A Tale of Two Learning Algorithms: Multiple Stream Random Walk and Asynchronous Gossip

Abstract

We investigate the relative performance of gossip and random walk-based decentralized learning algorithms across varying network topologies and data heterogeneities. To this end, we introduce "Multi-Walk," an asynchronous multi-stream random walk algorithm, and comprehensively compare its convergence against "Asynchronous Gossip" in terms of computational iterations, wall-clock time, and communication overhead. Our analysis demonstrates that Multi-Walk achieves superior iteration convergence in large-diameter graphs and consistently lower communication costs, except within small-diameter networks experiencing extreme data heterogeneity. Furthermore, both methods exhibit linear wall-clock speedups relative to the number of concurrent computations. These findings elucidate the inherent performance trade-offs between the two approaches, offering guidance on algorithm selection based on structural and distributional network conditions. Our codes are available for reproducibility.

1. Introduction

Decentralized learning is emerging as a crucial enabler for distributed intelligence in Next-Generation (NextG) wireless networks, such as 6G and edge computing systems. It mitigates traditional federated learning limitations, such as centralized communication bottlenecks and single points of failure (Tsitsiklis, 1984; Nedić & Ozdaglar, 2009; McMahan et al., 2023). While its two prominent paradigms, gossip and random walk-based algorithms, have been extensively studied (Boyd et al., 2006; Lian et al., 2017; Koloskova et al., 2019; Bertsekas, 1997; Ayache & Rouayheb, 2020; Sun et al., 2018; Needell et al., 2015), their relative performance across the varying graph topologies and data heterogeneities characteristic of dynamic wireless environments remains unclear. This work addresses this gap by comprehensively comparing their convergence rates, communication costs, and computational overheads, directly tackling the resource constraints of NextG systems.

Gossip algorithms advocate that nodes in a network iteratively update their models with Stochastic Gradient Descent (SGD) (Robbins, 1951; Bottou et al., 2018) and exchange the updated models with their neighbors, leading to global consensus over time. Gossip can employ synchronous communication (Lian et al., 2017; Koloskova et al., 2020), where nodes must wait for all nodes to update their model in each round. However, in heterogeneous wireless networks with straggler nodes or varying computation speeds (Kairouz et al., 2021), synchronous gossip results in significant idle times for fast nodes and creates bottlenecks (Chen et al., 2017). Asynchronous gossip algorithms (Baudet, 1978; Tsitsiklis et al., 1984; Recht et al., 2011) have been developed to leverage resources more effectively, allowing nodes to compute gradients using a stale model and communicate in an asynchronous manner, thereby eliminating the need to wait for all nodes (Lian et al., 2018; Nabli et al., 2023; Nadiradze et al., 2022; Bornstein et al., 2022; Even et al., 2024). Nevertheless, in both synchronous and asynchronous cases, gossip incurs high communication costs due to frequent, concurrent message exchanges among nodes, a major drawback for bandwidth-limited wireless links. The random walk-based learning algorithms suggest that one node at a time updates a model with its local data. The node then randomly selects a neighbor and sends the updated model to it. This neighbor becomes the next activated node and updates the model using its own local data. This process repeats until convergence. Random walk-based algorithms (Ayache & Rouayheb, 2020; Sun et al., 2018; Needell et al., 2015) are typically single stream, i.e., only one node updates the model at any given time, which leads to slow convergence. To evaluate these algorithmic trade-offs across diverse network conditions, our primary contributions are as follows:

Design of Multi-Walk. We introduce Multi-Walk, the first completely asynchronous, multi-stream random walk-based learning algorithm. By adjusting the number of walks, it enables a flexible trade-off between convergence speed and resource utilization, a critical capability for adaptive NextG deployments. Furthermore, we demonstrate that this approach achieves a linear speedup relative to the number of parallel streams.

Comprehensive analysis. We conduct an in-depth analysis of Multi-Walk and Asynchronous Gossip, comparing their convergence w.r.t iterations, wall-clock time, and communication overhead. Our results reveal a distinct struc-

tural and distributional trade-off: Multi-Walk excels in iteration complexity on large-diameter graphs with homogeneous data, whereas Asynchronous Gossip outperforms for small-diameter, highly non-iid networks. Additionally, while Asynchronous Gossip achieves a linear wall-clock speedup relative to the number of nodes, Multi-Walk consistently requires less communication overhead, except in small-diameter topologies with extreme data heterogeneity.

Empirical validation. We conduct experiments to validate our theoretical findings on the structural and distributional trade-offs of each algorithm. Overall, the experiments provide valuable insights into the performance trade-offs between gossip and random walk-based decentralized learning algorithms, offering guidance on algorithm selection for communication-restricted environments.

2. Related Work

Decentralized optimization algorithms typically utilize gossip-based mixing to propagate updates (Tsitsiklis, 1984; Nedić & Ozdaglar, 2009; Duchi et al., 2012; Yuan et al., 2015; Gholami & Seferoglu, 2024; Xiao & Boyd, 2003; Lian et al., 2017; Koloskova et al., 2020), inherently incurring a substantial communication overhead. While asynchronous techniques from federated learning (Agarwal & Duchi, 2011; Lian et al., 2015; Zheng et al., 2016; Feyzmahdavian & Johansson, 2021; Mishchenko et al., 2022; Koloskova et al., 2022) address synchronization bottlenecks, their decentralized extensions often rely on synchronous computation steps Assran & Rabbat (2021), neglect system delays (Nadiradze et al., 2022; Bornstein et al., 2022; Nabli et al., 2023), or use continuous communication loops that preclude theoretical overhead analysis Even et al. (2024). To benchmark Multi-Walk, we adopt the prominent asynchronous method AD-PSGD Lian et al. (2018) as Asynchronous Gossip. Furthermore, we strengthen its theoretical foundation with a comprehensive convergence proof that relaxes original assumptions regarding bounded computation delays and minimum iteration thresholds.

Alternatively, random walk-based approaches (Ayache & Rouayheb, 2020) typically utilize single-walk methods akin to SGD data sampling (Sun et al., 2018; Needell et al., 2015). However, existing analyses of multiple walks are often confined to convex settings, demand impractical full gradient computations, and rely on centralized synchronization Hendrikx (2023). To overcome these limitations, we introduce Multi-Walk, the first completely asynchronous, multi-stream random walk-based learning algorithm.

3. Setup and Algorithm Design

We model the underlying network topology with a connected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices (nodes) and $\mathcal{E}$ is the set edges. The vertex set contains V nodes, *i.e.,* $|\mathcal{V}| = V$. If node i is connected to node j through a communication link, then $\{i, j\}$ is in the edge set, *i.e.,* $\{i, j\} \in \mathcal{E}$. The set of the nodes that node i is connected to and can transmit data is called the neighbors of node i, and the neighbor set of node i is denoted by $\mathcal{N}_i$. Assume that the nodes in the network jointly minimize a d-dimensional function $f : \mathbb{R}^d \to \mathbb{R}$. The goal is to solve optimization problems where the elements of the objective function (i.e., the data used in machine learning tasks) are distributed across different nodes, $\min_{\mathbf{x}\in\mathbb{R}^d} \left[f(\mathbf{x}) := \frac{1}{V} \sum_{v\in\mathcal{V}} \left[f_v(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_v} F_v(\mathbf{x}, \xi) \right] \right]$. $F_v(\mathbf{x}, \xi) : \mathbb{R}^d \to \mathbb{R}$ is the loss function of $\mathbf{x}$ associated with data sample ξ at node v. The loss function on local dataset $\mathcal{D}_v$ at node v is $f_v(\mathbf{x})$. We provide a table of notations in App. A.

Algorithm 1 Asynchronous Multi-Walk with R walks

1: Start walk r at node $r-1$, which sets $\mathbf{x}_0^r = \mathbf{x}_0$, where $r \in \{1, \ldots, R\}$.
2: Node 0 initializes $\{u^r\}_{r\in\{1,\ldots,R\}}$ with $\mathbf{x}_0$.
3: Set $l = 1$, which is the last walk that visited Node 0.
4: **for** $t = 0$ to $T-1$ **do**
5: **if** Node v_t finishes the calculation of $\nabla F_{v_t}(\mathbf{x}_{t-\tau_t}^{r_t}, \xi_t)$ at point $\mathbf{x}_{t-\tau_t}^{r_t}$, which was transmitted to node v_t by one of its neighbors via walk r_t **then** iteration t is started and node v_t executes lines 6-12.
6: $\mathbf{x}_{t+1}^{r_t} = \mathbf{x}_{t-\tau_t}^{r_t} - \eta_t \nabla F_{v_t}(\mathbf{x}_{t-\tau_t}^{r_t}, \xi_t)$
7: **if** $v_t = 0$ **then**
8: $\mathbf{x}_{t+1}^{r_t} = u^l + \frac{1}{R}(\mathbf{x}_{t+1}^{r_t} - u^{r_t})$.
9: $u^{r_t} = \mathbf{x}_{t+1}^{r_t}$.
10: $l = r_t$.
11: Choose the next node based on matrix $\mathbf{P}$.
12: Send $\mathbf{x}_{t+1}^{r_t}$ to the next node via walk r_t.

Asynchronous Multi-Walk. Our novel Multi-Walk algorithm considers the standard asynchronous SGD for model updates. To achieve consensus, communication is performed using multiple walks. Multi-Walk algorithm is summarized in Algorithm 1. First, we assume that there are R walks over the graph, where $R \leq V$. Without loss of generality, we initialize walk r at node $r-1$ by setting $\mathbf{x}_0^r = \mathbf{x}_0$, where $r \in \{1, \ldots, R\}$. $\mathbf{x}_0$ is the global initial model. These nodes start computing the stochastic gradient at $\mathbf{x}_0$ using their local data. In order to mix the information among walks, we consider a dedicated node that we assume to be Node 0 without loss of generality.[1] We also define $\{u^r\}_{r\in\{1,\ldots,R\}}$, where u^r is a copy of walk r's model at the most recent instance when that walk was at Node 0. At Node 0, we initialize $\{u^r\}_{r\in\{1,\ldots,R\}}$ with $\mathbf{x}_0$ that will be used in the mixing. Assume that l is the last walk that visited node Node 0, and l is initialized with 1, *i.e.,* $l = 1$. Throughout the algorithm, each node receiving a model via a walk computes its gradient at its own pace, using its local data and the received model. On line 5, once a node (denoted as v_t) com-

[1] We note that Node 0 may become unavailable or fail due to underlying network conditions. This is addressed in Section 5.

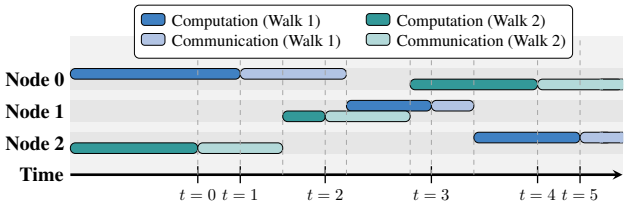


Figure 1: Example of MW in a 3-node network with $R = 2$.

Algorithm 2 Asynchronous Gossip (AD-PSGD)

1: Initialize local models $\mathbf{x}_t^v = \mathbf{x}_0$ in all nodes. All nodes start computing the stochastic gradient.
2: **for** $t = 0$ to $T-1$ **do**
3: Node v_t is randomly sampled from all nodes.
4: **if** Node v_t finishes computing the gradient at point $\mathbf{x}_{t-\tau_t}^{v_t}$, *i.e.,* $\nabla F_{v_t}(\mathbf{x}_{t-\tau_t}^{v_t}, \xi_t)$ **then** iteration t is started. Node v_t executes lines 5-7.
5: $\mathbf{x}_{t+\frac{1}{2}}^{v_t} = \mathbf{x}_t^{v_t} - \eta_t \nabla F_{v_t}(\mathbf{x}_{t-\tau_t}^{v_t}, \xi_t)$.
6: $\mathbf{x}_{t+1}^{v} = \sum_{i \in \mathcal{N}_v} p_{vi} \mathbf{x}_{t+\frac{1}{2}}^{i}$ (gossip averaging for all $v \in \mathcal{V}$ based on mixing matrix $\mathbf{P}$)
7: Start computing gradient at point $\mathbf{x}_{t+1}^{v_t}$.

pletes computing the gradient using the model received via walk r_t, iteration t is started. This model was last updated at iteration $t - \tau_t$ by a neighbor of v_t, or corresponds to the initial model $\mathbf{x}_0$. We note that only one gradient computation completion event happens in each iteration. On line 6, node v_t incorporates the computed gradient to update the model using the step size η_t. Note that communicating the model of walk r_t to node v_t and computing the gradient takes τ_t iterations. Now, if v_t is Node 0, we need to mix the current walk, r_t, with other walks. This is done in lines 8–10. On line 8, we incorporate the newly introduced updates of walk r_t, i.e., $(\mathbf{x}_{t+1}^{r_t} - u^{r_t})$, which have not been mixed before, into the latest model (u^l) with a weight of $\frac{1}{R}$. We update the last applied model of walk r_t (u^{r_t}) and the latest walk (l) on lines 9 and 10. Finally, node v_t chooses the next node based on the transition matrix $\mathbf{P}$ and sends the model. We note that $\mathbf{P}$ is the transition matrix of a Markov chain, representing each walk, where p_{ij} in row i and column j of $\mathbf{P}$ denotes the probability of choosing the next node as j given that the current node is i. Figure 1 depicts the operation of the Multi-Walk algorithm on a 3-node network employing two parallel walks ($R = 2$). Furthermore, the diagram visualizes the sequence of iterations over real time.

Asynchronous Gossip Algorithm. Based on Lian et al. (2018)[2], Algorithm 2 allows all nodes to compute gradients asynchronously. At iteration t, a randomly selected node v_t completes its delayed gradient computation $\nabla F_{v_t}(\mathbf{x}_{t-\tau_t}^{v_t}, \xi_t)$, evaluated at a stale model $\mathbf{x}_{t-\tau_t}^{v_t}$. Node v_t then updates its current model $\mathbf{x}_t^{v_t}$ using learning rate η_t, followed immediately by a gossip averaging step according to a mixing matrix $\mathbf{P}$ (with averaging weights p_{ij}). Finally, node v_t initiates its next gradient computation at the updated point $\mathbf{x}_{t+1}^{v_t}$.

[2]We use Asynchronous Gossip and AD-PSGD interchangeably.

4. Convergence Analysis

We use the following standard assumptions in our analysis:

1. **Smooth local loss.** $f_v(\mathbf{x})$ is differentiable and its gradient is L-Lipschitz for $v \in \mathcal{V}$, *i.e.,* $\|\nabla f_v(\mathbf{y}) - \nabla f_v(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.
2. **Bounded local variance.** The variance of the stochastic gradient is bounded for $v \in \mathcal{V}$, *i.e.,* $\mathbb{E}_{\xi \sim \mathcal{D}_v} \|\nabla F_v(\mathbf{x}, \xi) - \nabla f_v(\mathbf{x})\|^2 \leq \sigma^2$.
3. **Bounded diversity.** The diversity of the local loss functions are bounded for $v \in \mathcal{V}$, *i.e.,* $\|\nabla f_v(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \zeta^2$.
4. **Transition (mixing) matrix.** In Algorithm 1, $\mathbf{P}$ is the transition matrix of an irreducible and aperiodic Markov chain, representing each walk. In Algorithm 2, it defines the mixing step of the gossip averaging. Matrix $\mathbf{P}$ is doubly stochastic ($\mathbf{P1} = \mathbf{1}$, $\mathbf{1}^\top \mathbf{P} = \mathbf{1}^\top$) and the spectral gaps of $\mathbf{P}^\top \mathbf{P}$ and $\mathbf{P}$ are denoted by p and p', respectively.

4.1. Convergence rate w.r.t iterations

Theorem 4.1. ***Multi-Walk.*** *Let assumptions 1-4 hold, with a constant and small enough learning rate η (potentially depending on T), after T iterations of Algorithm 1, $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla f(\mathbf{x}_t^{r_t})\|^2$ is*

$$\mathcal{O}\left(\frac{FLRH}{T} + \frac{R\zeta^2}{p'T} + \sqrt{\frac{FL(\sigma^2+\zeta^2)}{T}} + \left(\frac{FLR\sqrt{V\sigma^2 + H^2\zeta^2}}{T}\right)^{\frac{2}{3}}\right), \quad (1)$$

where $F := f(\mathbf{x}_0) - f^*$, and H^2 is the second moment of the first return time to Node 0 for the Markov chain representing each walk.[3] □

Theorem 4.2. ***Asynchronous Gossip.*** *Let assumptions 1-4 hold, with a constant and small enough learning rate η (potentially depending on T), after T iterations of Algorithm 2, $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla f(\mathbf{x}_t^{v_t})\|^2$ is*

$$\mathcal{O}\left(\frac{FLV}{pT} + \sqrt{\frac{FL(\sigma^2+\zeta^2)}{T}} + \left(\frac{FLV\sqrt{\frac{\sigma^2}{p} + \frac{\zeta^2}{p^2}}}{T}\right)^{\frac{2}{3}}\right), \quad (2)$$

where $F := f(\mathbf{x}_0) - f^*$. □

The proofs are in APP. B and APP. C. The dominant term in both (1) and (2) is identically given by $\sqrt{\frac{FL(\sigma^2+\zeta^2)}{T}}$. Focusing on the next most significant term for comparison, in (1), this term is given by $\left(\frac{FLR\sqrt{V\sigma^2+H^2\zeta^2}}{T}\right)^{\frac{2}{3}}$, whereas in (2), it is $\left(\frac{FLV\sqrt{\frac{\sigma^2}{p}+\frac{\zeta^2}{p^2}}}{T}\right)^{\frac{2}{3}}$. Note that (1) includes a non-dominating term that describes the rate at which walks converge to their steady state. This term is related to the spectral

[3]Specifically, $H^2 = \mathbb{E}[h^2]$ denotes the second moment of the first return time, $h = \min\{\, k \geq 1 : X_k = 0 \mid X_0 = 0\}$, which measures the number of steps required for a walk's Markov chain to return to Node 0.

Table 1: Analysis in total transmitted bits (B).

ALGORITHM	CONVERGENCE RATE	COMP-COST
MULTI-WALK	$\mathcal{O}\left(\frac{FLRHm}{B}+\frac{R\zeta^2 m}{p'B}+\sqrt{\frac{FLm\left(\sigma^2+\zeta^2\right)}{B}}+\left(\frac{FLRm\sqrt{V\sigma^2+H^2\zeta^2}}{B}\right)^{\frac{2}{3}}\right)$	$\Theta(\frac{B}{m})$
ASYNCHRONOUS GOSSIP	$\mathcal{O}\left(\frac{FLVm\|\mathbf{P}\|_0}{pB}+\sqrt{\frac{FLm\|\mathbf{P}\|_0\left(\sigma^2+\zeta^2\right)}{B}}+\left(\frac{FLVm\|\mathbf{P}\|_0\sqrt{\frac{\sigma^2}{p}+\frac{\zeta^2}{p^2}}}{B}\right)^{\frac{2}{3}}\right)$	$\Theta(\frac{B}{m\|\mathbf{P}\|_0})$

Table 2: Analysis in wall-clock time (Z).

ALGORITHM	CONVERGENCE RATE	COMM-COST	COMP-COST
MULTI-WALK	$\mathcal{O}\left(\frac{FLHd}{Z}+\frac{\zeta^2 d}{p'Z}+\sqrt{\frac{FLd\left(\sigma^2+\zeta^2\right)}{RZ}}+\left(\frac{FLd\sqrt{\sigma^2V+\zeta^2H^2}}{Z}\right)^{\frac{2}{3}}\right)$	$\Theta(\frac{ZRm}{d})$	$\Theta(\frac{ZR}{d})$
ASYNCHRONOUS GOSSIP	$\mathcal{O}\left(\frac{FLd}{pZ}+\sqrt{\frac{FLd\left(\sigma^2+\zeta^2\right)}{VZ}}+\left(\frac{FLd\sqrt{\frac{\sigma^2}{p}+\frac{\zeta^2}{p^2}}}{Z}\right)^{\frac{2}{3}}\right)$	$\Theta(\frac{ZVm\|\mathbf{P}\|_0}{d})$	$\Theta(\frac{ZV}{d})$

gap of $\mathbf{P}$, represented by p'. In the following, we compare the dominant terms in the convergence rates of both algorithms in different settings.

Homogeneous Data Distribution. In iid setting ($\zeta = 0$), the differentiating factor in the second dominant term of the convergence rate is $\frac{V}{\sqrt{p}}$ for Asynchronous Gossip and $R\sqrt{V}$ for Multi-Walk. Specifically, Multi-Walk achieves faster iteration convergence on graphs where $p = \mathcal{O}(\frac{V}{R^2})$, whereas Asynchronous Gossip converges faster when $p = \Omega(\frac{V}{R^2})$. Notably, the network topology does not affect the performance of Multi-Walk in the iid setting; its convergence depends solely on the number of nodes and walks. Table 3 (Appendix G) compares the convergence rates and communication overheads of both algorithms across three topologies: cycle, 2D-torus, and complete graphs. As shown, Multi-Walk consistently outperforms Asynchronous Gossip across all three topologies, provided R is not excessively large for the complete graph.

Heterogeneous Data Distribution. In non-iid setting, the parameter ζ^2 is scaled by a factor of H^2 for Multi-Walk and p^2 for Asynchronous Gossip. We formally derive H^2 for both cycle and complete topologies using a Metropolis-Hastings transition matrix in Appendix E, and we summarize the theoretical comparison in Table 4 (Appendix G). As demonstrated in the table, Multi-Walk consistently achieves faster iteration convergence on cycle topologies. However, this advantage diminishes on networks with smaller diameters under high data heterogeneity.

4.2. Convergence rate w.r.t transmitted bits

Assume the model size is m bits. Each iteration of Algorithm 1 and 2 communicates one and $\|\mathbf{P}\|_0$ models, respectively. $\|\mathbf{P}\|_0$ denote the number of non-zero elements of mixing matrix $\mathbf{P}$.

Corollary 4.3. *Under the condition of Theorem 4.1, 4.2, we get the convergence rate of Algorithm 1, and 2 as shown in Table 1 where B represents total transmitted bits.*

The dominating term here is $\sqrt{\frac{FLm(\sigma^2+\zeta^2)}{B}}$ for Multi-Walk and $\sqrt{\frac{FLm\|\mathbf{P}\|_0(\sigma^2+\zeta^2)}{B}}$ for Asynchronous Gossip. Thus, we observe that Multi-Walk outperforms Asynchronous Gossip in terms of transmitted bits when the second dominating term is not comparable in magnitude. Intuitively, in every model transmission, Multi-Walk executes approximately one computation per model transmission, while Asynchronous Gossip performs $\|\mathbf{P}\|_0$ model transmissions per computation. Therefore, Multi-Walk is a better choice when there is a restriction on the amount of communicated bits. In the extreme noniid regime (large ζ), the second dominant term is proportional to $H^2\zeta^2$ for Multi-Walk, compared to $\frac{\zeta^2}{p^2}$ for Asynchronous Gossip. This suggests that Asynchronous Gossip is advantageous in highly non-iid settings with small graph diameters. In the extreme non-iid case, the second term becomes comparable to the leading term, and in small-diameter graphs, this term specifically favors Asynchronous Gossip. Taking a complete graph as an example, the terms simplify to $V^2\zeta^2$ for Multi-Walk versus ζ^2 for Asynchronous Gossip, a difference that significantly favors Asynchronous Gossip in such settings.

4.3. Convergence rate w.r.t wall-clock time

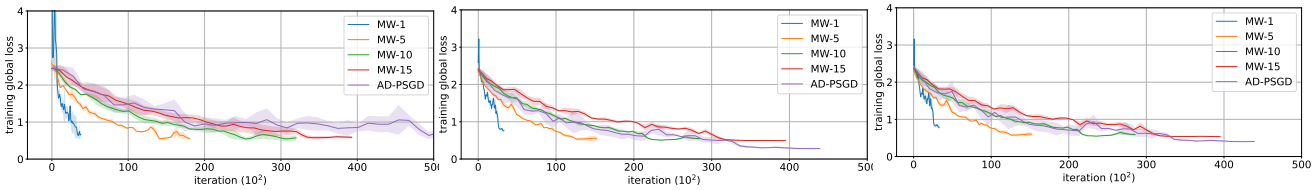
To evaluate convergence in terms of real time, let d denote the average computation and communication delay in the network per iteration (i.e., a single walk step in Algorithm 1 or a local update and gossip step in Algorithm 2).

Corollary 4.4. *Under the condition of Theorem 4.1, 4.2, we get the convergence rate of Algorithms 1 and 2 as shown in Table 2 where Z represents wall-clock time.*

The proof is in Appendix F.The dominating term here is $\sqrt{\frac{FLd(\sigma^2+\zeta^2)}{RZ}}$ for Multi-Walk and $\sqrt{\frac{FLd(\sigma^2+\zeta^2)}{VZ}}$ for Asynchronous Gossip. This highlights the advantage of Asynchronous Gossip when considering real-time performance.

5. Resilience of Multi-Walk against Node Failures

In Multi-Walk, node failures do not disrupt the system unless the node is actively processing a walk, which results in the loss of that specific walk's information, a straightforward ex-



(a) Cycle graph. (b) Complete graph. (c) Erdős–Rényi (0.3) graph.

Figure 2: Training loss of ResNet-20 on Cifar-10 on a 20-node graph with different topologies for $\alpha = 1.0$.

tension of the single-walk failure model (Egger et al., 2024). Furthermore, Node 0 is not a single point of failure; since all active streams maintain a valid, albeit slightly outdated, copy of the global model, any other node can seamlessly take the role of Node 0 to resume the aggregation process. To monitor Node 0 liveness, nodes exchange periodic heartbeat signals. Upon detecting a failure, nodes communicate resource metrics (e.g., bandwidth, memory, compute capacity) with their one-hop neighbors to collaboratively elect an optimal replacement. The newly elected Node 0 initializes its local copies $\{u^r\}_{r\in\{1,\dots,R\}}$ using parameters from the arriving walks, allowing the Multi-Walk algorithm to resume. Because Node 0 failures inevitably impact overall convergence time, we next analyze the algorithmic convergence behavior under such failure and recovery scenarios.

Let us assume there are E failures throughout the learning process, corresponding to $E+1$ different Node 0 selected until convergence. Let H_i^2 be the second moment of the first return time to Node 0 chosen after the i-th failure. Assumption 4 concerning the transition matrix must hold after each failure. Accordingly, we define p_i' as the spectral gap of the transition matrix $\mathbf{P}$ after the i-th failure, which may involve changes to the network topology. For this analysis, we assume that failures do not alter the global loss function and that no streams are lost.

Theorem 5.1. ***Multi-Walk with Node*** 0 ***failures.*** *Let assumptions 1-4 hold, with a constant and small enough learning rate η (potentially depending on T), after T iterations of Algorithm 1 with E Node 0 failures, $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{r_t})\|^2$ is*

$$\mathcal{O}\Bigg(\frac{FLRE\max_i H_i}{T}+\sum_{i=0}^{E}\frac{R\zeta^2}{p_i'T}+\sqrt{\frac{FL(\sigma^2+\zeta^2)}{T}} + \Big(\frac{FLR\sqrt{EV\sigma^2+E^2\zeta^2\max_i H_i^2}}{T}\Big)^{\frac{2}{3}}\Bigg), \tag{3}$$

where $F := f(\mathbf{x}_0) - f^*$.

The proof is in App. D. (3) demonstrates that Multi-Walk guarantees convergence to a stationary point despite Node 0 failures. However, these failures inherently degrade the convergence rate due to information loss and recovery time, an impact mathematically reflected by the penalty terms $E\max_i H_i$, the summation $\sum_{i=0}^{E}(\cdot)$, and $E^2\max_i H_i^2$.

6. Experiments

We empirically validate our theoretical findings by first analyzing the effects of network topology and data heterogeneity on convergence (Sections 6.1 and 6.2). We then evaluate the communication efficiency of Multi-Walk during an LLM fine-tuning task (Section 6.3), and finally demonstrate its resilience against Node 0 failures (Section 6.4).

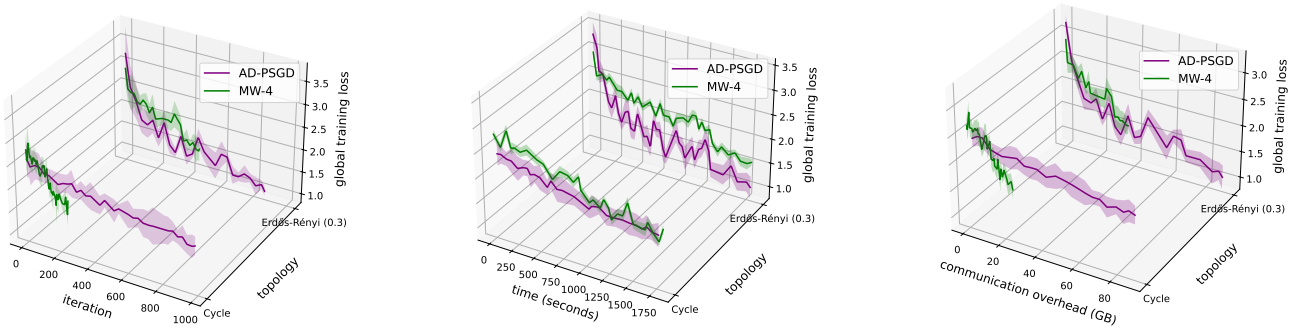
We use two machine learning tasks: (i) *Image classification* on CIFAR-10 (Krizhevsky, 2009) using ResNet-20 (He et al., 2015); and (ii) *LLM fine-tuning* of OPT-125M (Zhang et al., 2022) as a large language model on the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018). We repeat each experiment 10 times and present the error bars associated with the randomness of the optimization. In every figure, we include the average and standard deviation error bars. In the figures, we use "MW" as an abbreviation for MW. We have conducted the experiments on the National Resource Platform (NRP) (NRP) cluster. Detailed experimental results are provided in App. H. We use the Dirichlet distribution to create disjoint non-iid nodes (Lin et al., 2021). The degree of data heterogeneity is controlled by the distribution parameter α; the smaller α is, the more likely the nodes hold examples from only one class.

6.1. Homogeneous data distribution

Figure 8 presents the training loss of the image classification task in a graph of 20 nodes. We consider three topologies of cycle, complete, and Erdős–Rényi with connection probability of each pair of nodes being 0.3. The noniid-ness level for this experiment is set to $\alpha = 1$. We observe in Figure 8a (large diamater) that the convergence rate w.r.t iterations in cycle topology is faster for MW, regardless of the number of walks (R). On the other hand, when we decrease the diameter of the topology in Figure 8b (complete graph) and 8c (Erdős–Rényi 0.3), we observe that MW is superior as long as R is not very large (15 walks). These result are consistent with our theoretical result in section 4.

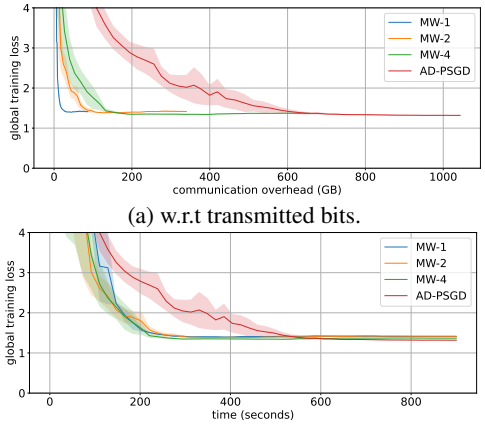
6.2. Heterogeneous data distribution

In Figure 3, we present the experimental results for the cycle and Erdős–Rényi (0.3) topologies in an extreme non-iid setting ($\alpha = 0.1$). In the cycle topology (which is characterized by a large diameter), MW remains faster in terms of both iterations (Figure 3a) and transmitted bits (Figure 3c), while successfully closing the gap in terms of wall-clock time (Figure 3b). However, MW is no longer superior in the small-diameter graph of Erdős–Rényi (0.3) under the extreme noniid scenario. This result is expected based on



(a) w.r.t iterations. (b) w.r.t wall-clock time. (c) w.r.t transmitted bits.

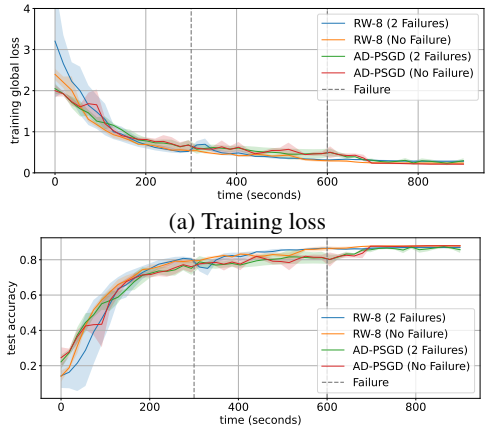
Figure 3: ResNet-20 on Cifar-10 across different network topologies (20-node) for extreme noniid level of $\alpha = 0.1$.



(a) w.r.t transmitted bits.

(b) w.r.t wall-clock time.

Figure 4: Fine-tuning OPT-125M on the MultiNLI corpus.



(a) Training loss

(b) Test Accuracy

Figure 5: Convergence behavior with two Node 0 failures.

the theoretical bounds presented in section 4. In the convergence rate of MW, the heterogeneity term ζ^2 is scaled by H^2, whereas in Asynchronous Gossip, it is scaled by $1/p^2$. In small-diameter graphs (e.g., complete graph), we have $H^2 = \mathcal{O}(V^2)$ and $p = 1$, meaning the impact of noniid data on MW ($\mathcal{O}(V^2)$) is far more severe than on Asynchronous Gossip ($\mathcal{O}(1)$). This confirms the degradation observed in our experiments. On the other hand, in large-diameter graphs like the cycle topology, we have $H^2 = \mathcal{O}(V^3)$ and $p = \Theta(1/V^2)$ (implying $1/p^2 = \mathcal{O}(V^4)$). Consequently, the impact of heterogeneity scales better for MW ($\mathcal{O}(V^3)$) compared to Asynchronous Gossip ($\mathcal{O}(V^4)$), explaining why MW retains its advantage in this setting.

6.3. Communication restricted settings

Fine-tuning large models like OPT-125M on MultiNLI drastically increases the per-round communication payload to 500 MB, compared to just 1.08 MB for smaller models like ResNet-20. As illustrated in Figure 4 for a 20-node Erdős–Rényi (0.3) graph, this increased size significantly impacts the total communication overhead. Consequently, Figure 4a demonstrates that MW requires only ~ 50 GB to converge, whereas Asynchronous Gossip consumes ~ 600 GB. Furthermore, MW achieves faster wall-clock convergence (Figure 4b). Although Asynchronous Gossip theoretically benefits from a linear speedup with more active nodes, its high volume of concurrent communications creates severe network congestion for large models. This congestion increases the average system delay (d), ultimately slowing down Asynchronous Gossip and highlighting MW as a highly promising solution for communication-restricted environments.

6.4. Resilience

As illustrated in Figure 5, ResNet-20 training on CIFAR-10 over a 20-node Erdős–Rényi (0.3) graph successfully converges despite two Node 0 failures (at 300s and 600s, indicated by gray dashed lines). Because the Node 0 primarily integrates information from individual walks, which inherently retain core model data, its loss minimally impacts the overall convergence rate. While Asynchronous Gossip demonstrates nearly identical resilience under the same failure sequence, it is important to note that losing nodes permanently removes their associated data partitions. This effectively alters the global objective function, causing the model to converge to a slightly different solution.

7. Conclusion

We presented a comprehensive analysis comparing gossip and random walk-based decentralized learning algorithms. Our findings reveal clear structural trade-offs that are crucial for the resource constraints of NextG systems.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A. and Duchi, J. C. Distributed delayed stochastic optimization. *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 5451–5452, 2011. URL https://api.semanticscholar.org/CorpusID:901118.

Assran, M. S. and Rabbat, M. G. Asynchronous gradient push. *IEEE Transactions on Automatic Control*, 66(1): 168–183, January 2021. ISSN 2334-3303. doi: 10.1109/tac.2020.2981035. URL http://dx.doi.org/10.1109/TAC.2020.2981035.

Ayache, G. and Rouayheb, S. Y. E. Private weighted random walk stochastic gradient descent. *IEEE Journal on Selected Areas in Information Theory*, 2:452–463, 2020. URL https://api.semanticscholar.org/CorpusID:221651576.

Baudet, G. M. Asynchronous iterative methods for multiprocessors. *J. ACM*, 25(2):226–244, apr 1978. ISSN 0004-5411. doi: 10.1145/322063.322067. URL https://doi.org/10.1145/322063.322067.

Bertsekas, D. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, November 1997. ISSN 1052-6234. doi: 10.1137/S1052623495287022.

Bornstein, M., Rabbani, T., Wang, E., Bedi, A. S., and Huang, F. Swift: Rapid decentralized federated learning via wait-free model communication, 2022. URL https://arxiv.org/abs/2210.14026.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning, 2018. URL https://arxiv.org/abs/1606.04838.

Boyd, S. P., Ghosh, A., Prabhakar, B., and Shah, D. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52:2508–2530, 2006. URL https://api.semanticscholar.org/CorpusID:2120244.

Chen, J., Pan, X., Monga, R., Bengio, S., and Jozefowicz, R. Revisiting distributed synchronous sgd, 2017. URL https://arxiv.org/abs/1604.00981.

Duchi, J. C., Agarwal, A., and Wainwright, M. J. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, March 2012. ISSN 1558-2523. doi: 10.1109/tac.2011.2161027. URL http://dx.doi.org/10.1109/TAC.2011.2161027.

Egger, M., Ayache, G., Bitar, R., Wachter-Zeh, A., and Rouayheb, S. E. Self-duplicating random walks for resilient decentralized learning on graphs. In *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, pp. 2960–2965, 2024. doi: 10.1109/GLOBECOM52923.2024.10901339.

Even, M., Koloskova, A., and Massoulie, L. Asynchronous SGD on graphs: a unified framework for asynchronous decentralized and federated optimization. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 64–72. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/even24a.html.

Feyzmahdavian, H. R. and Johansson, M. Asynchronous iterations in optimization: New sequence results and sharper algorithmic guarantees. *ArXiv*, abs/2109.04522, 2021. URL https://api.semanticscholar.org/CorpusID:237485562.

Gholami, P. and Seferoglu, H. Digest: Fast and communication efficient decentralized learning with local updates. *IEEE Transactions on Machine Learning in Communications and Networking*, 2:1456–1474, 2024. doi: 10.1109/TMLCN.2024.3354236.

Guruswami, V. Rapidly mixing markov chains: A comparison of techniques (a survey), 2016. URL https://arxiv.org/abs/1603.01512.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

Hendrikx, H. A principled framework for the design and analysis of token algorithms. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 470–489. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/hendrikx23a.html.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M.,

Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., et al. Advances and open problems in federated learning, 2021. URL https://arxiv.org/abs/1912.04977.

Koloskova, A., Stich, S., and Jaggi, M. Decentralized stochastic optimization and gossip algorithms with compressed communication. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3478–3487. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/koloskova19a.html.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized SGD with changing topology and local updates. 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. 5381–5393. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koloskova20a.html.

Koloskova, A., Stich, S. U., and Jaggi, M. Sharper convergence guarantees for asynchronous sgd for distributed and federated learning, 2022. URL https://arxiv.org/abs/2206.08307.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

Levin, D. A. and Peres, Y. Markov chains and mixing times: Second edition. 2017. URL https://api.semanticscholar.org/CorpusID:28640176.

Lian, X., Huang, Y., Li, Y., and Liu, J. Asynchronous parallel stochastic gradient for nonconvex optimization. *ArXiv*, abs/1506.08272, 2015. URL https://api.semanticscholar.org/CorpusID:21782.

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 *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:1467846.

Lian, X., Zhang, W., Zhang, C., and Liu, J. Asynchronous decentralized parallel stochastic gradient descent, 2018. URL https://arxiv.org/abs/1710.06952.

Lin, T., Karimireddy, S. P., Stich, S., and Jaggi, M. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6654–6665. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/lin21c.html.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data, 2023. URL https://arxiv.org/abs/1602.05629.

Mishchenko, K., Bach, F. R., Even, M., and Woodworth, B. E. Asynchronous sgd beats minibatch sgd under arbitrary delays. *ArXiv*, abs/2206.07638, 2022. URL https://api.semanticscholar.org/CorpusID:249674816.

Nabli, A., Belilovsky, E., and Oyallon, E. $\mathbf{A^2CiD^2}$: Accelerating asynchronous communication in decentralized deep learning, 2023. URL https://arxiv.org/abs/2306.08289.

Nadiradze, G., Sabour, A., Davies, P., Li, S., and Alistarh, D. Asynchronous decentralized sgd with quantized and local updates, 2022. URL https://arxiv.org/abs/1910.12308.

Nedić, A. and Ozdaglar, A. E. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54:48–61, 2009. URL https://api.semanticscholar.org/CorpusID:6489200.

Needell, D., Srebro, N., and Ward, R. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm, 2015. URL https://arxiv.org/abs/1310.5715.

NRP. National research platform (nrp). https://nationalresearchplatform.org.

Recht, B., Ré, C., Wright, S. J., and Niu, F. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Neural Information Processing Systems*, 2011. URL https://api.semanticscholar.org/CorpusID:6108215.

Robbins, H. E. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951. URL https://api.semanticscholar.org/CorpusID:16945044.

Stich, S. U. Local sgd converges fast and communicates little, 2019. URL https://arxiv.org/abs/1805.09767.

Sun, T., Sun, Y., and Yin, W. On markov chain gradient descent. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:54074144.

Tsitsiklis, J. N. *Problems in decentralized decision making and computation*. PhD thesis, Massachusetts Institute of Technology, 1984.

Tsitsiklis, J. N., Bertsekas, D. P., and Athans, M. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *1984 American Control Conference*, pp. 484–489, 1984. URL https://api.semanticscholar.org/CorpusID:17975552.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Xiao, L. and Boyd, S. P. Fast linear iterations for distributed averaging. *42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475)*, 5:4997–5002 Vol.5, 2003. URL https://api.semanticscholar.org/CorpusID:6001203.

Yuan, K., Ling, Q., and Yin, W. On the convergence of decentralized gradient descent, 2015. URL https://arxiv.org/abs/1310.7063.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M. T., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022. URL https://api.semanticscholar.org/CorpusID:248496292.

Zheng, S., Meng, Q., Wang, T., Chen, W., Yu, N., Ma, Z., and Liu, T.-Y. Asynchronous stochastic gradient descent with delay compensation. *ArXiv*, abs/1609.08326, 2016. URL https://api.semanticscholar.org/CorpusID:3713670.

A. Notation Table

$G = (\mathcal{V}, \xi)$	The graph representing the network
V	Number of nodes
$\mathcal{D}_v$	Local dataset at node v
$F_v(\mathbf{x}, \xi)$	Loss function of $\mathbf{x}$ associated with the data sample ξ at node v
$f(\mathbf{x})$	Global loss function of model $\mathbf{x}$
$f_v(\mathbf{x})$	Local loss function of model $\mathbf{x}$ on local dataset $\mathcal{D}_v$ at node v
f^*	$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$
$\mathbf{x}_0$	Initial model
T	Total number of iterations
η_t	Learning rate at iteration t
$\mathbf{x}_t^r$	Model of walk r at iteration t in MW Algorithm
$\mathbf{x}_t^v$	Local model of node v at iteration t in Asynchronous Gossip Algorithm
u^r	A copy of the model of walk r at the most recent instance when that walk was at Node 0 in MW Algorithm; to be kept at Node 0
l	The index of the latest walk visited Node 0 in MW Algorithm
$\mathbf{P}$	The transition matrix of each walk in MW, and in Asynchronous Gossip, it defines the mixing step of the gossip process
p_{ij}	The element in row i and column j of $\mathbf{P}$
p	The spectral gap of $\mathbf{P}^\top\mathbf{P}$
p'	The spectral gap of $\mathbf{P}$
m	Model size in bits
B	Total transmitted bits
Z	Wall-clock time
L	$f_v(\mathbf{x})$'s gradient is L-Lipschitz
σ^2	Upper Bound for local variance
ζ^2	Upper Bound for diversity
F	$f(\mathbf{x}_0) - f^*$
H^2	The second moment of the first return time to Node 0 for the Markov chain representing each walk
α	The degree of noniid-ness in the Dirichlet distribution is used to create disjoint noniid nodes; smaller values indicate a higher level of noniid-ness

B. Proof of Theorem 4.1

Motivated by (Stich, 2019), a virtual sequence $\{\tilde{\mathbf{x}}_t\}_{t\geq 0}$ is defined as follows.

$$\tilde{\mathbf{x}}_{t+1} = \tilde{\mathbf{x}}_t - \frac{\eta}{R}\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t}), \tag{4}$$

where we define $\hat{\tau}_t$ as the delay with which the gradient of the corresponding point $(\mathbf{x}_t^{r_t})$ will be computed. If we denote $t' = t + \hat{\tau}_t$, then it holds that $t' - \tau_{t'} = t$. We do not need to calculate this sequence in the algorithm explicitly and it is only used for the sake of analysis.

First, we illustrate how the virtual sequence, $\{\tilde{\mathbf{x}}_t\}_{t\geq 0}$, approaches to the optimal. Second, we depict that there is a little deviation from the virtual sequence in the actual iterates, $\mathbf{x}_t^{r_t}$. Finally, the convergence rate is proved.

Lemma B.1 (Descent Lemma for Multi-Walk)**.** *Under Assumptions 1, 2, 3, and learning rate $\eta \leq \frac{R}{6L}$, it holds that*

$$\mathbb{E}\, f(\tilde{\mathbf{x}}_{t+1}) \leq f(\tilde{\mathbf{x}}_t) - \frac{\eta}{4R}\|\nabla f(\mathbf{x}_t^{r_t})\|^2 + \frac{2c\eta}{R}\zeta^2(1-p')^{2|\mathcal{T}_{r_t}|} + \frac{3\eta^2L^2}{2R^2}\left(\sigma^2+\zeta^2\right) + \frac{\eta L^2}{2R}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2, \tag{5}$$

where $\mathcal{T}_{r_t} = \{t' \leq t : r_{t'} = r_t\}$.

Proof. Based on the definition of $\tilde{\mathbf{x}}_t$ and L-smoothness of $f(\mathbf{x})$ we have

$$\begin{aligned} f(\tilde{\mathbf{x}}_{t+1}) &= f(\tilde{\mathbf{x}}_t - \frac{\eta}{R}\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t})) && (6)\\ &\leq f(\tilde{\mathbf{x}}_t) + \frac{\eta}{R}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t})\rangle + \frac{\eta^2 L}{2R^2}\|\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t})\|^2. && (7)\end{aligned}$$

Lets take expectation of the second term on the right-hand side of (7).

$$\begin{aligned} &\frac{\eta}{R}\,\mathbb{E}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t})\rangle && (8)\\ &\quad= \frac{\eta}{R}\,\mathbb{E}_{v_t}\,\mathbb{E}_{\xi_{t+\hat{\tau}_t}}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t})\rangle && (9)\\ &\quad= \frac{\eta}{R}\,\mathbb{E}_{v_t}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla f_{v_t}(\mathbf{x}_t^{r_t})\rangle && (10)\\ &\quad= \frac{\eta}{R}\langle \nabla f(\tilde{\mathbf{x}}_t), -\,\mathbb{E}_{v_t}\,\nabla f_{v_t}(\mathbf{x}_t^{r_t})\rangle && (11)\\ &\quad= \frac{\eta}{R}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla f(\mathbf{x}_t^{r_t}) + \nabla f(\mathbf{x}_t^{r_t}) - \mathbb{E}_{v_t}\,\nabla f_{v_t}(\mathbf{x}_t^{r_t})\rangle && (12)\\ &\quad= \frac{\eta}{R}\underbrace{\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla f(\mathbf{x}_t^{r_t})\rangle}_{=:T_1} + \frac{\eta}{R}\underbrace{\langle \nabla f(\tilde{\mathbf{x}}_t), \nabla f(\mathbf{x}_t^{r_t}) - \mathbb{E}_{v_t}\,\nabla f_{v_t}(\mathbf{x}_t^{r_t})\rangle}_{=:T_2}. && (13)\end{aligned}$$

We estimate T_1 and T_2 separately.

$$T_1 = -\frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 - \frac{1}{2}\|\nabla f(\mathbf{x}_t^{r_t})\|^2 + \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t^{r_t})\|^2. \tag{14}$$

We also obtain

$$T_2 \leq \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 + \frac{1}{2}\|\,\mathbb{E}_{v_t}\left[\nabla f_{v_t}(\mathbf{x}_t^{r_t}) - f(\mathbf{x}_t^{r_t})\right]\|^2 \tag{15}$$

$$= \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 + \frac{1}{2}\|\sum_{v=1}^{V} P_v^t\left(\nabla f_{v_t}(\mathbf{x}_t^{r_t}) - f(\mathbf{x}_t^{r_t})\right)\|^2 \tag{16}$$

$$= \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 + \frac{1}{2}\|\sum_{v=1}^{V} (P_v^t - \pi_v)\left(\nabla f_{v_t}(\mathbf{x}_t^{r_t}) - f(\mathbf{x}_t^{r_t})\right)\|^2 \tag{17}$$

$$\leq \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 + \frac{1}{2}\left(\sum_{v=1}^{V} |P_v^t - \pi_v| \|\nabla f_{v_t}(\mathbf{x}_t^{r_t}) - f(\mathbf{x}_t^{r_t})\|\right)^2 \tag{18}$$

$$\leq \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 + \frac{1}{2}\zeta^2\left(\sum_{v=1}^{V} |P_v^t - \pi_v|\right)^2 \tag{19}$$

$$\leq \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 + \frac{1}{2}\zeta^2\left(2\|P^t - \pi\|_{TV}\right)^2 \tag{20}$$

$$\leq \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 + 2c\zeta^2(1-p')^{2|\mathcal{T}_{r_t}|}, \tag{21}$$

where (15) is based on the fact that for any $\lambda > 0$,

$$2\langle a, b\rangle \leq \lambda\|a\|^2 + \frac{1}{\lambda}\|b\|^2. \tag{22}$$

P_v^t shows the probability of being at node v at iteration t and π_v is the steady state distribution of node v. In (20) we have used the fact that the total variation distance between two probability distributions μ and ν on $\mathcal{X}$ satisfies

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)|. \tag{23}$$

(21) is based on the following well-known bound on the mixing time for a Markov chain (see, for example, Guruswami (2016); Levin & Peres (2017)).

$$\|P^t - \pi\|_{TV} \leq c(1-p')^{|\mathcal{T}_{r_t}|}, \tag{24}$$

where $\mathcal{T}_{r_t} = \{t' \leq t : r_{t'} = r_t\}$ is the set of all iteration on walk r_t. $(1-p')$ is the second largest eigenvalue of matrix $\mathbf{P}$ representing the irreducible aperiodic Markov chain of each walk and $c > 0$ is a constant.

So we get

$$\frac{\eta}{R}\,\mathbb{E}\langle\nabla f(\tilde{\mathbf{x}}_t), -\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t})\rangle \leq -\frac{\eta}{2R}\|\nabla f(\mathbf{x}_t^{r_t})\|^2 + \frac{\eta}{2R}\|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t^{r_t})\|^2 + \frac{2c\eta}{R}\zeta^2(1-p')^{2|\mathcal{T}_{r_t}|}. \tag{25}$$

Now we derive expectation of the last term on the right-hand side of (7).

$$\mathbb{E}\,\|\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t})\|^2 = \mathbb{E}\,\|\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t}) \pm \nabla f_{v_t}(\mathbf{x}_t^{r_t}) \pm \nabla f(\mathbf{x}_t^{r_t})\|^2 \tag{26}$$

$$\leq 3\,\mathbb{E}\,\|\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t}) - \nabla f_{v_t}(\mathbf{x}_t^{r_t})\|^2 + 3\,\mathbb{E}\,\|\nabla f_{v_t}(\mathbf{x}_t^{r_t}) - \nabla f(\mathbf{x}_t^{r_t})\|^2 + 3\|\nabla f(\mathbf{x}_t^{r_t})\|^2 \tag{27}$$

$$\leq 3\sigma^2 + 3\zeta^2 + 3\|\nabla f(\mathbf{x}_t^{r_t})\|^2, \tag{28}$$

where (27) is based on the following inequality.

$$\|\sum_{i=1}^{n}\mathbf{a}_i\|^2 \leq n\sum_{i=1}^{n}\|\mathbf{a}_i\|^2. \tag{29}$$

Combining these together and using L-smoothness to estimate $\|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t^{r_t})\|^2$ we obtain

$$\mathbb{E}\, f(\tilde{\mathbf{x}}_{t+1}) \leq f(\tilde{\mathbf{x}}_t) - \left(\frac{\eta}{2R} - \frac{3\eta^2 L}{2R^2}\right)\|\nabla f(\mathbf{x}_t^{r_t})\|^2 + \frac{\eta L^2}{2R}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2 + \frac{2c\eta}{R}\zeta^2(1-p')^{2|\mathcal{T}_{r_t}|} + \frac{3\eta^2 L}{2R^2}\left(\sigma^2 + \zeta^2\right). \tag{30}$$

Considering $\eta \leq \frac{R}{6L}$ we obtain

$$\mathbb{E}\, f(\tilde{\mathbf{x}}_{t+1}) \leq f(\tilde{\mathbf{x}}_t) - \frac{\eta}{4R}\|\nabla f(\mathbf{x}_t^{r_t})\|^2 + \frac{\eta L^2}{2R}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2 + \frac{2c\eta}{R}\zeta^2(1-p')^{2|\mathcal{T}_{r_t}|} + \frac{3\eta^2 L}{2R^2}\left(\sigma^2 + 2\zeta^2\right). \tag{31}$$

□

Lemma B.2 (Bounding Deviation for Multi-Walk)**.** *Under Assumptions 2, 3, 4, and learning rate $\eta \leq \frac{1}{7LH}$, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2 \leq 12V\sigma^2\eta^2 + 12H^2\zeta^2\eta^2 + \frac{1}{4L^2T}\sum_{t=1}^{T-1}\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{r_t})\|^2, \tag{32}$$

where H^2 is the second moment of the first return time to the Node 0.

Proof. First we define l_t^r as the last iteration before t when walk r has visited Node 0, *i.e.,* $l_t^r = \max\{t' \mid t' \leq t, r_t = r, v_t = 0\}$.

$$\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2 = \mathbb{E}\,\|\sum_{z=l^r_{l_t^{r_t}}, r_z \neq r_t}^{t-1} -\frac{\eta}{R}\nabla F_{v_z}(\mathbf{x}_z^{r_z}, \xi_{z+\hat{\tau}_z}) + \sum_{z=l_t^{r_t}, r_z = r_t}^{t-1}\left(1 - \frac{1}{R}\right)\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z}, \xi_{z+\hat{\tau}_z})\|^2 \tag{33}$$

$$\leq \frac{2}{R^2}\,\mathbb{E}\,\|\sum_{z=l^r_{l_t^{r_t}}, r_z \neq r_t}^{t-1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z}, \xi_{z+\hat{\tau}_z})\|^2 + 2\,\mathbb{E}\,\|\sum_{z=l_t^{r_t}, r_z = r_t}^{t-1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z}, \xi_{z+\hat{\tau}_z})\|^2 \tag{34}$$

$$\leq \underbrace{\frac{2}{R^2}\,\mathbb{E}\,\|\sum_{z\in U_t^1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z}, \xi_{z+\hat{\tau}_z})\|^2}_{:=T_1} + \underbrace{2\,\mathbb{E}\,\|\sum_{z\in U_t^2}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z}, \xi_{z+\hat{\tau}_z})\|^2}_{:=T_2}, \tag{35}$$

where $U_t^1 = \{l^r_{l_t^{r_t}} \leq z \leq t-1 \mid r_z \neq r_t\}$, and $U_t^2 = \{l_t^{r_t} \leq z \leq t-1 \mid r_z = r_t\}$.

We have

$$\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t}) = (\nabla F_{v_t}(\mathbf{x}_t^{r_t}, \xi_{t+\hat{\tau}_t}) - \nabla f_{v_t}(\mathbf{x}_t^{r_t})) + (\nabla f_{v_t}(\mathbf{x}_t^{r_t}) - \nabla f(\mathbf{x}_t^{r_t})) + \nabla f(\mathbf{x}_t^{r_t}). \tag{36}$$

So, based on (29) we can write

$$T_1 \leq \frac{6}{R^2}\,\mathbb{E}\Big(\|\sum_{z\in U_t^1}\eta\left(\nabla F_{v_z}(\mathbf{x}_z^{r_z}, \xi_{z+\hat{\tau}_z}) - \nabla f_{v_z}(\mathbf{x}_z^{r_z})\right)\|^2 + \|\sum_{z\in U_t^1}\eta\left(\nabla f_{v_z}(\mathbf{x}_z^{r_z}) - \nabla f(\mathbf{x}_z^{r_z})\right)\|^2 \tag{37}$$

$$+\|\sum_{z\in U_t^1}\eta\nabla f(\mathbf{x}_z^{r_z})\|^2\Big)$$

$$\leq \frac{6}{R^2}\,\mathbb{E}\left(\sum_{z\in U_t^1}\eta^2\sigma^2 + |U_t^1|\sum_{z\in U_t^1}\eta^2\zeta^2 + |U_t^1|\sum_{z\in U_t^1}\eta^2\|\nabla f(\mathbf{x}_z^{r_z})\|^2\right), \tag{38}$$

where in (38) we have applied (29) and the fact that for independent zero-mean random variables, we get a tighter bound as follows.

$$\mathbb{E}\,\|\sum_{i=1}^{n}\mathbf{a}_i\|^2 \leq \sum_{i=1}^{n}\mathbb{E}\,\|\mathbf{a}_i\|^2. \tag{39}$$

Averaging over T, we get

$$\frac{1}{T}\sum_{t=1}^{T-1} T_1 \leq \frac{6}{TR^2}\,\mathbb{E}\left(\sum_{t=1}^{T-1}\sum_{z\in U_t^1}\eta^2\sigma^2 + \sum_{t=1}^{T-1}|U_t^1|\sum_{z\in U_t^1}\eta^2\zeta^2 + \sum_{t=1}^{T-1}|U_t^1|\sum_{z\in U_t^1}\eta^2\|\nabla f(\mathbf{x}_z^{r_z})\|^2\right) \tag{40}$$

$$\leq \frac{6}{TR^2}\,\mathbb{E}\left(\sum_{t=1}^{T-1}|U_t^1|\eta^2\sigma^2 + \sum_{t=1}^{T-1}|U_t^1|^2\eta^2\zeta^2 + \sum_{t=1}^{T-1}|U_t^1|\sum_{z\in U_t^1}\eta^2\|\nabla f(\mathbf{x}_z^{r_z})\|^2\right) \tag{41}$$

$$\leq \frac{6}{TR^2}\,\mathbb{E}\left(\sum_{t=1}^{T-1}(R-1)h\eta^2\sigma^2 + \sum_{t=1}^{T-1}(R-1)^2h^2\eta^2\zeta^2 + (R-1)h\eta^2\sum_{t=1}^{T-1}\sum_{z\in U_t^1}\|\nabla f(\mathbf{x}_z^{r_z})\|^2\right) \tag{42}$$

$$\leq \frac{6}{TR^2}\,\mathbb{E}\left(\sum_{t=1}^{T-1}(R-1)h\eta^2\sigma^2 + \sum_{t=1}^{T-1}(R-1)^2h^2\eta^2\zeta^2 + (R-1)^2h^2\eta^2\sum_{t=1}^{T-1}\|\nabla f(\mathbf{x}_t^{r_t})\|^2\right) \tag{43}$$

$$\leq \frac{6}{TR^2}\left(\sum_{t=1}^{T-1}(R-1)\,V\eta^2\sigma^2 + \sum_{t=1}^{T-1}(R-1)^2\,H^2\eta^2\zeta^2 + \sum_{t=1}^{T-1}(R-1)^2\,H^2\eta^2\,\mathbb{E}\,\|\nabla f(\mathbf{x}_z^{r_z})\|^2\right) \tag{44}$$

$$\leq \frac{6}{T}\left(\sum_{t=1}^{T-1}\frac{V}{R}\eta^2\sigma^2 + \sum_{t=1}^{T-1}H^2\eta^2\zeta^2 + \sum_{t=1}^{T-1}H^2\eta^2\,\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{r_t})\|^2\right), \tag{45}$$

where in (42) and (43), we have used the fact that $|U_t^1|$ is upper bounded with $R-1$ times the first return time to Node 0 (h). Expectation of the first return time is $\frac{1}{\pi_0} = V$ and the second moment of this random variable is assumed H^2 that are applied in (44).

Following the same approach for T_2 and considering $|U_t^2|$ is upper bounded with the first return time to Node 0. we can get

$$\frac{1}{T}\sum_{t=1}^{T-1} T_2 \leq \frac{6}{T}\left(\sum_{t=1}^{T-1}V\eta^2\sigma^2 + \sum_{t=1}^{T-1}H^2\eta^2\zeta^2 + \sum_{t=1}^{T-1}H^2\eta^2\,\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{r_t})\|^2\right). \tag{46}$$

Putting these together, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2 \leq \frac{12}{T}\left(\sum_{t=1}^{T-1}V\eta^2\sigma^2 + \sum_{t=1}^{T-1}H^2\eta^2\zeta^2 + \sum_{t=1}^{T-1}H^2\eta^2\,\mathbb{E}\,\|\nabla f(\mathbf{x}_z^{r_z})\|^2\right) \tag{47}$$

$$\leq 12V\eta^2\sigma^2 + 12H^2\eta^2\zeta^2 + \frac{12H^2\eta^2}{T}\sum_{t=1}^{T-1}\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{r_t})\|^2. \tag{48}$$

Let $\eta \leq \frac{1}{7LH}$ to get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2 \leq 12V\sigma^2\eta^2 + 12H^2\zeta^2\eta^2 + \frac{1}{4L^2T}\sum_{t=1}^{T-1}\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{r_t})\|^2. \tag{49}$$

□

Now we complete the proof of Theorem 4.1. By multiplication of $\frac{4R}{\eta}$ in both sides and averaging over t in lemma B.1, we

get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t^{r_t})\right\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{4R}{\eta}\left(f(\tilde{\mathbf{x}}_t)-\mathbb{E}\,f\left(\tilde{\mathbf{x}}_{t+1}\right)\right)+\frac{1}{T}\sum_{t=0}^{T-1}8c\zeta^2(1-p')^{2|\mathcal{T}_{r_t}|}+\frac{6\eta L^2}{R}\left(\sigma^2+\zeta^2\right) \tag{50}$$

$$+\frac{1}{T}\sum_{t=0}^{T-1}2L^2\,\mathbb{E}\left\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{r_t}\right\|^2$$

$$\leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{4R}{\eta}\left(f(\tilde{\mathbf{x}}_t)-\mathbb{E}\,f\left(\tilde{\mathbf{x}}_{t+1}\right)\right)+\frac{1}{T}\sum_{t=0}^{T-1}8c\zeta^2(1-p')^{2|\mathcal{T}_{r_t}|}+\frac{6\eta L^2}{R}\left(\sigma^2+\zeta^2\right) \tag{51}$$

$$+\frac{1}{T}\sum_{t=0}^{T-1}2L^2\,\mathbb{E}\left\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{r_t}\right\|^2.$$

By replacing result of lemma B.2 and using $\sum_{t=0}^{T-1}(1-p')^{2|\mathcal{T}_{r_t}|}\leq\sum_{t=0}^{T-1}(1-p')^{|\mathcal{T}_{r_t}|}\leq R\sum_{t=0}^{T-1}(1-p')^t\leq\frac{R}{p'}$, then rearranging, we have

$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t^{r_t})\right\|^2 \leq \sum_{t=0}^{T-1}\frac{4R}{\eta}\left(f(\tilde{\mathbf{x}}_t)-\mathbb{E}\,f\left(\tilde{\mathbf{x}}_{t+1}\right)\right)+\frac{8cR\zeta^2}{p'T}+\frac{6\eta L^2}{R}\left(\sigma^2+\zeta^2\right)+24L^2\left(V\sigma^2+H^2\zeta^2\right)\eta^2 \tag{52}$$

Now, we state a lemma to obtain the final convergence rate based on (52).

Lemma B.3 (Similar to Lemma 16 in (Koloskova et al., 2020))**.** *For every non-negative sequence $\{r_t\}_{t\geq 0}$ and any parameters $d\geq 0, b\geq 0, c\geq 0, T\geq 0$, there exist a constant $\eta\leq\frac{1}{d}$, it holds*

$$\frac{1}{T\eta}\sum_{t=0}^{T-1}\left(r_t-r_{t+1}\right)+b\eta+c\eta^2\leq\frac{2\sqrt{br_0}}{\sqrt{T}}+2(\frac{r_0\sqrt{c}}{T})^{\frac{2}{3}}+\frac{dr_0}{T}. \tag{53}$$

Proof. By canceling the same terms in the telescopic sum, we get

$$\frac{1}{T\eta}\sum_{t=0}^{T-1}\left(r_t-r_{t+1}\right)+b\eta+c\eta^2\leq\frac{r_0}{T\eta}+b\eta+c\eta^2. \tag{54}$$

It is now followed by an η-tuning, the same way as in (Koloskova et al., 2020), which shows we need to choose $\eta=\min\{\frac{1}{d},\sqrt{\frac{r_0}{bT}},(\frac{r_0}{cT})^{\frac{1}{3}}\}$. □

Bounding the right hand side of inequality (52) with Lemma B.3 and considering that $\eta=\eta\leq\frac{1}{7LH}$, provides $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2$ is

$$\mathcal{O}\Bigg(\frac{(f(\mathbf{x}_0)-f^*)RLH}{T}+\frac{R\zeta^2}{p'T}+\frac{\sqrt{L(f(\mathbf{x}_0)-f^*)(\sigma^2+\zeta^2)}}{\sqrt{T}}+(\frac{RL(f(\mathbf{x}_0)-f^*)\sqrt{V\sigma^2+H^2\zeta^2}}{T})^{\frac{2}{3}}\Bigg). \tag{55}$$

This completes the proof of Theorem 4.1.

C. Proof of Theorem 4.2

For Async-Gossip algorithm, we define a virtual sequence $\{\tilde{\mathbf{x}}_t\}_{t\geq 0}$ as shown below.

$$\tilde{\mathbf{x}}_{t+1}=\tilde{\mathbf{x}}_t-\frac{\eta}{V}\nabla F_{v_t}(\mathbf{x}_t^{v_t},\xi_{t+\hat{\tau}_t}). \tag{56}$$

Lemma C.1 (Descent Lemma for Async-Gossip)**.** *Under Assumptions 1, 2, 3, and learning rate $\eta\leq\frac{V}{4L}$, it holds that*

$$\mathbb{E}\,f(\tilde{\mathbf{x}}_{t+1})\leq f(\tilde{\mathbf{x}}_t)-\frac{\eta}{4V}\|\nabla f(\mathbf{x}_t^{v_t})\|^2+\frac{\eta L^2}{2V}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{v_t}\|^2+\frac{\eta^2 L}{2V^2}\left(\sigma^2+2\zeta^2\right). \tag{57}$$

Proof. Based on the definition of $\tilde{\mathbf{x}}_t$ and L-smoothness of $f(\mathbf{x})$ we have

$$f(\tilde{\mathbf{x}}_{t+1}) = f(\tilde{\mathbf{x}}_t - \frac{\eta}{V}\nabla F_{v_t}(\mathbf{x}_t^{v_t}, \xi_{t+\hat{\tau}_t})) \tag{58}$$

$$\leq f(\tilde{\mathbf{x}}_t) + \frac{\eta}{V}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla F_{v_t}(\mathbf{x}_t^{v_t}, \xi_{t+\hat{\tau}_t})\rangle + \frac{\eta^2 L}{2V^2}\|\nabla F_{v_t}(\mathbf{x}_t^{v_t}, \xi_{t+\hat{\tau}_t})\|^2. \tag{59}$$

Lets take expectation of the second term on the right-hand side of (59)

$$\frac{\eta}{V}\,\mathbb{E}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla F_{v_t}(\mathbf{x}_t^{v_t}, \xi_{t+\hat{\tau}_t})\rangle \tag{60}$$

$$= \frac{\eta}{V}\,\mathbb{E}_{v_t}\,\mathbb{E}_{\xi_{t+\hat{\tau}_t}}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla F_{v_t}(\mathbf{x}_t^{v_t}, \xi_{t+\hat{\tau}_t})\rangle \tag{61}$$

$$= \frac{\eta}{V}\,\mathbb{E}_{v_t}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla f_{v_t}(\mathbf{x}_t^{v_t})\rangle \tag{62}$$

$$= \frac{\eta}{V}\langle \nabla f(\tilde{\mathbf{x}}_t), -\nabla f(\mathbf{x}_t^{v_t})\rangle \tag{63}$$

$$= -\frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 - \frac{1}{2}\|\nabla f(\mathbf{x}_t^{v_t})\|^2 + \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t^{v_t})\|^2 \tag{64}$$

$$\leq -\frac{1}{2}\|\nabla f(\mathbf{x}_t^{v_t})\|^2 + \frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t^{v_t})\|^2. \tag{65}$$

Now we derive expectation of the last term on the right-hand side of (59).

$$\mathbb{E}\,\|\nabla F_{v_t}(\mathbf{x}_t^{v_t}, \xi_{t+\hat{\tau}_t})\|^2 = \mathbb{E}\,\|\nabla F_{v_t}(\mathbf{x}_t^{v_t}, \xi_{t+\hat{\tau}_t}) \pm \nabla f_{v_t}(\mathbf{x}_t^{v_t}) \pm \nabla f(\mathbf{x}_t^{v_t})\|^2 \tag{66}$$

$$\leq \sigma^2 + 2\,\mathbb{E}\,\|\nabla f_{v_t}(\mathbf{x}_t^{v_t}) - \nabla f(\mathbf{x}_t^{v_t})\|^2 + 2\|\nabla f(\mathbf{x}_t^{v_t})\|^2 \tag{67}$$

$$\leq \sigma^2 + 2\zeta^2 + 2\|\nabla f(\mathbf{x}_t^{v_t})\|^2. \tag{68}$$

Combining these together and using L-smoothness to estimate $\|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t^{v_t})\|^2$ we obtain

$$\mathbb{E}\,f(\tilde{\mathbf{x}}_{t+1}) \leq f(\tilde{\mathbf{x}}_t) - \left(\frac{\eta}{2V} - \frac{\eta^2 L}{V^2}\right)\|\nabla f(\mathbf{x}_t^{v_t})\|^2 + \frac{\eta L^2}{2V}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{v_t}\|^2 + \frac{\eta^2 L}{2V^2}\left(\sigma^2 + 2\zeta^2\right). \tag{69}$$

Considering $\eta \leq \frac{V}{4L}$ we obtain

$$\mathbb{E}\,f(\tilde{\mathbf{x}}_{t+1}) \leq f(\tilde{\mathbf{x}}_t) - \frac{\eta}{4V}\|\nabla f(\mathbf{x}_t^{v_t})\|^2 + \frac{\eta L^2}{2V}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{v_t}\|^2 + \frac{\eta^2 L}{2V^2}\left(\sigma^2 + 2\zeta^2\right). \tag{70}$$

□

Lemma C.2 (Bounding Deviation for Async-Gossip)**.** *Under Assumptions 2, 3, 4, and learning rate $\eta \leq \frac{p}{14L}$, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{v_t}\|^2 \leq \frac{1}{4L^2}\sum_{z=0}^{T-1}\|\nabla f(\mathbf{x}_z^{v_z})\|^2 + \left(\frac{16\sigma^2}{p} + \frac{96\zeta^2}{p^2}\right)\sum_{t=0}^{T-1}\eta^2. \tag{71}$$

Proof. We will be using the following matrix notation.

$$\mathbf{X}_t := \left[\mathbf{x}_t^1, \ldots, \mathbf{x}_t^V\right] \in \mathbb{R}^{d\times V}, \tag{72}$$

$$\tilde{\mathbf{X}}_t := [\tilde{\mathbf{x}}_t, \ldots, \tilde{\mathbf{x}}_t] \in \mathbb{R}^{d\times V}, \tag{73}$$

$$\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t}) := \left[\nabla F_1(\mathbf{x}_t^1, \xi_{t+\hat{\tau}_t}), \ldots, \nabla F_V(\mathbf{x}_t^V, \xi_{t+\hat{\tau}_t})\right] \in \mathbb{R}^{d\times V}, \tag{74}$$

$$\partial f(\mathbf{X}_t) := \left[\nabla f_1(\mathbf{x}_t^1), \ldots, \nabla f_V(\mathbf{x}_t^V)\right] \in \mathbb{R}^{d\times V}. \tag{75}$$

Considering that v_t is uniformly random among all nodes, we have

$$V\,\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{v_t}\|^2 = \mathbb{E}\,\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|_F^2 \tag{76}$$

$$= \mathbb{E}\,\|\mathbf{X}_{t-1}\mathbf{W} - \eta\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t})\mathbf{W} - \tilde{\mathbf{X}}_t\|_F^2 \tag{77}$$

$$= \mathbb{E}\,\|\mathbf{X}_{t-1}\mathbf{W} - \eta\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t})\mathbf{W} - \tilde{\mathbf{X}}_{t-1} + \frac{\eta}{V}\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t})\|_F^2 \tag{78}$$

$$= \mathbb{E}\,\|\mathbf{X}_{t-1}\mathbf{W} - \tilde{\mathbf{X}}_{t-1} - \eta\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t})\left(\mathbf{W} - \frac{\mathbf{I}}{V}\right)\|_F^2 \tag{79}$$

$$\leq \mathbb{E}\,\|\mathbf{X}_{t-1}\mathbf{W} - \tilde{\mathbf{X}}_{t-1} - \eta\partial f(\mathbf{X}_t)\left(\mathbf{W} - \frac{\mathbf{I}}{V}\right)\|_F^2 \tag{80}$$

$$+\|\eta\left(\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t}) - \partial f(\mathbf{X}_t)\right)\left(\mathbf{W} - \frac{\mathbf{I}}{V}\right)\|_F^2,$$

where we used that $\mathbb{E}\,\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t}) = \partial f(\mathbf{X}_t)$. We can further separate the second term as the following.

$$V\,\mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{v_t}\|^2 \leq \mathbb{E}\,\|\mathbf{X}_{t-1}\mathbf{W} - \tilde{\mathbf{X}}_{t-1} - \eta\partial f(\mathbf{X}_t)\left(\mathbf{W} - \frac{\mathbf{I}}{V}\right)\|_F^2 \tag{81}$$

$$+2\eta^2\|\left(\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t}) - \partial f(\mathbf{X}_t)\right)\mathbf{W}\|_F^2 + 2\frac{\eta^2}{V^2}\|\left(\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t}) - \partial f(\mathbf{X}_t)\right)\|_F^2$$

$$\leq \mathbb{E}\,\|\mathbf{X}_{t-1}\mathbf{W} - \tilde{\mathbf{X}}_{t-1} - \eta\partial f(\mathbf{X}_t)\left(\mathbf{W} - \frac{\mathbf{I}}{V}\right)\|_F^2 \tag{82}$$

$$+2\eta^2\|\left(\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t}) - \partial f(\mathbf{X}_t)\right)\|_F^2 + 2\frac{\eta^2}{V^2}\|\left(\partial F(\mathbf{X}_t, \xi_{t+\hat{\tau}_t}) - \partial f(\mathbf{X}_t)\right)\|_F^2$$

$$\leq (1+\lambda)\,\mathbb{E}\,\|\mathbf{X}_{t-1}\mathbf{W} - \tilde{\mathbf{X}}_{t-1}\|_F^2 + \left(1+\lambda^{-1}\right)\mathbb{E}\,\|\eta\partial f(\mathbf{X}_t)\left(\mathbf{W} - \frac{\mathbf{I}}{V}\right)\|_F^2 + 2\eta^2 V\sigma^2 + 2\frac{\eta^2}{V^2}V\sigma^2 \tag{83}$$

$$\leq (1+\lambda)\,\mathbb{E}\,\|\mathbf{X}_{t-1}\mathbf{W} - \tilde{\mathbf{X}}_{t-1}\|_F^2 + 2\eta^2\left(1+\lambda^{-1}\right)\mathbb{E}\,\|\partial f(\mathbf{X}_t)\mathbf{W}\|_F^2 + \frac{2\eta^2\left(1+\lambda^{-1}\right)}{V^2}\,\mathbb{E}\,\|\partial f(\mathbf{X}_t)\|_F^2 \tag{84}$$

$$+4\eta^2 V\sigma^2$$

$$\leq (1+\lambda)\,\mathbb{E}\,\|\mathbf{X}_{t-1}\mathbf{W} - \tilde{\mathbf{X}}_{t-1}\|_F^2 + 4\eta^2\left(1+\lambda^{-1}\right)\mathbb{E}\,\|\partial f(\mathbf{X}_t)\|_F^2 + 4\eta^2 V\sigma^2 \tag{85}$$

$$\leq (1+\lambda)\,(1-p)\,\mathbb{E}\,\|\mathbf{X}_{t-1} - \tilde{\mathbf{X}}_{t-1}\|_F^2 + 4\eta^2\left(1+\lambda^{-1}\right)\underbrace{\mathbb{E}\,\|\partial f(\mathbf{X}_t)\|_F^2}_{:=T_1} + 4\eta^2 V\sigma^2. \tag{86}$$

(83) is based on the fact that for any $\lambda > 0$,

$$\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\lambda)\|\mathbf{a}\|^2 + (1+\lambda^{-1})\|\mathbf{b}\|^2. \tag{87}$$

We bound T_1 separately.

$$T_1 = \mathbb{E}\,\|\partial f(\mathbf{X}_t)\|_F^2 \tag{88}$$

$$= \mathbb{E}\sum_{v=1}^{V}\|\nabla f_v(\mathbf{x}_t^v)\|^2 \tag{89}$$

$$\leq \mathbb{E}\sum_{v=1}^{V}2\|\nabla f_v(\mathbf{x}_t^v) - \nabla f(\mathbf{x}_t^v)\|^2 + \mathbb{E}\sum_{v=1}^{V}2\|\nabla f(\mathbf{x}_t^v)\|^2 \tag{90}$$

$$\leq \mathbb{E}\sum_{v=1}^{V}2\zeta^2 + \mathbb{E}\sum_{v=1}^{V}2\|\nabla f(\mathbf{x}_t^v)\|^2 \tag{91}$$

$$= 2V\zeta^2 + 2V\,\mathbb{E}\,\mathbb{E}_{v_t}\,\|\nabla f(\mathbf{x}_t^{v_t})\|^2 \tag{92}$$

$$= 2V\zeta^2 + 2V\,\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{v_t})\|^2. \tag{93}$$

So, we get

$$\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{v_t}\|^2 \leq (1+\lambda)\,(1-p)\,\mathbb{E}\,\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}^{v_{t-1}}\|^2 + 8\eta^2\left(1+\lambda^{-1}\right)\left(\zeta^2 + \|\nabla f(\mathbf{x}_t^{v_t})\|^2\right) + 4\eta^2\sigma^2 \tag{94}$$

$$\leq \left(1-\frac{p}{2}\right)\mathbb{E}\,\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}^{v_{t-1}}\|^2 + \frac{24}{p}\eta^2\zeta^2 + \frac{24}{p}\eta^2\|\nabla f(\mathbf{x}_t^{v_t})\|^2 + 4\eta^2\sigma^2 \tag{95}$$

$$\leq \left(1-\frac{p}{2}\right)^{t-1}\mathbb{E}\,\|\tilde{\mathbf{x}}_0 - \mathbf{x}_0^{v_0}\|^2 + \frac{24\zeta^2}{p}\sum_{z=0}^{t-1}\eta^2\left(1-\frac{p}{2}\right)^{t-z} + \frac{24}{p}\sum_{z=0}^{t-1}\eta^2\left(1-\frac{p}{2}\right)^{t-z}\|\nabla f(\mathbf{x}_z^{v_z})\|^2 \tag{96}$$

$$+4\sigma^2\sum_{z=0}^{t-1}\eta^2\left(1-\frac{p}{2}\right)^{t-z}$$

$$\leq \frac{24\zeta^2}{p}\eta^2\sum_{z=0}^{t-1}\left(1-\frac{p}{2}\right)^{t-z} + \frac{24}{p}\eta^2\sum_{z=0}^{t-1}\left(1-\frac{p}{2}\right)^{t-z}\|\nabla f(\mathbf{x}_z^{v_z})\|^2 + 4\sigma^2\eta^2\sum_{z=0}^{t-1}\left(1-\frac{p}{2}\right)^{t-z} \tag{97}$$

$$\leq \frac{24}{p}\eta^2\sum_{z=0}^{t-1}\left(1-\frac{p}{2}\right)^{t-z}\|\nabla f(\mathbf{x}_z^{v_z})\|^2 + \left(\frac{8\sigma^2}{p} + \frac{48\zeta^2}{p^2}\right)\eta^2, \tag{98}$$

where we used $\lambda = \frac{p}{2}$ in (95).

Now by averaging over T and considering $\eta \leq \frac{p}{14L}$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{v_t}\|^2 \leq \frac{24}{pT}\sum_{t=0}^{T-1}\eta^2\sum_{z=0}^{t-1}\left(1-\frac{p}{2}\right)^{t-z}\|\nabla f(\mathbf{x}_z^{v_z})\|^2 + \left(\frac{8\sigma^2}{p} + \frac{48\zeta^2}{p^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\eta^2 \tag{99}$$

$$\leq \frac{24p}{196L^2T}\sum_{z=0}^{T-1}\|\nabla f(\mathbf{x}_z^{v_z})\|^2\sum_{t=j+1}^{T-1}\left(1-\frac{p}{2}\right)^{t-z} + \left(\frac{8\sigma^2}{p} + \frac{48\zeta^2}{p^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\eta^2 \tag{100}$$

$$\leq \frac{24p}{196L^2T}\sum_{z=0}^{T-1}\|\nabla f(\mathbf{x}_z^{v_z})\|^2\sum_{t=0}^{\infty}\left(1-\frac{p}{2}\right)^{t-z} + \left(\frac{8\sigma^2}{p} + \frac{48\zeta^2}{p^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\eta^2 \tag{101}$$

$$\leq \frac{48}{196L^2T}\sum_{z=0}^{T-1}\|\nabla f(\mathbf{x}_z^{v_z})\|^2 + \left(\frac{8\sigma^2}{p} + \frac{48\zeta^2}{p^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\eta^2 \tag{102}$$

$$\leq \frac{1}{4L^2T}\sum_{z=0}^{T-1}\|\nabla f(\mathbf{x}_z^{v_z})\|^2 + \left(\frac{8\sigma^2}{p} + \frac{48\zeta^2}{p^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\eta^2. \tag{103}$$

□

Now we complete the proof of Theorem 4.2. By multiplication of $\frac{4V}{\eta}$ in both sides and averaging over t in lemma C.1, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{v_t})\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{4V}{\eta}\left(f(\tilde{\mathbf{x}}_t) - \mathbb{E}\,f(\tilde{\mathbf{x}}_{t+1})\right) + \frac{4L\eta}{V}\left(\sigma^2 + 2\zeta^2\right) + \frac{1}{T}\sum_{t=0}^{T-1}2L^2\,\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{v_t}\|^2. \tag{104}$$

By replacing result of lemma C.2 and rearranging, we have

$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{v_t})\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{4V}{\eta}\left(f(\tilde{\mathbf{x}}_t) - \mathbb{E}\,f(\tilde{\mathbf{x}}_{t+1})\right) + \frac{4L\eta}{V}\left(\sigma^2 + 2\zeta^2\right) + 2L^2\eta^2\left(\frac{8\sigma^2}{p} + \frac{48\zeta^2}{p^2}\right). \tag{105}$$

Bounding the right hand side of inequality (105) with Lemma B.3 and considering that $\eta = \eta \leq \frac{p}{14L}$, provides $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\nabla f(\mathbf{x}_t^{v_t})\|^2$ is

$$\mathcal{O}\left(\frac{(f(\mathbf{x}_0) - f^*)VL}{pT} + \frac{\sqrt{L(f(\mathbf{x}_0) - f^*)(\sigma^2+\zeta^2)}}{\sqrt{T}} + \left(\frac{VL(f(\mathbf{x}_0) - f^*)\sqrt{\frac{\sigma^2}{p} + \frac{\zeta^2}{p^2}}}{T}\right)^{\frac{2}{3}}\right). \tag{106}$$

D. Proof of Theorem 5.1

Lemma D.1 (Bounding Deviation for Multi-Walk with Failure)**.** *Under Assumptions 2, 3, 4, and learning rate* $\eta \leq \frac{1}{15LE\max_i H_i}$*, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2 \leq 36V\sigma^2\eta^2 + 36H^2\zeta^2\eta^2 + \frac{1}{4L^2T}\sum_{i=0}^{E}\sum_{t=e_i}^{e_{i+1}-1}\mathbb{E}\,\|\nabla f^i(\mathbf{x}_t^{r_t})\|^2, \tag{107}$$

where H_i^2 *is the second moment of the first return time to Node* 0 *chosen after the* i*-th failure out of* E *failures. We assume* e_i *as the iteration of* i*-th failure, also* $e_0 = 0, e_{E+1} = T$*.*

Proof. Recall l_t^r as the last iteration before t when walk r has visited Node 0, *i.e.,* $l_t^r = \max\{t' \mid t' \leq t, r_t = r, v_t = 0\}$. We also define $d_t^r = \min\{t' \mid t' \geq t, r_t = r, v_t = 0\}$ and r_{e_i} as the first walk that reaches to the new Node 0 after e_i.

$$\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2 = \mathbb{E}\,\|\sum_{z=l_{l_t^{r_t}}^{r_t},r_z\neq r_t}^{t-1} -\frac{\eta}{R}\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z}) + \sum_{z=l_t^{r_t},r_z=r_t}^{t-1}\left(1-\frac{1}{R}\right)\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z}) \tag{108}$$

$$+\sum_{i=1}^{E}\sum_{r=1}^{R}\sum_{z=l_{e_i}^{r_t},r_z=r}^{d_{e_i}^r-1} -\frac{\eta}{R}\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z}) + \sum_{i=1}^{E}\sum_{z=l_{e_i}^{r_{e_i}},r_z=r_{e_i}}^{d_{e_i}^{r_{e_i}}-1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z})\|^2 \tag{109}$$

$$\leq \frac{4}{R^2}\mathbb{E}\,\|\sum_{z=l_{l_t^{r_t}}^{r_t},r_z\neq r_t}^{t-1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z})\|^2 + 4\,\mathbb{E}\,\|\sum_{z=l_t^{r_t},r_z=r_t}^{t-1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z})\|^2 \tag{110}$$

$$+\frac{4}{R^2}\mathbb{E}\,\|\sum_{i=1}^{E}\sum_{r=1}^{R}\sum_{z=l_{e_i}^{r_t},r_z=r}^{d_{e_i}^r-1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z})\|^2 + 4\,\mathbb{E}\,\|\sum_{i=1}^{E}\sum_{z=l_{e_i}^{r_{e_i}},r_z=r_{e_i}}^{d_{e_i}^{r_{e_i}}-1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z})\|^2 \tag{111}$$

$$\leq \underbrace{\frac{4}{R^2}\mathbb{E}\,\|\sum_{z\in U_t^1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z})\|^2}_{:=T_1} + \underbrace{4\,\mathbb{E}\,\|\sum_{z\in U_t^2}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z})\|^2}_{:=T_2} \tag{112}$$

$$+\underbrace{\frac{4}{R^2}\mathbb{E}\,\|\sum_{i=1}^{E}\sum_{r=1}^{R}\sum_{z=l_{e_i}^{r_t},r_z=r}^{d_{e_i}^r-1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z})\|^2}_{:=T_3} + \underbrace{4\,\mathbb{E}\,\|\sum_{i=1}^{E}\sum_{z=l_{e_i}^{r_{e_i}},r_z=r_{e_i}}^{d_{e_i}^{r_{e_i}}-1}\eta\nabla F_{v_z}(\mathbf{x}_z^{r_z},\xi_{z+\hat{\tau}_z})\|^2}_{:=T_4}, \tag{113}$$

where $U_t^1 = \{l_{l_t^{r_t}}^r \leq z \leq t-1 \mid r_z \neq r_t\}$, and $U_t^2 = \{l_t^{r_t} \leq z \leq t-1 \mid r_z = r_t\}$. In the same way as we bounded T_1 and T_2, in Lemma B.2, we can bound T_3 and T_4.

T_3 and T_4 is bounded by first and second moment of a random variable that the sum of two quantities: the first return time to old Node 0 (h_{i-1}) and the hitting time to the new Node 0. This hitting time is, in turn, upper-bounded by the first return time to new Node 0 (h_i). Expectation of the this random variable is bounded with $2V$ and the second moment of this random variable is bounded with twice the sum of the second moments of two random variables. So the second moment is $2(\max_i H_i^2)$.

Following the same approach as in Lemma B.2 and assuimg $\eta \leq \frac{1}{15LE\max_i H_i}$ to get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{r_t}\|^2 \leq 36EV\sigma^2\eta^2 + 36E^2\zeta^2\eta^2\max_i H_i^2 + \frac{1}{4L^2T}\sum_{i=0}^{E}\sum_{t=e_i}^{e_{i+1}-1}\mathbb{E}\,\|\nabla f^i(\mathbf{x}_t^{r_t})\|^2. \tag{114}$$

□

Now we complete the proof of Theorem 5.1. By multiplication of $\frac{4R}{\eta}$ in both sides and averaging over t in lemma B.1, we get

$$\frac{1}{T}\sum_{i=0}^{E}\sum_{t=e_i}^{e_{i+1}-1}\mathbb{E}\left\|\nabla f^i(\mathbf{x}_t^{r_t})\right\|^2 \le \frac{1}{T}\sum_{i=0}^{E}\sum_{t=e_i}^{e_{i+1}-1}\frac{4R}{\eta}\left(f^i(\tilde{\mathbf{x}}_t)-\mathbb{E}\,f^i(\tilde{\mathbf{x}}_{t+1})\right)+\frac{1}{T}\sum_{i=0}^{E}\sum_{t=e_i}^{e_{i+1}-1}8c\zeta^2(1-p_i')^{2|\mathcal{T}_{r_t}^i|} \tag{115}$$
$$+\frac{6\eta L^2}{R}\left(\sigma^2+\zeta^2\right)+\frac{1}{T}\sum_{t=0}^{T-1}2L^2\,\mathbb{E}\left\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{r_t}\right\|^2.$$

By replacing result of lemma D.1 and using $\sum_{i=0}^{E}\sum_{t=e_i}^{e_{i+1}-1}(1-p')^{2|\mathcal{T}_{r_t}^i|}\le\sum_{i=0}^{E}\sum_{t=0}^{T-1}(1-p_i')^{|\mathcal{T}_{r_t}^i|}\le\sum_{i=0}^{E}R\sum_{t=0}^{T-1}(1-p_i')^t\le\sum_{i=0}^{E}\frac{R}{p_i'}$, then rearranging, we have

$$\frac{1}{2T}\sum_{i=0}^{E}\sum_{t=e_i}^{e_{i+1}-1}\mathbb{E}\left\|\nabla f^i(\mathbf{x}_t^{r_t})\right\|^2 \le \frac{1}{T}\sum_{i=0}^{E}\sum_{t=e_i}^{e_{i+1}-1}\frac{4R}{\eta}\left(f^i(\tilde{\mathbf{x}}_t)-\mathbb{E}\,f^i(\tilde{\mathbf{x}}_{t+1})\right)+8c\zeta^2\sum_{i=0}^{E}\frac{R}{p_i'}+\frac{6\eta L^2}{R}\left(\sigma^2+\zeta^2\right) \tag{116}$$
$$+72EL^2\left(V\sigma^2+E\max_i H_i^2\zeta^2\right)\eta^2$$

We assume the failure of Node 0 does not change the objective function, *i.e.,* $f^0(\mathbf{x})=\cdots=f^E(\mathbf{x})=f(\mathbf{x})$. Then

$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t^{r_t})\right\|^2 \le \frac{1}{T}\sum_{t=0}^{T-1}\frac{4R}{\eta}\left(f(\tilde{\mathbf{x}}_t)-\mathbb{E}\,f\left(\tilde{\mathbf{x}}_{t+1}\right)\right)+8c\zeta^2\sum_{i=0}^{E}\frac{R}{p_i'}+\frac{6\eta L^2}{R}\left(\sigma^2+\zeta^2\right) \tag{117}$$
$$+72EL^2\left(V\sigma^2+E\max_i H_i^2\zeta^2\right)\eta^2$$

Bounding the right hand side of inequality (117) with Lemma B.3 and considering that $\eta=\eta\le\frac{1}{15LE\max_i H_i}$, provides $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t^{r_t})\right\|^2$ is

$$\mathcal{O}\Bigg(\frac{(f(\mathbf{x}_0)-f^*)RLE\max_i H_i}{T}+\sum_{i=0}^{E}\frac{R\zeta^2}{p_i'}+\frac{\sqrt{L(f(\mathbf{x}_0)-f^*)(\sigma^2+\zeta^2)}}{\sqrt{T}} \tag{118}$$
$$+\Big(\frac{RL(f(\mathbf{x}_0)-f^*)\sqrt{EV\sigma^2+E^2\zeta^2\max_i H_i^2}}{T}\Big)^{\frac{2}{3}}\Bigg).$$

E. Derivation of H^2

E.1. Complete graph under Metropolis–Hastings P

We have a complete graph on V vertices, labeled $0,1,\ldots,V-1$. Each vertex i has degree $\deg(i)=V-1$. The Metropolis–Hastings (MH) probability between two adjacent vertices (i,j) is

$$p_{ij} \;=\; \min\Big\{\frac{1}{\deg(i)+1},\,\frac{1}{\deg(j)+1}\Big\}.$$

Since $\deg(i)+1=V$ for every vertex i in a complete graph, it follows that

$$p_{ij} \;=\; \min\big\{\tfrac{1}{V},\,\tfrac{1}{V}\big\} \;=\; \tfrac{1}{V}.$$

Moreover, the leftover probability is also $\frac{1}{V}$ for staying in place (lazy step). Hence, from any state i, the chain picks each of the V vertices with probability $1/V$, including i itself.

Because each state is chosen uniformly at each step, independently of the past, the process $\{X_k\}_{k\ge0}$ is an iid sequence of $\mathrm{Uniform}\{0,\ldots,V-1\}$.

Define the first return time to state 0 by

$$h \;=\; \min\{\, k \geq 1 : X_k = 0 \;\mid\; X_0 = 0\}.$$

Since each X_k for $k \geq 1$ is uniformly distributed over $\{0, \ldots, V-1\}$, the probability that $X_k = 0$ is $1/V$, independent of previous steps. Thus, h is a $\mathrm{Geometric}(p = 1/V)$ random variable in the usual "first success" sense (with success probability $1/V$ each trial).

For a geometric random variable $Y \sim \mathrm{Geom}(p)$ (where $p = 1/V$), the second moment is a standard formula:

$$\mathbb{E}[Y^2] \;=\; \frac{2-p}{p^2}.$$

Plugging in $p = 1/V$ yields

$$H^2 \;=\; \mathbb{E}[h^2] \;=\; \frac{2-\frac{1}{V}}{\left(\frac{1}{V}\right)^2} \;=\; V^2\Big(2-\frac{1}{V}\Big) \;=\; 2V^2 - V.$$

Hence, under Metropolis-Hastings on the complete graph of V vertices, the first return time to state 0 has second moment $2\,V^2 - V$.

E.2. Cycle graph under Metropolis-Hastings P

Consider a cycle graph with V vertices labeled $0, 1, \ldots, V-1$ (indices mod V). Each vertex i has degree 2, so the Metropolis–Hastings (MH) transition rule gives

$$p_{i,i} \;=\; \frac{1}{3}, \quad p_{i,i+1} \;=\; \frac{1}{3}, \quad p_{i,i-1} \;=\; \frac{1}{3},$$

where addition/subtraction of indices is modulo V. Hence from each state i, the chain either stays put with probability $1/3$, or moves one step left or right (each with probability $1/3$).

Define

$$h \;=\; \min\{\, k \geq 1 : X_k = 0 \;\mid\; X_0 = 0\}.$$

Our goal is to derive $\mathbb{E}[h^2]$. To handle this systematically, for any initial state i, define the *first hitting time* of 0:

$$T_0 \;=\; \min\{\, k \geq 1 : X_k = 0\}.$$

And then set

$$m_i \;=\; \mathbb{E}[T_0 \mid X_0 = i], \quad M_i \;=\; \mathbb{E}[\,T_0^2 \mid X_0 = i].$$

In particular, $\mathbb{E}[h^2] = M_0$, since for $i = 0$, we interpret T_0 as the *first return time* to 0.

Recurrences for the First Moments (m_i). Based on the symmetry of the topology, we consider only half of the vertices, i.e., $2 \leq i \leq \lceil \frac{V}{2} \rceil$.

(a) m_0. Starting at 0, in one step:

- With probability $1/3$, we *stay* at 0, so the hitting time $T_0 = 1$ immediately.
- With probability $1/3$ each, we move to 1 or $V-1$. From such a neighbor, the expected time to hit 0 is $1 + m_1$ (by symmetry, m_1 is the same whether we step to 1 or $V-1$).

Thus

$$m_0 \;=\; \frac{1}{3}\cdot 1 \;+\; \frac{1}{3}\big(1+m_1\big) \;+\; \frac{1}{3}\big(1+m_1\big) \;=\; 1 + \frac{2}{3}\,m_1. \tag{119}$$

(b) m_1 **(separate expression).** From state 1:

- With probability $1/3$, we jump *directly* to 0. Then $T_0 = 1$ (not $1 + m_0$, because hitting 0 completes the journey right away).
- With probability $1/3$, we *stay* at 1. Then $T_0 = 1 + m_1$.
- With probability $1/3$, we move to 2. Then $T_0 = 1 + m_2$.

Hence

$$m_1 \;=\; \frac{1}{3}\cdot 1 \;+\; \frac{1}{3}\big(1+m_1\big) \;+\; \frac{1}{3}\big(1+m_2\big).$$

Simplify:

$$m_1 \;=\; 1+\frac{1}{3}\,m_1+\frac{1}{3}\,m_2 \;\Longrightarrow\; \frac{2}{3}\,m_1 = 1+\frac{1}{3}\,m_2 \;\Longrightarrow\; m_1 = \frac{3}{2}+\frac{1}{2}\,m_2. \tag{120}$$

(c) General m_i **for** $2 \le i \le \lceil \frac{V}{2} \rceil$**.** From state i, we have three possibilities (stay at i, move to $i+1$, or move to $i-1$). Each event occurs with probability $1/3$, and in each case we add 1 step plus the hitting time from the new state. Thus

$$m_i \;=\; \frac{1}{3}\big(1+m_i\big) \;+\; \frac{1}{3}\big(1+m_{i+1}\big) \;+\; \frac{1}{3}\big(1+m_{i-1}\big),$$

where indices are taken mod V. Rearranging gives

$$m_i \;=\; \frac{3+m_{i+1}+m_{i-1}}{2}. \tag{121}$$

Recurrences for the Second Moments (M_i)**.** Define $M_i = \mathbb{E}[T_0^2 \mid X_0 = i]$. We again do a first-step analysis.

(a) M_0**.** From state 0:

- With prob $1/3$, stay at 0 immediately: $T_0 = 1$, contributing 1^2.
- With prob $2/3$, move to a neighbor (1 or $V-1$), then $T_0 = 1 + T_0'$. Squaring, $(1+T_0')^2 = 1 + 2T_0' + (T_0')^2$, so $\mathbb{E}[(1+T_0')^2] = 1 + 2\,m_1 + M_1$.

Hence

$$M_0 \;=\; \frac{1}{3}\cdot 1^2 \;+\; \frac{2}{3}\left[1+2\,m_1+M_1\right] \;=\; 1+\frac{4}{3}\,m_1+\frac{2}{3}\,M_1. \tag{122}$$

(b) M_1**.** From state 1:

- With prob $1/3$, jump directly to 0: $T_0 = 1$, so contribution 1^2.
- With prob $1/3$, stay at 1: then $T_0 = 1 + T_0'$, so $\mathbb{E}[(1+T_0')^2] = 1 + 2\,m_1 + M_1$.
- With prob $1/3$, move to 2: then $T_0 = 1 + T_0''$, so $\mathbb{E}[(1+T_0'')^2] = 1 + 2\,m_2 + M_2$.

Thus

$$M_1 \;=\; \frac{1}{3}\cdot 1 \;+\; \frac{1}{3}\big[1+2\,m_1+M_1\big] \;+\; \frac{1}{3}\big[1+2\,m_2+M_2\big].$$

Simplifying leads to a linear relation among M_1, m_1, m_2, and M_2:

$$M_1 \;=\; 1+\frac{2}{3}\,m_1+\frac{2}{3}\,m_2+\frac{1}{3}\,M_1+\frac{1}{3}\,M_2 \tag{123}$$

$$= \frac{3}{2}+m_1+m_2+\frac{1}{2}\,M_2 \tag{124}$$

$$= 3m_1-\frac{3}{2}+\frac{1}{2}\,M_2. \tag{125}$$

(c) General M_i for $2 \le i \le \lceil \frac{V}{2} \rceil$. By the same logic:

$$M_i \;=\; \frac{1}{3}\big[1 + 2\,m_i + M_i\big] + \frac{1}{3}\big[1 + 2\,m_{i+1} + M_{i+1}\big] + \frac{1}{3}\big[1 + 2\,m_{i-1} + M_{i-1}\big],$$

with indices mod V. Rearrange to get

$$M_i \;=\; \frac{3}{2} \;+\; \big(m_i + m_{i+1} + m_{i-1}\big) \;+\; \frac{1}{2}\,\big(M_{i+1} + M_{i-1}\big) \tag{126}$$

$$\;=\; \frac{3}{2} \;+\; 3\big(m_i - 1\big) \;+\; \frac{1}{2}\,\big(M_{i+1} + M_{i-1}\big) \tag{127}$$

$$\;=\; 3m_i - \frac{3}{2} \;+\; \frac{1}{2}\,\big(M_{i+1} + M_{i-1}\big), \tag{128}$$

where we have used (121).

Solving the System. Altogether, we have:

$$\begin{cases} \text{(First moments)} \\ m_0 = 1 + \frac{2}{3}\,m_1, \\ m_1 = \frac{3}{2} + \frac{1}{2}\,m_2, \\ m_i = \dfrac{3 + m_{i+1} + m_{i-1}}{2}, \quad \text{for } 2 \le i \le \lceil \frac{V}{2} \rceil, \end{cases}$$

$$\begin{cases} \text{(Second moments)} \\ M_0 = 1 + \frac{4}{3}\,m_1 + \frac{2}{3}\,M_1, \\ M_1 = 3m_1 - \frac{3}{2} + \frac{1}{2}\,M_2 \\ M_i = 3m_i - \frac{3}{2} \;+\; \frac{1}{2}\,\big(M_{i+1} + M_{i-1}\big), \quad \text{for } 2 \le i \le \lceil \frac{V}{2} \rceil. \end{cases}$$

One can solve this $2\lceil \frac{V}{2} \rceil$-dimensional linear system to find $M_0 = \mathbb{E}[h^2]$.

Here, we assume that V is even (a similar approach can be used to derive the result for V being odd).

First, we solve for m_i, $0 \le i \le \frac{V}{2}$, starting from $i = \frac{V}{2}$ and using $m_{\frac{V}{2}-1} = m_{\frac{V}{2}+1}$, we get

$$m_{\frac{V}{2}} = \frac{3}{2} + m_{\frac{V}{2}-1}. \tag{129}$$

Putting it in the equation for $i = \frac{V}{2} - 1$, we obtain

$$m_{\frac{V}{2}-1} = \frac{3 + m_{\frac{V}{2}} + m_{\frac{V}{2}-2}}{2} \tag{130}$$

$$= \frac{3 + \frac{3}{2} + m_{\frac{V}{2}-1} + m_{\frac{V}{2}-2}}{2}. \tag{131}$$

By rearranging the terms, we derive

$$m_{\frac{V}{2}-1} = 3 + \frac{3}{2} + m_{\frac{V}{2}-2}. \tag{132}$$

By doing this, we observe the general relationship of

$$m_{\frac{V}{2}-i} = 3i + \frac{3}{2} + m_{\frac{V}{2}-i-1}, \tag{133}$$

where $0 \leq i \leq \frac{V}{2} - 2$. Putting $i = \frac{V}{2} - 2$, gives us

$$m_2 = \frac{3V}{2} - \frac{9}{2} + m_1. \tag{134}$$

So, we will reach to the following equations

$$\begin{cases} m_0 = 1 + \frac{2}{3}\, m_1, \\ m_1 = \frac{3}{2} + \frac{1}{2}\, m_2, \\ m_2 = \frac{3V}{2} - \frac{9}{2} + m_1, \end{cases}$$

which provides us with $m_0 = V, m_1 = \frac{3V}{2} + \frac{3}{2}$. Using (133) iteratively we get

$$\begin{aligned} m_{\frac{V}{2}-i} &= 3i + \frac{3}{2} + m_{\frac{V}{2}-i-1} && (135)\\ &= 3i + \frac{3}{2} + 3(i-1) + \frac{3}{2} + m_{\frac{V}{2}-i-2} && (136)\\ &= 3\left(i + (i-1) + \cdots + (\frac{V}{2} - 2)\right) + \frac{3}{2}(\frac{V}{2} - i) + m_1 && (137)\\ &= 3\frac{(\frac{V}{2} - 2 - i)(\frac{V}{2} - 2 + i)}{2} + \frac{3}{2}(\frac{V}{2} - i) + \frac{3V}{2} + \frac{3}{2} && (138)\\ &= \mathcal{O}(V^2). && (139) \end{aligned}$$

Now, we repeat the same approach for the second moment variables. starting from $i = \frac{V}{2}$ and using $M_{\frac{V}{2}-1} = M_{\frac{V}{2}+1}$ based on symmetry, we get

$$M_{\frac{V}{2}} = 3m_{\frac{V}{2}} - \frac{3}{2} + M_{\frac{V}{2}-1}. \tag{140}$$

Putting it in the equation for $i = \frac{V}{2} - 1$, we obtain

$$\begin{aligned} M_{\frac{V}{2}-1} &= 3m_{\frac{V}{2}-1} - \frac{3}{2} + \frac{1}{2}(M_{\frac{V}{2}} + M_{\frac{V}{2}-2}) && (141)\\ &= 3m_{\frac{V}{2}-1} - \frac{3}{2} + \frac{1}{2}(3m_{\frac{V}{2}} - \frac{3}{2} + M_{\frac{V}{2}-1} + M_{\frac{V}{2}-2}). && (142) \end{aligned}$$

By rearranging the terms, we derive

$$M_{\frac{V}{2}-1} = 6m_{\frac{V}{2}-1} + 3m_{\frac{V}{2}} - 3 - \frac{3}{2} + M_{\frac{V}{2}-2}. \tag{143}$$

By keep doing this, we observe the general relationship of

$$M_{\frac{V}{2}-i} = 6\left(m_{\frac{V}{2}-i} + \cdots + m_{\frac{V}{2}-1}\right) + 3m_{\frac{V}{2}} - 3i - \frac{3}{2} + M_{\frac{V}{2}-i-1}, \tag{144}$$

where $0 \leq i \leq \frac{V}{2} - 2$. Putting $i = \frac{V}{2} - 2$, gives us

$$M_2 = 6\left(\sum_{i=2}^{\frac{V}{2}-1} m_i\right) + 3m_{\frac{V}{2}} - 3(\frac{V}{2} - 2) - \frac{3}{2} + M_1. \tag{145}$$

Applying (145) in (125) provides

$$M_1 = 6\left(\sum_{i=1}^{\frac{V}{2}-1} m_i\right) + 3m_{\frac{V}{2}} - 3(\frac{V}{2} - 1) - \frac{3}{2}. \tag{146}$$

If we use this in (122) we obtain

$$H^2 \;=\; \mathbb{E}[h^2] = M_0 = 1 + \frac{4}{3}\, m_1 + \frac{2}{3}\, M_1 = \mathcal{O}(V^3), \tag{147}$$

this is due to the fact that we derived $m_i = \mathcal{O}(V^2)$ earlier.

Table 3: Comparison of the convergence rate and communication overhead in **iid** setting for Metropolis-Hastings $\mathbf{P}$.

TOPOLOGY	ALGORITHM	CONVERGENCE RATE	COMM-COST
CYCLE ($p = \Theta(\frac{1}{V^2})$)	MULTI-WALK	$\mathcal{O}\left(\frac{\sigma}{\sqrt{T}} + (\frac{R\sqrt{V\sigma^2}}{T})^{\frac{2}{3}}\right)$✓	$\Theta(T)$
	ASYNCHRONOUS GOSSIP	$\mathcal{O}\left(\frac{\sigma}{\sqrt{T}} + (\frac{V\sqrt{V^2\sigma^2}}{T})^{\frac{2}{3}}\right)$	$\Theta(VT)$
2D-TORUS ($p = \Theta(\frac{1}{V})$)	MULTI-WALK	$\mathcal{O}\left(\frac{\sigma}{\sqrt{T}} + (\frac{R\sqrt{V\sigma^2}}{T})^{\frac{2}{3}}\right)$✓	$\Theta(T)$
	ASYNCHRONOUS GOSSIP	$\mathcal{O}\left(\frac{\sigma}{\sqrt{T}} + (\frac{V\sqrt{V\sigma^2}}{T})^{\frac{2}{3}}\right)$	$\Theta(VT)$
COMPLETE ($p = 1$)	MULTI-WALK	$\mathcal{O}\left(\frac{\sigma}{\sqrt{T}} + (\frac{R\sqrt{V\sigma^2}}{T})^{\frac{2}{3}}\right)$[✓*if* $R = \mathcal{O}(\sqrt{V})$]	$\Theta(T)$
	ASYNCHRONOUS GOSSIP	$\mathcal{O}\left(\frac{\sigma}{\sqrt{T}} + (\frac{V\sqrt{\sigma^2}}{T})^{\frac{2}{3}}\right)$[✓*if* $R = \Omega(\sqrt{V})$]	$\Theta(V^2T)$

F. Wall Clock Time Convergence

In Algorithm 1, assume each walk performs one iteration (computation and communication) with a rate-$\frac{1}{d}$ exponential random variable, independent across walks and over time. The value of d is determined by the average computation and communication delay in the network. Thus, each walk does one iteration in Algorithm 1 according to a rate-$\frac{1}{d}$ Poisson process. Equivalently, this corresponds to all iterations in Algorithm 1 are according to a rate-$\frac{R}{d}$ Poisson process at times $\{Z_t\}_{t=0}^{T-1}$ where $\{Z_t - Z_{t-1}\}_{t=1}^{T-1}$, denoting the t-th iteration duration, are i.i.d. exponentials of rate $\frac{R}{d}$. Therefore, we have $\mathbb{E}[Z_t] = \frac{td}{R}$ and for any $\delta > 0$:

$$Pr\left(|Z_t - \frac{td}{R}| \geq \frac{\delta td}{R}\right) \leq 2\exp\left(\frac{-\delta^2 t}{2}\right). \tag{148}$$

This follows directly from Cramer's theorem (Boyd et al., 2006). Hence, by multiplying the terms obtained regarding iterations by $\frac{d}{R}$, we obtain the corresponding terms in real time. In other words, the convergence rate in Theorem 4.1 can be transformed to real time (Z) by substituting T with $\frac{RZ}{d}$. For Algorithm 2, we assume each node has a clock that ticks at the times of a rate-$\frac{1}{d}$ Poisson process. Here, the value of d is determined by the average computation and gossip communication delay for nodes. And the same result is valid by replacing R with V.

Corollary F.1. *Under the condition of Theorem 4.1, 4.2, we get the convergence rate of Algorithms 1 and 2 as shown in Table 2 where Z represents wall-clock time.*

G. Extended Theoretical Insights

Dominant terms. The dominant term in both (1) and (2) is identically given by $\sqrt{\frac{FL(\sigma^2+\zeta^2)}{T}}$. Focusing on the next most significant term for comparison, in (1), this term is given by $(\frac{FLR\sqrt{V\sigma^2+H^2\zeta^2}}{T})^{\frac{2}{3}}$, whereas in (2), it is $(\frac{FLV\sqrt{\frac{\sigma^2}{p}+\frac{\zeta^2}{p^2}}}{T})^{\frac{2}{3}}$. Note that (1) includes a non-dominating term that describes the rate at which walks converge to their steady state. This term is related to the spectral gap of $\mathbf{P}$, represented by p'. In the following, we compare the dominant terms in the convergence rates of both algorithms in different settings.

Homogeneous data distribution. In iid setting ($\zeta = 0$), the differentiating factor in the second dominant term of

Table 4: Comparison of the convergence rate and communication overhead in **noniid** setting for Metropolis-Hastings $\mathbf{P}$.

TOPOLOGY	ALGORITHM	CONVERGENCE RATE	COMM-COST
CYCLE ($p = \Theta(\frac{1}{V^2})$)	MULTI-WALK	$\mathcal{O}\left(\sqrt{\frac{\sigma^2+\zeta^2}{T}} + (\frac{R\sqrt{V\sigma^2+V^3\zeta^2}}{T})^{\frac{2}{3}}\right)$ ✓	$\Theta(T)$
	ASYNCHRONOUS GOSSIP	$\mathcal{O}\left(\sqrt{\frac{\sigma^2+\zeta^2}{T}} + (\frac{V\sqrt{V^2\sigma^2+V^4\zeta^2}}{T})^{\frac{2}{3}}\right)$	$\Theta(VT)$
2D-TORUS ($p = \Theta(\frac{1}{V})$)	MULTI-WALK	$\mathcal{O}\left(\sqrt{\frac{\sigma^2+\zeta^2}{T}} + (\frac{R\sqrt{V\sigma^2+H^2\zeta^2}}{T})^{\frac{2}{3}}\right)$	$\Theta(T)$
	ASYNCHRONOUS GOSSIP	$\mathcal{O}\left(\sqrt{\frac{\sigma^2+\zeta^2}{T}} + (\frac{V\sqrt{V\sigma^2+V^2\zeta^2}}{T})^{\frac{2}{3}}\right)$	$\Theta(VT)$
COMPLETE ($p = 1$)	MULTI-WALK	$\mathcal{O}\left(\sqrt{\frac{\sigma^2+\zeta^2}{T}} + (\frac{R\sqrt{V\sigma^2+V^2\zeta^2}}{T})^{\frac{2}{3}}\right)$	$\Theta(T)$
	ASYNCHRONOUS GOSSIP	$\mathcal{O}\left(\sqrt{\frac{\sigma^2+\zeta^2}{T}} + (\frac{V\sqrt{\sigma^2+\zeta^2}}{T})^{\frac{2}{3}}\right)$	$\Theta(V^2T)$

convergence rate is $\frac{V}{\sqrt{p}}$ for Asynchronous Gossip and $R\sqrt{V}$ for Multi-Walk. Specifically, for graphs with $p = \mathcal{O}(\frac{V}{R^2})$, Multi-Walk outperforms, while for $p = \Omega(\frac{V}{R^2})$, Asynchronous Gossip converges faster w.r.t iterations. It is interesting to observe that the graph's topology does not impact the performance of Multi-Walk in iid setting, and the only factors are the number of nodes and walks. We compare convergence rate and communication overhead for each algorithm in Table 3 across three different graph topologies, using the commonly employed Metropolis-Hastings matrix, $\mathbf{P}$, where $p_{ij} = p_{ji} = \min\left\{\frac{1}{\deg(i)+1}, \frac{1}{\deg(j)+1}\right\}, \quad \text{for } \{i,j\} \in \mathcal{E}$. Note that computation overhead is the same for both and equal to the number of iterations, *i.e.,* T, and we do not include that in the table. We observe that for both cycle and 2D-torus topologies, Multi-Walk outperforms Asynchronous Gossip in convergence rate. However, when the graph diameter decreases (i.e., p increases), such as in the case of a complete graph, Multi-Walk loses its advantage. It is important to note that Multi-Walk consistently maintains lower communication overhead; in each iteration, it involves at most one communication step, whereas Asynchronous Gossip activates multiple edges for mixing based on the graph topology

Heterogeneous data distribution. In non-iid setting, ζ^2 is multiplied by H^2 for Multi-Walk and by p^2 for Asynchronous Gossip. We derived H^2 for cycle and complete topologies with Metropolis-Hastings transition matrix in Appendix E, and the comparison is summarized in Table 4. We observe that for the cycle topology, Multi-Walk consistently demonstrates faster convergence in terms of iterations. However, this advantage diminishes as we move to topologies with smaller diameters. In complete topology, we observe that ζ^2 is multiplied by V^2 in Multi-Walk, whereas it is multiplied by 1 in Asynchronous Gossip. This indicates that, as we transition to increasingly non-iid settings in small-diameter topologies, Multi-Walk perform poorly.

H. Detailed Experimental Results

In this section, we validate our theoretical results through empirical experiments, which include the following: Section H.1 verifies the impact of network graph topology on the convergence rate. Section H.2 explores the impact of data heterogeneity on the convergence rate in two different graph topologies with small and large diameters. Section 6.3 evaluates the communication efficiency of Multi-Walk in bandwidth-constrained environments through an LLM fine-tuning task. Finally, Section 6.4 investigates the impact of Node 0 failure on MW to verify its resilience. We also compare this with Asynchronous Gossip, observing its performance when the same sequence of nodes fails.

We use two machine learning tasks: (i) *Image classification* on CIFAR-10 (Krizhevsky, 2009) using ResNet-20 (He et al.,

2015); and (ii) *LLM fine-tuning* of OPT-125M (Zhang et al., 2022) as a large language model on the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018). The details of the image classification and LLM fine-tuning tasks are specified in Table 5 and 6, respectively.

We repeat each experiment 10 times and present the error bars associated with the randomness of the optimization. In every figure, we include the average and standard deviation error bars. In the figures, we use "MW" as an abbreviation for MW.

Table 5: Default experimental settings for the image classification training

Dataset	CIFAR-10 (Krizhevsky, 2009), licensed under the MIT License
Architecture	ResNet-20 (He et al., 2015), licensed under the MIT License
Loss function	cross entropy
Accuracy objective	top-1 accuracy
Number of nodes	20
Topology	cycle, complete, Erdős–Rényi
Data distribution	iid (shuffled and split), non-iid (based on labels)
Local Steps τ	5
Optimizer	SGD with momentum
Batch size	32 per client
Momentum	0.9 (Nesterov)
Initial learning rate	0.05
Learning rate schedule	multiplied by 0.1 once after 75 and once after 90 percent of the training
Training time	15 minutes for $\alpha \in \{10, 1\}$ and 30 minutes for $\alpha = 0.1$
Weight decay	10^{-4}
Learning rate warm-up time	2 minutes
Repetitions	10
Reported metric	Mean and standard deviation (1-sigma error bars) of the aggregated model's training loss and accuracy, accounting for randomness in network conditions and algorithmic factors such as random walk-based next node selection and neighbor selection in gossip-based averaging.

We have conducted the experiments on the National Resource Platform (NRP) (NRP) cluster. Figure 6 provides a schematic representation of the network topology and node distribution used in our experiments. The setup consists of 20 nodes grouped into 5 geographic clusters labeled CA, NV, IA, IL, and KS, corresponding to the US states of California, Nevada, Iowa, Illinois, and Kansas, respectively. Within each cluster, nodes are connected locally, while additional links enable communication across clusters, implementing decentralized computation and communication patterns. As the NRP dynamically assigns resources for each run, the exact node distribution may differ slightly from what is shown in Figure 6. This variability in node assignment is one of the sources of randomness in network conditions, which we account for in our results by reporting error bars. All nodes are provisioned with 1 GPU each, along with 10 CPU cores and 80 GiB of memory. Additionally, each node mounts a 13 GiB in-memory volume for high-performance shared memory usage. The GPU type (e.g., A100, V100, etc.) is determined based on node and cluster assignments made by the National Resource Platform (NRP), which matches resource requests to suitable hardware across participating sites. This assignment introduces an element of randomness into our experiments, as the exact GPU model may vary between runs depending on resource availability. Most nodes are connected via high-speed research networks such as Science DMZs, with interconnect speeds ranging from 10G to 100G. This setup reflects a realistic decentralized learning environment over a wide-area network and introduces practical considerations like heterogeneous latency and bandwidth, which are difficult to model in simulation.

We use the Dirichlet distribution to create disjoint non-iid nodes (Lin et al., 2021). The degree of data heterogeneity is controlled by the distribution parameter α; the smaller α is, the more likely the nodes hold examples from only one class. Throughout the experiments, we use three levels of α; 10, 1, and 0.1.

In Figure 7, we include the effect of different values of α in creating disjoint noniid data from CIFAR-10 across nodes using

Table 6: Default experimental settings for the large language model fine-tuning

Dataset	Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018), released under the CC BY-SA 4.0 License
Architecture	OPT-125M (Zhang et al., 2022), released by Meta AI under the OPT License (a custom non-commercial research license)
Loss function	cross entropy
Number of nodes	20
Topology	Erdős–Rényi
Data distribution	iid (shuffled and split), non-iid (based on genre)
Local Steps τ	1
Optimizer	Adam
Batch size	16 sentences per client
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	10^{-8}
Initial learning rate	10^{-4}
Learning rate schedule	multiplied by 0.1 once after 75 and once after 90 percent of the training
Training time	15 minutes
Weight decay	10^{-4}
Learning rate warm-up time	2 minutes
Repetitions	10
Reported metric	Mean and standard deviation (1-sigma error bars) of the aggregated model's training loss and accuracy, accounting for randomness in network conditions and algorithmic factors such as random walk-based next node selection and neighbor selection in gossip-based averaging.

the Dirichlet distribution. We observe that as α decreases, the probability of each node containing data from only one class increases.

H.1. Graph topology

Figure 8 presents the training loss (left column) and test accuracy (right column) of the image classification task in a graph of 20 nodes. We consider three topologies of cycle, complete, and Erdős–Rényi with connection probability of each pair of nodes being 0.3. The noniid-ness level for this experiment is set to $\alpha = 1$. We observe in Figure 8a that the convergence rate w.r.t iterations in cycle topology is faster for MW, regardless of the number of walks (R). We also observe that as we decrease R, the convergence rate of MW w.r.t iterations improves. These are consistent with the theoretical results derived in section 4 and shown in Table 3 and 4. In the small-diameter topology shown in Figure 8b (a complete graph), we observe that MW is no longer superior across all numbers of walks; specifically, Asynchronous Gossip outperforms MW when 15 walks are used. This observation is consistent with the theoretical results indicating that in small-diameter graphs, there exists a specific threshold for the number of walks: below this threshold, Multi-Walk outperforms Asynchronous Gossip, whereas above it, performance degrades. Figure 8c presents the results for an Erdős–Rényi topology with the connection probability of 0.3. This topology, where each node is connected to every other node with a probability of 0.3, is a well-connected graph with a small diameter. We observe that the Erdős–Rényi graph results are quite similar to the complete graph.

H.2. Data heterogeneity (Noniid-ness)

In this section, we present experiments to evaluate the impact of data heterogeneity on convergence behavior in small and large diameter graphs. We provide comparisons across three domains: iterations, wall-clock time, and transmitted bits.

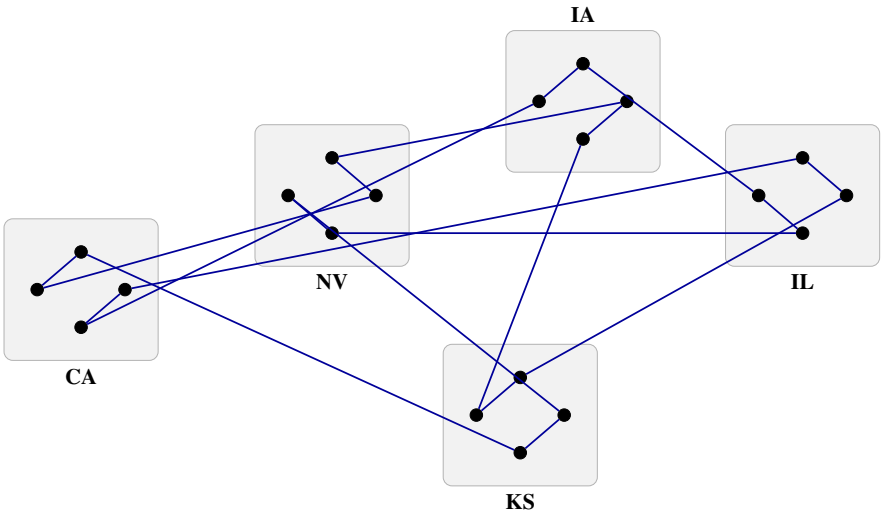
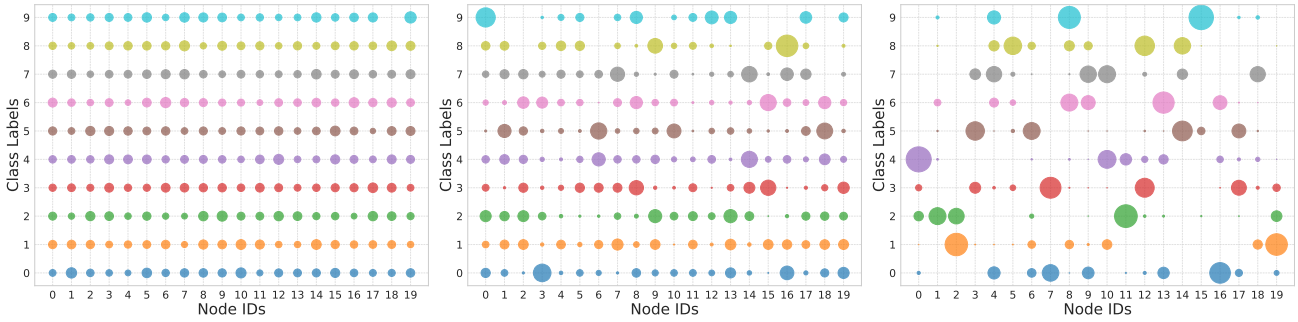


Figure 6: A 20-node decentralized system distributed across five geographic clusters (CA, NV, IA, IL, KS). The links depict the overlay network, which is configurable based on the desired graph topology.



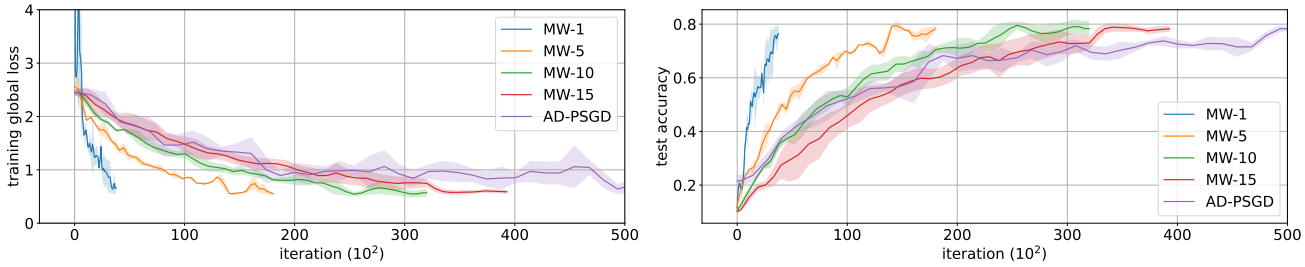
(a) $\alpha = 10$. (b) $\alpha = 1$. (c) $\alpha = 0.1$.

Figure 7: Data heterogeneity visualization for CIFAR-10 across a 20-node network using Dirichlet distributions with varying parameter α.

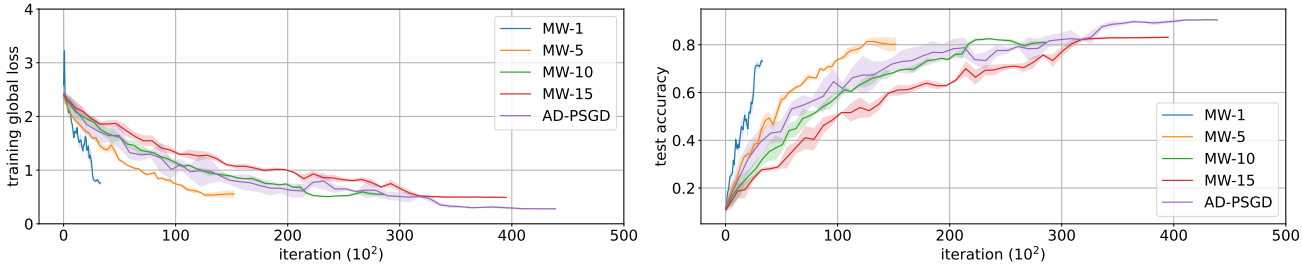
H.2.1. Convergence w.r.t iterations

Figure 9 illustrates the convergence behavior of two different topologies over iterations under varying levels of data heterogeneity. The left column (subfigures d, c, e) corresponds to a 20-node Erdős–Rényi topology ($p = 0.3$), while the right column (subfigures b, d, f) depicts a 20-node cycle topology. We observed in section H.1 that Erdős–Rényi (0.3) has a quite small diameter. In the first row (subfigures a, b), where $\alpha = 10$ and the data distribution is nearly iid, MW outperforms Asynchronous Gossip in terms of iterations across both graph topologies. We further observe that in this iid scenario, the impact of graph topology is minimal compared to settings with higher data heterogeneity (shown in the second and third rows). Decreasing α to 1 introduces a more noniid scenario, causing the performance of both MW and Asynchronous Gossip to degrade; however, even at this level of heterogeneity, MW continues to outperform in both topologies. We can go further and reduce α to 0.1 to get extreme non-iid data distribution (third row). Consistent with our theoretical results, MW is no longer superior in small-diameter graphs under extreme noniid scenarios, as verified in Figure 9e. Conversely, in the cycle topology (characterized by a large diameter), MW remains faster in terms of iterations.

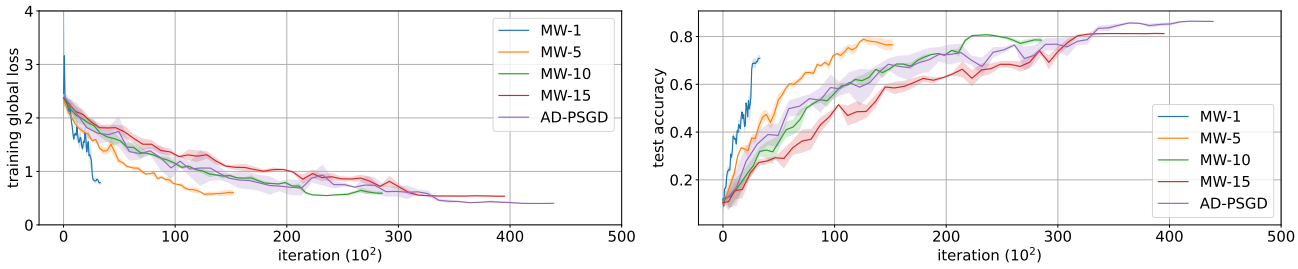
This result is expected based on the theoretical bounds presented in Table 4. In the convergence rate of MW, the heterogeneity term ζ^2 is scaled by H^2, whereas in Asynchronous Gossip, it is scaled by $1/p^2$. In small-diameter graphs (e.g., complete graph), we have $H^2 = \mathcal{O}(V^2)$ and $p = 1$, meaning the impact of noniid data on MW ($\mathcal{O}(V^2)$) is far more severe than on Asynchronous Gossip ($\mathcal{O}(1)$). This confirms the degradation observed in our experiments. On the other hand, in large-diameter graphs like the cycle topology, we have $H^2 = \mathcal{O}(V^3)$ and $p = \Theta(1/V^2)$ (implying $1/p^2 = \mathcal{O}(V^4)$). Consequently,



(a) Cycle graph.



(b) Complete graph.



(c) Erdős–Rényi (0.3) graph.

Figure 8: Comparison across different network topologies for a 20-node graph: Training loss (left column) and test accuracy (right column) of ResNet-20 on CIFAR-10.

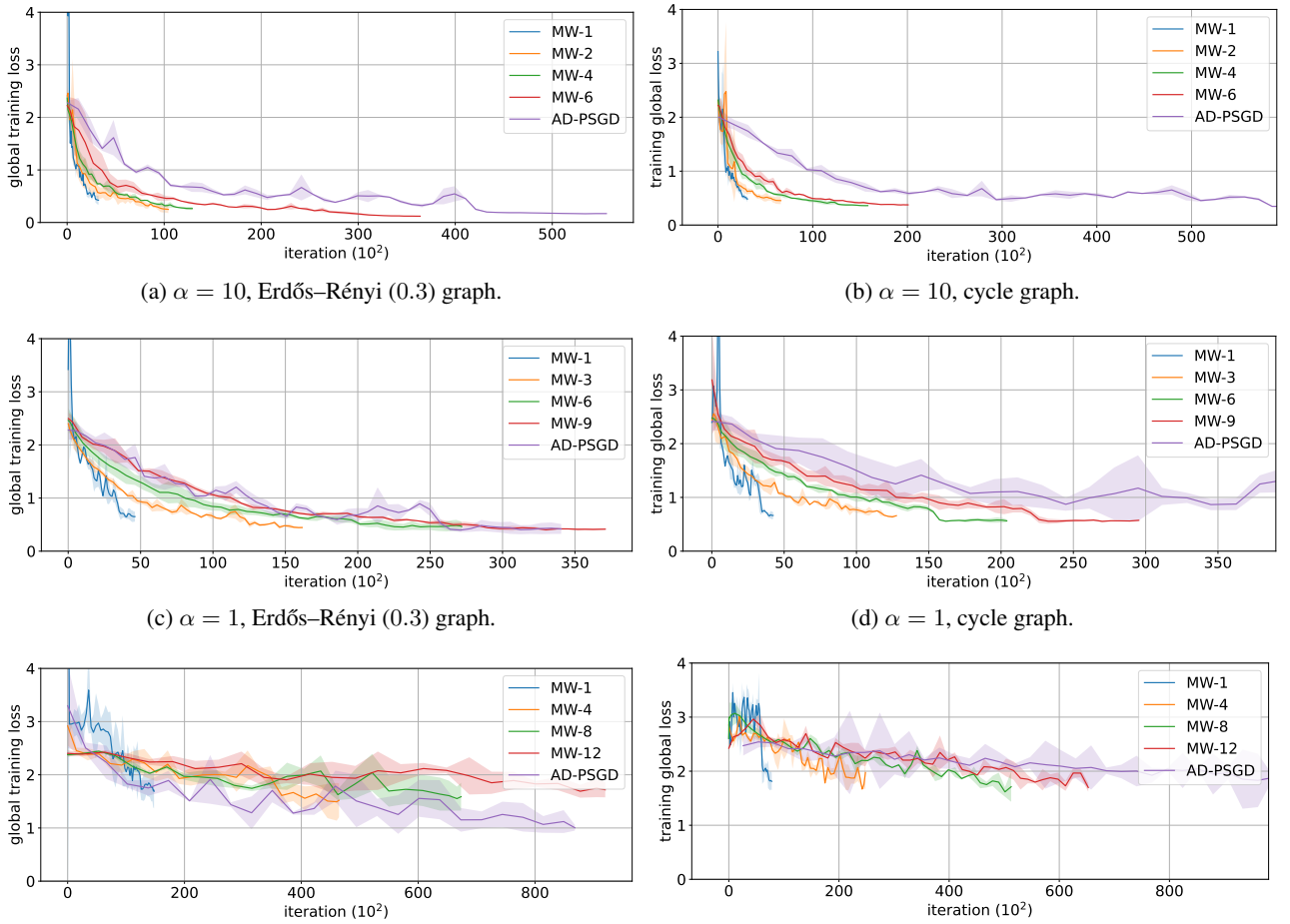
the impact of heterogeneity scales better for MW ($\mathcal{O}(V^3)$) compared to Asynchronous Gossip ($\mathcal{O}(V^4)$), explaining why MW retains its advantage in this setting.

H.2.2. Convergence w.r.t wall-clock time

Figure 10 shows convergence versus wall-clock time. We know that in the time domain, Asynchronous Gossip achieves a linear speed-up with the number of nodes compared to MW with a single walk. However, MW can improve its time-domain performance by increasing the number of walks, yielding a linear speed-up with respect to the walk count. In the first and second rows, we observe that increasing the number of walks to 2 or 3 is sufficient to catch up with Asynchronous Gossip. Conversely, as we increase heterogeneity to $\alpha = 0.1$, we observe that in the Erdős–Rényi (0.3) topology (which has a small diameter), the gap between Asynchronous Gossip and MW cannot be bridged even by increasing the number of walks. This reinforces the fact that Asynchronous Gossip performs better in small-diameter graphs under extreme heterogeneity.

H.2.3. Convergence w.r.t transmitted bits

Figure 11 shows convergence versus transmitted bits. In terms of communication overhead, we observe that in settings that are not extremely non-iid, where the second dominating term is negligible (due to small ζ), MW outperforms Asynchronous Gossip as predicted by the results in Table 1. This can be seen in the first ans second row in Figure 11. However, in the extreme non-iid setting of the third row, the value of ζ becomes too large that the second dominating term in Table 1 comes into play. In this term, the impact of noniid-ness in a graph topology with a small diameter (Erdős–Rényi (0.3)) significantly



(a) $\alpha = 10$, Erdős–Rényi (0.3) graph.

(b) $\alpha = 10$, cycle graph.

(c) $\alpha = 1$, Erdős–Rényi (0.3) graph.

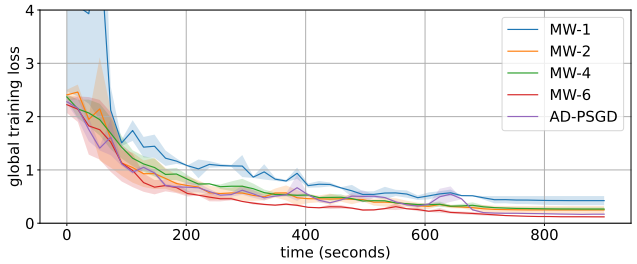
(d) $\alpha = 1$, cycle graph.

(e) $\alpha = 0.1$, Erdős–Rényi (0.3) graph.

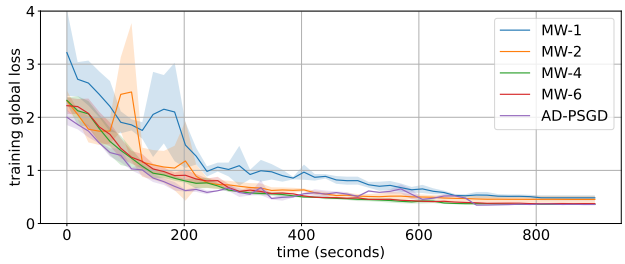
(f) $\alpha = 0.1$, cycle graph.

Figure 9: Comparison across different noniid-ness levels **w.r.t iterations** for a 20-node network with Erdős–Rényi (0.3) (left column) and cycle (right column) topology: Training loss for ResNet-20 on CIFAR-10.

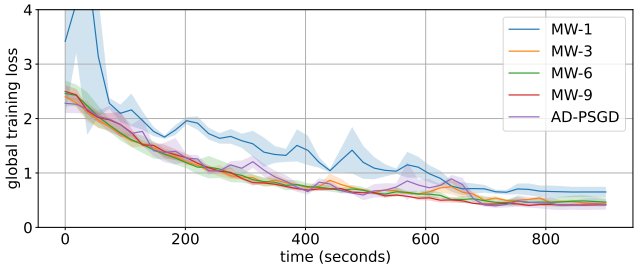
disfavors MW, which is evident from the observed results in Figure 11e. In Figure 11f, we again observe that, in contrast to small-diameter topologies, MW continues to outperform even under extreme noniid conditions. This is again predicted based on the theoretical results in section 4. Intuitively, in small-diameter graphs where connectivity is dense, Gossip algorithms propagate information across the network more efficiently than Random Walks. This rapid information spread is critical under extreme heterogeneity (noniid settings), as nodes rely on global information to maintain a trajectory toward the global minimum and avoid getting trapped in local optima.



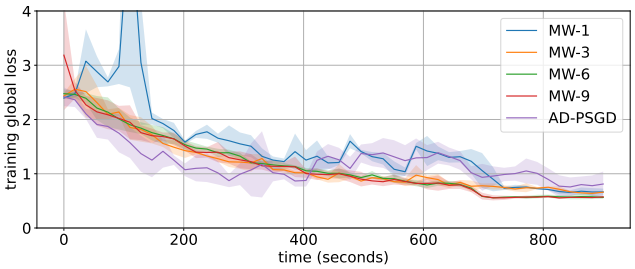
(a) $\alpha = 10$, Erdős–Rényi (0.3) graph.



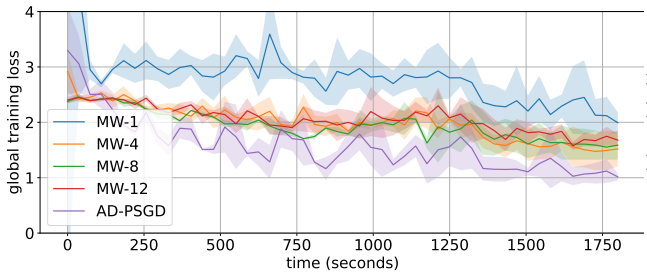
(b) $\alpha = 10$, cycle graph.



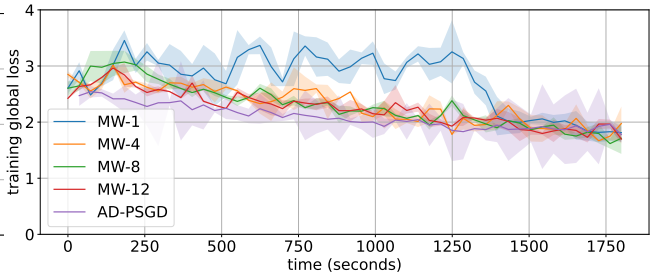
(c) $\alpha = 1$, Erdős–Rényi (0.3) graph.



(d) $\alpha = 1$, cycle graph.

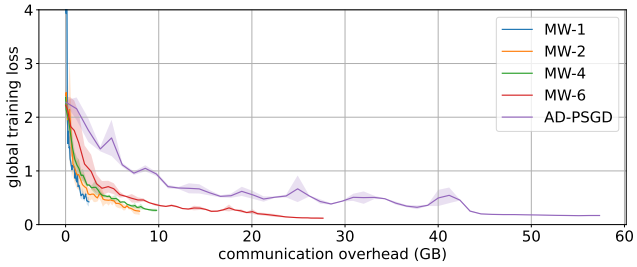


(e) $\alpha = 0.1$, Erdős–Rényi (0.3) graph.

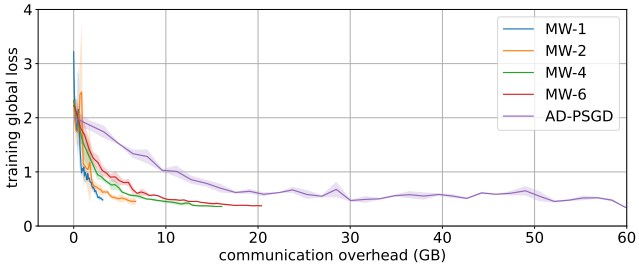


(f) $\alpha = 0.1$, cycle graph.

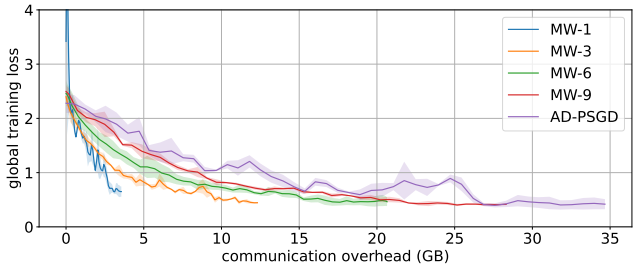
Figure 10: Comparison across different noniid-ness levels **w.r.t wall-clock time** for a 20-node network with Erdős–Rényi (0.3) (left column) and cycle (right column) topology: Training loss for ResNet-20 on CIFAR-10.



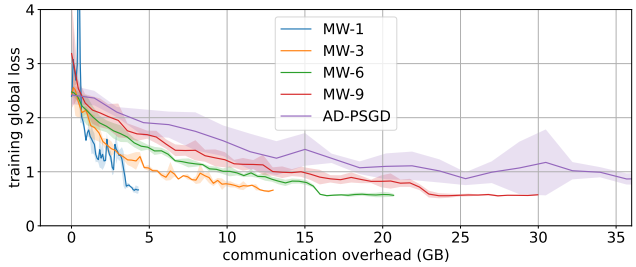
(a) $\alpha = 10$, Erdős–Rényi (0.3) graph.



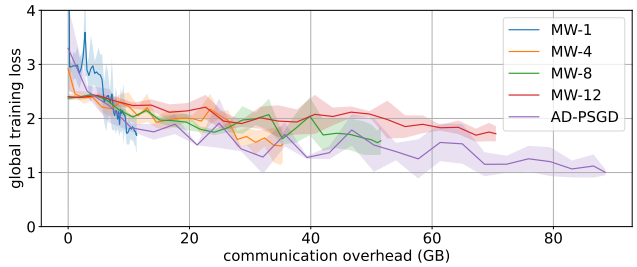
(b) $\alpha = 10$, cycle graph



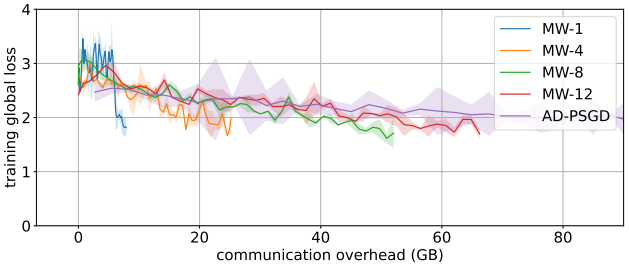
(c) $\alpha = 1$, Erdős–Rényi (0.3) graph.



(d) $\alpha = 1$, cycle graph.



(e) $\alpha = 0.1$, Erdős–Rényi (0.3) graph.



(f) $\alpha = 0.1$, cycle graph.

Figure 11: Comparison across different noniid-ness levels **w.r.t transmitted bits** for a 20-node network with Erdős–Rényi (0.3) (left column) and cycle (right column) topology: Training loss for ResNet-20 on CIFAR-10.