# **Hubs and Spokes Learning: Efficient and Scalable Collaborative Machine Learning**

Atul Sharma 1 Kavindu Kherath 1 Saurabh Bagchi 1 Somali Chaterji 1 Chaoyue Liu 1

## **Abstract**

We introduce the Hubs and Spokes Learning (HSL) framework, a novel paradigm for collaborative machine learning that combines the strengths of Federated Learning (FL) and Decentralized Learning (P2PL). HSL employs a two-tier communication structure that avoids the single point of failure inherent in FL and outperforms the state-of-the-art P2PL framework, Epidemic Learning Local (ELL). At equal communication budgets (total edges), HSL achieves higher performance than ELL, while at significantly lower communication budgets, it can match ELL's performance. For instance, with only 400 edges, HSL reaches the same test accuracy that ELL achieves with 1000 edges for 100 peers (spokes) on CIFAR-10, demonstrating its suitability for resource-constrained systems. HSL also achieves stronger consensus among nodes after mixing, resulting in improved performance with fewer training rounds. We substantiate these claims through rigorous theoretical analyses and extensive experimental results, showcasing HSL's practicality for large-scale collaborative learning.

# 1. Introduction

In modern machine learning, particularly in settings with edge devices, sensor networks, and large organizations, training models across distributed nodes presents significant challenges. These settings pose challenges like resource constraints, unreliable networks, and heterogeneous (noniid) data distributions across nodes (Nguyen et al., 2022; Wu et al., 2024). Federated Learning (FL) addresses these by employing a central server to aggregate model updates from clients who train locally (McMahan et al., 2017; Karimireddy et al., 2020; Ye et al., 2023). However, FL suffers from scalability bottlenecks and a single point of failure due to its reliance on the central server. Peer-to-Peer Learning (P2PL), or decentralized learning, offers an alternative by eliminating the central server and enabling nodes to directly exchange and aggregate model updates with their neighbors (Lian et al., 2017; Koloskova et al., 2020; Kong et al., 2021). Decentralized learning benefits from increasing the connectivity of nodes, often quantified by the parameter k, which represents the number of neighbors each node communicates with in a given round of training. Increasing k has been shown to improve convergence properties, as mathematically analyzed in the dynamic random graph-based framework, *Epidemic Learning* (De Vos et al., 2023), which is the state-of-the-art P2PL network. However, this improvement comes at the cost of higher total communication and computation, both of which are directly proportional to the total number of edges in the system. In a P2PL network with n nodes and degree k, the total number of edges scales linearly with n and with n. For large-scale networks, this renders fully decentralized approaches increasingly resource-intensive.

**Hubs and Spokes Learning (HSL):** We propose the *HSL* framework integrating the hierarchical structure of FL and the decentralized nature of P2PL. FL embodies high connectivity, while P2PL eliminates the single point of failure. Leveraging the strengths of both results in a scalable and resilient collaborative learning framework, HSL, illustrated in Figure 1, arranges a network into client-like spokes and server-like hubs. Spokes—nodes that hold private data and perform local training—communicate exclusively with hubs, while hubs form a peer-to-peer subnetwork, facilitating decentralized aggregation through gossiping. The spokes communicate exclusively with the hubs and not with other spokes forming a directed graph. This hierarchical design with multiple hubs as peers at the top helps mitigate FL's bottleneck while reducing the overall communication and computation costs in the system. A key limitation of fully decentralized methods is that maintaining strong model mixing at scale requires increasing node connectivity (k), which inflates the communication costs. In contrast, the three subnetworks in HSL—spokes to hubs, hubs-to-hubs, and hubs-to-spokes—increase graph connectivity without requiring spokes to maintain extensive connections. By leveraging a smaller number of hubs, HSL scales efficiently as more spokes join, preventing overload. Moreover, independent tuning of mixing levels at the hub and spoke layers offers flexibility in model propagation, improving convergence and robustness. Thus, HSL bridges the gap between

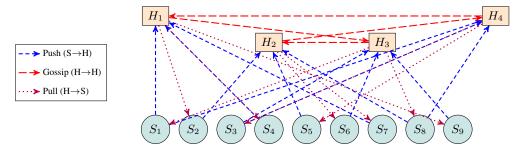


Figure 1. A snapshot of the Hubs and Spokes Learning (HSL) network with 9 spokes and 4 hubs with 25 directed edges, where connections dynamically change in each round, illustrating the three-stage communication process. In **Stage 1** (**Spoke-to-Hub Push**), hubs aggregate models from  $b_{hs} = 3$  randomly sampled spokes (depicted by blue dashed lines). In **Stage 2** (**Hub Gossip**), each hub exchanges models with  $b_{hh} = 1$  other hub (shown in red dashed lines) and averages the received models along with its own. Finally, in **Stage 3** (**Hub-to-Spoke Pull**), each spoke retrieves a model from  $b_{sh} = 1$  randomly selected hub (shown by purple dotted lines).

FL and fully decentralized methods by achieving efficient model mixing without increasing the communication burden on individual nodes. We make the following contributions:

- Framework Definition: We formalize the HSL design, which combines the hierarchical structure of FL with the decentralized communication of P2PL. By assigning distinct roles to hubs and spokes and enabling hubs to form a P2P subnetwork, HSL achieves efficient collaboration while mitigating FL's bottleneck and the high cost of full decentralization.
- Theoretical Analysis: We provide rigorous convergence guarantees for HSL, demonstrating that its two-tiered structure facilitates efficient information propagation and achieves asymptotic convergence under the standard assumptions of smoothness, bounded stochastic noise, and bounded heterogeneity.
- 3. Consensus Distance Bounds: We derive analytical bounds on the consensus distance ratio, quantifying the effectiveness of model mixing at different stages of HSL. Our framework enables independent tuning of hub and spoke budgets leading to improved mixing efficiency while keeping per-spoke communication costs low. By controlling the consensus distance ratio, a measure of mixing effectiveness, HSL achieves efficient model propagation without necessarily increasing spokes' communication budgets—offering a key advantage over fully decentralized P2PL frameworks.
- Empirical Validation: HSL achieves high accuracy even with constrained communication budgets, consistently outperforming or matching *EL Local*, the SOTA P2PL framework, while using fewer edges.

On both the CIFAR-10 and AG News datasets, HSL with just 400 edges achieves the same local test accuracy as ELL with 1000 edges for 100 spokes. Similarly,

for 200 spokes, HSL requires fewer than 600 edges to match ELL's performance with 3000 edges.

# 2. Background and Related Work

Consider a distributed learning setup with  $n_s$  nodes, each holding a private dataset  $\mathcal{D}_i$ . The nodes collaborate to train their models while maintaining privacy by exchanging model updates rather than raw data. Every node initializes the same model  $\mathbf{x}_0$  and follows a common training procedure. Let  $f^{(i)}(\mathbf{x})$  denote the local objective function optimized by node i over its private dataset  $\mathcal{D}_i$ , given by:

$$f^{(i)}(\mathbf{x}) := \mathbb{E}_{\xi \sim \mathcal{D}_i}[f(\mathbf{x}, \xi)],$$

where  $\xi$  is a randomly sampled mini-batch, and the expectation is taken over the data distribution  $\mathcal{D}_i$  of node i, and  $f(\mathbf{x}, \xi)$ . The global objective is formulated as the minimization of the average local objectives across all  $n_s$  nodes:

$$\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} \frac{1}{n_s} \sum_{i=1}^{n_s} f^{(i)}(\mathbf{x}),$$

where  $F(\mathbf{x})$  represents the global loss function to be minimized. We define the model state at iteration t using the matrix  $X_t$ , whose i-th row is  $\mathbf{x}_t^{(i)}$ :

$$X_t := \left[\mathbf{x}_t^{(1)}, \ \mathbf{x}_t^{(2)}, \ \dots, \ \mathbf{x}_t^{(n_s)}\right]^{\top} \in \mathbb{R}^{n_s \times d},$$

where d is the number of model parameters. Each node performs local training on mini-batches of  $\mathcal{D}_i$ , updating its model from  $\mathbf{x}_t^{(i)}$  to an intermediate state  $\mathbf{x}_{t'}^{(i)}$ , where t < t' < t+1. After local training, nodes exchange (or mix) their updated models following:

$$X_{t+1} = W_t X_{t'},$$
 (1)

where  $W_t \in \mathbb{R}^{n_s \times n_s}$  is the *mixing matrix* at iteration t. The mixing matrix  $W_t$  governs the information exchange, determining how nodes aggregate updates from their neighbors.

Federated Learning (FL): One approach to distributed training with decentralized data is *federated averaging* (McMahan et al., 2017), where a central server computes a weighted average of client updates, weighted by local data sizes, and broadcasts the global model. In the notation of (1), the mixing matrix  $W_t$  in FL round has identical rows, ensuring all clients receive the same update. While this star topology enables exact consensus with only  $n_s$  edges, it creates a single point of failure and burdens the central server as clients scale. A sampling scheme can reduce this load but slows training.

**Peer-to-Peer Learning (P2PL):** In P2PL, the nodes handle mixing themselves, optimizing the same global objective via Decentralized-SGD (D-SGD). The mixing matrix  $W_t$  encodes adjacency relationships, dictating how nodes "gossip". Unlike FL, which enforces exact consensus via a server, P2PL follows *inexact consensus*, meaning models  $\mathbf{x}^{(i)}$  are not necessarily identical at any given time. The *consensus distance* (CD) measures the average Euclidean distance of individual models from their mean:

$$CD_t = \frac{1}{n_s} \sum_{i=1}^{n_s} \left\| \mathbf{x}_t^{(i)} - \bar{\mathbf{x}}_t \right\|^2,$$
 (2)

where  $\bar{\mathbf{x}}_t = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{x}_t^{(i)}$  is the mean model at round t. Mixing reduces  $CD_t$ , aligning models across nodes. The CD ratio (CDR) quantifies mixing effectiveness:

$$CDR = \frac{CD_{t+1}}{CD_{t'}}. (3)$$

Lower CDR indicates stronger mixing and improved model alignment. Prior works (Lian et al., 2017) considered static mixing matrices, while (Assran et al., 2019; Koloskova et al., 2020; Kong et al., 2021) perform D-SGD for dynamic graphs, often assuming a doubly stochastic mixing matrix, which is impractical in full decentralization. Recent work has shown decentralized optimization with non-doubly stochastic matrices is viable, broadening mixing protocols. *Epidemic Learning (EL)* (De Vos et al., 2023), the SOTA approach, improves convergence by dynamically changing communication topologies. Each node sends updates to a random peer subset, outperforming static or structured methods in both speed and accuracy.

**Epidemic Learning (EL):** EL has two variants: *EL Oracle* and *EL Local*. EL Oracle enforces a k-regular dynamic graph, where every node maintains both in-degree and outdegree of exactly k, requiring a central coordinator, which conflicts with full decentralization. To address this, EL Local is designed to ensure decentralization, where each node maintains a fixed out-degree of k, sharing its model updates with k peers, while its in-degree can vary around the expected value k. Despite this relaxation, EL Local retains

EL's strong mixing and convergence properties, enabling decentralized learning without central coordination.

Other Approaches: Other works have explored architectures like hierarchical FL (Liu et al., 2020; Abad et al., 2020), which introduces intermediary aggregators but retains a single point of failure at the root server. Blockchain-based decentralized learning (Korkmaz et al., 2020; Qin et al., 2024) enhances consensus but incurs significant overhead. Additionally, works such as (Dhasade et al., 2023; Beltrán et al., 2023) address scalability challenges in collaborative machine learning networks, including communication bottlenecks and system efficiency as networks grow.

# 3. Hubs and Spokes Learning (HSL)

Now that we have established how FL and P2PL enable collaboration among nodes, we introduce our *hubs and spokes learning* (HSL) framework, which facilitates collaboration through intermediate *hubs*. In HSL, *spokes* are client-like nodes that perform local training, while *hubs* serve as intermediaries that aggregate and mix updates. Each spoke communicates exclusively with hubs—sending model updates and receiving aggregated models—without direct spoke-to-spoke communication. This structure, illustrated in Figure 1, improves efficiency by offloading mixing to hubs while decentralizing them ensures scalability and removes FL's single point of failure.

**Training in HSL:** At each round t, spoke i begins with its model  $x_t^{(i)}$  and performs a few local SGD rounds on its private loss function  $F_i(\cdot)$ , updating to  $x_{t+\frac{1}{4}}^{(i)}$ . The collection of spoke updates is:

$$X_{t+\frac{1}{4}} = \left[ x_{t+\frac{1}{4}}^{(1)}, x_{t+\frac{1}{4}}^{(2)}, \dots, x_{t+\frac{1}{4}}^{(n_s)} \right]^{\top} \in \mathbb{R}^{n_s \times d}.$$

Each communication round proceeds in three steps:

1. **Spoke-to-Hub Push:** Each hub k requests the updated models from  $b_{hs}$  spokes, forming the spoke-to-hub mixing matrix  $W_{hs} \in \mathbb{R}^{n_h \times n_s}$ . Concretely, let  $\mathcal{S}_k$  be the set of  $b_{hs}$  spokes from which hub k collects models. Hub k then aggregates these models as:

$$x_{t+\frac{2}{4}}^{(k)} = \frac{1}{b_{hs}} \sum_{i \in S_t} x_{t+\frac{1}{4}}^{(i)}, \quad X_{t+\frac{2}{4}} = W_{hs} X_{t+\frac{1}{4}},$$

where  $X_{t+\frac{2}{4}} \in \mathbb{R}^{n_h \times d}$  represents all hub models after this step. Here, each hub has a fixed in-degree of  $b_{hs}$ , as it collects models from the same number of spokes. The out-degree of spokes, however, may vary around an expected value of  $(n_h \cdot b_{hs})/n_s$ , as spokes share their models with multiple hubs.

2. **Hub-to-Hub Gossip:** The hubs form a peer-to-peer network and execute the gossip scheme from EL Local, where each hub shares its model with  $b_{hh}$  other hubs. While the out-degree is fixed at  $b_{hh}$ , the indegree varies around an average of  $b_{hh}$  since it depends on which hubs receive models. Let  $\mathcal{A}_k$  denote the set of hubs from which k receives models. The updated hub model is given by:

$$x_{t+\frac{3}{4}}^{(k)} = \frac{1}{|\mathcal{A}_k| + 1} \left( x_{t+\frac{2}{4}}^{(k)} + \sum_{m \in \mathcal{A}_k} x_{t+\frac{2}{4}}^{(m)} \right),$$
$$X_{t+\frac{3}{4}} = W_{hh} X_{t+\frac{2}{4}},$$

where  $W_{hh} \in \mathbb{R}^{n_h \times n_h}$  is the hub-hub mixing matrix.

3. **Hub-to-Spoke Pull:** In the final step, each spoke i queries a random set  $\mathcal{H}_i$  of  $b_{sh}$  hubs and averages their models:

$$x_{t+1}^{(i)} = \frac{1}{b_{sh}} \sum_{k \in \mathcal{H}_i} x_{t+\frac{3}{4}}^{(k)}, \quad X_{t+1} = W_{sh} X_{t+\frac{3}{4}},$$

where  $W_{sh} \in \mathbb{R}^{n_s \times n_h}$  is the hub-to-spoke mixing matrix. Each spoke has a fixed in-degree  $b_{sh}$ , while hubs may have a variable out-degree in this step, with an expected value of  $(n_s \cdot b_{sh})/n_h$ .

We summarize the training process in Algorithm 1. From the spokes' perspective, the three aggregation steps yield an *end-to-end* transformation:

$$X_{t+1} = (W_{sh} W_{hh} W_{hs}) X_{t+\frac{1}{4}} \equiv W_{hsl} X_{t+\frac{1}{4}},$$
 (4)

where  $W_{hsl}(\in \mathbb{R}^{n_s \times n_s}) = W_{sh} \, W_{hh} \, W_{hs}$  is the overall mixing matrix for one round of HSL. Although individual mixing matrices  $W_{sh}, \, W_{hh}$ , and  $W_{hs}$  may be sparse due to communication budget constraints (i.e.,  $b_{hs}, b_{hh}, b_{sh}$ ), the effective matrix  $W_{hsl} = W_{sh} \, W_{hh} \, W_{hs}$  is typically not sparse. Empirically, HSL exhibits a larger spectral gap than P2PL at comparable budgets, signifying stronger connectivity and more effective information propagation between spokes. The choice of hubs  $n_h$  in HSL balances individual and total communication budgets. Fewer hubs increase the load per hub, requiring each to maintain connections with more spokes. In the extreme case of unlimited individual budgets, HSL reduces to FL. Conversely, increasing  $n_h$  distributes load and improves fault tolerance but raises overall communication and computation costs.

Next, we analyze its convergence properties under standard assumptions in stochastic optimization, establishing theoretical guarantees on learning performance.

Algorithm 1 Hubs and Spokes Learning (HSL)

$$\begin{split} \textbf{Input:} \ & n_s, T, l, b_{hs}, b_{hh}, b_{sh}, \eta \\ \textbf{Initialize:} \ & x_0^{(i)} \leftarrow x_0, \quad \forall i \in [n_s] \\ \textbf{for} \ & t = 0 \ \text{to} \ T - 1 \ \textbf{do} \\ \textbf{Step 1:} \ & \textbf{Local Training} \\ & x_{t+\frac{1}{4}}^{(i)} \leftarrow \text{SGD}(x_t^{(i)}, \mathcal{D}_i, l, \eta) \end{split}$$

Step 2: Spoke-to-Hub Push

Each hub k samples  $S_k$  with  $|S_k| = b_{hs}$   $x_{t+\frac{2}{4}}^{(k)} = \frac{1}{b_{hs}} \sum_{i \in S_k} x_{t+\frac{1}{4}}^{(i)}$ 

Step 3: Hub-to-Hub Gossip

Each hub k samples  $\mathcal{A}_k$  with  $|\mathcal{A}_k| = b_{hh}$   $x_{t+\frac{3}{4}}^{(k)} = \frac{1}{|\mathcal{A}_k|+1} \left( x_{t+\frac{2}{4}}^{(k)} + \sum_{m \in \mathcal{A}_k} x_{t+\frac{2}{4}}^{(m)} \right)$ 

Step 4: Hub-to-Spoke Pull

Each spoke i samples  $\mathcal{H}_i$  with  $|\mathcal{H}_i| = b_{sh}$   $x_{t+1}^{(i)} = \frac{1}{b_{sh}} \sum_{k \in \mathcal{H}_i} x_{t+\frac{3}{2}}^{(k)}$ 

end for

**Output:**  $\{x_T^{(i)}\}_{i=1}^{n_s}$ 

# 4. Theoretical Analysis

## 4.1. Convergence of HSL

In this section, we analyze the convergence behavior of HSL under standard assumptions commonly used in stochastic first-order methods. Specifically, we assume the following:

1) **Smoothness:** capturing how the gradients change across the loss landscape, (2) **Stochastic Noise Bound:** controlling the randomness in gradient estimates due to mini-batch sampling, and (3) **Bounded Heterogeneity:** quantifying how much local objective functions deviate from the global objective. Formally, we state:

**Assumption 4.1** (Smoothness). For each  $i \in [n_s]$ , the function  $f^{(i)}: \mathbb{R}^d \to \mathbb{R}$  is differentiable, and there exists a constant  $L < \infty$  such that for all  $x, y \in \mathbb{R}^d$ :

$$\|\nabla f^{(i)}(y) - \nabla f^{(i)}(x)\| \le L\|y - x\|.$$

**Assumption 4.2** (Bounded Stochastic Noise). There exists a constant  $\sigma < \infty$  such that for all  $i \in [n]$  and  $x \in \mathbb{R}^d$ :

$$\mathbb{E}_{\xi \sim D^{(i)}} \big[ \|\nabla f(x,\xi) - \nabla f^{(i)}(x)\|^2 \big] \le \sigma^2.$$

**Assumption 4.3** (Bounded Heterogeneity). There exists a constant  $\mathcal{H} < \infty$  such that for all  $x \in \mathbb{R}^d$ :

$$\frac{1}{n_s} \sum_{i \in [n_s]} \|\nabla f^{(i)}(x) - \nabla F(x)\|^2 \le \mathcal{H}^2,$$

where  $F(\mathbf{x})$  denotes the average of the local objective functions  $f^{(i)}(x)$ . The term  $\sigma$  captures variance introduced by stochastic gradients, while  $\mathcal{H}$  quantifies the heterogeneity arising from the non-iid data distribution.

We now present our main convergence result.

**Theorem 4.4.** Consider Algorithm 1 under the above assumptions. Let the initial optimization gap be:

$$\Delta_0 := F(x_0) - \min_{x \in \mathbb{R}^d} F(x).$$

Then, for any  $T \ge 1$ , with  $n_s \ge 2$  spokes, a communication budget of  $b_{sh} \ge 1$ , and  $n_h \ge 2$  hubs with budgets  $b_{hs} \ge 1$ ,  $b_{hh} \ge 1$ , selecting the step size as:

$$\gamma \in \Theta\left(\min\left\{\sqrt{\frac{n_s\Delta_0}{TL((1+\beta')\sigma^2+\beta'\mathcal{H}^2)}}\right.\right.\right.$$

$$\sqrt[3]{\frac{\Delta_0}{TL^2\beta_{HSL}(\sigma^2+\mathcal{H}^2)}}, \frac{1}{L}\right\}\right).$$

we have

$$\frac{1}{n_s T} \sum_{t=0}^{T-1} \sum_{i=1}^{n_s} \mathbb{E}\left[\left\|\nabla F(x_t^{(i)})\right\|^2\right] 
\in \mathcal{O}\left(\sqrt{\frac{L\Delta_0}{Tn_s}}((1+\beta')\sigma^2 + \beta'\mathcal{H}^2)\right) 
+ \sqrt[3]{\frac{L^2\beta_{HSL}\Delta_0^2(\sigma^2 + \mathcal{H}^2)}{T^2}} + \frac{L\Delta_0}{T}\right).$$

where

$$\beta_{HSL} := \beta_{sh} \beta_{hh} \beta_{hs}$$
$$\beta' := \frac{1}{2} \left[ \beta_{HSL} + \frac{n_s}{n_h} \beta_{hs} \left( 1 + \beta_{hh} \right) \right]$$

and

$$\begin{split} \beta_{hs} &:= \frac{1}{b_{hs}} \left( 1 - \frac{b_{hs} - 1}{n_s - 1} \right) \\ \beta_{hh} &:= \frac{1}{b_{hh}} \left( 1 - \left( 1 - \frac{b_{hh}}{n_h - 1} \right)^{n_h} \right) - \frac{1}{n_h - 1} \\ \beta_{sh} &:= \frac{1}{b_{sh}} \left( 1 - \frac{b_{sh} - 1}{n_h - 1} \right) \end{split}$$

Each of  $\beta_{hs}$ ,  $\beta_{hh}$ ,  $\beta_{sh}$  represents an upper bound on the CDR for the respective mixing steps in HSL. For details, refer to Lemma A.2 in the Appendix.

# 4.2. Consensus Distance Ratio in HSL

Lemma A.2 relates the expected consensus distance (CD) before and after each mixing step in HSL. Specifically, the following inequalities hold:

$$\frac{\mathbb{E}[\mathrm{CD}_{t+\frac{2}{4}}]}{\mathbb{E}[\mathrm{CD}_{t+\frac{1}{4}}]} \leq \beta_{hs}, \frac{\mathbb{E}[\mathrm{CD}_{t+\frac{3}{4}}]}{\mathbb{E}[\mathrm{CD}_{t+\frac{2}{4}}]} \leq \beta_{hh}, \frac{\mathbb{E}[\mathrm{CD}_{t+1}]}{\mathbb{E}[\mathrm{CD}_{t+\frac{3}{4}}]} \leq \beta_{sh},$$

which combine to yield:  $\frac{\mathbb{E}[CD_{t+1}]}{\mathbb{E}[CD_{t+\frac{1}{4}}]} \leq \beta_{HSL}$ ,

These  $\beta$ -values represent upper bounds on the expected consensus distance ratio (CDR) at each stage of mixing and can be tuned by adjusting the budgets  $b_{hs}$ ,  $b_{hh}$ ,  $b_{sh}$ .

In *EL Local*, the CDR satisfies  $\beta_{EL} \in \mathcal{O}(1/k)$ , meaning that the only guaranteed way to improve mixing is by increasing the node budget k. In contrast, hubs improve mixing without increasing the burden on individual spokes, enabling a more efficient communication-cost trade-off.

# 5. Evaluation

Figure 2 illustrates the final spoke test-accuracy distributions and consensus distance ratios (CDR) for CIFAR-10 with 100 and 200 spokes. Similarly, Figures 3 and 6 present these results for AG News. We use candle plots to visualize final accuracy, where the candle body represents the interquartile range (25th–75th percentile) and wicks mark the minimum and maximum across the spokes. For each configuration shown in the candle plot, a corresponding CDR plot shows the ratio of the post-mixing consensus distance to the premixing distance in each round. Since this ratio is typically less than 1, we report its negative logarithm making higher values reflect stronger mixing efficiency.

**Experimental Setup:** To assess the effectiveness of HSL, we evaluate its performance on two machine learning tasks: image classification on CIFAR-10 (Krizhevsky & Hinton, 2009) and text classification on AG News (Zhang et al., 2015). In both datasets, data is distributed across  $n_s$  spokes in a non-iid manner using a Dirichlet distribution with  $\alpha = 1$ . Following our discussion on mixing effectiveness, we compare HSL against ELL, monitoring both test accuracy and consensus distance across 500 communication rounds.

For CIFAR-10, each spoke trains a simple CNN with two convolutional layers, a pooling layer, and two fully connected layers (4.2M parameters). Training is done using stochastic gradient descent (SGD) with a constant learning rate of 0.01, a batch size of 128, and three local mini-batch updates per communication round.

For AG News, each spoke trains a lightweight Transformer model with 12.9M parameters, consisting of an embedding layer, two Transformer encoder layers, and a final classification head. The vocabulary size is 95,812, and input sequences are padded to a maximum length of 207 tokens. Training is performed with SGD using a learning rate of 0.05, a batch size of 64, and five local iterations per round.

All experiments were conducted on NVIDIA A100 GPUs, with a maximum memory usage of 40 GB across all configurations. We evaluate multiple configurations of HSL and ELL and present the results in the following sections.

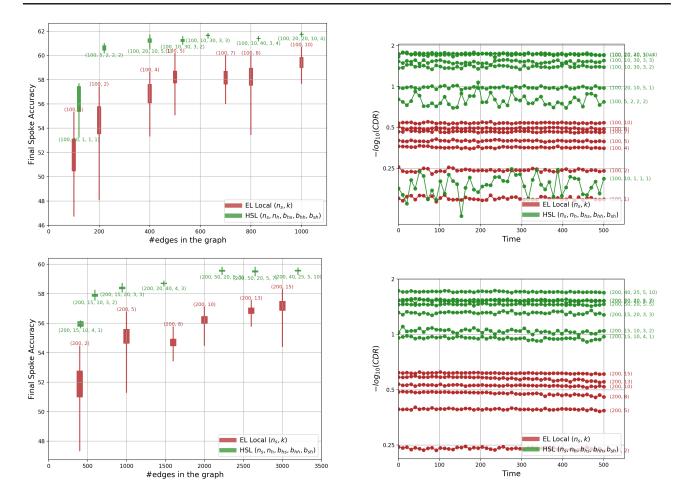


Figure 2. HSL vs. ELL on CIFAR-10 ( $n_s=100,200$ ). (Top)  $n_s=100$ : (Left) Final accuracy vs. total edges (budget). Candle bodies represent the interquartile range, and wicks indicate min/max values. HSL consistently achieves higher accuracy at lower budgets, demonstrating its efficiency and scalability. (Right) Mixing efficiency over 500 rounds, measured via  $-\log(\text{CDR})$ , where higher values indicate stronger mixing. HSL achieves superior mixing at all budgets, explaining its improved accuracy. (Bottom)  $n_s=200$ : HSL maintains its advantage, matching ELL's 3000-edge accuracy with only a third of the budget. The CDR plot further highlights HSL's superior mixing, where HSL with just 595 edges achieves better consensus than ELL with 3000 edges, reinforcing its scalability.

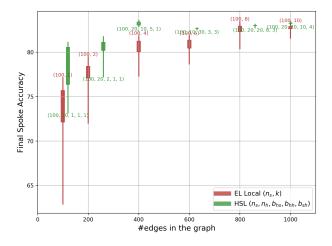
Comparison with ELL Under Varying Budgets: We compare HSL and ELL for 100 and 200 spokes on CIFAR-10 and AG News, varying the total communication budget (the number of directed edges in the graph). After 500 rounds of training, we report each spoke's final test accuracy against this budget. The total edge count serves as a fair metric because both the system's per-round communication (number of messages) and computation (volume of model aggregation) scale proportionally with the number of directed edges.

For ELL, the configuration is represented by the tuple  $(n_s, k)$ , where  $n_s$  is the number of spokes and k is the outdegree per spoke in each round. The total number of directed edges in this case is given by  $n_s \cdot k$ .

For HSL, the configuration is defined by the tuple

 $(n_s,n_h,b_{hs},b_{hh},b_{sh})$ , where  $n_s$  represents the number of spokes,  $n_h$  denotes the number of hubs, and  $b_{hs},b_{hh}$ , and  $b_{sh}$  represent the hub-spoke, hub-hub, and spoke-hub budgets, respectively. The total number of directed edges in HSL is calculated as  $n_h \cdot b_{hs} + n_s \cdot b_{sh} + n_h \cdot b_{hh}$ .

We observe that HSL consistently outperforms ELL across various budgets, datasets, and spoke counts  $(n_s=100,200).$  In Figure 2, the HSL configuration (100,5,2,2,2) uses just 220 edges  $(5\times 2+5\times 2+100\times 2)$  yet matches the performance of ELL (100,10) with 1000 edges. Although a 220-edge HSL graph with 5 hubs is highly sparse, it still maintains strong accuracy through efficient mixing. HSL's design enables effective information sharing even under tight budget constraints, allowing scalability to larger networks without a proportional rise in communication and computation costs. Only the extreme low-budget setting



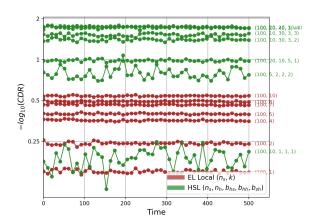


Figure 3. HSL vs. ELL on AG News ( $n_s = 100$ ). The left plot shows the final accuracy distribution, while the right plot presents the consensus distance ratio (CDR). HSL with only 400 edges matches ELL's performance with 1000 edges. The CDR plot continues to confirm the superior mixing efficiency of HSL over ELL.

 $(b_{hs}=b_{hh}=b_{sh}=1)$  shows a notably lower candle, but such configurations are rarely practical. By contrast, ELL's accuracy drops more sharply as edge count decreases. With a moderate budget ( $\geq 400$ )—letting each hub connect to at least  $b_{hs}=10$  spokes (approximately 10% of all spokes)—HSL surpasses ELL's best configuration at k=10 under a 1000-edge budget.

For  $n_s=200$ , HSL's advantage becomes even more pronounced, highlighting its cost-effectiveness for larger networks. While ELL struggles to exceed 58% accuracy even at 3000 edges, HSL attains comparable accuracy with only 945 edges. The CDR plots clarify HSL's superior mixing properties: except for the minimal 10-hub case  $(b_{hs}=b_{hh}=b_{sh}=1)$ , all HSL configurations achieve stronger mixing than ELL's 1000-edge setup. As a result, HSL consistently reduces model variance per round, leading to tighter final accuracy distributions, as reflected in its shorter candle plots.

Even on the AG News dataset, HSL maintains its efficiency, matching ELL's performance while requiring significantly fewer edges. —requiring just 400 compared to ELL's 1000 at  $n_s=100$  and only 410 versus 3000 at  $n_s=200$ , as shown in Figures 3 and 6.

Accuracy-Time Plot and Baseline Comparisons It is important to note that HSL not only achieves higher final test accuracy but also sustains this advantage throughout training. Figure 4 shows a typical test accuracy trajectory over training rounds for HSL and ELL at an equal budget of 400 edges for 100 spokes on the AG News dataset. For comparison, we also include FedAvg and two additional decentralized baselines—Torus and Erdős-Rényi(Erdos & Rényi, 1984) both configured with 400 directed edges.

FedAvg achieves the fastest convergence, leveraging exact consensus to ensure ideal mixing even with just 200 edges. This advantage arises from the presence of a centralized server that efficiently coordinates updates.

Torus represents a structured, low-degree topology where each node connects to a fixed set of neighbors in a 2D grid-like pattern. As shown in Figure 4, it reaches 70% accuracy, serving as a baseline for decentralized learning. EL Local and Erdős-Rényi provide competitive alternatives at equal budgets, as both leverage dynamically sampled random graphs for communication. However, they differ in degree distribution—EL Local maintains a fixed, uniform out-degree per node, whereas Erdős-Rényi constructs a random graph with edges assigned independently with a fixed probability.

HSL consistently outperforms the decentralized baselines and closely tracks FedAvg, even briefly surpassing it in the final rounds, as seen in Figure 4. This momentary advantage may, however, stem from inherent randomness in decentralized updates.

In summary, these experiments demonstrate that HSL outperforms or closely matches ELL and other decentralized baselines, all while using significantly fewer communication edges. The improved mixing in HSL is evident from higher test accuracy, reduced variance across spokes, and faster convergence in most settings. To further reinforce these findings, we extend our evaluation with a mathematical simulation that directly examines the mixing properties of HSL, ELL, and Erdos-Renyi serving as our baseline.

To further substantiate these findings, we complement our empirical evaluation with a mathematical simulation. Specifically, we analyze the mixing properties of HSL and ELL

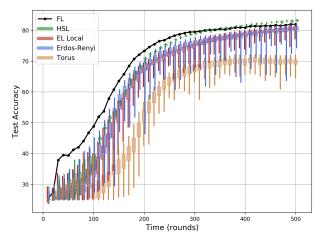


Figure 4. Test Accuracy vs. Training Rounds on AG News  $(n_s=100)$ . We compare FedAvg (200 edges) with HSL and other decentralized methods (400 edges). HSL consistently outperforms decentralized baselines and closely follows FedAvg. Torus reaches 70% accuracy, serving as a baseline, while EL Local and Erdős-Rényi exhibit similar trends due to their dynamic graphs.

through a random graph sampling process. At each round, we generate a fresh random graph according to the configuration and compute the effective mixing matrix. The spectral gap of this matrix, defined as the difference between the largest and second-largest eigenvalues of the transition matrix, serves as a widely accepted measure of graph connectivity (Lovász, 1993; Chung, 1997). A larger spectral gap indicates faster mixing and improved convergence properties in decentralized learning.

We repeat this process for 1000 rounds, averaging the spectral gap across all realizations. Figure 5 illustrates the results, conclusively demonstrating that HSL consistently achieves a significantly higher spectral gap than ELL for all comparable budgets between 0 and 2200, where both methods begin to converge. Notably, HSL not only converges faster but also attains a higher final spectral gap value, reinforcing its superior mixing efficiency and scalability.

## 6. Discussion

In this work, we introduced HSL as a scalable and resilient framework that merges the strengths of Federated Learning (FL) and decentralized Peer-to-Peer Learning (P2PL). By structuring communication into hubs and spokes, HSL enables efficient model mixing at the hub level while reducing the communication burden on individual spokes. Our empirical results validate this design: HSL consistently outperforms or matches ELL while requiring significantly fewer communication edges.

Several key insights emerge from our evaluation. First, the hierarchical structure with decentralized hubs at the top allows HSL to achieve efficient information propagation

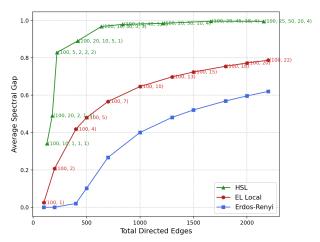


Figure 5. Average spectral gap variation of HSL, ELL, and Erdős-Rényi with total directed edges. The spectral gap was computed from the effective mixing matrix, sampled at each round for 1000 rounds with 100 spokes, and then averaged. Erdős-Rényi serves as a reference baseline for comparison. The results reaffirm HSL's superior mixing efficiency, even in mathematical simulations.

without the bottlenecks of FL or the increasing connectivity demands of fully decentralized methods. Second, HSL maintains its efficiency as network size scales, demonstrating strong performance even at larger spoke counts. Beyond these, two additional advantages stand out. The ability of HSL to deliver strong results even with low budgets and sparse connections suggests future strategies where only a subset of spokes participate in each update round, allowing others to conserve resources while maintaining overall model consistency. Additionally, the *receiver-driven* selection in HSL — where nodes independently choose whom to receive updates from, when hubs and spokes interact—enhances robustness, a property that provides natural resilience against targeted attacks, a promising avenue for future security-focused analyses.

HSL achieves strong mixing at lower budgets, making it well-suited for hybrid systems with both low-power edge devices and high-performance cloud servers. Overall, HSL bridges the gap between FL and fully decentralized learning, offering a scalable, resilient, and communication-efficient framework for large-scale collaborative learning systems.

# References

- Abad, M. S. H., Ozfatura, E., Gunduz, D., and Ercetin, O. Hierarchical federated learning across heterogeneous cellular networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8866–8870. IEEE, 2020.
- Assran, M., Loizou, N., Ballas, N., and Rabbat, M. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pp. 344–353. PMLR, 2019.
- Beltrán, E. T. M., Pérez, M. Q., Sánchez, P. M. S., Bernal, S. L., Bovet, G., Pérez, M. G., Pérez, G. M., and Celdrán, A. H. Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, 2023.
- Chung, F. R. K. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI, 1997.
- De Vos, M., Farhadkhani, S., Guerraoui, R., Kermarrec, A.-m., Pires, R., and Sharma, R. Epidemic learning: Boosting decentralized learning with randomized communication. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36132–36164. Curran Associates, Inc., 2023.
- Dhasade, A., Kermarrec, A.-M., Pires, R., Sharma, R., and Vujasinovic, M. Decentralized learning made easy with decentralizepy. In *Proceedings of the 3rd Workshop on Machine Learning and Systems*, pp. 34–41, 2023.
- Erdos, P. L. and Rényi, A. On the evolution of random graphs. *Transactions of the American Mathematical Society*, 286:257–257, 1984. URL https://api.semanticscholar.org/CorpusID:6829589.
- Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for federated learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.
- Koloskova, A., Stich, S. U., and Jaggi, M. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pp. 5301–5310. PMLR, 2020.
- Kong, L., Lin, T., Koloskova, A., Jaggi, M., and Stich, S. Consensus control for decentralized deep learning.

- In *International Conference on Machine Learning*, pp. 5686–5696. PMLR, 2021.
- Korkmaz, C., Kocas, H. E., Uysal, A., Masry, A., Ozkasap, O., and Akgun, B. Chain fl: Decentralized federated machine learning via blockchain. In 2020 Second international conference on blockchain computing and applications (BCCA), pp. 140–146. IEEE, 2020.
- Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neu*ral Information Processing Systems (NeurIPS 2017), pp. 5330–5340. Curran Associates, Inc., 2017.
- Liu, L., Zhang, J., Song, S., and Letaief, K. B. Client-edgecloud hierarchical federated learning. In *ICC* 2020-2020 *IEEE international conference on communications (ICC)*, pp. 1–6. IEEE, 2020.
- Lovász, L. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2(1-46):4, 1993.
- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the* 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), pp. 1273–1282. PMLR, 2017.
- Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., and Hwang, W.-J. Federated learning for smart healthcare: A survey. 55(3), February 2022. ISSN 0360-0300. doi: 10.1145/3501296.
- Qin, Z., Yan, X., Zhou, M., and Deng, S. Blockdfl: A blockchain-based fully decentralized peer-to-peer federated learning framework. In *Proceedings of the ACM on Web Conference 2024*, pp. 2914–2925, 2024.
- Wu, J., Dong, F., Leung, H., Zhu, Z., Zhou, J., and Drew, S. Topology-aware federated learning in edge computing: A comprehensive survey. *ACM Comput. Surv.*, 56(10), June 2024. ISSN 0360-0300. doi: 10.1145/3659205.
- Ye, R., Xu, M., Wang, J., Xu, C., Chen, S., and Wang, Y. Feddisco: Federated learning with discrepancy-aware collaboration. In *International Conference on Machine Learning*, pp. 39879–39902. PMLR, 2023.
- Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

## A. Useful Lemmas

In this section, we present several key lemmas that are instrumental in establishing the convergence results and analyzing the properties of HSL.

**Lemma A.1.** Average preservation in expectation: The average of the models across the network remains preserved in expectation through all stages of the process:

$$\mathbb{E}\left[\overline{x}_{t+1}\right] = \mathbb{E}\left[\overline{x}_{t+\frac{3}{2}}\right] = \mathbb{E}\left[\overline{x}_{t+\frac{1}{2}}\right] = \mathbb{E}\left[\overline{x}_{t+\frac{1}{2}}\right].$$

**Lemma A.2.** For each stage of mixing in HSL, the consensus distance is recursively bounded as follows:

## 1. Spoke-to-Hub Push:

$$\frac{1}{n_h^2} \sum_{\substack{i,j \in [n_h] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(j)} \right\|^2 \right] \leq \frac{\beta_{hs}}{n_s^2} \sum_{\substack{i,j \in [n_s] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)} \right\|^2 \right],$$

where

$$\beta_{hs} = \frac{1}{b_{hs}} \left[ 1 - \frac{b_{hs} - 1}{n_s - 1} \right].$$

## 2. Hub Gossip:

$$\frac{1}{n_h^2} \sum_{\substack{i,j \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{3}{4}}^{(i)} - x_{t+\frac{3}{4}}^{(j)} \right\|^2 \right] \ \leq \ \frac{\beta_{hh}}{n_h^2} \sum_{\substack{i,j \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(j)} \right\|^2 \right],$$

where

$$\beta_{hh} = \frac{1}{b_{hh}} \left( 1 - \left( 1 - \frac{b_{hh}}{n_h - 1} \right)^{n_h} \right) - \frac{1}{n_h - 1}.$$

#### 3. Hub-to-Spoke Pull:

$$\frac{1}{n_s^2} \sum_{\substack{i,j \in [n_s] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+1}^{(i)} - x_{t+1}^{(j)} \right\|^2 \right] \leq \frac{\beta_{sh}}{n_h^2} \sum_{\substack{i,j \in [n_h] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{3}{4}}^{(i)} - x_{t+\frac{3}{4}}^{(j)} \right\|^2 \right],$$

where

$$\beta_{sh} = \frac{1}{b_{sh}} \left[ 1 - \frac{b_{sh} - 1}{n_h - 1} \right].$$

4. Final Consensus Bound: Combining the above three stages, the consensus distance at  $x_{t+1}$  is bounded in terms of the distance at  $x_{t+\frac{1}{4}}$ :

$$\frac{1}{n_s^2} \sum_{\substack{i,j \in [n_s]\\i \neq j}} \mathbb{E} \left[ \left\| x_{t+1}^{(i)} - x_{t+1}^{(j)} \right\|^2 \right] \leq \beta_{HSL} \cdot \frac{1}{n_s^2} \sum_{\substack{i,j \in [n_s]\\i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)} \right\|^2 \right],$$

where

$$\beta_{HSL} = \beta_{hs} \cdot \beta_{hh} \cdot \beta_{sh}.$$

**Lemma A.3.** For each stage of aggregation in HSL, the expected deviation of the average model is bounded as follows:

# 1. Spoke-to-Hub Push:

$$\mathbb{E}\left[\left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right] = \frac{\beta_{hs}}{n_{s}n_{h}} \sum_{i \in [n_{s}]} \mathbb{E}\left[\left\|x_{t+\frac{1}{4}}^{(i)} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right].$$

2. Hub Gossip:

$$\mathbb{E}\left[\left\|x_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right] \leq \frac{\beta_{hh}}{n_{h}^{2}} \sum_{i \in [n_{h}]} \mathbb{E}\left[\left\|x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right].$$

3. Hub-to-Spoke Pull:

$$\mathbb{E}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right] = \frac{\beta_{sh}}{n_{s}n_{h}} \sum_{i \in [n_{h}]} \mathbb{E}\left[\left\|x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right].$$

**Lemma A.4.** The expected consensus distance and gradient variance across spokes are bounded as follows:

1. Consensus Distance Bound:

$$\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] \le 20 \frac{1 + 3\beta_{HSL}}{(1 - \beta_{HSL})^2} \beta_{HSL} \gamma^2 (\sigma^2 + \mathcal{H}^2).$$

2. Gradient Variance Bound:

$$\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| g_t^{(i)} - g_t^{(j)} \right\|^2 \right] \le 15(\sigma^2 + \mathcal{H}^2).$$

**Lemma A.5.** The expected gradient norm of the global objective satisfies the following upper bound:

$$\mathbb{E}\left[\left\|\nabla F(\bar{x}_{t})\right\|^{2}\right] \leq \frac{2}{\gamma} \mathbb{E}\left[F(\bar{x}_{t}) - F(\bar{x}_{t+1})\right] + \frac{L}{2n^{2}} \sum_{i,j} \mathbb{E}\left[\left\|x_{t}^{(i)} - x_{t}^{(j)}\right\|^{2}\right] + \frac{4L\gamma\sigma^{2}}{n} + \frac{4L}{\gamma} \mathbb{E}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2} + \left\|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2} + \left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right].$$

**Lemma A.6.** Variance decomposition: For any set of vectors  $\{x_t^{(i)}, i \in [n_s]\}$ ,

$$\frac{1}{n_s} \sum_{i} \left\| x_t^{(i)} - \bar{x}_t \right\|^2 = \frac{1}{2} \frac{1}{n_s^2} \sum_{\substack{i,j\\i \neq j}} \left\| x_t^{(i)} - x_t^{(j)} \right\|^2.$$

## **B. Proof of Theorem 1**

*Proof.* Recall that for any vectors  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ , Jensen's inequality (for the  $\ell_2$ -norm) states:

$$\|\mathbf{a} + \mathbf{b}\|^2 \le 2 \|\mathbf{a}\|^2 + 2 \|\mathbf{b}\|^2.$$

We apply this inequality with  $\mathbf{a} = \nabla F(\bar{x_t})$  and  $\mathbf{b} = \nabla F(x_t^{(i)}) - \nabla F(\bar{x_t})$ . For any  $i \in [n_s]$ , we obtain

$$\mathbb{E}\left[\left\|\nabla F(x_t^{(i)})\right\|^2\right] = \mathbb{E}\left[\left\|\nabla F(\bar{x}_t) + \left(\nabla F(x_t^{(i)}) - \nabla F(\bar{x}_t)\right)\right\|^2\right]$$

$$\leq 2\mathbb{E}\left[\left\|\nabla F(\bar{x}_t)\right\|^2\right] + 2\mathbb{E}\left[\left\|\nabla F(x_t^{(i)}) - \nabla F(\bar{x}_t)\right\|^2\right].$$

Using Assumption 4.1 (*Smoothness*), which implies  $\|\nabla F(x) - \nabla F(y)\| \le L \|x - y\|$ , we further bound the second term to obtain:

$$\mathbb{E}\left[\left\|\nabla F(x_t^{(i)})\right\|^2\right] \leq 2\,\mathbb{E}\left[\left\|\nabla F(\bar{x_t})\right\|^2\right] \; + \; 2\,L^2\mathbb{E}\left[\left\|x_t^{(i)} - \bar{x_t}\right\|^2\right].$$

Next, we average over all  $i \in [n_s]$ :

$$\frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{E} \left[ \left\| \nabla F(x_t^{(i)}) \right\|^2 \right] \ \leq \ 2 \, \mathbb{E} \left[ \left\| \nabla F(\bar{x_t}) \right\|^2 \right] \ + \ \frac{2L^2}{n_s} \sum_{i=1}^{n_s} \mathbb{E} \left[ \left\| x_t^{(i)} - \bar{x_t} \right\|^2 \right].$$

Finally, making use of Lemma A.6, which states

$$\frac{1}{n_s} \sum_{i=1}^{n_s} \left\| x_t^{(i)} - \bar{x_t} \right\|^2 = \frac{1}{2n_s^2} \sum_{i,j \in [n_s]} \left\| x_t^{(i)} - x_t^{(j)} \right\|^2,$$

we get

$$\frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{E}\left[ \left\| \nabla F(x_t^{(i)}) \right\|^2 \right] \leq 2 \, \mathbb{E}\left[ \left\| \nabla F(\bar{x_t}) \right\|^2 \right] \, + \, \frac{L^2}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right].$$

Bounding the first term on the RHS using Lemma A.5, we further obtain:

$$\frac{1}{n_{s}} \sum_{i=1}^{n_{s}} \mathbb{E}\left[\left\|\nabla F(x_{t}^{(i)})\right\|^{2}\right] \leq \frac{4}{\gamma} \mathbb{E}\left[F(\bar{x}_{t}) - F(\bar{x}_{t+1})\right] + \frac{2L^{2}}{n_{s}^{2}} \sum_{i,j \in [n_{s}]} \mathbb{E}\left[\left\|x_{t}^{(i)} - x_{t}^{(j)}\right\|^{2}\right] + \frac{8L\gamma\sigma^{2}}{n_{s}} + \frac{8L}{\gamma} \mathbb{E}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2} + \left\|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2} + \left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right]. \tag{5}$$

Using Lemma A.2, we also have:

$$\begin{split} & \mathbb{E}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^2\right] \leq \frac{\beta_{sh}\beta_{hh}\beta_{hs}}{2n_s^3} \sum_{i,j \in [n_s]} \mathbb{E}\left[\left\|x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)}\right\|^2\right] \\ & \mathbb{E}\left[\left\|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^2\right] \leq \frac{\beta_{hh}\beta_{hs}}{2n_hn_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[\left\|x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)}\right\|^2\right] \\ & \mathbb{E}\left[\left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^2\right] \leq \frac{\beta_{hs}}{2n_hn_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[\left\|x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)}\right\|^2\right] \end{split}$$

Adding the above inequalites,

$$\mathbb{E}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right] + \mathbb{E}\left[\left\|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right] + \mathbb{E}\left[\left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right] \leq \frac{\beta'}{n_{s}} \frac{1}{n_{s}^{2}} \sum_{i,j \in [n_{s}]} \mathbb{E}\left[\left\|x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)}\right\|^{2}\right]$$

$$\tag{6}$$

where

$$\frac{\beta'}{n_s} = \frac{\beta_{hs}}{2n_h} + \frac{\beta_{hh}\beta_{hs}}{2n_h} + \frac{\beta_{sh}\beta_{hh}\beta_{hs}}{2n_s}$$

Remember the partial update step  $x_{t+\frac{1}{4}}^{(i)} \triangleq x_t^{(i)} - \gamma \, g_t^{(i)}.$  Thus,

$$\mathbb{E}\left[\left\|x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)}\right\|^{2}\right] = \mathbb{E}\left[\left\|x_{t}^{(i)} - \gamma g_{t}^{(i)} - x_{t}^{(j)} + \gamma g_{t}^{(j)}\right\|^{2}\right] \\
\leq 2 \mathbb{E}\left[\left\|x_{t}^{(i)} - x_{t}^{(j)}\right\|^{2}\right] + 2 \gamma^{2} \mathbb{E}\left[\left\|g_{t}^{(i)} - g_{t}^{(j)}\right\|^{2}\right]. \tag{7}$$

where we make use of Young's inequality.

Substituting 7 and 6 into 5, we get:

$$\frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{E}\left[\left\|\nabla F(x_t^{(i)})\right\|^2\right] \leq \frac{4}{\gamma} \mathbb{E}\left[F(\bar{x}_t) - F(\bar{x}_{t+1})\right] + \frac{8L\gamma\sigma^2}{n_s} + \left(2L^2 + \frac{16L\beta'}{\gamma n_s}\right) \frac{1}{n_s^2} \sum_{i=1}^{n_s} \mathbb{E}\left[\left\|x_t^{(i)} - x_t^{(j)}\right\|^2\right] + \frac{16L\gamma\beta'}{n_s} \frac{1}{n_s^2} \sum_{i=1}^{n_s} \mathbb{E}\left[\left\|g_t^{(i)} - g_t^{(j)}\right\|^2\right] \quad (8)$$

From Remark C.1, we have:  $\beta_{HSL} \leq 1 - \frac{1}{e}$  Therefore,

$$20\frac{1+3\beta_{HSL}}{(1-\beta_{HSL})^2} \le 500$$

We substitute this in Lemma A.4 to get

$$\frac{1}{n_s^2} \sum_{i=1}^{n_s} \mathbb{E}\left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] \le 500 \beta_{HSL} \gamma^2 (\sigma^2 + \mathcal{H}^2)$$

From Lemma A.4, we also have,

$$\frac{1}{n_s^2} \sum_{i=1}^{n_s} \mathbb{E}\left[ \left\| g_t^{(i)} - g_t^{(j)} \right\|^2 \right] \le 15(\sigma^2 + \mathcal{H}^2)$$

Substituting this in 8, we obtain:

$$\frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{E}\left[ \left\| \nabla F(x_t^{(i)}) \right\|^2 \right] \leq \frac{4}{\gamma} \mathbb{E}\left[ F(\bar{x}_t) - F(\bar{x}_{t+1}) \right] + \left( 2L^2 + \frac{16L\beta'}{\gamma n_s} \right) 500 \beta_{HSL} \gamma^2 (\sigma^2 + \mathcal{H}^2) + \frac{8L\gamma \sigma^2}{n_s} + \frac{240}{n_s} L\gamma \beta_{HSL} (\sigma^2 + \mathcal{H}^2)$$

Taking the average over  $t \in 0, ..., T - 1$ , we obtain:

$$\frac{1}{n_s T} \sum_{t=0}^{T-1} \sum_{i=1}^{n_s} \mathbb{E}\left[\left\|\nabla F(x_t^{(i)})\right\|^2\right] \leq \frac{4}{T\gamma} \Delta_0 + \frac{\gamma}{n_s} \left(16L\beta' 500\beta_{HSL}(\sigma^2 + \mathcal{H}^2) + 8L\sigma^2 + 240L\beta_{HSL}(\sigma^2 + \mathcal{H}^2)\right) + \gamma^2 \left(2L^2 500\beta_{HSL}(\sigma^2 + \mathcal{H}^2)\right) \\
\leq \frac{4}{T\gamma} \Delta_0 + \frac{8L\gamma}{n_s} \left((1 + 663\beta')\sigma^2 + 663\beta'\mathcal{H}^2\right) + \gamma^2 \left(1000L^2\beta_{HSL}(\sigma^2 + \mathcal{H}^2)\right) \tag{9}$$

Here, we make use of the fact that since  $\beta_{HSL} \leq 1 - \frac{1}{e}$ ,  $1000\beta_{HSL} \leq 663$  Now, setting

$$\gamma = min \left\{ \sqrt{\frac{n_s \Delta_0}{2TL((1 + 663\beta')\sigma^2 + 663\beta'\mathcal{H}^2)}}, \sqrt[3]{\frac{\Delta_0}{250TL^2\beta_{HSL}(\sigma^2 + \mathcal{H}^2)}}, \frac{1}{20L} \right\}$$
(10)

we have

$$\frac{1}{\gamma} = max \left\{ \sqrt{\frac{2TL((1+663\beta')\sigma^2 + 663\beta'\mathcal{H}^2)}{n_s\Delta_0}}, \sqrt[3]{\frac{250TL^2\beta_{HSL}(\sigma^2 + \mathcal{H}^2)}{\Delta_0}}, 20L \right\}$$

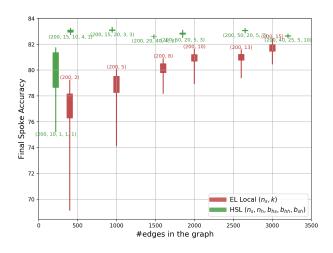
$$\leq \sqrt{\frac{2TL((1+663\beta')\sigma^2 + 663\beta'\mathcal{H}^2)}{n_s\Delta_0}} + \sqrt[3]{\frac{250TL^2\beta_{HSL}(\sigma^2 + \mathcal{H}^2)}{\Delta_0}} + 20L \tag{11}$$

Plugging equation 11 and equation 10 in equation 9 we obtain

$$\begin{split} \frac{1}{n_s T} \sum_{t=0}^{T-1} \sum_{i=1}^{n_s} \mathbb{E} \left[ \left\| \nabla F(x_t^{(i)}) \right\|^2 \right] &\leq 8 \sqrt{\frac{2L\Delta_0}{Tn_s}} (1 + 663\beta') \sigma^2 + 663\beta' \mathcal{H}^2 + 51 \sqrt[3]{\frac{L^2 \beta_{HSL} \Delta_0^2 (\sigma^2 + \mathcal{H}^2)}{T^2}} + \frac{80L\Delta_0}{T} \right] \\ &\in \mathcal{O} \left( \sqrt{\frac{L\Delta_0}{Tn_s}} ((1 + \beta') \sigma^2 + \beta' \mathcal{H}^2) + \sqrt[3]{\frac{L^2 \beta_{HSL} \Delta_0^2 (\sigma^2 + \mathcal{H}^2)}{T^2}} + \frac{L\Delta_0}{T} \right) \end{split}$$

Here, we use the simplification that  $\frac{2000}{250^{\frac{2}{3}}} < 51$ . This completes the derivation of the stated bound.

# C. Additional Results and Remarks



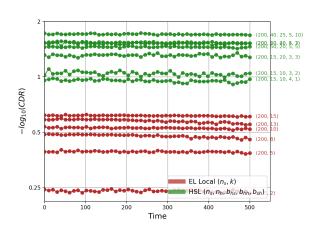


Figure 6. HSL vs. ELL on AG News ( $n_s = 200$ ). Final accuracy distribution on the left and consensus distance ratio (CDR) on the right. HSL with only 410 edges matches ELL's performance with 3000 edges. The CDR plot continues to confirm the superior mixing efficiency of HSL over ELL

Remark C.1. Note that  $\beta_{hh}$ , computed in Lemma A.2 is decreasing in  $b_{hh}$  and increasing in  $n_h$ , therefore, for any  $b_{hh} \ge 1$ , and  $n_h \ge 2$  we have

$$\begin{aligned} \beta_{hh} \Big|_{n < \infty} &\leq \lim_{n \to \infty} \beta_1 \\ &= \lim_{n \to \infty} \left( 1 - \left( 1 - \frac{1}{n-1} \right)^n - \frac{1}{n-1} \right) \\ &= 1 - \frac{1}{e}, \end{aligned}$$

where  $\beta_1$  is the  $\beta$  at  $b_{hh}=1$ . e is Euler's Number and we used the fact that  $\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n=\frac{1}{e}$ . We also have  $\beta_{hs}\leq 1$  and  $\beta_{sh}\leq 1$ . Multiplying these, we get  $\beta_{HSL}\leq 1-\frac{1}{e}$ . Remark C.2.

$$\beta_{hs} = \frac{1}{b_{hs}} \left( 1 - \frac{b_{hs} - 1}{n_s - 1} \right) \tag{12}$$

$$\beta_{hh} \le 1 - \frac{1}{e} \tag{13}$$

$$\beta_{hs} \le \frac{1}{b_{sh}} \left( 1 - \frac{b_{hs} - 1}{n_h - 1} \right) \tag{14}$$

By combining (12) and (14),

$$\beta_{hs}\beta_{sh} \le \frac{1}{b_{hs}b_{sh}} \frac{n_s - b_{hs}}{n_s - 1} \frac{n_h - b_h s}{n_h - 1}$$
$$= \frac{1}{b_{hs}b_{sh}} \frac{\left(1 - \frac{b_{hs}}{n_s}\right)}{\left(1 - \frac{1}{n_s}\right)} \frac{\left(1 - \frac{b_{sh}}{n_h}\right)}{\left(1 - \frac{1}{n_h}\right)}$$

If we have  $b_{sh}n_s \leq b_{hs}n_h$ 

$$\beta_{hs}\beta_{sh} \le \frac{n_s}{b_{hs}b_{sh}n_h} \frac{\left(1 - \frac{b_{hs}}{n_s}\right)}{\left(1 - \frac{1}{n_s}\right)} \frac{\left(1 - \frac{b_{hs}}{n_s}\right)}{\left(1 - \frac{1}{n_h}\right)}$$
$$= \frac{n_s}{n_h} \frac{\left(\frac{1}{b_{hs}} - \frac{1}{n_s}\right)^2}{\left(1 - \frac{1}{n_s}\right)\left(1 - \frac{1}{n_h}\right)}$$

This is decreasing in  $b_{hs}$ . For  $b_{sh}=1$ , (lowest spoke budget),  $\beta_{hs}=\frac{n_s}{n_h}$ 

$$\beta_{hs}\beta_{sh} \le \frac{n_s}{n_h} \frac{(n_h - 1)^2}{n_s^2} \frac{n_s}{n_s - 1} \frac{n_h}{n_h - 1}$$

$$= \frac{n_h - 1}{n_s - 1} < \frac{n_e}{n_s}$$

$$\beta_{HSL} \le \frac{n_h}{n_s} \left(1 - \frac{1}{e}\right)$$

Therefore, we guarantee lower upper bound on  $\beta_{HSL}$  as compared to  $\beta_{EL}$  under the condition where  $n_h \cdot b_{hs} \geq n_s \cdot b_{sh}$ 

# **D. Proof of Lemmas**

## D.1. Proof of Lemma A.1

Here, we prove the average preservation in expectation property of HSL.

$$\mathbb{E}\big[\overline{x}_{t+1}\big] = \mathbb{E}\big[\overline{x}_{t+\frac{3}{4}}\big] = \mathbb{E}\big[\overline{x}_{t+\frac{2}{4}}\big] = \mathbb{E}\big[\overline{x}_{t+\frac{1}{4}}\big].$$

*Proof.* From the system dynamics, we have:

$$X_{t+\frac{2}{4}} = W_{hs} X_{t+\frac{1}{4}},$$

where  $W_{hs}$  is independent of  $X_{t+\frac{1}{4}}$ . Taking expectations:

$$\mathbb{E}\left[X_{t+\frac{2}{4}}\right] = \mathbb{E}\left[W_{hs}\right]\mathbb{E}\left[X_{t+\frac{1}{4}}\right].$$

By the construction of  $W_{hs}$  as row-stochastic, we have:

$$\sum_{j=1}^{n_s} W_{hs}^{(i,j)} = 1 \quad \text{for all hubs } i.$$

Taking expectations and using symmetry (equal probability for all elements), let  $c = \mathbb{E}[W_{hs}^{(i,j)}]$ . Then:

$$\sum_{j=1}^{n_s} \mathbb{E}\big[W_{hs}^{(i,j)}\big] = 1 \implies n_s c = 1 \implies c = \frac{1}{n_s}.$$

Thus:

$$\mathbb{E}[W_{hs}] = \frac{1}{n_s} \mathbf{1}_{n_h} \mathbf{1}_{n_s}^T,$$

where  $\mathbf{1}_{n_h}$  and  $\mathbf{1}_{n_s}$  are column vectors of ones of dimension  $n_h$  and  $n_s$ , respectively.

Substituting, we get:

$$\mathbb{E}\left[X_{t+\frac{2}{4}}\right] = \frac{1}{n_s} \mathbf{1}_{n_s} \mathbf{1}_{n_s}^T \mathbf{1}_{n_s} \overline{x}_{t+\frac{1}{4}}$$
$$= \mathbf{1}_{n_h} \overline{x}_{t+\frac{1}{4}}$$
$$= \mathbb{E}\left[X_{t+\frac{1}{4}}\right]$$

Here we use the fact that  $\mathbf{1}_{n_s}^T \mathbf{1}_{n_s} = n_s$ .

Thus,  $\mathbb{E}ig[\overline{x}_{t+\frac{2}{4}}ig] = \mathbb{E}ig[\overline{x}_{t+\frac{1}{4}}ig]$ 

Now, consider the second stage of aggregation post hub gossip.

$$X_{t+\frac{3}{4}} = W_h X_{t+\frac{2}{4}},$$

Let  $n_h$  be the total number of hubs, and let  $A^{(i)} = |\mathcal{A}_k|$  denote the in-degree of the *i*-th hub, where the outdegree of every hub is fixed to  $b_{hs}$ . For any hub  $i \in [n_h]$ , define  $I_j^{(i)}$  as the indicator function denoting whether the *j*-th hub is connected to hub *i*. Then we claim:

$$\mathbb{E}\left[x_{t+\frac{3}{4}}^{(i)}\right] = \mathbb{E}\left[\frac{1}{A^{(i)}+1}\left(x_{t+\frac{2}{4}}^{(i)} + \sum_{j\in[n_h]\setminus\{i\}} \mathcal{I}_j^{(i)} x_{t+\frac{2}{4}}^{(j)}\right)\right].$$

First, we take a conditional expectation on  $A^{(i)}$ :

$$\begin{split} \mathbb{E}\big[x_{t+\frac{3}{4}}^{(i)}\big] &= \ \mathbb{E}\Big[\mathbb{E}\Big[\frac{1}{A^{(i)}+1}\Big(x_{t+\frac{2}{4}}^{(i)} \ + \ \sum_{j\in[n_h]\backslash\{i\}}\mathcal{I}_j^{(i)}\,x_{t+\frac{2}{4}}^{(j)}\,\Big|\ A^{(i)}\Big]\Big] \\ &= \ \mathbb{E}\Big[\frac{1}{A^{(i)}+1}\Big(x_{t+\frac{2}{4}}^{(i)} \ + \ \sum_{j\in[n_h]\backslash\{i\}}\mathbb{E}[\mathcal{I}_j^{(i)}\mid A^{(i)}]\,x_{t+\frac{2}{4}}^{(j)}\Big)\Big]. \end{split}$$

Since each of the other  $n_h - 1$  hubs has the same probability of sending its value to hub i, we have

$$\mathbb{E}\big[I_j^{(i)} \mid A^{(i)}\big] = \frac{A^{(i)}}{n_h - 1}.$$

Thus,

$$\mathbb{E}\left[x_{t+\frac{3}{4}}^{(i)}\right] = \mathbb{E}\left[\frac{1}{A^{(i)}+1}\left(x_{t+\frac{2}{4}}^{(i)} + \frac{A^{(i)}}{n_h-1} \sum_{j \in [n_h] \setminus \{i\}} x_{t+\frac{2}{4}}^{(j)}\right)\right]$$
$$= \mathbb{E}\left[\frac{1}{A^{(i)}+1}\left(x_{t+\frac{2}{4}}^{(i)} + \frac{A^{(i)}}{n_h-1}\left(n_h \bar{x}_{t+\frac{2}{4}} - x_{t+\frac{2}{4}}^{(i)}\right)\right)\right],$$

where  $\bar{x}_{t+\frac{2}{4}} = \frac{1}{n_h} \sum_{j=1}^{n_h} x_{t+\frac{2}{4}}^{(j)}$ . Let

$$p = \mathbb{E}\Big[\frac{A^{(i)}}{A^{(i)}+1}\Big].$$

Collecting terms, it follows that

$$\mathbb{E}\big[x_{t+\frac{3}{4}}^{(i)}\big] \ = \ \frac{p\,n_h}{n_h-1}\,\bar{x}_{t+\frac{2}{4}} \ + \ \Big(1-\frac{p\,n_h}{n_h-1}\Big)\,x_{t+\frac{2}{4}}^{(i)}.$$

Averaging over all  $i \in [n_h]$  gives

$$\mathbb{E}\big[\bar{x}_{t+\frac{3}{4}}\big] = \mathbb{E}\big[\bar{x}_{t+\frac{2}{4}}\big].$$

Now, we consider the **last step of aggregation** where the spokes aggregate models received from the hubs.

From the system dynamics, we have:

$$X_{t+1} = W_{sh} X_{t+\frac{3}{4}},$$

where  $W_{sh}$  is independent of  $X_{t+\frac{3}{4}}$ . Taking expectations:

$$\mathbb{E}[X_{t+1}] = \mathbb{E}[W_{sh}] \, \mathbb{E}[X_{t+\frac{3}{2}}].$$

By the construction of  $W_{sh}$  as row-stochastic, we have:

$$\sum_{j=1}^{n_h} W_{sh}^{(i,j)} = 1 \quad \text{for all spokes } i.$$

Taking expectations and using symmetry (equal probability for all elements), let

$$c = \mathbb{E}[W_{sh}^{(i,j)}].$$

Then,

$$\sum_{j=1}^{n_h} \mathbb{E}\big[W_{sh}^{(i,j)}\big] = 1 \quad \Longrightarrow \quad n_h \, c = 1 \quad \Longrightarrow \quad c = \frac{1}{n_h}.$$

Thus,

$$\mathbb{E}[W_{sh}] = \frac{1}{n_h} \, \mathbf{1}_{n_s} \, \mathbf{1}_{n_h}^T,$$

where  $\mathbf{1}_{n_s}$  and  $\mathbf{1}_{n_h}$  are column vectors of ones of dimension  $n_s$  and  $n_h$ , respectively.

Substituting into the expectation, we get:

$$\mathbb{E}[X_{t+1}] = \frac{1}{n_h} \mathbf{1}_{n_s} \mathbf{1}_{n_h}^T \mathbf{1}_{n_h} \overline{x}_{t+\frac{3}{4}} = \mathbf{1}_{n_s} \overline{x}_{t+\frac{3}{4}},$$

since  $\mathbf{1}_{n_h}^T \mathbf{1}_{n_h} = n_h$ .

Therefore,

$$\mathbb{E}\big[\overline{x}_{t+1}\big] = \mathbb{E}\big[\overline{x}_{t+\frac{3}{4}}\big].$$

Thus, we conclude that

$$\mathbb{E}\left[\overline{x}_{t+1}\right] = \mathbb{E}\left[\overline{x}_{t+\frac{3}{4}}\right] = \mathbb{E}\left[\overline{x}_{t+\frac{2}{4}}\right] = \mathbb{E}\left[\overline{x}_{t+\frac{1}{4}}\right].$$

# D.2. Proof of Lemma A.2

Stage One mixing: SPoke-to-Hub Push The models at the spokes after local training are denoted by  $x_{t+\frac{1}{4}}$ . Each hub randomly samples  $b_{hs}$  spokes, and the model transfer from spoke j to hub i is represented by the indicator function  $I_j^i$ . Hub i aggregates the  $b_{hs}$  collected models to produce  $x_{t+\frac{2}{4}}$ . This proof bounds the consensus distance after mixing to that before it, that is:

$$\frac{1}{n_h^2} \sum_{\substack{i,j \in [n_h] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(j)} \right\|^2 \right] \leq \frac{\beta_{hs}}{n_s^2} \sum_{\substack{i,j \in [n_s] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)} \right\|^2 \right].$$

where

$$\beta_{hs} = \frac{1}{b_{hs}} \left[ 1 - \frac{b_{hs} - 1}{n_s - 1} \right].$$

Proof.

$$\begin{split} \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \bigg[ \bigg\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{1}{4}} \bigg\|^2 \bigg] &= \frac{1}{n_h} \sum_{i} \mathbb{E} \bigg[ \bigg\| \frac{1}{b_{hs}} \sum_{j \in [n_s]} \mathcal{I}_{j}^{(i)} \, x_{t+\frac{1}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} \bigg\|^2 \bigg] \\ &= \frac{1}{n_h \, b_{hs}^2} \sum_{i} \mathbb{E} \bigg[ \bigg\| \sum_{j} \left( \mathcal{I}_{j}^{(i)} \, x_{t+\frac{1}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} \right) \bigg\|^2 \bigg] \\ &= \frac{1}{n_h \, b_{hs}^2} \sum_{i} \mathbb{E} \bigg[ \sum_{j} \mathcal{I}_{j}^{(i)} \, \bigg\| x_{t+\frac{1}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} \bigg\|^2 \bigg] \\ &+ \frac{1}{n_h \, b_{hs}^2} \sum_{i} \mathbb{E} \bigg[ \sum_{j} \sum_{k \neq j} \mathcal{I}_{j}^{(i)} \, \mathcal{I}_{k}^{(i)} \, \left\langle x_{t+\frac{1}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} - \bar{x}_{t+\frac{1}{4}} \right\rangle \bigg] \\ &= \frac{1}{n_h \, b_{hs}} \frac{1}{n_s} \sum_{i} \mathbb{E} \bigg[ \sum_{j} \left\| x_{t+\frac{1}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} \right\|^2 \bigg] + \frac{1}{n_h \, b_{hs}} \frac{b_{hs} - 1}{n_s \, (n_s - 1)} (-1) \sum_{i} \mathbb{E} \bigg[ \sum_{j} \left\| x_{t+\frac{1}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} \right\|^2 \bigg] \end{split}$$

where we utilize the fact that  $\mathbb{E}[\mathcal{I}_j^{(i)}] = \frac{b_{hs}}{n_s}$  and  $\mathbb{E}[\mathcal{I}_j^{(i)}\mathcal{I}_k^{(i)}] = \frac{b_{hs}}{n_s} \frac{b_{hs}-1}{n_s-1}$ ,

Observe that

$$\mathbb{E} \left[ \sum_{j} \left\| \boldsymbol{x}_{t+\frac{1}{4}}^{(j)} - \bar{\boldsymbol{x}}_{t+\frac{1}{4}} \right\|^2 \right]$$

is independent of i, therefore summing over all  $i \in [n_h]$  scales the entire expression by a factor of  $n_h$ . Thus,

$$\frac{1}{n_{h}} \sum_{j} \mathbb{E} \left[ \left\| x_{t+\frac{1}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} \right\|^{2} \right] = \frac{1}{n_{s} b_{hs}} \left( 1 - \frac{b_{hs} - 1}{n_{s} - 1} \right) \mathbb{E} \left[ \sum_{j} \left\| x_{t+\frac{1}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} \right\|^{2} \right] 
= \frac{\beta_{hs}}{n_{s}} \mathbb{E} \left[ \sum_{j} \left\| x_{t+\frac{1}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} \right\|^{2} \right].$$
(15)

where

$$\beta_{hs} = \frac{1}{b_{hs}} \left( 1 - \frac{b_{hs} - 1}{n_s - 1} \right)$$

Noting that as  $\bar{y}$  is the minimizer of  $g(z) := \frac{1}{n} \sum_{i \in [n]} \mathbb{E}\left[\left\|y^{(i)} - z\right\|^2\right]$ , we have

$$\frac{1}{n} \sum_{i \in [n]} \mathbb{E}\left[ \left\| y^{(i)} - \bar{y} \right\|^2 \right] \le \frac{1}{n} \sum_{i \in [n]} \mathbb{E}\left[ \left\| y^{(i)} - \bar{x} \right\|^2 \right]. \tag{16}$$

Substituting x to  $x_{t+\frac{1}{4}}$  and y to  $x_{t+\frac{2}{4}}$ , and n to  $n_h$ , and using (15), we obtain

$$\frac{1}{n_h} \sum_{i=1}^{n_h} \mathbb{E} \left[ \left\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right] \leq \frac{1}{n_h} \sum_{i=1}^{n_h} \mathbb{E} \left[ \left\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{1}{4}} \right\|^2 \right] = \frac{\beta_{hs}}{n_s} \sum_{i=1}^{n_s} \mathbb{E} \left[ \left\| x_{t+\frac{1}{4}}^{(i)} - \bar{x}_{t+\frac{1}{4}} \right\|^2 \right]. \tag{17}$$

Using Lemma A.6, we obtain,

$$\frac{1}{n_h^2} \sum_{\substack{i,j \in [n_h] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(j)} \right\|^2 \right] \le \frac{\beta_{hs}}{n_s^2} \sum_{\substack{i,j \in [n_s] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)} \right\|^2 \right]. \tag{18}$$

**Stage Two mixing: Hub Gossip** During the hub-gossip stage, every hub shares its models with  $b_{hh}$  other hubs, all having a constant outdegree. However, the indegree of the hubs is a variable  $A^{(i)}$ . This is exactly how nodes communicate in ELL. Then the consensus distance among the hub models after then gossip stage is bound by:

$$\frac{1}{n_h^2} \sum_{\substack{i,j\\i\neq j}} \mathbb{E}\left[ \left\| x_{t+\frac{3}{4}}^{(i)} - x_{t+\frac{3}{4}}^{(j)} \right\|^2 \right] \leq \beta_{hh} \cdot \frac{1}{n_h^2} \sum_{\substack{i,j\\i\neq j}} \mathbb{E}\left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(j)} \right\|^2 \right],$$

where

$$\beta_{hh} = \frac{1}{b_{hh}} \left( 1 - \left( 1 - \frac{b_{hh}}{n_h - 1} \right)^{n_h} \right) - \frac{1}{n_h - 1}.$$

Proof.

$$\begin{split} &\frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \left\| x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right] = \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \left\| \frac{1}{A^{(i)} + 1} \left( x_{t+\frac{2}{4}}^{(i)} + \sum_{j \in [n_h] \backslash \{i\}} \mathcal{I}_{j}^{(i)} x_{t+\frac{2}{4}}^{(j)} \right) - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right] \\ &= \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \mathbb{E} \left[ \left\| \frac{1}{A^{(i)} + 1} \left( x_{t+\frac{2}{4}}^{(i)} + \sum_{j} \mathcal{I}_{j}^{(i)} x_{t+\frac{2}{4}}^{(j)} \right) - \bar{x}_{t+\frac{2}{4}} \right\|^2 |A^{(i)}| \right] \right] \\ &= \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \mathbb{E} \left[ \left\| \frac{1}{A^{(i)} + 1} \left( (x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(i)}) + \sum_{j} \mathcal{I}_{j} (x_{\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}) \right) \right\|^2 |A^{(i)}| \right] \\ &= \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \frac{1}{A^{(i)} + 1} \mathbb{E} \left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(i)} \right\|^2 + \sum_{j} \mathcal{I}_{j} \left\| (x_{\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}) \right\|^2 \right] |A^{(i)}| \right] \\ &= \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \frac{1}{(A^{(i)} + 1)^2} \mathbb{E} \left[ 2 \sum_{j \neq i} \mathcal{I}_{j}^{(i)} \langle x_{\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \rangle + \sum_{j \neq i} \sum_{k \neq i, k \neq j} \mathcal{I}_{j}^{(i)} \mathcal{I}_{k}^{(i)} \langle x_{\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \rangle |A^{(i)}| \right] \right] \end{aligned}$$

Taking the expectation inside, we obtain

$$\begin{split} \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E}[\left|x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}\right|^2] &= \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E}\left[\frac{1}{(A^{(i)} + 1)^2} \left(\left\|x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}\right\|^2 + \sum_{j \neq i} \mathbb{E}[\mathcal{I}_j^{(i)}|A^{(i)}] \left\|x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}\right\|^2\right)\right] \\ &+ \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E}\left[\frac{1}{(A^{(i)} + 1)^2} \left(2\sum_{j \neq i} \mathbb{E}[\mathcal{I}_j^{(i)}|A^{(i)}] \langle x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}\rangle\right)\right] \\ &+ \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E}\left[\frac{1}{(A^{(i)} + 1)^2} \left(\sum_{j \neq i, k \neq i, k \neq j} \mathbb{E}[\mathcal{I}_j^{(i)} \mathcal{I}_k^{(i)}] \langle x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{2}{4}}^{(k)} - \bar{x}_{t+\frac{2}{4}}\rangle\right)\right]. \end{split}$$

Observe that  $\mathbb{E}[\mathcal{I}_{j}^{(i)}|A^{(i)}]$  represents the probability of node j selecting node i, given that a total of  $A^{(i)}$  nodes select i. Thus,

$$\mathbb{E}[\mathcal{I}_j^{(i)}|A^{(i)}] = \frac{A^{(i)}}{n_h - 1}$$

Similarly,  $\mathcal{I}_{i}^{(i)}\mathcal{I}_{k}^{(i)}$  equals 1 only when both j and k choose i, hence

$$\mathbb{E}\left[\mathcal{I}_{j}^{(i)}\mathcal{I}_{k}^{(i)}|A^{(i)}\right] = \frac{A^{(i)}(A^{(i)}-1)}{(n_{h}-1)(n_{h}-2)}.$$

Also, note that

$$\sum_{j \neq i} \langle x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \rangle = \langle x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, \sum_{j \neq i} (x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}) = - \left\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2$$

and

$$\begin{split} & \sum_{j \neq i} \sum_{k \neq i, k \neq j} \langle x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{2}{4}}^{(k)} - \bar{x}_{t+\frac{2}{4}} \rangle = \sum_{j \neq i} \left\langle x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}, \sum_{k \neq i, k \neq j} (x_{t+\frac{2}{4}}^{(k)} - \bar{x}_{t+\frac{2}{4}}) \right\rangle \\ & = \sum_{j \neq i} \left\langle x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}, (x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}) + (x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}) \right\rangle \\ & = \|x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}\|^2 - \sum_{j \neq i} \|x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}\|^2 \end{split}$$

Bringing everything together, we obtain

$$\begin{split} \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \left\| x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right] &= \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \frac{1}{(A^{(i)} + 1)^2} \left( \left\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 + \frac{A^{(i)}}{n_h - 1} \sum_{j \neq i} \left\| x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right) \right] \\ &+ \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \frac{1}{(A^{(i)} + 1)^2} \left( -\frac{2A^{(i)}}{n_h - 1} \left\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right) \right] \\ &+ \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \frac{1}{(A^{(i)} + 1)^2} \left( \frac{A^{(i)}(A^{(i)} - 1)}{(n_h - 1)(n_h - 2)} \left( \left\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 - \sum_{j \neq i} \left\| x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right) \right) \right] \\ &= \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \frac{1}{(A^{(i)} + 1)^2} \left( \frac{1}{(A^{(i)} + 1)^2} \left( 1 - \frac{2A^{(i)}}{n_h - 1} + \frac{A^{(i)}(A^{(i)} - 1)}{(n_h - 1)(n_h - 2)} \right) \right] \\ &+ \frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \frac{1}{(A^{(i)} + 1)^2} \left( \frac{A^{(i)}}{n_h - 1} - \frac{A^{(i)}(A^{(i)} - 1)}{(n_h - 1)(n_h - 2)} \right) \right] \sum_{i \neq i} \left\| x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \end{split}$$

Observe that, due to symmetry, the distribution of  $A^{(i)}$  is identical to that of  $A^{(j)}$  for any  $i, j \in [n_h]$ . Hence,

$$\begin{split} &\frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E} \left[ \left\| x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right] \\ &= \frac{1}{n_h} \sum_{i \in [n_h]} \left\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \mathbb{E} \left[ \frac{1}{(A^{(1)} + 1)^2} \left( 1 - \frac{2A^{(1)}}{n_h - 1} + \frac{A^{(1)}(A^{(1)} - 1)}{(n_h - 1)(n_h - 2)} + \frac{A^{(1)}}{n_h - 1} - \frac{A^{(1)}(A^{(1)} - 1)}{(n_h - 2)} \right) \right] \end{split}$$

Now note that

$$1 - \frac{2A^{(1)}}{n_h - 1} + \frac{A^{(1)}(A^{(1)} - 1)}{(n_h - 1)(n_h - 2)} + A^{(1)} - \frac{A^{(1)}(A^{(1)} - 1)}{n_h - 2} = 1 + A^{(1)} - \frac{A^{(1)^2} + A^{(1)}}{n_h - 1} = (1 + A^{(1)})\left(1 - \frac{A^{(1)}}{n_h - 1}\right)$$

Thus

$$\frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E}\left[ \left\| x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right] = \frac{1}{n_h} \sum_{i \in [n_h]} ||x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}||^2 \left[ \mathbb{E}\left[ \frac{1}{A^{(1)} + 1} \right] - \frac{1}{n_h - 1} \cdot \mathbb{E}\left[ \frac{A^{(1)}}{A^{(1)} + 1} \right] \right]$$

Observe that since each node  $j \neq 1$  independently and uniformly selects a set of  $b_{hh}$  nodes,  $A^{(1)}$  follows a binomial distribution with parameters  $n_h - 1$  and  $\frac{b_{hh}}{n_h - 1}$ . Thus, for  $b_{hh} > 0$ , we have

$$\mathbb{E}\left[\frac{1}{A^{(1)}+1}\right] = \sum_{k=0}^{n_h-1} \frac{1}{k+1} \binom{n_h-1}{k} \left(\frac{b_{hh}}{n_h-1}\right)^k \left(1 - \frac{b_{hh}}{n_h-1}\right)^{n_h-1-k}$$

$$= \frac{n_h-1}{b_{hh}n_h} \sum_{k=0}^{n_h-1} \binom{n_h}{k+1} \left(\frac{b_{hh}}{n_h-1}\right)^{k+1} \left(1 - \frac{b_{hh}}{n_h-1}\right)^{n_h-1-k}$$

$$= \frac{n_h-1}{b_{hh}n_h} \left(1 - \left(1 - \frac{b_{hh}}{n_h-1}\right)^{n_h}\right)$$

Also noting that

$$\mathbb{E}\left[\frac{A^{(1)}}{A^{(1)}+1}\right] = 1 - \mathbb{E}\left[\frac{1}{A^{(1)}+1}\right],$$

we obtain

$$\frac{1}{n_h} \sum_{i \in [n_h]} \mathbb{E}\left[ \left\| x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2 \right] = \left[ \frac{1}{b_{hh}} \left( 1 - \left( 1 - \frac{b_{hh}}{n_h - 1} \right)^{n_h} \right) - \frac{1}{n_h - 1} \right] \frac{1}{n_h} \sum_{i \in [n_h]} \left\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2$$
(19)

we obtain that

Noting that as  $\bar{y}$  is the minimizer of  $g(z) := \frac{1}{n} \sum_{i \in [n]} \mathbb{E} \left[ \|y^{(i)} - z\|^2 \right]$ , following the logic in (16), and using Lemma A.6, we convert the equality in (19) to the following inequality:

$$\frac{1}{n_h^2} \sum_{i,j \in [n_h]} \mathbb{E}\left[ \left\| x_{t+\frac{3}{4}}^{(i)} - x_{t+\frac{3}{4}}^{(j)} \right\|^2 \right] \leq \left[ \frac{1}{b_{hh}} \left( 1 - \left( 1 - \frac{b_{hh}}{n_h - 1} \right)^{n_h} \right) - \frac{1}{n_h - 1} \right] \frac{1}{n_h^2} \sum_{i,j \in [n_h]} \mathbb{E}\left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(j)} \right\|^2 \right]$$
(20)

which is the desired result.  $\Box$ 

Stage Three Mixing: Hub-to-Spoke Pull The models at the hubs after hub gossip are represented as  $x_{t+\frac{3}{4}}$ . Each spoke independently selects  $b_{sh}$  hubs at random, with the model transfer from hub j to spoke i indicated by the function  $\mathcal{I}_j^i$ . Spoke i then aggregates the  $b_{sh}$  received models to obtain  $x_{t+1}$ . This proof establishes an upper bound on the consensus distance among the spoke models after the final aggregation stage relative to its value before aggregation, that is,

$$\frac{1}{n_s^2} \sum_{\substack{i,j \in [n_s]\\i \neq j}} \mathbb{E} \left[ \left\| x_{t+1}^{(i)} - x_{t+1}^{(j)} \right\|^2 \right] \leq \frac{\beta_{sh}}{n_h^2} \sum_{\substack{i,j \in [n_h]\\i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{3}{4}}^{(i)} - x_{t+\frac{3}{4}}^{(j)} \right\|^2 \right].$$

where

$$\beta_{sh} = \frac{1}{b_{sh}} \left[ 1 - \frac{b_{sh} - 1}{n_h - 1} \right].$$

Proof.

$$\begin{split} \frac{1}{n_s} \sum_{i \in [n_s]} \mathbb{E} \bigg[ \bigg\| x_{t+1}^{(i)} - \bar{x}_{t+\frac{3}{4}} \bigg\|^2 \bigg] &= \frac{1}{n_s} \sum_i \mathbb{E} \bigg[ \bigg\| \frac{1}{b_{sh}} \sum_j \mathcal{I}_j^{(i)} x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{3}{4}} \bigg\|^2 \bigg] \\ &= \frac{1}{n_s b_{sh}^2} \sum_i \mathbb{E} \bigg[ \bigg\| \sum_j \bigg( \mathcal{I}_j^{(i)} x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{3}{4}} \bigg) \bigg\|^2 \bigg] \\ &= \frac{1}{n_s b_{sh}^2} \sum_i \mathbb{E} \bigg[ \sum_j \mathcal{I}_j^{(i)} \ \bigg\| x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{3}{4}} \bigg\|^2 \bigg] \\ &+ \frac{1}{n_s b_{sh}^2} \sum_i \mathbb{E} \bigg[ \sum_j \sum_{k \neq j} \mathcal{I}_j^{(i)} \mathcal{I}_k^{(i)} \left\langle x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{3}{4}} \right\rangle \bigg] \\ &= \frac{1}{n_s b_{sh}} \frac{1}{n_h} \sum_i \mathbb{E} \bigg[ \sum_j \bigg\| x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{3}{4}} \bigg\|^2 \bigg] + \frac{1}{n_s b_{sh}} \frac{b_{sh} - 1}{n_h (n_h - 1)} (-1) \sum_i \mathbb{E} \bigg[ \sum_j \bigg\| x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{3}{4}} \bigg\|^2 \bigg] \end{split}$$

where we utilize the fact that  $\mathbb{E}[\mathcal{I}_j^{(i)}] = \frac{b_{sh}}{n_h}$  and  $\mathbb{E}[\mathcal{I}_j^{(i)}\mathcal{I}_k^{(i)}] = \frac{b_{sh}}{n_h} \frac{b_{sh}-1}{n_h-1}$ ,

Just like for stage 1 aggregation, observe that

$$\mathbb{E}\left[\sum_{j}\left\|x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right]$$

is independent of i, therefore summing over all  $i \in [n_s]$  scales the entire expression by a factor of  $n_s$ . Thus,

$$\frac{1}{n_s} \sum_{i \in [n_s]} \mathbb{E} \left[ \left\| x_{t+1}^{(i)} - \bar{x}_{t+\frac{3}{4}} \right\|^2 \right] = \frac{1}{n_h b_{sh}} \left( 1 - \frac{b_{sh} - 1}{n_h - 1} \right) \mathbb{E} \left[ \sum_{j} \left\| x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{3}{4}} \right\|^2 \right] \\
= \frac{\beta_{hs}}{n_h} \mathbb{E} \left[ \sum_{j} \left\| x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{3}{4}} \right\|^2 \right]. \tag{21}$$

where  $\beta_{sh}$ :

$$\beta_{sh} = \frac{1}{b_{sh}} \left( 1 - \frac{b_{sh} - 1}{n_h - 1} \right).$$

Using the minimizer property described in (16), and applying Lemma A.6, we finally obtain:

$$\frac{1}{n_s^2} \sum_{\substack{i,j \in [n_s] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+1}^{(i)} - x_{t+1}^{(j)} \right\|^2 \right] \leq \frac{\beta_{sh}}{n_h^2} \sum_{\substack{i,j \in [n_h] \\ i \neq j}} \mathbb{E} \left[ \left\| x_{t+\frac{3}{4}}^{(i)} - x_{t+\frac{3}{4}}^{(j)} \right\|^2 \right]. \tag{22}$$

#### D.3. Proof of Lemma A.3

Stage 1 Spoke-to-Hub Push We have:

$$\mathbb{E}\left[\left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right] = \frac{\beta_{hs}}{n_{s}n_{h}} \sum_{i} \mathbb{E}\left[\left\|x_{t+\frac{1}{4}}^{(i)} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right]$$

*Proof.* Note that we can expand the norm as follows:

$$\mathbb{E}\left[\left\|\bar{y} - \bar{x}\right\|^{2}\right] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i}y^{(i)} - \bar{x}\right\|^{2}\right]$$

$$= \frac{1}{n^{2}}\sum_{i}\mathbb{E}\left[\left\|y^{(i)} - \bar{x}\right\|^{2}\right] + \frac{1}{n^{2}}\sum_{i\neq j}\mathbb{E}\left[\left\langle y^{(i)} - \bar{x}, y^{(j)} - \bar{x}\right\rangle\right]$$
(23)

For the first stage of communication from spokes to hubs, we denote  $x_{t+\frac{2}{4}}$  as y and  $x_{t+\frac{1}{4}}$  as x, replacing n with  $n_h$ . For the i-th hub, we have:

$$x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{1}{4}} = \frac{1}{b_{hs}} \sum_{k} \mathcal{I}_{k}^{(i)} (x_{t+\frac{1}{4}}^{(k)} - \bar{x}_{t+\frac{1}{4}})$$

where  $\mathcal{I}_k^{(i)}$  is an indicator function that represents the connectivity between hub i and spoke k. We can thus write the second term in (23) as

$$\frac{1}{n^{2}} \sum_{i \neq j} \mathbb{E} \left[ \langle y^{(i)} - \bar{x}, y^{(j)} - \bar{x} \rangle \right] = \frac{1}{n_{h}^{2}} \sum_{\substack{i \in [n_{h}] \\ i \neq j}} \mathbb{E} \left[ \langle x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{1}{4}}, x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{1}{4}} \rangle \right]$$

$$= \frac{2}{n_{h}^{2}} \sum_{i \neq j} \sum_{k \in [n_{s}]} \sum_{l \in [n_{s}]} \mathbb{E} \left[ \frac{\mathcal{I}_{k}^{(i)} \mathcal{I}_{l}^{(j)}}{b_{hs}^{2}} \langle x_{t+\frac{1}{4}}^{(k)} - \bar{x}_{t+\frac{1}{4}}, x_{t+\frac{1}{4}}^{(l)} - \bar{x}_{t+\frac{1}{4}} \rangle \right] \tag{24}$$

Now note that by symmetry, for any  $i, j \in [n_h]$ , we have,

$$\mathbb{E}\left[\mathcal{I}_k^{(i)}\mathcal{I}_l^{(j)}\right] = \mathbb{E}\left[\mathcal{I}_3^{(1)}\mathcal{I}_4^{(2)}\right]$$

This implies that all three terms in (24) can be written as,

$$\begin{split} c \cdot \mathbb{E} \left[ \sum_{k \in [n_s]} \sum_{l \in [n_s]} \left\langle x_{t+\frac{1}{4}}^{(k)} - \bar{x}_{t+\frac{1}{4}}, x_{t+\frac{1}{4}}^{(l)} - \bar{x}_{t+\frac{1}{4}} \right\rangle \right] \\ &= c \cdot \mathbb{E} \left[ \sum_{l} \left\| x_{t+\frac{1}{4}}^{(l)} - \bar{x}_{t+\frac{1}{4}} \right\|^2 + \sum_{k \neq l} \left\langle x_{t+\frac{1}{4}}^{(k)} - \bar{x}_{t+\frac{1}{4}}, x_{t+\frac{1}{4}}^{(l)} - \bar{x}_{t+\frac{1}{4}} \right\rangle \right] \\ &= c \cdot \mathbb{E} \left[ \sum_{l} \left\| x_{t+\frac{1}{4}}^{(l)} - \bar{x}_{t+\frac{1}{4}} \right\|^2 - \sum_{l} \left\| x_{t+\frac{1}{4}}^{(l)} - \bar{x}_{t+\frac{1}{4}} \right\|^2 \right] \\ &= 0 \end{split}$$

Therefore, from equation (23), we obtain

$$\mathbb{E}\left[\left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right] = \frac{1}{n_{h}^{2}} \sum_{i} \mathbb{E}\left[\left\|x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right]$$

$$= \frac{\beta_{hs}}{n_{s}n_{h}} \sum_{i} \mathbb{E}\left[\left\|x_{t+\frac{1}{4}}^{(i)} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right]$$
(25)

where we make use of (15).

**Stage 2 mixing: Hub Gossip** Here we prove that

$$\mathbb{E}\left[\left\|x_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right] \leq \frac{\beta_{hh}}{n_{h}^{2}} \sum_{i \in [n_{h}]} \left\|x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}$$

*Proof.* For the second stage of aggregation, we have:

$$\mathbb{E}\left[\left\|x_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right] = \mathbb{E}\left[\left\|\frac{1}{n_{h}} \sum_{i \in [n_{h}]} \left(x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}\right)\right\|^{2}\right] \\
= \frac{1}{n^{2}} \sum_{i \in [n_{h}]} \mathbb{E}\left[\left\|x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right] + \frac{1}{n_{h}^{2}} \sum_{i \neq j} \mathbb{E}\left[\left\langle x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}\right\rangle\right] \tag{26}$$

Now recall that

$$x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} = \frac{1}{A^{(i)} + 1} \left( (x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}) + \sum_{k \in [n_h] \setminus \{i\}} \mathcal{I}_k^{(i)} (x_{t+\frac{2}{4}}^{(k)} - \bar{x}_{t+\frac{2}{4}}) \right)$$

This implies that

$$\begin{split} \frac{1}{n_h^2} \sum_{i \neq j} \mathbb{E} \left[ \left\langle x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{3}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \right\rangle \right] &= \frac{1}{n_h^2} \sum_{i \in [n_h]} \sum_{j \neq i} \mathbb{E} \left[ \frac{1}{(A^{(i)} + 1)(A^{(j)} + 1)} \left\langle x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \right\rangle \right] \\ &+ \frac{2}{n_h^2} \sum_{i \in [n_h]} \sum_{j \neq i} \sum_{k \neq i, k \neq j} \mathbb{E} \left[ \frac{\mathcal{I}_k^{(i)}}{(A^{(i)} + 1)(A^{(j)} + 1)} \left\langle x_{t+\frac{2}{4}}^{(k)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \right\rangle \right] \\ &+ \frac{2}{n_h^2} \sum_{i \in [n_h]} \sum_{j \neq i} \sum_{k \neq i, k \neq j} \sum_{l \neq i, l \neq j, l \neq k} \mathbb{E} \left[ \frac{\mathcal{I}_k^{(i)} \mathcal{I}_l^{(j)}}{(A^{(i)} + 1)(A^{(j)} + 1)} \left\langle x_{t+\frac{2}{4}}^{(k)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{2}{4}}^{(l)} - \bar{x}_{t+\frac{2}{4}} \right\rangle \right] \end{split}$$

Now note that by symmetry, for any  $i, j \in [n_h]$ , we have

$$\mathbb{E}\left[\frac{1}{(A^{(i)}+1)(A^{(j)}+1)}\right] = \mathbb{E}\left[\frac{1}{(A^{(1)}+1)(A^{(2)}+1)}\right]$$

Similarly

$$\mathbb{E}\left[\frac{\mathcal{I}_k^{(i)}}{(A^{(i)}+1)(A^{(j)}+1)}\right] = \mathbb{E}\left[\frac{\mathcal{I}_3^{(1)}}{(A^{(1)}+1)(A^{(2)}+1)}\right]$$

and

$$\mathbb{E}\left[\frac{\mathcal{I}_k^{(i)}\mathcal{I}_l^{(j)}}{(A^{(i)}+1)(A^{(j)}+1)}\right] = \mathbb{E}\left[\frac{\mathcal{I}_3^{(1)}\mathcal{I}_4^{(2)}}{(A^{(1)}+1)(A^{(2)}+1)}\right]$$

This implies that all three terms in (27) can be written as

$$c\sum_{i\in[n]}\sum_{j\neq i}\langle x_{t+\frac{2}{4}}^{(i)}-\bar{x}_{t+\frac{2}{4}},x_{t+\frac{2}{4}}^{(j)}-\bar{x}_{t+\frac{2}{4}}\rangle$$

where c is a positive constant. We also have

$$\sum_{i \in [n_h]} \sum_{j \neq i} \langle x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}} \rangle = \sum_{i \in [n_h]} \left\langle x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}, \sum_{j \neq i} (x_{t+\frac{2}{4}}^{(j)} - \bar{x}_{t+\frac{2}{4}}) \right\rangle \\
= -\sum_{i \in [n_h]} \left\| x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}} \right\|^2.$$

Therefore all the terms in (27) are non-positive. Combining this with (26), we obtain that

$$\mathbb{E}\left[\left\|x_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right] \leq \frac{1}{n_{h}^{2}} \sum_{i \in [n_{h}]} \mathbb{E}\left[\left\|x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right]$$
$$\leq \frac{\beta_{hh}}{n_{h}} \cdot \frac{1}{n_{h}} \sum_{i \in [n_{h}]} \left\|x_{t+\frac{2}{4}}^{(i)} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}$$

where the second inequality uses (20). Combining this with Lemma A.6 then concludes the proof.

**Stage 3 mixing: Hub-to-Spoke Pull** Here we have:

$$\mathbb{E}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right] = \frac{\beta_{sh}}{n_{s}n_{h}} \sum_{i} \mathbb{E}\left[\left\|x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right]$$

*Proof.* Note that we can expand the norm as follows:

$$\mathbb{E}\left[\left\|\bar{y} - \bar{x}\right\|^{2}\right] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i}y^{(i)} - \bar{x}\right\|^{2}\right]$$

$$= \frac{1}{n^{2}}\sum_{i}\mathbb{E}\left[\left\|y^{(i)} - \bar{x}\right\|^{2}\right] + \frac{1}{n^{2}}\sum_{i\neq j}\mathbb{E}\left[\left\langle y^{(i)} - \bar{x}, y^{(j)} - \bar{x}\right\rangle\right]$$
(28)

For the third stage of communication from hubs to spokes, we denote  $x_{t+1}$  as y and  $x_{t+\frac{3}{4}}$  as x, replacing n with  $n_s$ . For the i-th spoke, we have:

$$x_{t+1}^{(i)} - \bar{x}_{t+\frac{3}{4}} = \frac{1}{b_{sh}} \sum_{k} \mathcal{I}_{k}^{(i)} (x_{t+\frac{3}{4}}^{(k)} - \bar{x}_{t+\frac{3}{4}})$$

where  $\mathcal{I}_k^{(i)}$  is an indicator function that represents the connectivity between spoke i and hub k. We can thus write the second term in (28) as

$$\frac{1}{n^{2}} \sum_{i \neq j} \mathbb{E} \left[ \langle y^{(i)} - \bar{x}, y^{(j)} - \bar{x} \rangle \right] = \frac{1}{n_{s}^{2}} \sum_{\substack{i \in [n_{s}] \\ i \neq j}} \mathbb{E} \left[ \langle x_{t+1}^{(i)} - \bar{x}_{t+\frac{3}{4}}, x_{t+1}^{(j)} - \bar{x}_{t+\frac{3}{4}} \rangle \right]$$

$$= \frac{2}{n_{s}^{2}} \sum_{\substack{i \neq j \\ k \in [n_{b}]}} \sum_{\substack{l \in [n_{b}] \\ l \in [n_{b}]}} \mathbb{E} \left[ \frac{\mathcal{I}_{k}^{(i)} \mathcal{I}_{l}^{(j)}}{b_{sh}^{2}} \langle x_{t+\frac{3}{4}}^{(k)} - \bar{x}_{t+\frac{3}{4}}, x_{t+\frac{3}{4}}^{(l)} - \bar{x}_{t+\frac{3}{4}} \rangle \right] \tag{29}$$

Now note that by symmetry, for any  $i, j \in [n_s]$ , we have,

$$\mathbb{E}\left[\mathcal{I}_k^{(i)}\mathcal{I}_l^{(j)}\right] = \mathbb{E}\left[\mathcal{I}_3^{(1)}\mathcal{I}_4^{(2)}\right]$$

This implies that all three terms in (29) can be written as,

$$\begin{split} c \cdot \mathbb{E} \left[ \sum_{k \in [n_h]} \sum_{l \in [n_h]} \left\langle x_{t+\frac{3}{4}}^{(k)} - \bar{x}_{t+\frac{3}{4}}, x_{t+\frac{3}{4}}^{(l)} - \bar{x}_{t+\frac{3}{4}} \right\rangle \right] \\ &= c \cdot \mathbb{E} \left[ \sum_{l} \left\| x_{t+\frac{3}{4}}^{(l)} - \bar{x}_{t+\frac{3}{4}} \right\|^2 + \sum_{k \neq l} \left\langle x_{t+\frac{3}{4}}^{(k)} - \bar{x}_{t+\frac{3}{4}}, x_{t+\frac{3}{4}}^{(l)} - \bar{x}_{t+\frac{3}{4}} \right\rangle \right] \\ &= c \cdot \mathbb{E} \left[ \sum_{l} \left\| x_{t+\frac{3}{4}}^{(l)} - \bar{x}_{t+\frac{3}{4}} \right\|^2 - \sum_{l} \left\| x_{t+\frac{3}{4}}^{(l)} - \bar{x}_{t+\frac{3}{4}} \right\|^2 \right] \\ &= 0 \end{split}$$

Therefore, from equation (28), we obtain

$$\mathbb{E}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right] = \frac{1}{n_{s}^{2}} \sum_{i} \mathbb{E}\left[\left\|x_{t+1}^{(i)} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right]$$

$$= \frac{\beta_{sh}}{n_{s} n_{h}} \sum_{i} \mathbb{E}\left[\left\|x_{t+\frac{3}{4}}^{(i)} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right]$$
(30)

where we make use of (21).

#### D.4. Proof of Lemma A.4

The expected consensus distance and gradient variance across spokes are bounded as follows:

## 1. Consensus Distance Bound:

$$\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] \le 20 \frac{1 + 3\beta_{HSL}}{(1 - \beta_{HSL})^2} \beta_{HSL} \gamma^2 (\sigma^2 + \mathcal{H}^2).$$

# 2. Gradient Variance Bound:

$$\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| g_t^{(i)} - g_t^{(j)} \right\|^2 \right] \le 15(\sigma^2 + \mathcal{H}^2).$$

*Proof.* For any  $i \in [n]$ , we have

$$g_t^{(i)} - g_t^{(j)} = g_t^{(i)} - \nabla f^{(i)} \left( x_t^{(i)} \right) + \nabla f^{(i)} \left( x_t^{(i)} \right) - \nabla f^{(i)} \left( \bar{x}_t \right) + \nabla f^{(j)} \left( \bar{x}_t \right)$$
$$- \nabla f^{(j)} \left( \bar{x}_t \right) + \nabla f^{(j)} \left( \bar{x}_t \right) - \nabla f^{(j)} \left( x_t^{(j)} \right) + \nabla f^{(j)} \left( x_t^{(j)} \right) - g_t^{(j)}$$

where  $g_t$  is the stochastic version of  $\nabla f^{(i)}\left(x_t^{(i)}\right)$ . Thus, using Jensens's inequality, we have,

$$\begin{aligned} \left\| g_{t}^{(i)} - g_{t}^{(j)} \right\|^{2} &\leq 5 \left\| g_{t}^{(i)} - \nabla f^{(i)} \left( x_{t}^{(i)} \right) \right\|^{2} + 5 \left\| \nabla f^{(i)} \left( x_{t}^{(i)} \right) - \nabla f^{(i)} \left( \bar{x}_{t} \right) \right\|^{2} \\ &+ 5 \left\| \nabla f^{(j)} \left( x_{t}^{(j)} \right) - \nabla f^{(j)} \left( \bar{x}_{t} \right) \right\|^{2} + 5 \left\| g_{t}^{(j)} - \nabla f^{(j)} \left( x_{t}^{(j)} \right) \right\|^{2} \\ &+ 5 \left\| \nabla f^{(i)} \left( \bar{x}_{t} \right) - \nabla f^{(j)} \left( \bar{x}_{t} \right) \right\|^{2} \end{aligned}$$

Taking the conditional expectation, we have,

$$\mathbb{E}_{t} \left[ \left\| g_{t}^{(i)} - g_{t}^{(j)} \right\|^{2} \right] \leq 5\mathbb{E}_{t} \left[ \left\| g_{t}^{(i)} - \nabla f^{(i)} \left( x_{t}^{(i)} \right) \right\|^{2} \right] + 5\mathbb{E}_{t} \left[ \left\| \nabla f^{(i)} \left( x_{t}^{(i)} \right) - \nabla f^{(i)} \left( \bar{x}_{t} \right) \right\|^{2} \right] \\
+ 5\mathbb{E}_{t} \left[ \left\| \nabla f^{(j)} \left( x_{t}^{(j)} \right) - \nabla f^{(j)} \left( \bar{x}_{t} \right) \right\|^{2} \right] + 5\mathbb{E}_{t} \left[ \left\| g_{t}^{(j)} - \nabla f^{(j)} \left( x_{t}^{(j)} \right) \right\|^{2} \right] \\
+ 5\mathbb{E}_{t} \left[ \left\| \nabla f^{(i)} \left( \bar{x}_{t} \right) - \nabla f^{(j)} \left( \bar{x}_{t} \right) \right\|^{2} \right] \tag{31}$$

Now by Assumption 4.2, we have,

$$\mathbb{E}_{t} \left[ \left\| g_{t}^{(i)} - \nabla f^{(i)} \left( x_{t}^{(i)} \right) \right\|^{2} \right] \leq \sigma^{2}$$
(32)

By Assumption 4.1, we have,

$$\mathbb{E}_{t}\left[\left\|\nabla f^{(i)}\left(x_{t}^{(i)}\right) - \nabla f^{(i)}\left(\bar{x}_{t}\right)\right\|^{2}\right] \leq L^{2}\mathbb{E}_{t}\left[\left\|x_{t}^{(i)} - \bar{x}_{t}\right\|^{2}\right]$$
(33)

Thus, by Assumption 4.3, and Lemma A.6, we obtain that,

$$\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}_t \left[ \left\| \nabla f^{(i)}(\bar{x}_t) - \nabla f^{(j)}(\bar{x}_t) \right\|^2 \right] \le 2\mathcal{H}^2$$
 (34)

Combining (31), (32), (33), and (34), and taking total expectation from both sides, we obtain that,

$$\frac{1}{n_s^2} \sum_{i,j \in [n_h]} \mathbb{E}\left[ \left\| g_t^{(i)} - g_t^{(j)} \right\|^2 \right] \le \frac{10L^2}{n_s} \sum_{i \in [n_s]} \mathbb{E}\left[ \left\| x_t^{(i)} - \bar{x}_t \right\|^2 \right] + 10\sigma^2 + 10\mathcal{H}^2$$
(35)

Now Lemma A.6 yields

$$\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| g_t^{(i)} - g_t^{(j)} \right\|^2 \right] \le \frac{5L^2}{n_s^2} \sum_{i \in [n_s]} \mathbb{E}\left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] + 10\sigma^2 + 10\mathcal{H}^2$$
(36)

We now analyze  $\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[\left\|x_t^{(i)} - x_t^{(j)}\right\|^2\right]$ . From Algorithm 1, recall that for all  $i \in [n]$ , we have  $x_{t+1/4}^{(i)} = x_t^{(i)} - \gamma g_t^{(i)}$ . We obtain for all  $i, j \in [n_s]$ , that

$$\mathbb{E}\left[\left\|x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)}\right\|^{2}\right] \leq \mathbb{E}\left[\left\|x_{t}^{(i)} - x_{t}^{(j)} - \gamma(g_{t}^{(i)} - g_{t}^{(j)})\right\|^{2}\right] \\
\leq (1+c)\mathbb{E}\left[\left\|x_{t}^{(i)} - x_{t}^{(j)}\right\|^{2}\right] + \left(1 + \frac{1}{c}\right)\gamma^{2}\mathbb{E}\left[\left\|g_{t}^{(i)} - g_{t}^{(j)}\right\|^{2}\right] \tag{37}$$

From (18), we have

$$\frac{1}{n_h^2} \sum_{i,j} \mathbb{E}\left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(j)} \right\|^2 \right] \leq \frac{\beta_{hs}}{n_s^2} \sum_{i,j} \mathbb{E}\left[ \left\| x_{t+\frac{1}{4}}^{(i)} - x_{t+\frac{1}{4}}^{(j)} \right\|^2 \right]$$

Combining with (37), we get

$$\frac{1}{n_h^2} \sum_{i,j} \mathbb{E} \left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(j)} \right\|^2 \right] \le (1+c) \frac{\beta_{hs}}{n_s^2} \sum_{i,j} \mathbb{E} \left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] + \left( 1 + \frac{1}{c} \right) \gamma^2 \frac{\beta_{hs}}{n_s^2} \sum_{i,j} \mathbb{E} \left[ \left\| g_t^{(i)} - g_t^{(j)} \right\|^2 \right]$$
(38)

From (20), we also have

$$\frac{1}{n_h^2} \sum_{i,j} \mathbb{E}\left[ \left\| x_{t+\frac{3}{4}}^{(i)} - x_{t+\frac{3}{4}}^{(j)} \right\|^2 \right] \leq \frac{\beta_{hh}}{n_h^2} \sum_{i,j} \mathbb{E}\left[ \left\| x_{t+\frac{2}{4}}^{(i)} - x_{t+\frac{2}{4}}^{(j)} \right\|^2 \right]$$

We substitute (38) here to get:

$$\frac{1}{n_h^2} \sum_{i,j} \mathbb{E}\left[ \left\| x_{t+\frac{3}{4}}^{(i)} - x_{t+\frac{3}{4}}^{(j)} \right\|^2 \right] \le (1+c) \frac{\beta_{hs}\beta_{hh}}{n_s^2} \sum_{i,j} \mathbb{E}\left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right]$$
(39)

$$+\left(1+\frac{1}{c}\right)\gamma^2 \frac{\beta_{hs}\beta_{hh}}{n_s^2} \sum_{i,j} \mathbb{E}\left[\left\|g_t^{(i)} - g_t^{(j)}\right\|^2\right] \tag{40}$$

Also, from (22), we obtain

$$\frac{1}{n_s^2} \sum_{i,j} \mathbb{E} \left[ \left\| x_{t+1}^{(i)} - x_{t+1}^{(j)} \right\|^2 \right] \leq \frac{\beta_{hs}}{n_s^2} \sum_{i,j} \mathbb{E} \left[ \left\| x_{t+\frac{3}{4}}^{(i)} - x_{t+\frac{3}{4}}^{(j)} \right\|^2 \right]$$

We combine this with (40) to get

$$\begin{split} \frac{1}{n_s^2} \sum_{i,j} \mathbb{E} \left[ \left\| x_{t+1}^{(i)} - x_{t+1} \right\|^2 \right] &\leq (1+c) \frac{\beta_{HSL}}{n_s^2} \sum_{i,j} \mathbb{E} \left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] \\ &+ \left( 1 + \frac{1}{c} \right) \gamma^2 \frac{\beta_{HSL}}{n_s^2} \sum_{i,j} \mathbb{E} \left[ \left\| g_t^{(i)} - g_t^{(j)} \right\|^2 \right] \end{split}$$

For  $c = \frac{1 - \beta_{HSL}}{4\beta_{HSL}}$ , we obtain that,

$$\begin{split} \frac{1}{n_{s}^{2}} \sum_{i,j \in [n_{s}]} \mathbb{E}\left[ \left\| x_{t+1}^{(i)} - x_{t+1}^{(j)} \right\|^{2} \right] &\leq \frac{1 + 3\beta_{HSL}}{4} \frac{1}{n_{s}^{2}} \sum_{i,j \in [n_{s}]} \mathbb{E}\left[ \left\| x_{t}^{(i)} - x_{t}^{(j)} \right\|^{2} \right] \\ &+ \frac{1 + 3\beta_{HSL}}{1 - \beta_{HSL}} \beta_{HSL} \gamma^{2} \frac{1}{n_{s}^{2}} \sum_{i,j \in [n_{s}]} \mathbb{E}\left[ \left\| g_{t}^{(i)} - g_{t}^{(j)} \right\|^{2} \right] \end{split}$$

Combining this with (36), we obtain that

$$\begin{split} \frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E} \left[ \left\| x_{t+1}^{(i)} - x_{t+1}^{(j)} \right\|^2 \right] &\leq \frac{1 + 3\beta_{HSL}}{4} \frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E} \left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] \\ &+ \frac{1 + 3\beta_{HSL}}{1 - \beta_{HSL}} \beta_{HSL} \gamma^2 \left( \frac{5L^2}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E} \left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] + 10\sigma^2 + 10\mathcal{H}^2 \right) \\ &= \left( \frac{1 + 3\beta_{HSL}}{4} + 5\frac{1 + 3\beta_{HSL}}{1 - \beta_{HSL}} \beta_{HSL} \gamma^2 L^2 \right) \frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E} \left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] \\ &+ \frac{1 + 3\beta_{HSL}}{1 - \beta_{HSL}} \beta_{HSL} \gamma^2 (10\sigma^2 + 10\mathcal{H}^2). \end{split}$$

Now note that from Remark C.1 we have  $\beta_{HSL} \leq 1 - \frac{1}{e}$  which implies that

$$\gamma^2 \le \frac{1}{(20L)^2} \le \frac{(1 - \beta_{HSL})^2}{20\beta_{HSL}(1 + 3\beta_{HSL})L^2}$$

Therefore,

$$\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| x_{t+1}^{(i)} - x_{t+1}^{(j)} \right\|^2 \right] \leq \frac{1 + \beta_{HSL}}{2} \frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] + \frac{1 + 3\beta_{HSL}}{1 - \beta_{HSL}} \beta_{HSL} \gamma^2 (10\sigma^2 + 10\mathcal{H}^2).$$

Unrolling the recursion, we obtain that,

$$\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| x_t^{(i)} - x_t^{(j)} \right\|^2 \right] \leq 20 \frac{1 + 3\beta_{HSL}}{(1 - \beta_{HSL})^2} \beta_{HSL} \gamma^2 (\sigma^2 + \mathcal{H}^2).$$

Combining this with (35), we obtain that,

$$\frac{1}{n_s^2} \sum_{i,j \in [n_s]} \mathbb{E}\left[ \left\| g_t^{(i)} - g_t^{(j)} \right\|^2 \right] \le 15(\sigma^2 + \mathcal{H}^2).$$

#### D.4.1. PROOF OF LEMMA A.5

The expected gradient norm of the global objective satisfies the following upper bound:

$$\mathbb{E}\left[\left\|\nabla F(\bar{x}_{t})\right\|^{2}\right] \leq \frac{2}{\gamma} \mathbb{E}\left[F(\bar{x}_{t}) - F(\bar{x}_{t+1})\right] + \frac{L}{2n^{2}} \sum_{i,j} \mathbb{E}\left[\left\|x_{t}^{(i)} - x_{t}^{(j)}\right\|^{2}\right] + \frac{4L\gamma\sigma^{2}}{n} + \frac{4L}{\gamma} \mathbb{E}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2} + \left\|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2} + \left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right].$$

*Proof.* Consider an arbitrary  $t \in [T]$ . Then from the smoothness property, we have:

$$F(\bar{x}_{t+1}) - F(\bar{x}_t) \leq \langle \bar{x}_{t+1} - \bar{x}_t, \nabla F(\bar{x}_t) \rangle + \frac{L}{2} ||\bar{x}_{t+1} - \bar{x}_t||^2$$

$$= \langle \bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}} + \bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}} + \bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}} + \bar{x}_{t+\frac{1}{4}} - \bar{x}_t, \nabla F(\bar{x}_t) \rangle$$

$$+ \frac{L}{2} ||\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}} + \bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}} + \bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}} + \bar{x}_{t+\frac{1}{4}} - \bar{x}_t ||^2.$$

$$(41)$$

From Lemma A.1, we have

$$\mathbb{E}\big[\overline{x}_{t+1}\big] = \mathbb{E}\big[\overline{x}_{t+\frac{3}{4}}\big] = \mathbb{E}\big[\overline{x}_{t+\frac{2}{4}}\big] = \mathbb{E}\big[\overline{x}_{t+\frac{1}{4}}\big].$$

Now, we take conditional expectation on (41) and use Lemma A.1 to get

$$\mathbb{E}_{t}\left[F(\bar{x}_{t+1}) - F(\bar{x}_{t})\right] \leq \langle \mathbb{E}_{t}\left[\bar{x}_{t+1} - \bar{x}_{t}\right], \nabla F(\bar{x}_{t})\rangle + \frac{L}{2}\mathbb{E}_{t}\left[\left||\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}} + \bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}} + \bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}} + \bar{x}_{t+\frac{1}{4}} - \bar{x}_{t}\right]^{2}\right] \\
\leq -\gamma\langle \overline{\nabla}F_{t}, \nabla F(\bar{x}_{t})\rangle + 2L\gamma^{2}\mathbb{E}_{t}\left[\left||\bar{g}_{t}|\right|^{2}\right] + 2L\,\mathbb{E}_{t}\left[\left|\left|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right|\right|^{2}\right] + 2L\,\mathbb{E}_{t}\left[\left|\left|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right|\right|^{2}\right] \\
+ 2L\,\mathbb{E}_{t}\left[\left|\left|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right|\right|^{2}\right] \tag{42}$$

where  $\overline{\nabla F}_t = \frac{1}{n_s} \sum_{i \in [n_s]} \nabla f^{(i)}(x_t^{(i)})$ ,  $\bar{g}_t = \frac{1}{n_s} \sum_{i \in [n_s]} g_t^{(i)}$ , and we make use of Jensen's inequality. Then we use the  $\mathbb{E}_t[\bar{g}_t] = \overline{\nabla F}_t$ , and  $\mathbb{E}_t[\|\bar{g}_t - \overline{\nabla F}_t\|^2] \leq \frac{\sigma^2}{n}$ 

$$\mathbb{E}_{t}\left[F(\bar{x}_{t+1}) - F(\bar{x}_{t})\right] \leq -\gamma \langle \overline{\nabla F}_{t}, \nabla F(\bar{x}_{t}) \rangle + 2L\gamma^{2} \mathbb{E}\left[\|\nabla F(\bar{x}_{t})\|^{2}\right] + 2L\frac{\gamma^{2}\sigma^{2}}{n} + 2L\,\mathbb{E}_{t}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right] \\
+ 2L\,\mathbb{E}_{t}\left[\left\|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right] + 2L\,\mathbb{E}_{t}\left[\left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right] \tag{43}$$

Then we use  $\gamma \leq \frac{1}{4L}$  to get

$$-\gamma \langle \overline{\nabla F}_t, \nabla F(\overline{x}_t) \rangle + 2L\gamma^2 ||\overline{\nabla F}_t||^2 \le \frac{\gamma}{2} (-2\langle \overline{\nabla F}_t, \nabla F(\overline{x}_t) \rangle + ||\overline{\nabla F}_t||^2)$$

$$= \frac{\gamma}{2} (-||\nabla F(\overline{x}_t)||^2 + ||\nabla F(\overline{x}_t) - \overline{\nabla F}_t||^2). \tag{44}$$

Combining (43) and (44), we obtain

$$\mathbb{E}_{t} \left[ F \left( \bar{x}_{t+1} \right) - F \left( \bar{x}_{t} \right) \right] \leq -\frac{\gamma}{2} \left\| \nabla F \left( \bar{x}_{t} \right) \right\|^{2} + \frac{\gamma}{2} \left\| \nabla F \left( \bar{x}_{t} \right) - \nabla \bar{F}_{t} \right\|^{2} + 2L \frac{\gamma^{2} \sigma^{2}}{n} \\
+ 2L \mathbb{E}_{t} \left[ \left\| \bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}} \right\|^{2} \right] + 2L \mathbb{E}_{t} \left[ \left\| \bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}} \right\|^{2} \right] + 2L \mathbb{E}_{t} \left[ \left\| \bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}} \right\|^{2} \right]$$

Taking total expectation, we obtain

$$\mathbb{E}_{t}\left[F\left(\bar{x}_{t+1}\right) - F\left(\bar{x}_{t}\right)\right] \leq -\frac{\gamma}{2}\mathbb{E}_{t}\left[\left\|\nabla F\left(\bar{x}_{t}\right)\right\|^{2}\right] + \frac{\gamma}{2}\mathbb{E}_{t}\left[\left\|\nabla F(\bar{x}_{t}) - \nabla \bar{F}_{t}\right\|^{2}\right] + 2L\frac{\gamma^{2}\sigma^{2}}{n} + 2L\mathbb{E}_{t}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right] + 2L\mathbb{E}_{t}\left[\left\|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right] + 2L\mathbb{E}_{t}\left[\left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right] \tag{45}$$

Now, note that

$$\mathbb{E}\left[\left\|\overline{\nabla F}_{t} - \nabla F(\overline{x}_{t})\right\|^{2}\right] = \mathbb{E}\left[\left\|\frac{1}{n_{s}}\sum_{i\in[n_{s}]}\nabla f^{(i)}(x_{t}^{(i)}) - \frac{1}{n_{s}}\sum_{i\in[n_{s}]}\nabla f^{(i)}(\overline{x}_{t})\right\|^{2}\right]$$

$$= \mathbb{E}\left[\left\|\frac{1}{n_{s}}\sum_{i\in[n_{s}]}\left(\nabla f^{(i)}(x_{t}^{(i)}) - \nabla f^{(i)}(\overline{x}_{t})\right)\right\|^{2}\right]$$

$$\leq \frac{1}{n_{s}}\sum_{i\in[n_{s}]}\mathbb{E}\left[\left\|\nabla f^{(i)}(x_{t}^{(i)}) - \nabla f^{(i)}(\overline{x}_{t})\right\|^{2}\right]$$

$$\leq \frac{L^{2}}{n_{s}}\sum_{i\in[n_{s}]}\mathbb{E}\left[\left\|x_{t}^{(i)} - \overline{x}_{t}\right\|^{2}\right].$$

$$\leq \frac{L^{2}}{2n_{s}^{2}}\sum_{i,j\in[n_{s}]}\mathbb{E}\left[\left\|x_{t}^{(i)} - x_{t}^{(j)}\right\|^{2}\right],$$

where we use Assumption 4.1 and Lemma A.6 in the above two inequalities. Now, substituting this back in (45).

$$\mathbb{E}_{t}\left[F\left(\bar{x}_{t+1}\right) - F\left(\bar{x}_{t}\right)\right] \leq -\frac{\gamma}{2}\mathbb{E}_{t}\left[\left\|\nabla F\left(\bar{x}_{t}\right)\right\|^{2}\right] + \frac{\gamma}{2}\frac{L^{2}}{2n_{s}^{2}}\sum_{i,j\in[n_{s}]}\mathbb{E}\left[\left\|x_{t}^{(i)} - x_{t}^{(j)}\right\|^{2}\right] + 2L\frac{\gamma^{2}\sigma^{2}}{n} + 2L\mathbb{E}_{t}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2}\right] + 2L\mathbb{E}_{t}\left[\left\|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2}\right] + 2L\mathbb{E}_{t}\left[\left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right]$$

Rearranging these terms, we get

$$\begin{split} \mathbb{E}\left[\left\|\nabla F(\bar{x}_{t})\right\|^{2}\right] &\leq \frac{2}{\gamma}\mathbb{E}\left[F(\bar{x}_{t}) - F(\bar{x}_{t+1})\right] + \frac{L}{2n^{2}}\sum_{i,j}\mathbb{E}\left[\left\|x_{t}^{(i)} - x_{t}^{(j)}\right\|^{2}\right] + \frac{4L\gamma\sigma^{2}}{n} \\ &+ \frac{4L}{\gamma}\mathbb{E}\left[\left\|\bar{x}_{t+1} - \bar{x}_{t+\frac{3}{4}}\right\|^{2} + \left\|\bar{x}_{t+\frac{3}{4}} - \bar{x}_{t+\frac{2}{4}}\right\|^{2} + \left\|\bar{x}_{t+\frac{2}{4}} - \bar{x}_{t+\frac{1}{4}}\right\|^{2}\right] \end{split}$$

## D.5. Proof of Lemma A.6

For any set  $\{x_t^{(i)}\}_{i\in[n]}$  of n vectors, we have

$$\frac{1}{n} \sum_{i \in [n]} \|x_t^{(i)} - \bar{x}_t\|^2 = \frac{1}{2n^2} \sum_{i,j \in [n]} \|x_t^{(i)} - x_t^{(j)}\|^2,$$

where  $\bar{x}_t = \frac{1}{n} \sum_{i \in [n]} x_t^{(i)}$ .

**Proof** 

$$\begin{split} \frac{1}{n^2} \sum_{i,j \in [n]} \|x_t^{(i)} - x_t^{(j)}\|^2 &= \frac{1}{n^2} \sum_{i,j \in [n]} \|(x_t^{(i)} - \bar{x}_t) - (x_t^{(j)} - \bar{x}_t)\|^2 \\ &= \frac{1}{n^2} \sum_{i,j \in [n]} \left[ \|x_t^{(i)} - \bar{x}_t\|^2 + \|x_t^{(j)} - \bar{x}_t\|^2 - 2\langle x_t^{(i)} - \bar{x}_t, x_t^{(j)} - \bar{x}_t \rangle \right] \\ &= \frac{2}{n} \sum_{i \in [n]} \|x_t^{(i)} - \bar{x}_t\|^2 - \frac{2}{n^2} \sum_{i \in [n]} \left\langle x_t^{(i)} - \bar{x}_t, \sum_{j \in [n]} (x_t^{(j)} - \bar{x}_t) \right\rangle. \end{split}$$

Noting that  $\sum_{j \in [n]} (x_t^{(j)} - \bar{x}_t) = 0$ , yields the desired result.