

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

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

(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:

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

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 $W^{(t)} \sim \mathcal{W}$

for $t = 0, \ldots, T-1$ **do**
  $X^{(t+1/3)} \leftarrow X^{(t)} - \eta_t \nabla F(X^{(t)}, \mathcal{B}^{(t)})$
  $X^{(t+2/3)} \leftarrow X^{(t+1/3)} W^{(t)}$
  if $t \equiv 0 \pmod{H}$ **then**
    sample $\mathcal{S}^{(t)} \subseteq \mathcal{V}$, $|\mathcal{S}^{(t)}| = K$
    build $W^{(t)}_{\text{S2(S/A)}}$ by Eq. (5) (S2S) or Eq. (6) (S2A)
    $X^{(t+1)} \leftarrow X^{(t+2/3)} W^{(t)}_{\text{S2(S/A)}}$
  else
    $X^{(t+1)} \leftarrow X^{(t+2/3)}$
return $X^{(T)}$

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

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

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

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

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.

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:

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

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

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

with magnitude (see Lemma 8 in Rodio et al. (2025)):

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

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 $\Pi W_{\text{S2A}} = \Pi$ and $W_{\text{S2A}}\Pi = W_{\text{S2A}} \neq \Pi$. This property eliminates disagreement since:

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

but introduces the broadcast-induced bias:

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

with magnitude (see Lemma 11 in Rodio et al. (2025)):

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

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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

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

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

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

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

To keep the analysis unified across convex and non-convex settings, we assume that the stochastic variance is uniformly bounded in $\boldsymbol{x}$ (Le Bars et al. 2023), although in the convex case it suffices 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 $\boldsymbol{x} \in \mathbb{R}^d$,*

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

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 in Rodio et al. (2025)).

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 $W_c \coloneqq W[\mathcal{V}_c, \mathcal{V}_c] \in \mathbb{R}^{n_c\times n_c}$ 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:

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

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 $X\in\mathbb{R}^{d\times n}$,*

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

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 $X \in \mathbb{R}^{d\times n}$,*

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

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 in Rodio et al. (2025)). 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 (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 $\bar{\zeta}^2_{intra}$, $\bar{\zeta}^2_{inter} > 0$ such that, for all $\boldsymbol{x} \in \mathbb{R}^d$:*

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

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

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 ready to present our main convergence results; all proofs are deferred to Rodio et al. (2025, Appendices B–C).

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: $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left(f(\bar{\boldsymbol{x}}^{(t)}) - f^\star\right) \le \epsilon$ *after*

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

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

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

where $R_0 := \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2$ and $f_0 := f(\boldsymbol{x}^{(0)}) - f^\star$ denote the initial errors, and $\mathcal{O}(\cdot)$ hides the numerical constants explicitly provided in Rodio et al. (2025, Appendix B).

Theorem 2 (Sampled-to-All)**.** *Under Assumptions 1–5, there exists a constant stepsize $\eta \le \frac{p}{8L}$ such that, for any target accuracy $\epsilon > 0$, Algorithm 1 (S2A) achieves*
Convex: $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left(f(\bar{\boldsymbol{x}}^{(t)}) - f^\star\right) \le \epsilon$ *after*

$$T \ge \mathcal{O}\Bigg(\frac{\bar{\sigma}^2}{n\epsilon^2} + \frac{n-K}{K(n-1)}\frac{\bar{\zeta}_{intra}^2}{Hp^2\epsilon^2} + \frac{n-K}{K(n-1)}\frac{H\bar{\zeta}_{inter}^2}{\epsilon^2} + \frac{\sqrt{L}\bar{\zeta}_{intra}}{p\epsilon^{3/2}} + \frac{\sqrt{L}H\bar{\zeta}_{inter}}{\epsilon^{3/2}} + \frac{L}{p\epsilon}\Bigg)R_0^2, \tag{23}$$

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

$$T \ge \mathcal{O}\Bigg(\frac{\bar{\sigma}^2}{n\epsilon^2} + \frac{n-K}{K(n-1)}\frac{\bar{\zeta}_{intra}^2}{Hp^2\epsilon^2} + \frac{n-K}{K(n-1)}\frac{H\bar{\zeta}_{inter}^2}{\epsilon^2} + \frac{\bar{\zeta}_{intra}}{p\epsilon^{3/2}} + \frac{H\bar{\zeta}_{inter}}{\epsilon^{3/2}} + \frac{1}{p\epsilon}\Bigg)Lf_0, \tag{24}$$

where $R_0 := \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2$ and $f_0 := f(\boldsymbol{x}^{(0)}) - f^\star$ denote the initial errors, and $\mathcal{O}(\cdot)$ hides the numerical constants explicitly provided in Rodio et al. (2025, Appendix C).

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.

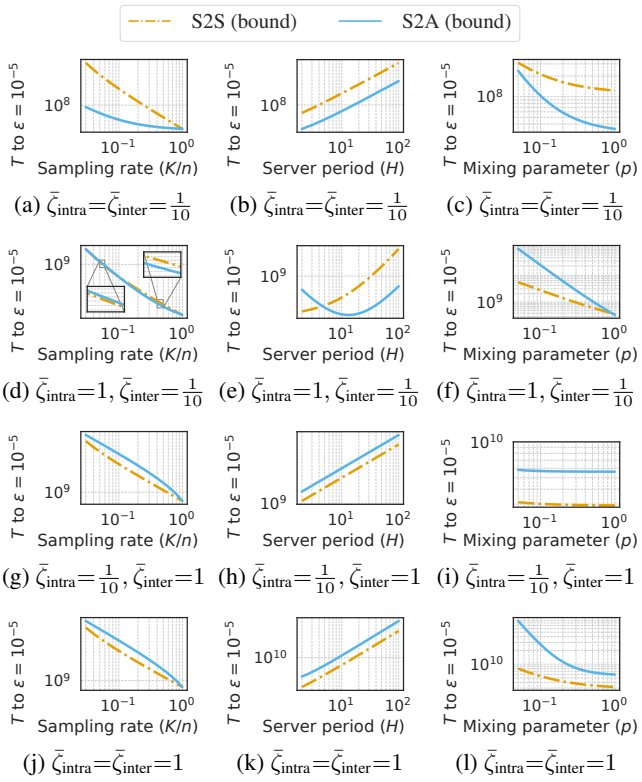


Figure 1: Convergence rates for S2S and S2A, comparing Eqs. (22)–(24) with $n=100$, $L=f_0=1$, $\bar{\sigma}=0$. Left column: Sampling rate (K/n) with $H=5$, $p=1$. Center column: Server period (H) with $K/n=0.2$, $p=1$. Right column: Mixing parameter (p) with $K/n=0.2$, $H=5$.

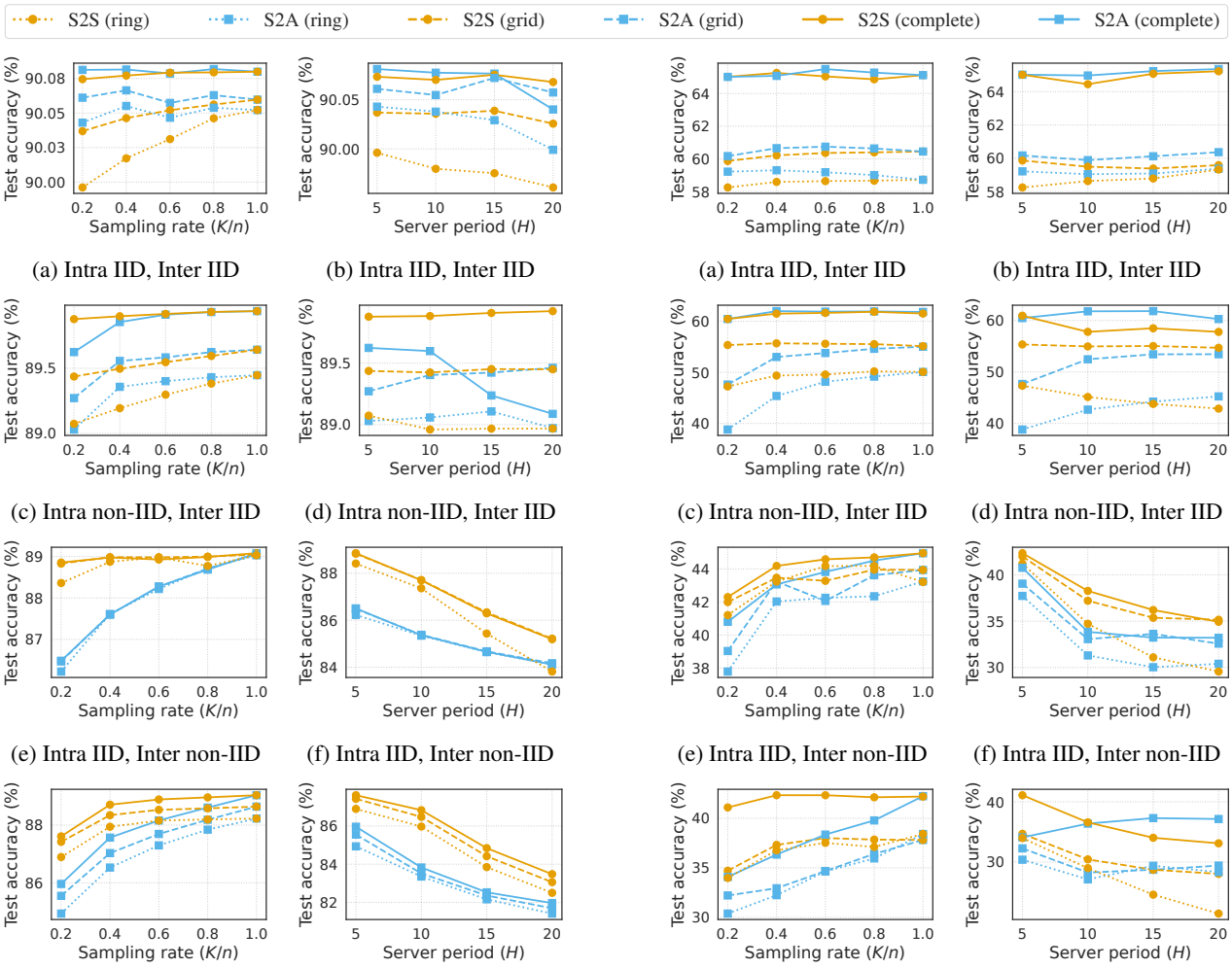
- When only *one* device is sampled ($K = 1$), $W_{\text{S2S}} = I$, S2S is unable to mix the sampled model across components, and the bounds in Eqs. (21)–(22) diverge. In contrast, S2A still broadcasts the (single) sampled model to all devices and thus converges, albeit at a slower rate.

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 $H \to \infty$, 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 multiped 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 period (center column), and mixing parameter (right column). We consider $n = 100$ devices, and set the parameters



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

(c) Intra non-IID, Inter IID (d) Intra non-IID, Inter IID

(e) Intra IID, Inter non-IID (f) Intra IID, Inter non-IID

(g) Intra non-IID, Inter non-IID (h) Intra non-IID, Inter non-IID

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

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

(c) Intra non-IID, Inter IID (d) Intra non-IID, Inter IID

(e) Intra IID, Inter non-IID (f) Intra IID, Inter non-IID

(g) Intra non-IID, Inter non-IID (h) Intra non-IID, Inter non-IID

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

$L = f_0 = 1$ 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. $\bar{\zeta}_{intra}$, $\bar{\zeta}_{inter}$ ***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. $\bar{\zeta}_{inter} \ll \bar{\zeta}_{intra}$***:*** 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. $\bar{\zeta}_{inter}$ ***is high:*** S2S converges faster for most values of K/n, H, and p, irrespective of $\bar{\zeta}_{\text{intra}}$ (Figs. 1(g–l)).

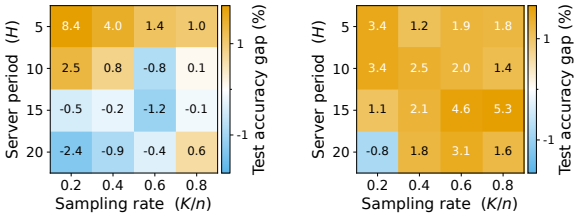
6 Experimental Results

We simulate a semi-decentralized FL system consisting of a central server and $n = 100$ devices partitioned into $C = 2$ 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).

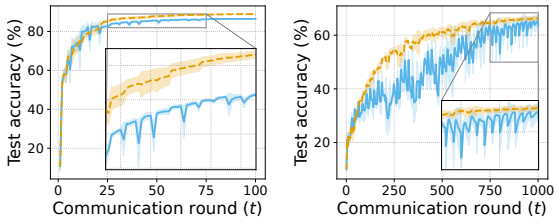
We introduce intra- and inter-component heterogeneity mimicking the constants $\bar{\zeta}_{\text{intra}}$ and $\bar{\zeta}_{\text{inter}}$ of Assumption 5:

- *Inter-component heterogeneity*. We partition the dataset



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

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



(a) MNIST, $H = 5$ (b) CIFAR-10, $H = 20$

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

across components either through an IID split (each component 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. Additional experimental details are given in Rodio et al. (2025, Appendix E).

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*.

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.

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. ProxSkip: 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.

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*.

Rodio, A.; Neglia, G.; Chen, Z. et al. 2025. A Unified Convergence Analysis for Semi-Decentralized Learning: Sampled-to-Sampled vs. Sampled-to-All Communication. arXiv:2511.11560.

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.

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