

Communication-Efficient Decentralized Learning: Topology Design and Optimization Methods



Yuki Takezawa

Graduate School of Informatics
Kyoto University

A thesis submitted for the degree of
Doctor of Philosophy

Abstract

The growing scale of modern machine learning has intensified the need for training paradigms that can efficiently utilize distributed data and computational resources. Among these paradigms, decentralized learning has attracted significant attention. Despite its inherent communication sparsity, decentralized learning still faces fundamental challenges: the exchange of high-dimensional model parameters often dominates the runtime, and convergence can significantly slow down as the number of nodes increases or when local datasets differ greatly across nodes. Thus, improving the communication efficiency—both in terms of per-round communication and the number of rounds required for convergence—remains a central problem in decentralized optimization. This thesis develops new communication topologies and optimization algorithms that advance the scalability and efficiency of decentralized learning.

First, we introduce the BASE- $(k + 1)$ GRAPH, a novel topology that achieves efficient parameter mixing while preventing communication costs from growing with the number of nodes. We further propose TELEPORTATION, a decentralized learning method that entirely removes the communication overhead induced by the large number of nodes and yields convergence rates that improve as the number of nodes increases.

Next, we address inefficiencies arising from data heterogeneity, a defining characteristic of decentralized systems. While classical methods such as Gradient Tracking guarantee convergence regardless of data heterogeneity, they fail to exploit the similarity among the datasets held by nodes. To remedy this, we propose Accelerated Stabilized Proximal Decentralized Optimization (ACCELERATED SPDO), an optimization method that benefits from inter-node dataset similarity. ACCELERATED SPDO reduces communication costs when datasets are similar and achieves communication complexity that is optimal up to logarithmic factors.

Overall, this thesis contributes new theoretical insights and practical methodologies that significantly enhance the communication efficiency of decentralized learning.

Acknowledgements

I would first like to express my sincere gratitude to Makoto Yamada. He was my supervisor during my Bachelor's and Master's programs, and I learned the fundamentals of conducting research from him. During this time, he taught me the excitement of research and the value of discussing research ideas with others. Without this experience, I would not have pursued a Ph.D. program. Even after I entered the Ph.D. program at Kyoto University and he moved to OIST, he frequently invited me to Okinawa, helped me expand my professional network, and collaborated with me on many research projects.

I am also grateful to my supervisor, Hisashi Kashima, and the dissertation committee members: Nobuo Yamashita and Kazuyuki Shudo. Their feedback was very useful and constructive in improving this thesis.

I am deeply grateful to Ryoma Sato, Han Bao, and Kenta Niwa. Ryoma was a Ph.D. student in the same lab and taught me many things as a senior student, including how to explore research topics and how to productively publish papers. Han was an assistant professor in the same lab. We shared similar research interests in the theoretical aspects of machine learning and had many discussions on various research projects. I was fortunate to have him in the same office, which gave me the opportunity to discuss research ideas with him in person. Kenta is a researcher at NTT Communication Science Laboratories and was a mentor when I joined the internship program at NTT. Through this internship, I was introduced to decentralized learning, and this experience ultimately led me to engage in decentralized learning during my Ph.D. program. After the internship, we continued our weekly discussions, and I am very grateful for his support.

During my Ph.D. program, I spent a total of six months in Sebastian Stich's lab at CISA Helmholtz Center for Information Security. It was my first time staying abroad for a long time, and the experience was truly exciting. He kindly spent a large amount of time on research discussions with me, and I got a great opportunity to discuss research topics in detail with many optimization experts, including him and his lab members. I am grateful to him for providing such an opportunity. I would also like to thank the lab members. Thanks to them, my life in St. Ingbert was full of joy.

I also thank all the people with whom I worked and from whom I learned a lot, in particular Eduard Gorbunov, Anastasia Koloskova, Xiaowen Jiang, and Kyohei

Atarashi.

Finally, I would like to thank my parents for their constant support throughout my Ph.D. program. I would also like to thank my friends—especially Naoki Matsumoto, Kotaro Matsuoka, and Naoyuki Okamoto. They are my fellow Ph.D. students at Kyoto University in the same cohort. I feel very fortunate to have shared my Ph.D. journey with them, as our friendship and daily conversations made this experience far more enjoyable.

Publications

This thesis is based on the following papers:

- Yuki Takezawa, Ryoma Sato, Han Bao, Kenta Niwa, Makoto Yamada,
“Beyond Exponential Graph: Communication-efficient Topologies for Decen-
tralized Learning via Finite-time Convergence”,
In Advances in Neural Information Processing Systems 2023
- Yuki Takezawa, Sebastian Stich,
“Scalable Decentralized Learning with Teleportation”,
In International Conference on Learning Representations 2025
- Yuki Takezawa, Xiaowen Jiang, Anton Rodomanov, Sebastian Stich,
“Exploiting Similarity for Computation and Communication-efficient Decentral-
ized Optimization”,
In International Conference on Machine Learning 2025

Contents

1	Introduction	1
1.1	Topology Design for Communication-efficient Decentralized Learning	2
1.2	Optimizer for Communication-efficient Decentralized Learning	3
1.3	Contributions beyond this Thesis	3
2	Setup and Assumptions	5
2.1	Notation	5
2.2	Problem Setting	5
2.3	Assumptions on Loss Functions	5
2.4	Assumptions on Network Topology	6
3	Communication-efficient Topologies for Decentralized Learning via Finite-time Convergence	7
3.1	Introduction	7
3.2	Related Work	9
3.3	Preliminary and Notation	10
3.4	Finite-time Sequence of Graphs	11
3.4.1	k -peer Hyper-hypercube Graph	11
3.4.2	Simple Base- $(k + 1)$ Graph	12
3.4.3	Base- $(k + 1)$ Graph	13
3.5	Decentralized SGD on Base- $(k + 1)$ Graph	16
3.6	Experiments	17
3.6.1	Consensus Rate	17
3.6.2	Decentralized Learning	18
3.7	Conclusion	20
4	Scalable Decentralized Learning with Teleportation	21
4.1	Introduction	21
4.2	Problem Setup	23
4.3	Proposed Method	24
4.3.1	TELEPORTATION	25
4.3.2	Convergence Analysis of TELEPORTATION	26
4.3.3	Efficient Hyperparameter Search for Number of Active Nodes	28

4.3.4	Convergence Analysis of TELEPORTATION with Efficient Hyperparameter Search	29
4.4	Related Work	30
4.5	Experiments	31
4.5.1	Synthetic Experiment	31
4.5.2	Neural Networks	32
4.5.3	Comparison under Heterogeneous Networks	33
4.6	Conclusion	34
5	Exploiting Similarity for Computation and Communication-efficient Decentralized Optimization	35
5.1	Introduction	35
5.2	Related Work	37
5.3	Problem Setup	39
5.4	Refined Analysis of PDO	40
5.4.1	Convergence Analysis	41
5.4.2	Discussion	42
5.5	Stabilized Proximal Decentralized Optimization Method	43
5.5.1	Convergence Analysis	44
5.6	Accelerated Stabilized Proximal Decentralized Optimization Method	45
5.6.1	Convergence Analysis	45
5.6.2	Discussion	46
5.6.3	Discussion on Federated Learning	48
5.7	Experiments	48
5.8	Conclusion	49
6	Conclusion and Future Directions	51
6.1	Future Directions	52
A	Useful Lemma	61
B	Appendix of Base-$(k + 1)$ Graph	63
B.1	Detailed Explanation of k -peer Hyperhypercube Graph	63
B.2	Simple Base- $(k + 1)$ Graph	65
B.3	Illustration of Topologies	66
B.3.1	Examples	66
B.3.2	Illustrative Comparison between Simple Base- $(k + 1)$ and Base- $(k + 1)$ Graphs	67
B.3.3	Additional Examples	68
B.3.4	1-peer Hypercube Graph and 1-peer Exponential Graph	72
B.4	Proof of Theorem 1	73
B.5	Convergence Rate of DSGD over Various Topologies	75
B.6	Additional Experiments	76
B.6.1	Comparison of Base- $(k + 1)$ and Simple Base- $(k + 1)$ Graphs	76

B.6.2	Consensus Rate	76
B.6.3	Decentralized Learning	77
B.7	Results with Other Neural Network Architecture	80
B.8	Hyperparameter Setting	80
C	Appendix of Teleportation	83
C.1	TELEPORTATION with Communication Overlap	83
C.2	Proof of Theorems 3 and 4	84
C.2.1	Notation	84
C.2.2	Main Proof	85
C.3	Proof of Theorem 5	94
C.4	Convergence Rate of Decentralized SGD with Various Topologies . .	96
C.5	Convergence Rate of Decentralized SGD with Client Sampling . . .	96
C.6	Additional Experiment	97
C.7	Experimental Setup	98
D	Appendix of Stabilized Proximal Decentralized Optimization	99
D.1	Comparison of Communication and Computational Complexities in Convex Case	99
D.2	Technical Lemmas	100
D.2.1	Useful Lemmas	100
D.3	Proof of Theorem 6	103
D.4	Proof of Theorem 8	110
D.5	Proof of Theorem 10	118
D.5.1	Useful Lemmas	118
D.5.2	Main Proof	119
D.5.3	Convex Case	129
D.5.4	Strongly Convex Case	134
D.6	Analysis of Computational Complexity	138
D.6.1	Useful Lemmas	138
D.6.2	Proof of Theorem 7	139
D.6.3	Proof of Theorem 9	139
D.6.4	Proof of Theorem 11	139
D.7	Experimental Setup	140
D.7.1	Setup for Adaptive Number of Local Steps	140
D.7.2	Hyperparameter Setting	140

1

Introduction

Machine learning has achieved remarkable results by leveraging vast amounts of data. Large-scale datasets such as C4 (Raffel et al., 2020) for natural language processing, ImageNet (Deng et al., 2009) for computer vision, and CLIP (Radford et al., 2021) have played a central role in this progress. The recent advances in machine learning have been driven by the ability to leverage large-scale datasets. However, there still remain vast amounts of data that have not yet been fully utilized. For instance, pictures and messages stored on personal smartphones, as well as large collections of clinical data maintained by medical institutions, are difficult to exploit due to concerns over privacy preservation. Distributed learning, such as federated learning and decentralized learning, offers a promising approach for leveraging such untapped data (Kairouz et al., 2021).

Distributed learning allows nodes to train machine learning models without sharing the training dataset with other nodes. The nodes correspond to personal smartphones or medical institutions and possess their own training datasets. During the training, nodes only exchange the model parameters with other nodes and do not share the training datasets. In this manner, distributed learning can inherently preserve the privacy of sensitive data while enabling model training by taking advantage of a substantially larger amount of data stored across different nodes.

Distributed learning can be broadly classified into federated learning and decentralized learning, based on the way communication is carried out. In federated learning, the presence of a central server is assumed, and nodes exchange model parameters via this central server. However, when the number of nodes becomes large, significant communication overhead arises between the central server and the nodes. By contrast, decentralized learning dispenses with a central server and instead assumes that nodes are interconnected through an underlying network topology, such as a ring. As each node communicates only with a limited number of neighboring nodes, decentralized learning achieves higher communication efficiency than federated learning. In this thesis, we focus on the study of decentralized learning.

Although decentralized learning is more communication-efficient than federated learning, communication efficiency is still important in decentralized learning.

Machine learning models typically contain a large number of parameters, and transmitting and receiving model parameters are generally more time-consuming than performing local computations at each node (Lian et al., 2017; Wang et al., 2019). Thus, reducing the communication costs is key to improving the overall efficiency of decentralized learning. The communication costs are determined by the communication that occurs in each round and the number of rounds required to reach the desired accuracy, i.e., convergence rate. While the sparse communication structure of decentralized learning reduces per-round communication costs, it can significantly degrade the convergence rate, especially when the number of nodes is large or when local datasets are highly heterogeneous (Koloskova et al., 2020). Motivated by these challenges, this thesis proposes novel communication topologies and optimization methods that improve the convergence rate of decentralized learning while maintaining low communication costs.

1.1 Topology Design for Communication-efficient Decentralized Learning

In Chapters 3 and 4, we design a novel communication pattern to improve the convergence rate of the decentralized learning methods. To reduce per-round communication costs, sparse topologies such as a ring are commonly adopted. While these sparse topologies are attractive from a per-round communication efficiency perspective, they typically lead to slow information propagation across the network. As a result, the convergence rate of decentralized learning methods deteriorates significantly, especially as the number of nodes increases. This trade-off between per-round communication efficiency and convergence rate motivates the development of new communication patterns that preserve fast convergence rate while maintaining low per-round communication costs.

In Chapter 3, we propose a novel graph called BASE- $(k+1)$ GRAPH. BASE- $(k+1)$ GRAPH requires nodes to communicate with k nodes in each round. By choosing a small k , e.g., 1, we can reduce the communication costs in each round. Despite such sparsity, we can prevent the degradation of the convergence rate since BASE- $(k+1)$ GRAPH efficiently mixes the parameters held by the nodes.

In Chapter 4, we propose a novel decentralized learning method called TELEPORTATION. BASE- $(k+1)$ GRAPH can prevent the degradation of the convergence rate even if the number of nodes is extremely large. However, its convergence rate still deteriorates as the number of nodes increases. To address this issue, we propose TELEPORTATION. TELEPORTATION is based on a totally different idea than BASE- $(k+1)$ GRAPH. For BASE- $(k+1)$ GRAPH, we consider ways to connect nodes so that we can efficiently mix the parameters, while for TELEPORTATION, we propose activating only a few nodes during training. By choosing an appropriate number of active nodes, TELEPORTATION can completely eliminate the degradation of the convergence rate caused by increasing the number of nodes, and the convergence rate is consistently improved as the number of nodes increases.

1.2 Optimizer for Communication-efficient Decentralized Learning

In Chapter 5, we study communication efficiency from an algorithmic perspective and propose new decentralized optimization methods that exploit similarities among local datasets. A fundamental characteristic of decentralized learning is that each node holds a different local dataset. In decentralized learning, nodes exchange model parameters with neighboring nodes without directly sharing the training datasets, enabling all nodes to train the same model. Therefore, decentralized learning methods require more communication costs when the nodes have significantly different datasets, which is called the heterogeneous setting.

A number of prior studies have addressed this issue (Lorenzo and Scutari, 2016; Pu and Nedic, 2021; Lin et al., 2021; Yuan et al., 2021; Takezawa et al., 2023a; Aketi et al., 2023). The most seminal decentralized learning method is Gradient Tracking (Lorenzo and Scutari, 2016; Pu and Nedic, 2021), which can converge to a stationary point with the same communication costs regardless of the data heterogeneity. While this robustness is appealing, it also implies that the same communication cost is incurred regardless of whether local datasets are nearly identical or completely different. From an intuitive standpoint, this is unsatisfactory: when nodes have highly similar datasets, decentralized learning should be possible with substantially less communication.

Motivated by this observation, in Chapter 5, we propose Stabilized Proximal Decentralized Optimization (SPDO) and Accelerated Stabilized Proximal Decentralized Optimization (ACCELERATED SPDO). These methods explicitly exploit similarity among local datasets and adapt their communication requirements accordingly. When data heterogeneity is small, SPDO and ACCELERATED SPDO achieve significantly reduced communication costs. Furthermore, we show that the communication complexity of ACCELERATED SPDO is optimal up to logarithmic factors, establishing its theoretical optimality in communication-efficient decentralized learning.

1.3 Contributions beyond this Thesis

In addition to the research detailed in this thesis, the author published the following papers. In the first paper, we propose a novel decentralized learning method that is robust to data heterogeneity. In the second paper, we investigate decentralized learning methods that utilize communication compression to minimize communication costs. In the third paper, we apply a sophisticated adaptive optimizer, called Muon, to the field of federated learning.

- Yuki Takezawa, Anastasia Koloskova, Xiaowen Jiang, Sebastian Stich, “FedMuon: Federated Learning with Bias-corrected LMO-based Optimization”, *In International Conference on Learning Representations 2026*

- Yuki Takezawa, Han Bao, Kenta Niwa, Ryoma Sato, Makoto Yamada,
“Momentum Tracking: Momentum Acceleration for Decentralized Deep Learning on Heterogeneous Data”,
In Transactions on Machine Learning Research 2023
- Yuki Takezawa, Kenta Niwa, Makoto Yamada,
“Communication Compression for Decentralized Learning with Operator Splitting Methods”,
In IEEE Transactions on Signal and Information Processing over Networks 2023

Beyond decentralized learning topics, the author worked on a variety of research topics. These papers are not included in this thesis as they are not connected to the main topic of decentralized learning.

- Yuki Takezawa, Kenta Niwa, Makoto Yamada,
“Necessary and Sufficient Watermark for Large Language Models”,
In Transactions on Machine Learning Research 2025
- Yuki Takezawa, Han Bao, Ryoma Sato, Kenta Niwa, Makoto Yamada,
“Parameter-free Clipped Gradient Descent Meets Polyak”,
In Advances in Neural Information Processing Systems 2024
- Yuki Takezawa, Ryoma Sato, Zornitsa Kozareva, Sujith Ravi, Makoto Yamada,
“Fixed Support Tree-Sliced Wasserstein Barycenter”,
In International Conference on Artificial Intelligence and Statistics 2022
- Yuki Takezawa, Ryoma Sato, Makoto Yamada,
“Supervised Tree-Wasserstein Distance”,
In Transactions on Machine Learning Research 2021

2

Setup and Assumptions

In this section, we introduce the problem setting that is the focus of this thesis.

2.1 Notation

For any $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$. The norm $\|\cdot\|$ without a subscript denotes the Euclidean norm, and $\|\cdot\|_F$ denotes Frobenius norm. $\mathbf{I}_d$ denotes the d -dimensional identity matrix, $\mathbf{0}$ denotes a vector with all zeros, and $\mathbf{1}$ denotes a vector with all ones.

2.2 Problem Setting

The problem we address is one where the loss functions are distributed among n nodes:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \left[f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}) \right], \quad f_i(\mathbf{x}) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i} [F_i(\mathbf{x}; \xi_i)], \quad (2.1)$$

where $\mathbf{x}$ is a model parameter, $\mathcal{D}_i$ is a training dataset that the i -th node has, ξ_i is a data sample following $\mathcal{D}_i$, and the local loss function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is defined as the expectation of $F_i(\cdot; \xi_i)$ over data sample ξ_i . The nodes are connected by a graph, in which $W_{ij} \in [0, 1]$ represents the edge weight between nodes i and j , and $W_{ij} > 0$ if and only if nodes i and j are connected.

2.3 Assumptions on Loss Functions

Following the previous work (Koloskova et al., 2020; Li et al., 2020; Murata and Suzuki, 2021; Tian et al., 2022; Sun et al., 2022; Lin et al., 2023; Khaled and Jin, 2023; Jiang et al., 2024b), we use the following assumptions on the loss functions.

Assumption 1. *There exists $f^* > -\infty$ that satisfies $f(\mathbf{x}) \geq f^*$ for any $\mathbf{x} \in \mathbb{R}^d$.*

Assumption 2. *There exists $L \geq 0$ that satisfies for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $i \in [n]$,*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|. \quad (2.2)$$

Assumption 3. *It holds that $\mathbb{E}_{\xi_i \sim \mathcal{D}_i}[\nabla F_i(\mathbf{x}; \xi_i)] = \nabla f_i(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$. Then, there exists $\sigma \geq 0$ that satisfies for any $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$,*

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i} \|\nabla F_i(\mathbf{x}; \xi_i) - \nabla f_i(\mathbf{x})\|^2 \leq \sigma^2. \quad (2.3)$$

The above three assumptions are quite common in the optimization literature, e.g., (Nesterov, 2018). Furthermore, since each node has a different loss function, we use the following assumption to measure their dissimilarity:

Assumption 4. *There exists $\zeta \geq 0$ that satisfies for any $\mathbf{x} \in \mathbb{R}^d$,*

$$\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \zeta^2. \quad (2.4)$$

The quantity of ζ is often called data heterogeneity, and a large ζ implies that nodes have different datasets. When ζ is large, it is called a heterogeneous setting, and when ζ is close to 0, it is called a homogeneous setting.

In Chapter 5, we further assume that the loss function is (strongly) convex.

Assumption 5. *There exists $\mu \geq 0$ such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and i , it holds that*

$$f_i(\mathbf{x}) \geq f_i(\mathbf{y}) + \langle \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2. \quad (2.5)$$

2.4 Assumptions on Network Topology

We use the following assumption for the topology.

Assumption 6. *Let λ_i be the i -th largest eigenvalue of $\mathbf{W} \in [0, 1]^{n \times n}$. $\mathbf{W}$ is a doubly stochastic matrix, and $p := 1 - \max\{|\lambda_2|, |\lambda_n|\}^2$ satisfies $p \in (0, 1]$. That is, it holds that*

$$\|\mathbf{X}\mathbf{W} - \bar{\mathbf{X}}\|_F^2 \leq (1 - p) \|\mathbf{X} - \bar{\mathbf{X}}\|_F^2,$$

for any $\mathbf{X} \in \mathbb{R}^{d \times n}$ where $\bar{\mathbf{X}} := \frac{1}{n} \mathbf{X} \mathbf{1} \mathbf{1}^\top$.

The above assumption holds when the topology is connected. The quantity of p is often called the spectral gap, and the quantity of $\sqrt{1 - p}$ is called the consensus rate. The spectral gap measures how well the topology is connected, and as the topology becomes sparse, p approaches zero. For instance, when the topology is fully-connected, i.e., $\mathbf{W} = \frac{1}{n} \mathbf{1} \mathbf{1}^\top$, $p = 1$. Additionally, we assume that the topology is undirected, i.e., $\mathbf{W} = \mathbf{W}^\top$, when necessary.

3

Communication-efficient Topologies for Decentralized Learning via Finite-time Convergence

Abstract

Many recent studies stated that the underlying network topology with a faster consensus rate (a.k.a. spectral gap) leads to a better convergence rate and accuracy for decentralized learning. However, a topology with a fast consensus rate, e.g., the exponential graph, generally has a large maximum degree, which incurs significant communication costs. Thus, seeking topologies with both a fast consensus rate and small maximum degree is important. In this chapter, we propose a novel topology combining both a fast consensus rate and small maximum degree called the BASE- $(k + 1)$ GRAPH. Unlike the existing topologies, the BASE- $(k + 1)$ GRAPH enables all nodes to reach the exact consensus after a finite number of iterations for any number of nodes and maximum degree k . Thanks to this favorable property, the BASE- $(k + 1)$ GRAPH endows Decentralized SGD with both a faster convergence rate and more communication efficiency than the exponential graph. We conducted experiments with various topologies, demonstrating that the BASE- $(k + 1)$ GRAPH enables various decentralized learning methods to achieve higher accuracy with better communication efficiency than the existing topologies.

3.1 Introduction

In decentralized learning, each node only needs to communicate with a few nodes, while this sparse communication characteristic may degrade the convergence rate and accuracy (Koloskova et al., 2020; Zhu et al., 2022). Specifically, the smaller the maximum degree of an underlying network topology is, the fewer the communication cost becomes (Song et al., 2022; Wang et al., 2019); meanwhile, the faster the

Table 3.1: Comparison among different topologies with n nodes. The definition of the consensus rate and finite-time convergence is shown in Sec. 3.3.

Topology	Consensus Rate	Maximum Degree	#Nodes n
Ring (Nedic et al., 2018)	$1 - O(n^{-2})$	2	$\forall n \in \mathbb{N}$
Torus (Nedic et al., 2018)	$1 - O(n^{-1})$	4	$\forall n \in \mathbb{N}$
Exp. (Ying et al., 2021)	$1 - O((\log_2(n))^{-1})$	$\lceil \log_2(n) \rceil$	$\forall n \in \mathbb{N}$
1-peer Exp. (Ying et al., 2021)	$O(\log_2(n))$ -finite time conv.	1	A power of 2
1-peer Hypercube (Shi et al., 2016)	$O(\log_2(n))$ -finite time conv.	1	A power of 2
Base-$(k+1)$ Graph (ours)	$O(\log_{k+1}(n))$ -finite time conv.	k	$\forall n \in \mathbb{N}$

consensus rate (a.k.a. spectral gap) of a topology is, the faster the convergence rate of decentralized learning becomes (Koloskova et al., 2020). Thus, developing a topology with both a fast consensus rate and small maximum degree is essential for decentralized learning. Table 3.1 summarizes the properties of various topologies. For instance, the ring and exponential graph are commonly used (Assran et al., 2019; Cyffers et al., 2022; Koloskova et al., 2019; Lu and De Sa, 2020). The ring is a communication-efficient topology because its maximum degree is two but its consensus rate quickly deteriorates as the number of nodes n increases (Nedic et al., 2018). The exponential graph has a fast consensus rate, which does not deteriorate much as n increases, but it incurs significant communication costs because its maximum degree increases as n increases (Ying et al., 2021). Thus, these topologies sacrifice either communication efficiency or consensus rate.

Recently, the 1-peer exponential graph (Ying et al., 2021) and 1-peer hypercube graph (Shi et al., 2016) were proposed as topologies that combine both a small maximum degree and fast consensus rate (see Sec. B.3.4 for examples). As Fig. 3.1 shows, in the ring and exponential graph, node parameters only reach the consensus asymptotically by repeating exchanges of parameters with neighbors. Contrarily, in the 1-peer exponential and 1-peer hypercube graphs, parameters reach the exact consensus after a finite number of iterations when n is a power of 2 (see Fig. B.12 in Sec. B.6.2). Thanks to this property of finite-time convergence, the 1-peer exponential and 1-peer hypercube graphs enable Decentralized SGD (Lian et al., 2017) to converge at the same convergence rate as the exponential graph when n is a power of 2, even though the maximum degree of the 1-peer exponential and 1-peer hypercube graphs is only one (Ying et al., 2021).

However, this favorable property only holds when n is a power of 2. When n is not a power of 2, the 1-peer hypercube graph cannot be constructed, and the 1-peer exponential graph only reaches the consensus asymptotically as well as the ring and exponential graph, as Fig. 3.1 illustrates. Thus, the 1-peer exponential and 1-peer hypercube graphs cannot enable Decentralized SGD to converge as fast as the exponential graph when n is not a power of 2. Moreover, even if n is a power of 2, the 1-peer hypercube and 1-peer exponential graphs still cannot enable Decentralized SGD to converge faster than the exponential graph.

In this chapter, we ask the following question: *Can we construct topologies that*

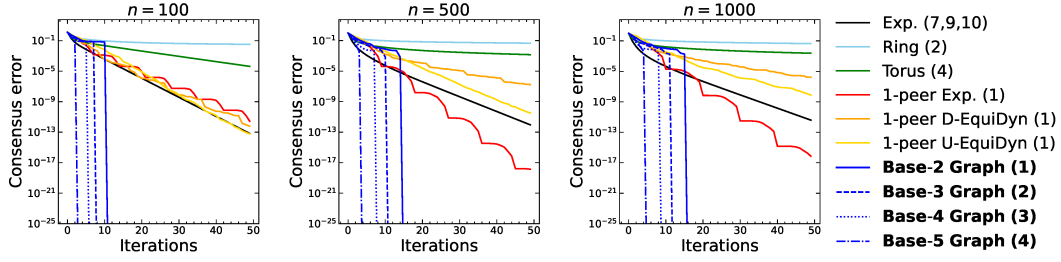


Figure 3.1: Comparison of consensus rate. See Sec. 3.6 for detailed experimental settings. The number in the bracket is the maximum degree.

provide Decentralized SGD with both a faster convergence rate and better communication efficiency than the exponential graph for any number of nodes? Our work provides the affirmative answer by proposing the BASE- $(k+1)$ GRAPH,¹ which is finite-time convergence for any number of nodes n and maximum degree k (see Fig. 3.1). Thanks to this favorable property, the BASE-2 GRAPH enables Decentralized SGD to converge faster than the ring and torus and as fast as the exponential graph for any n , while the BASE-2 GRAPH is more communication-efficient than the ring, torus, and exponential graph because its maximum degree is only one. Furthermore, when $2 \leq k < \lceil \log_2(n) \rceil$, the BASE- $(k+1)$ GRAPH enables Decentralized SGD to converge faster with fewer communication costs than the exponential graph because the maximum degree of the BASE- $(k+1)$ GRAPH is still less than that of the exponential graph. Experimentally, we compared the BASE- $(k+1)$ GRAPH with various existing topologies, demonstrating that the BASE- $(k+1)$ GRAPH enables various decentralized learning methods to more successfully reconcile accuracy and communication efficiency than the existing topologies.

3.2 Related Work

Many prior studies indicated that topologies with a fast consensus rate improve the accuracy of decentralized learning (Koloskova et al., 2020; Kong et al., 2021; Marfoq et al., 2020; Wang et al., 2019). For instance, Decentralized SGD and Gradient Tracking converge faster as the topology has a faster consensus rate (Lian et al., 2017; Takezawa et al., 2023a). Zhu et al. (2022) revealed that topology with a fast consensus rate improves the generalization bound of Decentralized SGD. Especially when the data distributions are statistically heterogeneous, the topology with a fast consensus rate prevents the parameters of each node from drifting away and can improve accuracy (Koloskova et al., 2020; Takezawa et al., 2023a). However, communication costs increase as the maximum degree increases (Song et al., 2022; Ying et al., 2021). Thus, developing a topology with a fast consensus rate and small maximum degree is important for decentralized learning.

¹Note that the maximum degree of the BASE- $(k+1)$ GRAPH is not $k+1$, but at most k .

3.3 Preliminary and Notation

Notation. A graph G is represented by (V, E) where V is a set of nodes and E is a set of edges. If G is a graph, $V(G)$ (resp. $E(G)$) denotes the set of nodes (resp. edges) of G . For any $a_1, \dots, a_n$, $(a_1, \dots, a_n)$ denotes the ordered set. An empty (ordered) set is denoted by $\emptyset$. For any $n, a \in \mathbb{N}$, $\text{mod}(a, n)$ is the remainder of dividing a by n .

Topology. Let G be an underlying network topology with n nodes, and $\mathbf{W} \in [0, 1]^{n \times n}$ be a mixing matrix associated with G . That is, W_{ij} is the weight of the edge (i, j) , and $W_{ij} > 0$ if and only if $(i, j) \in E(G)$. Most of the decentralized learning methods require $\mathbf{W}$ to be doubly stochastic (i.e., $\mathbf{W}\mathbf{1} = \mathbf{1}$ and $\mathbf{W}^\top \mathbf{1} = \mathbf{1}$) (Lian et al., 2017; Lin et al., 2021; Nedić et al., 2017; Tang et al., 2018). Then, the consensus rate of G is defined below.

Assumption 7. Let $\mathbf{W}$ be a mixing matrix associated with a graph G with n nodes. There exists $\rho \in [0, 1)$ such that for any $\mathbf{X} \in \mathbb{R}^{d \times n}$, it satisfies

$$\|\mathbf{X}\mathbf{W} - \bar{\mathbf{X}}\|_F^2 \leq \rho^2 \|\mathbf{X} - \bar{\mathbf{X}}\|_F^2,$$

where $\bar{\mathbf{X}} := \frac{1}{n} \mathbf{X} \mathbf{1} \mathbf{1}^\top$.

The quantity of ρ is called the consensus rate. Thanks to $\rho \in [0, 1)$, $\mathbf{x}_i$ asymptotically converge to consensus $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ by repeating parameter exchanges with neighbors. However, this does not mean that all nodes reach the exact consensus within a finite number of iterations except when $\rho = 0$, that is, when G is fully connected. Then, utilizing time-varying topologies, Ying et al. (2021) and Shi et al. (2016) aimed to obtain sequences of graphs that can make all nodes reach the exact consensus within finite iterations and proposed the 1-peer exponential and 1-peer hypercube graphs respectively (see Sec. B.3.4 for illustrations).

Definition 1. Let $(G^{(1)}, \dots, G^{(m)})$ be a sequence of graphs with the same set of nodes (i.e., $V(G^{(1)}) = \dots = V(G^{(m)})$). Let n be the number of nodes. Let $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(m)}$ be mixing matrices associated with $G^{(1)}, \dots, G^{(m)}$, respectively. Suppose that $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(m)}$ satisfy $\mathbf{X}\mathbf{W}^{(1)}\mathbf{W}^{(2)} \dots \mathbf{W}^{(m)} = \bar{\mathbf{X}}$ for any $\mathbf{X} \in \mathbb{R}^{d \times n}$, where $\bar{\mathbf{X}} = \frac{1}{n} \mathbf{X} \mathbf{1} \mathbf{1}^\top$. Then, $(G^{(1)}, \dots, G^{(m)})$ is called m -finite time convergence or an m -finite time convergent sequence of graphs.

Because Definition 1 assumes that $V(G^{(1)}) = \dots = V(G^{(m)})$ holds, we often write a sequence of graphs $(G^{(1)}, \dots, G^{(m)})$ as $(E(G^{(1)}), \dots, E(G^{(m)}))$ using a slight abuse of notation. Additionally, in the following section, we often abbreviate the weights of self-loops because they are uniquely determined due to the condition that the mixing matrix is doubly stochastic.

Algorithm 1 k -PEER HYPER-HYPERCUBE GRAPH $\mathcal{H}_k(V)$

```

1: Input: the set of nodes  $V := \{v_1, \dots, v_n\}$  and number of nodes  $n$ .
2: Decompose  $n$  as  $n = n_1 \times \dots \times n_L$  with minimum  $L$  such that  $n_l \in [k+1]$  for all  $l \in [L]$ .
3: for  $l \in [L]$  do
4:   Initialize  $b_i$  to zero for all  $i \in [n]$  and  $E^{(l)}$  to  $\emptyset$ .
5:   for  $i \in [n]$  do
6:     for  $m \in [n_l]$  do
7:        $j \leftarrow \text{mod}(i + m \times \prod_{l'=1}^{l-1} n_{l'} - 1, n) + 1$ .
8:       if  $b_i < n_l - 1$  and  $b_j < n_l - 1$  then
9:         Add edge  $(v_i, v_j)$  with weight  $\frac{1}{n_l}$  to  $E^{(l)}$  and  $b_i \leftarrow b_i + 1$ .
10: return  $(E^{(1)}, E^{(2)}, \dots, E^{(L)})$ .

```

3.4 Construction of Finite-time Convergent Sequence of Graphs

In this section, we propose the BASE- $(k+1)$ GRAPH, which is finite-time convergence for any number of nodes $n \in \mathbb{N}$ and maximum degree $k \in [n-1]$. Specifically, we consider the setting where node i has a parameter $\mathbf{x}_i$ and propose a graph sequence whose maximum degree is at most k that makes all nodes reach the exact consensus $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. To this end, we first propose the k -PEER HYPER-HYPERCUBE GRAPH, which is finite-time convergence when n does not have prime factors larger than $k+1$. Using it, we propose the SIMPLE BASE- $(k+1)$ GRAPH and BASE- $(k+1)$ GRAPH, which are finite-time convergence for any n .

3.4.1 k -peer Hyper-hypercube Graph

Before proposing the BASE- $(k+1)$ GRAPH, we first extend the 1-peer hypercube graph (Shi et al., 2016) to the k -peer setting and propose the k -PEER HYPER-HYPERCUBE GRAPH, which is finite-time convergence when the number of nodes n does not have prime factors larger than $k+1$ and is used as a component in the BASE- $(k+1)$ GRAPH. Let V be a set of n nodes. We assume that all prime factors of n are less than or equal to $k+1$. That is, there exists $n_1, n_2, \dots, n_L \in [k+1]$ such that $n = n_1 \times \dots \times n_L$. In this case, we can construct the L -finite time convergent sequence of graphs whose maximum degree is at most k . Using Fig. 3.2a, we explain

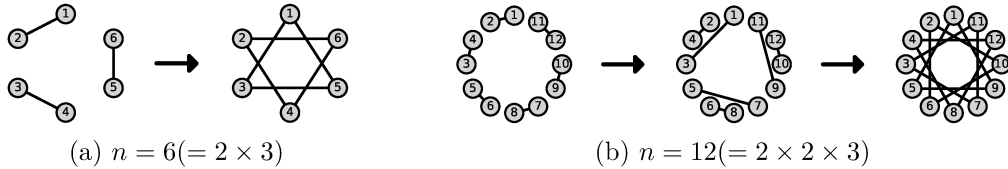


Figure 3.2: Illustration of the 2-PEER HYPER-HYPERCUBE GRAPH.

how all nodes reach the exact consensus. Let $G^{(1)}$ and $G^{(2)}$ denote the graphs in Fig. 3.2a from left to right, respectively. After the nodes exchange parameters with neighbors in $G^{(1)}$, nodes 1 and 2, nodes 3 and 4, and nodes 5 and 6 have the same parameter respectively. Then, after exchanging parameters in $G^{(2)}$, all nodes reach the exact consensus. We present the complete algorithms for constructing the k -PEER HYPER-HYPERCUBE GRAPH in Alg. 1.

3.4.2 Simple Base- $(k + 1)$ Graph

As described in Sec. 3.4.1, when n does not have prime factors larger than $k + 1$, we can easily make all nodes reach the exact consensus by the k -PEER HYPER-HYPERCUBE GRAPH. However, when n has prime factors larger than $k + 1$, e.g., when $(k, n) = (1, 5)$, the k -PEER HYPER-HYPERCUBE GRAPH cannot be constructed. In this section, we extend the k -PEER HYPER-HYPERCUBE GRAPH and propose the SIMPLE BASE- $(k + 1)$ GRAPH, which is finite-time convergence for any number of nodes n and maximum degree k . Note that the maximum degree of the SIMPLE BASE- $(k + 1)$ GRAPH is not $k + 1$, but at most k .

We present the pseudo-code for constructing the SIMPLE BASE- $(k + 1)$ GRAPH in Alg. 2. For simplicity, here we explain only the case when the maximum degree k is one. The case with $k \geq 2$ is explained in Sec. B.2. The SIMPLE BASE- $(k + 1)$ GRAPH mainly consists of the following four steps. The key idea is that splitting V into disjoint subsets to which the k -PEER HYPER-HYPERCUBE GRAPH is applicable.

Step 1. As in the base-2 number of n , we decompose n as $n = 2^{p_1} + \dots + 2^{p_L}$ in line 1, and then split V into disjoint subsets $V_1, \dots, V_L$ such that $|V_l| = 2^{p_l}$ for all $l \in [L]$ in line 3.²

Step 2. For all $l \in [L]$, we make all nodes in V_l obtain the average of parameters in V_l using the 1-PEER HYPER-HYPERCUBE GRAPH $\mathcal{H}_1(V_l)$ in line 11. Then, we initialize l' as one.

Step 3. Each node in $V_{l'+1} \cup \dots \cup V_L$ exchanges parameters with one node in $V_{l'}$ such that the average in $V_{l'}$ becomes equivalent to the average in V . We increase l' by one and repeat step 3 until $l' = L$. This procedure corresponds to line 15.

Step 4. For all $l \in [L]$, we make all nodes in V_l obtain the average in V_l using the 1-PEER HYPER-HYPERCUBE GRAPH $\mathcal{H}_1(V_l)$. Because the average in V_l is equivalent to the average in V after step 3, all nodes can reach the exact consensus. This procedure corresponds to line 25.

Using the example presented in Fig. 3.3, we explain Alg. 2 in more detail. Let $G^{(1)}, \dots, G^{(5)}$ denote the graphs in Fig. 3.3 from left to right, respectively. In step 1, we split $V := \{1, \dots, 5\}$ into $V_1 := \{1, \dots, 4\}$ and $V_2 := \{5\}$. In step 2, nodes in V_1 obtain the average in V_1 by exchanging parameters in $G^{(1)}$ and $G^{(2)}$. In step 3,

²Splitting V_l into $V_{l,1}, \dots, V_{l,a_l}$ becomes crucial when $k \geq 2$ (see Sec. B.2).

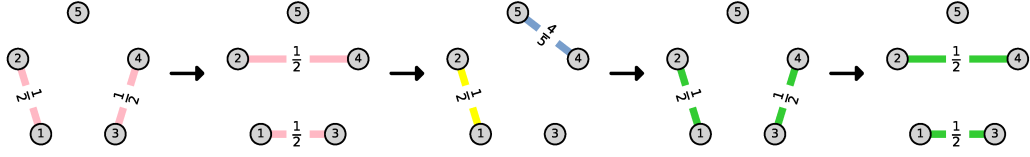


Figure 3.3: SIMPLE BASE-2 GRAPH with $n = 5 (= 2^2 + 1)$. The value on the edge is the edge weight, and the edges are colored in the same color as the line in Alg. 2 where they were added.

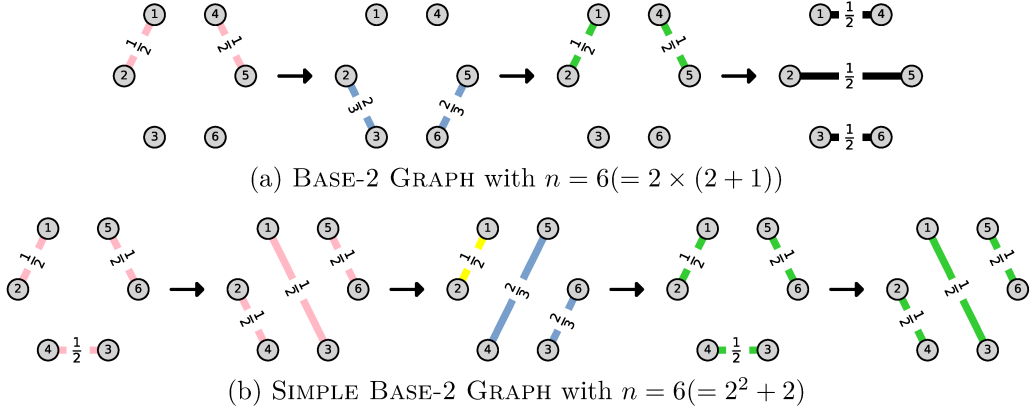


Figure 3.4: Comparison of SIMPLE BASE-2 GRAPH and BASE-2 GRAPH with $n = 6$. The value on the edge indicates the edge weight. The edges added in line 10 in Alg. 3 are colored black, and the edges added in line 6 are colored the same color as the line in Alg. 2 where they are added.

the average in V_1 becomes equivalent to the average in V by exchanging parameters in $G^{(3)}$. In step 4, nodes in V_1 can get the average in V by exchanging parameters in $G^{(4)}$ and $G^{(5)}$. Because node 5 also obtains the average in V after exchanging parameters in $G^{(3)}$, all nodes reach the exact consensus after exchanging parameters in $G^{(5)}$.

Note that edges added in lines 20 and 27 are not necessary if we only need to make all nodes reach the exact consensus. Nonetheless, these edges are effective in keeping the parameters of nodes close in value to each other in decentralized learning because the parameters are updated by gradient descent before the parameter exchange with neighbors. For instance, edge $(1, 2)$ in $G^{(3)}$, which is added in line 20, is not necessary for finite-time convergence because nodes 1 and 2 already have the same parameter after exchanging parameters in $G^{(1)}$ and $G^{(2)}$. We provide more examples in Sec. B.3.

3.4.3 Base- $(k + 1)$ Graph

The SIMPLE BASE- $(k + 1)$ GRAPH is finite-time convergence for any n and k , while the SIMPLE BASE- $(k + 1)$ GRAPH contains graphs that are not necessary for the

finite-time convergence and becomes a redundant sequence of graphs in some cases, e.g., $(k, n) = (1, 6)$ (see the example in Fig. 3.4b and a detailed explanation in Sec. B.3.2). To remove this redundancy, this section proposes the BASE- $(k + 1)$ GRAPH that can make all nodes reach the exact consensus after fewer iterations than the SIMPLE BASE- $(k + 1)$ GRAPH.

The pseudo-code for constructing the BASE- $(k + 1)$ GRAPH is shown in Alg. 3. The BASE- $(k + 1)$ GRAPH consists of the following three steps.

Algorithm 2 SIMPLE BASE- $(k + 1)$ GRAPH $\mathcal{A}_k^{\text{simple}}(V)$

```

1: Input: the set of nodes  $V$  and number of nodes  $n (= a_1(k + 1)^{p_1} + a_2(k + 1)^{p_2} + \dots + a_L(k + 1)^{p_L})$  such that  $p_1 > p_2 > \dots > p_L \geq 0$  and  $a_l \in [k]$  for all  $l \in [L]$ .
2: If all prime factors of  $n$  are less than or equal to  $k + 1$  then return  $\mathcal{H}_k(V)$ .
3: Split  $V$  into disjoint subsets  $V_1, \dots, V_L$  such that  $|V_l| = a_l(k + 1)^{p_l}$  for all  $l \in [L]$ . Then, for all  $l \in [L]$ , split  $V_l$  into disjoint subsets  $V_{l,1}, \dots, V_{l,a_l}$  such that  $|V_{l,a}| = (k + 1)^{p_l}$  for all  $a \in [a_l]$ .
4: Construct  $k$ -PEER HYPER-HYPERCUBE GRAPH  $\mathcal{H}_k(V_l)$  for all  $l \in [L]$  and  $m_1 = |\mathcal{H}_k(V_1)|$ .
5: Construct  $k$ -PEER HYPER-HYPERCUBE GRAPH  $\mathcal{H}_k(V_{l,a})$  for all  $l \in [L]$  and  $a \in [a_l]$ .
6: Initialize  $b_l$  as zero for all  $l \in [L]$ , and initialize  $m$  as zero.
7: while  $b_1 < |\mathcal{H}_k(V_{1,1})|$  do
8:    $m \leftarrow m + 1$  and  $E^{(m)} \leftarrow \emptyset$ .
9:   for  $l \in \{L, L - 1, \dots, 1\}$  do
10:    if  $m \leq m_1$  then
11:      Add  $E(\mathcal{H}_k(V_l)^{(m')})$  to  $E^{(m)}$  where  $m' = \text{mod}(m - 1, |\mathcal{H}_k(V_l)|) + 1$ .
12:    else if  $m < m_1 + l$  then
13:      for  $v \in V_l$  do
14:        Select isolated node  $u_1, \dots, u_{a_{m-m_1}}$  from  $V_{m-m_1,1}, \dots, V_{m-m_1,a_{m-m_1}}$ .
15:        Add edges  $(v, u_1), \dots, (v, u_{a_{m-m_1}})$  with weight  $\frac{|V_{m-m_1}|}{a_{m-m_1} \sum_{l'=m-m_1}^L |V_{l'}|}$  to  $E^{(m)}$ .
16:    else if  $m = m_1 + l$  and  $l \neq L$  then
17:      while There are two or more isolated nodes in  $V_l$  do
18:         $c \leftarrow$  the number of isolated nodes in  $V_l$ .
19:        Select  $\min\{k + 1, c\}$  isolated nodes  $V'$  in  $V_l$ .
20:        Add edges with weights  $\frac{1}{|V'|}$  to  $E^{(m)}$  such that  $V'$  compose the complete graph.
21:    else
22:       $b_l \leftarrow b_l + 1$ .
23:      if  $p_l \neq 0$  then
24:        for  $a \in [a_l]$  do
25:          Add  $E(\mathcal{H}_k(V_{l,a})^{(m')})$  to  $E^{(m)}$  where  $m' = \text{mod}(b_l - 1, |\mathcal{H}_k(V_{l,a})|) + 1$ .
26:        else
27:          Add  $E(\mathcal{H}_k(V_l)^{(m')})$  to  $E^{(m)}$  where  $m' = \text{mod}(b_l - 1, |\mathcal{H}_k(V_l)|) + 1$ .
28: return  $(E^{(1)}, E^{(2)}, \dots, E^{(m)})$ .
```

Algorithm 3 BASE- $(k + 1)$ GRAPH $\mathcal{A}_k(V)$

-
- 1: **Input:** the set of nodes V and number of nodes n .
 - 2: Decompose n as $n = p \times q$ such that p is a multiple of $2, 3, \dots, (k + 1)$ and q is prime to $2, 3, \dots, (k + 1)$.
 - 3: Split V into disjoint subsets $V_1, \dots, V_p$ such that $|V_l| = q$ for all l .
 - 4: Construct SIMPLE BASE- $(k + 1)$ GRAPH $\mathcal{A}_k^{\text{simple}}(V_l)$ for all $l \in [p]$.
 - 5: **for** $m \in \{1, 2, \dots, |\mathcal{A}_k^{\text{simple}}(V_1)|\}$ **do**
 - 6: $E^{(m)} \leftarrow \bigcup_{l \in [p]} E(\mathcal{A}_k^{\text{simple}}(V_l)^{(m)})$.
 - 7: Split V into disjoint subsets $U_1, \dots, U_q$ such that $|U_l| = p$ and $|U_l \cap V_{l'}| = 1$ for all l, l' .
 - 8: Construct k -PEER HYPER-HYPERCUBE GRAPH $\mathcal{H}_k(U_l)$ for all $l \in [q]$.
 - 9: **for** $m \in \{1, 2, \dots, |\mathcal{H}_k(U_1)|\}$ **do**
 - 10: $E^{(m+|\mathcal{A}_k^{\text{simple}}(V_1)|)} \leftarrow \bigcup_{l \in [q]} E(\mathcal{H}_k(U_l)^{(m)})$.
 - 11: $\mathcal{E} \leftarrow (E^{(1)}, E^{(2)}, \dots, E^{(|\mathcal{A}_k^{\text{simple}}(V_1)|+|\mathcal{H}_k(U_1)|)})$.
 - 12: **If** $|\mathcal{A}_k^{\text{simple}}(V)| < |\mathcal{E}|$ **then return** $\mathcal{A}_k^{\text{simple}}(V)$ **else return** $\mathcal{E}$.
-

Step 1. We decompose n as $p \times q$ such that p is a multiple of $2, \dots, (k + 1)$ and q is prime to $2, \dots, (k + 1)$, and split V into disjoint subsets $V_1, \dots, V_p$ such that $|V_l| = q$ for all $l \in [p]$.

Step 2. For all $l \in [p]$, we make all nodes in V_l reach the average in V_l by the SIMPLE BASE- $(k + 1)$ GRAPH $\mathcal{A}_k^{\text{simple}}(V_l)$. Then, we take p nodes from $V_1, \dots, V_p$ respectively and construct a set U_1 . Similarly, we construct $U_2, \dots, U_q$ such that $U_1, \dots, U_q$ are disjoint sets.

Step 3. For all $l \in [q]$, we make all nodes in U_l reach the average in U_l by the k -PEER HYPER-HYPERCUBE GRAPH $\mathcal{H}_k(U_l)$. Because the average in U_l is equivalent to the average in V after step 2, all nodes reach the exact consensus.

Using the example in Fig. 3.4a, we explain the BASE- $(k + 1)$ GRAPH in more detail. Let $G^{(1)}, \dots, G^{(4)}$ denote the graphs in Fig. 3.4a from left to right, respectively. In step 1, we split V into $V_1 := \{1, 2, 3\}$ and $V_2 := \{4, 5, 6\}$. In step 2, nodes in V_1 and nodes in V_2 have the same parameter respectively by exchanging parameters on $G^{(1)}, \dots, G^{(3)}$ because the subgraphs composed on V_1 and V_2 are same as the SIMPLE BASE- $(k + 1)$ GRAPH (see Fig. B.5a). Then, we construct $U_1 := \{1, 4\}$, $U_2 := \{2, 5\}$, and $U_3 := \{3, 6\}$. Finally, in step 3, all nodes reach the exact consensus by exchanging parameters in $G^{(4)}$.

Fig. 3.5 and Sec. B.6.1 compare the BASE- $(k + 1)$ GRAPH with the SIMPLE BASE- $(k + 1)$ GRAPH, demonstrating that the length of the BASE- $(k + 1)$ GRAPH is less than that of the SIMPLE BASE- $(k + 1)$ GRAPH in many cases. Moreover, Theorem 1 show the upper bound of the length of the SIMPLE BASE- $(k + 1)$ GRAPH and BASE- $(k + 1)$ GRAPH. The proof is provided in Sec. B.4.

Theorem 1. *For any number of nodes $n \in \mathbb{N}$ and maximum degree $k \in [n - 1]$, the length of the SIMPLE BASE- $(k + 1)$ GRAPH and BASE- $(k + 1)$ GRAPH is less than or equal to $2 \log_{k+1}(n) + 2$.*

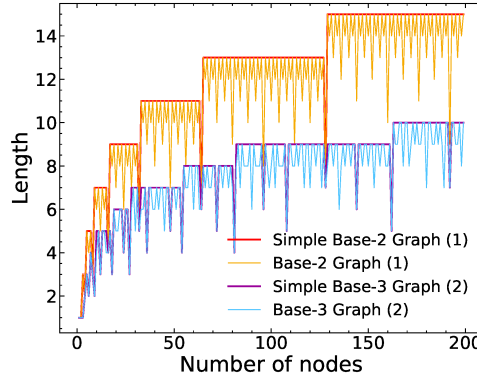


Figure 3.5: Comparison of length.

Corollary 1. *For any number of nodes $n \in \mathbb{N}$ and maximum degree $k \in [n - 1]$, the SIMPLE BASE- $(k + 1)$ GRAPH and BASE- $(k + 1)$ GRAPH are $\mathcal{O}(\log_{k+1}(n))$ -finite time convergence.*

Therefore, the BASE- $(k + 1)$ GRAPH is a powerful extension of the 1-peer exponential (Ying et al., 2021) and 1-peer hypercube graphs (Shi et al., 2016) because they are $\mathcal{O}(\log_2(n))$ -finite time convergence only if n is a power of 2 and their maximum degree cannot be set to any number other than 1.

3.5 Decentralized SGD on Base- $(k + 1)$ Graph

In this section, we verify the effectiveness of the BASE- $(k + 1)$ GRAPH for decentralized learning, demonstrating that the BASE- $(k + 1)$ GRAPH can endow Decentralized SGD with both a faster convergence rate and fewer communication costs than the existing topologies, including the ring, torus, and exponential graph. See Chapter 2 for the problem setting of decentralized learning.

We consider the case when Decentralized SGD (Lian et al., 2017), the most widely used decentralized learning method, is used as an optimization method. Let $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(m)}$ be mixing matrices of the BASE- $(k + 1)$ GRAPH. In Decentralized SGD on the BASE- $(k + 1)$ GRAPH, node i updates its parameter $\mathbf{x}_i$ as follows:

$$\mathbf{x}_i^{(r+1)} = \sum_{j=1}^n W_{ij}^{(1+\text{mod}(r,m))} \left(\mathbf{x}_j^{(r)} - \eta \nabla F_j(\mathbf{x}_j^{(r)}; \xi_j^{(r)}) \right), \quad (3.1)$$

where η is the learning rate. In this case, thanks to the property of finite-time convergence, Decentralized SGD on the BASE- $(k + 1)$ GRAPH converges at the following convergence rate.

Theorem 2. *Suppose that Assumptions 1-4 hold. Then, for any number of nodes $n \in \mathbb{N}$ and maximum degree $k \in [n - 1]$, there exists η such that $\bar{\mathbf{x}} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$*

generated by Eq. (3.1) satisfies $\frac{1}{R+1} \sum_{r=0}^R \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^{(r)}) \right\|^2 \leq \epsilon$ after

$$R = \mathcal{O} \left(\frac{\sigma^2}{n\epsilon^2} + \frac{\zeta \log_{k+1}(n) + \sigma \sqrt{\log_{k+1}(n)}}{\epsilon^{3/2}} + \frac{\log_{k+1}(n)}{\epsilon} \right) \cdot LF_0 \quad (3.2)$$

iterations, where $F_0 := f(\bar{\mathbf{x}}^{(0)}) - f^*$.

The above theorem follows immediately from Theorem 2 stated in Koloskova et al. (2020) and Corollary 1. The convergence rates of Decentralized SGD over commonly used topologies are summarized in Sec. B.5. From Theorem 2 and Sec. B.5, we can conclude that for any number of nodes n , the BASE-2 GRAPH enables Decentralized SGD to converge faster than the ring and torus and as fast as the exponential graph, although the maximum degree of the BASE-2 GRAPH is only one. Moreover, if we set the maximum degree k to the value between 2 to $\lceil \log_2(n) \rceil$, the BASE- $(k+1)$ GRAPH enables Decentralized SGD to converge faster than the exponential graph, even though the maximum degree of the BASE- $(k+1)$ GRAPH remains less than that of the exponential graph. It is worth noting that if we increase the maximum degree of the 1-peer exponential and 1-peer hypercube graphs (i.e., k -peer exponential and k -peer hypercube graphs with $k \geq 2$), these topologies cannot enable Decentralized SGD to converge faster than the exponential graph because these topologies are no longer finite-time convergence even when the number of nodes is a power of 2.

3.6 Experiments

In this section, we validate the effectiveness of the BASE- $(k+1)$ GRAPH. First, we experimentally verify that the BASE- $(k+1)$ GRAPH is finite-time convergence for any number of nodes in Sec. 3.6.1, and we verify the effectiveness of the BASE- $(k+1)$ GRAPH for decentralized learning in Sec. 3.6.2.

3.6.1 Consensus Rate

Setup. Let $x_i \in \mathbb{R}$ be the parameter that node i has, and let $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$. For each i , the initial value of x_i was drawn from Gaussian distribution with mean 0 and standard variance 1. Then, we evaluated how the consensus error $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ decreases when x_i is updated as $x_i \leftarrow \sum_{j=1}^n W_{ij} x_j$ where $\mathbf{W}$ is the mixing matrix associated with a given topology.

Results. Figs. 3.1 and 3.6 present how the consensus errors decrease on various topologies. The results indicate that the BASE- $(k+1)$ GRAPH reaches the exact consensus after a finite number of iterations, while the other topologies only reach the consensus asymptotically. Moreover, as the maximum degree k increases, the BASE- $(k+1)$ GRAPH reaches the exact consensus with fewer iterations. We also present the results when n is a power of 2 in Sec. B.6.2, demonstrating that the 1-peer exponential graph can reach the exact consensus as well as the BASE-2 GRAPH, but requires more iterations than the BASE-4 GRAPH.

3.6.2 Decentralized Learning

Next, we examine the effectiveness of the BASE- $(k + 1)$ GRAPH in decentralized learning.

Setup. We used three datasets, Fashion MNIST (Xiao et al., 2017), CIFAR- $\{10, 100\}$ (Krizhevsky, 2009), and used LeNet (LeCun et al., 1998) for Fashion MNIST and VGG-11 (Simonyan and Zisserman, 2015) for CIFAR- $\{10, 100\}$. Additionally, we

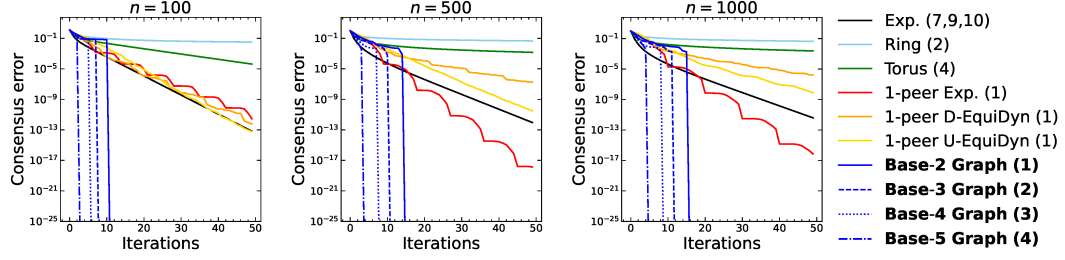


Figure 3.6: Comparison of consensus rates among various topologies. The number in the bracket indicates the maximum degree of a topology. Because the maximum degree of the exponential graph depends on n , the three numbers in the bracket indicate the maximum degree for each n .

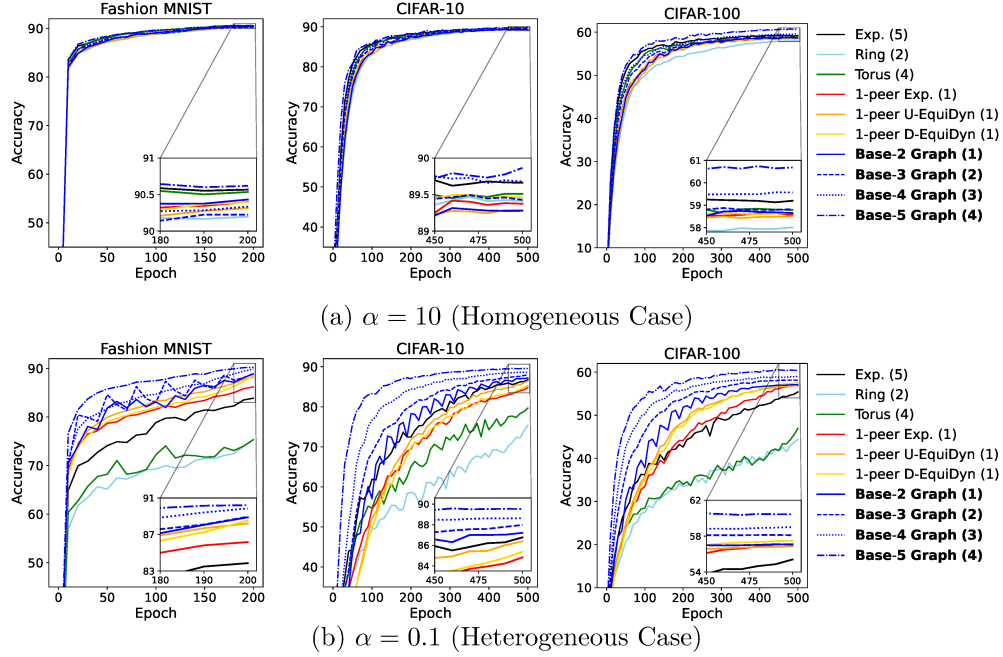


Figure 3.7: Test accuracy (%) of Decentralized SGD on various topologies with $n = 25$. The number in the bracket indicates the maximum degree of a topology. We also compared with dense variants of the 1-peer {U, D}-EquiDyn (Song et al., 2022) in Sec. B.6.3, showing the superior performance of the BASE- $(k + 1)$ GRAPH.

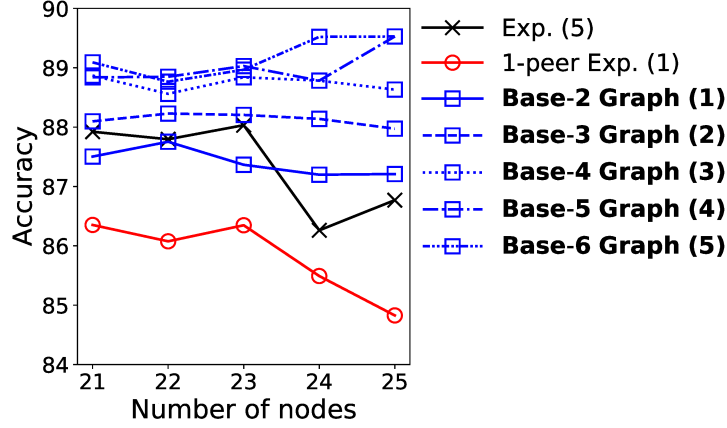


Figure 3.8: Test accuracy (%) of Decentralized SGD with CIFAR-10 and $\alpha = 0.1$.

present the results using ResNet-18 (He et al., 2016) in Sec. B.7. The learning rate was tuned by the grid search and we used the cosine learning rate scheduler (Loshchilov and Hutter, 2017). We distributed the training dataset to nodes by using Dirichlet distributions with hyperparameter α (Hsu et al., 2019), conducting experiments in both homogeneous and heterogeneous data distribution settings. As α approaches zero, the data distributions held by each node become more heterogeneous. We repeated all experiments with three different seed values and reported their averages. See Sec. B.8 for more detailed settings.

Results of Decentralized SGD on Various Topologies. We compared various topologies combined with the Decentralized SGD with momentum (Gao and Huang, 2020; Lian et al., 2017), showing the results in Fig. 3.7. From Fig. 3.7, the accuracy differences among topologies are larger as the data distributions are more heterogeneous. From Fig. 3.7b, the BASE- $\{2, 3, 4, 5\}$ GRAPH reach high accuracy faster than the other topologies. Furthermore, comparing the final accuracy, the final accuracy of the BASE-2 GRAPH is comparable to or higher than that of the existing topologies, including the exponential graph. Moreover, the final accuracy of the BASE- $\{3, 4, 5\}$ GRAPH is higher than that of all existing topologies. From Fig. 3.7a, the accuracy differences among topologies become small when $\alpha = 10$; however, the BASE-5 GRAPH still outperforms the other topologies. In Fig. 3.8, we present the results in cases other than $n = 25$, demonstrating that the BASE-2 GRAPH outperforms the 1-peer exponential graph and the BASE- $\{3, 4, 5\}$ GRAPH can consistently outperform the exponential and 1-peer exponential graphs for all n . In Sec. B.6.3, we show the learning curves and the comparison of the consensus rate when n is 21, 22, 23, 24, and 25.

Results of D^2 and QG-DSGDm on Various Topologies. The above results demonstrated that the BASE- $(k+1)$ GRAPH outperforms the existing topologies, especially when the data distributions are heterogeneous. Hence, we next compared the BASE- $(k+1)$ GRAPH with the existing topologies in the case where D^2 (Tang et al.,

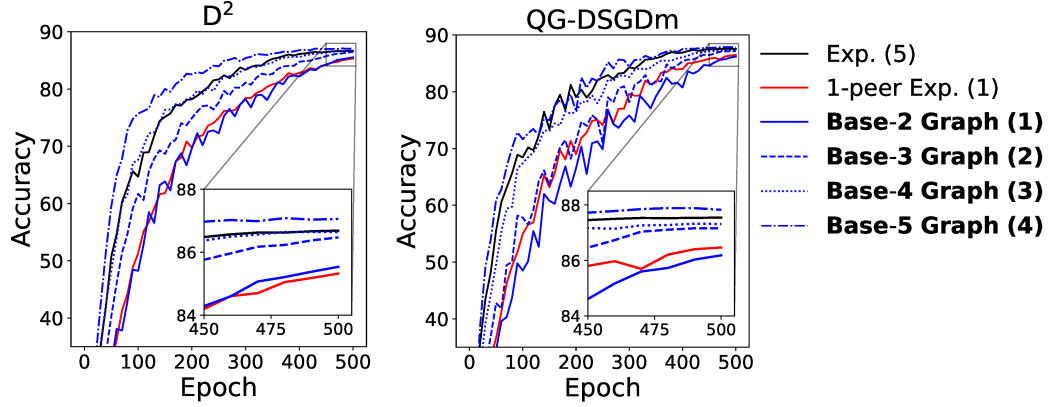


Figure 3.9: Test accuracy (%) of D^2 and QG-DSGDm with CIFAR-10, $n = 25$, and $\alpha = 0.1$.

2018) and QG-DSGDm (Lin et al., 2021), which are robust to data heterogeneity, are used as decentralized learning methods. From Fig. 3.9, the BASE-2 GRAPH can achieve comparable or higher accuracy than the 1-peer exponential graph, and the BASE-5 GRAPH consistently outperforms the exponential graph. Thus, the BASE- $(k + 1)$ GRAPH is useful not only for Decentralized SGD but also for D^2 and QG-DSGDm and then enables these methods to achieve a reasonable balance between accuracy and communication efficiency.

3.7 Conclusion

In this chapter, we propose the BASE- $(k + 1)$ GRAPH, a novel topology with both a fast consensus rate and small maximum degree. Unlike the existing topologies, the BASE- $(k + 1)$ GRAPH is finite-time convergence for any number of nodes and maximum degree k . Thanks to this favorable property, the BASE- $(k + 1)$ GRAPH enables Decentralized SGD to obtain both a faster convergence rate and more communication efficiency than the existing topologies, including the ring, torus, and exponential graph. Through experiments, we compared the BASE- $(k + 1)$ GRAPH with various existing topologies, demonstrating that the BASE- $(k + 1)$ GRAPH enables various decentralized learning methods to more successfully reconcile accuracy and communication efficiency than the existing topologies.

4

Scalable Decentralized Learning with Teleportation

Abstract

Decentralized SGD can run with low communication costs, but its sparse communication characteristics deteriorate the convergence rate, especially when the number of nodes is large. In decentralized learning settings, communication is assumed to occur on only a given topology, while in many practical cases, the topology merely represents a preferred communication pattern, and connecting to arbitrary nodes is still possible. Previous studies have tried to alleviate the convergence rate degradation in these cases by designing topologies with large spectral gaps. However, the degradation is still significant when the number of nodes is substantial. In this chapter, we propose TELEPORTATION. TELEPORTATION activates only a subset of nodes, and the active nodes fetch the parameters from previous active nodes. Then, the active nodes update their parameters by SGD and perform gossip averaging on a relatively small topology comprising only the active nodes. We show that by activating only a proper number of nodes, TELEPORTATION can completely alleviate the convergence rate degradation. Furthermore, we propose an efficient hyperparameter-tuning method to search for the appropriate number of nodes to be activated. Experimentally, we showed that TELEPORTATION can train neural networks more stably and achieve higher accuracy than Decentralized SGD.

4.1 Introduction

While each node communicates with a limited number of nodes in every round, the degradation of the convergence rate due to its sparse communication characteristics is a trade-off (Koloskova et al., 2020). Especially, when the number of nodes is large, the parameters held by each node are likely to be far away during the training, and

gossip averaging deteriorates the convergence rate. This makes it challenging for decentralized learning to scale to a large number of nodes.

In decentralized learning literature, it is commonly assumed that communication can only occur on a given topology. However, in many practical cases, the topology merely represents a preferred communication pattern, and connecting to arbitrary nodes is still possible, e.g., in a data center or the case where nodes are connected on the Internet. Since we can choose which topology to use in these cases, many prior studies designed topologies to reconcile the communication efficiency and convergence rate, attempting to alleviate the degradation of the convergence rate caused by a large number of nodes (Wang et al., 2019; Ying et al., 2021; Ding et al., 2023; Takezawa et al., 2023b). Specifically, the communication costs are determined by the maximum degree of the topology, and the convergence rate is determined by its spectral gap (Wang et al., 2019; Koloskova et al., 2020). Thus, the prior studies designed topologies with large spectral gaps and small maximum degrees and proposed performing gossip averaging on these topologies. However, the approximate average estimated by gossip averaging deviates from the exact average of all nodes as the number of nodes increases, and the convergence rate remains to degrade when the number of nodes is substantial.

In this chapter, we ask the following question: *Can we develop a decentralized learning method whose convergence rate does not degrade as the number of nodes increases?* Our work yielded an affirmative answer with the proposal of TELEPORTATION. TELEPORTATION can completely eliminate the degradation of the convergence rate caused by increasing the number of nodes, and the convergence rate is consistently improved as the number of nodes increases. Specifically, TELEPORTATION activates only a subset of nodes and initializes their parameters with the parameters of previous active nodes. Then, the active node updates its parameters by SGD and performs gossip averaging on a relatively small topology comprising only the active nodes. By activating only a proper number of nodes, gossip averaging can make the parameters of each node reach the consensus fast, and the parameters of each node can be prevented from being far away even when the total number of nodes is large. We show that by activating an appropriate number of nodes, the degradation of the convergence rate can be completely alleviated. Furthermore, we propose an efficient hyperparameter-tuning method to search for the proper number of active nodes. For an arbitrary number of nodes, the proposed hyperparameter-tuning method can find the appropriate number of nodes to be activated within $2T$ iterations in total, where T is the number of iterations to run TELEPORTATION once. Experimentally, we demonstrated that TELEPORTATION can converge faster than Decentralized SGD and train neural networks more stably.

Note: In this chapter, we consider a setting where any two nodes can exchange parameters as required, similar to previous studies on topology design for decentralized learning (Marfoq et al., 2020; Ying et al., 2021; Song et al., 2022; Takezawa et al., 2023b; Ding et al., 2023). A data center is an example, and Assran et al. (2019)

demonstrated that decentralized methods can train neural networks faster than All-Reduce. In addition to a data center, when nodes are connected to the Internet, any two nodes can communicate by specifying the IP address. TELEPORTATION is not applicable in the presence of a pair of nodes that cannot communicate, as in wireless sensor networks.

4.2 Problem Setup

Notation: A graph G is represented by (V, E) where V is a set of nodes and E is a set of edges. For simplicity, we also write $(\{1, \dots, n\}, E)$ to represent a graph G where n is the number of nodes.

Problem Setting: In this chapter, we assume that the following assumptions are satisfied:

Assumption 1. *There exists $f^* > -\infty$ that satisfies $f(\mathbf{x}) \geq f^*$ for any $\mathbf{x} \in \mathbb{R}^d$.*

Assumption 2. *There exists $L \geq 0$ that satisfies for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $i \in [n]$,*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|. \quad (2.2)$$

Assumption 3. *It holds that $\mathbb{E}_{\xi_i \sim \mathcal{D}_i}[\nabla F_i(\mathbf{x}; \xi_i)] = \nabla f_i(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$. Then, there exists $\sigma \geq 0$ that satisfies for any $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$,*

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i} \|\nabla F_i(\mathbf{x}; \xi_i) - \nabla f_i(\mathbf{x})\|^2 \leq \sigma^2. \quad (2.3)$$

Assumption 4. *There exists $\zeta \geq 0$ that satisfies for any $\mathbf{x} \in \mathbb{R}^d$,*

$$\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \zeta^2. \quad (2.4)$$

Decentralized SGD: One of the most fundamental decentralized learning methods to solve Eq. (2.1) is Decentralized SGD (Lian et al., 2017). Let $\mathbf{x}_i \in \mathbb{R}^d$ denote the i -th node parameter. Once the topology comprising n nodes $G_n = (\{1, \dots, n\}, E)$ is specified, Decentralized SGD updates $\mathbf{x}_i$ as follows:

$$\mathbf{x}_i^{(t+1)} = \sum_{j=1}^n W_{ij} \left(\mathbf{x}_j^{(t)} - \eta \nabla F_j(\mathbf{x}_j^{(t)}; \xi_j^{(t)}) \right),$$

where $\eta > 0$ is the step size and $W_{ij} \in [0, 1]$ is the weight of edge $(i, j) \in E$ that satisfies $\sum_{i=1}^n W_{ij} = \sum_{j=1}^n W_{ij} = 1$. The following assumption is commonly used to represent how well the topology G_n is connected (Yuan et al., 2021; Koloskova et al., 2020, 2021; Aketi et al., 2023).

Assumption 8. Let λ_i is the i -th largest eigenvalue of $\mathbf{W} \in [0, 1]^{n \times n}$. $\mathbf{W}$ is a doubly stochastic matrix, and $p_n := 1 - \max\{|\lambda_2|, |\lambda_n|\}^2$ satisfies $p_n \in (0, 1]$. That is, it holds that

$$\|\mathbf{X}\mathbf{W} - \bar{\mathbf{X}}\|_F^2 \leq (1 - p_n) \|\mathbf{X} - \bar{\mathbf{X}}\|_F^2,$$

for any $\mathbf{X} \in \mathbb{R}^{d \times n}$ where $\bar{\mathbf{X}} := \frac{1}{n} \mathbf{X} \mathbf{1} \mathbf{1}^\top$.

We write p_n to emphasize that p_n depends on n . A small p_n indicates that the topology is poorly connected, and a large p_n indicates that the topology is well connected. Generally, p_n approaches zero as the number of node n increases. For instance, when a ring is used as the underlying topology, p_n is $\Omega(n^{-2})$, which rapidly decreases as n increases (Nedic et al., 2018). Under these assumptions, Decentralized SGD converges with the following rate.

Proposition 1 (Koloskova et al. (2020)). Suppose that Assumptions 1, 2, 3, 4, and 8 hold. Let $\{\{\mathbf{x}_i^{(t)}\}_{i=1}^n\}_{t=0}^T$ denote the parameters generated by Decentralized SGD. Then, there exists the step size η that satisfies:

$$\begin{aligned} & \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|^2 \\ & \leq \mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{L^2 r_0^2 (p_n \sigma^2 + \zeta^2) (1 - p_n)}{T^2 p_n^2} \right)^{\frac{1}{3}} + \frac{Lr_0}{Tp_n} \right), \end{aligned} \quad (4.1)$$

where $r_0 := f(\bar{\mathbf{x}}^{(0)}) - f^*$ and $\bar{\mathbf{x}}^{(t)} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(t)}$.

As the number of nodes n increases, the first term $\mathcal{O}(\sqrt{\frac{Lr_0\sigma^2}{nT}})$ is improved, while the second and third terms degrade since p_n reaches zero. Thus, as the number of nodes n increases substantially, the second and third terms dominate the convergence rate, and the convergence rate deteriorates. In many practical cases, G_n merely represents a preferred communication pattern, and any two nodes can communicate as needed, as introduced in Sec. 4.1. Since we can choose which topology to use in these cases, many recent studies have attempted to alleviate the convergence rate degradation by proposing topologies with large p_n (Ying et al., 2021; Ding et al., 2023; Takezawa et al., 2023b). However, their p_n still approach zero as n increases, and the convergence rate degrades when n is substantially large. Table C.1 in Sec. C.4 lists the convergence rates of Decentralized SGD with various topologies.

4.3 Proposed Method

In this section, we propose TELEPORTATION, which can completely eliminate the degradation of the convergence rate caused by increasing the number of nodes. In the remainder of this section, we describe TELEPORTATION in Sec. 4.3.1 and analyze its convergence rate in Sec. 4.3.2. Then, we propose an efficient hyperparameter search method in Sec. 4.3.3 and analyze its convergence rate in Sec. 4.3.4.

Algorithm 4 Simple version of TELEPORTATION

```

1: Input: the number of nodes  $n$ , set of nodes  $V_n$ , number of active nodes  $k$ , total number
   of iteration  $T$ , step size  $\eta$ , and topology comprising  $k$  active nodes  $G_k = (\{1, \dots, k\}, E)$ .
2: for  $t \in \{0, 1, \dots, T\}$  do
3:   sample next active  $k$  nodes  $V_{\text{active}}^{(t)}$  from  $V_n$  without replacement, and assign
    $\{1, 2, \dots, k\}$  to variables  $\{\text{token\_id}_i^{(t)} \mid v_i \in V_{\text{active}}^{(t)}\}$  randomly without overlap.
4:   for  $i \in \{1, 2, \dots, n\}$  in parallel do
5:     if  $v_i \in V_{\text{active}}^{(t)}$  thena
6:       if  $t > 0$  then
7:         Let  $v_j \in V_{\text{active}}^{(t-1)}$  denote the node such that  $\text{token\_id}_j^{(t-1)} = \text{token\_id}_i^{(t)}$ .
8:         receive  $\mathbf{x}_j^{(t)}$  from  $v_j \in V_{\text{active}}^{(t-1)}$ , and  $\mathbf{x}_i^{(t)} \leftarrow \mathbf{x}_j^{(t)}$ .
9:          $\mathbf{y}_i^{(t)} \leftarrow \mathbf{x}_i^{(t)} - \eta \nabla F_i(\mathbf{x}_i^{(t)}; \xi_i^{(t)})$ .
10:        receive  $\mathbf{y}_j^{(t)}$  from  $v_j \in \{v_j \in V_{\text{active}}^{(t)} \mid (\text{token\_id}_i^{(t)}, \text{token\_id}_j^{(t)}) \in E\}$ .
11:         $\mathbf{x}_i^{(t+1)} \leftarrow \sum_{v_j \in V_{\text{active}}^{(t)}} W_{\text{token\_id}_i^{(t)}, \text{token\_id}_j^{(t)}} \mathbf{y}_j^{(t)}$ .

```

^aThe update rule of the parameters of the inactive nodes is not described because the parameters held by the inactive nodes are discarded and initialized with the parameters of the other active nodes when they are activated in line 8. The parameters of inactive nodes do not affect the behavior of active node parameters.

4.3.1 Teleportation

p_n represents the speed at which the gossip averaging makes the parameter of each node $\{\mathbf{x}_i\}_{i=1}^n$ reach the average $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. As the number of nodes n increases, averaging all node parameters becomes difficult with gossip averaging, and p_n reaches zero. Thus, as n increases, the parameters of each node $\{\mathbf{x}_i\}_{i=1}^n$ comes to be apart during the training, and the convergence rate deteriorates, as discussed in Sec. 4.2.

To prevent the parameters in each node $\{\mathbf{x}_i\}_{i=1}^n$ from being away when the number of nodes n is large, we propose updating the parameters of only a small subset of nodes and performing gossip averaging on a small topology. We refer to the proposed method as TELEPORTATION.¹ Let $k \in \{1, \dots, n\}$ denote the number of active nodes and $G_k = (\{1, \dots, k\}, E)$ be the topology comprising k active nodes. Essentially, TELEPORTATION consists of the following three steps:

- (1) We sample k active nodes $V_{\text{active}}^{(t)}$ from V_n without replacement and assign $\{1, 2, \dots, k\}$ to variables $\{\text{token_id}_i^{(t)} \mid v_i \in V_{\text{active}}^{(t)}\}$ without overlap.²
- (2) The active node $v_i \in V_{\text{active}}^{(t)}$ fetches the parameters from the previous active node $v_j \in V_{\text{active}}^{(t-1)}$ whose $\text{token_id}_j^{(t-1)}$ stores the same value as that of $\text{token_id}_i^{(t)}$.

¹We named our proposed method TELEPORTATION to reflect the manner in which the active nodes copy their parameters to the next active nodes.

²The first step can be performed in a decentralized manner by sharing the same seed value to sample active nodes before starting the training.

Then, the active node v_i initializes its parameters with the fetched parameter and updates them by SGD.

- (3) The active node $v_i \in V_{\text{active}}^{(t)}$ exchanges its parameters with its neighbors $\{v_j \in V_{\text{active}}^{(t)} \mid (\text{token_id}_i^{(t)}, \text{token_id}_j^{(t)}) \in E\}$ and computes the weighted average with them.

We show the pseudo-code and illustration in Alg. 4 and Fig. 4.1. In the above implementation, the first step does not start until the third step is completed. We can solve this issue so that gossip averaging in the third step is performed on the next active node v_j , and the first and third steps can be executed simultaneously. We show the optimized version in Sec. C.1.

In TELEPORTATION, gossip averaging is performed on a relatively small topology G_k , which can make the parameters of k active nodes reach the average fast. Therefore, using the proper number of active nodes k , we can mitigate the degradation of the convergence rate by increasing the number of nodes n . In the next section, we analyze the convergence rate and show that our proposed method successfully circumvents the negative effect of increasing the number of nodes n .

Comparison with Client Sampling: TELEPORTATION randomly samples a subset of nodes and activates them, but TELEPORTATION is different from the client sampling (Liu et al., 2022). TELEPORTATION can prevent the parameters of nodes from being away by using the small k because gossip averaging is performed on the small topology G_k . In contrast to TELEPORTATION, client sampling incites the parameters of nodes to drift away because gossip averaging is performed on the topology comprising all nodes, and nodes can exchange parameters only when two neighboring nodes are sampled simultaneously, which makes it more difficult for gossip averaging to make the parameters reach the consensus. Thus, client sampling cannot alleviate the degradation of the convergence rate caused by a large number of nodes n and degrades the convergence rate as well as the vanilla Decentralized SGD (see Sec. C.5 for more detailed discussion).

4.3.2 Convergence Analysis of Teleportation

We assume that the topology comprising k active nodes G_k satisfies the following assumption.

Assumption 9. Let λ_i is the i -th largest eigenvalue of $\mathbf{W} \in [0, 1]^{k \times k}$. $\mathbf{W}$ is a doubly stochastic matrix, and $p_k := 1 - \max\{|\lambda_2|, |\lambda_k|\}^2$ satisfies $p_k \in (0, 1]$. That is, it holds that

$$\|\mathbf{X}\mathbf{W} - \bar{\mathbf{X}}\|_F^2 \leq (1 - p_k) \|\mathbf{X} - \bar{\mathbf{X}}\|_F^2, \quad (4.2)$$

for any $\mathbf{X} \in \mathbb{R}^{d \times k}$ where $\bar{\mathbf{X}} := \frac{1}{k} \mathbf{X} \mathbf{1} \mathbf{1}^\top$.

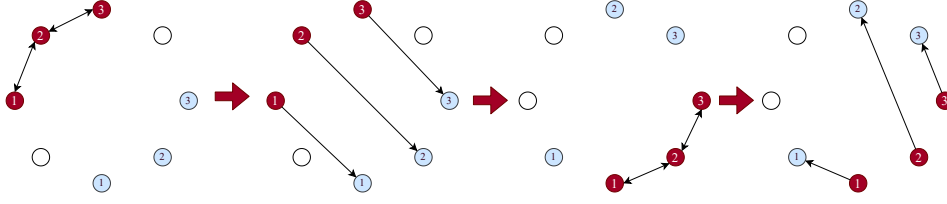


Figure 4.1: Illustration of Alg. 4 with $n = 8$ and $k = 3$. We use a line as the topology consisting of active nodes $G_k = (\{1, 2, 3\}, \{(1, 1), (1, 2), (2, 2), (2, 3), (3, 3)\})$. The black nodes represent active nodes, and the number written on the node is $\text{token_id}_i^{(t)}$. The blue nodes represent the next active nodes, and the number on the node is $\text{token_id}_i^{(t+1)}$. The first and third graphs from the left represent the communication in line 12, and the other graphs represent the communication in line 8.

Note that $\mathbf{W}$ is a $k \times k$ matrix, and p_k depends on the number of active nodes k , but is independent of the total number of nodes n . Under Assumptions 1, 2, 3, 4, and 9, Theorem 3 provides the convergence rate of TELEPORTATION for any number of active nodes k .

Theorem 3. Suppose that Assumptions 1, 2, 3, 4, and 9 hold. Let $\{\{\mathbf{x}_i^{(t)}\}_{i=1}^n\}_{t=0}^T$ denote the parameters generated by Alg. 4, and suppose that $\{\mathbf{x}_i^{(0)}\}_{i=1}^n$ is initialized with the same parameter $\bar{\mathbf{x}}^{(0)}$. Then, for any number of active nodes $k \in \{1, 2, \dots, n\}$, there exists the step size η such that $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\mathcal{O} \left(\sqrt{\frac{Lr_0(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}{kT}} + \left(\frac{L^2 r_0^2 (\sigma^2 + \zeta^2) (1 - p_k)}{T^2 p_k} \right)^{\frac{1}{3}} + \frac{Lr_0}{Tp_k} \right), \quad (4.3)$$

where $\bar{\mathbf{x}}_{\text{active}}^{(t)} := \frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \mathbf{x}_i^{(t)}$ and $r_0 := f(\bar{\mathbf{x}}^{(0)}) - f^*$.

By tuning the number of active nodes k , we obtain the following theorem. To determine the proper number of active nodes k , the explicit form of p_k is necessary. Here, we show the rates when we use a ring and exponential graph (Ying et al., 2021) as G_k .

Theorem 4. Suppose that Assumptions 1, 2, 3, 4, and 9 hold. Let $\{\{\mathbf{x}_i^{(t)}\}_{i=1}^n\}_{t=0}^T$ be the parameters generated by Alg. 4, and suppose that $\{\mathbf{x}_i^{(0)}\}_{i=1}^n$ is initialized with the same parameter $\bar{\mathbf{x}}^{(0)}$.

Ring: Suppose that the active nodes are connected by a ring, i.e., $p_k = \Omega(k^{-2})$. Then, if we set the number of active nodes k as follows:

$$k = \max \left\{ 1, \min \left\{ \left\lceil \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right)^{\frac{1}{7}} \right\rceil, n \right\} \right\},$$

there exists η such that $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{Lr_0(\sigma^2 + \zeta^2)^{\frac{3}{4}}}{T} \right)^{\frac{4}{7}} + \left(\frac{Lr_0(\sigma^2 + \zeta^2)^{\frac{2}{5}}}{T} \right)^{\frac{5}{7}} + \frac{Lr_0}{T} \right),$$

where $\bar{\mathbf{x}}_{\text{active}}^{(t)} := \frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \mathbf{x}_i^{(t)}$ and $r_0 := f(\bar{\mathbf{x}}^{(0)}) - f^*$.

Exp. Graph: Suppose that the active nodes are connected by an exponential graph, i.e., $p_k = \Omega(\log_2^{-1}(k))$. Then, if we set the number of active nodes k as follows:

$$k = \max \left\{ 1, \min \left\{ \left\lceil \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right)^{\frac{1}{3}} \right\rceil, \left\lceil \frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right\rceil, n \right\} \right\},$$

there exists η such that $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\begin{aligned} \mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{L^2 r_0^2 (\sigma^2 + \zeta^2)}{T^2} \log_2 \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right)^{\frac{1}{3}} \right)^{\frac{1}{3}} \right. \\ \left. + \frac{Lr_0}{T} \log_2 \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right) + \frac{Lr_0}{T} \right). \end{aligned}$$

All proofs are presented in Sec. C.2. The first term in Eq. (4.3) is worse than the first term in Eq. (4.1), but Theorem 4 indicates that the first term becomes the same as that of Decentralized SGD by carefully tuning k . For both cases, only the first term depends on the number of nodes n , which is improved as n increases. Therefore, TELEPORTATION can completely eliminate the negative effect by increasing the number of nodes n , and its convergence rate is consistently improved as n increases. As discussed in Sec. 4.2, several previous studies have attempted to avoid convergence rate degradation caused by large n by designing topologies with large spectral gaps (Ying et al., 2021; Song et al., 2022; Takezawa et al., 2023b). However, no prior topology, except for a complete graph, can completely remove this degradation. By comparing with the complete graph, TELEPORTATION is more communication efficient. For instance, if a ring is used as G_k , each node only needs to communicate with three nodes per iteration. Therefore, TELEPORTATION is the first method that does not suffer from large n without sacrificing the communication efficiency.

4.3.3 Efficient Hyperparameter Search for Number of Active Nodes

TELEPORTATION can eliminate the convergence rate degradation caused by a large number of nodes n , while the number of active nodes $k \in \{1, \dots, n\}$ needs to be tuned carefully. If k is tuned by grid search, it requires nT iterations in total. This hyperparameter-tuning becomes expensive as n increases. To alleviate this issue, we further extend TELEPORTATION by proposing an efficient hyperparameter-tuning

method for the number of active nodes k . The key idea for efficient hyperparameter-tuning is based on the following lemma.

Lemma 1. *Let $k^* \in \{1, 2, 3, \dots, n\}$ be the optimal number of active nodes. If $k^* < n$, there exists $k \in \{1, 2, 4, 8, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}\}$ that satisfies $\frac{k^*}{4} < k \leq k^*$. Furthermore, it holds that $\sum_{i=0}^{\lfloor \log_2(n+1) \rfloor - 1} 2^i \leq n$.*

Lemma 1 shows that for any optimal number of active nodes $k^* \leq n - 1$, a similar number of active nodes is contained in $\{1, 2, 2^2, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}\}$. This suggests that we only need to search from $\{1, 2, 2^2, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}, n\}$ to obtain the appropriate k and we do not need run Alg. 4 for all $k \in \{1, 2, 3, \dots, n\}$. Moreover, Lemma 1 indicates that Alg. 4 can be run with various $k \in \{1, 2, 2^2, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}\}$ in parallel because the total number of active nodes is less than or equal to n . Based on these observations, we propose an efficient hyperparameter-tuning method. The pseudo-code is shown in Alg. 5.

4.3.4 Convergence Analysis of Teleportation with Efficient Hyperparameter Search

Under the same assumptions as in Theorem 4, Theorem 5 provides the convergence rate of Alg. 5. The proof is deferred to Sec. C.3.

Theorem 5. *Suppose that Assumptions 1, 2, 3, 4, and 9 hold. Let $\{\{\mathbf{x}_{k,i}^{(t)}\}_{v_i \in V_{\text{active}}^{(t)}}\}_{t=0}^T$ denote the parameters of active nodes generated by Alg. 4 when the number of active nodes is set to k , and we define*

$$\mathcal{K} := \{1, 2, 2^2, 2^3, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}, n\}.$$

Then, suppose that the parameters are initialized with the same parameter $\bar{\mathbf{x}}^{(0)}$.

Ring: *If the active nodes are connected by a ring, i.e., $p_k = \Omega(k^{-2})$, there exists*

Algorithm 5 Efficient hyperparameter search for TELEPORTATION.

- 1: **Input:** the total number of iteration $2T$, and the number of nodes n .
- 2: run Alg. 4 for T iterations with number of active nodes n , and let $\{\{\mathbf{x}_{n,i}^{(t)}\}_{i=1}^n\}_{t=0}^T$ denote the generated parameters.
- 3: **for** $k \in \{1, 2, 2^2, 2^3, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}\}$ **in parallel do**
- 4: run Alg. 4 for T iterations with number of active nodes k .
- 5: let $\{\{\mathbf{x}_{k,i}^{(t)}\}_{v_i \in V_{\text{active}}^{(t)}}\}_{t=0}^T$ denote the parameters generated by Alg. 4 with k active nodes.
- 6: **return** the best^b parameters among $k \in \{1, 2, 2^2, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}, n\}$.

^bTheoretically, we select the parameters with the smallest gradient norm as in Theorem 4. In practice, we can select k that achieves the best accuracy on the validation datasets as in the vanilla grid search.

η such that $\min_{k \in \mathcal{K}} \left(\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{active,k}^{(t)})\|^2 \right)$ is bounded from above by

$$\mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{Lr_0(\sigma^2 + \zeta^2)^{\frac{3}{4}}}{T} \right)^{\frac{4}{7}} + \left(\frac{Lr_0(\sigma^2 + \zeta^2)^{\frac{2}{5}}}{T} \right)^{\frac{5}{7}} + \frac{Lr_0}{T} \right),$$

where $r_0 := f(\bar{\mathbf{x}}^{(0)}) - f^*$ and $\bar{\mathbf{x}}_{active,k}^{(t)} := \frac{1}{k} \sum_{v_i \in V_{active}^{(t)}} \mathbf{x}_{k,i}^{(t)}$.

Exp. Graph: If the active nodes are connected by an exponential graph, i.e., $p_k = \Omega(\log_2^{-1} k)$, there exists η such that $\min_{k \in \mathcal{K}} \left(\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{active,k}^{(t)})\|^2 \right)$ is bounded from above by

$$\begin{aligned} \mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{L^2 r_0^2 (\sigma^2 + \zeta^2)}{T^2} \log_2 \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right)^{\frac{1}{3}} \right)^{\frac{1}{3}} \right. \\ \left. + \frac{Lr_0}{T} \log_2 \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right) + \frac{Lr_0}{T} \right). \end{aligned}$$

Theorem 5 shows that Alg. 5 can achieve exactly the same convergence rate as the one shown in Theorem 4 (i.e., the convergence rate with optimal k). Algorithm 5 requires only $2T$ iterations to find the proper number of active nodes, whereas grid search requires nT iterations in total. Thus, Alg. 5 can find the appropriate number of active nodes more efficiently than the vanilla grid search.

4.4 Related Work

Token Algorithm: Token algorithms (Johansson et al., 2007; Dorfman and Levy, 2022; Even, 2023; Hendrikx, 2023) are a variant of decentralized learning methods different from consensus-based decentralized learning methods, such as Decentralized SGD. In contrast to the consensus-based decentralized learning methods, a parameter called token randomly walks on the topology and is updated on each node. Since token algorithms have only one parameter, they do not suffer from the issue of parameters held by each node drifting away. However, their convergence rates cannot achieve linear speedup as that in Decentralized SGD (Even, 2023). In TELEPORTATION, k parameters randomly move from previous active nodes to the next active nodes and are updated on the active nodes. Then, by carefully selecting k , TELEPORTATION can alleviate the issue of parameter drifting without sacrificing linear speedup property.

Client Sampling: Client sampling is widely studied in federated learning to reduce the communication costs between the central server and nodes (McMahan et al., 2017; Fraboni et al., 2021; Wu et al., 2023). McMahan et al. (2017) proposed sampling a subset of nodes randomly. Cho et al. (2020) proposed selecting nodes according to the loss values. Chen et al. (2022) and Wang et al. (2023) proposed sampling

nodes according to their gradient norms. In decentralized learning, Liu et al. (2022) studied client sampling and analyzed the convergence rate. However, as discussed in Sec. 4.3.1, the client sampling cannot alleviate the degradation of the convergence rate caused by a large number of nodes.

4.5 Experiments

4.5.1 Synthetic Experiment

Setting: We followed the experimental setting in Koloskova et al. (2020) and set the number of nodes n to 100 and loss function as $f_i(\mathbf{x}) := \frac{1}{2} \|\mathbf{A}_i(\mathbf{x} - \mathbf{b}_i)\|^2$ with $\mathbf{A}_i := \frac{i}{\sqrt{n}} \mathbf{I}_d$ and $d = 50$. $\mathbf{b}_i$ was drawn from $\mathcal{N}(\mathbf{0}, \frac{\zeta^2}{i^2} \mathbf{I}_d)$ for each node. We defined the stochastic gradient as $\nabla F_i(\mathbf{x}; \xi) := \nabla f_i(\mathbf{x}) + \epsilon$ where ϵ was drawn from $\mathcal{N}(\mathbf{0}, \frac{\sigma^2}{d} \mathbf{I}_d)$ at each iteration. We used a ring and Base-2 Graph (Takezawa et al., 2023b) as the

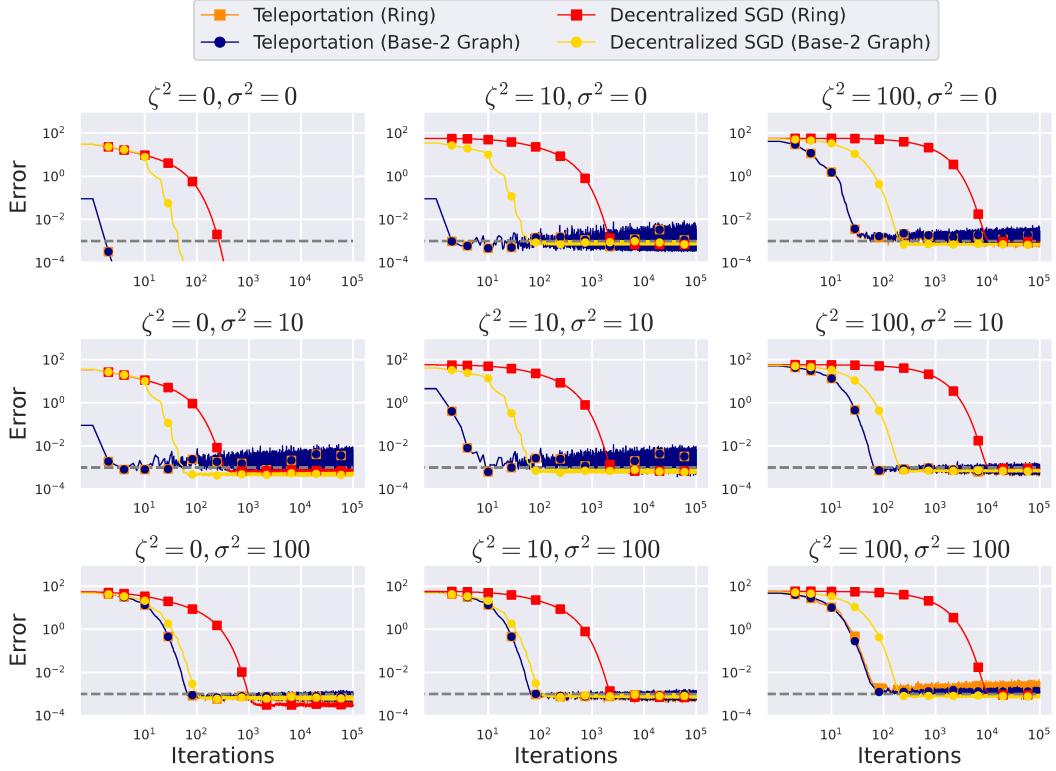


Figure 4.2: Convergence of the error to the target accuracy 0.001 for different stochastic noise σ^2 and heterogeneity ζ^2 . We plotted $\frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \|\mathbf{x}_i^{(t)} - \mathbf{x}^*\|^2$ and $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(t)} - \mathbf{x}^*\|^2$ as the error for TELEPORTATION and Decentralized SGD, respectively. TELEPORTATION consistently reached the target accuracy faster than Decentralized SGD.

topology. The ring is one of the most commonly used topologies for decentralized learning. The Base-2 Graph is the state-of-the-art topology for Decentralized SGD, and Takezawa et al. (2023b) demonstrated that the Base-2 Graph can be superior to the various topologies, e.g., an exponential graph, 1-peer exponential graph (Ying et al., 2021), and EquiTopo (Song et al., 2022). Note that the Base-2 Graph assumes that any two nodes can communicate as in TELEPORTATION. For each setting, we tuned the step size to reach the target accuracy as few iterations as possible. See Sec. C.7 for a more detailed setting.

Results: We depict the results in Fig. 4.2. For all cases, TELEPORTATION required fewer iterations to reach the target accuracy than Decentralized SGD. By comparing Decentralized SGD with a ring, TELEPORTATION converged up to three orders of magnitude faster than Decentralized SGD. Even in the case when the state-of-the-art topology, Base-2 Graph, is used, TELEPORTATION converged up to 10 times faster than Decentralized SGD. Table 4.1 lists the number of active nodes k selected by Alg. 5. Although the total number of nodes n is 100, a small number of nodes k was selected. Therefore, activating only a few nodes can prevent the parameters from being far away and lead to a faster convergence rate than that of Decentralized SGD.

4.5.2 Neural Networks

Setting: We used Fashion MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky, 2009) as datasets and used LeNet (LeCun et al., 1998) and VGG (Simonyan and Zisserman, 2015) as neural networks, respectively. To use the momentum in TELEPORTATION, the momentum is copied from the previous active nodes to the next active nodes, as well as the parameters. We set the number of nodes n to 25 and distributed the data to nodes using Dirichlet distribution with parameter α (Hsu et al., 2019). As α approaches zero, each node comes to have a different dataset. We repeated all experiments with three different seed values, reporting the averages. See Sec. C.7 for a more detailed setting.

Results: We depict the results in Fig. 4.3. When $\alpha = 0.1$, TELEPORTATION outperformed Decentralized SGD and trained the neural networks more stably. When $\alpha = 10.0$, all comparison methods achieved a competitive accuracy. By

Table 4.1: The number of active nodes k selected by Alg. 5 in Fig. 4.2. The left value is the number of active nodes when the topology is a ring, and the right value is the number when the topology is the Base-2 Graph.

	$\zeta^2 = 0$	$\zeta^2 = 10$	$\zeta^2 = 100$
$\sigma^2 = 0$	1 / 1	1 / 1	4 / 4
$\sigma^2 = 10$	1 / 1	1 / 1	8 / 8
$\sigma^2 = 100$	8 / 8	8 / 8	4 / 16

comparing the results with $\alpha = 0.1$ and 10.0, the accuracy curves of Decentralized SGD became unstable in the heterogeneous setting, whereas the accuracy curves of TELEPORTATION were stable in both cases. This is because Decentralized SGD suffers from client drift, while TELEPORTATION can suppress it by initializing the parameters of active nodes by using the parameters of other nodes.

4.5.3 Comparison under Heterogeneous Networks

Next, we examine the effectiveness of TELEPORTATION in terms of wallclock time. In TELEPORTATION, active nodes send the parameters to the next active nodes, which are randomly sampled from the entire set of nodes. Thus, the communication may happen between distant nodes, e.g., nodes located in distant regions. In this section, we evaluate TELEPORTATION under the heterogeneous networks where communication costs differ between pairs of nodes.

Setting: We used Fashion MNIST and LeNet as with Sec. 4.5.2. To simulate a heterogeneous network, we added $\tau \times \text{mod}(|i - j|, 24) \mu\text{s}$ delay when nodes i and j communicate, where $\text{mod}(a, b)$ is the remainder of dividing a by b . We constructed a ring by connecting node i to node $(1 + \text{mod}(i, 25))$. This ring is the optimal topology

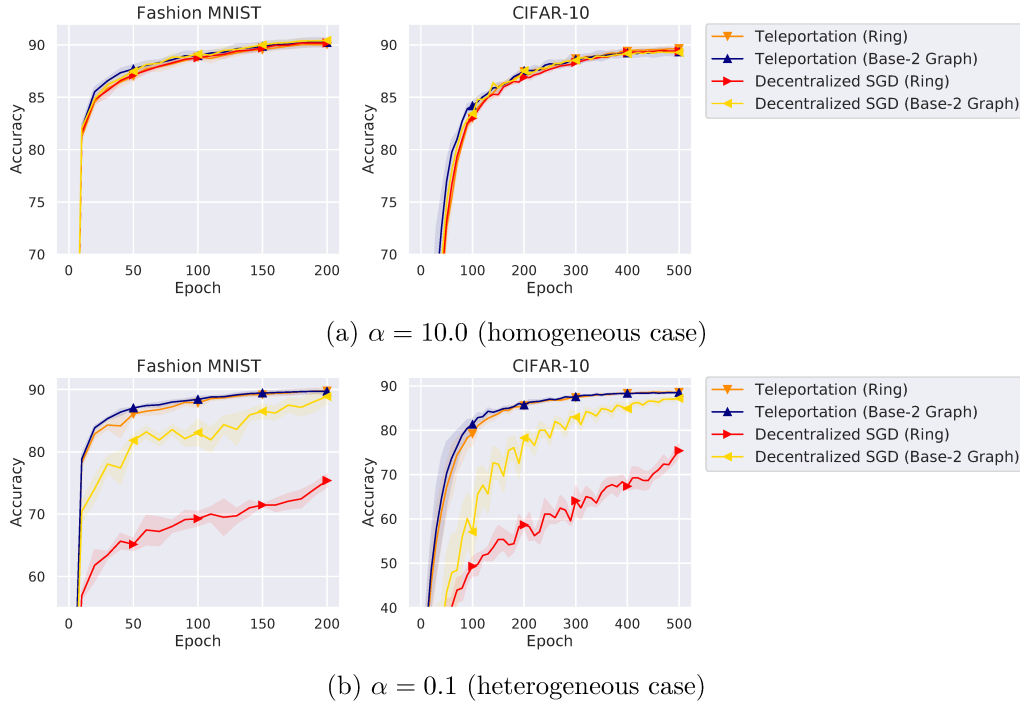


Figure 4.3: Test accuracy of TELEPORTATION and Decentralized SGD for different heterogeneity. All methods achieved competitive accuracy in the homogeneous case, while TELEPORTATION outperformed Decentralized SGD in the heterogeneous case.

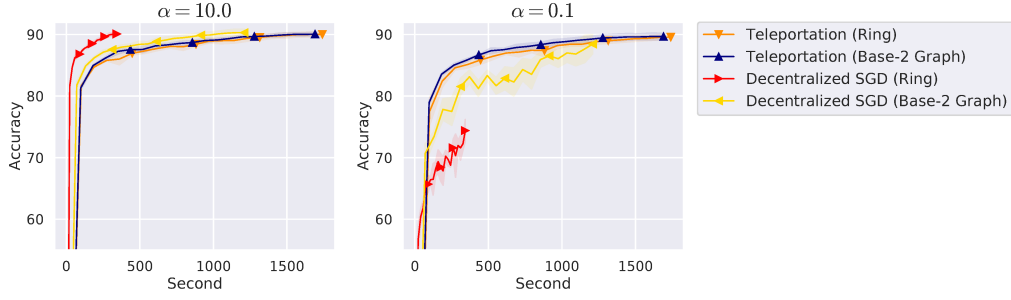


Figure 4.4: Test accuracy of TELEPORTATION and Decentralized SGD under the heterogeneous networks with $\tau = 5$. Decentralized SGD with the ring reached a high accuracy faster than the other methods in the homogeneous case, while TELEPORTATION reached a high accuracy faster in the heterogeneous case. Note that the number of epochs was set the same for all methods.

in terms of the communication delay, and the communication in Decentralized SGD with the ring was delayed by $\tau \mu s$, whereas the communication in TELEPORTATION was delayed by at most $24\tau \mu s$. The communication in Decentralized SGD with the Base-2 Graph was also delayed by at most $24\tau \mu s$ since the Base-2 Graph assumes that any two nodes can communicate, and communication occurs between various nodes.

Results: We depict the results with $\tau = 5$ in Fig. 4.4. In the Appendix, we also show the results with $\tau = 0$ in Fig. C.3. Decentralized SGD with the ring finished the training faster than Decentralized SGD with the Base-2 Graph and TELEPORTATION. However, when $\alpha = 0.1$, the accuracy of Decentralized SGD with the ring increased very slowly, and TELEPORTATION reached a high accuracy faster than Decentralized SGD. When $\alpha = 10.0$, Decentralized SGD with the ring reached a high accuracy faster than the other methods. Therefore, Decentralized SGD with a sparse topology is the preferred method when the data distributions are homogeneous, while when the data distributions are heterogeneous, TELEPORTATION can be a preferred method even if the network is heterogeneous since the methods that can prevent the parameter drifting are necessary to achieve high accuracy.

4.6 Conclusion

In this chapter, we propose TELEPORTATION, which activates a subset of nodes and performs gossip averaging on a relatively small topology comprising only active nodes. We showed that TELEPORTATION converges to the stationary point without suffering from a large number of nodes by activating a proper number of nodes. Furthermore, we proposed an efficient hyperparameter tuning method to search for this appropriate number of nodes. We experimentally investigated the effectiveness of TELEPORTATION, demonstrating that TELEPORTATION can converge faster than Decentralized SGD and train neural networks more stably in the heterogeneous case.

5

Exploiting Similarity for Computation and Communication-efficient Decentralized Optimization

Abstract

Reducing communication complexity is critical for efficient decentralized optimization. The proximal decentralized optimization (PDO) framework is particularly appealing, as methods within this framework can exploit functional similarity among nodes to reduce communication rounds. Specifically, when local functions at different nodes are similar, these methods achieve faster convergence with fewer communication steps. However, existing PDO methods often require highly accurate solutions to subproblems associated with the proximal operator, resulting in significant computational overhead. In this chapter, we propose the Stabilized Proximal Decentralized Optimization (SPDO) method, which achieves state-of-the-art communication and computational complexities within the PDO framework. Additionally, we refine the analysis of existing PDO methods by relaxing subproblem accuracy requirements and leveraging average functional similarity. Experimental results demonstrate that SPDO significantly outperforms existing methods.

5.1 Introduction

One of the key techniques to reduce the communication complexity is for each node to perform multiple local updates before the communication. The concept of multiple local updates was initially applied for Decentralized Gradient Descent (Nedic and Ozdaglar, 2009; Lian et al., 2017) and Gradient Tracking (Lorenzo and Scutari, 2016; Pu and Nedic, 2021). However, increasing the number of local steps requires a smaller learning rate, which ultimately requires the same communication cost to achieve the desired accuracy level (Koloskova et al., 2020; Liu et al., 2024). This is because,

Table 5.1: Summary of communication and computational complexities in the strongly-convex setting. The results in the convex case are summarized in Sec. D.1. “# communication” represents the number of times that each node exchanges parameters with neighboring nodes to reach the target accuracy ϵ , and “# computation” represents the number of gradient oracle queries required by each node. Since Tian et al. (2022) only showed the convergence rate of Accelerated Exact/Inexact-PDO with $\mu \leq \delta_{\max}$, we show the convergence rates under the assumption that $\mu \leq \delta_{\max}$ and $\mu \leq \delta$ for simplicity.

Algorithm	Reference	# Communication	# Computation
Gradient Tracking	Koloskova et al. (2021)	$\tilde{\mathcal{O}}\left(\frac{L}{\mu(1-\rho)^2} \log\left(\frac{1}{\epsilon}\right)\right)^{(a)}$	$\tilde{\mathcal{O}}\left(\frac{L}{\mu(1-\rho)^2} \log\left(\frac{1}{\epsilon}\right)\right)^{(a)}$
Exact-PDO (SONATA)	Sun et al. (2022), Th. 6 ^(b)	$\tilde{\mathcal{O}}\left(\frac{\delta}{\mu(1-\rho)} \log\left(\frac{1}{\epsilon}\right)\right)^{(b)}$	n.a. ^(c)
Inexact-PDO	Th. 6, 7	$\tilde{\mathcal{O}}\left(\frac{\delta}{\mu(1-\rho)} \log\left(\frac{1}{\epsilon}\right)\right)$	$\tilde{\mathcal{O}}\left(\frac{\sqrt{\delta L}}{\mu(1-\rho)} \log\left(\frac{1}{\epsilon}\right) \log \log\left(\frac{1}{\epsilon}\right)\right)$
Stabilized-PDO [new]	Th. 8, 9	$\tilde{\mathcal{O}}\left(\frac{\delta}{\mu(1-\rho)} \log\left(\frac{1}{\epsilon}\right)\right)$	$\tilde{\mathcal{O}}\left(\frac{\sqrt{\delta L}}{\mu(1-\rho)} \log\left(\frac{1}{\epsilon}\right)\right)$
Accelerated SONATA	Tian et al. (2022)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{\delta_{\max}}{\mu(1-\rho)}} \log\left(\frac{1}{\epsilon}\right)\right)^{(d)}$	n.a. ^(c)
Inexact Accelerated SONATA	Tian et al. (2022)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{\delta_{\max}}{\mu(1-\rho)}} \log\left(\frac{1}{\epsilon}\right)\right)^{(d)}$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{\delta_{\max} L}{\mu(1-\rho)}} \left(\log\left(\frac{1}{\epsilon}\right)\right)^2\right)^{(d)}$
Accelerated Stabilized-PDO [new]	Th. 10, 11	$\tilde{\mathcal{O}}\left(\sqrt{\frac{\delta}{\mu(1-\rho)}} \log\left(\frac{1}{\epsilon}\right)\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu(1-\rho)}} \log\left(\frac{1}{\epsilon}\right)\right)$

(a) It holds that $\delta \leq L$, and $\delta \ll L$ holds when nodes have similar datasets.

(b) Sun et al. (2022) showed that the communication complexity of SONATA is $\tilde{\mathcal{O}}\left(\frac{\delta_{\max}}{\mu(1-\rho)} \log\left(\frac{1}{\epsilon}\right)\right)$, while our new theorem, Theorem 6, shows that the dependence on $\delta_{\max}$ can be reduced to δ .

(c) “n.a.” represents that we need to solve the subproblem exactly.

(d) δ is smaller than $\delta_{\max}$ and can be at most $\sqrt{n}$ times smaller than $\delta_{\max}$.

in Decentralized Gradient Descent, each node tries to minimize its own local loss function by performing gradient descent multiple times, but fully minimizing its local loss function is not desirable. Gradient Tracking also has the same issue, and this inconsistency hinders the use of multiple local updates to reduce communication complexity. Thus, to take advantage of multiple local updates, we need to resolve this inconsistency.

An effective approach to address this challenge is to develop decentralized optimization methods based on the proximal-point method, where nodes solve the subproblem associated with the proximal operator in their local updates (termed PDO framework in this work). Scutari and Sun (2019) demonstrated that combining proximal-point updates with gradient tracking yields an efficient decentralized optimization scheme, known as SONATA. This method effectively exploits the similarity of local loss functions across nodes to significantly reduce communication complexity (Sun et al., 2022; Tian et al., 2022). However, achieving such low communication complexity requires solving the proximal subproblem with high accuracy at each node, leading to suboptimal computational overhead during local updates.

In this chapter, we pose the following question: *Can we develop a proximal decentralized optimization (PDO) method that achieves low communication complexity without sacrificing computational efficiency?* We provide an affirmative answer by introducing the STABILIZED PROXIMAL DECENTRALIZED OPTIMIZATION (SPDO)

method. SPDO improves upon existing approaches in several key aspects: (i) it requires less accurate solutions to local subproblems, resulting in better computational efficiency, and (ii) it leverages the *average* functional similarity across nodes, leading to more effective optimization. Furthermore, (iii) the accelerated variant of our method, ACCELERATED-SPDO, achieves state-of-the-art communication and computational complexities among decentralized optimization methods, as summarized in Table 5.1. More specifically, we make the following contributions.

- We propose Accelerated-SPDO, which can attain the best communication and computational complexities among the existing methods.
- We also propose non-accelerated methods, called SPDO, which can attain the best communication and computational complexities among the existing non-accelerated methods.
- To clarify the difference between (Accelerated-)SPDO and the existing method, we provide a refined convergence analysis of PDO, which contains Exact-PDO, also known as SONATA (Sun et al., 2022), as a special instance. Specifically, Exact-PDO assumed that the subproblem was exactly solved, but PDO allows us to solve it approximately. We then show that SPDO can solve the subproblem more coarsely than Inexact-PDO.
- The existing analysis of Exact-PDO showed that the communication complexity depends on the maximum dissimilarity of local functions $\delta_{\max}$ (Sun et al., 2022), while we prove that it depends only on the average dissimilarity δ , which can lead to significantly lower communication costs.
- Moreover, the prior papers on PDO analyzed the convergence rate only in the strongly-convex case, while we analyze it in both convex and strongly-convex cases.

Notation: We use $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to hide polylogarithmic factors.

5.2 Related Work

Decentralized Optimization: The most basic decentralized optimization method is Decentralized Gradient Descent (Nedic and Ozdaglar, 2009; Lian et al., 2017; Koloskova et al., 2020), while due to the data heterogeneity, Decentralized Gradient Descent cannot achieve the linear convergence rate in the strongly-convex case. Many researchers have tried to achieve the linear convergence rate, proposing EXTRA (Shi et al., 2015), NIDS (Li et al., 2019), Gradient Tracking (Lorenzo and Scutari, 2016; Pu and Nedic, 2021; Yuan et al., 2019; Takezawa et al., 2023a), and so on (Scaman et al., 2017; Kovalev et al., 2021). These methods can achieve the linear convergence rate, while they still involve a certain amount of communication cost even though nodes have very similar loss functions. The concept of second-order

Algorithm 6 Proximal Decentralized Optimization Method (PDO)

-
- 1: **Input:** The number of rounds R , and hyperparameters M and λ .
 - 2: **for** $r = 0, 1, \dots, R - 1$ **do**
 - 3: **for** $i = 1, \dots, n$ in parallel **do**
 - 4: $\mathbf{x}_i^{(r+\frac{1}{2})} \approx \arg \min_{\mathbf{x}} F_{i,r}(\mathbf{x})$ where $F_{i,r}(\mathbf{x}) := f_i(\mathbf{x}) + \langle \mathbf{h}_i^{(r)}, \mathbf{x} \rangle + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}_i^{(r)}\|^2$.
 - 5: $\mathbf{x}_i^{(r+1)} = \text{MULTIGOSSIP}(\mathbf{x}_i^{(r+\frac{1}{2})}, \mathbf{W}, M, i)$.
 - 6: $\mathbf{h}_i^{(r+1)} = \text{MULTIGOSSIP}(\mathbf{h}_i^{(r)} + \nabla f_i(\mathbf{x}_i^{(r+1)}), \mathbf{W}, M, i) - \nabla f_i(\mathbf{x}_i^{(r+1)})$.
-

similarity was introduced to capture this phenomenon, in which δ and $\delta_{\max}$ measure the similarity among the loss functions held by nodes (see Definition 2 and Assumption 10) (Karimireddy et al., 2021; Murata and Suzuki, 2021; Kovalev et al., 2022). δ and $\delta_{\max}$ are smaller than L , where L is the smoothness of the loss function (see Assumption 2), and δ and $\delta_{\max}$ approach zero when the local loss functions come to be similar. To exploit this similarity, many variants of the proximal decentralized optimization method have been proposed, such as Network-DANE (Li et al., 2020), SONATA (Sun et al., 2022), and Accelerated SONATA (Tian et al., 2022).

Proximal-point Methods: In this section, we briefly introduce Proximal-point Method. Applying Proximal-point Method to minimize f yields the following update rule:

$$\mathbf{x}^{(r+1)} = \arg \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ F_r(\mathbf{x}) := f(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}^{(r)}\|^2 \right\},$$

where $\lambda > 0$ is hyperparameter. For the vanilla Proximal-point Method, we need to solve the above subproblem exactly, which is infeasible in general. Many researchers have tried to avoid the exact proximal evaluations (Rockafellar, 1976; Solodov and Svaiter, 1999, 2001; Monteiro and Svaiter, 2013). Rockafellar (1976) analyzed the condition to maintain the same convergence rate, showing Proximal-point Method can attain the same convergence rate when $\|\nabla F_r(\mathbf{x}^{(r+1)})\| \leq \mathcal{O}(\frac{\lambda}{r} \|\mathbf{x}^{(r+1)} - \mathbf{x}^{(r)}\|)$ is satisfied. That is, we need to solve the subproblem more accurately as the number of iterations increases. To mitigate this issue, Solodov and Svaiter (1999) proposed Hybrid Projection Proximal-point Method, which can achieve the same convergence rate if $\|\nabla F_r(\mathbf{x}^{(r+1)})\| \leq \mathcal{O}(\lambda \|\mathbf{x}^{(r+1)} - \mathbf{x}^{(r)}\|)$ is satisfied. This condition is significantly weaker than the condition for the vanilla Proximal-point Method.

The accelerated Proximal-point Method (Güler, 1992) also has the same issue as Proximal-point Method, and it requires a more accurate subproblem solution as the number of iterations increases. Based on Hybrid Projection Proximal-point Method, Monteiro and Svaiter (2013) proposed its accelerated method and relaxed the condition for an approximate solution of the subproblem.

A recent line of work developed distributed versions of the Proximal-point Method (Shamir and Srebro, 2014; Li et al., 2020; Sun et al., 2022; Jiang et al., 2024a,b). We will further discuss these approaches in Sec. 5.6 below.

Algorithm 7 Multiple Gossip Averaging

```

1: function MULTIGOSSIP( $\{\mathbf{a}_i\}_{i=1}^n, \mathbf{W}, M, i$ )
2:    $\mathbf{a}_i^{(0)} = \mathbf{a}_i$  for all  $i$ .
3:   for  $m = 0, \dots, M - 1$  do
4:      $\mathbf{a}_i^{(m+1)} = \sum_{j=1}^n W_{ij} \mathbf{a}_j^{(r)}$ .
5:   return  $\mathbf{a}_i^{(M)}$ 

```

5.3 Problem Setup

In this chapter, we assume the following:

Assumption 1. *There exists $f^* > -\infty$ that satisfies $f(\mathbf{x}) \geq f^*$ for any $\mathbf{x} \in \mathbb{R}^d$.*

Assumption 2. *There exists $L \geq 0$ that satisfies for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $i \in [n]$,*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|. \quad (2.2)$$

Assumption 5. *There exists $\mu \geq 0$ such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and i , it holds that*

$$f_i(\mathbf{x}) \geq f_i(\mathbf{y}) + \langle \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2. \quad (2.5)$$

To measure the similarity among the loss functions held by nodes, we use the following notation as in the prior work (Jiang et al., 2024b,a).

Definition 2 (Similarity). *Under Assumption 2, let $\delta \geq 0$ be the smallest number such that it holds that*

$$\frac{1}{n} \sum_{i=1}^n \|\nabla h_i(\mathbf{x}) - \nabla h_i(\mathbf{y})\|^2 \leq \delta^2 \|\mathbf{x} - \mathbf{y}\|^2, \quad (5.1)$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ where $h_i(\mathbf{x}) := f(\mathbf{x}) - f_i(\mathbf{x})$.

Remark 1. *If Assumption 2 holds, there exists δ such that $\delta \leq L$ and Eq. (5.1) holds.*

Apart from Definition 2, Assumption 10 is also commonly used in the existing literature (Tian et al., 2022; Sun et al., 2022). While the existence of δ follows from Assumption 2, Assumption 10 further needs to assume that f_i is twice differentiable. Furthermore, δ can be much smaller than $\delta_{\max}$ (see Remark 2). Thus, we use Definition 2 instead of Assumption 10 in this chapter.

Assumption 10. *There exists $\delta_{\max} \geq 0$ such that for any $\mathbf{x} \in \mathbb{R}^d$ and i , it holds that*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f_i(\mathbf{x})\| \leq \delta_{\max}. \quad (5.2)$$

Remark 2. *It holds that $\delta \leq \delta_{\max}$. Moreover, δ can be at most $\sqrt{n}$ times smaller than $\delta_{\max}$.*

Then, we assume that edge weights $\{W_{ij}\}_{ij}$ satisfy the following assumption, in which ρ represents how well the graph is connected.

Assumption 11. *For any i and j , $W_{ij} = W_{ji}$ and $\sum_{i=1}^n W_{ij} = 1$. Then, there exists $\rho \in [0, 1)$ such that for any $\mathbf{X} \in \mathbb{R}^{d \times n}$, it holds that*

$$\|\mathbf{X}\mathbf{W} - \bar{\mathbf{X}}\|_F^2 \leq \rho^2 \|\mathbf{X} - \bar{\mathbf{X}}\|_F^2, \quad (5.3)$$

where $\bar{\mathbf{X}} := \frac{1}{n} \mathbf{X} \mathbf{1} \mathbf{1}^\top$.

Evaluation Metric: As we introduced in Sec. 5.1, the communication is often more expensive than the local computation. Thus, the primary goal of developing decentralized optimization methods is to reduce communication complexity. If methods can achieve the same communication complexity, the method with less computational complexity is superior.

5.4 Refined Analysis of Proximal Decentralized Optimization Method

In this section, we briefly explain the framework of Proximal Decentralized Optimization (PDO) methods and provide the convergence analysis. PDO contains SONATA (Sun et al., 2022) as a special instance when the proximal subproblem is solved exactly. The proofs are deferred to Sec. D.3 and D.6.2.

In decentralized optimization, each node would like to minimize f , but node i can access only the local loss function f_i . For PDO, each node approximates f as follows:

$$\begin{aligned} f(\mathbf{x}) &= f_i(\mathbf{x}) + h_i(\mathbf{x}) \\ &\approx f_i(\mathbf{x}) + \langle \nabla h_i(\mathbf{x}_i^{(r)}), \mathbf{x} \rangle + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}_i^{(r)}\|^2, \end{aligned} \quad (5.4)$$

where $\lambda \geq 0$ is the hyperparameter and we approximate $h_i(\mathbf{x})$ by linear approximation. Furthermore, PDO cannot compute $\nabla h_i(\mathbf{x}_i^{(r)})$ since node i can exchange parameters only with its neighbors. Thus, PDO estimates $\nabla h_i(\mathbf{x}_i^{(r)})$ by gradient tracking. We show the pseudo-code in Alg. 6. In line 4, PDO solves the subproblem associated with Eq. (5.4), and in lines 5 and 6, PDO estimates $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(r+\frac{1}{2})}$ and $\nabla h_i(\mathbf{x}_i^{(r+1)})$ by gossip averaging and gradient tracking.

Arbitrary Approximate Averaging Operators: While Alg. 6 is demonstrated using Alg. 7 as a subroutine, it is important to emphasize that the PDO framework is not limited to this specific mechanism. Instead, PDO can be combined with any arbitrary averaging operator that achieves a sufficient level of accuracy.

Algorithm 8 Stabilized Proximal Decentralized Optimization Method (SPDO)

-
- 1: **Input:** The number of rounds R , and hyperparameters M and λ .
 - 2: **for** $r = 0, 1, \dots, R - 1$ **do**
 - 3: **for** $i = 1, \dots, n$ **in parallel do**
 - 4: $\mathbf{x}_i^{(r+1)} \approx \arg \min_{\mathbf{x}} F_{i,r}(\mathbf{x})$ where $F_{i,r}(\mathbf{x}) := f_i(\mathbf{x}) + \langle \mathbf{h}_i^{(r)}, \mathbf{x} \rangle + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{v}_i^{(r)}\|^2$.
 - 5: $\mathbf{v}_i^{(r+\frac{1}{2})} = \arg \min_{\mathbf{v}} \left\{ \langle \nabla f_i(\mathbf{x}_i^{(r+1)}) + \mathbf{h}_i^{(r)}, \mathbf{v} \rangle + \frac{\mu}{2} \|\mathbf{v} - \mathbf{x}_i^{(r+1)}\|^2 + \frac{\lambda}{2} \|\mathbf{v} - \mathbf{v}_i^{(r)}\|^2 \right\}^a$.
 - 6: $\mathbf{v}_i^{(r+1)} = \text{MULTIGOSSIP} \left(\mathbf{v}_i^{(r+\frac{1}{2})}, \mathbf{W}, M, i \right)$.
 - 7: $\mathbf{h}_i^{(r+1)} = \text{MULTIGOSSIP} \left(\mathbf{h}_i^{(r)} + \nabla f_i(\mathbf{v}_i^{(r+1)}), \mathbf{W}, M, i \right) - \nabla f_i(\mathbf{v}_i^{(r+1)})$.
-

^aThis update rule can be rewritten as $\mathbf{v}_i^{(r+\frac{1}{2})} = \frac{1}{\mu+\lambda} (\mu \mathbf{x}_i^{(r+1)} + \lambda \mathbf{v}_i^{(r)} - \nabla f_i(\mathbf{x}_i^{(r+1)}) - \mathbf{h}_i^{(r)})$.

For instance, it is sufficient to use a single gossip averaging step (i.e., $M = 1$) if the spectral gap parameter ρ satisfies the condition

$$\rho \leq \frac{\delta}{6L}, \quad (5.5)$$

as we prove in Lemma 18 in the appendix. This flexibility extends to other averaging mechanisms, provided they meet the same error guarantees.

In many practical scenarios, the choice $M = 1$ may not suffice, as it requires the network topology to be sufficiently dense to ensure adequate communication accuracy. When the network is sparse, a common approach is to perform multiple iterations of gossip averaging (i.e., Alg. 7) to improve the averaging accuracy. Specifically, if gossip averaging is performed M times, it holds that:

$$\sum_{i=1}^n \|\tilde{\mathbf{x}}_i - \bar{\mathbf{x}}\|^2 \leq \rho^{2M} \sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2,$$

where $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\tilde{\mathbf{x}}_i\}_{i=1}^n$ are the input and output of Alg. 7, and $\bar{\mathbf{x}} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. Thus, if we perform gossip averaging more than $M = \frac{1}{1-\rho} \log(\frac{6L}{\delta})$ times, ρ^M is less than or equal to $\frac{\delta}{6L}$, and then PDO works for arbitrary ρ .

5.4.1 Convergence Analysis

We now present the convergence analysis of PDO. The proofs are deferred to Sec. D.3 and D.6.2.

Theorem 6. *Consider Alg. 6. Suppose that Assumptions 5, 2, and 11 hold, and $\mathbf{h}_i^{(0)} = \nabla h_i(\mathbf{x}_i^{(0)})$, $\mathbf{x}_i^{(0)} = \bar{\mathbf{x}}^{(0)}$, and*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\delta(4\delta + \mu)}{4(r+1)(r+2)} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2. \quad (5.6)$$

Strongly-convex Case: *Suppose that $\mu > 0$. When $\lambda = 4\delta$ and $M \geq \frac{1}{1-\rho} \log(\frac{6L}{\delta})$, it holds that $\frac{1}{W_R} \sum_{r=0}^{R-1} w^{(r)} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) \leq \epsilon$ after $R = \mathcal{O}(\frac{\mu+\delta}{\mu} \log(\frac{\mu \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}))$*

rounds where $w^{(r)} := (1 + \frac{\mu}{4\delta})^r$ and $W_R := \sum_{r=0}^{R-1} w^{(r)}$. Thus, it requires at most

$$\mathcal{O}\left(\frac{\mu + \delta}{\mu(1-\rho)} \log\left(\frac{L}{\delta}\right) \log\left(\frac{\mu\|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}\right)\right)$$

communication where $\bar{\mathbf{x}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(r)}$.

Convex Case: Suppose that $\mu = 0$. When $\lambda = 4\delta$ and $M \geq \frac{1}{1-\rho} \log(\frac{6L}{\delta})$, it holds that $\frac{1}{R} \sum_{r=0}^{R-1} f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*) \leq \epsilon$ after $R = \mathcal{O}(\frac{\delta\|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon})$ rounds. Thus, it requires at most

$$\mathcal{O}\left(\frac{\delta\|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{(1-\rho)\epsilon} \log\left(\frac{L}{\delta}\right)\right)$$

communication.

Any optimization algorithm can be used to approximately solve the subproblem so that Eq. (5.6) is satisfied. For instance, if we use Nesterov's Accelerated Gradient Descent (Nesterov, 2018), we can achieve the following result.

Theorem 7. Consider Alg. 6. Suppose that Assumptions 5, 2, and 11 hold, and we use the same initial parameters, λ , and M as in Theorem 6. Then, if we use Nesterov's Accelerated Gradient Descent with initial parameter $\mathbf{x}_i^{(r)}$ to approximately solve the subproblem in line 4, each node requires at most

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu + \delta}} \log\left(\frac{L^2(r+2)^2}{\delta(\delta + \mu)}\right)\right)$$

iterations to satisfy Eq. (5.6).

5.4.2 Discussion

New Analysis with Inexact Subproblem Solution: The framework of PDO was initially introduced by Li et al. (2020) and Sun et al. (2022), and Sun et al. (2022) proposed SONATA. However, they assumed that the subproblem was exactly solved, which is infeasible in general. Our new theorem significantly relaxes this condition and shows that PDO can achieve the same communication complexity without solving the subproblem exactly as long as Eq. (5.6) is satisfied. To clarify this difference from Sun et al. (2022), we refer to PDO with the inexact subproblem solution as **Inexact-PDO**.

Reducing Dependence on $\delta_{\max}$ to δ : Sun et al. (2022) showed that SONATA can achieve $\tilde{\mathcal{O}}(\frac{\delta_{\max}}{\mu(1-\rho)} \log(\frac{1}{\epsilon}))$ communication complexity in the strongly-convex case. It suggests that the maximum dissimilarity $\delta_{\max}$ plays a significant role in the convergence behavior. However, Theorem 6 shows that the communication complexity can be improved to $\tilde{\mathcal{O}}(\frac{\delta}{\mu(1-\rho)} \log(\frac{1}{\epsilon}))$. This improvement indicates a more favorable dependence on the dissimilarity between functions $\{f_i\}_i$, which can lead to significantly lower communication costs since δ can be at most $\sqrt{n}$ times smaller than $\delta_{\max}$.

New Analysis in Convex Case: SONATA was analyzed only in the strongly-convex case (Sun et al., 2022). Our theorem is the first to provide the analysis in both convex and strongly-convex cases.

5.5 Stabilized Proximal Decentralized Optimization Method

Theorem 6 shows that PDO can achieve the best communication complexity among the existing non-accelerated decentralized methods. However, Eq. (5.6) implies that PDO needs to solve the subproblem more accurately as the number of rounds increases, which involves the suboptimal computational complexity. In this section, we propose **Stabilized Proximal Decentralized Optimization (SPDO)**, which can achieve the same communication complexity while enjoying a better computational complexity than PDO.

For simplicity, we consider that case when $\mu = 0$ and $M = 1$. To relax the condition for an approximate subproblem solution, we use the idea of Hybrid Projection Proximal-point Method (Solodov and Svaiter, 1999). Straightforwardly adapting Hybrid Projection-proximal Point Method to PDO yields the following update rules:¹

$$\begin{aligned} \mathbf{x}_i^{(r+1)} &\approx \arg \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ f_i(\mathbf{x}) + \langle \mathbf{h}_i^{(r)}, \mathbf{x} \rangle + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{v}_i^{(r)}\|^2 \right\}, \\ \mathbf{v}_i^{(r+\frac{1}{2})} &= \arg \min_{\mathbf{v} \in \mathbb{R}^d} \left\{ \langle \nabla f_i(\mathbf{x}_i^{(r+1)}), \mathbf{v} \rangle + \frac{\lambda}{2} \|\mathbf{v} - \mathbf{v}_i^{(r)}\|^2 \right\}, \\ \mathbf{v}_i^{(r+1)} &= \sum_{j=1}^n W_{ij} \mathbf{v}_j^{(r+\frac{1}{2})}, \\ \mathbf{h}_i^{(r+1)} &= \sum_{j=1}^n W_{ij} \left(\mathbf{h}_j^{(r)} + \nabla f_j(\mathbf{v}_j^{(r+1)}) \right) - \nabla f_i(\mathbf{v}_i^{(r+1)}). \end{aligned}$$

However, we found that the above update rules do not work. In the above update rule, $\mathbf{v}_i^{(r+\frac{1}{2})} \neq \mathbf{x}_i^{(r+1)}$ even though the subproblem is solved exactly. That is, the above update rules do not recover PDO when the subproblem is exactly solved. This discrepancy can hinder the convergence of the parameters to the optimal solution. Intuitively, when we solve the subproblem exactly, $\mathbf{v}_i^{(r+\frac{1}{2})}$ should equal to $\mathbf{x}_i^{(r+1)}$ since the above technique is not necessary. Motivated by this intuition, we modify the update rule of $\mathbf{v}_i^{(r+\frac{1}{2})}$ as in Alg. 8, in which the modification is highlighted. We refer to this method as **Stabilized Proximal Decentralized Optimization (SPDO)**. As a special instance of SPDO, we recover PDO. When the subproblem is exactly solved, i.e., $\nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) = \mathbf{0}$, it holds

$$\nabla f_i(\mathbf{x}_i^{(r+1)}) + \mathbf{h}_i^{(r)} + \lambda (\mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r)}) = \mathbf{0},$$

¹Note that $\arg \min_{\mathbf{v}}(\cdot)$ has the closed-form solution.

and with this choice of $\mathbf{h}_i^{(r)} = \lambda(\mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r)}) - \nabla f_i(\mathbf{x}_i^{(r+1)})$, Alg. 8 recovers PDO.

5.5.1 Convergence Analysis

In this section, we show the convergence analysis of SPDO. The proofs are deferred to Sec. D.4 and D.6.3. See also Lemma 22 as the result with $M = 1$.

Theorem 8. *Consider Alg. 8. Suppose that Assumptions 5, 2, and 11 hold. Then, suppose that $\mathbf{h}_i^{(0)} = \nabla h_i(\mathbf{v}_i^{(0)})$, $\mathbf{v}_i^{(0)} = \bar{\mathbf{v}}^{(0)}$, and*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) \right\|^2 \leq \frac{\lambda^2}{10} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2. \quad (5.7)$$

Strongly-convex Case: Suppose that $\mu > 0$. When $\lambda = 20\delta$ and $M \geq \frac{1}{1-\rho} \log(\frac{5L}{\delta})$, it holds that $\frac{1}{W_R} \sum_{r=0}^{R-1} w^{(r)} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) \leq \epsilon$ after $R = \mathcal{O}(\frac{\mu+\delta}{\mu} \log(\frac{\mu \|\bar{\mathbf{v}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}))$ rounds where $w^{(r)} := (1 + \frac{\mu}{20\delta})^r$ and $W_R := \sum_{r=0}^{R-1} w^{(r)}$. Thus, it requires at most

$$\mathcal{O} \left(\frac{\mu + \delta}{\mu(1-\rho)} \log \left(\frac{L}{\delta} \right) \log \left(\frac{\mu \|\bar{\mathbf{v}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon} \right) \right)$$

communication where $\bar{\mathbf{v}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i^{(r)}$.

Convex Case: Suppose that $\mu = 0$. When $\lambda = 20\delta$ and $M \geq \frac{1}{1-\rho} \log(\frac{5L}{\delta})$, it holds that $\frac{1}{R} \sum_{r=1}^R f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*) \leq \epsilon$ after $R = \mathcal{O}(\frac{\delta \|\bar{\mathbf{v}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon})$ rounds. Thus, it requires at most

$$\mathcal{O} \left(\frac{\delta \|\bar{\mathbf{v}}^{(0)} - \mathbf{x}^*\|^2}{(1-\rho)\epsilon} \log \left(\frac{L}{\delta} \right) \right)$$

communication.

Theorem 9. *Consider Alg. 8. Suppose that Assumptions 5, 2, and 11 hold, and we use the same initial parameters, λ , and M as in Theorem 8. If we use the algorithm proposed in Remark 1 in Nesterov et al. (2018) with initial parameter $\mathbf{v}_i^{(r)}$ to approximately solve the subproblem in line 4, each node requires at most $\mathcal{O}(\sqrt{\frac{L}{\delta}})$ iterations to satisfy Eq. (5.7).*

Remark 3. *Suppose that the same assumptions hold as in Theorem 9. Then, if we use Nesterov's Accelerated Gradient Descent with initial parameter $\mathbf{v}_i^{(r)}$ to approximately solve the subproblem in line 4, each node requires at most $\mathcal{O}(\sqrt{\frac{L}{\mu+\delta}} \log(\frac{L}{\delta}))$ iterations to satisfy Eq. (5.7).*

Discussion: Eq. (5.6) shows that Inexact-PDO needs to solve the subproblem more accurately as the number of rounds increases, whereas Eq. (5.7) indicates that SPDO does not. Thus, as shown in Theorem 6, 7, 8, and 9, SPDO can attain the same communication complexity as Inexact-PDO while enjoying a better computational complexity.

Algorithm 9 Accelerated Stabilized Proximal Decentralized Optimization Method (Accelerated-SPDO)

- 1: **Input:** The number of rounds R , and hyperparameters γ , M , and λ .
 - 2: $A_0 = 0$ and $B_0 = 1$.
 - 3: **for** $r = 0, 1, \dots, R - 1$ **do**
 - 4: Find $a_{r+1} > 0$ that satisfies $\lambda = \frac{(A_r + a_{r+1})B_r}{a_{r+1}^2}$.
 - 5: $A_{r+1} = A_r + a_{r+1}$.
 - 6: $B_{r+1} = B_r + \mu a_{r+1}$.
 - 7: **for** $i = 1, \dots, n$ **in parallel do**
 - 8: $\mathbf{y}_i^{(r)} = \frac{A_r \mathbf{x}_i^{(r)} + a_{r+1} \mathbf{v}_i^{(r)}}{A_{r+1}}$.
 - 9: **if** $r \geq 1$ **then**
 - 10: $\mathbf{h}_i^{(r)} = \text{FASTGOSSIP} \left(\mathbf{h}_i^{(r-1)} + \nabla f_i(\mathbf{y}_i^{(r)}), \mathbf{W}, M, \gamma, i \right) - \nabla f_i(\mathbf{y}_i^{(r)})$.
 - 11: $\mathbf{x}_i^{(r+\frac{1}{2})} \approx \arg \min_{\mathbf{x}} F_{i,r}(\mathbf{x})$ where $F_{i,r}(\mathbf{x}) := f_i(\mathbf{x}) + \langle \mathbf{h}_i^{(r)}, \mathbf{x} \rangle + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{y}_i^{(r)}\|^2$.
 - 12: $\mathbf{v}_i^{(r+\frac{1}{2})} = \arg \min_{\mathbf{v}} \left\{ a_{r+1} \left(\langle \nabla f_i(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{h}_i^{(r)}, \mathbf{v} \rangle + \frac{\mu}{2} \|\mathbf{v} - \mathbf{x}_i^{(r+\frac{1}{2})}\|^2 \right) + \frac{B_r}{2} \|\mathbf{v} - \mathbf{v}_i^{(r)}\|^2 \right\}^a$.
 - 13: $\mathbf{x}_i^{(r+1)} = \text{FASTGOSSIP} \left(\mathbf{x}_i^{(r+\frac{1}{2})}, \mathbf{W}, M, \gamma, i \right)$.
 - 14: $\mathbf{v}_i^{(r+1)} = \text{FASTGOSSIP} \left(\mathbf{v}_i^{(r+\frac{1}{2})}, \mathbf{W}, M, \gamma, i \right)$.
-

^aThis update rule can be rewritten as $\mathbf{v}_i^{(r+\frac{1}{2})} = \frac{1}{B_{r+1}} (B_r \mathbf{v}_i^{(r)} + \mu a_{r+1} \mathbf{x}_i^{(r+\frac{1}{2})} - a_{r+1} (\nabla f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \mathbf{h}_i^{(r)}))$.

5.6 Accelerated Stabilized Proximal Decentralized Optimization Method

In this section, we propose **Accelerated-SPDO**, which can attain the best communication and computational complexities among the existing methods. We present the pseudo-code in Alg. 9. We use Monteiro and Svaiter acceleration (Monteiro and Svaiter, 2013) and accelerated gossip averaging (Liu and Morse, 2011) in Accelerated-SPDO.

5.6.1 Convergence Analysis

We now provide the convergence analysis of Accelerated-SPDO. The proofs are deferred to Sec. D.5 and D.6.4.

Theorem 10. *Consider Alg. 9. Suppose that Assumptions 5, 2, and 11 hold. Then, suppose that $\mathbf{h}_i^{(0)} = \nabla h_i(\mathbf{v}_i^{(0)})$, $\mathbf{x}_i^{(0)} = \mathbf{v}_i^{(0)} = \bar{\mathbf{v}}^{(0)}$, and*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\lambda^2}{352} \sum_{i=1}^n \left\| \mathbf{y}_i^{(r)} - \mathbf{x}_i^{(r+\frac{1}{2})} \right\|^2. \quad (5.8)$$

Algorithm 10 Fast Gossip Averaging

```

1: function FASTGOSSIP( $\{\mathbf{a}_i\}_{i=1}^n, \mathbf{W}, M, \gamma, i$ )
2:    $\mathbf{a}_i^{(0)} = \mathbf{a}_i$  and  $\mathbf{a}_i^{(-1)} = \mathbf{a}_i$  for all  $i$ .
3:   for  $m = 0, \dots, M-1$  do
4:      $\mathbf{a}_i^{(m+1)} = (1 + \gamma) \sum_{j=1}^n W_{ij} \mathbf{a}_j^{(m)} - \gamma \mathbf{a}_i^{(m-1)}$ .
5:   return  $\mathbf{a}_i^{(M)}$ .

```

Strongly-convex Case: Suppose that $\mu > 0$. When $\lambda = 96\delta$, $\gamma = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1+\rho^2}}$, and $M \geq \frac{4}{\sqrt{1-\rho}} \log(\frac{18L^2(192\delta+\mu)}{\mu\delta^2})$, it holds that $f(\bar{\mathbf{x}}^{(R)}) - f(\mathbf{x}^*) \leq \epsilon$ holds after $R = \mathcal{O}(\sqrt{\frac{\mu+\delta}{\mu}} \log(1 + \sqrt{\frac{\min\{\mu, \delta\} \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}}))$ rounds. Thus, it requires at most

$$\mathcal{O}\left(\sqrt{\frac{\mu+\delta}{\mu(1-\rho)}} \log\left(\frac{L^2(\mu+\delta)}{\mu\delta^2}\right) \log\left(1 + \sqrt{\frac{\min\{\mu, \delta\} \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}}\right)\right)$$

communication where $\bar{\mathbf{x}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(r)}$.

Convex Case: Suppose that $\mu = 0$. When $\lambda = 208\delta$, $\gamma = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1+\rho^2}}$, and $M \geq \frac{3}{2\sqrt{1-\rho}} \log(\max\{\frac{12L}{\delta}, \frac{20384\delta \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}\})$, it holds that $f(\bar{\mathbf{x}}^{(R)}) - f(\mathbf{x}^*) \leq \epsilon$ holds after $R = \mathcal{O}(\sqrt{\frac{\delta \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}})$ rounds. Thus, it requires at most

$$\mathcal{O}\left(\sqrt{\frac{\delta \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{(1-\rho)\epsilon}} \log\left(\max\left\{\frac{L}{\delta}, \frac{\delta \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}\right\}\right)\right)$$

communication.

Theorem 11. Consider Alg. 9. Suppose that Assumptions 5, 2, and 11 hold, and we use the same initial parameters, γ , λ , and M as in Theorem 10. Then, if we use the algorithm proposed in Remark 1 in Nesterov et al. (2018) with initial parameter $\mathbf{y}_i^{(r)}$ to approximately solve the subproblem in line 12, it requires at most $\mathcal{O}(\sqrt{\frac{L}{\delta}})$ iterations to satisfy Eq. (5.8).

Remark 4. Suppose that the same assumptions hold as in Theorem 11. Then, if we use Nesterov's Accelerated Gradient Descent with initial parameter $\mathbf{y}_i^{(r)}$ to approximately solve the subproblem in line 12, it requires at most $\mathcal{O}(\sqrt{\frac{L}{\mu+\delta}} \log(\frac{L}{\delta}))$ iterations to satisfy Eq. (5.8).

5.6.2 Discussion

In this section, we compare Accelerated-SPDO with the existing state-of-the-art method, (Inexact) Accelerated SONATA (Tian et al., 2022). Tian et al. (2022)

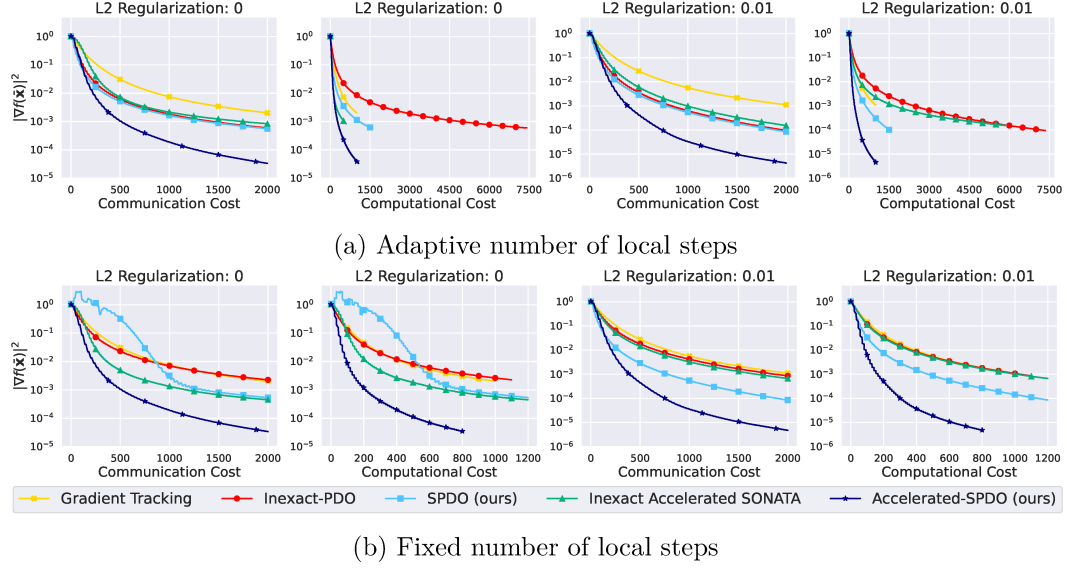


Figure 5.1: Convergence of the gradient norm with $\alpha = 0.1$. In (a), we ran gradient descent until the condition for an approximate subproblem solution was satisfied for all methods except for Gradient Tracking. For Gradient Tracking, we ran gradient descent for 10 times. In (b), we ran gradient descent for 10 times to approximately solve the subproblem. See Sec. D.7 for a more detailed setting.

assumed that f_i is strongly convex and $\mu \leq \delta_{\max}$. Thus, in the following, we discuss the case when f_i is strongly-convex, $\mu \leq \delta$, and $\mu \leq \delta_{\max}$. See Table 5.1 for the results.

Computational Complexity: Accelerated-SPDO can achieve the best computational complexity among the existing decentralized optimization methods. Accelerated SONATA needs to solve the subproblem exactly, which is infeasible in general. Inexact Accelerated SONATA solves the subproblem approximately and can achieve $\tilde{O}((\log(\frac{1}{\epsilon}))^2)$ computational complexity. Compared with these results, Accelerated-SPDO can achieve $\tilde{O}(\log(\frac{1}{\epsilon}))$ computational complexity. This is because Inexact Accelerated SONATA requires a more precise subproblem solution as the number of rounds increases, while Accelerated-SPDO does not.

Communication Complexity: Accelerated-SPDO can achieve the best communication complexity among the existing methods. In the worst case, δ can be equal to $\delta_{\max}$, and Accelerated-SPDO attains the same communication complexity as (Inexact) Accelerated SONATA. However, δ can be significantly smaller than $\delta_{\max}$. In this case, Accelerated-SPDO can achieve better communication complexity.

Acceleration in Convex Case: (Inexact) Accelerated SONATA has been shown to accelerate the rate only in the strongly-convex case, whereas Accelerated-SPDO

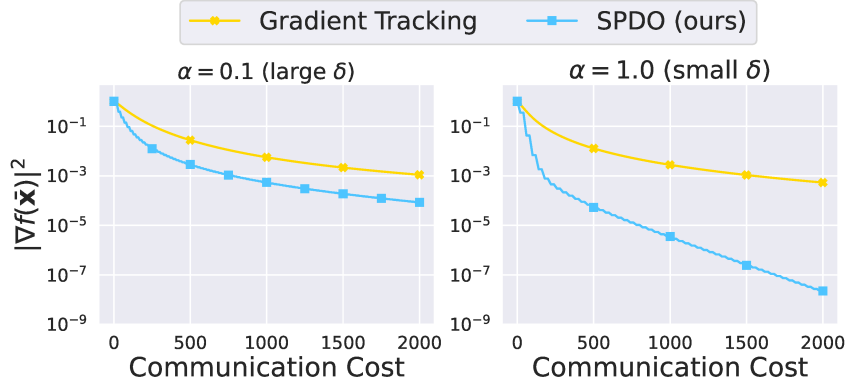


Figure 5.2: Convergence of the gradient norm with different α . We set the coefficient of L2 regularization to 0.01. For SPDO, we approximately solve the subproblem as in Fig. 5.1a.

can accelerate it in both convex and strongly-convex cases.

5.6.3 Discussion on Federated Learning

In this section, we discuss the relationship between our PDO-based methods and the existing methods developed for federated learning. Decentralized optimization generalizes federated learning, and when the underlying network is a complete graph, decentralized learning becomes equivalent to federated learning. PDO recovers DANE (Shamir et al., 2014) when the graph is fully-connected (i.e., $W_{ij} = \frac{1}{n}$). SPDO and Accelerated-SPDO contain S-DANE and Accelerated-S-DANE (Jiang et al., 2024b) as a special instance, respectively. Note that when the graph is fully-connected, $\rho = 0$ and multiple gossip averaging is not necessary. Furthermore, since $\sum_{i=1}^n \mathbf{h}_i^{(r)} = \mathbf{0}$, the technique we proposed in Sec. 5.4.2 is also not required (see the discussion in Sec. 5.4.2). However, in decentralized optimization, the network is generally not fully connected; hence, these techniques become essential to ensure convergence.

5.7 Experiments

Experimental Setup: We used MNIST (Lecun et al., 1998) and logistic loss with L2 regularization. We set the coefficient of L2 regularization to 0 and 0.01 and the number of nodes n to 25. We used the ring as the underlying network topology and distributed the data to nodes by using Dirichlet distribution with parameter $\alpha > 0$ (Hsu et al., 2019). For Inexact-PDO, Inexact Accelerated SONATA, SPDO, and (4) Accelerated-SPDO, we used gradient descent with stepsize η to solve the subproblem approximately. We set the number of communications to 2000 for all methods and tuned other hyperparameters, e.g., λ , M , and η , to minimize the norm of the last gradient for each method. See Sec. D.7 for a more detailed setting.

Results: Fig. 5.1a indicates Accelerated-SPDO can achieve the best communication and computational complexities. SPDO achieved the same communication complexity as Inexact-PDO while enjoying a better computation complexity. These results are consistent with Theorem 6 and 8. In Fig. 5.1b, we ran gradient descent for 10 times to solve the subproblem approximately. Accelerated-SPDO achieved the best communication and computational complexities as in Fig. 5.1a. The results indicate that SPDO outperforms Inexact-PDO in both communication and computational complexities. This is because the condition for an approximate subproblem solution Eq. (5.6) is not satisfied when the number of rounds is sufficiently large, which makes Inexact-PDO need to communicate more.

Effect of Similarity δ : Figure 5.2 compares the convergence behavior with different α . As α increases, nodes come to have similar datasets. The convergence behaviors of Gradient Tracking were almost the same in both cases, whereas SPDO can converge faster as α increases. This observation was consistent with the comparison in Table 5.1. Gradient Tracking is not affected by data heterogeneity, while it still requires a certain amount of communication even though nodes have very similar functions. In contrast, SPDO can exploit the similarity of local functions and reduce the communication complexity when nodes have similar functions.

5.8 Conclusion

In this chapter, we develop a communication and computation-efficient decentralized optimization method. To achieve this, we first introduce the proximal decentralized optimization (PDO) method and analyze the communication and computational complexities. We then propose SPDO, which can achieve the same communication complexity as PDO while enjoying a better computational complexity. Finally, we propose Accelerated-SPDO, which can achieve the best communication and computational complexities among the existing methods. Throughout the numerical experiments, we confirmed that our proposed method can outperform the existing methods as shown by our analysis.

6

Conclusion and Future Directions

This thesis investigated communication efficiency in decentralized learning and proposed novel network topologies and optimization methods to substantially reduce communication costs. By jointly addressing communication structure and algorithm design, this thesis contributed to improving the scalability and practicality of decentralized learning.

In Chapters 3 and 4, we addressed the issue that the existing decentralized learning methods require huge communication costs when the number of nodes increases. In Chapter 3, we proposed a novel topology, BASE- $(k + 1)$ GRAPH. BASE- $(k + 1)$ GRAPH is the finite-time convergence for any number of nodes and maximum degree, and can mix the parameters held by each node efficiently. Thanks to this property, BASE- $(k + 1)$ GRAPH can suppress the increase in the total number of communication rounds required for training, even when the number of nodes becomes large. In Chapter 4, we proposed TELEPORTATION. As we discussed in Chapter 3, BASE- $(k + 1)$ GRAPH is more communication-efficient than the existing topologies, such as 1-peer exponential graph, although the convergence rate of Decentralized SGD still degrades when the number of nodes is substantially large. TELEPORTATION can totally mitigate this issue by activating only a few nodes, and its convergence rate can be consistently improved as the number of nodes increases.

In Chapter 5, we proposed a novel optimization method, called ACCELERATED SPDO, to reduce communication costs. One of the most important characteristics of decentralized learning is that each node has a different dataset. Many optimization methods, including Gradient Tracking, have been proposed to address this issue, and their convergence rates are not affected by the data heterogeneity. However, Gradient Tracking requires the same communication costs for both cases, where nodes have the same datasets and have totally different datasets. ACCELERATED SPDO addresses this issue and reduces the communication costs by exploiting the similarity among the datasets held by nodes. Then, we showed that the communication complexity of ACCELERATED SPDO is optimal up to logarithmic factors.

6.1 Future Directions

This thesis opens up several promising directions for future research on communication-efficient decentralized learning.

Further Advances in Communication Patterns: First, further advances in communication topology design remain an important research direction. In particular, establishing fundamental lower bounds on the convergence rate or communication complexity would provide valuable insights. For example, TELEPORTATION can converge even when the number of nodes tends to infinity, yet it remains unclear whether TELEPORTATION is optimal.

Heterogeneous Communication Speeds: Second, it is important to consider the case where communication speeds differ across pairs of nodes. In this thesis, we assume that the communication speeds are the same for any pair of nodes, and the per-round communication costs are governed by the maximum degree of an underlying network topology. Relaxing these assumptions and explicitly modeling pairwise communication delays or bandwidth constraints is an important step toward bridging the gap between theory and real-world decentralized methods.

Optimizer for Neural Networks: The optimization methods proposed in this thesis, including PDO, SPDO, and ACCELERATED SPDO, are currently limited to convex objective functions. This limitation is not merely technical, as these methods explicitly exploit convexity in their design. Extending these algorithms to non-convex settings, incorporating stochastic gradients, and adapting them to the training of deep neural networks are highly important and challenging future directions. We believe that the principles underlying SPDO and ACCELERATED SPDO provide valuable insights for developing communication-efficient optimization methods beyond the convex regime.

Overall, this thesis represents an important step toward a deeper understanding of communication-efficient decentralized learning. By combining novel algorithmic ideas with rigorous theoretical analysis, it sheds light on both the potential and the limitations of decentralized optimization. We hope that the ideas and techniques developed herein will inspire further research in this area and contribute to the design of decentralized learning algorithms that are both theoretically sound and practically effective.

Bibliography

- Aketi, S. A., Hashemi, A., and Roy, K. (2023). Global update tracking: A decentralized learning algorithm for heterogeneous data. In *Advances in Neural Information Processing Systems*.
- Assran, M., Loizou, N., Ballas, N., and Rabbat, M. (2019). Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*.
- Bubeck, S. (2015). Convex optimization: Algorithms and complexity. In *Foundations and Trends in Machine Learning*.
- Chen, W., Horváth, S., and Richtárik, P. (2022). Optimal client sampling for federated learning. In *Transactions on Machine Learning Research*.
- Cho, Y. J., Wang, J., and Joshi, G. (2020). Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. In *arXiv*.
- Cyffers, E., Even, M., Bellet, A., and Massoulié, L. (2022). Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging. In *Advances in Neural Information Processing Systems*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Ding, L., Jin, K., Ying, B., Yuan, K., and Yin, W. (2023). DSGD-CECA: Decentralized SGD with communication-optimal exact consensus algorithm. In *International Conference on Machine Learning*.
- Dorfman, R. and Levy, K. Y. (2022). Adapting to mixing time in stochastic optimization with Markovian data. In *International Conference on Machine Learning*.
- Even, M. (2023). Stochastic gradient descent under Markovian sampling schemes. In *International Conference on Machine Learning*.
- Fraboni, Y., Vidal, R., Kamení, L., and Lorenzi, M. (2021). Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *International Conference on Machine Learning*.

- Gao, H. and Huang, H. (2020). Periodic stochastic gradient descent with momentum for decentralized training. In *arXiv*.
- Güler, O. (1992). New proximal point algorithms for convex minimization. In *SIAM Journal on Optimization*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Hendriks, H. (2023). A principled framework for the design and analysis of token algorithms. In *International Conference on Artificial Intelligence and Statistics*.
- Hsu, T.-M. H., Qi, H., and Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. In *arXiv*.
- Jiang, X., Rodomanov, A., and Stich, S. U. (2024a). Federated optimization with doubly regularized drift correction. In *International Conference on Machine Learning*.
- Jiang, X., Rodomanov, A., and Stich, S. U. (2024b). Stabilized proximal-point methods for federated optimization. In *Advances in Neural Information Processing Systems*.
- Johansson, B., Rabi, M., and Johansson, M. (2007). A simple peer-to-peer algorithm for distributed optimization in sensor networks. In *IEEE Conference on Decision and Control*.
- 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., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2021). Advances and open problems in federated learning. In *Foundations and Trends in Machine Learning*.
- Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. (2021). Mime: Mimicking centralized stochastic algorithms in federated learning. In *arXiv*.
- Khaled, A. and Jin, C. (2023). Faster federated optimization under second-order similarity. In *International Conference on Learning Representations*.

- Koloskova, A., Lin, T., and Stich, S. U. (2021). An improved analysis of gradient tracking for decentralized machine learning. In *Advances in Neural Information Processing Systems*.
- Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. (2020). A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*.
- Koloskova, A., Stich, S., and Jaggi, M. (2019). Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*.
- Kong, L., Lin, T., Koloskova, A., Jaggi, M., and Stich, S. (2021). Consensus control for decentralized deep learning. In *International Conference on Machine Learning*.
- Kovalev, D., Beznosikov, A., Borodich, E. D., Gasnikov, A., and Scutari, G. (2022). Optimal gradient sliding and its application to optimal distributed optimization under similarity. In *Advances in Neural Information Processing Systems*.
- Kovalev, D., Koloskova, A., Jaggi, M., Richtarik, P., and Stich, S. (2021). A linearly convergent algorithm for decentralized optimization: Sending less bits for free! In *International Conference on Artificial Intelligence and Statistics*.
- Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. In *Technical report*.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *IEEE*.
- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *IEEE*.
- Li, B., Cen, S., Chen, Y., and Chi, Y. (2020). Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. In *Journal of Machine Learning Research*.
- Li, Z., Shi, W., and Yan, M. (2019). A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. In *IEEE Transactions on Signal Processing*.
- Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*.
- Lin, D., Han, Y., Ye, H., and Zhang, Z. (2023). Stochastic distributed optimization under average second-order similarity: Algorithms and analysis. In *Advances in Neural Information Processing Systems*.

- Lin, T., Karimireddy, S. P., Stich, S., and Jaggi, M. (2021). Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In *International Conference on Machine Learning*.
- Liu, J. and Morse, A. (2011). Accelerated linear iterations for distributed averaging. In *Annual Reviews in Control*.
- Liu, Y., Lin, T., Koloskova, A., and Stich, S. U. (2024). Decentralized gradient tracking with local steps. In *Optimization Methods and Software*.
- Liu, Z., Koloskova, A., Jaggi, M., and Lin, T. (2022). Decentralized stochastic optimization with client sampling. In *Optimization for Machine Learning*.
- Lorenzo, P. D. and Scutari, G. (2016). Next: In-network nonconvex optimization. In *IEEE Transactions on Signal and Information Processing over Networks*.
- Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.
- Lu, Y. and De Sa, C. (2020). Moniqua: Modulo quantized communication in decentralized SGD. In *International Conference on Machine Learning*.
- Marfoq, O., Xu, C., Neglia, G., and Vidal, R. (2020). Throughput-optimal topology design for cross-silo federated learning. In *Advances in Neural Information Processing Systems*.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*.
- Monteiro, R. D. C. and Svaiter, B. F. (2013). An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. In *SIAM Journal on Optimization*.
- Murata, T. and Suzuki, T. (2021). Bias-variance reduced local SGD for less heterogeneous federated learning. In *International Conference on Machine Learning*.
- Nedic, A., Olshevsky, A., and Rabbat, M. G. (2018). Network topology and communication-computation tradeoffs in decentralized optimization. In *IEEE*.
- Nedić, A., Olshevsky, A., and Shi, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. In *SIAM Journal on Optimization*.
- Nedic, A. and Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. In *IEEE Transactions on Automatic Control*.
- Nesterov, Y. (2018). Lectures on convex optimization. In *Springer*.

- Nesterov, Y., Gasnikov, A. V., Guminov, S., and Dvurechensky, P. E. (2018). Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. In *Optimization Methods and Software*.
- Pu, S. and Nedic, A. (2021). Distributed stochastic gradient tracking methods. In *Mathematical Programming*.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research*.
- Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. In *SIAM Journal on Control and Optimization*.
- Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. (2017). Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning*.
- Scutari, G. and Sun, Y. (2019). Distributed nonconvex constrained optimization over time-varying digraphs. In *Mathematical Programming*.
- Shamir, O. and Srebro, N. (2014). Distributed stochastic optimization and learning. In *Annual Allerton Conference on Communication, Control, and Computing*.
- Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *International Conference on Machine Learning*.
- Shi, G., Li, B., Johansson, M., and Johansson, K. H. (2016). Finite-time convergent gossiping. In *IEEE Transactions on Networking*.
- Shi, W., Ling, Q., Wu, G., and Yin, W. (2015). Extra: An exact first-order algorithm for decentralized consensus optimization. In *SIAM Journal on Optimization*.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.
- Solodov, M. and Svaiter, B. (1999). A hybrid projection-proximal point algorithm. In *Journal of Convex Analysis*.
- Solodov, M. V. and Svaiter, B. F. (2001). A unified framework for some inexact proximal point algorithms. In *Numerical Functional Analysis and Optimization*.

- Song, Z., Li, W., Jin, K., Shi, L., Yan, M., Yin, W., and Yuan, K. (2022). Communication-efficient topologies for decentralized learning with $O(1)$ consensus rate. In *Advances in Neural Information Processing Systems*.
- Sun, Y., Scutari, G., and Daneshmand, A. (2022). Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation. In *SIAM Journal on Optimization*.
- Takezawa, Y., Bao, H., Niwa, K., Sato, R., and Yamada, M. (2023a). Momentum tracking: Momentum acceleration for decentralized deep learning on heterogeneous data. In *Transactions on Machine Learning Research*.
- Takezawa, Y., Sato, R., Bao, H., Niwa, K., and Yamada, M. (2023b). Beyond exponential graph: Communication-efficient topologies for decentralized learning via finite-time convergence. In *Advances in Neural Information Processing Systems*.
- Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. (2018). D^2 : Decentralized training over decentralized data. In *International Conference on Machine Learning*.
- Tian, Y., Scutari, G., Cao, T., and Gasnikov, A. (2022). Acceleration in distributed optimization under similarity. In *International Conference on Artificial Intelligence and Statistics*.
- Wang, J., Sahu, A. K., Yang, Z., Joshi, G., and Kar, S. (2019). MATCHA: Speeding up decentralized SGD via matching decomposition sampling. In *Indian Control Conference*.
- Wang, L., Guo, Y., Lin, T., and Tang, X. (2023). DELTA: Diverse client sampling for fasting federated learning. In *Advances in Neural Information Processing Systems*.
- Wu, F., Guo, S., Qu, Z., He, S., Liu, Z., and Gao, J. (2023). Anchor sampling for federated learning with partial client participation. In *International Conference on Machine Learning*.
- Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. In *arXiv*.
- Ying, B., Yuan, K., Chen, Y., Hu, H., Pan, P., and Yin, W. (2021). Exponential graph is provably efficient for decentralized deep training. In *Advances in Neural Information Processing Systems*.
- Yuan, K., Chen, Y., Huang, X., Zhang, Y., Pan, P., Xu, Y., and Yin, W. (2021). DecentLaM: Decentralized momentum sgd for large-batch deep training. In *International Conference on Computer Vision*.
- Yuan, K., Huang, X., Chen, Y., Zhang, X., Zhang, Y., and Pan, P. (2022). Revisiting optimal convergence rate for smooth and non-convex stochastic decentralized optimization. In *Advances in Neural Information Processing Systems*.

- Yuan, K., Ying, B., Zhao, X., and Sayed, A. H. (2019). Exact diffusion for distributed optimization and learning—part i: Algorithm development. In *IEEE Transactions on Signal Processing*.
- Zhu, T., He, F., Zhang, L., Niu, Z., Song, M., and Tao, D. (2022). Topology-aware generalization of decentralized SGD. In *International Conference on Machine Learning*.

A

Useful Lemma

We often use the following lemmas in the proofs.

Lemma 2. *For any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, it holds that for any $\alpha > 0$*

$$|\langle \mathbf{a}, \mathbf{b} \rangle| \leq \frac{1}{2\alpha} \|\mathbf{a}\|^2 + \frac{\alpha}{2} \|\mathbf{b}\|^2. \quad (\text{A.1})$$

Lemma 3. *For any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $\gamma > 0$, it holds that*

$$\|\mathbf{a} + \mathbf{b}\|^2 \leq (1 + \gamma) \|\mathbf{a}\|^2 + (1 + \gamma^{-1}) \|\mathbf{b}\|^2. \quad (\text{A.2})$$

Lemma 4. *For any $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \in \mathbb{R}^d$, it holds that*

$$\|\bar{\mathbf{a}}\|^2 \leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{a}_i\|^2, \quad (\text{A.3})$$

where $\bar{\mathbf{a}} := \frac{1}{n} \sum_{i=1}^n \mathbf{a}_i$.

B

Appendix of Base- $(k + 1)$ Graph

B.1 Detailed Explanation of k -peer Hyperhypercube Graph

In this section, we explain Alg. 1 in more detail. The k -PEER HYPER-HYPERCUBE GRAPH mainly consists of the following five steps.

Step 1. Decompose n as $n = n_1 \times \cdots \times n_L$ with minimum L such that $n_l \in [k + 1]$ for all $l \in [L]$.

Step 2. If $L = 1$, we make all nodes obtain the average of parameters in V by using the complete graph. If $L \geq 2$, we split V into disjoint subsets $V_1, \cdots, V_{n_L}$ such that $|V_l| = \frac{n}{n_L}$ for all $l \in [n_L]$ and continue to step 3.

Step 3. For all $l \in [n_L]$, we make all nodes in V_l obtain the average of parameters in V_l by using the k -PEER HYPER-HYPERCUBE GRAPH $\mathcal{H}_k(V_l)$.

Step 4. We take n_L nodes from $V_1, \cdots, V_{n_L}$ respectively and construct a set U_1 . Similarly, we construct $U_2, \cdots, U_{n_L}$ such that $U_1, \cdots, U_{n_L}$ are disjoint sets.

Step 5. For all $l \in [n_L]$, we make all nodes in U_l obtain the average of parameters in U_l by using the complete graph. Because the average of parameter U_l is equivalent to the average in V after step 4, all nodes reach the exact consensus.

When $n \leq k + 1$, the k -PEER HYPER-HYPERCUBE GRAPH becomes the complete graph because of step 2. When $n > k + 1$, we decompose n in step 1 and construct the k -PEER HYPER-HYPERCUBE GRAPH recursively in step 3. Thus, the k -PEER HYPER-HYPERCUBE GRAPH can make all nodes reach the exact consensus by the sequence of L graphs.

Using the example provided in Fig. B.1, we explain the k -PEER HYPER-HYPERCUBE GRAPH in a more detailed manner. When $n = 12$, we decompose 12 as $2 \times 2 \times 3$. In step 2, we split $V := \{1, \cdots, 12\}$ into $V_1 := \{1, \cdots, 4\}$, $V_2 := \{5, \cdots, 8\}$, and $V_3 := \{9, \cdots, 12\}$. Step 3 corresponds to the first two graphs in Fig. B.1b. As shown

in Fig. B.1a, the subgraphs consisting of V_1 , V_2 , and V_3 in the first two graphs in Fig. B.1b are equivalent to the k -PEER HYPER-HYPERCUBE GRAPH with the number of nodes 4. Thus, all nodes reach the exact consensus by exchanging parameters in Fig. B.1b.

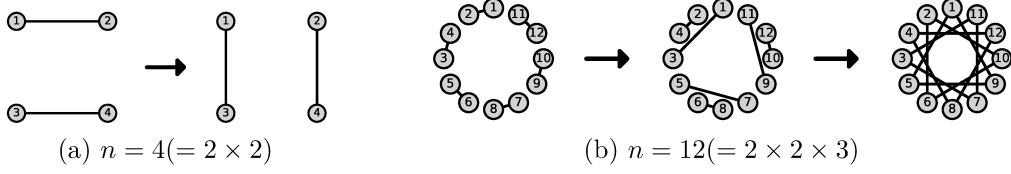


Figure B.1: Illustration of the 2-PEER HYPER-HYPERCUBE GRAPH. In Fig. B.1a, all edge weights are $\frac{1}{2}$. In Fig. B.1b, edge weights are $\frac{1}{2}$ in the first two graphs and $\frac{1}{3}$ in the last graph.

B.2 Detailed Explanation of Simple Base- $(k + 1)$ Graph with $k \geq 2$

In Sec. 3.4.2, we explain Alg. 2 only in the case where maximum degree k is one. In this section, we explain the details of Alg. 2 in the case with $k \geq 2$.

The SIMPLE BASE- $(k + 1)$ GRAPH mainly consists of the following five steps.

Step 1. As in the base- $(k + 1)$ number of n , we decompose n as $n = a_1(k + 1)^{p_1} + \dots + a_L(k + 1)^{p_L}$ in line 1, and then split V into disjoint subsets $V_1, \dots, V_L$ such that $|V_l| = a_l(k + 1)^{p_l}$ for all $l \in [L]$.

Step 2. For all $l \in [L]$, we split V_l into disjoint subsets $V_{l,1}, \dots, V_{l,a_l}$ such that $|V_{l,a}| = (k + 1)^{p_l}$ for all $a \in [a_l]$ in line 3.

Step 3. For all $l \in [L]$, we make all nodes in V_l obtain the average of parameters in V_l using the k -PEER HYPER-HYPERCUBE GRAPH $\mathcal{H}_k(V_l)$ in line 11. Then, we initialize l' as one.

Step 4. Each node in $V_{l'+1} \cup \dots \cup V_L$ exchange parameters with $a_{l'}$ nodes in $V_{l'} (= V_{l',1} \cup \dots \cup V_{l',a_{l'}})$ such that the average in $V_{l',a}$ becomes equivalent to the average in V for all $a \in [a_{l'}]$. We increase l' by one and repeat step 4 until $l' = L$. This procedure corresponds to line 15.

Step 5. For all $l \in [L]$ and $a \in [a_l]$, we make all nodes in $V_{l,a}$ obtain the average in $V_{l,a}$ using the k -PEER HYPER-HYPERCUBE GRAPH $\mathcal{H}_k(V_{l,a})$. Since the average in $V_{l,a}$ is equivalent to the average in V after step 4, all nodes reach the exact consensus. This procedure corresponds to line 25.

The major difference compared with the case where $k = 1$ is step 4. In the case where $k = 1$, each node in $V_{l'+1} \cup \dots \cup V_L$ exchange parameters with one node in $V_{l'}$ such that the average in $V_{l'}$ becomes equivalent to that in V , while in the case where $k \geq 2$, each node in $V_{l'+1} \cup \dots \cup V_L$ exchange parameters such that the average in $V_{l',a}$ becomes equivalent to that in V for all $a \in [a_{l'}]$. Thanks to this step, we can make all nodes reach the exact consensus using k -PEER HYPER-HYPERCUBE GRAPH $\mathcal{H}_k(V_{l,a})$ instead of $\mathcal{H}_k(V_l)$ in step 5, and we can reduce the length of a graph sequence.

Using the example provided in Fig. B.2, we explain Alg. 2 in a more detailed manner. Let $G^{(1)}, \dots, G^{(4)}$ denote the graphs depicted in Fig. B.2 from left to right, respectively. First, we split $V := \{1, \dots, 7\}$ into $V_1 := \{1, \dots, 6\}$ and $V_2 := \{7\}$, and then split V_1 into $V_{1,1} := \{1, 2, 3\}$ and $V_{1,2} := \{4, 5, 6\}$. In step 3, all nodes in V_1 obtain the same parameter by exchanging parameters in $G^{(1)}$ and $G^{(2)}$. In step 4, the average in $V_{1,1}$, that in $V_{1,2}$, and that in V_2 become the same as the average in all nodes V by exchanging parameters in $G^{(3)}$. Thus, in step 5, all nodes reach the exact consensus by exchanging parameters in $G^{(4)}$.

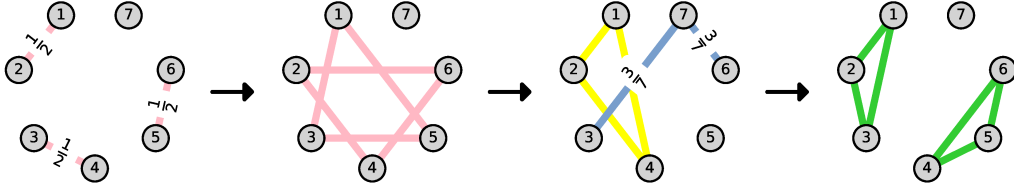


Figure B.2: $k = 2, n = 7 (= 2 \times 3 + 1)$. The value on the edge indicates the edge weight. For simplicity, we omit the edge value when it is $\frac{1}{3}$.

B.3 Illustration of Topologies

B.3.1 Examples

Fig. B.3 shows the examples of the SIMPLE BASE- $(k + 1)$ GRAPH. Using these examples, we explain how all nodes reach the exact consensus.

We explain the case depicted in Fig. B.3a. Let $G^{(1)}, G^{(2)}, G^{(3)}$ denote the graphs depicted in Fig. B.3a from left to right, respectively. First, we split $V := \{1, \dots, 5\}$ into $V_1 := \{1, 2, 3\}$ and $V_2 := \{4, 5\}$, and then split V_2 into $V_{2,1} := \{4\}$ and $V_{2,2} := \{5\}$. After exchanging parameters in $G^{(1)}$, nodes in V_1 and nodes in V_2 have the same parameter respectively. Then, after exchanging parameters in $G^{(2)}$, the average in V_1 , that in $V_{2,1}$, and that in $V_{2,2}$ become same as the average in V . Thus, by exchanging parameters in $G^{(3)}$, all nodes reach the exact consensus. Note that edge $(4, 5)$ in $G^{(3)}$, which is added in line 27 in Alg. 2, is not necessary for all nodes to reach the exact consensus because nodes 4 and 5 already have the same parameter after exchanging parameters in $G^{(2)}$; however, it is effective in decentralized learning as we explained in Sec. 3.4.2.

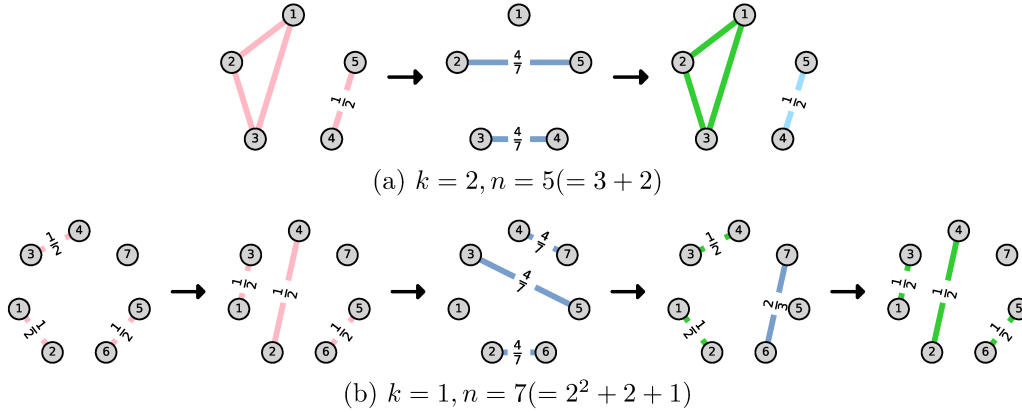


Figure B.3: Illustration of the SIMPLE BASE- $(k + 1)$ GRAPH. The edge is colored in the same color as the line of Alg. 2 where the edge is added. The value on the edge indicates the edge weight. For simplicity, we omit the edge value when it is $\frac{1}{3}$.

B.3.2 Illustrative Comparison between Simple Base- $(k + 1)$ and Base- $(k + 1)$ Graphs

In this section, we provide an example of the SIMPLE BASE- $(k + 1)$ GRAPH, explaining the reason why the length of the BASE- $(k + 1)$ GRAPH is less than that of the SIMPLE BASE- $(k + 1)$ GRAPH.

Let $G^{(1)}, \dots, G^{(5)}$ denote the graphs depicted in Fig. B.4 from left to right, respectively. $(G^{(1)}, G^{(2)}, G^{(3)}, G^{(4)}, G^{(5)})$ is finite-time convergence, but $(G^{(1)}, G^{(2)}, G^{(3)}, G^{(5)})$ is also finite-time convergence because after exchanging parameters in $G^{(3)}$, nodes 3 and 4 already have the same parameters. Then, using the technique proposed in Sec. 3.4.3, we can remove such unnecessary graphs contained in the SIMPLE BASE- $(k + 1)$ GRAPH (see Fig. 3.4a). Consequently, the BASE- $(k + 1)$ GRAPH can make all nodes reach the exact consensus faster than the SIMPLE BASE- $(k + 1)$ GRAPH.

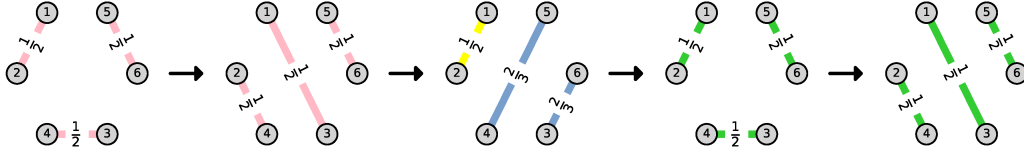


Figure B.4: Illustration of the SIMPLE BASE-2 GRAPH with $n = 6 (= 2^2 + 2)$. The edge is colored in the same color as the line of Alg. 2 where the edge is added.

B.3.3 Additional Examples

Simple Base- $(k + 1)$ Graph

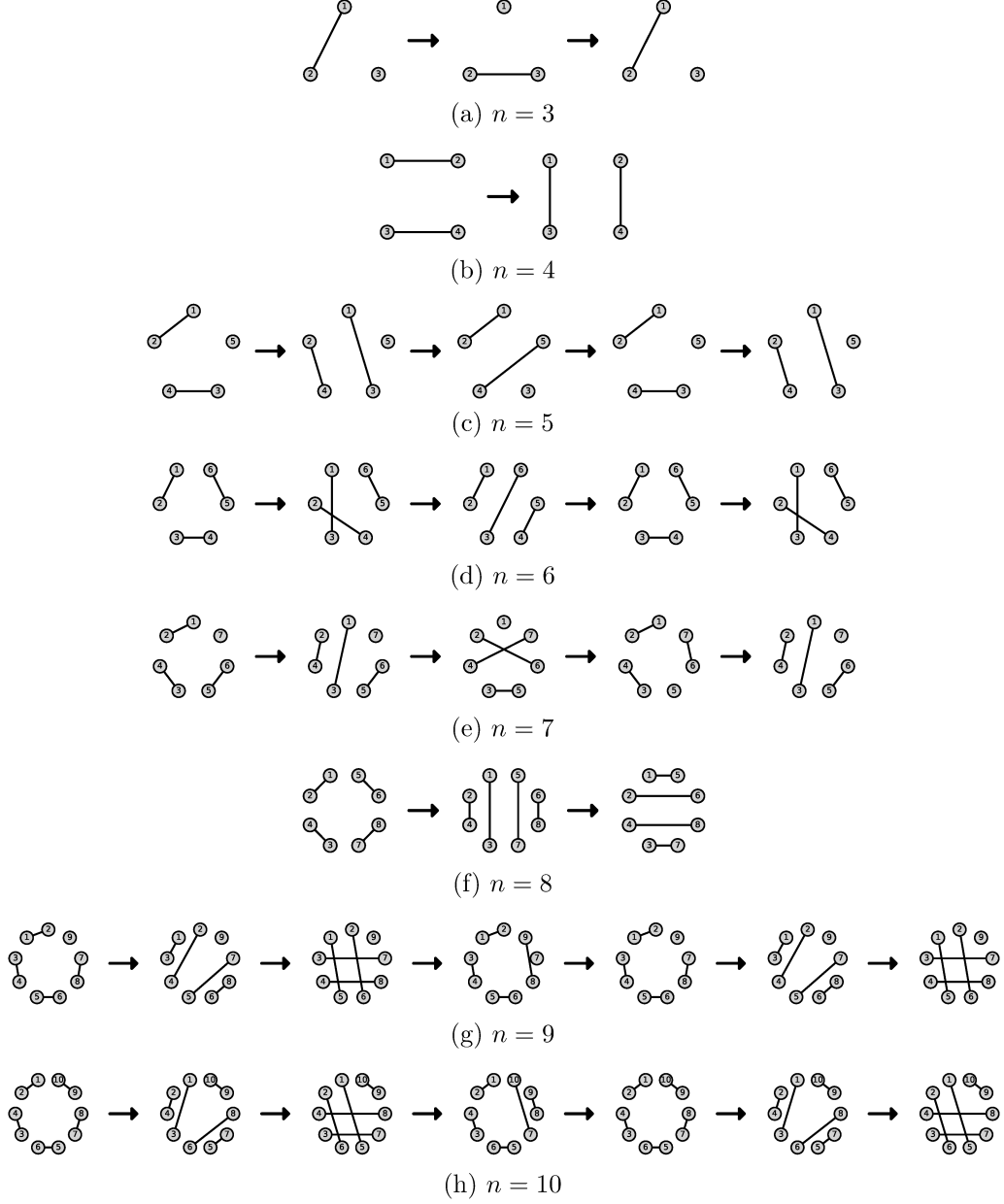


Figure B.5: Illustration of the SIMPLE BASE-2 GRAPH with the various numbers of nodes.

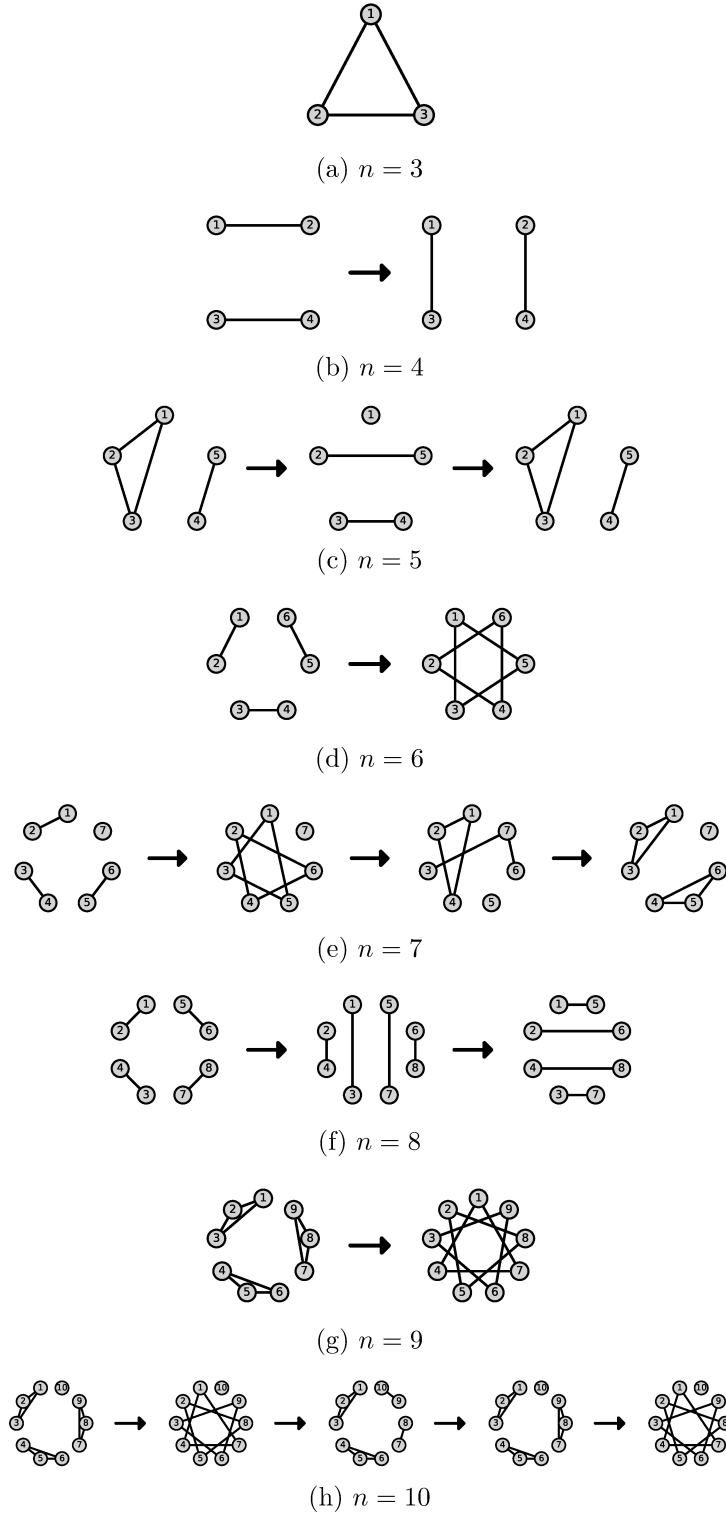


Figure B.6: Illustration of the SIMPLE BASE-3 GRAPH with the various numbers of nodes.

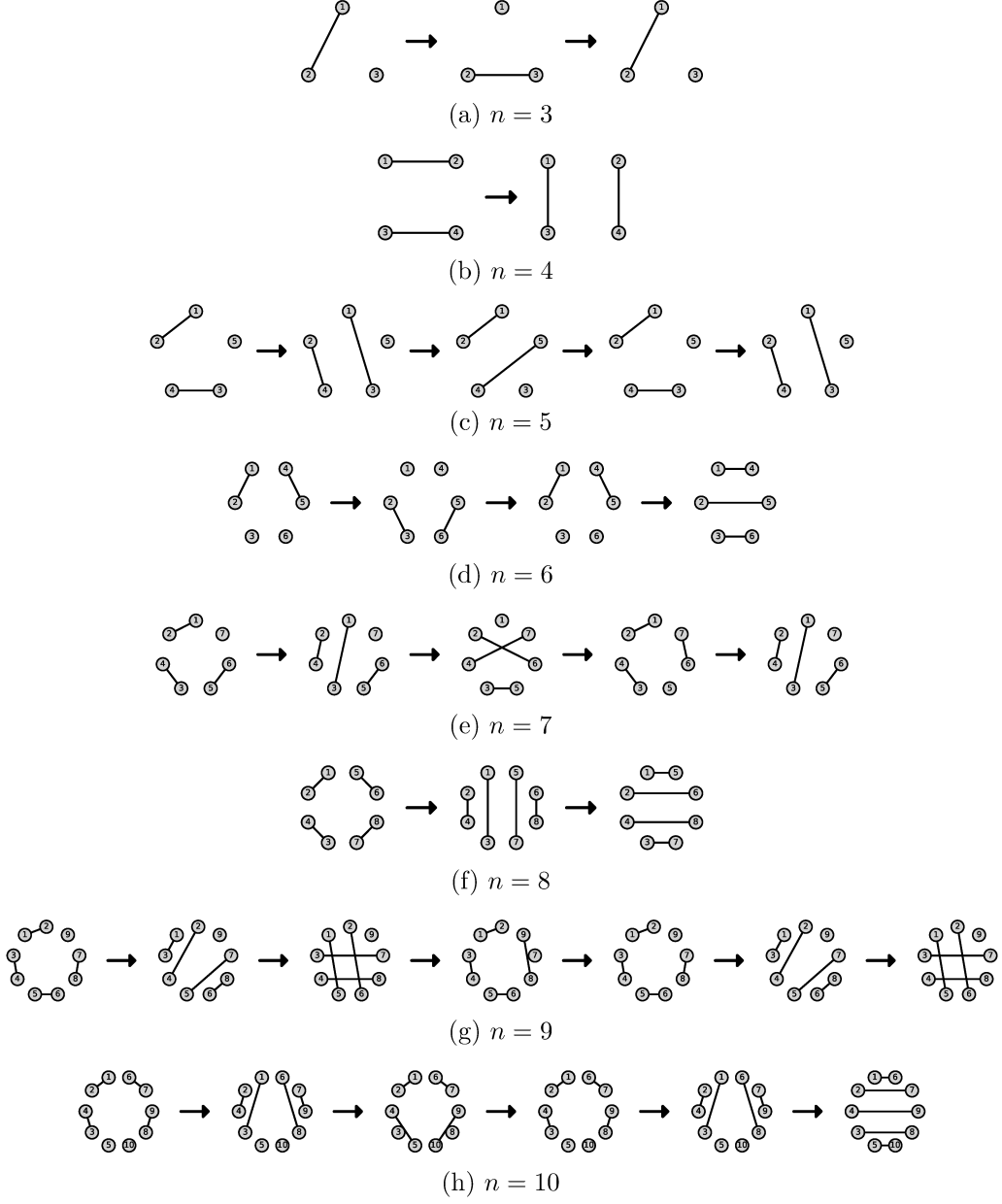
Base- $(k + 1)$ Graph

Figure B.7: Illustration of the BASE-2 GRAPH with the various numbers of nodes.

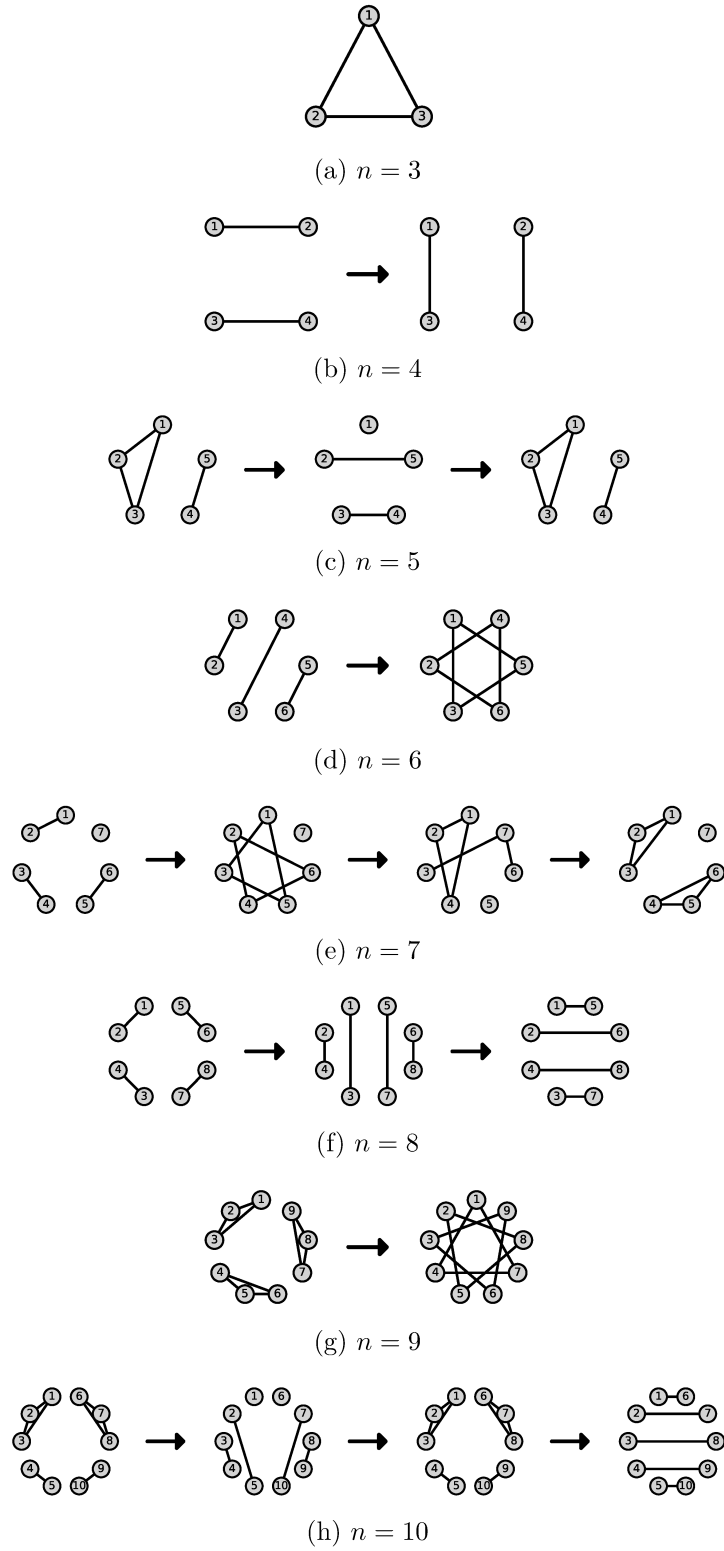


Figure B.8: Illustration of the BASE-3 GRAPH with the various numbers of nodes.

B.3.4 1-peer Hypercube Graph and 1-peer Exponential Graph

For completeness, we provide examples of the 1-peer hypercube Shi et al. (2016) and 1-peer exponential graphs Ying et al. (2021) in Figs. B.10 and B.9, respectively.

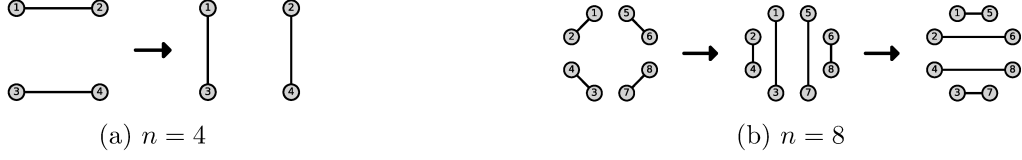


Figure B.9: Illustration of the 1-peer hypercube graph. All edge weights are 0.5.

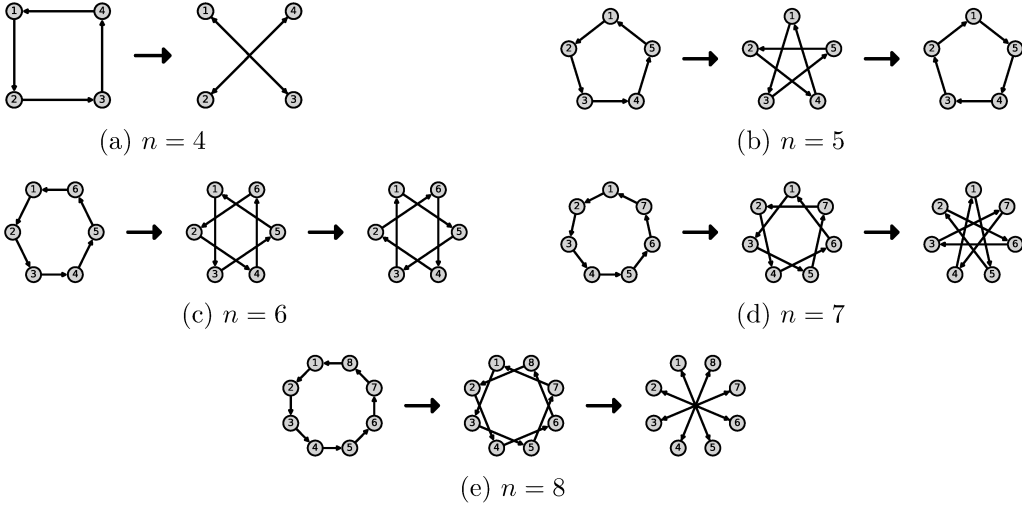


Figure B.10: Illustration of the 1-peer exponential graph. All edge weights are 0.5.

B.4 Proof of Theorem 1

Lemma 5 (Length of k -PEER HYPER-HYPERCUBE GRAPH). *Suppose that all prime factors of the number of nodes n are less than or equal to $k + 1$. Then, for any number of nodes $n \in \mathbb{N}$ and maximum degree $k \in [n - 1]$, the length of the k -PEER HYPER-HYPERCUBE GRAPH is less than or equal to $\max\{1, 2\log_{k+2}(n)\}$.*

Proof. We assume that n is decomposed as $n = n_1 \times \cdots \times n_L$ with minimum L where $n_l \in [k + 1]$ for all $l \in [L]$. Without loss of generality, we suppose $n_1 \leq n_2 \leq \cdots \leq n_L$. Then, for any $i \neq j$, it holds that $n_i \times n_j \geq k + 2$ because if $n_i \times n_j \leq k + 1$ for some i and j , this contradicts the assumption that L is minimum.

When L is even, we have

$$n = (n_1 \times n_2) \times \cdots \times (n_{L-1} \times n_L) \geq (k + 2)^{\frac{L}{2}}.$$

Then, we get $L \leq 2\log_{k+2}(n)$.

Next, we discuss the case when L is odd. When $L \geq 3$, $n_L \geq \sqrt{k + 2}$ holds because $n_{L-2} \times n_{L-1} \geq k + 2$. Thus, we get

$$n = (n_1 \times n_2) \times \cdots \times (n_{L-2} \times n_{L-1}) \times n_L \geq (k + 2)^{\frac{L-1}{2}} \times n_L \geq (k + 2)^{\frac{L}{2}}.$$

Then, we get $L \leq 2\log_{k+2}(n)$ when $L \geq 3$.

Thus, given the case when $L = 1$, the length of the k -PEER HYPER-HYPERCUBE GRAPH is less than or equal to $\max\{1, 2\log_{k+2}(n)\}$. $\square$

Lemma 6 (Length of SIMPLE BASE- $(k + 1)$ GRAPH). *For any number of nodes $n \in \mathbb{N}$ and maximum degree $k \in [n - 1]$, the length of the SIMPLE BASE- $(k + 1)$ GRAPH is less than or equal to $2\log_{k+1}(n) + 2$.*

Proof. When all prime factors of n are less than or equal to $k + 1$, the SIMPLE BASE- $(k + 1)$ GRAPH is equivalent to the k -PEER HYPER-HYPERCUBE GRAPH and the statement holds from Lemma 5. In the following, we consider the case when there exists a prime factor of n that is larger than $k + 1$. Note that because when $L = 1$ (i.e., $n = a_1 \times (k + 1)^{p_1}$), all prime factors of n are less than or equal to $k + 1$, we only need to consider the case when $L \geq 2$. We have the following inequality:

$$\begin{aligned} \log_{k+1}(n) &= \log_{k+1}(a_1(k + 1)^{p_1} + \cdots + a_L(k + 1)^{p_L}) \\ &\geq p_1 + \log_{k+1}(a_1) \\ &\geq p_1. \end{aligned}$$

Then, because $|V_1| = a_1 \times (k + 1)^{p_1}$, it holds that $m_1 = |\mathcal{H}_k(V_1)| \leq 1 + p_1 \leq \log_{k+1}(n) + 1$. Similarly, it holds that $|\mathcal{H}_k(V_{1,1})| = p_1 \leq \log_{k+1}(n)$ because $|V_{1,1}| = (k + 1)^{p_1}$. In Alg. 2, the update rule $b_1 \leftarrow b_1 + 1$ in line 22 is executed for the first time when $m = m_1 + 2$ because $L \geq 2$. Thus, the length of the SIMPLE BASE- $(k + 1)$ GRAPH is at most $m_1 + |\mathcal{H}_k(V_{1,1})| + 1 \leq 2\log_{k+1}(n) + 2$. This concludes the statement. $\square$

Lemma 7 (Length of BASE- $(k + 1)$ GRAPH). *For any number of nodes $n \in \mathbb{N}$ and maximum degree $k \in [n - 1]$, the length of the BASE- $(k + 1)$ GRAPH is less than or equal to $2\log_{k+1}(n) + 2$.*

Proof. The statement follows immediately from Lemma 6 and line 12 in Alg. 3. $\square$

B.5 Convergence Rate of DSGD over Various Topologies

Table B.1 lists the convergence rates of DSGD over various topologies. These convergence rates can be immediately obtained from Theorem 2 stated in Koloskova et al. (2020) and consensus rate of the topology. As seen from Table B.1, the BASE-2 GRAPH enables DSGD to converge faster than the ring and torus and as fast as the exponential graph for any number of nodes, although the maximum degree of the BASE-2 GRAPH is only one. Moreover, for any number of nodes, the BASE- $(k+1)$ GRAPH with $2 \leq k < \lceil \log_2(n) \rceil$ enables DSGD to converge faster than the exponential graph, even though the maximum degree of the BASE- $(k+1)$ GRAPH remains to be less than that of the exponential graph.

Table B.1: Convergence rates and maximum degrees of DSGD over various topologies.

Topology	Convergence Rate	Maximum Degree	#Nodes n
Ring (Nedic et al., 2018)	$\mathcal{O}\left(\frac{\sigma^2}{n\epsilon^2} + \frac{\zeta n^2 + \sigma n}{\epsilon^{3/2}} + \frac{n^2}{\epsilon}\right) \cdot LF_0$	2	$\forall n \in \mathbb{N}$
Torus (Nedic et al., 2018)	$\mathcal{O}\left(\frac{\sigma^2}{n\epsilon^2} + \frac{\zeta n + \sigma\sqrt{n}}{\epsilon^{3/2}} + \frac{n}{\epsilon}\right) \cdot LF_0$	4	$\forall n \in \mathbb{N}$
Exp. (Ying et al., 2021)	$\mathcal{O}\left(\frac{\sigma^2}{n\epsilon^2} + \frac{\zeta \log_2(n) + \sigma\sqrt{\log_2(n)}}{\epsilon^{3/2}} + \frac{\log_2(n)}{\epsilon}\right) \cdot LF_0$	$\lceil \log_2(n) \rceil$	$\forall n \in \mathbb{N}$
1-peer Exp. (Ying et al., 2021)	$\mathcal{O}\left(\frac{\sigma^2}{n\epsilon^2} + \frac{\zeta \log_2(n) + \sigma\sqrt{\log_2(n)}}{\epsilon^{3/2}} + \frac{\log_2(n)}{\epsilon}\right) \cdot LF_0$	1	A power of 2
1-peer Hypercube (Shi et al., 2016)	$\mathcal{O}\left(\frac{\sigma^2}{n\epsilon^2} + \frac{\zeta \log_2(n) + \sigma\sqrt{\log_2(n)}}{\epsilon^{3/2}} + \frac{\log_2(n)}{\epsilon}\right) \cdot LF_0$	1	A power of 2
Base-$(k+1)$ Graph (ours)	$\mathcal{O}\left(\frac{\sigma^2}{n\epsilon^2} + \frac{\zeta \log_{k+1}(n) + \sigma\sqrt{\log_{k+1}(n)}}{\epsilon^{3/2}} + \frac{\log_{k+1}(n)}{\epsilon}\right) \cdot LF_0$	k	$\forall n \in \mathbb{N}$

B.6 Additional Experiments

B.6.1 Comparison of Base- $(k + 1)$ and Simple Base- $(k + 1)$ Graphs

Fig. B.11 shows the length of the SIMPLE BASE- $(k + 1)$ GRAPH and BASE- $(k + 1)$ GRAPH. The results indicate that for all k , the length of the BASE- $(k + 1)$ GRAPH is less than the length of the SIMPLE BASE- $(k + 1)$ GRAPH in many cases.

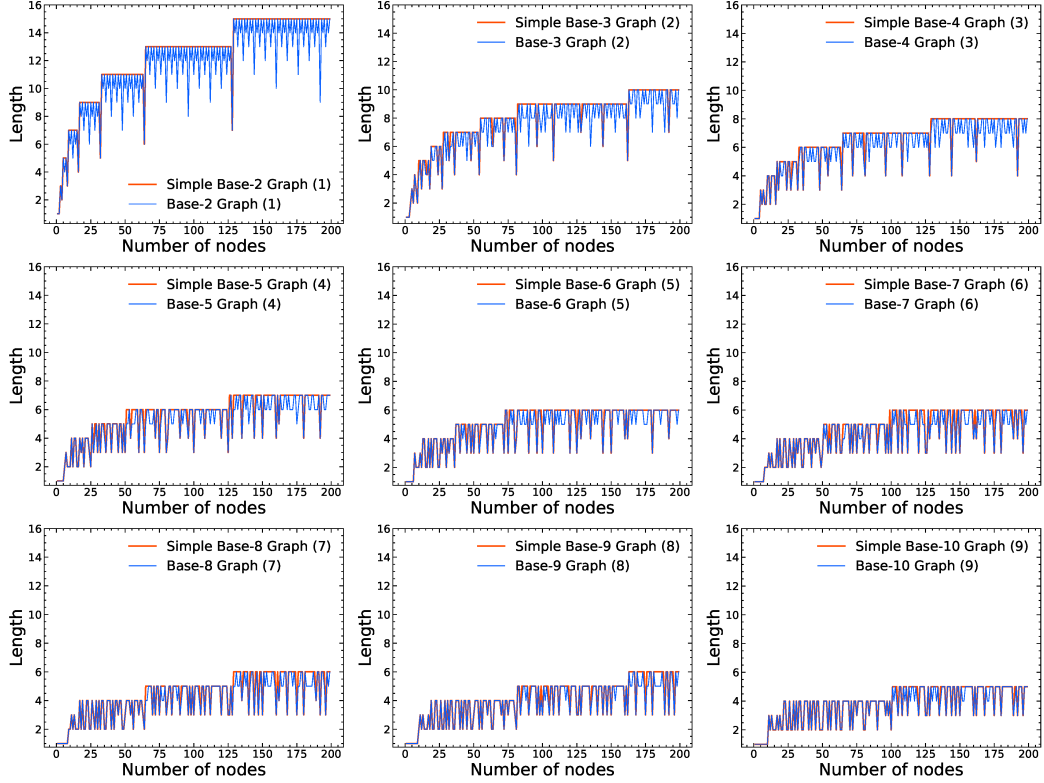


Figure B.11: Comparison of the length of the SIMPLE BASE- $(k + 1)$ GRAPH and BASE- $(k + 1)$ GRAPH.

B.6.2 Consensus Rate

In Fig. B.12, we demonstrate how consensus error decreases on various topologies when the number of nodes n is a power of 2. The results indicate that the BASE-2 GRAPH and 1-peer exponential graph can reach the exact consensus after the same finite number of iterations and reach the consensus faster than other topologies. Note that the BASE-2 GRAPH is equivalent to the 1-peer hypercube graph when n is a power of 2.

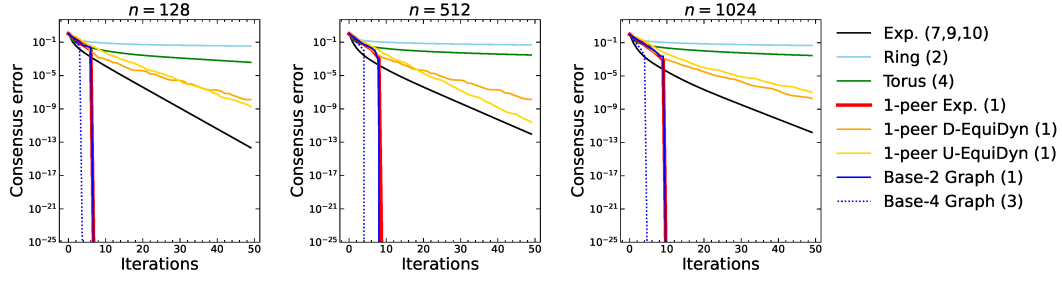


Figure B.12: Comparison of consensus rates among different topologies when the number of nodes n is a power of 2. Because the BASE- $\{3, 5\}$ GRAPH are the same as the BASE- $\{2, 4\}$ GRAPH, respectively, when n is a power of 2, we omit the results of the BASE- $\{3, 5\}$ GRAPH.

B.6.3 Decentralized Learning

Comparison of Base- $(k + 1)$ Graph and EquiStatic

In this section, we compared the BASE- $(k + 1)$ GRAPH with the $\{U, D\}$ -EquiStatic Song et al. (2022). The $\{U, D\}$ -EquiStatic are dense variants of the 1-peer $\{U, D\}$ -EquiDyn, and their maximum degree can be set as hyperparameters. We evaluated the $\{U, D\}$ -EquiStatic varying their maximum degrees; the results are presented in Fig. B.13. In both cases with $\alpha = 10$ and $\alpha = 0.1$, the BASE-2 GRAPH can achieve comparable or higher final accuracy than all $\{U, D\}$ -EquiStatic, and the BASE- $\{3, 4, 5\}$ GRAPH outperforms all $\{U, D\}$ -EquiStatic. Thus, the BASE- $(k + 1)$ GRAPH is superior to the $\{U, D\}$ -EquiStatic from the perspective of achieving a balance between accuracy and communication efficiency.

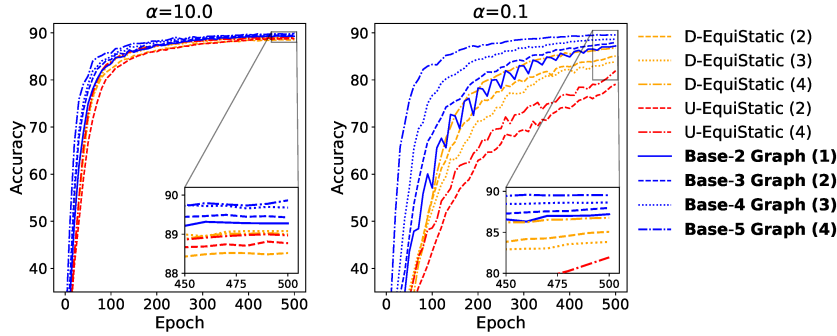


Figure B.13: Test accuracy (%) of DSGD with CIFAR-10 and $n = 25$. The number in the bracket is the maximum degree of a topology.

Comparison with Various Number of Nodes

In this section, we evaluated the effectiveness of the BASE- $(k + 1)$ GRAPH when varying the number of nodes n . Fig. B.15 presents the learning curves, and Fig B.14

shows how consensus error decreases when n is 21, 22, 23, 24, and 25. From Fig. B.15, the BASE-2 GRAPH consistently outperforms the 1-peer exponential graph and can achieve a final accuracy comparable to that of the exponential graph. Furthermore, the BASE- $\{3, 4, 5\}$ GRAPH can consistently outperform the exponential graph, even though the maximum degree of the BASE- $\{3, 4, 5\}$ GRAPH is less than that of the exponential graph.

Fig. B.16 presents the learning curve for $n = 16$. When the number of nodes is a power of two, the 1-peer exponential graph is also finite-time convergence, and the 1-peer exponential graph and BASE-2 GRAPH achieve competitive accuracy.

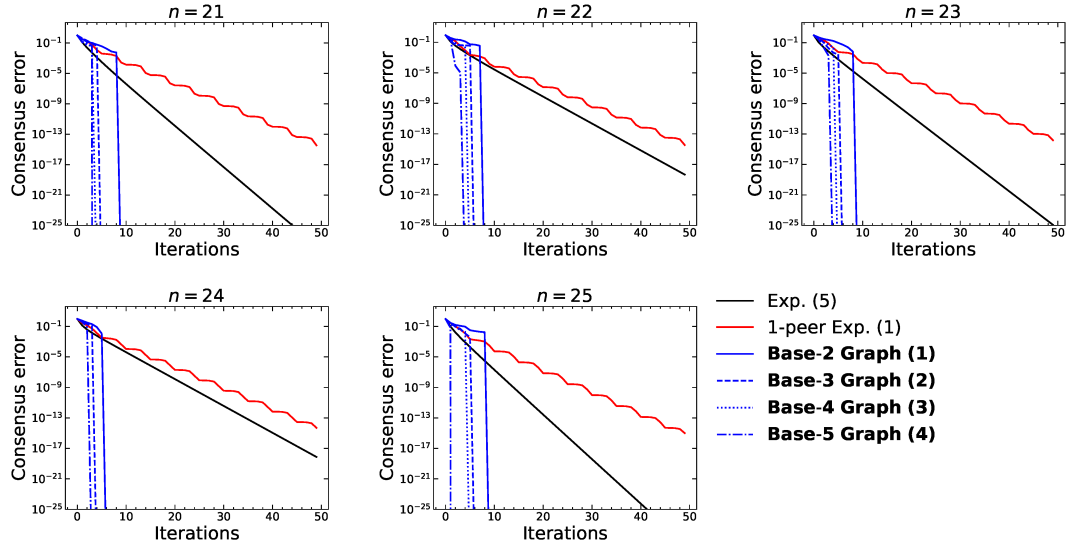


Figure B.14: Comparison of consensus rates among different topologies. The number in the bracket denotes the maximum degree of a topology. We omit the results of the BASE-5 GRAPH when $n = 24$ because the BASE-5 GRAPH and BASE-4 GRAPH are equivalent when $n = 24$.

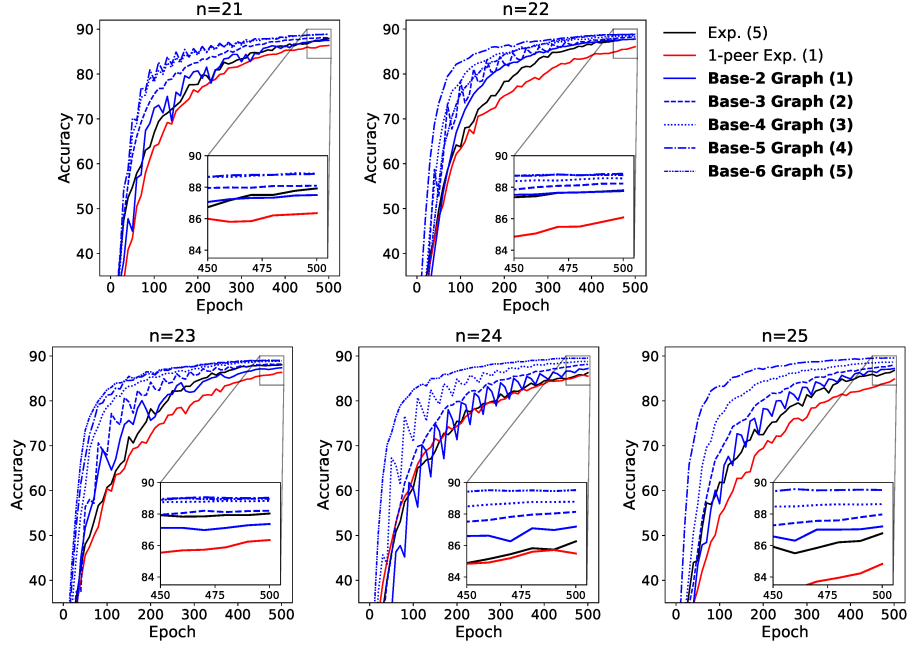


Figure B.15: Test accuracy (%) of DSGD with CIFAR-10 and $\alpha = 0.1$. The number in the bracket denotes the maximum degree of a topology. When $n = 24$, we omit the results of the BASE-5 GRAPH because the BASE-5 GRAPH and BASE-4 GRAPH are equivalent. When $n = 25$, we omit the results of the BASE-6 GRAPH because the BASE-6 GRAPH and BASE-5 GRAPH are equivalent.

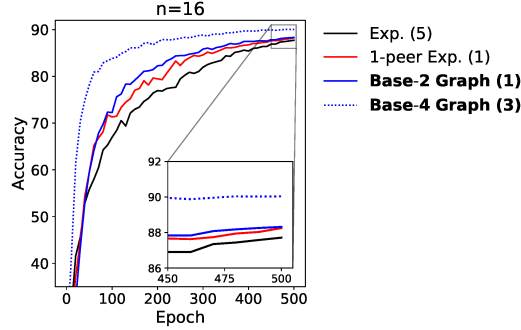


Figure B.16: Test accuracy (%) of DSGD with CIFAR-10 and $n = 16$. The number in the bracket is the maximum degree of a topology. We omit the results of the BASE-3 GRAPH and BASE-5 GRAPH because these graphs are equivalent to the BASE-2 GRAPH and BASE-4 GRAPH, respectively.

B.7 Results with Other Neural Network Architecture

In this section, we evaluate the effectiveness of the BASE- $(k + 1)$ GRAPH with other neural network architecture. Fig. B.17 shows the accuracy when we use ResNet-18 He et al. (2016). The results are consistent with the results with VGG-11 shown in Sec. 3.6.

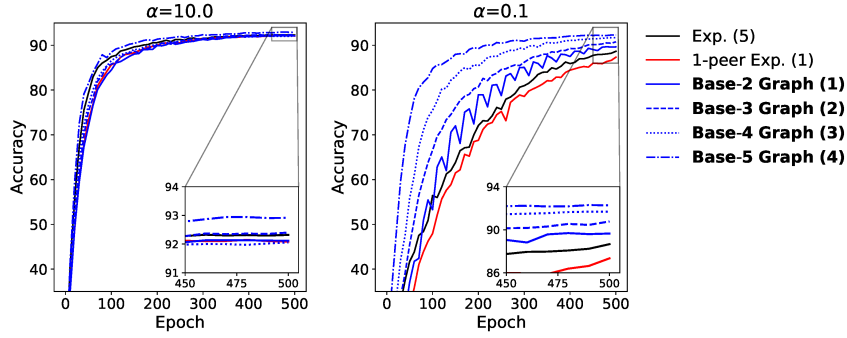


Figure B.17: Test accuracy (%) of DSGD with $n = 25$, CIFAR-10, and ResNet. The number in the bracket denotes the maximum degree of a topology.

B.8 Hyperparameter Setting

Tables B.2 and B.3 list the detailed hyperparameter settings used in Secs. 3.6 and B.6.3. We ran all experiments on a server with eight Nvidia RTX 3090 GPUs.

Table B.2: Hyperparameter settings for Fashion MNIST with LeNet.

Dataset	Fashion MNIST
Neural network architecture	LeNet (LeCun et al., 1998)
Data augmentation	RandomCrop of PyTorch
Step size	Grid search over $\{0.1, 0.01, 0.001\}$.
Momentum	0.9
Batch size	32
Step size scheduler	Cosine decay
Step size warmup	10 epochs
The number of epochs	200

Table B.3: Hyperparameter settings for CIFAR- $\{10, 100\}$ with $\{\text{VGG-11}, \text{ResNet-18}\}$.

Dataset	CIFAR- $\{10, 100\}$
Neural network architecture	$\{\text{VGG-11 (Simonyan and Zisserman, 2015)}, \text{ResNet-18 (He et al., 2016)}\}$
Data augmentation	RandomCrop , RandomHorizontalFlip , RandomErasing of PyTorch
Step size	Grid search over $\{0.1, 0.01, 0.001\}$.
Momentum	0.9
Batch size	32
Step size scheduler	Cosine decay
Step size warmup	10 epochs
The number of epochs	500

C

Appendix of Teleportation

C.1 Teleportation with Communication Overlap

In the implementation shown in Alg. 4, line 7 in Alg. 4 does not start until line 12 is completed. We can modify Alg. 4 so that the exchanges of parameters in lines 7 and 12 are performed simultaneously. We show the pseudo-code and illustration in Alg. 11 and Fig. C.1. The communication in lines 7 and 12 in Alg. 4 corresponds to that in lines 9 and 13 in Alg. 11.

Algorithm 11 TELEPORTATION

- 1: **Input:** the number of nodes n , set of nodes V_n , number of active nodes k , total number of iteration T , step size η , and topology comprising k active nodes $\mathcal{G}_k = (\{1, \dots, k\}, E)$.
- 2: sample active k nodes $V_{\text{active}}^{(0)}$ from V_n without replacement.
- 3: assign $\{1, 2, \dots, k\}$ to variables $\{\text{token_id}_i^{(0)} \mid v_i \in V_{\text{active}}^{(0)}\}$ randomly without overlap.
- 4: **for** $t \in \{0, 1, \dots, T\}$ **do**
- 5: sample next active k nodes $V_{\text{active}}^{(t+1)}$ from V_n without replacement, and assign $\{1, 2, \dots, k\}$ to variables $\{\text{token_id}_i^{(t+1)} \mid v_i \in V_{\text{active}}^{(t+1)}\}$ randomly without overlap.
- 6: **for** $i \in \{1, 2, \dots, n\}$ **in parallel do**
- 7: **if** $v_i \in V_{\text{active}}^{(t)}$ **then**^c
- 8: $\mathbf{y}_i^{(t)} = \mathbf{x}_i^{(t)} - \eta \nabla F_i(\mathbf{x}_i^{(t)}; \xi_i^{(t)})$.
- 9: send $\mathbf{y}_i^{(t)}$ to $v_j \in \{v_j \in V_{\text{active}}^{(t+1)} \mid (\text{token_id}_i^{(t)}, \text{token_id}_j^{(t+1)}) \in E\}$.
- 10: **if** $v_i \in V_{\text{active}}^{(t+1)}$ **then**
- 11: receive $\mathbf{y}_j^{(t)}$ from $v_j \in \{v_j \in V_{\text{active}}^{(t)} \mid (\text{token_id}_i^{(t+1)}, \text{token_id}_j^{(t)}) \in E\}$.
- 12: $\mathbf{x}_i^{(t+1)} = \sum_{v_j \in V_{\text{active}}^{(t)}} W_{\text{token_id}_i^{(t+1)}, \text{token_id}_j^{(t)}} \mathbf{y}_j^{(t)}$.

^cThe update rule of the parameters of the inactive nodes is not described because the parameters held by the inactive nodes are discarded and initialized with the parameters of the other active nodes when they are activated in line 13. The parameters of inactive nodes do not affect the behavior of active node parameters.

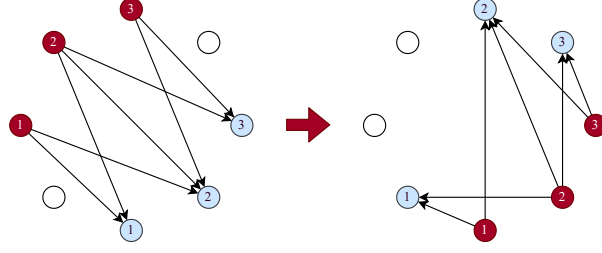


Figure C.1: Illustration of Alg. 11 with $n = 8$ and $k = 3$. We use a line as the topology consisting of active nodes $\mathcal{G}_k = (\{1, 2, 3\}, \{(1, 1), (1, 2), (2, 2), (2, 3), (3, 3)\})$. The black nodes represent active nodes, and the number written on the node is $\text{token_id}_i^{(t)}$. The blue nodes represent the next active nodes, and the number on the node is $\text{token_id}_i^{(t+1)}$.

C.2 Proof of Theorems 3 and 4

C.2.1 Notation

In this section, we introduce the notation used in the proof. TELEPORTATION has n parameters, $\{\mathbf{x}_i^{(t)}\}_{i=1}^n$, while the parameters of inactive nodes are not used and do not affect the behavior of the parameters of active nodes. To simplify the notation, we introduce the variables $\{\mathbf{z}_m^{(t)}\}_{m=1}^k$ that correspond to the active nodes parameters.

From line 5 in Alg. 11, $\text{text_id}_i^{(t)}$ has a different value for each active node $v_i \in V_{\text{active}}^{(t)}$. That is, there is one-to-one correspondence between $\{i \mid v_i \in V_{\text{active}}^{(t)}\}$ and $\{1, 2, \dots, k\}$. Let $g^{(t)} : \{i \mid v_i \in V_{\text{active}}^{(t)}\} \rightarrow \{1, 2, \dots, k\}$ be the function that takes node index i and returns $\text{token_id}_i^{(t)}$. Since $g^{(t)}$ is a bijection function, there exists an inverse function $(g^{(t)})^{-1} : \{1, 2, \dots, k\} \rightarrow \{i \mid v_i \in V_{\text{active}}^{(t)}\}$. Using this inverse function, we define variable $\text{node_id}_m^{(t)}$ as follows:

$$\text{node_id}_m^{(t)} := (g^{(t)})^{-1}(m),$$

for any $m \in \{1, 2, \dots, k\}$. $\text{node_id}_m^{(t)}$ stores the node index i whose $\text{token_id}_i^{(t)}$ stores m . Using this notation, we define $\mathbf{z}_m^{(t)} \in \mathbb{R}^d$ as follows:

$$\mathbf{z}_m^{(t)} = \mathbf{x}_{\text{node_id}_m^{(t)}}^{(t)},$$

for all $m \in \{1, 2, \dots, k\}$ and t . From the definition of $\mathbf{z}_m^{(t)}$, we can get its update rule

as follows:

$$\begin{aligned}
\mathbf{z}_m^{(t+1)} &= \mathbf{x}_{\text{node_id}_m^{(t+1)}}^{(t)} \\
&= \sum_{v_j \in V_{\text{active}}^{(t)}} W_{\text{token_id}_{\text{node_id}_m^{(t+1)}}, \text{token_id}_j^{(t)}} \left(\mathbf{x}_j^{(t)} - \eta \nabla F_j(\mathbf{x}_j^{(t)}; \xi_j^{(t)}) \right) \\
&= \sum_{v_j \in V_{\text{active}}^{(t)}} W_{m, \text{token_id}_j^{(t)}} \left(\mathbf{x}_j^{(t)} - \eta \nabla F_j(\mathbf{x}_j^{(t)}; \xi_j^{(t)}) \right) \\
&= \sum_{l=1}^k W_{m,l} \left(\mathbf{x}_{\text{node_id}_l^{(t)}}^{(t)} - \eta \nabla F_{\text{node_id}_l^{(t)}}(\mathbf{x}_{\text{node_id}_l^{(t)}}^{(t)}; \xi_{\text{node_id}_l^{(t)}}^{(t)}) \right) \\
&= \sum_{l=1}^k W_{m,l} \left(\mathbf{z}_l^{(t)} - \eta \nabla F_{\text{node_id}_l^{(t)}}(\mathbf{z}_l^{(t)}; \xi_{\text{node_id}_l^{(t)}}^{(t)}) \right).
\end{aligned}$$

In the next section, we analyze the convergence behavior of $\{\mathbf{z}_m^{(t)}\}_{m=1}^k$. Note that sets $\{\mathbf{z}_m^{(t)}\}_{m=1}^k$ and $\{\mathbf{x}_i^{(t)} \mid v_i \in V_{\text{active}}^{(t)}\}$ are equivalent, and their averages are equivalent, i.e.,

$$\frac{1}{k} \sum_{m=1}^k \mathbf{z}_m^{(t)} = \frac{1}{k} \sum_{m=1}^k \mathbf{x}_{\text{node_id}_m^{(t)}} = \frac{1}{k} \sum_{i=1}^n \mathbf{x}_i^{(t)} \mathbf{1}_{v_i \in V_{\text{active}}^{(t)}} = \frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \mathbf{x}_i^{(t)}, \quad (\text{C.1})$$

where $\mathbf{1}_{v_i \in V_{\text{active}}^{(t)}}$ is an indicator function. Furthermore, let $\mathbf{Z} \in \mathbb{R}^{d \times k}$, $\mathbf{G} \in \mathbb{R}^{d \times k}$, and $\nabla f(\mathbf{Z}) \in \mathbb{R}^{d \times k}$ as follows:

$$\begin{aligned}
\mathbf{Z}^{(t)} &:= (\mathbf{z}_1^{(t)}, \mathbf{z}_2^{(t)}, \dots, \mathbf{z}_k^{(t)}), \\
\mathbf{G}^{(t)} &:= \left(\nabla F_{\text{node_id}_1^{(t)}}(\mathbf{z}_1^{(t)}; \xi_{\text{node_id}_1^{(t)}}^{(t)}), \dots, \nabla F_{\text{node_id}_k^{(t)}}(\mathbf{z}_k^{(t)}; \xi_{\text{node_id}_k^{(t)}}^{(t)}) \right), \\
\nabla f(\mathbf{Z}^{(t)}) &:= (\nabla f(\mathbf{z}_1^{(t)}), \nabla f(\mathbf{z}_2^{(t)}), \dots, \nabla f(\mathbf{z}_k^{(t)})).
\end{aligned}$$

By using the above notation, Alg. 11 can be rewritten as follows:

$$\mathbf{Z}^{(t+1)} = \left(\mathbf{Z}^{(t)} - \eta \mathbf{G}^{(t)} \right) \mathbf{W}^\top.$$

C.2.2 Main Proof

Lemma 8. *Suppose that Assumptions 2, 3, 4, and 9 hold. If the step size satisfies $\eta \leq \frac{1}{4L}$, then it holds that*

$$\begin{aligned}
&\mathbb{E}f(\bar{\mathbf{z}}^{(t+1)}) \\
&\leq \mathbb{E}f(\bar{\mathbf{z}}^{(t)}) - \frac{\eta}{4} \mathbb{E} \left\| \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 + L^2 \eta \Xi^{(t)} + \frac{L\sigma^2}{2k} \eta^2 + \frac{L\zeta^2}{2k} \left(1 - \frac{k-1}{n-1} \right) \eta^2,
\end{aligned}$$

where $\bar{\mathbf{z}}^{(t)} := \frac{1}{k} \sum_{i=1}^k \mathbf{z}_i^{(t)}$ and $\Xi^{(t)} := \frac{1}{k} \sum_{i=1}^k \mathbb{E} \left\| \mathbf{z}_i^{(t)} - \bar{\mathbf{z}}^{(t)} \right\|^2$.

Proof. By calculating the average of the update rule of $\mathbf{z}_i^{(t)}$, we obtain

$$\begin{aligned}\bar{\mathbf{z}}^{(t+1)} &= \frac{1}{k} \sum_{m=1}^k \sum_{l=1}^k W_{m,l} \left(\mathbf{z}_l^{(t)} - \eta \nabla F_{\text{node_id}_l^{(t)}}(\mathbf{z}_l^{(t)}; \xi_{\text{node_id}_l^{(t)}}^{(t)}) \right) \\ &= \bar{\mathbf{z}}^{(t)} - \frac{\eta}{k} \sum_{l=1}^k \nabla F_{\text{node_id}_l^{(t)}}(\mathbf{z}_l^{(t)}; \xi_{\text{node_id}_l^{(t)}}^{(t)}),\end{aligned}$$

where we use $\sum_{m=1}^k W_{m,l} = 1$. Using Assumption 2, we get

$$\begin{aligned}\mathbb{E}_{t+1} f(\bar{\mathbf{z}}^{(t+1)}) &= \mathbb{E}_{t+1} f \left(\bar{\mathbf{z}}^{(t)} - \frac{\eta}{k} \sum_{j=1}^k \nabla F_{\text{node_id}_j^{(t)}}(\mathbf{z}_j^{(t)}; \xi_{\text{node_id}_j^{(t)}}^{(t)}) \right) \\ &\leq f(\bar{\mathbf{z}}^{(t)}) - \eta \left\langle \nabla f(\bar{\mathbf{z}}^{(t)}), \frac{1}{k} \sum_{j=1}^k \mathbb{E}_{t+1} \nabla F_{\text{node_id}_j^{(t)}}(\mathbf{z}_j^{(t)}; \xi_{\text{node_id}_j^{(t)}}^{(t)}) \right\rangle \\ &\quad + \frac{L\eta^2}{2} \mathbb{E}_{t+1} \left\| \frac{1}{k} \sum_{j=1}^k \nabla F_{\text{node_id}_j^{(t)}}(\mathbf{z}_j^{(t)}; \xi_{\text{node_id}_j^{(t)}}^{(t)}) \right\|^2 \\ &= f(\bar{\mathbf{z}}^{(t)}) - \underbrace{\eta \left\langle \nabla f(\bar{\mathbf{z}}^{(t)}), \frac{1}{k} \sum_{j=1}^k \nabla f(\mathbf{z}_j^{(t)}) \right\rangle}_{T_1} \\ &\quad + \underbrace{\frac{L\eta^2}{2} \mathbb{E}_{t+1} \left\| \frac{1}{k} \sum_{j=1}^k \nabla F_{\text{node_id}_j^{(t)}}(\mathbf{z}_j^{(t)}; \xi_{\text{node_id}_j^{(t)}}^{(t)}) \right\|^2}_{T_2},\end{aligned}$$

where we use the fact that active nodes are randomly selected, i.e., $\text{node_id}_j^{(t)}$ is randomly assigned to $\{1, 2, \dots, n\}$, in the last equality. T_1 and T_2 are bounded as follows:

$$\begin{aligned}-T_1 &= -\left\| \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 + \left\langle \nabla f(\bar{\mathbf{z}}^{(t)}), \frac{1}{k} \sum_{j=1}^k \nabla f(\mathbf{z}_j^{(t)}) - \nabla f(\bar{\mathbf{z}}^{(t)}) \right\rangle \\ &\leq -\frac{1}{2} \left\| \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 + \frac{1}{2} \left\| \frac{1}{k} \sum_{j=1}^k \nabla f(\mathbf{z}_j^{(t)}) - \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 \\ &\leq -\frac{1}{2} \left\| \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 + \frac{1}{2k} \sum_{j=1}^k \left\| \nabla f(\mathbf{z}_j^{(t)}) - \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 \\ &\leq -\frac{1}{2} \left\| \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 + \frac{L^2}{2k} \sum_{j=1}^k \left\| \mathbf{z}_j^{(t)} - \bar{\mathbf{z}}^{(t)} \right\|^2.\end{aligned}$$

$$\begin{aligned}
T_2 &= \mathbb{E}_{t+1} \left\| \frac{1}{k} \sum_{j=1}^k \nabla f_{\text{node_id}_j^{(t)}}(\mathbf{z}_j^{(t)}) \right\|^2 \\
&\quad + \mathbb{E}_{t+1} \left\| \frac{1}{k} \sum_{j=1}^k \nabla F_{\text{node_id}_j^{(t)}}(\mathbf{z}_j^{(t)}; \xi_{\text{node_id}_j^{(t)}}^{(t)}) - \nabla f_{\text{node_id}_j^{(t)}}(\mathbf{z}_j^{(t)}) \right\|^2 \\
&\leq \mathbb{E}_{t+1} \left\| \frac{1}{k} \sum_{j=1}^k \nabla f_{\text{node_id}_j^{(t)}}(\mathbf{z}_j^{(t)}) \right\|^2 + \frac{\sigma^2}{k} \\
&\leq \left\| \frac{1}{k} \sum_{j=1}^k \nabla f(\mathbf{z}_j^{(t)}) \right\|^2 + \mathbb{E}_{t+1} \left\| \frac{1}{k} \sum_{j=1}^k \nabla f_{\text{node_id}_j^{(t)}}(\mathbf{z}_j^{(t)}) - \nabla f(\mathbf{z}_j^{(t)}) \right\|^2 + \frac{\sigma^2}{k} \\
&\leq \left\| \frac{1}{k} \sum_{j=1}^k \nabla f(\mathbf{z}_j^{(t)}) \right\|^2 + \frac{\zeta^2}{k} \left(1 - \frac{k-1}{n-1}\right) + \frac{\sigma^2}{k} \\
&\leq 2 \left\| \frac{1}{k} \sum_{j=1}^k \nabla f(\mathbf{z}_j^{(t)}) - \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 + 2 \|\nabla f(\bar{\mathbf{z}}^{(t)})\|^2 + \frac{\zeta^2}{k} \left(1 - \frac{k-1}{n-1}\right) + \frac{\sigma^2}{k} \\
&\leq \frac{2L^2}{k} \sum_{j=1}^k \|\mathbf{z}_j^{(t)} - \bar{\mathbf{z}}^{(t)}\|^2 + 2 \|\nabla f(\bar{\mathbf{z}}^{(t)})\|^2 + \frac{\zeta^2}{k} \left(1 - \frac{k-1}{n-1}\right) + \frac{\sigma^2}{k},
\end{aligned}$$

where we use the fact that the active nodes $V_{\text{active}}^{(t)}$ are sampled from V_n without replacement in the third inequality. Using the above inequalities, we obtain

$$\begin{aligned}
\mathbb{E}_{t+1} f(\bar{\mathbf{z}}) &\leq f(\bar{\mathbf{z}}^{(t)}) - \frac{\eta}{2} \|\nabla f(\bar{\mathbf{z}}^{(t)})\|^2 + \frac{L^2 \eta}{2k} \sum_{j=1}^k \|\mathbf{z}_j^{(t)} - \bar{\mathbf{z}}^{(t)}\|^2 \\
&\quad + \frac{L^3 \eta^2}{k} \sum_{j=1}^k \|\mathbf{z}_j^{(t)} - \bar{\mathbf{z}}^{(t)}\|^2 + L\eta^2 \|\nabla f(\bar{\mathbf{z}}^{(t)})\|^2 + \frac{L\zeta^2}{2k} \left(1 - \frac{k-1}{n-1}\right) \eta^2 + \frac{L\sigma^2}{2k} \eta^2.
\end{aligned}$$

Using $\eta \leq \frac{1}{4L}$, we obtain the desired result. $\square$

Lemma 9. Suppose that Assumptions 2, 3, 4, and 9 hold. If the step size satisfies $\eta \leq \frac{p_k}{\sqrt{24L}}$, then it holds that

$$\Xi^{(t+1)} \leq (1 - \frac{p_k}{4}) \Xi^{(t)} + \frac{6(1-p_k)\eta^2}{p_k} \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(t)})\|^2 + (1-p_k)(\sigma^2 + \zeta^2)\eta^2,$$

where $\bar{\mathbf{z}}^{(t)} := \frac{1}{k} \sum_{i=1}^k \mathbf{z}_i^{(t)}$ and $\Xi^{(t)} := \frac{1}{k} \sum_{i=1}^k \mathbb{E} \|\mathbf{z}_i^{(t)} - \bar{\mathbf{z}}^{(t)}\|^2$.

Proof.

$$\begin{aligned}
& \mathbb{E}_{t+1} \left\| \mathbf{Z}^{(t+1)} - \bar{\mathbf{Z}}^{(t+1)} \right\|_F^2 \\
&= \mathbb{E}_{t+1} \left\| (\mathbf{Z}^{(t)} - \eta \mathbf{G}^{(t)}) \mathbf{W}^\top - (\bar{\mathbf{Z}}^{(t)} - \eta \bar{\mathbf{G}}^{(t)}) \right\|_F^2 \\
&\leq (1 - p_k) \mathbb{E}_{t+1} \left\| (\mathbf{Z}^{(t)} - \eta \mathbf{G}^{(t)}) - (\bar{\mathbf{Z}}^{(t)} - \eta \bar{\mathbf{G}}^{(t)}) \right\|_F^2 \\
&\leq (1 - p_k) \mathbb{E}_{t+1} \left\| (\mathbf{Z}^{(t)} - \eta \mathbf{G}^{(t)}) - \bar{\mathbf{Z}}^{(t)} \right\|_F^2 \\
&\leq (1 - p_k) \left\| (\mathbf{Z}^{(t)} - \eta \nabla f(\mathbf{Z}^{(t)})) - \bar{\mathbf{Z}}^{(t)} \right\|_F^2 + (1 - p_k) \eta^2 \mathbb{E}_{t+1} \left\| \mathbf{G}^{(t)} - f(\mathbf{Z}^{(t)}) \right\|^2 \\
&\leq (1 - p_k) \left\| (\mathbf{Z}^{(t)} - \eta \nabla f(\mathbf{Z}^{(t)})) - \bar{\mathbf{Z}}^{(t)} \right\|_F^2 + (1 - p_k) k(\sigma^2 + \zeta^2) \eta^2 \\
&\leq (1 - \frac{p_k}{2}) \left\| \mathbf{Z}^{(t)} - \bar{\mathbf{Z}}^{(t)} \right\|_F^2 + \frac{3(1 - p_k) \eta^2}{p_k} \underbrace{\left\| \nabla f(\mathbf{Z}^{(t)}) \right\|_F^2}_T + (1 - p_k) k(\sigma^2 + \zeta^2) \eta^2.
\end{aligned}$$

T is bounded as follows:

$$\begin{aligned}
T &\leq 2 \left\| \nabla f(\mathbf{Z}^{(t)}) - \nabla f(\bar{\mathbf{Z}}^{(t)}) \right\|_F^2 + 2 \left\| \nabla f(\bar{\mathbf{Z}}^{(t)}) \right\|_F^2 \\
&\leq 2L^2 \left\| \mathbf{Z}^{(t)} - \bar{\mathbf{Z}}^{(t)} \right\|_F^2 + 2k \left\| \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2,
\end{aligned}$$

where $\bar{\mathbf{Z}}^{(t)} := \frac{1}{k} \mathbf{Z}^{(t)} \mathbf{1} \mathbf{1}^\top$. Then, we get

$$\begin{aligned}
& \mathbb{E}_{t+1} \left\| \mathbf{Z}^{(t+1)} - \bar{\mathbf{Z}}^{(t+1)} \right\|_F^2 \\
&\leq (1 - \frac{p_k}{2}) \left\| \mathbf{Z}^{(t)} - \bar{\mathbf{Z}}^{(t)} \right\|_F^2 + \frac{6(1 - p_k) L^2 \eta^2}{p_k} \left\| \mathbf{Z}^{(t)} - \bar{\mathbf{Z}}^{(t)} \right\|_F^2 \\
&\quad + \frac{6(1 - p_k) k \eta^2}{p_k} \left\| \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 + (1 - p_k) k(\sigma^2 + \zeta^2) \eta^2.
\end{aligned}$$

Using $\eta \leq \frac{p_k}{\sqrt{24}L}$, we obtain the desired result. $\square$

Lemma 10. Suppose that Assumptions 2, 3, 4, and 9 hold, and $\{\mathbf{x}_i^{(0)}\}_{i=1}^n$ is initialized with the same parameter $\bar{\mathbf{x}}^{(0)}$. If the step size satisfies $\eta \leq \frac{p_k}{\sqrt{24}L}$, then it holds that

$$\frac{1}{T+1} \sum_{t=0}^T \Xi^{(t)} \leq \frac{24(1 - p_k)}{p_k^2(T+1)} \eta^2 \sum_{t=0}^T \mathbb{E} \left\| \nabla f(\bar{\mathbf{z}}^{(t)}) \right\|^2 + \frac{4(1 - p_k)(\sigma^2 + \zeta^2)}{p_k} \eta^2,$$

where $\bar{\mathbf{z}}^{(t)} := \frac{1}{k} \sum_{i=1}^k \mathbf{z}_i^{(t)}$ and $\Xi^{(t)} := \frac{1}{k} \sum_{i=1}^k \mathbb{E} \left\| \mathbf{z}_i^{(t)} - \bar{\mathbf{z}}^{(t)} \right\|^2$.

Proof. From Lemma 9, we get

$$\begin{aligned}
& \Xi^{(t+1)} \\
& \leq \frac{6(1-p_k)\eta^2}{p_k} \sum_{l=0}^t \left(1 - \frac{p_k}{4}\right)^{t-l} \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(l)})\|^2 + (1-p_k)(\sigma^2 + \zeta^2)\eta^2 \sum_{l=0}^t \left(1 - \frac{p_k}{4}\right)^{t-l} \\
& \leq \frac{6(1-p_k)\eta^2}{p_k} \sum_{l=0}^t \left(1 - \frac{p_k}{4}\right)^{t-l} \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(l)})\|^2 + \frac{4(1-p_k)(\sigma^2 + \zeta^2)}{p_k} \eta^2.
\end{aligned}$$

By summing up the above inequality from $t = 0$ to $T - 1$, we obtain

$$\begin{aligned}
& \frac{1}{T+1} \sum_{t=1}^T \Xi^{(t)} \\
& \leq \frac{6(1-p_k)\eta^2}{p_k(T+1)} \sum_{t=1}^T \sum_{l=0}^{t-1} \left(1 - \frac{p_k}{4}\right)^{t-l} \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(l)})\|^2 + \frac{4(1-p_k)(\sigma^2 + \zeta^2)}{p_k} \eta^2 \\
& = \frac{6(1-p_k)\eta^2}{p_k(T+1)} \sum_{l=0}^{T-1} \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(l)})\|^2 \sum_{t=l+1}^T \left(1 - \frac{p_k}{4}\right)^{t-l} + \frac{4(1-p_k)(\sigma^2 + \zeta^2)}{p_k} \eta^2 \\
& = \frac{6(1-p_k)\eta^2}{p_k(T+1)} \sum_{l=0}^{T-1} \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(l)})\|^2 \sum_{t'=1}^{T-l} \left(1 - \frac{p_k}{4}\right)^{t'} + \frac{4(1-p_k)(\sigma^2 + \zeta^2)}{p_k} \eta^2 \\
& \leq \frac{24(1-p_k)\eta^2}{p_k^2(T+1)} \sum_{l=0}^{T-1} \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(l)})\|^2 + \frac{4(1-p_k)(\sigma^2 + \zeta^2)}{p_k} \eta^2.
\end{aligned}$$

Using $\Xi^{(0)} = 0$ and $\mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(T)})\|^2 \geq 0$, we obtain the desired result. $\square$

Lemma 11. Suppose that Assumptions 2, 3, 4, and 9 hold, and $\{\mathbf{x}_i^{(0)}\}_{i=1}^n$ is initialized with the same parameter $\bar{\mathbf{x}}^{(0)}$. If the step size satisfies $\eta \leq \frac{p_k}{\sqrt{192L}}$, then it holds that

$$\begin{aligned}
& \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2 \\
& \leq \frac{8(f(\bar{\mathbf{x}}^{(0)}) - f^*)}{(T+1)\eta} + \frac{32L^2(\sigma^2 + \zeta^2)(1-p_k)}{p_k} \eta^2 + \frac{4L}{k} \left(\sigma^2 + \left(1 - \frac{k-1}{n-1}\right) \zeta^2 \right) \eta,
\end{aligned} \tag{C.2}$$

where $\bar{\mathbf{x}}_{\text{active}}^{(t)} := \frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \mathbf{x}_i^{(t)}$.

Proof. Rearranging Lemma 8, we obtain

$$\begin{aligned}
& \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(t)})\|^2 \\
& \leq \frac{4(\mathbb{E} f(\bar{\mathbf{z}}^{(t)}) - \mathbb{E} f(\bar{\mathbf{z}}^{(t+1)}))}{\eta} + 4L^2 \Xi^{(t)} + \frac{2L\sigma^2}{k} \eta + \frac{2L\zeta^2}{k} \left(1 - \frac{k-1}{n-1}\right) \eta.
\end{aligned}$$

Using Lemma 10, we get

$$\begin{aligned}
& \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(t)})\|^2 \\
& \leq \frac{4(f(\bar{\mathbf{z}}^{(0)}) - f^*)}{(T+1)\eta} + \frac{4L^2}{T+1} \Xi^{(t)} + \frac{2L\sigma^2}{k} \eta + \frac{2L\zeta^2}{k} \left(1 - \frac{k-1}{n-1}\right) \eta \\
& \leq \frac{4(f(\bar{\mathbf{z}}^{(0)}) - f^*)}{(T+1)\eta} + \frac{96L^2(1-p_k)}{p_k^2(T+1)} \eta^2 \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{z}}^{(t)})\|^2 \\
& \quad + \frac{16L^2(1-p_k)(\sigma^2 + \zeta^2)}{p_k} \eta^2 + \frac{2L\sigma^2}{k} \eta + \frac{2L\zeta^2}{k} \left(1 - \frac{k-1}{n-1}\right) \eta.
\end{aligned}$$

Using $\eta \leq \frac{p_k}{\sqrt{192L}}$ and $\bar{\mathbf{x}}_{\text{active}}^{(t)} = \bar{\mathbf{z}}^{(t)}$ (see Eq. (C.1)), we obtain the desired result. $\square$

Lemma 12. Suppose that Assumptions 2, 3, 4, and 9 hold, and $\{\mathbf{x}_i^{(0)}\}_{i=1}^n$ is initialized with the same parameter $\bar{\mathbf{x}}^{(0)}$. There exists a step size η that satisfies

$$\begin{aligned}
& \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2 \\
& \leq \mathcal{O} \left(\sqrt{\frac{Lr_0(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}{kT}} + \left(\frac{L^2r_0^2(\sigma^2 + \zeta^2)(1-p_k)}{T^2p_k} \right)^{\frac{1}{3}} + \frac{Lr_0}{Tp_k} \right),
\end{aligned} \tag{C.3}$$

where $\bar{\mathbf{x}}_{\text{active}}^{(t)} := \frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \mathbf{x}_i^{(t)}$ and $r_0 := f(\bar{\mathbf{x}}^{(0)}) - f^*$.

Proof. Choosing the step size as follows:

$$\eta = \min \left\{ \sqrt{\frac{2kr_0}{L(T+1)(\sigma^2 + \zeta^2)}}, \left(\frac{r_0p_k}{4L^2(T+1)(1-p_k)(\sigma^2 + \zeta^2)} \right)^{\frac{1}{3}}, \frac{p_k}{\sqrt{192L}} \right\},$$

we have three cases:

- When $\eta = \sqrt{\frac{2kr_0}{L(T+1)(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}}$, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\sqrt{\frac{128Lr_0(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}{k(T+1)}} + \left(\frac{2048L^2r_0^2(1-p_k)(\sigma^2 + \zeta^2)}{(T+1)^2p_k} \right)^{\frac{1}{3}}.$$

- When $\eta = \left(\frac{r_0p_k}{4L^2(T+1)(1-p_k)(\sigma^2 + \zeta^2)} \right)^{\frac{1}{3}}$, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\sqrt{\frac{32Lr_0(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}{k(T+1)}} + \left(\frac{16384L^2r_0^2(1-p_k)(\sigma^2 + \zeta^2)}{(T+1)^2p_k} \right)^{\frac{1}{3}}.$$

- When $\eta = \frac{p_k}{\sqrt{192L}}$, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\begin{aligned} & \frac{\sqrt{12288}Lr_0}{(T+1)p_k} + \sqrt{\frac{32Lr_0(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}{k(T+1)}} \\ & + \left(\frac{2048L^2r_0^2(1-p_k)(\sigma^2 + \zeta^2)}{(T+1)^2p_k} \right)^{\frac{1}{3}}. \end{aligned}$$

□

Lemma 13. Suppose that Assumptions 2, 3, 4, and 9 hold, and $\{\mathbf{x}_i^{(0)}\}_{i=1}^n$ is initialized with the same parameter $\bar{\mathbf{x}}^{(0)}$. Then, if the active nodes are connected by a ring, i.e., $p_k = \Omega(k^{-2})$, there exists a step size $\eta > 0$ and the number of active nodes $k \in \{1, 2, \dots, n\}$ that satisfies

$$\begin{aligned} & \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2 \\ & \leq \mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{Lr_0(\sigma^2 + \zeta^2)^{\frac{3}{4}}}{T} \right)^{\frac{4}{7}} + \left(\frac{Lr_0(\sigma^2 + \zeta^2)^{\frac{2}{5}}}{T} \right)^{\frac{5}{7}} + \frac{Lr_0}{T} \right), \end{aligned}$$

where $\bar{\mathbf{x}}_{\text{active}}^{(t)} := \frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \mathbf{x}_i^{(t)}$ and $r_0 := f(\bar{\mathbf{x}}^{(0)}) - f^*$.

Proof. From Lemma 12, we obtain

$$\begin{aligned} & \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2 \\ & \leq \mathcal{O} \left(\sqrt{\frac{Lr_0(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}{kT}} + \left(\frac{L^2r_0^2(\sigma^2 + \zeta^2)k^2}{T^2} \right)^{\frac{1}{3}} + \frac{Lr_0k^2}{T} \right). \end{aligned} \quad (\text{C.4})$$

For simplicity, let A , B , and C denote as follows:

$$A = \frac{Lr_0(\sigma^2 + \zeta^2)}{T}, \quad B = \left(\frac{L^2r_0^2(\sigma^2 + \zeta^2)}{T^2} \right)^{\frac{1}{3}}, \quad C = \frac{Lr_0}{T}.$$

Using these notations, we obtain

$$\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2 \leq \mathcal{O} \left(\sqrt{\frac{A}{k}} + Bk^{\frac{2}{3}} + Ck^2 \right). \quad (\text{C.5})$$

We choose k as follows:

$$k = \max \left\{ 1, \min \left\{ \left\lceil \left(\frac{A^3}{B^6} \right)^{\frac{1}{7}} \right\rceil, n \right\} \right\}.$$

Note that it holds that $k \in \{1, 2, 3, \dots, n\}$.

When $r_0(\sigma^2 + \zeta^2) = 0$, it holds that $k = 1$, $A = 0$, and $B = 0$. In this case, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\mathcal{O}\left(\frac{Lr_0}{T}\right).$$

When $r_0(\sigma^2 + \zeta^2) > 0$, it holds that

$$k = \min \left\{ \left\lceil \left(\frac{A^3}{B^6} \right)^{\frac{1}{7}} \right\rceil, n \right\}.$$

In this case, we have three cases:

- When $k = \left\lceil \left(\frac{A^3}{B^6} \right)^{\frac{1}{7}} \right\rceil$, we have

$$\sqrt{\frac{A}{k}} = \mathcal{O}\left(A^{\frac{2}{7}} B^{\frac{3}{7}}\right), \quad Bk^{\frac{2}{3}} = \mathcal{O}\left(A^{\frac{2}{7}} B^{\frac{3}{7}}\right), \quad Ck^2 = \mathcal{O}\left(CA^{\frac{6}{7}} B^{-\frac{12}{7}}\right).$$

By using the above inequalities, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\mathcal{O}\left(A^{\frac{2}{7}} B^{\frac{3}{7}} + A^{\frac{6}{7}} B^{-\frac{12}{7}} C\right).$$

- When $k = n$, we have

$$\sqrt{\frac{Lr_0(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}{kT}} = \sqrt{\frac{Lr_0\sigma^2}{nT}},$$

$$Bk^{\frac{2}{3}} \leq \mathcal{O}\left(A^{\frac{2}{7}} B^{\frac{3}{7}}\right), \quad Ck^2 \leq \mathcal{O}\left(CA^{\frac{6}{7}} B^{-\frac{12}{7}}\right),$$

where we use $k \leq \left\lceil \left(\frac{A^3}{B^6} \right)^{\frac{1}{7}} \right\rceil$. By using the above inequalities, we have

$$\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2 \leq \mathcal{O}\left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + A^{\frac{2}{7}} B^{\frac{3}{7}} + A^{\frac{6}{7}} B^{-\frac{12}{7}} C\right).$$

By summarizing the above inequalities, we obtain the desired results. $\square$

Lemma 14. Suppose that Assumptions 2, 3, 4, and 9 hold, and $\{\mathbf{x}_i^{(0)}\}_{i=1}^n$ is initialized with the same parameter $\bar{\mathbf{x}}^{(0)}$. Then, if the active nodes are connected by

an exponential graph, i.e., $p_k = \Omega(\log_2^{-1} k)$, there exists a step size $\eta > 0$ and the number of active nodes $k \in \{1, 2, \dots, n\}$ that satisfies

$$\begin{aligned} & \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2 \\ & \leq \mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{L^2 r_0^2 (\sigma^2 + \zeta^2)}{T^2} \log_2 \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right)^{\frac{1}{3}} \right)^{\frac{1}{3}} \right. \\ & \quad \left. + \frac{Lr_0}{T} \log_2 \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right) + \frac{Lr_0}{T} \right), \end{aligned}$$

where $\bar{\mathbf{x}}_{\text{active}}^{(t)} := \frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \mathbf{x}_i^{(t)}$ and $r_0 := f(\bar{\mathbf{x}}^{(0)}) - f^*$.

Proof. From Lemma 12, we obtain

$$\begin{aligned} & \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2 \\ & \leq \mathcal{O} \left(\sqrt{\frac{Lr_0(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}{kT}} + \left(\frac{L^2 r_0^2 (\sigma^2 + \zeta^2) \log_2(k)}{T^2} \right)^{\frac{1}{3}} + \frac{Lr_0 \log_2(k)}{T} \right). \end{aligned}$$

For simplicity, let A , B , and C denote as follows:

$$A = \frac{Lr_0(\sigma^2 + \zeta^2)}{T}, \quad B = \left(\frac{L^2 r_0^2 (\sigma^2 + \zeta^2)}{T^2} \right)^{\frac{1}{3}}, \quad C = \frac{Lr_0}{T}.$$

Using these notations, we obtain

$$\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2 \leq \mathcal{O} \left(\sqrt{\frac{A}{k}} + B \log_2^{\frac{1}{3}}(k) + C \log_2(k) \right).$$

We choose k as follows:

$$k = \max \left\{ 1, \min \left\{ \left\lceil \frac{A}{B^2} \right\rceil, \left\lceil \frac{A}{C^2} \right\rceil, n \right\} \right\}.$$

When $r_0(\sigma^2 + \zeta^2) = 0$, it holds that $k = 1$, $A = 0$, and $B = 0$. In this case, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\mathcal{O} \left(\frac{Lr_0}{T} \right).$$

When $r_0(\sigma^2 + \zeta^2) > 0$, it holds that

$$k = \min \left\{ \left\lceil \frac{A}{B^2} \right\rceil, \left\lceil \frac{A}{C^2} \right\rceil, n \right\}.$$

In this case, we have three cases:

- When $k = \left\lceil \frac{A}{B^2} \right\rceil$, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\mathcal{O} \left(B \log_2^{\frac{1}{3}} \left(\frac{A}{B^2} \right) + C \log_2 \left(\frac{A}{C^2} \right) \right).$$

- When $k = \left\lceil \frac{A}{C^2} \right\rceil$, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\mathcal{O} \left(B \log_2^{\frac{1}{3}} \left(\frac{A}{B^2} \right) + C \log_2 \left(\frac{A}{C^2} \right) \right).$$

- When $k = n$, $\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active}}^{(t)})\|^2$ is bounded from above by

$$\mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + B \log_2^{\frac{1}{3}} \left(\frac{A}{B^2} \right) + C \log_2 \left(\frac{A}{C^2} \right) \right).$$

By summarizing the above inequalities, we obtain the desired result. $\square$

C.3 Proof of Theorem 5

Lemma 1. *For any $k^* < n$, there exists $k \in \{1, 2, 4, 8, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}\}$ that satisfies $\frac{k^*}{4} < k \leq k^*$. Furthermore, it holds that $\sum_{i=0}^{\lfloor \log_2(n+1) \rfloor - 1} 2^i \leq n$.*

Proof. We define $\mathcal{K} := \{1, 2, 4, 8, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}\}$. We have two cases:

- When $1 \leq k^* < 2^{\lfloor \log_2(n+1) \rfloor}$, there exists $k \in \mathcal{K}$ that satisfies $k \leq k^* < 2k$.
- When $2^{\lfloor \log_2(n+1) \rfloor} \leq k^* < n$, it holds that $k < k^*$ for all $k \in \mathcal{K}$. Then, we have

$$k^* < n < 4 \times 2^{\lfloor \log_2(n+1) \rfloor - 1}.$$

Thus, there exists $k \in \mathcal{K}$ that satisfies $k < k^* < 4k$.

By combining the above two cases, we obtain the desired result. $\square$

Lemma 11. *Suppose that Assumptions 1, 2, 3, 4, and 9 hold. Let $\{\{\mathbf{x}_{k,i}^{(t)}\}_{v_i \in V_{\text{active}}^{(t)}}\}_{t=0}^T$ denote the parameters of active nodes generated by Alg. 11 when the number of active nodes is set to k and define*

$$\mathcal{K} := \{1, 2, 2^2, 2^3, \dots, 2^{\lfloor \log_2(n+1) \rfloor - 1}, n\}.$$

Then, suppose that the parameters are initialized with the same parameter $\bar{\mathbf{x}}^{(0)}$.

Ring: If the active nodes are connected by a ring, i.e., $p_k = \Omega(k^{-2})$, there exists η such that $\min_{k \in \mathcal{K}} \left(\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active},k}^{(t)})\|^2 \right)$ is bounded from above by

$$\mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{Lr_0(\sigma^2 + \zeta^2)^{\frac{3}{4}}}{T} \right)^{\frac{4}{7}} + \left(\frac{Lr_0(\sigma^2 + \zeta^2)^{\frac{2}{3}}}{T} \right)^{\frac{3}{5}} + \frac{Lr_0}{T} \right),$$

where $r_0 := f(\bar{\mathbf{x}}_{\text{active},k}^{(0)}) - f^*$ and $\bar{\mathbf{x}}_{\text{active},k}^{(t)} := \frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \mathbf{x}_{k,i}$.

Exp. Graph: If the active nodes are connected by an exponential graph, i.e., $p_k = \Omega(\log_2^{-1} k)$, there exists η such that $\min_{k \in \mathcal{K}} \left(\frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}_{\text{active},k}^{(t)})\|^2 \right)$ is bounded from above by

$$\mathcal{O} \left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{L^2 r_0^2 (\sigma^2 + \zeta^2)}{T^2} \log_2 \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right)^{\frac{1}{3}} \right)^{\frac{1}{3}} + \frac{Lr_0}{T} \log_2 \left(\frac{T(\sigma^2 + \zeta^2)}{Lr_0} \right) + \frac{Lr_0}{T} \right).$$

Proof. By combining Lemmas 1, 13, and 14, we obtain the statement. $\square$

C.4 Convergence Rate of Decentralized SGD with Various Topologies

Table C.1: Convergence rates of DSGD over various topologies.

Topology	Convergence Rate
Ring (Nedic et al., 2018)	$\mathcal{O}\left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{L^2r_0^2n^2(\sigma^2+n^2\zeta^2)}{T^2}\left(1-\frac{1}{n^2}\right)\right)^{\frac{1}{3}} + \frac{Lr_0n^2}{T}\right)$
Torus (Nedic et al., 2018)	$\mathcal{O}\left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{L^2r_0^2n(\sigma^2+n\zeta^2)}{T^2}\left(1-\frac{1}{n}\right)\right)^{\frac{1}{3}} + \frac{Lr_0n}{T}\right)$
Exponential Graph (Ying et al., 2021)	$\mathcal{O}\left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{L^2r_0^2\log_3(n)(\sigma^2+\log_2(n)\zeta^2)}{T^2}\left(1-\frac{1}{\log_2(n)}\right)\right)^{\frac{1}{3}} + \frac{Lr_0\log_2(n)}{T}\right)$
Base-2 Graph (Takezawa et al., 2023b)	$\mathcal{O}\left(\sqrt{\frac{Lr_0\sigma^2}{nT}} + \left(\frac{L^2r_0^2\log_2(n)(\sigma^2+\log_2(n)\zeta^2)}{T^2}\right)^{\frac{1}{3}} + \frac{Lr_0\log_2(n)}{T}\right)$

C.5 Convergence Rate of Decentralized SGD with Client Sampling

Proposition 2 (Liu et al. (2022)). *Suppose that Assumptions 1, 2, 3, 4, and 8 hold. Let k be the number of active nodes and $\{\{\mathbf{x}_i^{(t)}\}_{i=1}^n\}_{t=0}^T$ denote the parameters generated by Decentralized SGD with client sampling. Then, there exists the step size η that satisfies:*

$$\begin{aligned} & \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|^2 \\ & \leq \mathcal{O}\left(\sqrt{\frac{Lr_0(\sigma^2 + (1 - \frac{k-1}{n-1})\zeta^2)}{kT}} + \left(\frac{n}{k} \frac{Lr_0(\sigma\sqrt{p_n} + \zeta^2)}{Tp_n}\right)^{\frac{2}{3}} + \frac{Lr_0}{Tp_n}\right), \end{aligned}$$

where $r_0 := f(\bar{\mathbf{x}}^{(0)}) - f^*$ and $\bar{\mathbf{x}}^{(t)} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(t)}$.

Remark 5. *The convergence rate in Proposition 2 consistently deteriorates as k decreases.*

Unlike TELEPORTATION, the convergence rate shown in Proposition 2 depends on p_n , and the second and third terms degrade as n increases.

C.6 Additional Experiment

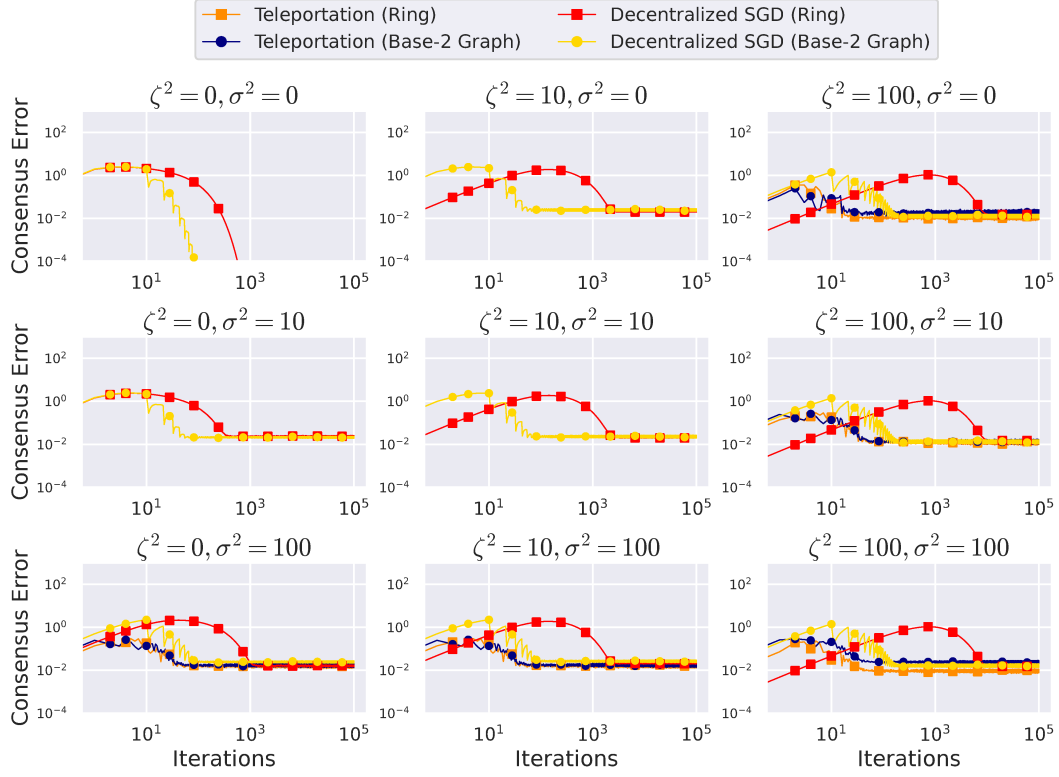


Figure C.2: The behavior of consensus error $\frac{1}{k} \sum_{v_i \in V_{\text{active}}^{(t)}} \|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}_{\text{active}}^{(t)}\|^2$ for different stochastic noise σ^2 and heterogeneity ζ^2 . The experimental setup was the same as in Fig. 4.2. Note that when the optimal number of active nodes k is one, we did not plot the consensus error. See Table 4.1.

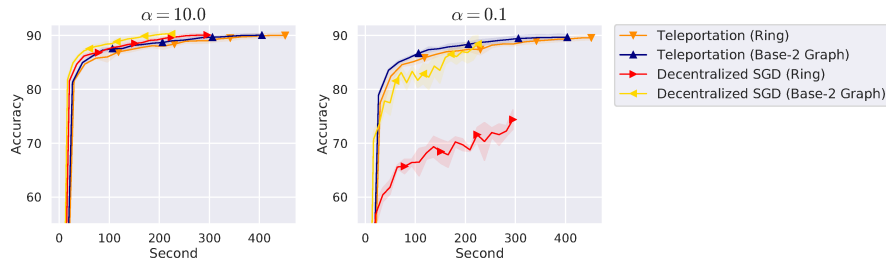


Figure C.3: Test accuracy of TELEPORTATION and Decentralized SGD under heterogeneous networks with $\tau = 0$. The experimental setup was the same as in Fig. 4.4.

C.7 Experimental Setup

Table C.2: Experimental setups for Fig. 4.2.

Step size	Grid search over $\{0.1, 0.075, 0.05, 0.025, 0.01, \dots, 0.0001\}$
Computational resources	AMD Epyc 7702 CPU or Intel Xeon Gold 6230 CPU

Table C.3: Experimental setups for Fashion MNIST in Fig. 4.3.

Model	LeNet
Step size	Grid search over $\{0.1, 0.01, 0.001\}$
Batch size	32
Momentum	0.9
Epoch	200
Computational resources	Titan $\times 8$

Table C.4: Experimental setups for CIFAR-10 in Fig. 4.3.

Model	VGG
Step size	Grid search over $\{0.1, 0.01, 0.001\}$
Scheduler	Cosine decay
Batch size	32
Momentum	0.9
Epoch	500
Computational resources	A6000 $\times 8$ or RTX 3090 $\times 8$

Table C.5: Experimental setups for Figs. 4.4 and C.3.

Model	LeNet
Step size	Grid search over $\{0.1, 0.01, 0.001\}$
Batch size	32
Momentum	0.9
Epoch	200
Computational resources	A6000 $\times 8$

D

Appendix of Stabilized Proximal Decentralized Optimization

D.1 Comparison of Communication and Computational Complexities in Convex Case

Table D.1: Summary of communication communication and computational complexities in the convex setting.

Algorithm	Reference	# Communication	# Computation	Assumptions
Gradient Tracking	Koloskova et al. (2021)	$\tilde{\mathcal{O}}\left(\frac{Ld_0^2}{(1-\rho)^2\epsilon}\right)^{(a)}$	$\tilde{\mathcal{O}}\left(\frac{Ld_0^2}{(1-\rho)^2\epsilon}\right)^{(a)}$	5, 2, 11
Exact-PDO (SONATA)	Theorems 6, 7 ^(b)	$\tilde{\mathcal{O}}\left(\frac{\delta d_0^2}{(1-\rho)\epsilon}\right)$	n.a. ^(c)	5, 2, 11
Inexact-PDO	Theorems 6, 7	$\tilde{\mathcal{O}}\left(\frac{\delta d_0^2}{(1-\rho)\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sqrt{\delta}Ld_0^2}{(1-\rho)\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)^{(b)}$	5, 2, 11
Stabilized-PDO [new]	Theorems 8, 9	$\tilde{\mathcal{O}}\left(\frac{\delta d_0^2}{(1-\rho)\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sqrt{\delta}Ld_0^2}{(1-\rho)\epsilon}\right)^{(b)}$	5, 2, 11
Accelerated SONATA	Tian et al. (2022)	-	n.a. ^(c)	-
Inexact Accelerated SONATA	Tian et al. (2022)	-	-	-
Accelerated Stabilized-PDO [new]	Theorems 10, 11	$\tilde{\mathcal{O}}\left(\sqrt{\frac{\delta d_0^2}{(1-\rho)\epsilon}}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{Ld_0^2}{(1-\rho)\epsilon}}\right)$	5, 2, 11

(a) It holds that $\delta \leq 2L$, and $\delta \ll 2L$ holds when nodes have similar datasets.

(b) Sun et al. (2022) did not analyze the communication complexity of SONATA in the convex case, while our new theorem, Theorem 6, is the first to analyze it.

(c) “n.a.” represents that we need to solve the subproblem exactly.

(d) δ is smaller than $\delta_{\max}$ and can be at most $\sqrt{n}$ times smaller than $\delta_{\max}$.

D.2 Technical Lemmas

D.2.1 Useful Lemmas

Lemma 12. For any $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^2$ and $\mathbf{b} \in \mathbb{R}$, it holds that

$$\frac{1}{n} \sum_{i=1}^n \|\mathbf{a}_i - \mathbf{b}\|^2 = \|\bar{\mathbf{a}} - \mathbf{b}\|^2 + \frac{1}{n} \sum_{i=1}^n \|\mathbf{a}_i - \bar{\mathbf{a}}\|^2. \quad (\text{D.1})$$

Proof. See Lemma 14 in Jiang et al. (2024a). $\square$

Lemma 13. Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ be vectors in $\mathbb{R}^d$ and $\mathbf{W} \in \mathbb{R}^{n \times n}$ be a matrix that satisfies $\sum_{i=1}^n W_{ij} = 1$. Let $\tilde{\mathbf{x}}_i := \sum_{j=1}^n W_{ij} \mathbf{x}_j$. Then, it holds that

$$\sum_{i=1}^n \mathbf{x}_i = \sum_{i=1}^n \tilde{\mathbf{x}}_i.$$

Proof. We have

$$\sum_{i=1}^n \tilde{\mathbf{x}}_i = \sum_{i=1}^n \sum_{j=1}^n W_{ij} \mathbf{x}_j = \sum_{j=1}^n \left(\sum_{i=1}^n W_{ij} \right) \mathbf{x}_j = \sum_{j=1}^n \mathbf{x}_j.$$

This concludes the statement. $\square$

Lemma 14. For any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and any matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ that satisfies Assumption 11, it holds that

$$\|(\mathbf{W} - \mathbf{I})\mathbf{X}\|_F \leq 4 \|\mathbf{X}\|_F. \quad (\text{D.2})$$

Proof. We have

$$\begin{aligned} \|(\mathbf{W} - \mathbf{I})\mathbf{X}\|_F &\leq \left\| \left(\mathbf{W} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \right) \mathbf{X} \right\|_F + \left\| \left(\frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top - \mathbf{I} \right) \mathbf{X} \right\|_F \\ &\stackrel{(5.3)}{\leq} (\rho + 1) \left\| \left(\frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top - \mathbf{I} \right) \mathbf{X} \right\|_F \\ &\leq 2 \left\| \mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \right\|_{\text{op}} \|\mathbf{X}\|_F, \end{aligned}$$

where $\mathbf{1}_n$ is n -dimensional vector with all ones and we use $\rho \leq 1$ in the last inequality. Let $D(1 - \frac{1}{n}, 1 - \frac{1}{n}) \in \mathbb{R}$ be a closed disc centered at $1 - \frac{1}{n}$ with radius $1 - \frac{1}{n}$. From Gershgorin circle theorem, every eigenvalue of $\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top$ lies within the discs $D(1 - \frac{1}{n}, 1 - \frac{1}{n})$. This implies that the absolute values of eigenvalues are less than 2. This concludes the statement. $\square$

Lemma 15. Let $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function that satisfies Assumptions 2 and $\mathbf{W} \in \mathbb{R}^{n \times n}$ be a matrix that satisfies Assumption 11. Let $\{\{\mathbf{x}_i^{(r)}\}_{i=1}^n\}_{r=0}^R$ be a set of vectors in $\mathbb{R}^d$ and define the update rule of $\{\mathbf{h}_i^{(r)}\}_{i=1}^n$ as follows:

$$\mathbf{h}_i^{(r+1)} = \sum_{j=1}^n W_{ij} \left(\mathbf{h}_j^{(r)} + \nabla f_j(\mathbf{x}_j^{(r)}) \right) - \nabla f_i(\mathbf{x}_i^{(r)}).$$

Then, when $\sum_{i=1}^n \mathbf{h}_i^{(0)} = \mathbf{0}$, it holds that for all r

$$\sum_{i=1}^n \mathbf{h}_i^{(r)} = \mathbf{0}, \quad (\text{D.3})$$

and

$$\mathcal{E}^{(r+1)} \leq 4\rho^2 \mathcal{E}^{(r)} + 4\rho^2 \delta^2 \left\| \bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{x}}^{(r)} \right\|^2 + 32L^2 \Xi_{\mathbf{x}}^{(r+1)}, \quad (\text{D.4})$$

where $\mathcal{E}^{(r)} := \frac{1}{n} \sum_{i=1}^n \|\mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{x}}^{(r)})\|^2$ and $\Xi_{\mathbf{x}}^{(r+1)} := \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)}\|^2$.

Proof. We have

$$\begin{aligned} \sum_{i=1}^n \mathbf{h}_i^{(r+1)} &= \sum_{i=1}^n \left(\sum_{j=1}^n W_{ij} \left(\mathbf{h}_j^{(r)} + \nabla f_j(\mathbf{x}_j^{(r+1)}) \right) - \nabla f_i(\mathbf{x}_i^{(r+1)}) \right) \\ &= \sum_{j=1}^n \left(\sum_{i=1}^n W_{ij} \right) \left(\mathbf{h}_j^{(r)} + \nabla f_j(\mathbf{x}_j^{(r+1)}) \right) - \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(r+1)}) \\ &= \sum_{i=1}^n \mathbf{h}_i^{(r)}. \end{aligned}$$

Since $\sum_{i=1}^n \mathbf{h}_i^{(0)} = \mathbf{0}$, it holds that $\sum_{i=1}^n \mathbf{h}_i^{(r)} = \mathbf{0}$ for any r . For the second statement,

we have

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r+1)} - \nabla h_i(\bar{\mathbf{x}}^{(r+1)}) \right\|^2 \\
& \stackrel{(A.1)}{\leq} \frac{2}{n} \sum_{i=1}^n \left\| \sum_{j=1}^n W_{ij} \left(\mathbf{h}_j^{(r)} + \nabla f_j(\bar{\mathbf{x}}^{(r+1)}) \right) - \nabla f(\bar{\mathbf{x}}^{(r+1)}) \right\|^2 \\
& \quad + \frac{2}{n} \sum_{i=1}^n \left\| \sum_{j=1}^n (W_{ij} - \mathbb{1}_{i=j}) \left(\nabla f_j(\mathbf{x}_j^{(r+1)}) - \nabla f_j(\bar{\mathbf{x}}^{(r+1)}) \right) \right\|^2 \\
& \leq \frac{2}{n} \sum_{i=1}^n \left\| \sum_{j=1}^n W_{ij} \left(\mathbf{h}_j^{(r)} + \nabla f_j(\bar{\mathbf{x}}^{(r+1)}) \right) - \nabla f(\bar{\mathbf{x}}^{(r+1)}) \right\|^2 \\
& \quad + \frac{32}{n} \sum_{i=1}^n \left\| \nabla f_i(\mathbf{x}_i^{(r+1)}) - \nabla f_i(\bar{\mathbf{x}}^{(r+1)}) \right\|^2 \\
& \leq \frac{2}{n} \sum_{i=1}^n \left\| \sum_{j=1}^n W_{ij} \left(\mathbf{h}_j^{(r)} + \nabla f_j(\bar{\mathbf{x}}^{(r+1)}) \right) - \nabla f(\bar{\mathbf{x}}^{(r+1)}) \right\|^2 \\
& \quad + \frac{32L^2}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2,
\end{aligned}$$

where $\mathbb{1}_{i=j}$ is an indicator function. Using Eq. (D.3), we obtain

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r+1)} - \nabla h_i(\bar{\mathbf{x}}^{(r+1)}) \right\|^2 \\
& \stackrel{(5.3)}{\leq} \frac{2\rho^2}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{x}}^{(r+1)}) \right\|^2 + \frac{32L^2}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\
& \stackrel{(A.2)}{\leq} \frac{4\rho^2}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{x}}^{(r)}) \right\|^2 + \frac{4\rho^2}{n} \sum_{i=1}^n \left\| \nabla h_i(\bar{\mathbf{x}}^{(r+1)}) - \nabla h_i(\bar{\mathbf{x}}^{(r)}) \right\|^2 \\
& \quad + \frac{32L^2}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\
& \stackrel{(5.1)}{\leq} \frac{4\rho^2}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{x}}^{(r)}) \right\|^2 + 4\delta^2\rho^2 \left\| \bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\
& \quad + \frac{32L^2}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2.
\end{aligned}$$

This concludes the statement. $\square$

D.3 Proof of Theorem 6

Lemma 16. *Suppose that Assumptions 5 and 11 hold, and $\sum_{i=1}^n \mathbf{h}_i^{(0)} = \mathbf{0}$ and $M = 1$. Then, it holds that*

$$\begin{aligned}
f(\mathbf{x}^\star) + \frac{\lambda}{2n} \sum_{i=1}^n \|\mathbf{x}_i^{(r)} - \mathbf{x}^\star\|^2 + \frac{1}{2\delta} \mathcal{E}^{(r)} \\
\geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+1)} - \mathbf{x}^\star\|^2 \\
+ \left[\frac{1}{\rho^2} \left(\frac{\mu + \lambda}{2} - \delta \right) - \frac{\mu + \lambda}{2} \right] \Xi^{(r+1)} \\
+ \left(\frac{\lambda}{4} - \frac{\delta}{2} \right) \|\bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{x}}^{(r)}\|^2 + \frac{\lambda}{4n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)}\|^2 \\
+ \frac{1}{n} \sum_{i=1}^n \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}^\star - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle,
\end{aligned} \tag{D.5}$$

where $\Xi^{(r)} := \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)}\|^2$ and $\mathcal{E}^{(r)} := \frac{1}{n} \sum_{i=1}^n \|\mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{x}}^{(r)})\|^2$.

Proof. We have

$$\begin{aligned}
f_i(\mathbf{x}^\star) + \frac{\lambda}{2} \|\mathbf{x}_i^{(r)} - \mathbf{x}^\star\|^2 \\
\stackrel{(2.5)}{\geq} f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \left\langle \nabla f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \lambda \left(\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right), \mathbf{x}^\star - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle \\
+ \frac{\lambda}{2} \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)}\|^2 + \frac{\mu + \lambda}{2} \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star\|^2 \\
\stackrel{(2.5)}{\geq} f_i(\bar{\mathbf{x}}^{(r+1)}) + \left\langle \nabla f_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\rangle \\
+ \left\langle \nabla f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \lambda \left(\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right), \mathbf{x}^\star - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle \\
+ \frac{\lambda}{2} \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)}\|^2 + \frac{\mu + \lambda}{2} \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star\|^2,
\end{aligned}$$

where we use the fact that $f_i(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}_i^{(r)}\|^2$ is $(\mu + \lambda)$ -strongly convex in the first inequality. From the definition of $F_{i,r}(\cdot)$, we have

$$\nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) = \nabla f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \mathbf{h}_i^{(r)} + \lambda \left(\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right).$$

Using the above equality, we obtain

$$\begin{aligned}
& f_i(\mathbf{x}^\star) + \frac{\lambda}{2} \left\| \mathbf{x}_i^{(r)} - \mathbf{x}^\star \right\|^2 \\
& \geq f_i(\bar{\mathbf{x}}^{(r+1)}) + \left\langle \nabla f_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\rangle \\
& \quad + \left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star \right\rangle + \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}^\star - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle \\
& \quad + \frac{\lambda}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 + \frac{\mu + \lambda}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star \right\|^2.
\end{aligned}$$

By summing up the above inequality from $i = 1$ to $i = n$, we obtain

$$\begin{aligned}
& f(\mathbf{x}^\star) + \frac{\lambda}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \mathbf{x}^\star \right\|^2 \\
& \geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{1}{n} \sum_{i=1}^n \left[\left\langle \nabla f_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\rangle \right. \\
& \quad + \left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star \right\rangle + \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}^\star - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle \\
& \quad \left. + \frac{\lambda}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 + \frac{\mu + \lambda}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star \right\|^2 \right] \\
& = f(\bar{\mathbf{x}}^{(r+1)}) + \frac{1}{n} \sum_{i=1}^n \left[\left\langle \mathbf{h}_i - \nabla h_i(\bar{\mathbf{x}}^{(r)}), \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\rangle \right. \\
& \quad + \left\langle \nabla h_i(\bar{\mathbf{x}}^{(r)}) - \nabla h_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\rangle \\
& \quad + \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}^\star - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle \\
& \quad \left. + \frac{\lambda}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 + \frac{\mu + \lambda}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star \right\|^2 \right],
\end{aligned}$$

where we use Eq. (D.3) in the last equality. Then, we have

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star \right\|^2 \\
& \stackrel{(D.1)}{=} \left\| \bar{\mathbf{x}}^{(r+1)} - \mathbf{x}^\star \right\|^2 + \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\
& \stackrel{(D.1)}{=} \frac{1}{n} \sum_{i=1}^n \left(\left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 - \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 + \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right) \\
& \stackrel{(5.3)}{\geq} \frac{1}{n} \sum_{i=1}^n \left(\left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 + \left(\frac{1}{\rho^2} - 1 \right) \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right),
\end{aligned}$$

where we use the fact that $\sum_{i=1}^n \mathbf{x}_i^{(r+\frac{1}{2})} = \sum_{i=1}^n \mathbf{x}_i^{(r+1)}$ in the first equality (see Lemma 13). Using the above inequality, we have

$$\begin{aligned}
& f(\mathbf{x}^*) + \frac{\lambda}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \mathbf{x}^* \right\|^2 \\
& \geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{1}{n} \sum_{i=1}^n \left[\left\langle \mathbf{h}_i - \nabla h_i(\bar{\mathbf{x}}^{(r)}), \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\rangle \right. \\
& \quad + \left\langle \nabla h_i(\bar{\mathbf{x}}^{(r)}) - \nabla h_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\rangle + \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}^* - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle \\
& \quad \left. + \frac{\lambda}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 + \frac{\mu + \lambda}{2} \left(\frac{1}{\rho^2} - 1 \right) \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right].
\end{aligned}$$

Then, using Eq. (A.1) with $\alpha = \delta$, we obtain

$$\begin{aligned}
& f(\mathbf{x}^*) + \frac{\lambda}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \mathbf{x}^* \right\|^2 \\
& \stackrel{(5.1,5.3)}{\geq} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\
& \quad + \frac{1}{n} \sum_{i=1}^n \left[-\frac{1}{2\delta} \left\| \mathbf{h}_i - \nabla h_i(\bar{\mathbf{x}}^{(r)}) \right\|^2 - \frac{\delta}{2} \left\| \bar{\mathbf{x}}^{(r)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right. \\
& \quad + \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}^* - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle + \frac{\lambda}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 \\
& \quad \left. + \left[\frac{\mu + \lambda}{2} \left(\frac{1}{\rho^2} - 1 \right) - \frac{\delta}{\rho^2} \right] \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right] \\
& \stackrel{(A.3)}{\geq} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\
& \quad + \frac{1}{n} \sum_{i=1}^n \left[-\frac{1}{2\delta} \left\| \mathbf{h}_i - \nabla h_i(\bar{\mathbf{x}}^{(r)}) \right\|^2 + \left(\frac{\lambda}{4} - \frac{\delta}{2} \right) \left\| \bar{\mathbf{x}}^{(r)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right. \\
& \quad + \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}^* - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle + \frac{\lambda}{4} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 \\
& \quad \left. + \left[\frac{\mu + \lambda}{2} \left(\frac{1}{\rho^2} - 1 \right) - \frac{\delta}{\rho^2} \right] \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right].
\end{aligned}$$

This concludes the statement. $\square$

Lemma 17. Suppose that Assumptions 5, 2 and 11 hold, $\sum_{i=1}^n \mathbf{h}_i^{(0)} = \mathbf{0}$, $M = 1$, and

$$\rho \leq \frac{\delta}{6L}.$$

Then, when $\lambda = 4\delta$, it holds that

$$\begin{aligned} f(\mathbf{x}^*) &+ \frac{2\delta}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \mathbf{x}^* \right\|^2 + \frac{1}{\delta} \mathcal{E}^{(r)} \\ &\geq f(\bar{\mathbf{x}}^{(r+1)}) + \left(1 + \frac{\mu}{4\delta}\right) \frac{2\delta}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{1}{\delta} \left(1 + \frac{\mu}{4\delta}\right) \mathcal{E}^{(r+1)} \\ &\quad + \frac{\delta}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 + \frac{1}{n} \sum_{i=1}^n \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}^* - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle. \end{aligned} \quad (\text{D.6})$$

Proof. Since $\delta \leq L$, $\rho \leq \frac{\delta}{6L}$ implies that $\rho \leq \frac{1}{4}$. Using $\rho \leq \frac{1}{4}$ and Lemma 15, we obtain

$$\mathcal{E}^{(r+1)} \leq \rho \mathcal{E}^{(r)} + \delta^2 \rho \left\| \bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{x}}^{(r)} \right\|^2 + 32L^2 \Xi^{(r+1)}, \quad (\text{D.7})$$

Using $\lambda = 4\delta$ and Eq. (D.5) + $\frac{1}{\delta}(1 + \frac{\mu}{4\delta}) \times$ Eq. (D.7), we obtain

$$\begin{aligned} f(\mathbf{x}^*) &+ \frac{4\delta}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \mathbf{x}^* \right\|^2 + \left(\frac{1}{2\delta} + \frac{\rho}{\delta} \left(1 + \frac{\mu}{4\delta}\right) \right) \mathcal{E}^{(r)} \\ &\geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + 4\delta}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{1}{\delta} \left(1 + \frac{\mu}{4\delta}\right) \mathcal{E}^{(r+1)} \\ &\quad + \left[\frac{1}{\rho^2} \left(\frac{\mu + 4\delta}{2} - \delta \right) - \frac{\mu + 4\delta}{2} - \frac{32L^2}{\delta} \left(1 + \frac{\mu}{4\delta}\right) \right] \Xi^{(r+1)} \\ &\quad + \left(\frac{\delta}{2} - \delta\rho \left(1 + \frac{\mu}{4\delta}\right) \right) \left\| \bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\ &\quad + \frac{\delta}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 + \frac{1}{n} \sum_{i=1}^n \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}^* - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle. \end{aligned}$$

Using the assumption on ρ , we have

$$\begin{aligned} \frac{1}{2\delta} + \frac{\rho}{\delta} \left(1 + \frac{\mu}{4\delta}\right) &\leq \frac{1}{\delta}, \\ \frac{1}{\rho^2} \left(\frac{\mu + 4\delta}{2} - \delta \right) - \frac{\mu + 4\delta}{2} - \frac{32L^2}{\delta} \left(1 + \frac{\mu}{4\delta}\right) &\geq 0, \\ \frac{\delta}{2} - \delta\rho \left(1 + \frac{\mu}{4\delta}\right) &\geq 0, \end{aligned}$$

where we use $\mu \leq L$ and $\delta \leq L$. This concludes the statement. $\square$

Lemma 18. Suppose that Assumptions 5, 2 and 11 hold, $M = 1$, $\mathbf{x}_i^{(0)} = \bar{\mathbf{x}}^{(0)}$, $\mathbf{h}_i^{(0)} := \nabla f(\bar{\mathbf{x}}^{(0)}) - \nabla f_i(\bar{\mathbf{x}}^{(0)})$, and

$$\begin{aligned} \rho &\leq \frac{\delta}{6L}, \\ \sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 &\leq e_{r+1} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2. \end{aligned}$$

Then, when $\lambda = 4\delta$ and $\sum_{r=1}^R e_r \leq \frac{\delta(4\delta+\mu)}{4}$, we obtain

$$\begin{aligned} \frac{1}{\sum_{r=1}^R \frac{1}{q^r}} \sum_{r=1}^R \frac{f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*)}{q^r} &\leq \frac{\mu}{(1 + \frac{\mu}{4\delta})^R - 1} \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2 \\ &\leq \frac{4\delta}{R} \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2. \end{aligned}$$

Proof. From Eq. (D.6), we have

$$\begin{aligned} &f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \\ &\quad + \left(1 + \frac{\mu}{4\delta}\right) \left[\frac{2\delta}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+1)} - \mathbf{x}^*\|^2 + \frac{1}{\delta} \mathcal{E}^{(r+1)} \right] + \frac{\delta}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 \\ &\leq \left[\frac{2\delta}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(r)} - \mathbf{x}^*\|^2 + \frac{1}{\delta} \mathcal{E}^{(r)} \right] + \frac{1}{n} \sum_{i=1}^n \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^* \right\rangle \\ &\leq \left[\frac{2\delta}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(r)} - \mathbf{x}^*\|^2 + \frac{1}{\delta} \mathcal{E}^{(r)} \right] \\ &\quad + \sqrt{\left(\frac{1}{n} \sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \right) \left(\frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^* \right\|^2 \right)}, \end{aligned}$$

where we use Cauchy–Schwarz inequality in the last inequality. Using $\mathcal{E}^{(0)} = 0$ and dividing the above inequality by $\frac{4\delta+\mu}{2}$, we obtain

$$\begin{aligned} &\frac{2}{4\delta + \mu} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{2\delta}{4\delta + \mu} \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2 \\ &\leq \frac{4\delta}{4\delta + \mu} \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \mathbf{x}^* \right\|^2 \\ &\quad + \frac{2}{4\delta + \mu} \sqrt{\left(\frac{1}{n} \sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \right) \left(\frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^* \right\|^2 \right)}, \end{aligned}$$

Using the following notation,

$$\begin{aligned} A_{r+1} &:= \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^* \right\|^2, \\ B_{r+1} &:= \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2, \\ C_{r+1} &:= \frac{1}{n} \sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2, \\ q &:= \frac{4\delta}{4\delta + \mu}, \end{aligned}$$

we can simplify the above inequality as follows:

$$\frac{2}{4\delta + \mu} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) + A_{r+1} + \frac{2\delta}{4\delta + \mu} B_{r+1} \leq qA_r + \frac{2}{4\delta + \mu} \sqrt{C_{r+1}A_{r+1}}.$$

Dividing the above inequality by q^{r+1} and summing up the above inequality from $r = 0$ to $r = R - 1$, we get

$$\frac{2}{4\delta + \mu} \sum_{r=1}^R \frac{f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*)}{q^r} + \frac{A_R}{q^R} + \frac{2\delta}{4\delta + \mu} \sum_{r=1}^R \frac{B_r}{q^r} \leq A_0 + \frac{2}{4\delta + \mu} \sum_{r=1}^R \frac{\sqrt{C_r A_r}}{q^r}.$$

Define Q_R as follows:

$$Q_R := A_0 + \frac{2}{4\delta + \mu} \sum_{r=1}^R \frac{\sqrt{C_r A_r}}{q^r}.$$

The above inequality implies that $A_R \leq q^R Q_R$ for any R . Using this inequality, we get

$$\begin{aligned} Q_{R+1} - Q_R &= \frac{2}{4\delta + \mu} \frac{\sqrt{C_{R+1}A_{R+1}}}{q^{R+1}} \\ &\leq \frac{2}{4\delta + \mu} \sqrt{\frac{C_{R+1}Q_{R+1}}{q^{R+1}}}. \end{aligned}$$

Since we now get the recursive inequality about Q_R , we obtain the following by using Lemma 12 in Jiang et al. (2024a).

$$Q_R \leq 2A_0 + \frac{8}{(4\delta + \mu)^2} \left(\sum_{r=1}^R \sqrt{\frac{C_r}{q^r}} \right)^2.$$

Then, by using $\sum_{i=1}^n \|\nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})})\|^2 \leq e_{r+1} \sum_{i=1}^n \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)}\|^2$, we have

$$\begin{aligned} &\frac{2}{4\delta + \mu} \sum_{r=1}^R \frac{f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*)}{q^r} + \frac{A_R}{q^R} + \frac{2\delta}{4\delta + \mu} \sum_{r=1}^R \frac{B_r}{q^r} \\ &\leq 2A_0 + \frac{8}{(4\delta + \mu)^2} \left(\sum_{r=1}^R \sqrt{\frac{e_r B_r}{q^r}} \right)^2 \\ &\leq 2A_0 + \frac{8}{(4\delta + \mu)^2} \left(\sum_{r=1}^R e_r \right) \left(\sum_{r=1}^R \frac{B_r}{q^r} \right), \end{aligned}$$

where we use Cauchy-Schwarz inequality in the last inequality. Using the assumption that $\sum_{r=1}^R e_r \leq \frac{\delta(4\delta + \mu)}{4}$, we get

$$\frac{1}{4\delta + \mu} \sum_{r=1}^R \frac{f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*)}{q^r} \leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(0)} - \mathbf{x}^*\|^2.$$

Thus, we have

$$\begin{aligned} \frac{1}{\sum_{r=1}^R \frac{1}{q^r}} \sum_{r=1}^R \frac{f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*)}{q^r} &\leq \frac{\mu}{(1 + \frac{\mu}{4\delta})^R - 1} \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(0)} - \mathbf{x}^* \right\|^2 \\ &\leq \frac{4\delta}{R} \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(0)} - \mathbf{x}^* \right\|^2. \end{aligned}$$

This concludes the statement. $\square$

Lemma 19. Suppose that Assumptions 5, 2 and 11 hold, and $\mathbf{x}_i^{(0)} = \bar{\mathbf{x}}^{(0)}$, $\mathbf{h}_i^{(0)} := \nabla f(\bar{\mathbf{x}}^{(0)}) - \nabla f_i(\bar{\mathbf{x}}^{(0)})$, and

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\delta(4\delta + \mu)}{4(r+1)(r+2)} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)} \right\|^2.$$

Then, when $\lambda = 4\delta$ and $M \geq \frac{\log(\frac{6L}{\delta})}{1-\rho}$, we obtain

$$\frac{1}{\sum_{r=1}^R \frac{1}{q^r}} \sum_{r=1}^R \frac{f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*)}{q^r} \leq \frac{\mu}{(1 + \frac{\mu}{4\delta})^R - 1} \left\| \bar{\mathbf{x}}^{(0)} - \mathbf{x}^* \right\|^2 \leq \frac{4\delta}{R} \left\| \bar{\mathbf{x}}^{(0)} - \mathbf{x}^* \right\|^2.$$

Proof. When $M \geq \frac{\log(\frac{6L}{\delta})}{1-\rho}$, it holds that

$$\rho^M \leq \frac{\delta}{6L},$$

where we use $\log(x) \geq 1 - \frac{1}{x}$. Furthermore, it holds that

$$\sum_{r=1}^R \frac{\delta(4\delta + \mu)}{4r(r+1)} = \frac{\delta(4\delta + \mu)}{4} \sum_{r=1}^R \left(\frac{1}{r} - \frac{1}{r+1} \right) \leq \frac{\delta(4\delta + \mu)}{4}.$$

Thus, from Lemma 18, we obtain the desired result. $\square$

D.4 Proof of Theorem 8

Lemma 20. *Suppose that Assumptions 5 and 11 hold and the following inequality is satisfied:*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) \right\|^2 \leq \frac{\lambda^2}{10} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2, \quad (\text{D.8})$$

Then, when $\sum_{i=1}^n \mathbf{h}_i^{(0)} = \mathbf{0}$ and $M = 1$, it holds that

$$\begin{aligned} f(\mathbf{x}^*) &+ \frac{\lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 + \frac{1}{2\delta} \mathcal{E}^{(r)} + 2\delta \Xi^{(r)} \\ &\geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{\lambda + \mu}{2} \left(\frac{1}{\rho^2} - 1 \right) \Xi^{(r+1)} \\ &\quad + \frac{\lambda}{8} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 + \left(\frac{\lambda}{12} - \frac{3\delta}{2} \right) \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2, \end{aligned} \quad (\text{D.9})$$

where $\Xi^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2$ and $\mathcal{E}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{v}}^{(r)}) \right\|^2$.

Proof. We have

$$\begin{aligned} f(\mathbf{x}^*) &+ \frac{\lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left(f_i(\mathbf{x}^*) + \frac{\lambda}{2} \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \right) \\ &\stackrel{(2.5)}{\geq} \frac{1}{n} \sum_{i=1}^n \left(f_i(\mathbf{x}_i^{(r+1)}) + \left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}), \mathbf{x}^* - \mathbf{x}_i^{(r+1)} \right\rangle \right. \\ &\quad \left. + \frac{\mu}{2} \left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{\lambda}{2} \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \right) \\ &\stackrel{(2.5)}{\geq} \frac{1}{n} \sum_{i=1}^n \left(f_i(\bar{\mathbf{x}}^{(r+1)}) + \left\langle \nabla f_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\rangle + \left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}), \mathbf{x}^* - \mathbf{x}_i^{(r+1)} \right\rangle \right. \\ &\quad \left. + \frac{\mu}{2} \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 + \frac{\mu}{2} \left\| \mathbf{x}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{\lambda}{2} \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \right). \end{aligned}$$

Using the following equation,

$$\sum_{i=1}^n \left\langle \nabla f_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\rangle = \sum_{i=1}^n \left\langle -\nabla h_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\rangle,$$

we get

$$\begin{aligned}
& f(\mathbf{x}^\star) + \frac{\lambda}{2n} \sum_{i=1}^n \|\mathbf{v}_i^{(r)} - \mathbf{x}^\star\|^2 \\
& \geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{1}{n} \sum_{i=1}^n \left[\left\langle -\nabla h_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\rangle \right. \\
& \quad \left. + \underbrace{\left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}), \mathbf{x}^\star - \mathbf{x}_i^{(r+1)} \right\rangle + \frac{\mu}{2} \|\mathbf{x}_i^{(r+1)} - \mathbf{x}^\star\|^2 + \frac{\lambda}{2} \|\mathbf{v}_i^{(r)} - \mathbf{x}^\star\|^2}_{\mathcal{T}_i^\star} \right. \\
& \quad \left. + \frac{\mu}{2} \|\mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)}\|^2 \right].
\end{aligned}$$

Note that $h_i(\mathbf{x}) := f(\mathbf{x}) - f_i(\mathbf{x})$. Using Eq. (D.3), we obtain

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \mathcal{T}_i^\star & \stackrel{(D.3)}{=} \frac{1}{n} \sum_{i=1}^n \left[\left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}) + \mathbf{h}_i^{(r)}, \mathbf{x}^\star \right\rangle + \left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}), -\mathbf{x}_i^{(r+1)} \right\rangle \right. \\
& \quad \left. + \frac{\mu}{2} \|\mathbf{x}_i^{(r+1)} - \mathbf{x}^\star\|^2 + \frac{\lambda}{2} \|\mathbf{v}_i^{(r)} - \mathbf{x}^\star\|^2 \right].
\end{aligned}$$

We define $G_{i,r}(\mathbf{v}) := \left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}) + \mathbf{h}_i^{(r)}, \mathbf{v} \right\rangle + \frac{\mu}{2} \|\mathbf{v} - \mathbf{x}_i^{(r+1)}\|^2 + \frac{\lambda}{2} \|\mathbf{v} - \mathbf{v}_i^{(r)}\|^2$. By using that fact that $G_{i,r}$ is $(\mu + \lambda)$ -strongly convex, we get

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \mathcal{T}_i^\star \\
& \stackrel{(2.5)}{\geq} \frac{1}{n} \sum_{i=1}^n \left[\left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}) + \mathbf{h}_i^{(r)}, \mathbf{v}_i^{(r+\frac{1}{2})} \right\rangle + \left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}), -\mathbf{x}_i^{(r+1)} \right\rangle \right. \\
& \quad \left. + \frac{\mu}{2} \|\mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r+\frac{1}{2})}\|^2 + \frac{\lambda}{2} \|\mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})}\|^2 + \frac{\lambda + \mu}{2} \|\mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star\|^2 \right] \\
& \stackrel{(D.1)}{=} \frac{1}{n} \sum_{i=1}^n \left[\left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}) + \mathbf{h}_i^{(r)}, \mathbf{v}_i^{(r+\frac{1}{2})} \right\rangle + \left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}), -\mathbf{x}_i^{(r+1)} \right\rangle \right. \\
& \quad \left. + \frac{\mu}{2} \|\mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r+\frac{1}{2})}\|^2 \right. \\
& \quad \left. + \frac{\lambda}{2} \|\mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})}\|^2 + \frac{\lambda + \mu}{2} \|\mathbf{v}_i^{(r+\frac{1}{2})} - \bar{\mathbf{v}}^{(r+1)}\|^2 + \frac{\lambda + \mu}{2} \|\bar{\mathbf{v}}^{(r+1)} - \mathbf{x}^\star\|^2 \right] \\
& \stackrel{(D.1)}{=} \frac{1}{n} \sum_{i=1}^n \left[\left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}) + \mathbf{h}_i^{(r)}, \mathbf{v}_i^{(r+\frac{1}{2})} \right\rangle + \left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}), -\mathbf{x}_i^{(r+1)} \right\rangle \right. \\
& \quad \left. + \frac{\mu}{2} \|\mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r+\frac{1}{2})}\|^2 + \frac{\lambda}{2} \|\mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})}\|^2 + \frac{\lambda + \mu}{2} \|\mathbf{v}_i^{(r+\frac{1}{2})} - \bar{\mathbf{v}}^{(r+1)}\|^2 \right. \\
& \quad \left. + \frac{\lambda + \mu}{2} \|\mathbf{v}_i^{(r+1)} - \mathbf{x}^\star\|^2 - \frac{\lambda + \mu}{2} \|\mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)}\|^2 \right].
\end{aligned}$$

Using the above inequality, we obtain

$$\begin{aligned}
& f(\mathbf{x}^*) + \frac{\lambda}{2n} \sum_{i=1}^n \|\mathbf{v}_i^{(r)} - \mathbf{x}^*\|^2 \\
& \geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \|\mathbf{v}_i^{(r+1)} - \mathbf{x}^*\|^2 \\
& \quad + \frac{1}{n} \sum_{i=1}^n \left[\left\langle -\nabla h_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\rangle + \frac{\mu}{2} \left\| \mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 \right. \\
& \quad + \frac{\lambda}{2} \left\| \mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 + \frac{\lambda + \mu}{2} \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 - \frac{\lambda + \mu}{2} \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 \\
& \quad \left. + \underbrace{\left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}) + \mathbf{h}_i^{(r)}, \mathbf{v}_i^{(r+\frac{1}{2})} \right\rangle + \left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}), -\mathbf{x}_i^{(r+1)} \right\rangle}_{\mathcal{T}_i^{**}} \right].
\end{aligned}$$

From the definition of $F_{i,r}$, we have

$$\nabla F_{i,r}(\mathbf{x}) = \nabla f_i(\mathbf{x}) + \mathbf{h}_i^{(r)} + \lambda(\mathbf{x} - \mathbf{v}_i^{(r)}).$$

By using the above equality, $\mathcal{T}_i^{**}$ can be rewritten as follows:

$$\begin{aligned}
\mathcal{T}_i^{**} &= \left\langle \nabla f_i(\mathbf{x}_i^{(r+1)}), \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r+1)} \right\rangle + \left\langle \mathbf{h}_i^{(r)}, \mathbf{v}_i^{(r+\frac{1}{2})} \right\rangle \\
&= \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) + \lambda(\mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)}), \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r+1)} \right\rangle + \left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+1)} \right\rangle \\
&= \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}), \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\rangle + \lambda \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2 + \left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+1)} \right\rangle \\
& \quad + \underbrace{\left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) + \lambda(\mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)}), \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\rangle}_{\mathcal{T}_i^{***}}.
\end{aligned}$$

Then, $\mathcal{T}_i^{***}$ can be bounded as follows:

$$\begin{aligned}
\mathcal{T}_i^{***} &\stackrel{(A.1)}{\geq} -\frac{2}{3\lambda} \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) + \lambda(\mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)}) \right\|^2 - \frac{3\lambda}{8} \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2 \\
&= -\frac{2}{3\lambda} \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) \right\|^2 - \frac{4}{3} \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}), \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\rangle \\
& \quad - \frac{2\lambda}{3} \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2 - \frac{3\lambda}{8} \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2.
\end{aligned}$$

Using the above inequality, we get

$$\begin{aligned}
\mathcal{T}_i^{\star\star} &\geq \frac{\lambda}{3} \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2 + \left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+1)} \right\rangle - \frac{2}{3\lambda} \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) \right\|^2 \\
&\quad - \frac{1}{3} \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}), \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\rangle - \frac{3\lambda}{8} \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2 \\
&\stackrel{(A.1)}{\geq} \frac{\lambda}{6} \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2 + \left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+1)} \right\rangle - \frac{5}{6\lambda} \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) \right\|^2 \\
&\quad - \frac{3\lambda}{8} \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2.
\end{aligned}$$

Using the assumption that $\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) \right\|^2 \leq \frac{\lambda^2}{10} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2$, we get

$$\sum_{i=1}^n \mathcal{T}_i^{\star\star} \geq \sum_{i=1}^n \left[\frac{\lambda}{12} \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2 + \left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+1)} \right\rangle - \frac{3\lambda}{8} \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2 \right].$$

Using the above inequalities, we get

$$\begin{aligned}
&f(\mathbf{x}^\star) + \frac{\lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^\star \right\|^2 \\
&\geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 \\
&\quad + \frac{1}{n} \sum_{i=1}^n \left[\frac{\mu}{2} \left\| \mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 + \frac{\lambda}{8} \left\| \mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 \right. \\
&\quad + \frac{\lambda + \mu}{2} \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 - \frac{\lambda + \mu}{2} \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 \\
&\quad \left. + \frac{\lambda}{12} \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2 + \underbrace{\left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+1)} \right\rangle + \left\langle -\nabla h_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\rangle}_{\mathcal{T}_i^{\star\star\star}} \right]
\end{aligned}$$

$\mathcal{T}_i^{****}$ can be bounded by blow as follows:

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \mathcal{T}_i^{****} \\
& \stackrel{(D.3)}{=} \frac{1}{n} \sum_{i=1}^n \left\langle \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+1)} - \bar{\mathbf{v}}^{(r)} \right\rangle \\
& = \frac{1}{n} \sum_{i=1}^n \left[\left\langle \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{v}}^{(r)}), \mathbf{x}_i^{(r+1)} - \bar{\mathbf{v}}^{(r)} \right\rangle \right. \\
& \quad \left. + \left\langle \nabla h_i(\bar{\mathbf{v}}^{(r)}) - \nabla h_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+1)} - \bar{\mathbf{v}}^{(r)} \right\rangle \right] \\
& \stackrel{(A.1)}{\geq} \frac{1}{n} \sum_{i=1}^n \left[-\frac{1}{2\delta} \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{v}}^{(r)}) \right\|^2 - \delta \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{v}}^{(r)} \right\|^2 \right. \\
& \quad \left. - \frac{1}{2\delta} \left\| \nabla h_i(\bar{\mathbf{v}}^{(r)}) - \nabla h_i(\bar{\mathbf{x}}^{(r+1)}) \right\|^2 \right] \\
& \stackrel{(5.1)}{\geq} \frac{1}{n} \sum_{i=1}^n \left[-\frac{1}{2\delta} \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{v}}^{(r)}) \right\|^2 - \delta \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{v}}^{(r)} \right\|^2 - \frac{\delta}{2} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right] \\
& \stackrel{(A.2)}{\geq} \frac{1}{n} \sum_{i=1}^n \left[-\frac{1}{2\delta} \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{v}}^{(r)}) \right\|^2 - 2\delta \left\| \mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r)} \right\|^2 - 2\delta \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2 \right. \\
& \quad \left. - \frac{\delta}{2} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right].
\end{aligned}$$

Using the above inequality, we obtain

$$\begin{aligned}
& f(\mathbf{x}^*) + \frac{\lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 + \frac{1}{2\delta n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{v}}^{(r)}) \right\|^2 + \frac{2\delta}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2 \\
& \geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\
& \quad + \frac{1}{n} \sum_{i=1}^n \left[\frac{\mu}{2} \left\| \mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 + \frac{\lambda}{8} \left\| \mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 \right. \\
& \quad \left. + \frac{\lambda + \mu}{2} \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 - \frac{\lambda + \mu}{2} \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 \right. \\
& \quad \left. + \left(\frac{\lambda}{12} - 2\delta \right) \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2 - \frac{\delta}{2} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right] \\
& \stackrel{(5.3)}{\geq} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{\lambda + \mu}{2n} \left(\frac{1}{\rho^2} - 1 \right) \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 \\
& \quad + \frac{1}{n} \sum_{i=1}^n \left[\frac{\mu}{2} \left\| \mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 + \frac{\lambda}{8} \left\| \mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 \right. \\
& \quad \left. + \left(\frac{\lambda}{12} - 2\delta \right) \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2 - \frac{\delta}{2} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right].
\end{aligned}$$

Using Eq. (D.3), we get the desired result. $\square$

Lemma 21. *Suppose that Assumptions 5 and 11 hold and the following inequality is satisfied:*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) \right\|^2 \leq \frac{\lambda^2}{10} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2. \quad (\text{D.10})$$

Then, when $\lambda = 20\delta$, $\sum_{i=1}^n \mathbf{h}_i^{(0)} = \mathbf{0}$, $M = 1$, and $\rho \leq \frac{\delta}{5L}$, it holds that

$$\begin{aligned} & f(\mathbf{x}^\star) + \frac{10\delta}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^\star \right\|^2 + \frac{1}{\delta} \mathcal{E}^{(r)} + 2\delta \Xi^{(r)} \\ & \geq f(\bar{\mathbf{x}}^{(r+1)}) + \left(1 + \frac{\mu}{20\delta}\right) \left[\frac{10\delta}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 + \frac{1}{\delta} \mathcal{E}^{(r+1)} + 2\delta \Xi^{(r+1)} \right], \end{aligned} \quad (\text{D.11})$$

where $\Xi^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2$ and $\mathcal{E}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{v}}^{(r)}) \right\|^2$.

Proof. Since $\delta \leq L$, $\rho \leq \frac{\delta}{5L}$ implies $\rho \leq \frac{1}{4}$. Thus, using Lemma 15, we obtain

$$\rho \mathcal{E}^{(r)} \geq \mathcal{E}^{(r+1)} - \rho \delta^2 \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{v}}^{(r)} \right\|^2 - 32L^2 \Xi^{(r+1)}. \quad (\text{D.12})$$

By calculating Eq. (D.9) + Eq. (D.12) $\times \frac{1}{\delta} (1 + \frac{\mu}{\lambda})$, we obtain

$$\begin{aligned} & f(\mathbf{x}^\star) + \frac{\lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^\star \right\|^2 + \frac{1}{\delta} \left(\frac{1}{2} + \rho \left(1 + \frac{\mu}{\lambda}\right) \right) \mathcal{E}^{(r)} + 2\delta \Xi^{(r)} \\ & \geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{\mu + \lambda}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 + \frac{1}{\delta} \left(1 + \frac{\mu}{\lambda}\right) \mathcal{E}^{(r+1)} \\ & \quad + \left(1 + \frac{\mu}{\lambda}\right) \left(\frac{\lambda}{2} \left(\frac{1}{\rho^2} - 1 \right) - \frac{32L^2}{\delta} \right) \Xi^{(r+1)} \\ & \quad + \left(\frac{\lambda}{8} - \left(1 + \frac{\mu}{\lambda}\right) \rho \delta \right) \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 + \left(\frac{\lambda}{12} - \frac{3\delta}{2} \right) \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2, \end{aligned}$$

By inserting $\lambda = 20\delta$, we obtain

$$\begin{aligned} & f(\mathbf{x}^\star) + \frac{10\delta}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^\star \right\|^2 + \frac{1}{\delta} \left(\frac{1}{2} + \rho \left(1 + \frac{\mu}{20\delta}\right) \right) \mathcal{E}^{(r)} + 2\delta \Xi^{(r)} \\ & \geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{10\delta}{n} \left(1 + \frac{\mu}{20\delta}\right) \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 + \frac{1}{\delta} \left(1 + \frac{\mu}{20\delta}\right) \mathcal{E}^{(r+1)} \\ & \quad + \left(1 + \frac{\mu}{20\delta}\right) \left(10\delta \left(\frac{1}{\rho^2} - 1 \right) - \frac{32L^2}{\delta} \right) \Xi^{(r+1)} \\ & \quad + \left(\frac{5\delta}{2} - \left(1 + \frac{\mu}{10\delta}\right) \rho \delta \right) \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2. \end{aligned}$$

$\rho \leq \frac{\delta}{5L}$ implies that

$$\begin{aligned} \frac{1}{2} + \rho \left(1 + \frac{\mu}{20\delta}\right) &\leq 1, \\ 10\delta \left(\frac{1}{\rho^2} - 1\right) - \frac{32L^2}{\delta} &\geq 2\delta, \\ \frac{5\delta}{2} - \left(1 + \frac{\mu}{20\delta}\right) \rho\delta &\geq 0, \end{aligned}$$

where we use $\mu \leq L$ and $\mu \leq L$. Using the above inequalities, we get

$$\begin{aligned} f(\mathbf{x}^\star) + \frac{10\delta}{n} \sum_{i=1}^n \|\mathbf{v}_i^{(r)} - \mathbf{x}^\star\|^2 + \frac{1}{\delta} \mathcal{E}^{(r)} + 2\delta \Xi^{(r)} \\ \geq f(\bar{\mathbf{x}}^{(r+1)}) + \frac{10\delta}{n} \left(1 + \frac{\mu}{20\delta}\right) \sum_{i=1}^n \|\mathbf{v}_i^{(r+1)} - \mathbf{x}^\star\|^2 \\ + \frac{1}{\delta} \left(1 + \frac{\mu}{20\delta}\right) \mathcal{E}^{(r+1)} + 2\delta \left(1 + \frac{\mu}{20\delta}\right) \Xi^{(r+1)}. \end{aligned}$$

Thus, we can get the desired result. $\square$

Lemma 22. *Suppose that Assumptions 5 and 11 hold and the following inequality is satisfied:*

$$\sum_{i=1}^n \|\nabla F_{i,r}(\mathbf{x}_i^{(r+1)})\|^2 \leq \frac{\lambda^2}{10} \sum_{i=1}^n \|\mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)}\|^2. \quad (\text{D.13})$$

Then, when $\lambda = 20\delta$, $M = 1$, $\mathbf{v}_i^{(0)} = \bar{\mathbf{v}}^{(0)}$, $\mathbf{h}_i^{(0)} = \nabla f(\bar{\mathbf{v}}^{(0)}) - \nabla f_i(\bar{\mathbf{v}}^{(0)})$, and $\rho \leq \frac{\delta}{5L}$, we obtain

$$\begin{aligned} \frac{1}{W_R} \sum_{r=0}^{R-1} w^{(r)} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^\star) \right) &\leq \frac{\mu}{2((1 + \frac{\mu}{20\delta})^R - 1)} \|\bar{\mathbf{v}}^{(0)} - \mathbf{x}^\star\|^2 \\ &\leq \frac{10\delta}{R} \|\bar{\mathbf{v}}^{(0)} - \mathbf{x}^\star\|^2, \end{aligned}$$

where $w^{(r)} := (1 + \frac{\mu}{20\delta})^r$ and $W_R := \sum_{r=0}^{R-1} w^{(r)}$.

Proof. From Eq. (D.11), we have

$$\begin{aligned} f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^\star) \\ \leq 10\delta \left[\frac{1}{n} \sum_{i=1}^n \|\mathbf{v}_i^{(r)} - \mathbf{x}^\star\|^2 + \frac{1}{10\delta^2} \mathcal{E}^{(r)} + \frac{1}{5} \Xi^{(r)} \right] \\ - 10\delta \left(1 + \frac{\mu}{20\delta}\right) \left[\frac{1}{n} \sum_{i=1}^n \|\mathbf{v}_i^{(r+1)} - \mathbf{x}^\star\|^2 + \frac{1}{10\delta^2} \mathcal{E}^{(r+1)} + \frac{1}{5} \Xi^{(r+1)} \right]. \end{aligned}$$

We define $w^{(r)} := (1 + \frac{\mu}{20\delta})^r$. We obtain

$$\begin{aligned} & \frac{1}{W_R} \sum_{r=0}^{R-1} w^{(r)} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) \\ & \leq \frac{10\delta}{W_R} \left[\frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(0)} - \mathbf{x}^* \right\|^2 + \frac{1}{10\delta^2} \mathcal{E}^{(0)} + \frac{1}{5} \Xi^{(0)} \right], \end{aligned}$$

where $W_R := \sum_{r=0}^{R-1} w^{(r)}$. Using $\mathcal{E}^{(0)} = 0$ and $\Xi^{(0)} = 0$, we obtain

$$\frac{1}{W_R} \sum_{r=0}^{R-1} w^{(r)} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) \leq \frac{10\delta}{W_R} \left\| \bar{\mathbf{v}}^{(0)} - \mathbf{x}^* \right\|^2.$$

Using $W_R = \frac{20\delta}{\mu} ((1 + \frac{\mu}{20\delta})^R - 1)$, we obtain

$$\begin{aligned} \frac{1}{W_R} \sum_{r=0}^{R-1} w^{(r)} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) & \leq \frac{\mu}{2((1 + \frac{\mu}{20\delta})^R - 1)} \left\| \bar{\mathbf{v}}^{(0)} - \mathbf{x}^* \right\|^2 \\ & \leq \frac{10\delta}{R} \left\| \bar{\mathbf{v}}^{(0)} - \mathbf{x}^* \right\|^2. \end{aligned}$$

□

Lemma 23. Suppose that Assumptions 5, 2, and 11 hold and the following inequality is satisfied:

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+1)}) \right\|^2 \leq \frac{\lambda^2}{10} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}_i^{(r+1)} \right\|^2. \quad (\text{D.14})$$

When $\mathbf{v}_i^{(0)} = \bar{\mathbf{v}}^{(0)}$, $\mathbf{h}_i^{(0)} = \nabla f(\bar{\mathbf{v}}^{(0)}) - \nabla f_i(\bar{\mathbf{v}}^{(0)})$, $\lambda = 20\delta$, and $M \geq \frac{\log(\frac{5L}{\delta})}{1-\rho}$ it holds that

$$\begin{aligned} \frac{1}{W_R} \sum_{r=0}^{R-1} w^{(r)} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) & \leq \frac{\mu}{2((1 + \frac{\mu}{20\delta})^R - 1)} \left\| \bar{\mathbf{v}}^{(0)} - \mathbf{x}^* \right\|^2 \\ & \leq \frac{10\delta}{R} \left\| \bar{\mathbf{v}}^{(0)} - \mathbf{x}^* \right\|^2, \end{aligned}$$

where $w^{(r)} := (1 + \frac{\mu}{20\delta})^r$ and $W_R := \sum_{r=0}^{R-1} w^{(r)}$.

Proof. When $M \geq \frac{1}{1-\rho} \log(\frac{5L}{\delta})$, it holds that

$$\rho^M \leq \frac{\delta}{5L},$$

where we use $\log(x) \geq 1 - \frac{1}{x}$. From Lemma 22, we obtain the desired result. □

D.5 Proof of Theorem 10

D.5.1 Useful Lemmas

Lemma 24. *When $\mu \leq 4\lambda$, it holds that for any $r \geq 0$*

$$A_r \geq \frac{1}{4\mu} \left[\left(1 + \sqrt{\frac{\mu}{4\lambda}}\right)^r - \left(1 - \sqrt{\frac{\mu}{4\lambda}}\right)^r \right]^2 \geq \frac{r^2}{4\lambda}.$$

Otherwise, for any $r \geq 1$, it holds that

$$A_r \geq \frac{1}{4\lambda} \left(1 + \sqrt{\frac{\mu}{4\lambda}}\right)^{2(r-1)}.$$

Proof. See Lemma 12 in Jiang et al. (2024b). □

Lemma 25. *When $\mu = 0$, it holds that for any $0 \leq r \leq R$,*

$$A_r \leq \frac{r^2}{\lambda}. \quad (\text{D.15})$$

Proof. Denote $C_r := \sqrt{A_r}$. From the definition of A_r and a_r , it holds that

$$C_{r+1}^2 = \lambda(C_{r+1}^2 - C_r^2)^2 = \lambda(C_{r+1} - C_r)^2(C_{r+1} + C_r)^2 \geq \lambda(C_{r+1} - C_r)^2 C_{r+1}^2.$$

Thus, we have

$$C_{r+1} - C_r \geq \sqrt{\frac{1}{\lambda}}.$$

This concludes the statement. □

Lemma 26. *It holds that for any $r \geq 1$,*

$$2 \left(1 + \frac{\mu}{\lambda}\right)^{-1} \left(1 + \sqrt{1 + \frac{4\lambda}{\mu}}\right)^{-1} \leq \frac{A_r}{a_{r+1}} \leq \frac{2\lambda}{\mu}. \quad (\text{D.16})$$

Furthermore, when $\mu = 0$, it holds that for any $r \geq 1$,

$$\frac{r}{5} \leq \frac{A_r}{a_{r+1}} \leq r. \quad (\text{D.17})$$

Note that the inequalities on the right-hand side also hold when $r = 0$ since $\frac{A_0}{a_1} = 0$.

Proof. From the definition of a_{r+1} , we have

$$\begin{aligned} a_{r+1} &= \frac{1 + \mu A_r + \sqrt{(1 + \mu A_r)^2 + 4\lambda A_r(1 + \mu A_r)}}{2\lambda} \geq \frac{\mu A_r}{2\lambda}, \\ a_{r+1} &= \frac{1 + \mu A_r + \sqrt{(1 + \mu A_r)^2 + 4\lambda A_r(1 + \mu A_r)}}{2\lambda} \leq \frac{1 + \mu A_r}{2\lambda} \left(1 + \sqrt{1 + \frac{4\lambda}{\mu}}\right). \end{aligned}$$

When $\mu = 0$, we have

$$\begin{aligned} a_{r+1} &= \frac{1 + \sqrt{1 + 4\lambda A_r}}{2\lambda} \geq \frac{A_r}{\sqrt{\lambda A_r}} \geq \frac{A_r}{r}, \\ a_{r+1} &= \frac{1 + \sqrt{1 + 4\lambda A_r}}{2\lambda} \leq A_r \sqrt{\frac{5}{\lambda A_r}} \leq \frac{5A_r}{r}, \end{aligned}$$

where we use Lemma 24 and 25. This concludes the statement. $\square$

Lemma 27. *For all $r \geq 0$, it holds that*

$$\frac{a_{r+2}}{a_{r+1}} \leq \frac{1}{2} \left(1 + \frac{\mu}{\lambda}\right) \left(1 + \frac{2\lambda}{\mu}\right) \left(1 + \sqrt{1 + \frac{4\lambda}{\mu}}\right). \quad (\text{D.18})$$

Furthermore, when $\mu = 0$, it holds that for all $r \geq 0$,

$$1 \leq \frac{a_{r+2}}{a_{r+1}} \leq 4. \quad (\text{D.19})$$

Proof. From Lemma 26, we get

$$\begin{aligned} a_{r+2} &\stackrel{(\text{D.16})}{\leq} \frac{A_{r+1}}{2} \left(1 + \frac{\mu}{\lambda}\right) \left(1 + \sqrt{1 + \frac{4\lambda}{\mu}}\right) \\ &= \frac{A_r + a_{r+1}}{2} \left(1 + \frac{\mu}{\lambda}\right) \left(1 + \sqrt{1 + \frac{4\lambda}{\mu}}\right) \\ &\stackrel{(\text{D.16})}{\leq} \frac{a_{r+1}}{2} \left(1 + \frac{2\lambda}{\mu}\right) \left(1 + \frac{\mu}{\lambda}\right) \left(1 + \sqrt{1 + \frac{4\lambda}{\mu}}\right). \end{aligned}$$

When $\mu = 0$, we get

$$\begin{aligned} \frac{a_{r+1}}{a_r} &= \sqrt{\frac{A_{r+1}}{A_r}} \leq \sqrt{\frac{(r+1)^2}{\frac{r^2}{4}}} = \frac{2(r+1)}{r} \leq 4, \\ \frac{a_{r+1}}{a_r} &= \sqrt{\frac{A_{r+1}}{A_r}} \geq 1, \end{aligned}$$

where we use $A_{r+1} = A_r + a_{r+1} \geq A_r$. This concludes the statement. $\square$

D.5.2 Main Proof

Lemma 28. *Suppose that Assumptions 5 and 11 hold and the following inequality is satisfied:*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\lambda^2}{352} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2.$$

Then, when $M = 1$, it holds that

$$\begin{aligned}
& A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \\
& + \frac{A_{r+1}}{\delta} \mathcal{E}^{(r)} + A_r \left(\frac{33\lambda}{16} + \delta \right) \Xi_{\mathbf{x}}^{(r)} + \delta a_{r+1} \Xi_{\mathbf{v}}^{(r)} \\
& \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{\mu A_{r+1}}{2\rho^2} \Xi_{\mathbf{x}}^{(r+1)} \\
& + \frac{B_{r+1}}{2} \left(\frac{1}{\rho^2} - 1 \right) \Xi_{\mathbf{v}}^{(r+1)} + \frac{A_{r+1}}{n} \left(\frac{\lambda}{32} - 2\delta \right) \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 \\
& + \frac{B_r}{12n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2 + \frac{\mu a_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2,
\end{aligned} \tag{D.20}$$

where $\Xi_{\mathbf{x}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2$, $\Xi_{\mathbf{v}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2$, and $\mathcal{E}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{y}}^{(r)}) \right\|^2$.

Proof. We have

$$\begin{aligned}
& A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \\
& \stackrel{(2.5)}{\geq} \frac{A_{r+1}}{n} \sum_{i=1}^n f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \frac{1}{n} \sum_{i=1}^n \left\langle \nabla f_i(\mathbf{x}_i^{(r)}), A_r \bar{\mathbf{x}}^{(r)} - A_{r+1} \mathbf{x}_i^{(r+\frac{1}{2})} + a_{r+1} \mathbf{x}^* \right\rangle \\
& + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 + \frac{\mu A_r}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r)} \right\|^2 + \frac{\mu a_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^* \right\|^2 \\
& = \frac{A_{r+1}}{n} \sum_{i=1}^n f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \frac{1}{n} \sum_{i=1}^n \left\langle \nabla f_i(\mathbf{x}_i^{(r)}), A_r \bar{\mathbf{x}}^{(r)} - A_{r+1} \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle \\
& + \frac{a_{r+1}}{n} \sum_{i=1}^n \left(\left\langle \nabla f_i(\mathbf{x}_i^{(r)}) + \mathbf{h}_i^{(r)}, \mathbf{x}^* \right\rangle + \frac{\mu}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}^* \right\|^2 \right) \\
& + \frac{\mu A_r}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r)} \right\|^2 + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \\
& \stackrel{(a)}{\geq} \frac{A_{r+1}}{n} \sum_{i=1}^n f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \frac{1}{n} \sum_{i=1}^n \left\langle \nabla f_i(\mathbf{x}_i^{(r)}), A_r \bar{\mathbf{x}}^{(r)} - A_{r+1} \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle \\
& + \frac{a_{r+1}}{n} \sum_{i=1}^n \left(\left\langle \nabla f_i(\mathbf{x}_i^{(r)}) + \mathbf{h}_i^{(r)}, \mathbf{v}_i^{(r+\frac{1}{2})} \right\rangle + \frac{\mu}{2} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 \right) \\
& + \frac{\mu A_r}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r)} \right\|^2 + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 \\
& + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{x}^* \right\|^2,
\end{aligned}$$

where we use the fact that $a_{r+1}(\langle \nabla f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \mathbf{h}_i^{(r)}, \mathbf{v} \rangle + \frac{\mu}{2} \|\mathbf{v} - \mathbf{x}_i^{(r+\frac{1}{2})}\|^2) + \frac{B_r}{2} \|\mathbf{v} - \mathbf{v}_i^{(r)}\|^2$ is B_{r+1} -strongly convex in (a). By using the above inequalities,

$$\begin{aligned}
& \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{x}^\star \right\|^2 \\
& \stackrel{(D.1)}{=} \frac{B_{r+1}}{2} \left\| \bar{\mathbf{v}}^{(r+1)} - \mathbf{x}^\star \right\|^2 + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 \\
& \stackrel{(D.1)}{=} \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 - \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 \\
& \quad + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 \\
& \stackrel{(5.3)}{\geq} \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 + \frac{B_{r+1}}{2n} \left(\frac{1}{\rho^2} - 1 \right) \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2,
\end{aligned}$$

we get

$$\begin{aligned}
& A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^\star) + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^\star \right\|^2 \\
& \geq \frac{A_{r+1}}{n} \sum_{i=1}^n f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \frac{A_{r+1}}{n} \sum_{i=1}^n \left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 \\
& \quad + \underbrace{\frac{1}{n} \sum_{i=1}^n \left\langle \nabla f_i(\mathbf{x}_i^{(r)}) + \mathbf{h}_i^{(r)}, A_r \bar{\mathbf{x}}^{(r)} - A_{r+1} \mathbf{x}_i^{(r+\frac{1}{2})} + a_{r+1} \mathbf{v}_i^{(r+\frac{1}{2})} \right\rangle}_{\mathcal{T}^\star} \\
& \quad + \frac{\mu A_r}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\
& \quad + \frac{\mu a_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 \\
& \quad + \frac{B_{r+1}}{2n} \left(\frac{1}{\rho^2} - 1 \right) \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2.
\end{aligned}$$

From the definition of $F_{i,r}$, we have

$$\nabla F_{i,r}(\mathbf{x}) = \nabla f_i(\mathbf{x}) + \mathbf{h}_i^{(r)} + \lambda(\mathbf{x} - \mathbf{y}_i^{(r)}).$$

Using the above equality, we get

$$\begin{aligned}
\mathcal{T}^\star &= \frac{1}{n} \sum_{i=1}^n \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) + \lambda \left(\mathbf{y}_i^{(r)} - \mathbf{x}_i^{(r+\frac{1}{2})} \right), \right. \\
&\quad \left. A_r \bar{\mathbf{x}}^{(r)} - A_{r+1} \mathbf{x}_i^{(r+\frac{1}{2})} + a_{r+1} \mathbf{v}_i^{(r+\frac{1}{2})} \right\rangle \\
&= \frac{1}{n} \sum_{i=1}^n \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) + \lambda \left(\mathbf{y}_i^{(r)} - \mathbf{x}_i^{(r+\frac{1}{2})} \right), \right. \\
&\quad \left. A_r \left(\bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right) + A_{r+1} \left(\mathbf{y}_i^{(r)} - \mathbf{x}_i^{(r+\frac{1}{2})} \right) + a_{r+1} \left(\mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right) \right\rangle \\
&= \frac{1}{n} \sum_{i=1}^n \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \right. \\
&\quad \left. A_r \left(\bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right) + A_{r+1} \left(\mathbf{y}_i^{(r)} - \mathbf{x}_i^{(r+\frac{1}{2})} \right) + a_{r+1} \left(\mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right) \right\rangle \\
&\quad + \frac{\lambda}{n} \sum_{i=1}^n \left\langle \mathbf{y}_i^{(r)} - \mathbf{x}_i^{(r+\frac{1}{2})}, A_r \left(\bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right) + a_{r+1} \left(\mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right) \right\rangle \\
&\quad + \frac{\lambda A_{r+1}}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 \\
&\stackrel{(A.1)}{\geq} \frac{1}{n} \sum_{i=1}^n \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \right. \\
&\quad \left. A_r \left(\bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right) + A_{r+1} \left(\mathbf{y}_i^{(r)} - \mathbf{x}_i^{(r+\frac{1}{2})} \right) + a_{r+1} \left(\mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right) \right\rangle \\
&\quad - \frac{2\lambda A_r}{n} \sum_{i=1}^n \left\| \bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right\|^2 - \frac{B_r}{3n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2 \\
&\quad + \frac{\lambda A_{r+1}}{8n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2.
\end{aligned}$$

We have

$$\begin{aligned}
A_r \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right\rangle &\stackrel{(A.1)}{\geq} -\frac{4A_r}{\lambda} \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \\
&\quad - \frac{\lambda A_r}{16} \left\| \bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right\|^2, \\
A_{r+1} \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{y}_i^{(r)} - \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle &\stackrel{(A.1)}{\geq} -\frac{4A_{r+1}}{\lambda} \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \\
&\quad - \frac{\lambda A_{r+1}}{16} \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2,
\end{aligned}$$

$$a_{r+1} \left\langle \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}), \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\rangle \stackrel{(A.1)}{\geq} -\frac{3A_{r+1}}{\lambda} \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 - \frac{B_r}{12} \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2,$$

where we use $\lambda = A_{r+1}B_r a_{r+1}^{-2}$ in the last inequality. By using the above inequalities, we get

$$\begin{aligned} \mathcal{T}^* &\geq -\frac{11A_{r+1}}{\lambda} \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 - \frac{33\lambda A_r}{16n} \sum_{i=1}^n \left\| \bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right\|^2 \\ &\quad - \frac{5B_r}{12n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2 + \frac{\lambda A_{r+1}}{16n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2. \end{aligned}$$

Then, using $\frac{1}{n} \sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\lambda^2}{352n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2$, we obtain

$$\begin{aligned} \mathcal{T}^* &\geq \frac{\lambda A_{r+1}}{32n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 - \frac{33\lambda A_r}{16n} \sum_{i=1}^n \left\| \bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right\|^2 \\ &\quad - \frac{5B_r}{12n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2. \end{aligned}$$

We have

$$\begin{aligned} &A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \\ &\geq \frac{A_{r+1}}{n} \sum_{i=1}^n f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) + \frac{A_{r+1}}{n} \sum_{i=1}^n \left\langle \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+\frac{1}{2})} \right\rangle + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\ &\quad + \frac{\lambda A_{r+1}}{32n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 - \frac{33\lambda A_r}{16n} \sum_{i=1}^n \left\| \bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)} \right\|^2 \\ &\quad + \frac{\mu A_r}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\ &\quad + \frac{\mu a_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 + \frac{B_r}{12n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 \\ &\quad + \frac{B_{r+1}}{2n} \left(\frac{1}{\rho^2} - 1 \right) \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2. \end{aligned}$$

Using the following inequality,

$$\begin{aligned} \frac{A_{r+1}}{n} \sum_{i=1}^n f_i(\mathbf{x}_i^{(r+\frac{1}{2})}) &\geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{A_{r+1}}{n} \sum_{i=1}^n \left\langle \nabla f_i(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\rangle \\ &\quad + \frac{\mu A_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\|^2. \end{aligned}$$

we get

$$\begin{aligned}
& A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{B_r}{2n} \sum_{i=1}^n \|\mathbf{v}_i^{(r)} - \mathbf{x}^*\|^2 \\
& \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \underbrace{\frac{A_{r+1}}{n} \sum_{i=1}^n \left\langle \nabla f_i(\bar{\mathbf{x}}^{(r+1)}) + \mathbf{h}_i^{(r)}, \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\rangle}_{\mathcal{T}^{**}} \\
& \quad + \frac{\mu A_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 + \frac{B_{r+1}}{2n} \sum_{i=1}^n \|\mathbf{v}_i^{(r+1)} - \mathbf{x}^*\|^2 \\
& \quad + \frac{\lambda A_{r+1}}{32n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 - \frac{33\lambda A_r}{16n} \sum_{i=1}^n \|\bar{\mathbf{x}}^{(r)} - \mathbf{x}_i^{(r)}\|^2 \\
& \quad + \frac{\mu A_r}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\
& \quad + \frac{\mu a_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 + \frac{B_r}{12n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2 \\
& \quad + \frac{B_{r+1}}{2n} \left(\frac{1}{\rho^2} - 1 \right) \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2.
\end{aligned}$$

Using the following inequality, we obtain

$$\begin{aligned}
\mathcal{T}^{**} &= \frac{A_{r+1}}{n} \sum_{i=1}^n \left\langle \nabla f_i(\bar{\mathbf{x}}^{(r+1)}) + \mathbf{h}_i^{(r)} - \nabla f(\bar{\mathbf{x}}^{(r+1)}), \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{y}}^{(r)} \right\rangle \\
&\stackrel{(A.1)}{\geq} -\frac{A_{r+1}}{2\delta n} \sum_{i=1}^n \left\| \nabla f_i(\bar{\mathbf{x}}^{(r+1)}) + \mathbf{h}_i^{(r)} - \nabla f(\bar{\mathbf{x}}^{(r+1)}) \right\|^2 \\
&\quad - \frac{\delta A_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{y}}^{(r)} \right\|^2 \\
&\stackrel{(A.2), (5.1)}{\geq} -\delta A_{r+1} \left\| \bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{y}}^{(r)} \right\|^2 - \frac{A_{r+1}}{\delta n} \sum_{i=1}^n \left\| \nabla f_i(\bar{\mathbf{y}}^{(r)}) + \mathbf{h}_i^{(r)} - \nabla f(\bar{\mathbf{y}}^{(r)}) \right\|^2 \\
&\quad - \frac{\delta A_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{y}}^{(r)} \right\|^2 \\
&\stackrel{(A.3)}{\geq} -\frac{2\delta A_{r+1}}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 - \frac{A_{r+1}}{\delta} \mathcal{E}^{(r)} - \frac{\delta A_{r+1}}{n} \sum_{i=1}^n \left\| \mathbf{y}_i^{(r)} - \bar{\mathbf{y}}^{(r)} \right\|^2 \\
&\geq -\frac{2\delta A_{r+1}}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 - \frac{A_{r+1}}{\delta} \mathcal{E}^{(r)} - \frac{\delta A_r}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\
&\quad - \frac{\delta a_{r+1}}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2,
\end{aligned}$$

where we use Jensen's inequality in the last inequality. Using the above inequality, we get

$$\begin{aligned}
& A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{B_r}{2n} \sum_{i=1}^n \|\mathbf{v}_i^{(r)} - \mathbf{x}^*\|^2 + \frac{A_{r+1}}{\delta} \mathcal{E}^{(r)} \\
& \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{B_{r+1}}{2n} \sum_{i=1}^n \|\mathbf{v}_i^{(r+1)} - \mathbf{x}^*\|^2 \\
& \quad - \frac{\delta a_{r+1}}{n} \sum_{i=1}^n \|\mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)}\|^2 + \frac{\mu A_{r+1}}{2n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)}\|^2 \\
& \quad + \frac{A_{r+1}}{n} \left(\frac{\lambda}{32} - 2\delta \right) \sum_{i=1}^n \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)}\|^2 \\
& \quad - \left(\frac{33\lambda A_r}{16n} + \frac{\delta A_r}{n} \right) \sum_{i=1}^n \|\mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)}\|^2 + \frac{\mu A_r}{2n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r)}\|^2 \\
& \quad + \frac{\mu a_{r+1}}{2n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r+\frac{1}{2})}\|^2 + \frac{B_r}{12n} \sum_{i=1}^n \|\mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)}\|^2 \\
& \quad + \frac{B_{r+1}}{2n} \left(\frac{1}{\rho^2} - 1 \right) \sum_{i=1}^n \|\mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)}\|^2 \\
& \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{B_{r+1}}{2n} \sum_{i=1}^n \|\mathbf{v}_i^{(r+1)} - \mathbf{x}^*\|^2 \\
& \quad - \frac{\delta a_{r+1}}{n} \sum_{i=1}^n \|\mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)}\|^2 + \frac{\mu A_{r+1}}{2\rho^2 n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)}\|^2 \\
& \quad + \frac{A_{r+1}}{n} \left(\frac{\lambda}{32} - 2\delta \right) \sum_{i=1}^n \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)}\|^2 \\
& \quad - \left(\frac{33\lambda A_r}{16n} + \frac{\delta A_r}{n} \right) \sum_{i=1}^n \|\mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)}\|^2 + \frac{\mu A_r}{2n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r)}\|^2 \\
& \quad + \frac{\mu a_{r+1}}{2n} \sum_{i=1}^n \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r+\frac{1}{2})}\|^2 + \frac{B_r}{12n} \sum_{i=1}^n \|\mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)}\|^2 \\
& \quad + \frac{B_{r+1}}{2n} \left(\frac{1}{\rho^2} - 1 \right) \sum_{i=1}^n \|\mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)}\|^2
\end{aligned}$$

By summarizing the above inequality, we obtain

$$\begin{aligned}
& A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 + \frac{A_{r+1}}{\delta} \mathcal{E}^{(r)} \\
& \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\
& \quad + \frac{A_{r+1}}{n} \left(\frac{\lambda}{32} - 2\delta \right) \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 \\
& \quad + \frac{B_r}{12n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2 + \frac{B_{r+1}}{2n} \left(\frac{1}{\rho^2} - 1 \right) \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \bar{\mathbf{v}}^{(r+1)} \right\|^2 \\
& \quad - \left(\frac{33\lambda A_r}{16n} + \frac{\delta A_r}{n} \right) \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2 - \frac{\delta a_{r+1}}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2 \\
& \quad + \frac{\mu A_{r+1}}{2\rho^2 n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 + \frac{\mu a_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2.
\end{aligned}$$

This concludes the statement. $\square$

Lemma 29. *Suppose that Assumptions 2 and 5.3 hold. Then, when $M = 1$, it holds that*

$$\begin{aligned}
4\rho^2 \mathcal{E}^{(r)} & \geq \mathcal{E}^{(r+1)} - 8\rho^2 \delta^2 \frac{a_{r+2}}{A_{r+2}} \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\
& \quad - 8\rho^2 \delta^2 \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 - 32L^2 \frac{A_{r+1}}{A_{r+2}} \Xi_{\mathbf{x}}^{(r+1)} - 32L^2 \frac{a_{r+2}}{A_{r+2}} \Xi_{\mathbf{v}}^{(r+1)},
\end{aligned} \tag{D.21}$$

where $\Xi_{\mathbf{x}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2$, $\Xi_{\mathbf{v}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2$, and $\mathcal{E}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{y}}^{(r)}) \right\|^2$.

Proof. From Lemma 15, we have

$$\mathcal{E}^{(r+1)} \leq 4\rho^2 \mathcal{E}^{(r)} + 4\rho^2 \delta^2 \left\| \bar{\mathbf{y}}^{(r+1)} - \bar{\mathbf{y}}^{(r)} \right\|^2 + 32L^2 \Xi_{\mathbf{y}}^{(r+1)}, \tag{D.22}$$

Using $\bar{\mathbf{y}}^{(r+1)} = \frac{A_{r+1}\bar{\mathbf{x}}^{(r+1)} + a_{r+2}\bar{\mathbf{v}}^{(r+1)}}{A_{r+2}}$, we have

$$\begin{aligned}
\mathcal{E}^{(r+1)} &\leq 4\rho^2\mathcal{E}^{(r)} + 8\rho^2\delta^2 \left(\frac{a_{r+2}}{A_{r+2}}\right)^2 \left\|\bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)}\right\|^2 \\
&\quad + 8\rho^2\delta^2 \left\|\bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{y}}^{(r)}\right\|^2 + 32L^2 \frac{A_{r+1}}{A_{r+2}} \Xi_{\mathbf{x}}^{(r+1)} + 32L^2 \frac{a_{r+2}}{A_{r+2}} \Xi_{\mathbf{v}}^{(r+1)} \\
&\leq 4\rho^2\mathcal{E}^{(r)} + 8\rho^2\delta^2 \frac{a_{r+2}}{A_{r+2}} \left\|\bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)}\right\|^2 \\
&\quad + 8\rho^2\delta^2 \left\|\bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{y}}^{(r)}\right\|^2 + 32L^2 \frac{A_{r+1}}{A_{r+2}} \Xi_{\mathbf{x}}^{(r+1)} + 32L^2 \frac{a_{r+2}}{A_{r+2}} \Xi_{\mathbf{v}}^{(r+1)} \\
&\leq 4\rho^2\mathcal{E}^{(r)} + 8\rho^2\delta^2 \frac{a_{r+2}}{A_{r+2}} \left\|\bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)}\right\|^2 \\
&\quad + 8\rho^2\delta^2 \frac{1}{n} \sum_{i=1}^n \left\|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)}\right\|^2 + 32L^2 \frac{A_{r+1}}{A_{r+2}} \Xi_{\mathbf{x}}^{(r+1)} + 32L^2 \frac{a_{r+2}}{A_{r+2}} \Xi_{\mathbf{v}}^{(r+1)}.
\end{aligned}$$

This concludes the statement. $\square$

Lemma 30. *Suppose that Assumptions 5 and 11 hold. Then, when $M = 1$, it holds that*

$$\begin{aligned}
&\left\|\bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)}\right\|^2 \\
&\leq \frac{A_r}{A_{r+1}} \left\|\bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r)}\right\|^2 + \frac{4A_{r+1}}{a_{r+1}} \frac{1}{n} \sum_{i=1}^n \left\|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)}\right\|^2 \\
&\quad + \frac{2A_{r+1}}{a_{r+1}} \frac{1}{n} \sum_{i=1}^n \left\|\mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)}\right\|^2.
\end{aligned} \tag{D.23}$$

Proof. For all $\gamma_1, \gamma_2 > 0$, it holds that

$$\begin{aligned}
&\left\|\bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)}\right\|^2 \\
&\stackrel{(A.2)}{\leq} (1 + \gamma_1) \left\|\bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r+1)}\right\|^2 + \left(1 + \frac{1}{\gamma_1}\right) \left\|\bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{v}}^{(r)}\right\|^2 \\
&\stackrel{(A.2)}{\leq} (1 + \gamma_1)(1 + \gamma_2) \left\|\bar{\mathbf{v}}^{(r)} - \bar{\mathbf{y}}^{(r)}\right\|^2 \\
&\quad + (1 + \gamma_1) \left(1 + \frac{1}{\gamma_2}\right) \left\|\bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{y}}^{(r)}\right\|^2 + \left(1 + \frac{1}{\gamma_1}\right) \left\|\bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{v}}^{(r)}\right\|^2 \\
&\leq (1 + \gamma_1)(1 + \gamma_2) \left(\frac{A_r}{A_{r+1}}\right)^2 \left\|\bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r)}\right\|^2 \\
&\quad + (1 + \gamma_1) \left(1 + \frac{1}{\gamma_2}\right) \left\|\bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{y}}^{(r)}\right\|^2 + \left(1 + \frac{1}{\gamma_1}\right) \left\|\bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{v}}^{(r)}\right\|^2.
\end{aligned}$$

Substituting $\gamma_1 = \gamma_2 = \min\{1, \sqrt{\frac{A_{r+1}}{A_r}} - 1\}$, we obtain

$$\begin{aligned} (1 + \gamma_1)(1 + \gamma_2) &\leq \frac{A_{r+1}}{A_r}, \\ \left(1 + \frac{1}{\gamma_1}\right) &= \max \left\{ 2, \frac{\sqrt{A_{r+1}}(\sqrt{A_{r+1}} + \sqrt{A_r})}{A_{r+1} - A_r} \right\} \\ &\leq \max \left\{ 2, \frac{2A_{r+1}}{a_{r+1}} \right\} = \frac{2A_{r+1}}{a_{r+1}}, \\ (1 + \gamma_1) \left(1 + \frac{1}{\gamma_2}\right) &\leq \max \left\{ 4, \frac{4A_{r+1}}{a_{r+1}} \right\} = \frac{4A_{r+1}}{a_{r+1}}, \end{aligned}$$

where we use $A_{r+1} \geq A_r$ and $A_{r+1} \geq a_{r+1}$. Using the above inequalities, we obtain the statement. $\square$

Lemma 31. *Suppose that Assumption 11 holds. Then, when $M = 1$, it holds that*

$$A_{r+1}\Xi_x^{(r+1)} \leq \rho A_r \Xi_x^{(r)} + \frac{\rho^2 A_{r+1}}{1 - \rho} \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 + \frac{\rho^2 a_{r+1}}{1 - \rho} \Xi_v^{(r)}, \quad (\text{D.24})$$

where $\Xi_x^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2$ and $\Xi_v^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2$.

Proof. It holds that for all $\gamma > 0$,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\ &\stackrel{(5.3)}{\leq} \frac{\rho^2}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\ &\stackrel{(D.1)}{=} \frac{\rho^2}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \bar{\mathbf{y}}^{(r)} \right\|^2 - \rho^2 \left\| \bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{y}}^{(r)} \right\|^2 \\ &\stackrel{(A.2)}{\leq} \frac{(1 + \gamma)\rho^2}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 + \left(1 + \frac{1}{\gamma}\right) \frac{\rho^2}{n} \sum_{i=1}^n \left\| \mathbf{y}_i^{(r)} - \bar{\mathbf{y}}^{(r)} \right\|^2 \\ &\quad - \rho^2 \left\| \bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{y}}^{(r)} \right\|^2 \\ &\leq \frac{(1 + \gamma)\rho^2}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 + \left(1 + \frac{1}{\gamma}\right) \frac{\rho^2 A_r}{n A_{r+1}} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\ &\quad + \left(1 + \frac{1}{\gamma}\right) \frac{\rho^2 a_{r+1}}{n A_{r+1}} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2 - \rho^2 \left\| \bar{\mathbf{x}}^{(r+1)} - \bar{\mathbf{y}}^{(r)} \right\|^2, \end{aligned}$$

where we use Jensen's inequality in the last inequality. Substituting $\gamma = \frac{\rho}{1-\rho}$, we obtain the statement. $\square$

D.5.3 Convex Case

Lemma 32. *Suppose that Assumptions 5, 2, and 11 hold with $\mu = 0$, and the following inequality is satisfied:*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\lambda^2}{352} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2.$$

Then, when $\lambda = 208\delta$, $M = 1$, and

$$\rho \leq \min \left\{ \frac{9}{1600}, \frac{\delta}{6L}, \frac{5}{48(R+1)^3} \right\},$$

it holds that

$$\begin{aligned} & A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \\ & + \frac{2A_{r+1}}{\delta} \mathcal{E}^{(r)} + 990\delta A_r \Xi_{\mathbf{x}}^{(r)} + \delta a_{r+1} \Xi_{\mathbf{v}}^{(r)} + \frac{R\rho}{24(R+1)^2} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\ & \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\ & + \frac{2A_{r+2}}{\delta} \mathcal{E}^{(r+1)} + 990\delta A_{r+1} \Xi_{\mathbf{x}}^{(r+1)} + \delta a_{r+2} \Xi_{\mathbf{v}}^{(r+1)} \\ & + \frac{R\rho}{24(R+1)^2} \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2, \end{aligned}$$

where $\Xi_{\mathbf{x}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2$, $\Xi_{\mathbf{v}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2$, and $\mathcal{E}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{y}}^{(r)}) \right\|^2$.

Proof. Using $\rho \leq \frac{1}{4}$, we can simplify Lemmas 29 and 31 as follows:

$$\begin{aligned} \rho \mathcal{E}^{(r)} & \geq \mathcal{E}^{(r+1)} - 8\rho^2 \delta^2 \frac{a_{r+2}}{A_{r+2}} \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\ & - 8\rho^2 \delta^2 \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 \\ & - 32L^2 \frac{A_{r+1}}{A_{r+2}} \Xi_{\mathbf{x}}^{(r+1)} - 32L^2 \frac{a_{r+2}}{A_{r+2}} \Xi_{\mathbf{v}}^{(r+1)}, \end{aligned} \tag{D.25}$$

$$\begin{aligned} \rho A_r \Xi_{\mathbf{x}}^{(r)} & \geq A_{r+1} \Xi_{\mathbf{x}}^{(r+1)} - \frac{4}{3} \rho^2 A_{r+1} \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 \\ & - \frac{4}{3} \rho^2 a_{r+1} \Xi_{\mathbf{v}}^{(r)}. \end{aligned} \tag{D.26}$$

Using Lemma 26, we can simplify Lemma 30 as follows:

$$\begin{aligned} \frac{R}{R+1} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2 &\geq \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 - \frac{4A_{r+1}}{a_{r+1}} \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 \\ &\quad - \frac{2A_{r+1}}{a_{r+1}} \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2. \end{aligned} \quad (\text{D.27})$$

From Eq. (D.20) + $\frac{2A_{r+2}}{\delta} \times$ Eq. (D.25) + $\frac{\rho}{24(R+1)} \times$ Eq. (D.27) + $\frac{560\delta}{\rho} \times$ Eq. (D.26) and $\lambda = 208\delta$, we obtain

$$\begin{aligned} &A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 + \left(1 + \frac{2A_{r+2}}{A_{r+1}} \rho \right) \frac{A_{r+1}}{\delta} \mathcal{E}^{(r)} \\ &\quad + 990\delta A_r \Xi_{\mathbf{x}}^{(r)} + \delta a_{r+1} \Xi_{\mathbf{v}}^{(r)} + \frac{R\rho}{24(R+1)^2} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\ &\geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{2A_{r+2}}{\delta} \mathcal{E}^{(r+1)} \\ &\quad + \left(\frac{560\delta}{\rho} - \frac{32L^2}{\delta} \right) A_{r+1} \Xi_{\mathbf{x}}^{(r+1)} \\ &\quad + \left[\frac{1}{2a_{r+2}} \left(\frac{1}{\rho^2} - 1 \right) - \frac{32L^2}{\delta} - \frac{2240a_{r+1}}{3a_{r+2}} \delta \rho \right] a_{r+2} \Xi_{\mathbf{v}}^{(r+1)} \\ &\quad + \left(\frac{\rho}{24(R+1)} - 16a_{r+2}\rho^2\delta \right) \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\ &\quad + \left[\frac{9}{2}\delta - \frac{16A_{r+2}}{A_{r+1}}\rho^2\delta - \frac{\rho}{6a_{r+1}(R+1)} - \frac{2240}{3}\delta\rho \right] \frac{A_{r+1}}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 \\ &\quad + \left(\frac{1}{12} - \frac{A_{r+1}}{12a_{r+1}(R+1)}\rho \right) \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2, \end{aligned}$$

where we use that fact that $B_r = 1$ for all r when $\mu = 0$. From Lemmas 25, 26, 26, and 27, we have

$$\begin{aligned} \frac{A_{r+1}}{a_{r+1}} &\stackrel{(D.17)}{\leq} 1 + R, \quad \frac{A_{r+2}}{A_{r+1}} \stackrel{(D.17)}{\leq} 4, \quad \frac{a_{r+1}}{a_{r+2}} \stackrel{(D.19)}{\leq} 1, \\ \frac{1}{a_{r+2}} &\stackrel{(D.19)}{\leq} \frac{1}{a_1} = 208\delta, \quad \frac{1}{a_{r+1}} \stackrel{(D.19)}{\leq} \frac{1}{a_1} = 208\delta, \quad a_{r+2} \stackrel{(D.17), (D.15)}{\leq} \frac{R+1}{40\delta}. \end{aligned}$$

Using the above inequalities, we obtain

$$\begin{aligned}
& A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 + (1 + 8\rho) \frac{A_{r+1}}{\delta} \mathcal{E}^{(r)} \\
& + 990\delta A_r \Xi_{\mathbf{x}}^{(r)} + \delta a_{r+1} \Xi_{\mathbf{v}}^{(r)} + \frac{R\rho}{24(R+1)^2} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\
& \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 + \frac{2A_{r+2}}{\delta} \mathcal{E}^{(r+1)} \\
& + \left(\frac{560\delta}{\rho} - \frac{32L^2}{\delta} \right) A_{r+1} \Xi_{\mathbf{x}}^{(r+1)} \\
& + \left[104\delta \left(\frac{1}{\rho^2} - 1 \right) - \frac{32L^2}{\delta} - 750\delta\rho \right] a_{r+2} \Xi_{\mathbf{v}}^{(r+1)} \\
& + \left(\frac{\rho}{24(R+1)} - \frac{2(R+1)}{5}\rho^2 \right) \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\
& + \left[\frac{9}{2}\delta - 16\rho\delta - \frac{35}{(R+1)}\delta\rho - \frac{2240}{3}\delta\rho \right] \frac{A_{r+1}}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 \\
& + \left(\frac{1}{12} - \frac{1}{12}\rho \right) \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2,
\end{aligned}$$

Then, using the assumption on ρ , we have

$$\begin{aligned}
& 1 + 8\rho \leq 2, \\
& \frac{560\delta}{\rho} - \frac{32L^2}{\delta} \geq 900\delta, \\
& 104\delta \left(\frac{1}{\rho^2} - 1 \right) - \frac{32L^2}{\delta} - 750\delta\rho \geq \delta, \\
& \frac{\rho}{24(R+1)} - \frac{2(R+1)}{5}\rho^2 \geq \frac{R\rho}{24(R+1)^2}, \\
& \frac{9}{2}\delta - 16\delta\rho - \frac{35}{R+1}\delta\rho - \frac{2240}{3}\delta\rho \geq 0.
\end{aligned}$$

Thus, we can simplify the above inequality as follows:

$$\begin{aligned}
& A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \\
& + \frac{2A_{r+1}}{\delta} \mathcal{E}^{(r)} + 990\delta A_r \Xi_{\mathbf{x}}^{(r)} + \delta a_{r+1} \Xi_{\mathbf{v}}^{(r)} + \frac{R\rho}{24(R+1)^2} \left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2 \\
& \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\
& + \frac{2A_{r+2}}{\delta} \mathcal{E}^{(r+1)} + 990\delta A_{r+1} \Xi_{\mathbf{x}}^{(r+1)} + \delta a_{r+2} \Xi_{\mathbf{v}}^{(r+1)} \\
& + \frac{R\rho}{24(R+1)^2} \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2.
\end{aligned}$$

Thus, we obtain the desired result. $\square$

Lemma 33. *Suppose that Assumptions 5, 2, and 11 hold with $\mu = 0$, and the following inequality is satisfied:*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\lambda^2}{352} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2.$$

Then, when $\lambda = 208\delta$, $M = 1$, and

$$\rho \leq \min \left\{ \frac{9}{1600}, \frac{\delta}{6L}, \frac{5}{48(R+1)^3} \right\},$$

it holds that $f(\bar{\mathbf{x}}^{(R)}) - f(\mathbf{x}^*) \leq \epsilon$ after

$$R = \mathcal{O} \left(\sqrt{\frac{\delta \left\| \bar{\mathbf{x}}^{(0)} - \mathbf{x}^* \right\|^2}{\epsilon}} \right)$$

rounds.

Proof. From Lemma 32, we have

$$\begin{aligned}
& A_{r+1} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) \\
& \leq A_r \left(f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*) \right) + \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 - \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\
& + \frac{2A_{r+1}}{\delta} \mathcal{E}^{(r)} - \frac{2A_{r+2}}{\delta} \mathcal{E}^{(r+1)} + 990\delta \left(A_r \Xi_{\mathbf{x}}^{(r)} - A_{r+1} \Xi_{\mathbf{x}}^{(r+1)} \right) \\
& + \delta \left(a_{r+1} \Xi_{\mathbf{v}}^{(r)} - a_{r+2} \Xi_{\mathbf{v}}^{(r+1)} \right) \\
& + \frac{R(1-\rho)}{24(R+1)^2} \left(\left\| \bar{\mathbf{v}}^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2 - \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \right).
\end{aligned}$$

Telescoping the sum from $r = 0$ to $r = R - 1$ and using $\Xi_{\mathbf{x}}^{(0)} = \Xi_{\mathbf{v}}^{(0)} = \mathcal{E}^{(0)} = 0$ and $\mathbf{v}_i^{(0)} = \mathbf{x}_i^{(0)}$, we have

$$f(\bar{\mathbf{x}}^{(R)}) - f(\mathbf{x}^*) \leq \frac{1}{2A_R} \|\bar{\mathbf{v}}^{(0)} - \mathbf{x}^*\|^2.$$

Using Lemma 24, we obtain the desired result. $\square$

Lemma 34. *Suppose that Assumptions 5, 2, and 11 hold with $\mu = 0$, and the following inequality is satisfied:*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\lambda^2}{352} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2.$$

Then, when $\lambda = 208\delta$, there exists γ and $M = \tilde{\Omega}(\frac{1}{\sqrt{1-\rho}})$ such that it holds that $f(\bar{\mathbf{x}}^{(R)}) - f(\mathbf{x}^*) \leq \epsilon$ after

$$R = \mathcal{O} \left(\sqrt{\frac{\delta \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}} \right)$$

rounds.

Proof. Define $\tilde{\mathbf{W}}^{(m)}$ as follows:

$$\begin{aligned} \tilde{\mathbf{W}}^{(-1)} &= \mathbf{I} \\ \tilde{\mathbf{W}}^{(0)} &= \mathbf{I} \\ \tilde{\mathbf{W}}^{(m+1)} &= (1 + \gamma) \tilde{\mathbf{W}}^{(m)} \mathbf{W} - \gamma \tilde{\mathbf{W}}^{(m-1)}. \end{aligned}$$

Then, the output of Alg. 10 is equivalent to

$$\sum_{j=1}^n \tilde{\mathbf{W}}_{ij}^{(M)} \mathbf{a}_j.$$

From Proposition 1 in Yuan et al. (2022), there exists γ such that it holds that for any $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \in \mathbb{R}^d$

$$\sum_{i=1}^n \left\| \sum_{j=1}^n \tilde{W}_{ij}^{(M)} \mathbf{a}_j - \bar{\mathbf{a}} \right\| \leq \sqrt{2} \left(1 - \sqrt{1-\rho} \right)^M \sum_{i=1}^n \|\mathbf{a}_i - \bar{\mathbf{a}}\|,$$

where $\bar{\mathbf{a}} := \frac{1}{n} \sum_{i=1}^n \mathbf{a}_i$. Thus, there exists $M = \tilde{\Omega}(\frac{1}{\sqrt{1-\rho}})$ that satisfies the condition on ρ in Lemma 33. From Lemma 33, we obtain the desired result. $\square$

D.5.4 Strongly Convex Case

Lemma 35. *Suppose that Assumptions 5 and 11 hold and the following inequality is satisfied:*

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\lambda^2}{352} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2.$$

Then, when $\lambda = 96\delta$, $M = 1$ and

$$\begin{aligned} \rho \leq \min & \left[\sqrt{\frac{\mu\delta}{1728L^2}}, \sqrt{\frac{48\delta^2\mu}{260L^2(192\delta + \mu)}}, \right. \\ & \left(4 + 2 \left(1 + \frac{\mu}{96\delta} \right) \left(1 + \sqrt{\frac{384\delta}{\mu}} \right) \right)^{-1}, \\ & \left. \frac{\mu}{4\delta} \left(\left(1 + \frac{96\delta}{\mu} \right) \left(1 + \frac{192\delta}{\mu} \right) \left(1 + \sqrt{1 + \frac{384\delta}{\mu}} \right) \right)^{-1} \right], \end{aligned}$$

it holds that

$$\begin{aligned} & A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^*) + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 \\ & + \frac{2A_{r+1}}{\delta} \mathcal{E}^{(r)} + 200\delta A_r \Xi_{\mathbf{x}}^{(r)} + \delta a_{r+1} \Xi_{\mathbf{v}}^{(r)} \\ & \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\ & + \frac{2A_{r+2}}{\delta} \mathcal{E}^{(r+1)} + 200\delta A_{r+1} \Xi_{\mathbf{x}}^{(r+1)} + \delta a_{r+2} \Xi_{\mathbf{v}}^{(r+1)}, \end{aligned} \tag{D.28}$$

where $\Xi_{\mathbf{x}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r)} - \bar{\mathbf{x}}^{(r)} \right\|^2$, $\Xi_{\mathbf{v}}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \bar{\mathbf{v}}^{(r)} \right\|^2$, and $\mathcal{E}^{(r)} := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{h}_i^{(r)} - \nabla h_i(\bar{\mathbf{y}}^{(r)}) \right\|^2$.

Proof. From Lemma 15, $\rho \leq \frac{1}{4}$ and $\bar{\mathbf{y}}^{(r+1)} = \frac{A_{r+1}\bar{\mathbf{x}}^{(r+1)} + a_{r+2}\bar{\mathbf{v}}^{(r+1)}}{A_{r+2}}$, we have

$$\begin{aligned} \rho \mathcal{E}^{(r)} & \geq \mathcal{E}^{(r+1)} - 2\rho\delta^2 \frac{a_{r+2}}{A_{r+2}} \left\| \bar{\mathbf{v}}^{(r+1)} - \bar{\mathbf{x}}^{(r+1)} \right\|^2 \\ & - 2\rho\delta^2 \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 - 32L^2 \frac{A_{r+1}}{A_{r+2}} \Xi_{\mathbf{x}}^{(r+1)} - 32L^2 \frac{a_{r+2}}{A_{r+2}} \Xi_{\mathbf{v}}^{(r+1)}, \end{aligned} \tag{D.29}$$

From Eq. (D.20) + $\frac{2A_{r+2}}{\delta} \times$ Eq. (D.29) and $\lambda = 96\delta$, we have

$$\begin{aligned}
& A_r f(\bar{\mathbf{x}}^{(r)}) + a_{r+1} f(\mathbf{x}^\star) \\
& + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^\star \right\|^2 + \frac{A_{r+1}}{\delta} \left(1 + \frac{2A_{r+2}\rho}{A_{r+1}} \right) \mathcal{E}^{(r)} \\
& + 200\delta A_r \Xi_{\mathbf{x}}^{(r)} + \delta a_{r+1} \Xi_{\mathbf{v}}^{(r)} \\
& \geq A_{r+1} f(\bar{\mathbf{x}}^{(r+1)}) + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^\star \right\|^2 + \frac{2A_{r+2}}{\delta} \mathcal{E}^{(r+1)} \\
& + \delta A_{r+1} \left[\frac{\mu}{2\delta\rho^2} - \frac{64L^2}{\delta^2} \right] \Xi_{\mathbf{x}}^{(r+1)} \\
& + \delta a_{r+2} \left[\frac{B_{r+1}}{2\delta a_{r+2}} \left(\frac{1}{\rho^2} - 1 \right) - \frac{64L^2}{\delta^2} \right] \Xi_{\mathbf{v}}^{(r+1)} \\
& + \frac{B_r}{12n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r)} \right\|^2 \\
& + \frac{\delta}{n} \left[A_{r+1} - 4\rho A_{r+2} \right] \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2 \\
& + \frac{1}{n} \left[\frac{\mu a_{r+1}}{2} - 4\rho\delta a_{r+2} \right] \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{v}_i^{(r+\frac{1}{2})} \right\|^2.
\end{aligned}$$

Using Lemmas 26, and 27, $\delta \leq 2L$, and the condition of ρ , we have

$$\begin{aligned}
1 + \frac{2A_{r+2}}{A_{r+1}}\rho &= 1 + 2 \left(1 + \frac{a_{r+2}}{A_{r+1}} \right) \rho \leq 2, \\
\frac{\mu}{2\delta\rho^2} - \frac{64L^2}{\delta^2} &\geq 200, \\
\frac{B_{r+1}}{2\delta a_{r+2}} \left(\frac{1}{\rho^2} - 1 \right) - \frac{64L^2}{\delta^2} &= \frac{48a_{r+2}}{A_{r+2}} \left(\frac{1}{\rho^2} - 1 \right) - \frac{64L^2}{\delta^2} \geq 1, \\
A_{r+1} - 4\rho A_{r+2} &\geq 0, \\
\frac{\mu a_{r+1}}{2} - 4\rho\delta a_{r+2} &\geq 0.
\end{aligned}$$

Using the above inequalities, we obtain the desired result. $\square$

Lemma 36. Suppose that Assumptions 5, 2, and 11 hold and the following inequality is satisfied:

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\lambda^2}{352} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2,$$

Then, when $\lambda = 96\delta$, $M = 1$, and

$$\rho \leq \min \left[\sqrt{\frac{\mu\delta}{1728L^2}}, \sqrt{\frac{48\delta^2\mu}{260L^2(192\delta + \mu)}}, \left(4 + 2 \left(1 + \frac{\mu}{96\delta} \right) \left(1 + \sqrt{\frac{384\delta}{\mu}} \right) \right)^{-1}, \frac{\mu}{4\delta} \left(\left(1 + \frac{96\delta}{\mu} \right) \left(1 + \frac{192\delta}{\mu} \right) \left(1 + \sqrt{1 + \frac{384\delta}{\mu}} \right) \right)^{-1} \right],$$

it holds that $f(\bar{\mathbf{x}}^{(R)}) - f(\mathbf{x}^*) \leq \epsilon$ after

$$R = \mathcal{O} \left(\sqrt{\frac{\delta + \mu}{\mu}} \log \left(1 + \sqrt{\frac{\min\{\mu, \delta\} \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}} \right) \right)$$

rounds.

Proof. We have

$$\begin{aligned} & A_r \left(f(\bar{\mathbf{x}}^{(r)}) - f(\mathbf{x}^*) \right) + \frac{B_r}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r)} - \mathbf{x}^* \right\|^2 + \frac{2A_{r+1}}{\delta} \mathcal{E}^{(r)} \\ & \quad + 200\delta A_r \Xi_x^{(r)} + \delta a_{r+1} \Xi_v^{(r)} \\ & \geq A_{r+1} \left(f(\bar{\mathbf{x}}^{(r+1)}) - f(\mathbf{x}^*) \right) + \frac{B_{r+1}}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(r+1)} - \mathbf{x}^* \right\|^2 \\ & \quad + \frac{2A_{r+2}}{\delta} \mathcal{E}^{(r+1)} + 200\delta A_{r+1} \Xi_x^{(r+1)} + \delta a_{r+2} \Xi_v^{(r+1)}. \end{aligned}$$

Telescoping the sum from $r = 0$ to $r = R - 1$ and using $\Xi_v^{(0)} = \Xi_x^{(0)} = \mathcal{E}^{(0)} = 0$, we have

$$A_R \left(f(\bar{\mathbf{x}}^{(R)}) - f(\mathbf{x}^*) \right) \leq \frac{B_0}{2n} \sum_{i=1}^n \left\| \mathbf{v}_i^{(0)} - \mathbf{x}^* \right\|^2 = \frac{1}{2} \left\| \bar{\mathbf{x}}^{(0)} - \mathbf{x}^* \right\|^2.$$

From Lemma 24, we obtain

$$f(\bar{\mathbf{x}}^{(R)}) - f(\mathbf{x}^*) \leq \begin{cases} \frac{2\mu}{\left[\left(1 + \sqrt{\frac{\mu}{384\delta}} \right)^R - 1 \right]^2} \left\| \bar{\mathbf{x}}^{(0)} - \mathbf{x}^* \right\|^2 & \text{if } \mu \leq 384\delta, \\ \frac{96\delta}{4^{R-2}} \left\| \bar{\mathbf{x}}^{(0)} - \mathbf{x}^* \right\|^2 & \text{if } \mu > 384\delta \end{cases},$$

for any $R \geq 1$. Then, we obtain the desired result. $\square$

Lemma 37. Suppose that Assumptions 5, 2, and 11 hold and the following inequality is satisfied:

$$\sum_{i=1}^n \left\| \nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})}) \right\|^2 \leq \frac{\lambda^2}{352} \sum_{i=1}^n \left\| \mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)} \right\|^2,$$

Then, when $\lambda = 96\delta$, there exists γ and $M = \tilde{\Omega}(\frac{1}{\sqrt{1-\rho}})$ such that it holds that $f(\bar{\mathbf{x}}^{(R)}) - f(\mathbf{x}^*) \leq \epsilon$ after

$$R = \mathcal{O} \left(\sqrt{\frac{\delta + \mu}{\mu}} \log \left(1 + \sqrt{\frac{\min\{\mu, \delta\} \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2}{\epsilon}} \right) \right)$$

rounds.

Proof. Define $\tilde{\mathbf{W}}^{(m)}$ as follows:

$$\begin{aligned} \tilde{\mathbf{W}}^{(-1)} &= \mathbf{I} \\ \tilde{\mathbf{W}}^{(0)} &= \mathbf{I} \\ \tilde{\mathbf{W}}^{(m+1)} &= (1 + \gamma) \tilde{\mathbf{W}}^{(m)} \mathbf{W} - \gamma \tilde{\mathbf{W}}^{(m-1)}. \end{aligned}$$

Then, the output of Alg. 10 is equivalent to

$$\sum_{j=1}^n \tilde{\mathbf{W}}_{ij}^{(M)} \mathbf{a}_j.$$

From Proposition 1 in Yuan et al. (2022), there exists γ such that it holds that for any $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \in \mathbb{R}^d$

$$\sum_{i=1}^n \left\| \sum_{j=1}^n \tilde{W}_{ij}^{(M)} \mathbf{a}_j - \bar{\mathbf{a}} \right\|^2 \leq \sqrt{2} \left(1 - \sqrt{1-\rho} \right)^M \sum_{i=1}^n \|\mathbf{a}_i - \bar{\mathbf{a}}\|^2,$$

where $\bar{\mathbf{a}} := \frac{1}{n} \sum_{i=1}^n \mathbf{a}_i$. Thus, there exists $M = \tilde{\Omega}(\frac{1}{\sqrt{1-\rho}})$ that satisfies the condition on ρ in Lemma 36. From Lemma 36, we obtain the desired result.

D.6 Analysis of Computational Complexity

D.6.1 Useful Lemmas

Lemma 38. *Let $F : \mathbb{R}^d \rightarrow \mathbb{R}$ be a μ -strongly convex function and $\mathbf{x}^*$ be a minimizer of F . If the following is satisfied for some $\alpha \leq \mu$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$*

$$\|\nabla F(\mathbf{x})\| \leq \alpha \|\mathbf{y} - \mathbf{x}^*\|,$$

it holds that

$$\|\nabla F(\mathbf{x})\| \leq \frac{\alpha\mu}{\mu - \alpha} \|\mathbf{x} - \mathbf{y}\|.$$

Proof. We have

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\| &\geq \|\mathbf{y} - \mathbf{x}^*\| - \|\mathbf{x} - \mathbf{x}^*\| \\ &\geq \|\mathbf{y} - \mathbf{x}^*\| - \frac{1}{\mu} \|\nabla F(\mathbf{x})\|, \end{aligned}$$

where we use the strongly-convexity of F in the last inequality. Using $\|\nabla F(\mathbf{x})\| \leq \alpha \|\mathbf{y} - \mathbf{x}^*\|$, we obtain the desired result. $\square$

Lemma 39. *Let $F : \mathbb{R}^d \rightarrow \mathbb{R}$ be a μ -strongly convex and L -smooth function and $\mathbf{x}^*$ be a minimizer of F . If we use Nesterov's Accelerated Gradient Descent with initial parameter $\mathbf{x}^{(0)} \in \mathbb{R}^d$, it holds that for any $\beta > 0$*

$$\|\nabla F(\mathbf{x}^{(T)})\| \leq \beta \|\mathbf{x}^{(0)} - \mathbf{x}^{(T)}\|$$

after

$$T = \mathcal{O} \left(\sqrt{\frac{L}{\mu}} \log \left(\frac{L(\mu + L)(\mu + \beta)^2}{\mu^2 \beta^2} \right) \right)$$

iterations where $\mathbf{x}^{(T)}$ is the output of Nesterov's Accelerated Gradient Descent.

Proof. From Theorem 3.18 in Bubeck (2015), it holds that

$$F(\mathbf{x}^{(T)}) - F(\mathbf{x}^*) \leq \frac{\mu + L}{2} \|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 \exp \left(-T \sqrt{\frac{\mu}{L}} \right).$$

Using the L -smoothness, we have

$$\|\nabla F(\mathbf{x}^{(T)})\|^2 \leq L(\mu + L) \|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 \exp \left(-T \sqrt{\frac{\mu}{L}} \right).$$

Thus, when $T \geq \sqrt{\frac{L}{\mu}} \log \left(\frac{L(\mu + L)}{\alpha^2} \right)$, it holds that

$$\|\nabla F(\mathbf{x}^{(T)})\| \leq \alpha \|\mathbf{x}^{(0)} - \mathbf{x}^*\|.$$

Thus, using Lemma 38, we obtain the desired result. $\square$

D.6.2 Proof of Theorem 7

Lemma 40. *Consider Alg. 6. Suppose that Assumptions 5, 2 and 11 hold, and $\mathbf{h}_i^{(0)} = \nabla f(\mathbf{x}_i^{(0)}) - \nabla f_i(\mathbf{x}_i^{(0)})$, $\mathbf{x}_i^{(0)} = \bar{\mathbf{x}}^{(0)}$, and we use the same λ and M as in Lemma 19. Then, if we use Nesterov's Accelerated Gradient Descent with initial parameter $\mathbf{x}_i^{(r)}$ to approximately solve the subproblem in line 4, each node requires at most*

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu + \delta}} \log\left(\frac{L^2(r+2)^2}{\delta(\delta + \mu)}\right)\right)$$

iterations to satisfy Eq. (5.6).

Proof. $F_{i,r}$ is $\Omega(\mu + \delta)$ -strongly convex and $\mathcal{O}(L + \delta)$ -smooth. Thus, using Lemma 39, $\mu \leq L$, and $\delta \leq 2L$, we obtain the desired result. $\square$

D.6.3 Proof of Theorem 9

Lemma 41. *Consider Alg. 8. Suppose that Assumptions 5, 2, and 11 hold, and $\mathbf{h}_i^{(0)} = \nabla f(\mathbf{v}_i^{(0)}) - \nabla f_i(\mathbf{v}_i^{(0)})$, $\mathbf{v}_i^{(0)} = \bar{\mathbf{v}}^{(0)}$, and we use the same λ and M as in Lemma 23. Then, if we use Nesterov's Accelerated Gradient Descent with initial parameter $\mathbf{v}_i^{(r)}$ to approximately solve the subproblem in line 4, each node requires at most*

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu + \delta}} \log\left(\frac{L}{\delta}\right)\right)$$

iterations to satisfy Eq. (5.7).

Proof. $F_{i,r}$ is $\Omega(\mu + \delta)$ -strongly convex and $\mathcal{O}(L + \delta)$ -smooth. Thus, using Lemma 39, $\mu \leq L$, and $\delta \leq 2L$, we obtain the desired result. $\square$

D.6.4 Proof of Theorem 11

Lemma 42. *Consider Alg. 9. Suppose that Assumptions 5, 2, and 11 hold, and $\mathbf{h}_i^{(0)} = \nabla f(\mathbf{v}_i^{(0)}) - \nabla f_i(\mathbf{v}_i^{(0)})$, $\mathbf{v}_i^{(0)} = \bar{\mathbf{v}}^{(0)}$, and we use the same γ , M , and λ as Theorem 10. Then, if we use Nesterov's Accelerated Gradient Descent with initial parameter $\mathbf{y}_i^{(r)}$ to approximately solve the subproblem in line 12, each node requires at most*

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu + \delta}} \log\left(\frac{L}{\delta}\right)\right)$$

iterations to satisfy Eq. (5.8).

Proof. $F_{i,r}$ is $\Omega(\mu + \delta)$ -strongly convex and $\mathcal{O}(L + \delta)$ -smooth. Thus, using Lemma 39, $\mu \leq L$, and $\delta \leq 2L$, we obtain the desired result. $\square$

D.7 Experimental Setup

D.7.1 Setup for Adaptive Number of Local Steps

In Fig. 5.1a, we ran gradient descent until the following condition was satisfied.

- **Inexact SONATA:** $\|\nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})})\| \leq \frac{\lambda(\lambda+\mu)}{(r+1)(r+2)} \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{x}_i^{(r)}\|$.
- **Stabilized SONATA:** $\|\nabla F_{i,r}(\mathbf{x}_i^{(r+1)})\| \leq \lambda \|\mathbf{x}_i^{(r+1)} - \mathbf{v}_i^{(r)}\|$.
- **Inexact Accelerated SONATA:** $\|\nabla F_{i,r}(\mathbf{x}_i^{(r+1)})\| \leq (1 - \sqrt{\frac{\mu}{\beta}})^r (\frac{16}{17})^t \|\mathbf{x}_i^{(r+1)} - \mathbf{x}_i^{(r)}\|$ where t is the number of inner loop.
- **Accelerated Stabilized SONATA:** $\|\nabla F_{i,r}(\mathbf{x}_i^{(r+\frac{1}{2})})\| \leq \lambda \|\mathbf{x}_i^{(r+\frac{1}{2})} - \mathbf{y}_i^{(r)}\|$.

D.7.2 Hyperparameter Setting

In Fig. 5.1, we tuned the hyperparameters by grid search. Specifically, we ran grid search over the hyperparameters listed in Table D.2 and chose the hyperparameter that minimizes the norm of the last gradient.

Table D.2: Experimental setups. Note that for Inexact Accelerated SONATA, δ is the coefficient of the additional L2 regularization in the subproblem, and it is different from Definition 2.

Gradient Tracking	η	{0.05, 0.01, 0.005, 0.001}
	M	{10, 20, 30}
Inexact SONATA	η	{0.05, 0.01, 0.005, 0.001}
	λ	{1, 10}
	M	{10, 20, 30}
(Accelerated) Stabilized SONATA	μ	We set the same value as the coefficient of L2 regularization.
	η	{0.05, 0.01, 0.005, 0.001}
	λ	{1, 10}
	M	{10, 20, 30}
Inexact Accelerated SONATA	η	{0.05, 0.01, 0.005, 0.001}
	β	{1, 10}
	M	{10, 20, 30}
	The number of inner loop T	{1, 2, 3, 4, 5}
	The coefficient of the regularizer δ	{0.1, 1, 10}

□