Efficient Algorithms for Distributed Networks with Applications in Communication and Learning

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Hadi Reisizadeh

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Prof. Soheil Mohajer

August, 2024

Acknowledgements

Completing this PhD thesis has been a significant milestone in my life, and it would not have been possible without the support and encouragement of many people.

First and foremost, I would like to express my deepest gratitude to my Ph.D. advisor, Professor Soheil Mohajer. Your guidance, mentorship, and unwavering support have been invaluable. Your insightful feedback, constructive criticism, and encouragement have shaped this thesis and significantly contributed to my growth as a researcher. Thank you for believing in my potential and pushing me to strive for excellence. Your dedication and commitment to my success have been truly inspiring.

My PhD journey would not have been as fulfilling without the fruitful collaborations I enjoyed with an extraordinarily talented group of mentors and colleagues. My first collaborative PhD project was a joint effort with Prof. Mohammad Ali Maddah-Ali. As a leading expert in the field, I was fortunate to learn the fundamentals of research and presentation from him. I am also deeply grateful to Prof. Behrouz Touri from UCSD. Our collaboration sparked my initial interest in optimization theory and machine learning. His ability to generate brilliant ideas and his remarkable attention to the intricate technical details of proofs have been invaluable. Over the past few years, he has been an incredible source of support and guidance.

I want to extend my gratitude to Prof. Yusuf Saad, Prof. Behrouz Touri, and Mohammad Ali Maddah-Ali for agreeing to serve on my committee. Their comments and feedback were invaluable in enhancing the quality and presentation of my thesis. I also want to thank my University of Minnesota professors for their passion and dedication to teaching. In particular, I am grateful to Professors Mingyi Hong, Steven Wu, Zhaosong Lu, and Mehmet Akcakaya, whose teachings provided me with the necessary foundation for conducting research in Electrical Engineering and Computer Science.

I am incredibly fortunate to have met many wonderful individuals and friends during my graduate school years. I shared memorable moments with Farina Mirbagheri, Mehran Elyasi, Adel Elmahdy, Alireza Sharbafchi, Hamidreza Aliakbarikhouei, Seyed Amirhossein Hosseini, Ali Ghoreyshi, Rasoul Faraji, and Reza Zamani.

I want to express my deepest gratitude to the most important people in my life: my family. Their unwavering love, support, and encouragement have been the bedrock of everything I've accomplished. To my amazing siblings, Amorhossein and Shadi, thank you for being such cherished gifts. A heartfelt thanks goes especially to my incredible mother, Fatemeh Khabazian. Your constant support and belief in me have fueled my dreams and instilled a confidence that I carry with me every day. The sacrifices you've made have shaped me into the person I am today, and for that, I am eternally grateful.

Dedication

This dissertation is dedicated to my beloved mother.

Abstract

Distributed networks connect independent nodes to act as one powerful system. These spread-out machines share resources (data, storage, processing) to handle large tasks efficiently. Such networks have various applications in communication and learning systems including distributed caching and distributed machine learning. Distributed caching improves performance by storing frequently accessed data closer to users, reducing retrieval times and network load. Distributed machine learning leverages the combined computing power of multiple machines to train complex models on massive datasets, leading to faster and more scalable results.

However, this distributed nature introduces its own set of challenges, including (1) communication efficiency due to the exchange of large amounts of data and noisy communication links, (2) time-varying networks because of the dynamic nature of networks, and (3) privacy concerns due to the presence of eavesdropper nodes.

This dissertation dives into two crucial distributed systems: distributed caching and distributed machine learning. The goal is to make them more efficient, reliable, and adaptable by tackling key challenges. To achieve this, the research leverages advanced techniques from various fields, including distributed optimization, statistical learning theory, probability theory, and communication and coding theory.

In the first part of the thesis, we study a distributed caching setup in a fast-fading environment where each user has storage with a fraction of the transmitter's files. We focus on uncoded cache placement and the challenges of assigning signal levels without knowing channel conditions. We study the asymptotic behavior of the system and characterize the maximum achievable source rate for various scenarios, and finally provide an upper bound. We also propose a scheme using linear programming to show the characterization's looseness for the general case.

Moving on to the second part, we consider distributed machine learning paradigms, focusing on methods to mitigate communication costs, time-varying links, and privacy concerns. To address the first two challenges, we propose a two-time scale algorithm. In the proposed method, one time scale suppresses the imperfect incoming information from neighboring agents, while the other time scale operates on the gradients of local cost

functions. To resolve the privacy concern, we employ a differential privacy mechanism. We also support our theoretical results in both parts with significant improvements in numerical experiments.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Distributed systems represent a collection of autonomous computers geographically separated but connected through a network. These individual systems, often referred to as nodes, collaborate to achieve a unified objective. This collaboration involves the sharing of resources, data, and processes across the network, enabling the system to function as a single, coherent entity from the user's perspective. The core principle of distributed systems lies in the division of labor. Complex tasks are partitioned and distributed amongst the various nodes, allowing for parallel execution and enhanced efficiency compared to a single-machine approach. This paradigm fosters scalability, enabling the system to grow by adding more nodes to manage increasing demands. Distributed systems underpin a vast array of modern applications. *Distributed caching* systems improve performance by storing frequently accessed data closer to users, reducing network traffic and response times. *Distributed machine learning* algorithms leverage the computational power of multiple machines to train complex models on massive datasets, enabling faster and more accurate predictions. These are just a few examples of how distributed systems revolutionize the way we compute and interact with information.

1.1 Distributed Caching

The global video streaming market is booming, fueled by the COVID-19 pandemic, and projected to reach a staggering $330 billion by 2030 [1]. Distributed caching, a cornerstone of modern computing, plays a vital role in this growth by enhancing performance,

reducing latency, and managing data efficiently across vast, distributed networks.

With mobile data traffic dominated by video content, caching is emerging as a critical technology for next-generation wireless systems (Figure 1.1). Caching operates in two phases, namely, placement (prefetching) phase and delivery (fetching) phase. In the placement phase, users (or a central server) strategically duplicate packets of popular files in their local memory during off-peak hours. But, in wireless communication networks, the signal strength varies over time due to multiple factors. This implies that the transmitter requires some form of knowledge of the wireless channel conditions, often referred to as channel state information (CSI), at the transmitter. However, sending the CSI from the receivers to the transmitter over a feedback link is costly.

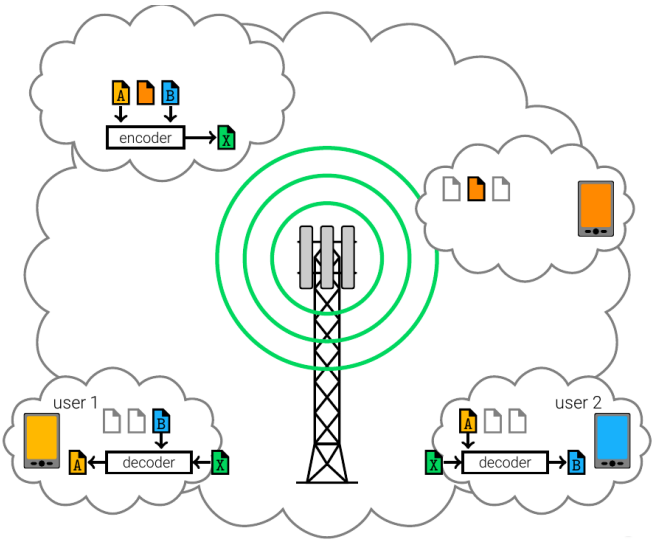

Figure 1.1: Mobile Edge Caching

In Chapter 2, we study a communication model over a wireless, time-varying, and fast-fading broadcast channel where each user is equipped with a cache. The transmitter has some source rate per channel for each file in its database. A fixed fraction of each file is available in the cache of each user where we focus on a class of pure cache placement schemes. After the completion of this phase, each user requests a file. Then, the transmitter forms broadcasting messages such that each user can decode his desired file. The main challenge for the transmitter is to assign the signal levels to the

(broadcasting) messages intended for each user without having access to the realization of the channels, which consists of the number of bits delivered to each user. Note that a fast-fading environment leads us to study the behavior of the channel where the transmitter has a certain source rate for generating bits of each file. We present the maximum achievable source rate for different regimes of cache size, by providing an explicit channel (level) allocation for each sub-message and proving a matching theoretical (upper) bound. Moreover, we investigate the maximum achievable source rate for a class of communication model for an arbitrary number of users. Then, we provide a theoretical analysis for the source rate. Finally, we develop an achievable scheme to show the looseness of the characterization for the general case.

1.2 Distributed Machine Learning

The multi-agent networked systems arise in various applications such as sensor networks, distributed network resource allocations, multi-robot control, computer games, computer vision, and especially large-scale machine learning and artificial intelligence, for which decentralized solutions offer promising results. Funding for artificial intelligence companies in the United States has increased exponentially in recent years, growing from a little under 300 million dollars in 2011 to around 16.5 billion dollars in 2019 [2]. Further, this project provides new techniques to train learning algorithms for medical and health databases. There is a shortage of labeled data available in the healthcare domain, and even if it is available, healthcare data is commonly distributed and needs to be aggregated at a centralized storage site so that deep learning models can be trained. However, most of the healthcare centers and laws at the country level, such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), are rightfully protective of the data and do not allow free sharing of data across computer networks and national boundaries. Distributed machine learning, with its unique privacy approach, can very effectively overcome the greatest obstacle facing AI adoption in health care today. We no longer need to choose between patient privacy and the utility of the data to society. We can now achieve privacy and utility simultaneously by exploiting distributed large-scale machine learning algorithms [3]. Federated and decentralized learning are two popular

paradigms that are widely used to perform learning tasks in a distributed fashion (Figure 1.2). In federated learning, the network consists of a central server that utilizes a number of worker nodes capable of performing computation tasks at the edge of the network. In practice, the presence of a server in the learning phase may be costly or even infeasible. Distributed optimization is a paradigm to mitigate this barrier when there is no central computing node in the network. In this setting, each computing node has its own estimate, which will be updated throughout the algorithm based on the local data points.

However, the distributed machine learning paradigm suffers from serious practical challenges:

(1) **Extensive Communication**: A distributed learning network is potentially comprised of millions of devices, e.g. smartphones, IoT devices. Training on such a massive network induces a dramatically heavy communication burden over the bandwidth-limited network.

(2) **Time-varying Links**: The devices experience unreliable data transmission due to infrequent interference and the unavailability of some participating devices during model updates. This may cause significant degradation of the learned model performance.

(3) **Privacy**: Even though decentralized algorithms keep sensitive data on individual devices, they can still leak information, making them vulnerable to both external eavesdropping and honest-but-curious attacks.

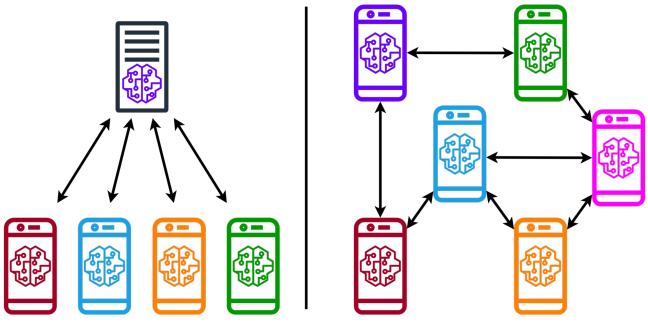

Figure 1.2: Federated and Decentralized Learning Paradigms

Several techniques have been developed to reduce the communication overhead while maintaining the accuracy and efficiency of the learning process:

- **Quantization.** One of the most widely used methods, quantization reduces the number of bits required to represent model updates or gradients by approximating their values. By using fewer bits per value, quantization significantly lowers the amount of data that needs to be transmitted, thus reducing communication costs. Different quantization schemes, such as fixed-point and stochastic quantization, allow for varying levels of precision depending on the needs of the application [4–6].

- **Sparsification.** Sparsification techniques focus on reducing the number of non-zero elements in the data being transmitted. By only sending the most significant components of the gradients or updates, sparsification drastically cuts down the volume of transmitted information. Methods like top-k sparsification, where only the top k largest components are sent, ensure that the most impactful data is communicated, preserving the learning performance while minimizing bandwidth usage [7–9].

- **Local Updates with Periodic Communication.** Instead of transmitting data after every iteration, nodes perform several local updates before communicating with the central server or neighboring nodes. This reduces the frequency of communication, thus lowering the overall communication cost. Techniques like local SGD or federated averaging leverage this approach to balance computation and communication effectively [10].

- **Adaptive Communication Strategies.** Adaptive methods dynamically adjust the communication parameters based on the current state of the learning process and network conditions. For instance, nodes might use more bits or communicate more frequently when the gradients are large or when communication links are stable, and reduce communication during periods of minor updates or noisy and time-varying links [11–13].

In Chapter 3, we introduce an adaptive quantization method for federated and decentralized learning to reduce communication costs. In both setups, we present adaptive

bit allocation schemes that allow nodes to optimize their bandwidth usage with minimal communication overhead. We demonstrate that these schemes improve the speed of convergence compared to uniform bit allocation methods, especially when data distribution among nodes is skewed. Extensive simulations on various datasets support our theoretical findings.

In Chapter 4, we explore decentralized learning over a time-varying network. Previous approaches to solving distributed optimization problems typically rely on single time-scale algorithms, where each agent conducts gradient descent with a diminishing or constant step-size based on the average estimate of the agents in the network. However, the exact exchange of information required to compute these averages can create significant communication overhead. We consider nodes only receive a *noisy* version of their neighbors' information to address the communication concern. We propose a two time-scale decentralized algorithm where one time-scale adjusts the imperfect incoming information, while the other applies to local cost function gradients. We characterize the convergence rates for strongly convex and non-convex loss functions with appropriate step-size choices. Finally, we identify sufficient conditions on the step-size sequences that ensure almost sure convergence of the agents' states to an optimal solution in the case of convex cost functions. The effectiveness of our algorithm is validated by the simulation results.

To address these privacy concerns, several techniques have been developed to protect the data and the learning process. These techniques aim to ensure that individual data points or sensitive information cannot be inferred by an adversary, even if they have access to the exchanged information. Key techniques include:

- **Differential Privacy.** Differential privacy is a robust mathematical framework that provides formal guarantees on the privacy of individual data points. In distributed machine learning, differential privacy is implemented by adding carefully calibrated noise to the model updates or gradients before they are shared. This noise masks the contribution of any single data point, ensuring that an adversary cannot infer specific information about individual datasets from the aggregated updates. Techniques like differential privacy with federated averaging or local differential privacy provide varying levels of privacy protection depending on the amount of noise added [14–16].

- **Secure Multi-Party Computation (SMPC).** SMPC allows multiple parties to jointly compute a function over their inputs while keeping those inputs private. In the context of distributed learning, SMPC techniques enable nodes to perform collaborative model training without revealing their local data to each other or to a central server. By using cryptographic protocols, SMPC ensures that the intermediate results and final model updates remain confidential, providing strong privacy guarantees [17–19].

- **Homomorphic Encryption.** Homomorphic encryption is a form of encryption that allows computations to be performed on encrypted data without decrypting it. This means that data can remain encrypted throughout the training process, and only the final model results are decrypted. In distributed learning, homomorphic encryption ensures that even if the data or gradients are intercepted during transmission, they cannot be understood without the decryption key, thus protecting the privacy of the data [20, 21].

In Chapter 5, we tackle a decentralized learning scenario to solve a convex optimization problem. We propose an accelerated distributed Nesterov gradient descent method that incorporates differential privacy by perturbing local variables with Laplace noise before sharing them with neighboring agents. This ensures differential privacy against eavesdropping and honest-but-curious attacks. Our chosen noise parameters guarantee convergence to the objective function's minimizer. We provide the convergence rate that outperforms existing decentralized learning methods with differential privacy under the same privacy budget. Simulation results confirm the effectiveness of our algorithm.

Part I

Algorithms for Distributed Caching

Chapter 2

Cache-Aided K-User Broadcast Channels with State Information at Receivers

We study a K-user coded-caching broadcast problem in a joint source-channel coding framework. The transmitter observes a database of files that are being generated at a certain rate per channel use, and each user has a cache, which can store a fixed fraction of the generated symbols. In the delivery phase, the transmitter broadcasts a message so that the users can decode their desired files using the received signal and their cache content. The communication between the transmitter and the receivers happens over a (deterministic) *time-varying* erasure broadcast channel, and the channel state information is only available to the users. We characterize the maximum achievable source rate for the 2-user and the degraded K-user problems. We provide an upper bound for any caching strategy's achievable source rates. Finally, we present a linear programming formulation to show that the upper bound is not a sharp characterization. Closing the gap between the achievable rate and the optimum rate remains open.

2.1 Introduction

The number of active users of video streaming applications such as Netflix, YouTube, HBO, etc., is growing rapidly. Coded caching is a promising strategy to overcome this rapidly growing traffic load of networks during their peak traffic time by duplicating parts of the content in the caches distributed across the network. A caching system operates in two phases: (i) a placement (pre-fetching) phase, where each user has access to the database of the transmitter and stores some packets from the database during the off-load time, and (ii) a delivery (fetching) phase, during which each user demands a file from the database, and the transmitter broadcasts a signal over a (noisy) channel to all the users (receivers), such that each user can decode his desired file from his cache content and his received signal. Moreover, in this phase, the network is congested, and the transmitter exploits the content of users to serve their requested files.

In practice, assuming a perfect broadcast channel fails, especially for the wireless communication setup. Therefore, we are dealing with a random time-varying channel between the transmitter and the users. In this chapter, to model the randomness of the channel, we consider a binary deterministic version of a time-varying memoryless fading broadcast channel when the transmitter is serving users. However, the pre-fetching phase takes over the noiseless links. We study such a caching problem in a joint source-channel coding framework and analyze the limitations of the source rate for the transmitter.

Related Works. Coded caching schemes are proposed under the perfect channel assumption for the delivery phase and the uncoded cache placement, where the placement performs on pure packets of the files for the centralized [22] and the decentralized settings [23]. It is shown that a significant gain can be achieved by sending coded packets and simultaneously serving multiple users. A distributed source coding problem is presented in [24] to study the cache-aided networks. The database is viewed as a discrete memoryless source and the users' requests as side information that is available everywhere except at the cache encoder. The inner and outer bounds on the fundamental trade-off of cache memory size and update rate are provided. For file selection networks with uniform requests, the derived bounds recover the rates established by [22,23]. The exact trade-off between the memory and load of delivery is characterized [25] for the

uncoded placement. The coded caching problem has also been studied in various setups, including online caching [26], device-to-device caching [27,28], caching with nonuniform demands [29,30], coded cache placement [31–34].

All of the aforementioned works assume that the delivery phase takes over a perfect channel. However, in practice, we are dealing with noisy broadcast channels, especially for wireless communication systems. For the wireless setup, various types of channel models have been studied, such as cache-aided interference channels [35–37], caching on broadcast channels [38,39], erasure and fading channels [40–42], and channels with delayed feedback with channel state information [43,44]. The cache-aided communications problem is modeled as a joint cache-channel coding problem in [38]. The delivery phase takes place over a memoryless erasure broadcast channel. It is shown that using unequal cache sizes and joint cache-channel coding improves system efficiency when the users experience different channel qualities. The capacity-memory trade-off of the K-user broadcast channel is studied when each user is equipped with a cache. It is optimal to assign all the cache memory to the weakest user for the small total cache size. On the other hand, for the large cache size, it is optimal to assign a positive portion of the cache to each user where weaker users have access to a larger cache memory than stronger users. Another wireless communication model is considered in [43] where a K-antenna transmitter communicates to K single receiver antenna. It is shown that the combination of caching with a rate-splitting broadcast approach can reduce the need for channel state information at the transmitter.

Note that in a fast-fading environment sending the channel state information (CSI) from the receiver to the transmitter over a feedback link is difficult. So, it is more reasonable to consider broadcast channels with no CSI. The ergodic capacity region of a K-user binary deterministic version of the time-varying memoryless fading broadcast channel (K-DTVBC) introduced by [45] is studied in [46,47]. Depending on the instantaneous channel strength, each user only receives the most significant bits of the transmit signal. Using the insight from the K-DTVBC model, an outer bound to the Gaussian fading BC capacity region is derived.

Contributions. In this work, we study a communication model over the K-DTVBC for n channel use where each user is equipped with a cache. The transmitter has some source rate per channel for each file in its database. A fixed fraction of each file is available

in each user's cache. Here, we focus on a class of uncoded cache placement schemes. After the completion of this phase, each user demands a file. Then, the transmitter forms broadcasting messages such that each user can decode his desired file. The main challenge for the transmitter is to assign the signal levels to the (broadcasting) messages intended for each user without having access to the realization of the channels, which consists of the number of bits delivered to each user. Note that a fast-fading environment needs coding for reliable communication, where the capacity of the channel is achievable using channel codes with sufficiently large block lengths. Thus, we study the asymptotic behavior of the system, where we allow the size of messages in the pre-fetching and fetching phases will grow with the communication block length. This leads us to deal with a joint source-channel coding problem where the transmitter has a certain source rate per channel use. We characterize the maximum achievable source rate for the two-user and the degraded K-user problems. Then, we provide an upper bound for the source rate. Finally, we discuss an achievable scheme with the linear programming (LP) formulation to show the looseness of the characterization for $K > 2$.

Notation. Throughout this chapter, we denote the set of integers $\{1, 2, \ldots, N\}$ by $[N]$ and the set of non-negative real numbers by $\mathbb{R}^+$. For a binary vector of length B, i.e., $X \in \mathbb{F}_2^B$, and a pair of integers $a < b$, we use the short hand notation $X(a:b)$ to denote $[X(a), X(a+1), \ldots, X(b)]$. We use $(a, b]$ to refer to the interval $(a, b] \coloneqq \{x \in \mathbb{R} : a < x \leq b\}$, and its scaled and shifted version is defined as $\alpha + \beta(a, b] \coloneqq (\alpha + \beta a, \alpha + \beta b]$. For a set of real numbers $\mathcal{I}$, we use $|\mathcal{I}|$ to denote its Lebesgue measure, e.g., $|(a, b]| \coloneqq b - a$ denotes the length of the interval. The all-ones and all-zeros vectors are defined as $\mathbf{1}_n \coloneqq (1, 1, \ldots, 1) \in \mathbb{R}^{n\times 1}$ and $\mathbf{0}_n \coloneqq (0, 0, \ldots, 0) \in \mathbb{R}^{n\times 1}$, respectively. For a real number $x \in \mathbb{R}$, we denote its floor and ceiling by $\lfloor x \rfloor$ and $\lceil x \rceil$, respectively. The fractional part of x is denote by $\{x\} \coloneqq x - \lfloor x \rfloor$. Finally, for $n, k \in \mathbb{Z}$, the binomial coefficient is defined as $\binom{n}{k} \coloneqq \frac{n!}{k!(n-k)!}$, if $0 \leq k \leq n$, and $\binom{n}{k} \coloneqq 0$, otherwise.

2.2 Problem Formulation

In this section, we first introduce the K-DTVBC, which is the core of this work. Then, we discuss the joint source-channel coding problem studied in this chapter.

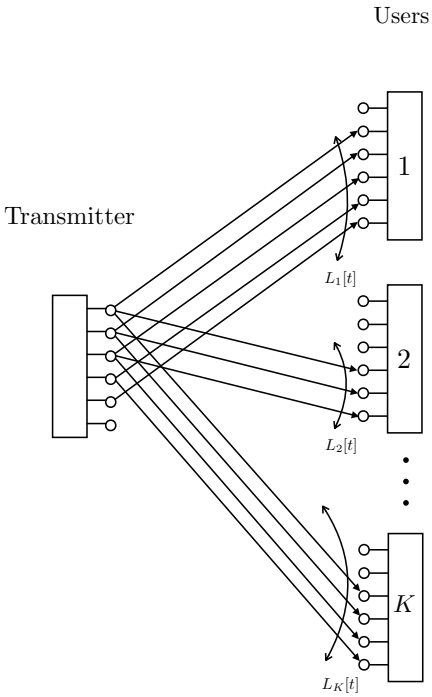

Figure 2.1: A K-user binary deterministic version of the time-varying memoryless fading broadcast channel. The transmitter only knows the statistics, but not the realizations of the generated i.i.d. random sequence $\{L_k[t]\}_{t=1}^{n}$.

2.2.1 Channel Model

We are interested in a time-varying broadcast channel, where a transmitter aims at sending one message to each of the K users. We consider the K-DTVBC introduced by [45] as shown in Figure 2.1. The channel is modeled by

$$Y_{k,t} = D^{B-L_k[t]}X_t = X_t(1 : L_k[t]), \qquad k \in [K], \tag{2.1}$$

where $X_t, Y_{k,t} \in \mathbb{F}_2^B$ for $k \in [K]$, and D is a $B \times B$ shift matrix, given by

$$D = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}.$$

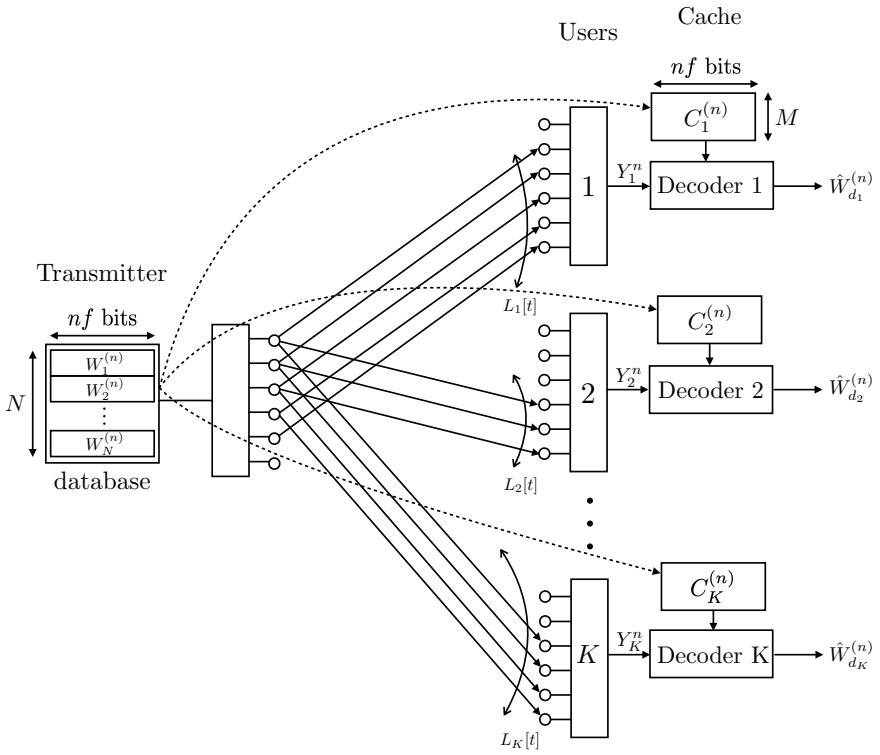

Figure 2.2: A transmitter containing N files of size nf bits each is connected through a K-DTVBC to users each with a cache of size nMf bits.

Here $L_k[t]$ with $0 \le L_k[t] \le B$ determines the number of bits delivered to user k at time t. The channel state at user k, i.e., $\{L_k[t] : t = 1, \ldots, n\}$, is an i.i.d. random sequence generated according to some probability mass function (PMF) $P_{L_k}(\ell) \coloneqq \mathbb{P}[L_k = \ell]$. Intuitively, sending a message X_t of length B bits over a channel with parameters $L_k[t]$, the receiver only receives the $L_k[t]$ most significant bits (MSBs) of X_t, and the remaining bits will be erased. This operation can be modeled as the multiplication of the message X_t by $D^{B-L_k[t]}$, where D is the shift matrix. We assume that the channel state information is casually known only to the receivers. However, the transmitter only knows the channel statistics $P_{L_k}(\ell)$, but not the channel realizations.

Definition 2.1. We denote the complementary cumulative distribution function (CCDF) of L_k for any $\ell \in [B]$ by

$$\overline{F}_{L_k}(\ell) \coloneqq \mathbb{P}[L_k \ge \ell].$$

For notational simplicity, let

$$\overline{\boldsymbol{F}}_{L_k} \coloneqq \begin{bmatrix} \overline{F}_{L_k}(1) \\ \vdots \\ \overline{F}_{L_k}(B) \end{bmatrix},$$

for each user $k \in [K]$.

Definition 2.2. The random variable L_k is *stochastically* larger than L_v, if $\overline{F}_{L_k}(\ell) \geq \overline{F}_{L_v}(\ell)$ for every $\ell \in [B]$ and we denote it by $L_k \geq_{\mathsf{st}} L_v$.

The K-DTVBC channel model for wireless communication simplifies analysis compared to the Gaussian model while still capturing the important features of the problem. This model focuses on signal interactions rather than background noise since networks often operate in interference-limited scenarios. The deterministic model operates on a finite-field, makes it simpler, and provides a complete characterization of network capacity. The insights gained from the deterministic analysis can be applied to find approximately optimal communication schemes for Gaussian relay networks. The analysis of deterministic networks not only guides coding schemes for Gaussian channels but also offers useful proof techniques. The capacity region of the K-DTVBC is derived in [47]. In this work, we focus on a cache-aided version of this problem, where the users are equipped with a cache that can pre-fetch part of the messages. In contrast, to [47], where the capacity region is characterized, we are interested in the symmetric rate, as it is standard to consider equal file sizes in file delivery systems.

In the majority of the existing literature on coded caching, a perfect channel is assumed between the transmitter and the users. Hence, the focus is on minimizing the design of the placement and delivery phases to minimize the load on the perfect channel [22,23,25,27,28]. Here, we are dealing with a fast fading channel which requires coding for reliable communication. Hence, we allow for a large code length and study the asymptotic behavior of the channel. Consequently, the size of the message(s) will grow with the communication block length. This leads to a joint source-channel coding problem [48–50]. More precisely, we consider a communication scenario over n *channel uses*, where the transmitter has a library of N files, each of size nf bits. Each user is equipped with a cache that can pre-fetch up to nMf bits (before the actual request of the user is revealed), and the goal is to send one requested file to each user reliably. We

are interested in characterizing the maximum source rate f for which, and for sufficiently large block length n, a reliable communication scheme can be devised. A similar joint source-channel coding approach is used to study the original coded caching problem with common rate and side information in [24]. Further details of the cache model are discussed in the next section.

2.2.2 Joint Source-channel Coding Framework

Let us consider a communication scenario over the K-DTVBC for n channel uses. The transmitter has some source rate $f \in \mathbb{R}^+$ per channel use that generates N files, namely, $W_i^{(n)}$ for $i \in [N]$. This means the transmitter has access to a database of N mutually independent files $W_1^{(n)}, \ldots, W_N^{(n)}$ each of size nf bits, i.e.,

$$W_i^{(n)} \in \{1, 2, \ldots, 2^{nf}\}, \quad i \in [N].$$

We assume each user k is equipped with a cache, which can pre-fetch part of the files. The size of the content is proportional to the communication block length. More precisely, we assume that user k has a cache $C_k^{(n)}$ of size nMf bits, for $k \in [K]$. In the placement phase, the cache memory of each user is filled with *uncoded* bits of the files; that is, the content of the cache $C_k^{(n)}$ can be partitioned into raw (uncoded) bits of the files.

Definition 2.3. A *caching strategy* $\mathfrak{C}$ for a normalized cache size $\mu = M/N$ and a network with K users consists of K collections of intervals in $(0, 1]$. More precisely, $\mathfrak{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K)$ where

- $\mathbf{c}_k = \bigcup_{\ell \in [N_k]} \mathcal{I}_{k,\ell}$,
- N_k is a finite positive integer number for every $k \in [K]$,
- $\mathcal{I}_{k,\ell} = (a_{k,\ell}, b_{k,\ell}] \subseteq (0, 1]$ where $b_{k,\ell} \leq a_{k,\ell+1}$ for every $\ell \in [K-1]$, and
- $\sum_{\ell \in [N_k]} |\mathcal{I}_{k,\ell}| = \mu$, for every $k \in [K]$.

For a file $W = (W(1), W(2), \ldots, W(F)) \in \mathbb{F}_2^F$ of length F bits, we define

$$W(\mathbf{c}_k) := \bigcup_{\ell \in [N_k]} \{W(\lceil a_{k,\ell} F \rceil + 1), \cdots, W(\lfloor b_{k,\ell} F \rfloor)\}.$$

For a given source rate f, block length n, family of files $\left\{W_i^{(n)}\right\}$, and caching placement strategy $\mathfrak{C}$, the cache content of user $k \in [K]$ is given by

$$\begin{aligned} C_k^{(n)} &:= \bigcup_{i\in[N]} C_{k,i}^{(n)} = \bigcup_{i\in[N]} W_i^{(n)}(\mathbf{c}_k) \\ &= \bigcup_{i\in[N]} \bigcup_{\ell\in[N_k]} \left\{W_i^{(n)}(\lceil nfa_{k,\ell}\rceil+1), \ldots, W_i^{(n)}(\lfloor nfb_{k,\ell}\rfloor)\right\}. \end{aligned} \tag{2.2}$$

This implies that

$$\begin{aligned} H\left(C_{k,i}^{(n)}\right) &\le \sum_{\ell\in[N_k]} \left(\lfloor nfb_{k,\ell}\rfloor - \lceil nfa_{k,\ell}\rceil\right) \\ &\le \sum_{\ell\in[N_k]} nf|\mathcal{I}_{k,\ell}| = \mu nf. \end{aligned} \tag{2.3}$$

Therefore, we get

$$\begin{aligned} H\left(C_k^{(n)}\right) &= H\left(C_{k,1}^{(n)}, C_{k,2}^{(n)}, \ldots, C_{k,N}^{(n)}\right) \\ &\le H\left(C_{k,1}^{(n)}\right) + H\left(C_{k,2}^{(n)}\right) + \cdots + H\left(C_{k,N}^{(n)}\right) \\ &\le \sum_{i=1}^{N} n\mu f = nMf. \end{aligned}$$

Moreover, from the definition of the cache content in (2.2) and the independence of files, we can write

$$\begin{aligned} &H\left(C_{k,i}^{(n)} \middle| W_i^{(n)}\right) = 0, \\ &I\left(C_{k,j}^{(n)}; W_i^{(n)}\right) = 0, \quad j \neq i. \end{aligned}$$

We define $\mathbf{c}_{\mathcal{S}} := \bigcup_{u\in\mathcal{S}} \mathbf{c}_u = \bigcup_{u\in\mathcal{S}} \bigcup_{\ell\in[N_u]} \mathcal{I}_{u,\ell}$ for every $\mathcal{S} \subseteq [K]$ and the *caching tuple* $\boldsymbol{\mu} := (\mu_{\mathcal{S}} : \mathcal{S} \subseteq [K])$ where $\mu_{\mathcal{S}} := |\mathbf{c}_{\mathcal{S}}|$. We also use $C_{\mathcal{S},i}^{(n)}$ to refer to the collection of all the parts of file i cached by the users in the subset $\mathcal{S} \subseteq [K]$, i.e., $C_{\mathcal{S},i}^{(n)} = \bigcup_{u\in\mathcal{S}} C_{u,i}^{(n)}$. Therefore, we have

$$H\left(C_{\mathcal{S},i}^{(n)}\right) \le \mu_{\mathcal{S}} nf, \tag{2.4}$$

for every $i \in [N]$. After the completion of the placement phase, each user requests one of the N files, where all files are equally likely to be requested. We denote $d_k \in [N]$ as the index of the file requested by user $k \in [K]$ and the sequence of all requests

by $\boldsymbol{d} = (d_1, \ldots, d_K)$. Once the requests are revealed to the transmitter, it forms a broadcasting message $X^n = (X_1, X_2, \ldots, X_n) = \psi_{\boldsymbol{d}}^{(n)}\left(W_1^{(n)}, \ldots, W_N^{(n)}; C_{[K]}^{(n)}\right)$, where

$$\psi_{\boldsymbol{d}}^{(n)}: \{1, 2, \ldots, 2^{nf}\}^N \times \{1, 2, \ldots, 2^{nMf}\}^K \to \{1, 2, \ldots, 2^B\}^n,$$

and transmits X_t over the broadcast channel during the tth channel use of the delivery phase, for $t = 1, \ldots, n$. Upon receiving $Y_k^n = (Y_{k,1}, Y_{k,2}, \ldots, Y_{k,n})$, user $k \in [K]$ should be able to decode its desired file using its cache content $C_k^{(n)}$ and the received message Y_k^n (see Figure 4.5), i.e.,

$$\hat{W}_{d_k}^{(n)} = \phi_k^{(n)}\left(Y_k^n, C_k^{(n)}\right).$$

Here, we define the overall decoding error probability as $P_e^{(n)} := \sum_{k=1}^K \mathbb{P}\left[\hat{W}_{d_k}^{(n)} \neq W_{d_k}^{(n)}\right]$.

Definition 2.4. For a given caching strategy $\mathfrak{C}$ and a (distinct) request profile $\boldsymbol{d}$, a source rate $f(\mathfrak{C}, \boldsymbol{d})$ is called achievable if there exists a sequence of encoding and decoding functions $\left\{\left(\psi^{(n)}, \phi_1^{(n)}, \ldots, \phi_K^{(n)}\right)\right\}_n$, for which $P_e^{(n)} \to 0$ as n grows.

Here, our goal is to characterize the maximum achievable source rate $f(\mathfrak{C}, \boldsymbol{d})$ for a given K-DTVBC with channel statistics, $\overline{\boldsymbol{F}}_{L_k}$ for $k \in [K]$. Note that the cache placement is fixed prior to the users' demands, and we are not designing the cache contents of users based on the requested files.

For every subset of users $\mathcal{S} \subseteq [K]$ and file index $i \in [N]$, we define

$$W_{i,\mathcal{S}}^{(n)} = \bigcap_{k \in \mathcal{S}} W_i^{(n)}(\mathbf{c}_k) = W_i^{(n)}\left(\bigcap_{k \in \mathcal{S}} \mathbf{c}_k\right),$$

to be the sections of file $W_i^{(n)}$ which are cached at all users in $\mathcal{S}$.

Next, inspired by the central cache placement strategy of [22], we introduce the *central caching strategy* $\mathfrak{C}^{\mathsf{cent}}$. For a subset $\mathcal{S} \subseteq [K]$ with $|\mathcal{S}| = s$, let $\chi(\mathcal{S}) \in \left\{1, 2, \ldots, \binom{K}{s}\right\}$ be the rank of $\mathcal{S}$ among all subsets of $[K]$ of size s, according to the *lexicographical order*.

Definition 2.5. For every $\mathcal{S} \subseteq [K]$ with $|\mathcal{S}| = s$, define

$$\mathcal{J}_{\mathcal{S}} := \left(\frac{\chi(\mathcal{S})}{\binom{K}{s}}, \frac{\chi(\mathcal{S}) + 1}{\binom{K}{s}}\right].$$

Then, for a network with K users and normalized cache size $\mu \in [0,1]$, we define the central caching strategy $\mathfrak{C}^{\mathsf{cent}} := (\mathbf{c}_1^{\mathsf{cent}}, \ldots, \mathbf{c}_K^{\mathsf{cent}})$ where

$$\mathbf{c}_k := \left(\bigcup_{\substack{\mathcal{S}\subseteq[K] \\ |\mathcal{S}|=\lfloor \mu K \rfloor \\ \mathcal{S}\ni k}} (1-\lambda)\mathcal{J}_{\mathcal{S}} \right) \cup \left(\bigcup_{\substack{\mathcal{T}\subseteq[K] \\ |\mathcal{T}|=\lfloor \mu K \rfloor+1 \\ \mathcal{T}\ni k}} ((1-\lambda)+\lambda\mathcal{J}_{\mathcal{T}}) \right),$$

and[1] $\lambda = \{\mu K\}$.

Note that for any set of users $\mathcal{Q} \subseteq [K]$, we have

$$\begin{aligned}
\mu_{\mathcal{Q}}^{\mathsf{cent}} &= \left|\mathbf{c}_{\mathcal{Q}}^{\mathsf{cent}}\right| \\
&= \left| \bigcup_{k\in\mathcal{Q}} \mathbf{c}_k^{\mathsf{cent}} \right| \\
&= (1-\lambda)\left(1 - \frac{1}{\binom{K}{\lfloor \mu K \rfloor}} \sum_{\substack{\mathcal{S}\subseteq[K] \\ |\mathcal{S}|=\lfloor \mu K \rfloor \\ \mathcal{S}\cap\mathcal{Q}=\varnothing}} 1 \right) \\
&\quad + \lambda\left(1 - \frac{1}{\binom{K}{\lfloor \mu K \rfloor+1}} \sum_{\substack{\mathcal{T}\subseteq[K] \\ |\mathcal{T}|=\lfloor \mu K \rfloor+1 \\ \mathcal{T}\cap\mathcal{Q}=\varnothing}} 1 \right) \\
&= (1-\lambda)\left(1 - \frac{\binom{K-|\mathcal{Q}|}{\lfloor \mu K \rfloor}}{\binom{K}{\lfloor \mu K \rfloor}} \right) + \lambda\left(1 - \frac{\binom{K-|\mathcal{Q}|}{\lfloor \mu K \rfloor+1}}{\binom{K}{\lfloor \mu K \rfloor+1}} \right). \qquad (2.5)
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\left| \bigcap_{k\in\mathcal{Q}} \mathbf{c}_k^{\mathsf{cent}} \right| &= (1-\lambda)\frac{1}{\binom{K}{\lfloor \mu K \rfloor}} \sum_{\substack{\mathcal{S}\subseteq[K] \\ |\mathcal{S}|=\lfloor \mu K \rfloor \\ \mathcal{Q}\subseteq\mathcal{S}}} 1 + \lambda\frac{1}{\binom{K}{\lfloor \mu K \rfloor+1}} \sum_{\substack{\mathcal{T}\subseteq[K] \\ |\mathcal{T}|=\lfloor \mu K \rfloor+1 \\ \mathcal{Q}\subseteq\mathcal{T}}} 1 \\
&= (1-\lambda)\frac{\binom{K-|\mathcal{Q}|}{\lfloor \mu K \rfloor-|\mathcal{Q}|}}{\binom{K}{\lfloor \mu K \rfloor}} + \lambda\frac{\binom{K-|\mathcal{Q}|}{\lfloor \mu K \rfloor+1-|\mathcal{Q}|}}{\binom{K}{\lfloor \mu K \rfloor+1}}. \qquad (2.6)
\end{aligned}$$

[1] Note that $\mu K = (1-\lambda)\lfloor \mu K \rfloor + \lambda\lceil \mu K \rceil$

2.3 Main Results

In this section, we present the main results of this chapter, organized according to the level of generalization of the setting.

We first characterize the maximum achievable source rate for the 2-user DTVBC.

Theorem 2.1 (Two-User (Non-Degraded) BC)**.** *For a 2-DTVBC with a caching strategy $\mathfrak{C}$, a distinct request profile $\boldsymbol{d}$, and $\mu \leq \frac{1}{2}$, any achievable source rate is upper bounded by*

$$f^\star = \min\left\{\min_{\omega\geq 1}\frac{\omega R_1(\omega)+R_2(\omega)}{\omega(1-\mu)+(1-2\mu)}, \min_{0\leq\omega\leq 1}\frac{R_1(\omega)+\frac{1}{\omega}R_2(\omega)}{(1-2\mu)+\frac{1}{\omega}(1-\mu)}\right\}. \tag{2.7}$$

Moreover, if $\frac{1}{2} \leq \mu \leq 1$, any achievable source rate is upper bounded by

$$f^\star = \min\left\{\frac{\sum_{\ell=1}^{B}\overline{F}_{L_1}(\ell)}{1-\mu}, \frac{\sum_{\ell=1}^{B}\overline{F}_{L_2}(\ell)}{1-\mu}\right\}, \tag{2.8}$$

where

$$\begin{aligned} R_1(\omega) &:= \sum_{\ell\in\mathcal{L}_1(\omega)}\overline{F}_{L_1}(\ell), \\ R_2(\omega) &:= \sum_{\ell\in\mathcal{L}_2(\omega)}\overline{F}_{L_2}(\ell), \end{aligned} \tag{2.9}$$

$\mathcal{L}_1(\omega) := \{\ell : \omega\overline{F}_{L_1}(\ell) \geq \overline{F}_{L_2}(\ell)\}$, and $\mathcal{L}_2(\omega) := \{\ell : \omega\overline{F}_{L_1}(\ell) < \overline{F}_{L_2}(\ell)\}$. Moreover, the source rates in (2.7) and (2.8) are achievable for the central caching strategy $\mathfrak{C}^{\mathsf{cent}}$.

The proof of Theorem 2.1 is provided in Section 2.5.

In the upper bound presented in Theorem 2.1, we intuitively enhance both the cache and channel strengths for each user and compare the two possible settings.

Now, let us consider a more general setting where a network is serving K users. In a K user setting, we can characterize the maximum achievable source rate when the channels from the transmitter to the users are degraded (see Definition 2.2). In the following theorem, we provide an LP optimization problem for the maximum achievable source rate of the degraded K-DTVBC with the caching strategy $\mathfrak{C}^{\mathsf{cent}}$.

Theorem 2.2 (K-User Degraded BC)**.** *For a degraded K-DTVBC $L_K \geq_{\mathsf{st}} \cdots \geq_{\mathsf{st}} L_1$ and a normalized cache sizes μ satisfying $K\mu \in \mathbb{N}$, with the central caching strategy $\mathfrak{C}^{\mathsf{cent}}$ and a distinct request profile $\boldsymbol{d}$, the maximum achievable source rate is given by*

$$\max_{\{z_{\ell,k}\}} \ \bar{f}, \tag{2.10}$$

$$s.t. \quad \left(1-\mu_{[k]}^{\mathsf{cent}}\right)\bar{f} \leq \sum_{\ell=1}^{B} z_{\ell,k}\overline{F}_{L_k}(\ell), \quad \forall k \in [K],$$

$$z_{\ell,k} \geq 0, \quad \forall k \in [K], \forall \ell \in [B], \tag{2.11}$$

$$\sum_{k=1}^{K} z_{\ell,k} \leq 1, \quad \forall \ell \in [B].$$

The proof of Theorem 2.2 is presented in two parts. The proof of achievability part is presented in Section 2.6, and its converse proof is provided in Section 2.8. We note that the achievability proof of Theorem 2.2 is based on the LP-based method which is discussed in Section 2.4. Next, we will now present an illustrative example of the degraded K-DTVBC with three users ($k = 3$). This example will be helpful in establishing the notation and following the proof techniques.

$$\mathbf{A}_\pi = \left[\begin{array}{ccccc|c} \overline{\boldsymbol{F}}_{L_{\pi(1)}} & \mathbf{0}_B & \cdots & \mathbf{0}_B & \mathbf{0}_B & \left(\mu_{\pi(1)}-1\right)\boldsymbol{I} \\ \mathbf{0}_B & \overline{\boldsymbol{F}}_{L_{\pi(2)}} & \cdots & \mathbf{0}_B & \mathbf{0}_B & \left(\mu_{\pi([2])}-1\right)\boldsymbol{I} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_B & \mathbf{0}_B & \cdots & \mathbf{0}_B & \overline{\boldsymbol{F}}_{L_{\pi(K)}} & \left(\mu_{\pi([K])}-1\right)\boldsymbol{I} \\ \hline \mu_{\pi([2])}-1 & 1-\mu_{\pi(1)} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mu_{\pi([K])}-1 & 1-\mu_{\pi([K-1])} & 0 \end{array}\right], \tag{2.15}$$

Example 2.1 (*Degraded Channel Case*)**.** We consider a network with $K = 3$ users $N = 3$ files, namely $\{W_1, W_2, W_3\}$, and $B = 3$ signal levels. The channel statistics of the three users are given by the cumulative distribution functions as

$$\overline{F}_{L_1} = [0.5, 0.4, 0.3]^T,$$
$$\overline{F}_{L_2} = [0.7, 0.5, 0.4]^T,$$
$$\overline{F}_{L_3} = [0.9, 0.6, 0.5]^T.$$

That is, the first user receives the top level with probability 0.5, but the bits sent over all three $B = 3$ levels are delivered at this same user with probability 0.3. Considering the caching strategy $\mathfrak{C}^{\mathsf{cent}}$ with a caching factor of $\mu = 1/3$, it can be defined as follows

$$\mathbf{c}_1 = (0, 1/3], \quad \mathbf{c}_2 = (1/3, 2/3], \quad \mathbf{c}_3 = (2/3, 1].$$

Hence, the placement strategy $\mathfrak{C}^{\mathsf{cent}}$ implies that the cached parts of the files at different users are *disjoint*, i.e., $W_i^{(n)} = \bigcup_{k=1}^{3} W_{i,k}^{(n)}$ for every $i \in [N]$ where $W_{i,k}^{(n)}$ the part of file $W_i^{(n)}$ cached *exactly* by user k.

Without loss of generality, assume user k is interested in file $W_k^{(n)}$, for $k \in \{1, 2, 3\}$. Here, we have $\mu_{\{1\}}^{\mathsf{cent}} = 1/3$, $\mu_{\{1,2\}}^{\mathsf{cent}} = 2/3$, and $\mu_{\{1,2,3\}}^{\mathsf{cent}} = 1$. The coefficients $[\mathbf{z}]_{(\ell,k)} := z_{\ell,k}$ that provide the optimum solution of (2.10) are give by

$$\begin{matrix} (1,1) & (1,2) & (1,3) & (2,1) & (2,2) & (2,3) & (3,1) & (3,2) & (3,3) \end{matrix}$$
$$\mathbf{z} = \begin{bmatrix} 0.37 & 0.63 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}^T,$$

with the optimum source rate

$$f^\star = \frac{\sum_{\ell=1}^{3} z_{\ell,k} \overline{F}_{L_1}(\ell)}{1 - \mu_{\{1\}}^{\mathsf{cent}}} = \frac{\sum_{\ell=1}^{3} z_{\ell,k} \overline{F}_{L_2}(\ell)}{1 - \mu_{\{1,2\}}^{\mathsf{cent}}} = 1.326.$$

Finally, we can present our result for the most general case, which is the (non-degraded) K-DTVBC. The following theorem provides an upper bound for the source rate of the K-DTVBC with *any* given caching strategy $\mathfrak{C}$.

Theorem 2.3 (K-User (Non-Degraded) BC)**.** *Any achievable source rate of the K-DTVBC for a given cache placement strategy $\mathfrak{C}$ and a (distinct) request profile $\boldsymbol{d}$ is upper-bounded by*

$$f(\mathfrak{C}, \boldsymbol{d}) \le f^\star(\mathfrak{C}, \boldsymbol{d}) := \min_{\boldsymbol{\omega} \ge 0} \frac{\sum_{\ell=1}^{B} \max_k \omega_{\pi(k)} \overline{F}_{L_{\pi(k)}}(\ell)}{\sum_{k=1}^{K} \omega_{\pi(k)} \left(1 - \mu_{\pi([k])}\right)}, \tag{2.12}$$

where $\boldsymbol{\omega} := (\omega_1, \ldots, \omega_K) \in [0, \infty)^K$, $\pi([k]) := \{\pi(1), \ldots, \pi(k)\}$, and $\pi : [K] \to [K]$ is the permutation that sorts $\boldsymbol{\omega}$ in the non-increasing order.

The proof of Theorem 2.3 is presented in Section 2.7.

Note that the upper bound in the theorem is given as a min-max problem, for which the evaluation of the optimum point can be computationally challenging. In the

following proposition, we show that the upper bound in (2.12) can be evaluated by solving $K!$ linear programming problems, each corresponding to a permutation of users. Hence, denoting t_{LP} as the run time for each LP, the complexity of evaluation of the upper-bound in (2.12) is $K! \times t_{\mathsf{LP}}$.

Proposition 2.1. The min-max problem in (2.12) is equivalent to

$$f^\star(\mathfrak{C}, \boldsymbol{d}) = \min_{\pi \in \Pi} f_\pi^\star(\mathfrak{C}, \boldsymbol{d}), \tag{2.13}$$

where

$$\begin{aligned} f_\pi^\star(\mathfrak{C}, \boldsymbol{d}) \coloneqq \min_{\mathbf{x} \in \mathbb{R}^{K+B}} \ & [\mathbf{0}_K^T, \mathbf{1}_B^T]\mathbf{x}, \\ \text{s.t.} \quad & \mathbf{A}_\pi \mathbf{x} \leq \mathbf{0}, \\ & \mathbf{b}\mathbf{x} = 1, \\ & -\mathbf{x} \leq \mathbf{0}, \end{aligned} \tag{2.14}$$

and Π is the set of all permutations over $[K]$. The matrix $\mathbf{A}_\pi \in \mathbb{R}^{(KB+K-1)\times(K+B)}$ is given in (2.15) at the top of this page. Moreover,

$$\mathbf{b} = [\mathbf{1}_K^T, \mathbf{0}_B^T] \in \mathbb{R}^{1\times(K+B)}, \tag{2.16}$$

and $\boldsymbol{I} \in \mathbb{R}^{B\times B}$ is the identity matrix. Note that for each permutation $\pi \in \Pi$, the LP problem in (2.14) consists of $K+B$ variables and $K(B+2)+B$ constraints.

The proof of Proposition 2.1 is provided in Section 2.7.

2.4 An Achievable Scheme: LP Formulation

In this section, we provide an achievable scheme, which is based on Linear programming. This scheme is optimum for the degraded broadcast channels (as claimed in Theorem 2.3). However, in an illustrative example, we show that there is a gap between the achievable rate of the proposed scheme and the upper bound in (2.12). This implies that either the achievable scheme is not optimum, or the upper bound is not tight. Consequently, the exact characterization of the optimum source rate of a non-degraded K-DTVBC with $K > 2$ remains as an open problem for future works.

Similar to [22], we focus on specific normalized cache sizes μ such that $t := K\mu = KM/N \in \mathbb{N}$. Let us assume that each user $k \in [K]$ requests file $W_k^{(n)}$. The delivery scheme of [22] consists of broadcasting coded packets to serve multiple users simultaneously. Each coded packet is intended for a group of users $\mathcal{S} \subseteq \{1,\dots,K\}$ with $|\mathcal{S}| = t+1$. More precisely, we have

$$W_{\mathcal{S}}^{(n)} = \bigoplus_{k\in\mathcal{S}} W_{d_k,\mathcal{S}\setminus\{k\}}^{(n)}.$$

We aim at sending each coded packet $W_{\mathcal{S}}^{(n)}$ to all users $k \in \mathcal{S}$. To this end, we devise a bit allocation strategy $\mathbf{y} = \{y_{\ell,\mathcal{S}} : \ell \in [B], \mathcal{S} \subseteq [K], |\mathcal{S}| = t+1\}$, where $0 \le y_{\ell,\mathcal{S}} \le 1$ is a variable indicating the fraction of time that level ℓ of the channel is used to transmit coded message $W_{\mathcal{S}}^{(n)}$. Note that a signal level ℓ can be shared between *multiple* coded message $W_{\mathcal{S}}^{(n)}$. For a feasible allocation policy $\mathbf{y}$, in each level, the sum of time fractions allocated to all coded messages should not exceed 1, that is,

$$\sum_{\substack{\mathcal{S}\subseteq[K]\\|\mathcal{S}|=t+1}} y_{\ell,\mathcal{S}} \le 1, \qquad \forall \ell \in [B]. \tag{2.17}$$

To ensure successful decoding of the sub-message $W_{\mathcal{S}}^{(n)}$ by every user in $k \in \mathcal{S}$, it is necessary to assign sufficiently large values to the coefficients $y_{\ell,\mathcal{S}}$. Recall that for a given common source rate f, the rate of $W_{\mathcal{S}}^{(n)}$ is given by $f/\binom{K}{t}$. Then, a coded message $W_{\mathcal{S}}^{(n)}$ is decodable at user $k \in \mathcal{S}$ if

$$\sum_{\ell=1}^{B} \overline{F}_{L_k}(\ell) y_{\ell,\mathcal{S}} \ge \frac{f}{\binom{K}{t}}. \tag{2.18}$$

$$\mathbf{G} = \begin{array}{c} \\ (1,\{2\}) \\ (2,\{1\}) \\ (1,\{3\}) \\ (3,\{1\}) \\ (2,\{3\}) \\ (3,\{2\}) \end{array} \begin{array}{c} \begin{array}{ccccccccccc} (1,\{1,2\}) & (1,\{1,3\}) & (1,\{2,3\}) & (2,\{1,2\}) & (2,\{1,3\}) & (2,\{2,3\}) & (3,\{1,2\}) & (3,\{1,3\}) & (3,\{2,3\}) & m \end{array} \\ \begin{bmatrix} -0.9 & 0 & 0 & -0.3 & 0 & 0 & -0.3 & 0 & 0 & 0.33 \\ -0.7 & 0 & 0 & -0.4 & 0 & 0 & -0.4 & 0 & 0 & 0.33 \\ 0 & -0.9 & 0 & 0 & -0.3 & 0 & 0 & -0.3 & 0 & 0.33 \\ 0 & -0.5 & 0 & 0 & -0.5 & 0 & 0 & -0.5 & 0 & 0.33 \\ 0 & 0 & -0.7 & 0 & 0 & -0.4 & 0 & 0 & -0.4 & 0.33 \\ 0 & 0 & -0.5 & 0 & 0 & -0.5 & 0 & 0 & -0.5 & 0.33 \end{bmatrix} \end{array}, \tag{2.22}$$

$$\mathbf{H} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccccccccc} (1,\{1,2\}) & (1,\{1,3\}) & (1,\{2,3\}) & (2,\{1,2\}) & (2,\{1,3\}) & (2,\{2,3\}) & (3,\{1,2\}) & (3,\{1,3\}) & (3,\{2,3\}) & m \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}. \tag{2.23}$$

$$\mathbf{y} = \begin{array}{c} \begin{array}{cccccccccc} (1,\{1,2\}) & (1,\{1,3\}) & (1,\{2,3\}) & (2,\{1,2\}) & (2,\{1,3\}) & (2,\{2,3\}) & (3,\{1,2\}) & (3,\{1,3\}) & (3,\{2,3\}) & m \end{array} \\ \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 & \frac{1}{12} & 0 & \frac{11}{12} & 0 & \frac{2}{3} & \frac{1}{3} & \frac{3}{2} \end{bmatrix}^T \end{array}. \tag{2.24}$$

So, we have an optimization problem given by

$$\begin{aligned} f_{\mathsf{LP}} \coloneqq \max \ & f \\ \text{s.t. } & (2.17) - (2.18) \\ & y_{\ell,\mathcal{S}} \geq 0, \quad \forall \ell \in [B], \forall \mathcal{S} \subseteq [K], |\mathcal{S}| = t+1. \end{aligned} \tag{2.19}$$

We can write the optimization problem as a linear program. Thus, we define a vector $\mathbf{y}$ of length $m = B\binom{K}{t+1}+1$, forming by stacking all variables in $\{y_{\ell,\mathcal{S}} : \ell \in [B], \mathcal{S} \subseteq [K], |\mathcal{S}| = t+1\}$ along with f at the very last position. The first $B\binom{K}{t+1}$ entries of $\mathbf{y}$ are labeled by pairs $(\ell, \mathcal{S})$ and we set $\mathbf{y}_m = -f$ as the last entry of $\mathbf{y}$.

To write the constraint in (2.17) in matrix form, we can define a matrix $\mathbf{H} \in \mathbb{R}^{B \times m}$, where its rows indexed by $\ell \in [B]$, and its columns are labeled similar to $\mathbf{y}$. Moreover, we have

$$\mathbf{H}_{b,(\ell,\mathcal{S})} = \begin{cases} 1 & \text{if } b = \ell \\ 0 & \text{if } b \neq \ell, \end{cases}$$
$$\mathbf{H}_{b,m} = 0, \qquad \forall b \in [B].$$

Thus, the constraint in (2.17) is equivalent to $\mathbf{H}\mathbf{y} \leq \mathbf{1}$.

Similarly, to write the constraint in (2.18), we define a matrix $\mathbf{G} \in \mathbb{R}^{K\binom{K-1}{t} \times m}$. The columns of $\mathbf{G}$ are labeled similar the entries of $\mathbf{y}$, and each row in $\mathbf{G}$ is labeled by a pair $(k, \mathcal{T})$ where $\mathcal{T} \subseteq [K] \setminus \{k\}$, and $|\mathcal{T}| = t$. An entry at row $(k, \mathcal{T})$ and column $(\ell, \mathcal{S})$

is given by

$$\mathbf{G}_{(k,\mathcal{T}),(\ell,\mathcal{S})} = \begin{cases} -\overline{F}_{L_k}(\ell) & \text{if } \mathcal{S} = \mathcal{T} \cup \{k\}, \\ 0 & \text{otherwise,} \end{cases}$$

and the entries in the mth column are

$$\mathbf{G}_{(k,\mathcal{T}),m} = \frac{1}{\binom{K}{t}}, \qquad \forall (k,\mathcal{T}).$$

With this, the constraint (2.18) is reduced to $\mathbf{G}\mathbf{y} \leq \mathbf{0}$.

We can conclude the following proposition by rephrasing the coding scheme devised above and its constraints in a linear form.

Proposition 2.2. For any K-DTVBC with cache placement strategy $\mathfrak{C}$ and a (distinct) request profile $\boldsymbol{d}$, the source rate f_{LP} given by

$$\begin{aligned} f_{\mathsf{LP}} = \min\ & [\mathbf{0}_{m-1}^T, -1]\mathbf{y} \\ \text{s.t.}\quad & \mathbf{G}\mathbf{y} \leq \mathbf{0} \\ & \mathbf{H}\mathbf{y} \leq \mathbf{1} \\ & -\mathbf{y} \leq \mathbf{0}, \end{aligned} \tag{2.20}$$

is achievable.

In the following example, we evaluate f_{LP} by solving (2.20) for a non-degraded broadcast channel with $K = 3$ users. We also solve the LP in (2.12) and show that the achievable rate and the upper bound do not match. This shows that our result does not provide an exact characterization for the maximum source rate, when the channel is not degraded, and the number of users is more than 2.

Example 2.2 (*Non-Degraded Channel Case*)**.** Consider a network with $K = 3$ users $N = 3$ files, namely $\{W_1, W_2, W_3\}$, and $B = 3$ transmit levels. The channel statistics of the three users are given by the CCDFs as

$$\begin{aligned} \overline{F}_{L_1} &= [0.9, 0.3, 0.3]^T, \\ \overline{F}_{L_2} &= [0.7, 0.4, 0.4]^T, \\ \overline{F}_{L_3} &= [0.5, 0.5, 0.5]^T. \end{aligned}$$

Consider the caching strategy $\mathfrak{C}^{\mathsf{cent}}$ with $\mu = 1/3$, i.e.,

$$\mathbf{c}_1 = (0, 1/3], \quad \mathbf{c}_2 = (1/3, 2/3], \quad \mathbf{c}_3 = (2/3, 1].$$

Again, we assume that user k is interested in file $W_k^{(n)}$, for $k \in \{1, 2, 3\}$. Here, the LP in (2.20) is given by

$$\begin{aligned} f_{\mathsf{LP}} = \min \ & [\mathbf{0}_{1\times(m-1)}, -1]\mathbf{y} \\ \text{s.t.} \quad & \mathbf{G}\mathbf{y} \le \mathbf{0} \\ & \mathbf{H}\mathbf{y} \le \mathbf{1} \\ & -\mathbf{y} \le \mathbf{0}, \end{aligned} \tag{2.21}$$

where the matrices $\mathbf{G}$ and $\mathbf{H}$ are given in (2.22) and (2.23), at the top of this page.

The optimum solution of the LP in (2.21) is also presented in (2.24) where $\mathbf{y}_{(\ell,\mathcal{S})}$ indicates the fraction of time that the transmitter uses signal level ℓ to send a coded message $W_{\mathcal{S}}^{(n)}$. The transmitter has to send coded messages $W_{\{1,2\}}^{(n)} = W_{1,\{2\}}^{(n)} \oplus W_{2,\{1\}}^{(n)}$, $W_{\{1,3\}}^{(n)} = W_{1,\{3\}}^{(n)} \oplus W_{3,\{1\}}^{(n)}$, and $W_{\{2,3\}}^{(n)} = W_{2,\{3\}}^{(n)} \oplus W_{3,\{2\}}^{(n)}$. Note that the source rate of $f_{\mathsf{LP}} = \mathbf{y}_m = 3/2$ is achievable. Therefore, since the rate of each coded message is $1/3$ of the rate of the original files, the source rate for each coded message is $f_{\mathsf{LP}}/3 = 1/2$.

The transmission scheme devised by (2.24) suggests that $W_{\{1,2\}}^{(n)}$ will be broadcast over the top level ($\ell = 1$) for $2/3$ fraction of time, and the second level ($\ell = 2$) for $1/12$ fraction. Thus, user 1 is able to decode the message $W_{\{1,2\}}^{(n)}$, as

$$0.9 \times \frac{2}{3} + 0.3 \times \frac{1}{12} = \frac{5}{8} \ge \frac{1}{2} = \frac{f_{\mathsf{LP}}}{3}.$$

Similarly, user 2 decodes the message, since

$$0.7 \times \frac{2}{3} + 0.4 \times \frac{1}{12} = \frac{1}{2} \ge \frac{1}{2} = \frac{f_{\mathsf{LP}}}{3}.$$

A similar argument holds for decodability of $W_{\{1,3\}}^{(n)}$ at users 1 and 3, as well as decodability of $W_{\{2,3\}}^{(n)}$ at users 2 and 3.

Next, we evaluate the upper-bound $f^\star(\mathfrak{C}^{\mathsf{cent}}, \boldsymbol{d})$. We solve the LP problems in (2.12) for all possible permutations. The bound corresponding to each permutation is given in Table 2.1.

π	$f_\pi^\star(\mathfrak{C}^{\mathsf{cent}}, \boldsymbol{d})$
$(3,2,1)$	1.66
$(3,1,2)$	1.76
$(2,3,1)$	**1.61**
$(2,1,3)$	1.62
$(1,3,2)$	1.73
$(1,2,3)$	1.64

Table 2.1: The upper bound on the source rate for each permutation with the caching strategy $\mathfrak{C}^{\mathsf{cent}}$ and the normalized cache size $\mu = 1/3$.

Therefore, we get

$$\begin{aligned}\hat{f} &:= f^\star(\mathfrak{C}^{\mathsf{cent}}, \boldsymbol{d}) = \min_{\pi\in\Pi} f_\pi^\star(\mathfrak{C}^{\mathsf{cent}}, \boldsymbol{d}) \\ &= \min\{1.66, 1.76, \mathbf{1.61}, 1.62, 1.73, 1.64\} = 1.61,\end{aligned}$$

where the minimum value is obtained for $\pi^\star = (2,3,1)$ with $\boldsymbol{\omega}^\star = (0, 1.25, 1)$. Clearly, we have $f_{\mathsf{LP}} = 1.5 < 1.61 = \hat{f}$, and there is a gap between the achievable rate and the upper bound. ◇

2.5 Proof of Theorem 2.1

In this section, we present the proof of Theorem 2.1. To do this, we first provide the converse proof and then discuss the achievability part. For the converse proof, we first enhance the channel to achieve a degraded broadcast channel. Next, we introduce a lemma that allows us to establish an upper bound on the maximum achievable source rate for a physically degraded BC. By applying this lemma to the degraded channel we have obtained, we are able to characterize the maximum achievable source rate.

2.5.1 Converse

The converse proof of Theorem 2.1 is based on the result of [46]. We need to construct a degraded broadcast channel. In this regard, we replace L_2, the channel of User 2, with

an enhanced channel $\widetilde{L}_2$. For a given $\omega \geq 1$, we define

$$\overline{F}_{\widetilde{L}_2}(\ell) := \min\left[1, \max\left[\overline{F}_{L_2}(\ell), \omega \overline{F}_{L_1}(\ell)\right]\right], \quad \ell \in [B], \tag{2.25}$$

which is the CCDF of $\widetilde{L}_2$. Moreover, we define $\widetilde{Y}_2 := D^{B-\widetilde{L}_2} X = X(1:\widetilde{L}_2)$. Hence, we have a degraded broadcast channel, i.e., $X \leftrightarrow \widetilde{Y}_2 \leftrightarrow Y_1$.

Next, we derive an upper bound on the maximum achievable source rate for a physically degraded BC. We refer to Section 2.10 for the proof of Lemma 2.1.

Lemma 2.1. Consider a physically degraded memoryless BC described by $P_{Y_1,Y_2|X}$ for a given cache placement strategy $\mathfrak{C}$ and a distinct request profile $\boldsymbol{d}$. Then, any achievable source rate $f(\mathfrak{C}, \boldsymbol{d})$ satisfies

$$f(\mathfrak{C}, \boldsymbol{d}) \leq \frac{I(U_1; Y_1)}{1-\mu}, \tag{2.26}$$

$$f(\mathfrak{C}, \boldsymbol{d}) \leq \frac{I(X; Y_2|U_1)}{1-2\mu}. \tag{2.27}$$

for some U_1 satisfying $U_1 \leftrightarrow X \leftrightarrow Y_2 \leftrightarrow Y_1$.

Now, we are ready for the proof of Theorem 2.1. We can use Lemma 2.1 for the degraded channel obtained by the enhancement procedure in (2.25). Let U_1 be the random variable satisfying the claim of Lemma 2.1, and form a Markov chain $U_1 \leftrightarrow X \leftrightarrow \widetilde{Y}_2 \leftrightarrow Y_1$. Using the fact that CSI is available at the receivers, for the terms in (2.26) and (2.27), we can write

$$\begin{aligned}
I(U_1; Y_1, L_1) &= I(U_1; X(1:L_1), L_1) \\
&= \sum_{j=1}^{B} \mathbb{P}_{L_1}(j) I(U; X(1:j), L_1 = j) \\
&= \sum_{j=1}^{B} \left[\mathbb{P}_{L_1}(j) \sum_{\ell=1}^{j} I(U_1; X(\ell)|X(1:\ell-1)) \right] \\
&= \sum_{\ell=1}^{B} \left[I(U_1; X(\ell)|X(\ell-1)) \sum_{j=1}^{\ell} \mathbb{P}_{L_1}(j) \right] \\
&= \sum_{\ell=1}^{B} \overline{F}_{L_1}(\ell) I(U_1; X(\ell)|X(\ell-1)) \\
&= \sum_{\ell=1}^{B} \overline{F}_{L_1}(\ell) \Big[H(X(\ell)|X(1:\ell-1)) - H(X(\ell)|X(1:\ell-1)), U_1) \Big]
\end{aligned} \tag{2.28}$$

Similarly, we have

$$\begin{aligned}
I(X;\widetilde{Y},\widetilde{L}_2|U_1) &= I(X;X(1:\widetilde{L}_2),\widetilde{L}_2|U_1)\\
&= I(X;\widetilde{L}_2|U_1) + I(X;X(1:\widetilde{L}_2)|U_1,\widetilde{L}_2)\\
&\overset{(a)}{=} \sum_{j=1}^{B} \mathbb{P}_{\widetilde{L}_2}(j) I(X;X(1:j)|U_1,\widetilde{L}_2=j)\\
&= \sum_{j=1}^{B}\left[\mathbb{P}_{\widetilde{L}_2}(j)\sum_{\ell=1}^{j} I(X;X(\ell)|X(1:\ell-1),U_1)\right]\\
&= \sum_{\ell=1}^{B}\left[I(X;X(\ell)|X(\ell-1),U_1)\sum_{j=1}^{\ell}\mathbb{P}_{\widetilde{L}_2}(j)\right]\\
&= \sum_{\ell=1}^{B}\overline{F}_{\widetilde{L}_2}(\ell) I(X;X(\ell)|X(\ell-1),U_1)\\
&= \sum_{\ell=1}^{B}\overline{F}_{\widetilde{L}_2}(\ell)\big[H(X(\ell)|X(1:\ell-1),U_1) - H(X(\ell)|X,X(1:\ell-1)),U_1)\big]\\
&= \sum_{\ell=1}^{B}\overline{F}_{\widetilde{L}_2}(\ell) H(X(\ell)|X(1:\ell-1),U_1), \qquad (2.29)
\end{aligned}$$

where (a) follows from the fact that $\widetilde{L}_2$ is independent from U_1 and X. Therefore, from Lemma 2.1, we get

$$\omega(1-\mu)f(\mathfrak{C},\boldsymbol{d}) \le \omega\sum_{\ell=1}^{B}\overline{F}_{L_1}(\ell)\big[H(X(\ell)|X(1:\ell-1)) - H(X(\ell)|X(1:\ell-1)),U_1)\big], \qquad (2.30)$$

and

$$(1-2\mu)f(\mathfrak{C},\boldsymbol{d}) \le \sum_{\ell=1}^{B}\overline{F}_{\widetilde{L}_2}(\ell) H(X(\ell)|X(1:\ell-1),U_1). \qquad (2.31)$$

Taking the sum of the two inequalities in (2.30) and (2.31), we arrive at

$$\begin{aligned}
&(\omega(1-\mu)+(1-2\mu))f(\mathfrak{C},\boldsymbol{d})\\
&\le \omega\sum_{\ell=1}^{B}\overline{F}_{L_1}(\ell)H(X(\ell)|X(1:\ell-1)) + \sum_{\ell=1}^{B}(\tilde{g}(\ell)-\omega)\overline{F}_{L_1}(\ell)H(X(\ell)|X(1:\ell-1),U_1), \qquad (2.32)
\end{aligned}$$

where $\tilde{g}(\ell) := \overline{F}_{\widetilde{L}_2}(\ell)/\overline{F}_{L_1}(\ell)$. The summands of the first summation in (2.32) will be maximized by an i.i.d. Bernoulli random variable choice for $X_1,\dots,X_B$. Moreover, the

terms in the second summation can be maximized if

$$H(X(\ell)|X(1{:}\ell{-}1),U_1)=\begin{cases}1 & \tilde{g}(\ell)>\omega,\\ 0 & \tilde{g}(\ell)\le\omega,\end{cases}$$

which can be satisfied by an optimum choice for U_1, given by

$$U_1=\{X(\ell)|\tilde{g}(\ell)\le\omega\}.$$

Hence, for (2.32) we can write

$$\begin{aligned}
&(\omega(1-\mu)+(1-2\mu))f(\mathfrak{C},\boldsymbol{d})\\
&\le\omega\sum_{\ell:\tilde{g}(\ell)\le\omega}\overline{F}_{L_1}(\ell)+\sum_{\ell:\tilde{g}(\ell)>\omega}\tilde{g}(\ell)\overline{F}_{L_1}(\ell)\\
&\overset{(a)}{=}\omega\sum_{\ell:\tilde{g}(\ell)\le\omega}\overline{F}_{L_1}(\ell)+\sum_{\ell:\tilde{g}(\ell)>\omega}\overline{F}_{\widetilde{L}_2}(\ell)\\
&\overset{(b)}{=}\omega\sum_{\ell:\tilde{g}(\ell)\le\omega}\overline{F}_{L_1}(\ell)+\sum_{\ell:\tilde{g}(\ell)>\omega}\overline{F}_{L_2}(\ell)\\
&=\omega R_1(\omega)+R_2(\omega),
\end{aligned}\tag{2.33}$$

where $R_1(\omega)$ and $R_2(\omega)$ are defined in (2.9). We note that in the chain of inequalities in (2.33), the step (a) follows from $\tilde{g}(\ell)\overline{F}_{L_1}(\ell)=\overline{F}_{L_2}(\ell)$ and $\{\ell|\tilde{g}(\ell)>\omega\}=\{\ell|g(\ell)>\omega\}$ and (b) holds since $\overline{F}_{\widetilde{L}_2}(\ell)=\overline{F}_{L_2}(\ell)$ whenever $g(\ell)>\omega$. Dividing both sides of (2.33) by $\omega(1{-}\mu)+(1{-}2\mu)$ and minimizing over all $\omega\ge1$, we arrive at the the first minimization in (2.7).

For $0\le\omega\le1$, we can repeat the steps in (2.25) through (2.33) by swapping the labels of the users and replacing ω by $\frac{1}{\omega}$. Under these reversed labels, we now enhance the channel of User 1 and get the second minimization in (2.7).

Finally, we characterize the maximum achievable source rate when $\mu>\frac{1}{2}$. Starting (2.26) and using (2.28), we can write

$$\begin{aligned}
(1-\mu)f(\mathfrak{C},\boldsymbol{d})&\le I(U_1;Y_1,L_1)\\
&=\sum_{\ell=1}^{B}\overline{F}_{L_1}(\ell)\big[H(X(\ell)|X(1:\ell-1))-H(X(\ell)|X(1{:}\ell{-}1)),U_1)\big]\\
&\le\sum_{\ell=1}^{B}\overline{F}_{L_1}(\ell).
\end{aligned}\tag{2.34}$$

Similarly, by swapping the labels of the users, we can repeat the steps in (2.34) that leads us to

$$(1-\mu)f(\mathfrak{C},\boldsymbol{d}) \le \sum_{\ell=1}^{B} \overline{F}_{L_2}(\ell).$$

This completes the converse proof.

2.5.2 Achievability

The achievability proof of Theorem 2.1 is based on a linear scheme in which the transmitter only broadcasts a raw and linear combination of the messages. To show that $f^\star$ is achievable, we split the messages into three messages, including two private messages (one for each user) and a common message, which is intended for both users. Then, we allocate the (signal) levels in $[B]$ to each of these messages and prove that both users can decode their desired file.

Without loss of generality, we assume the source rate $f^\star$ is obtained by the first minimization in (2.7) and (2.8) for $0 \le \mu \le \frac{1}{2}$ and $\frac{1}{2} \le \mu \le 1$, respectively. Now, we present the achievability proof for each regime of μ.

$0 \le \mu \le \frac{1}{2}$:

Consider the central cache placement of Definition 2.5. Let us denote the file requested by user k by W_{d_k} for $k \in [2]$. Recall that user 1 needs $W^{(n)}_{d_1}$, which is partitioned into $\left(W^{(n)}_{d_1,\varnothing}, W^{(n)}_{d_1,\{1\}}, W^{(n)}_{d_1,\{2\}}, W^{(n)}_{d_1,\{1,2\}}\right)$. Note that since $\mu \le \frac{1}{2}$, from (2.6) we can conclude that $|\mathbf{c}^{\mathsf{cent}}_{\{1,2\}}| = 0$, and hence $W^{(n)}_{d_1,\{1,2\}} = \varnothing$. While user 1 has $W^{(n)}_{d_1,\{1\}}$ in its cache, the subfiles $W^{(n)}_{d_1,\varnothing}$ and $W^{(n)}_{d_1,\{2\}}$, need to be delivered. Similarly, the user 2 will be served be by sending $W^{(n)}_{d_2,\varnothing}$, $W^{(n)}_{d_2,\{1\}}$. Instead of sending these message separately, we send *individual* messages $W^{(n)}_{d_1,\varnothing}$ and $W^{(n)}_{d_2,\varnothing}$, as well as the *common* message $W^{(n)}_{d_1,\{2\}} \oplus W^{(n)}_{d_2,\{1\}}$. We aim to send each individual message to the intended user and the common message to both receivers. To this end, we need to allocate the *levels* and *time* among the messages. We first define $g(\ell) \coloneqq \overline{F}_{L_2}(\ell)/\overline{F}_{L_1}(\ell)$ for each level $\ell \in [B]$, and sort all the B levels of the channel in an non-decreasing order according to $g(\cdot)$. This leads to a one-to-one mapping $\lambda : [B] \to [B]$ that sorts the level, and thus, $g(\lambda(1)) \le g(\lambda(2)) \le \cdots \le g(\lambda(B-1)) \le g(\lambda(B))$. For notational simplicity, we rename the levels and define $\ell_i \coloneqq \lambda(i)$ and $\gamma_i \coloneqq g(\ell_i)$, for every $i \in [B]$. We clearly

have $\gamma_1 \le \gamma_2 \le \cdots \le \gamma_B$. We refer to Figure 2.3 for clarification. Our proposed level (and time) allocation scheme is parameterized by $(u, v; \alpha, \beta)$, where $u, v \in [B]$ with $u \le v$, and $0 < \alpha, \beta \le 1$: We use levels $\{\ell_1, \ell_2 \ldots, \ell_{u-1}\}$ for the entire communication block and level ℓ_u for an α fraction of time to send the individual message $W^{(n)}_{d_1,\varnothing}$. Similarly, the individual message $W^{(n)}_{d_2,\varnothing}$ will be sent on levels $\{\ell_{v+1}, \ldots, \ell_B\}$ for the entire communication block and on level ℓ_v for β fraction of time. The remaining levels (including the remaining $(1-\alpha)$ fraction of ℓ_u and $(1-\beta)$ fraction of ℓ_v) will be used to send the common message. Such a delivery strategy can support any source rate f that satisfies

$$\begin{aligned}
&\sum_{i<u} \overline{F}_{L_1}(\ell_i) + \alpha \overline{F}_{L_1}(\ell_u) \ge \frac{1}{n}\left|W^{(n)}_{d_1,\varnothing}\right| = (1-2\mu)f,\\
&\sum_{i>v} \overline{F}_{L_2}(\ell_i) + \beta \overline{F}_{L_2}(\ell_v) \ge \frac{1}{n}\left|W^{(n)}_{d_2,\varnothing}\right| = (1-2\mu)f,\\
&(1-\alpha)\overline{F}_{L_1}(\ell_u) + \sum_{u<i<v} \overline{F}_{L_1}(\ell_i) + (1-\beta)\overline{F}_{L_1}(\ell_v) \ge \frac{1}{n}\left|W^{(n)}_{d_1,\{2\}} \oplus W^{(n)}_{d_2,\{1\}}\right| = \mu f, \qquad (2.35)\\
&(1-\alpha)\overline{F}_{L_2}(\ell_u) + \sum_{u<i<v} \overline{F}_{L_2}(\ell_i) + (1-\beta)\overline{F}_{L_2}(\ell_v) \ge \frac{1}{n}\left|W^{(n)}_{d_1,\{2\}} \oplus W^{(n)}_{d_2,\{1\}}\right| = \mu f.
\end{aligned}$$

It is easy to verify that constraints in (2.35) are feasible if and only if the constraints

$$\begin{aligned}
&\sum_{i<u} \overline{F}_{L_1}(\ell_i) + \alpha \overline{F}_{L_1}(\ell_u) \ge (1-2\mu)f,\\
&(1-\alpha)\overline{F}_{L_2}(\ell_u) + \sum_{i>u} \overline{F}_{L_2}(\ell_i) \ge (1-\mu)f\\
&\sum_{i>v} \overline{F}_{L_2}(\ell_i) + \beta \overline{F}_{L_2}(\ell_v) \ge (1-2\mu)f,\\
&\sum_{i<v} \overline{F}_{L_1}(\ell_i) + (1-\beta)\overline{F}_{L_1}(\ell_v) \ge (1-\mu)f,
\end{aligned} \qquad (2.36)$$

are satisfied. Note that we can optimize the allocation parameters $(u, v; \alpha, \beta)$. Moreover, the first and the second constraints in (2.36) only depend on (u, α), and the third and the fourth constraints only depend on (v, β). These motivate defining

$$f_1(u,\alpha) := \min\left(\frac{1}{1-2\mu}\left[\sum_{i<u} \overline{F}_{L_1}(\ell_i) + \alpha \overline{F}_{L_1}(\ell_u)\right], \frac{1}{1-\mu}\left[\sum_{i>u} \overline{F}_{L_2}(\ell_i) + (1-\alpha)\overline{F}_{L_2}(\ell_u)\right]\right), \qquad (2.37)$$

$$f_2(v,\beta) := \min\left(\frac{1}{1-\mu}\left[\sum_{i<v} \overline{F}_{L_1}(\ell_i) + (1-\beta)\overline{F}_{L_1}(\ell_v)\right], \frac{1}{1-2\mu}\left[\sum_{i>v} \overline{F}_{L_2}(\ell_i) + \beta \overline{F}_{L_2}(\ell_v)\right]\right). \qquad (2.38)$$

Our goal would be to maximize $\min(f_1(u,\alpha), f_2(v,\beta))$. The following lemma formally presents the properties of the optimum solution of $f_1(u,\alpha)$ and $f_2(v,\beta)$. We show that the maximum of $\min(f_1(u,\alpha), f_2(v,\beta))$ over the choice of $(u,v;\alpha,\beta)$ meets the upper bound of source rate in (2.7). This completes the achievability proof for the regime $0 \le \mu \le \frac{1}{2}$

Lemma 2.2. Consider a 2-DTVBC with a distinct request profile $\boldsymbol{d}$, a normalized cache size $\mu \le \frac{1}{2}$, and the central caching strategy $\mathfrak{C}^{\mathsf{cent}}$. Let $f_1^\star := \max_{u,\alpha} f_1(u,\alpha)$ and $f_2^\star := \max_{v,\beta} f_2(v,\beta)$, where $f_1(u,\alpha)$ and $f_2(v,\beta)$ are defined in (2.37) and (2.38), respectively. Then, the following properties hold:

(i) The source rate $\min(f_1^\star, f_2^\star)$ is achievable;

(ii) If $(u^\star,\alpha^\star) := \arg\max f_1(u,\alpha)$ be the maximizer of f_1 and $(v^\star,\beta^\star) := \arg\max f_2(v,\beta)$ be the maximizer of f_2, then we have $u^\star \le v^\star$;

(iii) If $f_1^\star \le f_2^\star$ then $g(\ell_{u^\star}) \le 1$. Alternatively, if $f_1^\star \ge f_2^\star$, then we have $g(\ell_{v^\star}) \ge 1$;

(iv) For $f^\star$ defined in (2.7), we have $f^\star \le \min(f_1^\star, f_2^\star)$.

The proof of Lemma 2.2 is presented in Section 2.10. It is worth noting that parts (i) and (iv) of the lemma above immediately yield the achievability proof of Theorem 2.1 for $0 \le \mu \le \frac{1}{2}$.

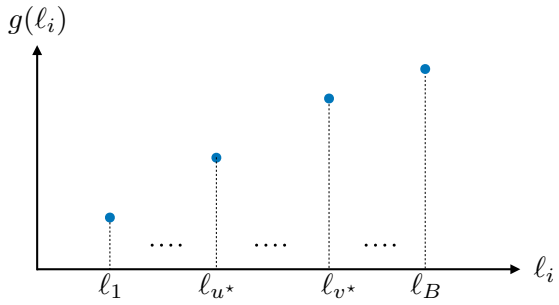

Figure 2.3: Sorting the signal levels of a 2-user system according to their ratio.

$\frac{1}{2} \le \mu \le 1$:

We need to show that the minimum attained in (2.8) is achievable. Let $\boldsymbol{d} = (d_1, d_2)$ be the demand profile. The file $W_{d_1}^{(n)}$ is partitioned into $\left(W_{d_1,\varnothing}^{(n)}, W_{d_1,\{1\}}^{(n)}, W_{d_1,\{2\}}^{(n)}, W_{d_1,\{1,2\}}^{(n)}\right)$. Note that since $\mu \geq \frac{1}{2}$, from (2.6) we can conclude that $|\mathbf{c}_{\varnothing}^{\mathsf{cent}}| = 0$, and hence $W_{d_1,\varnothing}^{(n)} = \varnothing$. This means user 1 has $W_{d_1,\{1\}}^{(n)}$ and $W_{d_1,\{1,2\}}^{(n)}$ in its cache, and only $W_{d_1,\{2\}}^{(n)}$ need to be delivered over the channel. Similarly, user 2 will be served by $W_{d_2,\{1\}}^{(n)}$. The transmitter only needs to multicast a common message $W_{d_1,\{2\}} \oplus W_{d_2,\{1\}}$ to both users over the signal levels in $[B]$. The size of this common message is $(1-\mu)\, f^\star$. Hence, the maximum achievable source rate is given the minimum of the capacities of the channels to two users. For user i, the maximum rate is

$$f^\star = \frac{\sum_{\ell=1}^{B} \overline{F}_{L_i}(\ell)}{1-\mu}.$$

Taking the minimum over $i \in \{1,2\}$, we get the rate in (2.8). This completes the proof of the theorem. □

2.6 Achievability Proof of Theorem 2.2

The achievability proof of Theorem 2.2 is based on the achievable scheme in Section 2.4 and Proposition 2.2. We consider a specific μ for which $t = K\mu = KM/N \in \mathbb{N}$. Therefore, from (2.5), we have

$$\mu_{[k]}^{\mathsf{cent}} = \frac{\binom{K-k}{t}}{\binom{K}{t}}, \tag{2.39}$$

for every $k \in [K]$. Now, we show that every set $\{z_{\ell,k}\}$ that satisfies (2.11) provides a feasible set of $\{y_{\ell,\mathcal{S}}\}$ for the achievable scheme of Proposition 2.2. Recall that we assume the channel is degraded (i.e., $L_K \geq_{\mathsf{st}} \cdots \geq_{\mathsf{st}} L_1$). Let $k_{\mathsf{w}}(\mathcal{S}) \coloneqq \min_{k\in\mathcal{S}} k$ be the index of the *weakest* user in the set $\mathcal{S}$. For every $\mathcal{S} \subseteq [K]$ with $|\mathcal{S}| = t+1$, we set

$$y_{\ell,\mathcal{S}} = \frac{1}{\binom{K-k_{\mathsf{w}}(\mathcal{S})}{t}} z_{\ell,k_{\mathsf{w}}(\mathcal{S})}. \tag{2.40}$$

First, note that $y_{\ell,\mathcal{S}} \geq 0$ for every $\ell \in [B]$.

Next, we have

$$\begin{aligned}
\sum_{\substack{\mathcal{S}\subseteq[K]\\|\mathcal{S}|=t+1}} y_{\ell,\mathcal{S}} &\overset{\text{(a)}}{=} \sum_{\substack{\mathcal{S}\subseteq[K]\\|\mathcal{S}|=t+1}} \frac{z_{\ell,k_{\mathsf{w}}(\mathcal{S})}}{\binom{K-k_{\mathsf{w}}(\mathcal{S})}{t}} \\
&= \sum_{k=1}^{K} \sum_{\substack{\mathcal{S}\subseteq[K]\\|\mathcal{S}|=t+1\\k_{\mathsf{w}}(\mathcal{S})=k}} \frac{z_{\ell,k}}{\binom{K-k}{t}} \\
&= \sum_{k=1}^{K} \frac{z_{\ell,k}}{\binom{K-k}{t}} \sum_{\substack{\mathcal{S}\subseteq[K]\\|\mathcal{S}|=t+1\\k_{\mathsf{w}}(\mathcal{S})=k}} 1 \\
&\overset{\text{(b)}}{=} \sum_{k=1}^{K} \frac{z_{\ell,k}}{\binom{K-k}{t}} \binom{K-k}{t} \\
&= \sum_{k=1}^{K} z_{\ell,k} \overset{\text{(c)}}{\leq} 1,
\end{aligned} \tag{2.41}$$

where (a) follows from (2.40), in (b) we used the fact that any $\mathcal{S}$ satisfying $\mathcal{S} \subseteq [K]$, $|\mathcal{S}| = t+1$ and $k_{\mathsf{w}}(\mathcal{S}) = k$ should be of the form of $\mathcal{S} = \{k\} \cup \mathcal{T}$, where $\mathcal{T} \subseteq \{k+1, k+2, \ldots, K\}$, with $|\mathcal{T}| = t$. Hence, the number of such $\mathcal{S}$'s is $\binom{K-k}{t}$. Lastly, (c) follows from the last constraint in (2.11). Thus, (2.41) shows that the $\{y_{\ell,\mathcal{S}}\}$ introduced above satisfies (2.17).

Furthermore, the degradedness of the channel implies that

$$\begin{aligned}
\sum_{\ell=1}^{B} \overline{F}_{L_k}(\ell) y_{\ell,\mathcal{S}} &\overset{\text{(a)}}{\geq} \sum_{\ell=1}^{B} \overline{F}_{L_{k_{\mathsf{w}}(\mathcal{S})}}(\ell) y_{\ell,\mathcal{S}} \\
&= \frac{1}{\binom{K-k_{\mathsf{w}}(\mathcal{S})}{t}} \sum_{\ell=1}^{B} \overline{F}_{L_{k_{\mathsf{w}}(\mathcal{S})}}(\ell) z_{\ell,k_{\mathsf{w}}(\mathcal{S})} \\
&\overset{\text{(b)}}{\geq} \frac{1}{\binom{K-k_{\mathsf{w}}(\mathcal{S})}{t}} \left(1 - \mu^{\mathsf{cent}}_{[k_{\mathsf{w}}(\mathcal{S})]}\right) \bar{f} \\
&\overset{\text{(c)}}{\geq} \frac{1}{\binom{K-k_{\mathsf{w}}(\mathcal{S})}{t}} \frac{\binom{K-k_{\mathsf{w}}(\mathcal{S})}{t}}{\binom{K}{t}} \bar{f} \\
&= \frac{\bar{f}}{\binom{K}{t}},
\end{aligned} \tag{2.42}$$

for every $k \in \mathcal{S}$. Here, (a) holds since for every user $k \in \mathcal{S}$ and every $\ell \in [B]$, we have $\overline{F}_{L_k}(\ell) \geq \overline{F}_{L_{k_{\mathsf{w}}(\mathcal{S})}}(\ell)$, (b) follows from the first constraint in (2.11), and we used (2.39)

with $k = k_{\mathsf{w}}(\mathcal{S})$ in (c). This shows that the $\{y_{\ell,\mathcal{S}}\}$ sequence introduced in (2.40) satisfies (2.18).

We proved that the $\{y_{\ell,\mathcal{S}}\}$ introduced in (2.40) satisfies all constraints of the LP problem in Proposition 2.2. In other words, every $\{z_{\ell,k}\}$ satisfying (2.11) provides a feasible solution $\{y_{\ell,\mathcal{S}}\}$ for the (achievable) optimization method in (2.19). This, together with Proposition 2.2 (that any feasible solution of $\{y_{\ell,\mathcal{S}}\}$ leads to an achievable source rate) completes the achievability proof of Theorem 2.2. □

2.7 Proof of Theorem 2.3 and Proposition 2.1

In this section, we provide the proof of Theorem 2.3 and Proposition 2.1. First, we present some auxiliary lemmas whose proofs are provided in Section 2.10.

2.7.1 Preliminary results

The proof of Theorem 2.3 is built based on the result of [47], in which the rate region of an erasure K-user broadcast channel is characterized. We first need to enhance the channels to convert the network to a degraded broadcast channel. To this end, we will replace L_k, the channel of User k, by a *stronger* channel $\widetilde{L}_k$, so that the channel of user k statistically degrades that of user $k-1$, for $k = 2, 3, \ldots, K$. More precisely, for a given weight vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K) \in [0, \infty)^K$ with *sorted* entries $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_K$, we define

$$\overline{F}_{\widetilde{L}_k}(\ell) := \min\left[1, \max\left(\overline{F}_{L_k}(\ell), \frac{\omega_{k-1}}{\omega_k}\overline{F}_{\widetilde{L}_{k-1}}(\ell)\right)\right], \tag{2.43}$$

for every $\ell \in [B]$ and $k \in \{2, 3, \ldots, K\}$. with an initialization given by $\overline{F}_{\widetilde{L}_1}(\ell) = \overline{F}_{L_1}(\ell)$, for every $\ell \in [B]$.

The following lemma demonstrates some of the properties of the enhanced channel, which will be useful in the proof of Theorem 2.3.

Lemma 2.3. The CCDF of $\widetilde{L}_k$ providing in (2.43) has the following properties

(i) If $\overline{F}_{\widetilde{L}_k}(\ell) = 1$, then

$$\overline{F}_{\widetilde{L}_u}(\ell) = 1,$$

for every $u \geq k$.

(ii) If $\omega_k \overline{F}_{\widetilde{L}_k}(\ell) > \omega_{k-1}\overline{F}_{\widetilde{L}_{k-1}}(\ell)$, then

$$\omega_k \overline{F}_{L_k}(\ell) = \omega_k \overline{F}_{\widetilde{L}_k}(\ell) > \omega_{k-1}\overline{F}_{\widetilde{L}_{k-1}}(\ell) \geq \cdots \geq \omega_1 \overline{F}_{\widetilde{L}_1}(\ell).$$

(iii) If $\omega_k \overline{F}_{\widetilde{L}_k}(\ell) < \omega_{k-1}\overline{F}_{\widetilde{L}_{k-1}}(\ell)$, then

$$\omega_K \overline{F}_{\widetilde{L}_K}(\ell) \leq \cdots \leq \omega_k \overline{F}_{\widetilde{L}_k}(\ell) < \omega_{k-1}\overline{F}_{\widetilde{L}_{k-1}}(\ell).$$

(iv) The maximum of the weighted channel parameters satisfy

$$\max_k \omega_k \overline{F}_{\widetilde{L}_k}(\ell) = \max_k \omega_k \overline{F}_{L_k}(\ell),$$

for every $\ell \in [B]$.

The following corollary is based on the properties presented in Lemma 2.3 and provides a better understanding of the enhancement procedure. Note that for a given $\ell \in [B]$, the quantity $\omega_k \overline{F}_{\widetilde{L}_k}(\ell)$ may be equal for $k \neq k'$. Hence, $\arg\max_k \omega_k \overline{F}_{\widetilde{L}_k}(\ell)$ is a set, with possibly many elements.

Corollary 2.1. Let $k^\star := \min \arg\max_k \omega_k \overline{F}_{\widetilde{L}_k}(\ell)$ and $u^\star := \max \arg\max_k \omega_k \overline{F}_{\widetilde{L}_k}(\ell)$ for any fixed level $\ell \in [B]$. Then, we can decompose the set of users $[K]$ into the following non-overlapping subsets

$$[K] = \{1, \ldots, k^\star - 1\} \cup \{k^\star, \ldots, u^\star\} \cup \{u^\star + 1, \ldots, K\}.$$

Then, we arrive at the below properties

(i) The sequence $\{\omega_k \overline{F}_{L_k}(\ell)\}_{k=1}^{k^\star - 1}$ is non-decreasing, i.e.,

$$\omega_1 \overline{F}_{\widetilde{L}_1}(\ell) \leq \cdots \leq \omega_{k^\star - 1}\overline{F}_{\widetilde{L}_{k^\star - 1}}(\ell) < \omega_{k^\star}\overline{F}_{\widetilde{L}_{k^\star}}(\ell).$$

(ii) The sequence $\{\omega_k \overline{F}_{L_k}(\ell)\}_{k=k^\star}^{u^\star}$ satisfies

$$\omega_{k^\star} \overline{F}_{L_{k^\star}}(\ell) = \omega_{k^\star} \overline{F}_{\widetilde{L}_{k^\star}}(\ell) = \cdots = \omega_{u^\star} \overline{F}_{\widetilde{L}_{u^\star}}(\ell).$$

(iii) The sequence $\{\omega_k \overline{F}_{L_k}(\ell)\}_{k=u^\star+1}^{K}$ is non-increasing, i.e.,

$$\omega_{u^\star} \overline{F}_{\widetilde{L}_{u^\star}}(\ell) > \omega_{u^\star+1} \overline{F}_{\widetilde{L}_{u^\star+1}}(\ell) \geq \cdots \geq \omega_K \overline{F}_{\widetilde{L}_K}(\ell).$$

The proof of Corollary 2.1 is presented in Section 2.10. Using the properties discussed in Corollary 2.1, we can visualize the behavior of the enhanced channels $\overline{F}_{\widetilde{L}_k}(\ell)$ and their weighted versions $\omega_k \overline{F}_{\widetilde{L}_k}(\ell)$ as in Figure 2.4.

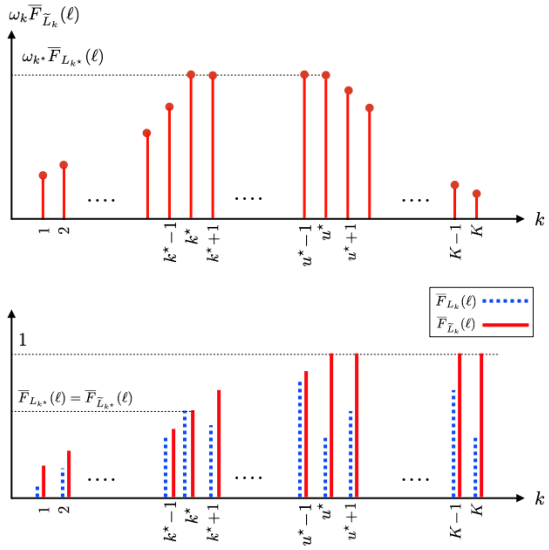

Figure 2.4: Channel enhancement: The behavior of the $\omega_k \overline{F}_{\widetilde{L}_k}(\ell)$ for the enhanced deterministic broadcast channel (top), and comparison of the channel parameters before and after enhancement (bottom).

In the next lemma, we provide an important property of the cache placement strategy.

Lemma 2.4. For a given caching strategy and mutually independent files $W_1^{(n)}, \ldots, W_N^{(n)}$, we get

$$I\left(W_i; C_{\mathcal{S}}^{(n)}\right) \leq n\mu_{\mathcal{S}} f, \quad \mathcal{S} \subseteq [K]. \tag{2.44}$$

Finally, we provide the extension of Lemma 2.1 to the K-users system in the following result.

Lemma 2.5. If a source rate $f(\mathfrak{C}, \boldsymbol{d})$ for a given caching strategy $\mathfrak{C}$ and a distinct request profile $\boldsymbol{d}$ is achievable on a physically degraded broadcast channel, i.e.,

$$X \leftrightarrow Y_K \leftrightarrow \cdots \leftrightarrow Y_1,$$

then $f(\mathfrak{C}, \boldsymbol{d})$ satisfies

$$f(\mathfrak{C}, \boldsymbol{d}) \le \frac{I(U_1; Y_1)}{1 - \mu_{[1]}}, \tag{2.45}$$

$$f(\mathfrak{C}, \boldsymbol{d}) \le \frac{I(U_k; Y_k | U_{k-1})}{1 - \mu_{[k]}}, \quad k \in [2 : K-1] \tag{2.46}$$

$$f(\mathfrak{C}, \boldsymbol{d}) \le \frac{I(X; Y_K | U_{K-1})}{1 - \mu_{[K]}}, \tag{2.47}$$

for some random variables $(U_1, U_2, \ldots, U_{K-1})$ that form a Markov chain

$$X \leftrightarrow U_{K-1} \leftrightarrow \cdots \leftrightarrow U_1.$$

The proof of Lemma 2.5 is presented in Section 2.10.

Remark 2.1. The global capacity of a broadcast channel depends on the underlying transition probability $\mathbb{P}(Y_1, \ldots, Y_K | X)$ only through its marginal conditional probabilities $\mathbb{P}(Y_1|X), \ldots, \mathbb{P}(Y_K|X)$. Therefore, the claim of Lemma 2.5 also applies to stochastically degraded BCs.

Now, we are ready to present the proof of Theorem 2.3.

2.7.2 An upper-bound on the achievable source rate

The main steps of the proof of Theorem 2.3 are twofold: We first enhance and replace the arbitrary $L_1, \ldots, L_K$ by the *degraded* $\widetilde{L}_1, \ldots, \widetilde{L}_K$ and then by exploiting Lemma 2.5, we derive an upper-bound on the achievable source rate.

Proof of Theorem 2.3. We first note that for arbitrary channels $L_1, \ldots, L_K$, the K-DTVBC is not degraded. Hence, we recursively enhance the channel of the users to obtain a set of degraded channels. In this regard, we consider a weight vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K) \in [0, \infty)^K$ with sorted entries $\omega_1 \ge \omega_2 \ge \cdots \ge \omega_K$. We define the enhanced channel output of user k as $\widetilde{Y}_k = D^{B - \widetilde{L}_k} X = X(1 : \widetilde{L}_k)$ where $\widetilde{L}_k$ is a random variable drawn according to $\overline{F}_{\widetilde{L}_k}$, given by (2.43), independent of all other users. Since $\omega_{k-1} \ge \omega_k$, from (2.43) we have $\overline{F}_{\widetilde{L}_k}(\ell) \ge \overline{F}_{\widetilde{L}_{k-1}}(\ell)$. Thus, from [47, Lemma 1], we can conclude that the enhanced broadcast channel is (stochastically) degraded. Now, we can use Lemma 2.5 for the degraded channel obtained by the enhancement procedure.

Let $U_1, \ldots, U_{K-1}$ be the random variables satisfying the claim of the lemma, and form a Markov chain $X \leftrightarrow U_{K-1} \leftrightarrow \cdots \leftrightarrow U_1$. Then, given the fact that CSI is available at the receivers, the terms in Lemma 2.5 for the deterministic channel of interest will be simplified to

$$
\begin{aligned}
& I(U_k; \widetilde{Y}_k, \widetilde{L}_k | U_{k-1}) \\
&= I(U_k; X(1:\widetilde{L}_k), \widetilde{L}_k | U_{k-1}) \\
&= I(U_k; \widetilde{L}_k | U_{k-1}) + I(U_k; X(1:\widetilde{L}_k), | U_{k-1}, \widetilde{L}_k) \\
&\overset{(a)}{=} \sum_{j=1}^{B} \mathbb{P}_{\widetilde{L}_k}(j) I(U_k; X(1:j) | U_{k-1}, \widetilde{L}_k = j) \\
&= \sum_{j=1}^{B} \left[\mathbb{P}_{\widetilde{L}_k}(j) \sum_{\ell=1}^{j} I(U_k; X(\ell) | X(1:\ell-1), U_{k-1}) \right] \\
&= \sum_{\ell=1}^{B} \left[I\left(U_k; X(\ell) | X(1:\ell-1), U_{k-1}\right) \cdot \sum_{j=\ell}^{B} \mathbb{P}_{\widetilde{L}_k}(j) \right] \\
&= \textstyle\sum_{\ell=1}^{B} \overline{F}_{\widetilde{L}_k}(\ell) I(U_k; X(\ell) | X(1{:}\ell-1) | U_{k-1}) \\
&= \sum_{\ell=1}^{B} \overline{F}_{\widetilde{L}_k}(\ell) \left[H\left(X(\ell) | X(1{:}\ell-1), U_{k-1}\right) - H\left(X(\ell) | X(1{:}\ell-1), U_{k-1}, U_k\right) \right] \\
&\overset{(b)}{=} \sum_{\ell=1}^{B} \overline{F}_{\widetilde{L}_k}(\ell) \left[H(X(\ell) | X(1{:}\ell-1), U_{k-1}) - H(X(\ell) | X(1{:}\ell-1), U_k) \right] \\
&= \sum_{\ell=1}^{B} \overline{F}_{\widetilde{L}_k}(\ell) Q_{\ell,k}, \qquad (2.48)
\end{aligned}
$$

where (a) follows from the fact that $\widetilde{L}_k$ is independent from U_{k-1} and U_k, (b) holds due to $U_{k-1} \leftrightarrow U_k \leftrightarrow X(\ell)$ and $Q_{\ell,k} \coloneqq H(X(\ell) | X(1{:}\ell-1), U_{k-1}) - H(X(\ell) | X(1{:}\ell-1), U_k)$. Therefore, from Lemma 2.5 we have

$$
\begin{aligned}
f(\mathfrak{C}, \boldsymbol{d}) \cdot \left(1 - \mu_{[k]}\right) &\le I(U_k; \widetilde{Y}_k, \widetilde{L}_k | U_{k-1}) \\
&= \sum_{\ell=1}^{B} \overline{F}_{\widetilde{L}_k}(\ell) Q_{\ell,k}, \qquad (2.49)
\end{aligned}
$$

for $k \in [K]$. Taking a weighted sum of (2.49) with coefficients $\{\omega_k\}$, we arrive at

$$f(\mathfrak{C}, \boldsymbol{d}) \sum_{k=1}^{K} \omega_k \left(1 - \mu_{[k]}\right) \leq \sum_{k=1}^{K} \sum_{\ell=1}^{B} \omega_k \overline{F}_{\widetilde{L}_k}(\ell) \cdot Q_{\ell,k}$$
$$= \sum_{\ell=1}^{B} \sum_{k=1}^{K} \omega_k \overline{F}_{\widetilde{L}_k}(\ell) \cdot Q_{\ell,k}. \tag{2.50}$$

Note that for each $\ell \in [B]$ we have

$$\begin{aligned} \sum_{k=1}^{K} Q_{\ell,k} &= \sum_{k=1}^{K} [H(X(\ell)|X(1{:}\ell{-}1), U_{k-1}) - H(X(\ell)|X(1{:}\ell{-}1), U_k)] \\ &= H\left(X(\ell)|X(1{:}\ell{-}1), U_0\right) - H\left(X(\ell)|X(1{:}\ell{-}1), U_K\right) \\ &\leq H(X(\ell)) \leq 1, \end{aligned} \tag{2.51}$$

where we define $U_0 = \varnothing$ as a dummy variable and $U_K = X$. Therefore, using (2.51) for each $\ell \in [B]$ we can write

$$\sum_{k=1}^{K} \omega_k \overline{F}_{\widetilde{L}_k}(\ell) \cdot Q_{\ell,k} \leq \left(\max_k \omega_k \overline{F}_{\widetilde{L}_k}(\ell)\right) \cdot \sum_{k=1}^{K} Q_{\ell,k}$$
$$\leq \max_k \omega_k \overline{F}_{\widetilde{L}_k}(\ell). \tag{2.52}$$

Thus, plugging (2.52) into (2.50) we get

$$f(\mathfrak{C}, \boldsymbol{d}) \sum_{k=1}^{K} \omega_k \sum_{k=1}^{K} \omega_k \left(1 - \mu_{[k]}\right) \leq \max_k \sum_{\ell=1}^{B} \max_k \omega_k \overline{F}_{\widetilde{L}_k}(\ell)$$
$$= \sum_{\ell=1}^{B} \max_k \omega_k \overline{F}_{L_k}(\ell), \tag{2.53}$$

where the last equality follows from Lemma 2.3-(iv). Dividing both sides of (2.53) by $\sum_{k=1}^{K} \omega_k \left(1 - \mu_{[k]}\right)$, we arrive at

$$f(\mathfrak{C}, \boldsymbol{d}) \leq \frac{\sum_{\ell=1}^{B} \max_k \omega_k \overline{F}_{L_k}(\ell)}{\sum_{k=1}^{K} \omega_k \left(1 - \mu_{[k]}\right)}. \tag{2.54}$$

Now, we examine the upper bound for an arbitrary $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k) \in [0, \infty)^K$. Let π be a permutation that sorts the vector $\boldsymbol{\omega}$ in a non-increasing order, i.e., $\omega_{\pi(1)} \geq \cdots \geq \omega_{\pi(K)}$. Now, applying the enhancement in (2.43), we arrive at a set of (statistically) degraded channels,

$$X \leftrightarrow \widetilde{Y}_{\pi(K)} \leftrightarrow \widetilde{Y}_{\pi(K-1)} \leftrightarrow \cdots \leftrightarrow \widetilde{Y}_{\pi(2)} \leftrightarrow \widetilde{Y}_{\pi(1)}.$$

Repeating the argument above, we arrive at (2.54) for a permuted version of the variables, that is,

$$f(\mathfrak{C}, \boldsymbol{d}) \leq \frac{\sum_{\ell=1}^{B} \max_k \omega_{\pi(k)} \overline{F}_{L_{\pi(k)}}(\ell)}{\sum_{k=1}^{K} \omega_{\pi(k)} \big(1 - \mu_{\pi([k])}\big)}. \tag{2.55}$$

By minimizing the right hand side of (2.55) over all non-negative vectors $\boldsymbol{\omega}$, we get the desired bound, i.e.,

$$\begin{aligned} f(\mathfrak{C}, \boldsymbol{d}) &\leq f^{\star}(\mathfrak{C}, \boldsymbol{d}) \\ &= \min_{\boldsymbol{\omega} \geq 0} \frac{\sum_{\ell=1}^{B} \max_k \omega_{\pi(k)} \overline{F}_{L_{\pi(k)}}(\ell)}{\sum_{k=1}^{K} \omega_{\pi(k)} \big(1 - \mu_{\pi([k])}\big)}. \end{aligned} \tag{2.56}$$

This completes the proof of the theorem. □

2.7.3 An LP Representation

The main step in the proof of Proposition 2.1 is to define a new weight vector $\boldsymbol{\sigma}$, which allows us to transform the upper-bound in (2.56) into a linear form.

Proof of Proposition 2.1. Using $\omega_{\pi(1)} \geq \cdots \geq \omega_{\pi(K)} \geq 0$ and starting from (2.56), we can write

$$\begin{aligned} f^{\star}(\mathfrak{C}, \boldsymbol{d}) &= \min_{\boldsymbol{\omega} \geq 0} \frac{\sum_{\ell=1}^{B} \max_k \omega_{\pi(k)} \overline{F}_{L_{\pi(k)}}(\ell)}{\sum_{k=1}^{K} \omega_{\pi(k)} \big(1 - \mu_{\pi([k])}\big)} \\ &= \min_{\pi \in \Pi} \min_{\boldsymbol{\omega} \in \Omega_{\pi}} \frac{\sum_{\ell=1}^{B} \max_k \omega_{\pi(k)} \overline{F}_{L_{\pi(k)}}(\ell)}{\sum_{k=1}^{K} \omega_{\pi(k)} \big(1 - \mu_{\pi([k])}\big)}, \end{aligned} \tag{2.57}$$

where $\Omega_{\pi} := \big\{\boldsymbol{\omega} : \omega_{\pi(1)} \geq \cdots \geq \omega_{\pi(K)} \geq 0\big\}$. Now, we fix some $\pi \in \Pi$ and focus on the inner minimization in (2.57), i.e.,

$$f_{\pi}^{\star}(\mathfrak{C}, \boldsymbol{d}) = \min_{\boldsymbol{\omega} \in \Omega_{\pi}} \frac{\sum_{\ell=1}^{B} \max_k \omega_{\pi(k)} \overline{F}_{L_{\pi(k)}}(\ell)}{\sum_{k=1}^{K} \omega_{\pi(k)} \big(1 - \mu_{\pi([k])}\big)}. \tag{2.58}$$

We define $\sigma_k := \omega_{\pi(k)} \big(1 - \mu_{\pi([k])}\big) / \sum_{u=1}^{K} \omega_{\pi(u)} \big(1 - \mu_{\pi([u])}\big)$ for every $k \in [K]$. We note that the vector $\boldsymbol{\sigma} := (\sigma_1, \ldots, \sigma_K)$ satisfies the following conditions:

(a) Since $\omega_{\pi(k)} \geq 0$ and $\mu_{\pi([k])} \leq 1$ for every $k \in [K]$, we have $\sigma_k \geq 0$ for every $k \in [K]$;

(b) We have

$$\sum_{k=1}^{K} \sigma_k = \sum_{k=1}^{K} \frac{\omega_{\pi(k)} \left(1 - \mu_{\pi([k])}\right)}{\sum_{u=1}^{K} \omega_{\pi(u)} \left(1 - \mu_{\pi([u])}\right)} = 1;$$

(c) Using $\omega_{\pi(k-1)} \geq \omega_{\pi(k)}$, we get

$$\frac{\sigma_{k-1}}{1 - \mu_{\pi([k-1]),}} \geq \frac{\sigma_k}{1 - \mu_{\pi([k])}},$$

or equivalently,

$$\sigma_{k-1} \cdot \left(1 - \mu_{\pi([k])}\right) \geq \sigma_k \cdot \left(1 - \mu_{\pi([k-1])}\right),$$

for every $k \in \{2, \cdots, K\}$.

We define Σ_π as the set of all vectors $\boldsymbol{\sigma}$ satisfying three conditions in (a)-(c).

Note that for every vector $\boldsymbol{\omega} \in \Omega_\pi$, there is a vector $\boldsymbol{\sigma} \in \Sigma_\pi$ and vice versa. Applying this change of variables in (2.57), we arrive at

$$\begin{aligned} f_\pi^\star(\mathfrak{C}, \boldsymbol{d}) &= \min_{\boldsymbol{\omega} \in \Omega_\pi} \sum_{\ell=1}^{B} \max_k \frac{\omega_{\pi(k)} \left(1 - \mu_{\pi([k])}\right)}{\sum_{u=1}^{K} \omega_{\pi(u)} \left(1 - \mu_{\pi([u])}\right)} \frac{\overline{F}_{L_{\pi(k)}}(\ell)}{1 - \mu_{\pi([k])}} \\ &= \min_{\boldsymbol{\sigma} \in \Sigma_\pi} \sum_{\ell=1}^{B} \max_k \sigma_k \frac{\overline{F}_{L_{\pi(k)}}(\ell)}{1 - \mu_{\pi([k])}}. \end{aligned} \tag{2.59}$$

Let us consider each summand in (2.59). For a given $\pi \in \Pi$ and $\boldsymbol{\sigma} \in \Sigma_\pi$, the ℓth term in the summation is

$$\max_k \sigma_k \frac{\overline{F}_{L_{\pi(k)}}(\ell)}{1 - \mu_{\pi([k])}} = \min \left\{ \theta_\ell : \theta_\ell \geq \sigma_k \frac{\overline{F}_{L_{\pi(k)}}(\ell)}{1 - \mu_{\pi([k])}}, k \in [K] \right\}. \tag{2.60}$$

Let us define

$$\Theta_\pi^{\boldsymbol{\sigma}} := \left\{ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_B) : \theta_\ell \geq \sigma_k \frac{\overline{F}_{L_{\pi(k)}}(\ell)}{1 - \mu_{\pi([k])}}, k \in [K] \right\}. \tag{2.61}$$

Then, (2.60) can be written as

$$\max_k \sigma_k \frac{\overline{F}_{L_{\pi(k)}}(\ell)}{1 - \mu_{\pi([k])}} = \min_{\boldsymbol{\theta} \in \Theta_\pi^{\boldsymbol{\sigma}}} \theta_\ell. \tag{2.62}$$

Note that the conditions on each θ_ℓ in (2.61) only depend on $\{\overline{F}_{L_{\pi(k)}}(\ell)\}_k$, and hence, for $\ell \neq \ell'$, the conditions on θ_ℓ and $\theta_{\ell'}$ are independent of each other. In other words, $\Theta_\pi^{\boldsymbol{\sigma}}$ is an orthant with an offset in $\mathbb{R}^B$. Hence, the minimum of the summation of $\{\theta_\ell\}_{\ell=1}^B$ and the summation of the minimum of $\{\theta_\ell\}_{\ell=1}^B$ are equivalent, i.e.,

$$\sum_{\ell=1}^{B} \min_{\boldsymbol{\theta}\in\Theta_\pi^{\boldsymbol{\sigma}}} \theta_\ell = \min_{\boldsymbol{\theta}\in\Theta_\pi^{\boldsymbol{\sigma}}} \sum_{\ell=1}^{B} \theta_\ell. \tag{2.63}$$

Combining (2.62) and (2.63), we arrive at

$$\sum_{\ell=1}^{B} \max_k \sigma_k \frac{\overline{F}_{L_{\pi(k)}}(\ell)}{1-\mu_{\pi([k])}} = \min_{\boldsymbol{\theta}\in\Theta_\pi^{\boldsymbol{\sigma}}} \sum_{\ell=1}^{B} \theta_\ell. \tag{2.64}$$

Plugging (2.64) into (2.59), we get

$$f_\pi^\star(\mathfrak{C}, \boldsymbol{d}) = \min_{\boldsymbol{\sigma}\in\Sigma_\pi} \min_{\boldsymbol{\theta}\in\Theta_\pi^{\boldsymbol{\sigma}}} \sum_{\ell=1}^{B} \theta_\ell. \tag{2.65}$$

Let $\mathbf{x} \coloneqq [\boldsymbol{\sigma}, \boldsymbol{\theta}]$. It is important to note that the objective function and the constraints on the optimization problem (2.65) are linear in $\mathbf{x}$. More precisely, using matrices $\mathbf{A}$ and $\mathbf{b}$ defined in and (2.15) in (2.16), conditions (a), (b), and (c) on vector $\boldsymbol{\sigma} \in \Sigma_\pi$ can be translated into $-\mathbf{x} \leq 0$, $\mathbf{b}\mathbf{x} = 1$, and the lower $K-1$ rows of $\mathbf{A}_\pi \mathbf{x} \leq 0$, respectively. Moreover, the constraints on $\boldsymbol{\theta} \in \Theta_\pi^{\boldsymbol{\sigma}}$ in (2.61) can be expressed as the top KB rows of $\mathbf{A}_\pi \mathbf{x} \leq 0$. Therefore, we can rewrite (2.65) as

$$\begin{aligned} f^\star(\mathfrak{C}, \boldsymbol{d}) = \min_{\mathbf{x}\in\mathbb{R}^{K+B}} \; & [\mathbf{0}_K^T, \mathbf{1}_B^T]\mathbf{x} \\ \text{s.t.} \; & \mathbf{A}_\pi \mathbf{x} \leq \mathbf{0}, \\ & \mathbf{b}\mathbf{x} = 1, \\ & -\mathbf{x} \leq \mathbf{0}. \end{aligned}$$

This completes the proof of the proposition. □

2.8 Converse Proof of Theorem 2.2

The converse proof of Theorem 2.2 is derived directly from the proof of Theorem 2.3 where no channel enhancement is required, i.e., $\widetilde{L}_k = L_k$ for every $k \in [K]$. Since $L_K \geq_{\mathsf{st}} \cdots \geq_{\mathsf{st}} L_1$ the K-DTVBC is degraded, we can repeat the steps (2.48) in through (2.49)

with no further channel enhancement, i.e., $\overline{F}_{\widetilde{L}_k}(\ell) = \overline{F}_{L_k}(\ell)$ for every $k \in [K]$ and $\ell \in [B]$. Hence, for the cache placement strategy $\mathfrak{C}^{\mathsf{cent}}$ and its caching tuple $\boldsymbol{\mu}^{\mathsf{cent}}$ we can write

$$f(\mathfrak{C}, \boldsymbol{d}) \cdot \left(1 - \mu_{[k]}^{\mathsf{cent}}\right) \leq \sum_{\ell=1}^{B} \overline{F}_{L_k}(\ell) Q_{\ell,k}, \quad \forall k \in [K], \tag{2.66}$$

where

$$Q_{\ell,k} = H(X(\ell)|X(1:\ell-1), U_{k-1}) - H(X(\ell)|X(1:\ell-1), U_k)$$

for the Markov chain $U_{k-1} \leftrightarrow U_k \leftrightarrow X(\ell)$. It is easy to verify that $Q_{\ell,k} \geq 0$. Moreover, from (2.51), we have

$$\sum_{k=1}^{K} Q_{\ell,k} \leq 1, \quad \forall \ell \in [B]. \tag{2.67}$$

From (2.66) and (2.67), we can write

$$\begin{aligned} f(\mathfrak{C}, \boldsymbol{d}) \leq \max \ & \bar{f} \\ \text{s.t.} \quad & \bar{f} \cdot \left(1 - \mu_{[k]}^{\mathsf{cent}}\right) \leq \sum_{\ell=1}^{B} \overline{F}_{L_k}(\ell) Q_{\ell,k}, \quad \forall k \in [K], \\ & \sum_{k=1}^{K} Q_{\ell,k} \leq 1, \quad \forall \ell \in [B]. \end{aligned} \tag{2.68}$$

Noting the LP problems in (2.10) with constraints (2.11) and (2.68) are equivalent, we arrive at the claim of Theorem 2.2. This completes the proof of the theorem. □

2.9 Concluding Remarks

In this work, we studied a K-user coded-caching problem in a joint source-channel coding framework by providing each user a cache. The transmitter has a certain rate for all files per channel use, and a fraction of the bits/symbols are available in each user's cache. After this, each user requests a file from the database where the transmitter needs to satisfy users' demands over the K-DTVBC. The receivers have only access to the channel state information. We characterized the maximum achievable source rate for the 2-DTVBC and the degraded K-DTVBC. Then, we provided an upper bound for the source rate with any caching strategy $\mathfrak{C}$. Finally, we presented an achievable scheme with the LP formulation to show that the upper bound is not a sharp characterization.

2.10 Proof of Auxiliary Lemmas

In this section, we provide the proofs of lemmas. Note that, in order to avoid repetition, we provide the proof of Lemma 2.1 after the proof of Lemma 2.5, since most of the techniques used in the latter are also applied in the former.

Proof of Lemma 2.2. The first part of the lemma is an immediate consequence of the level-allocation in (2.35) and (2.36). Before we prove the other claims of the lemma, consider the LHS of (2.37) and note that the first term in the minimization increases with respect to both u and α, while the second term decreases with u and α. Hence, the minimum of two terms is maximized when two terms are equal. Thus, we can write

$$f_1^\star = \frac{1}{1-2\mu}\left[\sum_{i<u^\star} \overline{F}_{L_1}(\ell_i) + \alpha^\star \overline{F}_{L_1}(\ell_{u^\star})\right] = \frac{1}{1-\mu}\left[\sum_{i>u^\star} \overline{F}_{L_2}(\ell_i) + (1-\alpha^\star)\overline{F}_{L_2}(\ell_{u^\star})\right]. \tag{2.69}$$

Similarly, for (2.38) we get

$$f_2^\star = \frac{1}{1-\mu}\left[\sum_{i<v^\star} \overline{F}_{L_1}(\ell_i) + (1-\beta^\star)\overline{F}_{L_1}(\ell_{v^\star})\right] = \frac{1}{1-2\mu}\left[\sum_{i>v^\star} \overline{F}_{L_2}(\ell_i) + \beta^\star \overline{F}_{L_2}(\ell_{v^\star})\right]. \tag{2.70}$$

Now, to prove the second and the third claims of the lemma, we can distinguish two cases, depending on whether $f_1^\star \le f_2^\star$ or $f_1^\star \ge f_2^\star$. Let us start with the first case. Since $\mu > 0$, we can write

$$\begin{aligned}
\sum_{i\le v^\star} \overline{F}_{L_1}(\ell_i) &> \frac{1-2\mu}{1-\mu}\left[\sum_{i\le v^\star} \overline{F}_{L_1}(\ell_i)\right] \\
&\overset{(a)}{\ge} \frac{1-2\mu}{1-\mu}\left[\sum_{i<v^\star} \overline{F}_{L_1}(\ell_i) + (1-\beta^\star)\overline{F}_{L_1}(\ell_{v^\star})\right] \\
&\overset{(b)}{\ge} (1-2\mu) f_2^\star \\
&\overset{(c)}{\ge} (1-2\mu) f_1^\star \overset{(d)}{=} \sum_{i<u^\star} \overline{F}_{L_1}(\ell_i) + \alpha^\star \overline{F}_{L_1}(\ell_{u^\star}) \overset{(e)}{\ge} \sum_{i<u^\star} \overline{F}_{L_1}(\ell_i),
\end{aligned} \tag{2.71}$$

where (a) holds since $\beta^\star \le 1$, (b) follows from the first equality in (2.70), (c) is due to assuming $f_1^\star \le f_2^\star$, (d) follows from the first equality in (2.69), and (e) holds since $\alpha^\star \ge 0$. Then, (2.71) implies that $v^\star \ge u^\star$. For the second case with $f_1^\star \ge f_2^\star$ and $\mu > 0$,

we can write

$$\begin{aligned}\sum_{i\geq u^\star}\overline{F}_{L_2}(\ell_i) &> \frac{1-2\mu}{1-\mu}\left[\sum_{i\geq u^\star}\overline{F}_{L_2}(\ell_i)\right]\\ &\overset{(a)}{\geq} \frac{1-2\mu}{1-\mu}\left[\sum_{i> u^\star}\overline{F}_{L_2}(\ell_i)+(1-\alpha^\star)\overline{F}_{L_2}(\ell_{u^\star})\right]\\ &\overset{(b)}{=} (1-2\mu)f_1^\star\\ &\overset{(c)}{\geq} (1-2\mu)f_2^\star \overset{(d)}{=} \sum_{i>v^\star}\overline{F}_{L_2}(\ell_i)+\beta^\star\overline{F}_{L_2}(\ell_{u^\star}) \overset{(e)}{\geq} \sum_{i>v^\star}\overline{F}_{L_2}(\ell_i),\end{aligned} \tag{2.72}$$

where (a) holds for $\alpha^\star \leq 1$, (b) follows from the second equality in (2.69), (c) is true since we assumed $f_1^\star \geq f_2^\star$, (d) follows from the second equality in (2.70), and (e) holds for $\beta^\star > 0$. From (2.72), it can be observed that $v^\star \geq u^\star$, as claimed in part (ii) of the lemma.

We prove the third claim assuming that $f_1^\star \leq f_2^\star$. Note that the proof for the other case is very similar. The proof is by contradiction. Assume $g(\ell_{u^\star}) > 1$. Then, since the levels are sorted with respect to $g(\cdot)$, we have $1 < g(\ell_{u^\star}) \leq \cdots \leq g(\ell_B)$, which implies $\overline{F}_{L_2}(\ell_i) > \overline{F}_{L_1}(\ell_i)$ for every $i > u^\star$. Thus, we have

$$\begin{aligned}&(1-\alpha^\star)\overline{F}_{L_2}(\ell_{u^\star})+\sum_{u^\star<i<v^\star}\overline{F}_{L_2}(\ell_i)+\beta^\star\overline{F}_{L_2}(\ell_{v^\star})+(1-\beta^\star)\overline{F}_{L_2}(\ell_{v^\star})+\sum_{i>v^\star}\overline{F}_{L_2}(\ell_i)\\ &= (1-\alpha^\star)\overline{F}_{L_2}(\ell_{u^\star})+\sum_{i>u^\star}\overline{F}_{L_2}(\ell_i)\\ &\overset{(a)}{=} (1-\mu)f_1^\star\\ &\overset{(b)}{\leq} (1-\mu)f_2^\star\\ &\overset{(c)}{=} \sum_{i<v^\star}\overline{F}_{L_1}(\ell_i)+(1-\beta^\star)\overline{F}_{L_1}(\ell_{v^\star})\\ &= \sum_{i<u^\star}\overline{F}_{L_1}(\ell_i)+\alpha^\star\overline{F}_{L_1}(\ell_{u^\star})+(1-\alpha^\star)\overline{F}_{L_1}(\ell_{u^\star})+\sum_{u^\star<i<v^\star}\overline{F}_{L_1}(\ell_i)+(1-\beta^\star)\overline{F}_{L_1}(\ell_{v^\star}),\end{aligned} \tag{2.73}$$

where (a) follows from the second equality in (2.69), (b) holds as $f_1^\star \leq f_2^\star$, and step (c) follows from the first equality in (2.70). Subtracting the RHS of (2.73) from its LHS,

we arrive at

$$
\begin{aligned}
0 \le & \left[\sum_{i<u^\star} \overline{F}_{L_1}(\ell_i) + \alpha^\star \overline{F}_{L_1}(\ell_{u^\star}) + (1-\alpha^\star)\overline{F}_{L_1}(\ell_{u^\star}) \right. \\
& \left. + \sum_{u^\star<i<v^\star} \overline{F}_{L_1}(\ell_i) + (1-\beta^\star)\overline{F}_{L_1}(\ell_{v^\star}) \right] \\
& - \left[(1-\alpha^\star)\overline{F}_{L_2}(\ell_{u^\star}) + \sum_{u^\star<i<v^\star} \overline{F}_{L_2}(\ell_i) \right. \\
& \left. + \beta^\star \overline{F}_{L_2}(\ell_{v^\star}) + (1-\beta^\star)\overline{F}_{L_2}(\ell_{v^\star}) + \sum_{i>v^\star} \overline{F}_{L_2}(\ell_i) \right] \qquad (2.74) \\
= & \left[\sum_{i<u^\star} \overline{F}_{L_1}(\ell_i) + \alpha^\star \overline{F}_{L_1}(\ell_{u^\star}) \right] \\
& - \left[\sum_{i>v^\star} \overline{F}_{L_2}(\ell_i) + \beta^\star \overline{F}_{L_2}(\ell_{v^\star}) \right] \\
& + \left[(1-\alpha^\star)\left(\overline{F}_{L_1}(\ell_{u^\star}) - \overline{F}_{L_2}(\ell_{u^\star}) \right) \right] \\
& + \left[(1-\beta^\star)\left(\overline{F}_{L_1}(\ell_{v^\star}) - \overline{F}_{L_2}(\ell_{v^\star}) \right) \right] \\
& + \left[\sum_{u^\star<i<v^\star} \left(\overline{F}_{L_1}(\ell_i) - \overline{F}_{L_2}(\ell_i) \right) \right] \\
\overset{(a)}{<} & \left[\sum_{i<u^\star} \overline{F}_{L_1}(\ell_i) + \alpha^\star \overline{F}_{L_1}(\ell_{u^\star}) \right] - \left[\sum_{i>v^\star} \overline{F}_{L_2}(\ell_i) + \beta^\star \overline{F}_{L_2}(\ell_{v^\star}) \right] \\
\overset{(b)}{=} & (1-2\mu) f_1^\star - (1-2\mu) f_2^\star = (1-2\mu)(f_1^\star - f_2^\star), \qquad (2.75)
\end{aligned}
$$

where (a) holds since $\overline{F}_{L_1}(\ell_i) < \overline{F}_{L_2}(\ell_i)$ for $i > u^\star$, and (b) follows from the second equalities in (2.69) and (2.70). Then, (2.74) implies that $f_1^\star > f_2^\star$, which is in contradiction with the assumption that $f_1^\star \le f_2^\star$. Hence, we can conclude that $g(\ell_{u^\star}) \le 1$. This completes the proof of part (iii).

Finally, we can prove part (iv) of the lemma. First, assume that $f_1^\star \le f_2^\star$. From (2.69),

we can write

$$\frac{1}{g(\ell_{u^\star})}(1-\mu)f_1^\star = \frac{1}{g(\ell_{u^\star})}\left[\sum_{i>u^\star}\overline{F}_{L_2}(\ell_i)+(1-\alpha^\star)\overline{F}_{L_2}(\ell_{u^\star})\right]$$
$$\overset{(a)}{=} \frac{1}{g(\ell_{u^\star})}\sum_{i>u^\star}\overline{F}_{L_2}(\ell_i)+(1-\alpha^\star)\overline{F}_{L_1}(\ell_{u^\star}),$$

where in (a) follow from $g(\ell_{u^\star}) = \overline{F}_{L_2}(\ell_{u^\star})/\overline{F}_{L_1}(\ell_{u^\star})$. Moreover, from (2.69) we have

$$(1-2\mu)f_1^\star = \sum_{i<u^\star}\overline{F}_{L_1}(\ell_i) + \alpha^\star\overline{F}_{L_1}(\ell_{u^\star}).$$

Combining these two equations, we arrive at

$$\begin{aligned} f_1^\star &= \frac{\sum_{i\le u^\star}\overline{F}_{L_1}(\ell_i) + \frac{1}{g(\ell_{u^\star})}\sum_{i>u^\star}\overline{F}_{L_2}(\ell_i)}{(1-2\mu)+\frac{1}{g(\ell_{u^\star})}(1-\mu)} \\ &\overset{(a)}{=} \frac{R_1(g(\ell_{u^\star})) + \frac{1}{g(\ell_{u^\star})}R_2(g(\ell_{u^\star}))}{(1-2\mu)+\frac{1}{g(\ell_{u^\star})}(1-\mu)} \\ &= \left.\frac{R_1(\omega)+\frac{1}{\omega)}R_2(\omega)}{(1-2\mu)+\frac{1}{\omega}(1-\mu)}\right|_{\omega=g(\ell_{u^\star})} \\ &\overset{(b)}{\ge} \min_{\omega\le 1}\frac{R_1(\omega)+\frac{1}{\omega)}R_2(\omega)}{(1-2\mu)+\frac{1}{\omega}(1-\mu)}, \end{aligned} \tag{2.76}$$

where in (a) we have $R_1(g(\ell_{u^\star}))$ and $R_1(g(\ell_{u^\star}))$ as given in (2.9). Moreover,

$$\begin{aligned} \mathcal{L}_1(g(\ell_{u^\star})) &= \{\ell : g(\ell_{u^\star})\overline{F}_{L_1}(\ell) \ge \overline{F}_{L_2}(\ell)\} \\ &= \{\ell : g(\ell_{u^\star}) \ge g(\ell)\} = \{i : i \le u^\star\}, \end{aligned}$$

and $\mathcal{L}_2(g(\ell_{u^\star})) = \{i : i > u^\star\}$ as indicated in the statement of Theorem 2.1. Also, note that from part (iii), we have $g(\ell_{u^\star}) \le 1$, which justifies the inequality in (b).

Now, we consider the case that $f_2^\star \le f_1^\star$. From (2.70), we can write

$$g(\ell_{v^\star})(1-\mu)f_2^\star = g(\ell_{v^\star})\left[\sum_{i<v^\star}\overline{F}_{L_1}(\ell_i)+(1-\beta^\star)\overline{F}_{L_1}(\ell_{v^\star})\right].$$

Further, we get

$$\begin{aligned} (1-2\mu)f_2^\star &= \sum_{i>v^\star}\overline{F}_{L_2}(\ell_i) + \beta^\star\overline{F}_{L_2}(\ell_{v^\star}) \\ &\overset{(a)}{=} \sum_{i>v^\star}\overline{F}_{L_2}(\ell_i) + g(\ell_{v^\star})\beta^\star\overline{F}_{L_1}(\ell_{v^\star}), \end{aligned}$$

where (a) follows from $g(\ell_{v^\star}) = \overline{F}_{L_2}(\ell_{v^\star})/\overline{F}_{L_1}(\ell_{v^\star})$. Combining these two equations, we have

$$\begin{aligned} f_2^\star &= \frac{g(\ell_{v^\star})\sum_{i\le v^\star}\overline{F}_{L_1}(\ell_i) + \sum_{i>v^\star}\overline{F}_{L_2}(\ell_i)}{g(\ell_{v^\star})(1-\mu)+(1-2\mu)} \\ &= \frac{g(\ell_{v^\star})R_1(g(\ell_{v^\star})) + R_2(g(\ell_{v^\star}))}{g(\ell_{v^\star})(1-\mu)+(1-2\mu)} \\ &= \left.\frac{\omega R_1(x\omega)+R_2(\omega)}{\omega(1-\mu)+(1-2\mu)}\right|_{\omega=g(\ell_{v^\star})} \\ &\ge \min_{0\le\omega\le 1}\frac{\omega R_1(x\omega)+R_2(\omega)}{\omega(1-\mu)+(1-2\mu)}. \end{aligned} \tag{2.77}$$

Here, $R_1(\omega)$ and $R_2(\omega)$ are defined as in (2.9). Moreover, we have

$$\begin{aligned} \mathcal{L}_1(g(\ell_{v^\star})) &= \{\ell : g(\ell_{v^\star})\overline{F}_{L_1}(\ell) \ge \overline{F}_{L_2}(\ell)\} \\ &= \{\ell : g(\ell_{v^\star}) \ge g(\ell)\} = \{i : i \le v^\star\}, \end{aligned}$$

and $\mathcal{L}_1(g(\ell_{v^\star})) = \{i : i > v^\star\}$. It is worth noting that, the last inequality in (2.77) holds since $g(\ell_{v^\star}) \ge 1$, as shown in part (iii) of the lemma.

Combining (2.76) and (2.77), we arrive at $\min(f_1^\star, f_2^\star) \ge f^\star$, for $f^\star$ defined in (2.7). This completes the proof of part (iv). □

Proof of Lemma 2.3. In order to prove (i), we show that $\overline{F}_{\widetilde{L}_k}(\ell) = 1$ implies $\overline{F}_{\widetilde{L}_{k+1}}(\ell) = 1$. From (2.43), we have

$$\begin{aligned} \overline{F}_{\widetilde{L}_{k+1}}(\ell) &= \min\left[1, \max\left(\overline{F}_{L_{k+1}}(\ell), \frac{\omega_k}{\omega_{k+1}}\overline{F}_{\widetilde{L}_k}(\ell)\right)\right] \\ &= \min\left[1, \max\left(\overline{F}_{L_{k+1}}(\ell), \frac{\omega_k}{\omega_{k+1}}\right)\right] \\ &\overset{(a)}{=} \min\left[1, \frac{\omega_k}{\omega_{k+1}}\right] \overset{(b)}{=} 1, \end{aligned}$$

where (a) and (b) hold because $\overline{F}_{L_{k+1}}(\ell) \le 1 \le \frac{\omega_k}{\omega_{k+1}}$, for a non-increasing sequence $\omega_1 \ge \omega_2 \ge \cdots \ge \omega_K$. This implies that $\overline{F}_{\widetilde{L}_u}(\ell) = 1$ for every $u \ge k$.

Next, we prove part (ii). First, assume that $\omega_k\overline{F}_{L_k}(\ell) < \omega_{k-1}\overline{F}_{\widetilde{L}_{k-1}}(\ell)$. Then, we can write

$$\begin{aligned} \overline{F}_{\widetilde{L}_k}(\ell) &= \min\left[1, \max\left(\overline{F}_{L_k}(\ell), \frac{\omega_{k-1}}{\omega_k}\overline{F}_{\widetilde{L}_{k-1}}(\ell)\right)\right] \\ &\le \max\left(\overline{F}_{L_k}(\ell), \frac{\omega_{k-1}}{\omega_k}\overline{F}_{\widetilde{L}_{k-1}}(\ell)\right) = \frac{\omega_{k-1}}{\omega_k}\overline{F}_{\widetilde{L}_{k-1}}(\ell), \end{aligned}$$

which implies $\omega_k \overline{F}_{\widetilde{L}_k}(\ell) \le \omega_{k-1}\overline{F}_{\widetilde{L}_{k-1}}(\ell)$, which is in contradiction with the assumption of part (ii). Hence, we have $\omega_k \overline{F}_{L_k}(\ell) \ge \omega_{k-1}\overline{F}_{\widetilde{L}_{k-1}}(\ell)$. This together with (2.43) leads to

$$\begin{aligned}\overline{F}_{\widetilde{L}_k}(\ell) &= \min\left[1, \max\left(\overline{F}_{L_k}(\ell), \frac{\omega_{k-1}}{\omega_k}\overline{F}_{\widetilde{L}_{k-1}}(\ell)\right)\right] \\ &= \min\left[1, \overline{F}_{L_k}(\ell)\right] = \overline{F}_{L_k}(\ell),\end{aligned}$$

where the last equality follows from $\overline{F}_{L_k}(\ell) \le 1$. This shows the first equality in part (ii).

Then, assume $\overline{F}_{\widetilde{L}_u}(\ell) = 1$ for some $u < k$. From part (i), we get

$$\overline{F}_{\widetilde{L}_u}(\ell) = \cdots = \overline{F}_{\widetilde{L}_{k-1}}(\ell) = \overline{F}_{\widetilde{L}_k}(\ell) = 1.$$

This together with the fact that $\omega_{k-1} \ge \omega_k$ implies $\omega_{k-1}\overline{F}_{\widetilde{L}_{k-1}}(\ell) \ge \omega_k \overline{F}_{\widetilde{L}_k}(\ell)$ which contradicts with the assumption of part (ii). Hence, we have $\overline{F}_{\widetilde{L}_u}(\ell) < 1$ for every $u < k$. Using this fact, we get

$$\begin{aligned}\overline{F}_{\widetilde{L}_u}(\ell) &= \min\left[1, \max\left(\overline{F}_{L_u}(\ell), \frac{\omega_{u-1}}{\omega_u}\overline{F}_{\widetilde{L}_{u-1}}(\ell)\right)\right] \\ &= \max\left(\overline{F}_{L_u}(\ell), \frac{\omega_{u-1}}{\omega_u}\overline{F}_{\widetilde{L}_{u-1}}(\ell)\right) \\ &\ge \frac{\omega_{u-1}}{\omega_u}\overline{F}_{\widetilde{L}_{u-1}}(\ell),\end{aligned}$$

which results in $\omega_u \overline{F}_{\widetilde{L}_u}(\ell) \ge \omega_{u-1}\overline{F}_{\widetilde{L}_{u-1}}(\ell)$ for every $u < k$, or equivalently

$$\omega_k \overline{F}_{\widetilde{L}_k}(\ell) > \omega_{k-1}\overline{F}_{\widetilde{L}_{k-1}}(\ell) \ge \omega_{k-2}\overline{F}_{\widetilde{L}_{k-2}}(\ell) \ge \cdots \ge \omega_1 \overline{F}_{\widetilde{L}_1}(\ell).$$

This completes the proof of part (ii).

In order to prove part (iii), we first note that

$$\begin{aligned}\frac{\omega_{k-1}}{\omega_k}\overline{F}_{\widetilde{L}_{k-1}}(\ell) &\overset{(a)}{>} \overline{F}_{\widetilde{L}_k}(\ell) \\ &= \min\left[1, \max\left(\overline{F}_{L_k}(\ell), \frac{\omega_{k-1}}{\omega_k}\overline{F}_{\widetilde{L}_{k-1}}(\ell)\right)\right] \\ &\overset{(b)}{=} \min\left[1, \frac{\omega_{k-1}}{\omega_k}\overline{F}_{\widetilde{L}_{k-1}}(\ell)\right],\end{aligned} \tag{2.78}$$

where both (a) and (b) follow from the assumption of part (iii). Then, (2.78) implies $\overline{F}_{\widetilde{L}_k}(\ell) = 1$. This, from part (i) of the lemma, we get

$$\overline{F}_{\widetilde{L}_u}(\ell) = 1, \quad u \ge k.$$

This along with $\omega_1 \geq \cdots \geq \omega_K$ leads to

$$\omega_K \overline{F}_{\widetilde{L}_K}(\ell) \leq \cdots \leq \omega_k \overline{F}_{\widetilde{L}_k}(\ell) < \omega_{k-1} \overline{F}_{\widetilde{L}_{k-1}}(\ell),$$

which is the claim of part (iii).

Next, we prove part (iv) of the lemma. Fix some $\ell \in [B]$, and define

$$s^\star := \max\left\{j : \omega_j \overline{F}_{\widetilde{L}_j}(\ell) > \omega_{j-1} \overline{F}_{\widetilde{L}_{j-1}}(\ell)\right\}.$$

In the following, we first show that

$$\omega_{s^\star} \overline{F}_{\widetilde{L}_{s^\star}}(\ell) = \max_k \omega_k \overline{F}_{\widetilde{L}_k}(\ell). \tag{2.79}$$

The definition of $s^\star$ implies $\omega_u \overline{F}_{\widetilde{L}_u}(\ell) \leq \omega_{u+1} \overline{F}_{\widetilde{L}_{u+1}}(\ell)$, for every $u > s^\star$, leading to

$$\omega_{s^\star} \overline{F}_{\widetilde{L}_{s^\star}}(\ell) \geq \omega_{s^\star+1} \overline{F}_{\widetilde{L}_{s^\star+1}}(\ell) \geq \cdots \geq \omega_K \overline{F}_{\widetilde{L}_K}(\ell). \tag{2.80}$$

Moreover, Since $\omega_{s^\star} \overline{F}_{\widetilde{L}_{s^\star}}(\ell) > \omega_{s^\star-1} \overline{F}_{\widetilde{L}_{s^\star-1}}(\ell)$, from part (ii) of the lemma we have

$$\omega_1 \overline{F}_{\widetilde{L}_1}(\ell) \leq \cdots \leq \omega_{s^\star-1} \overline{F}_{\widetilde{L}_{s^\star-1}}(\ell) < \omega_{s^\star} \overline{F}_{\widetilde{L}_{s^\star}}(\ell) = \omega_{s^\star} \overline{F}_{L_{s^\star}}(\ell). \tag{2.81}$$

Combining (2.80) and (2.81), we can conclude (2.79). Furthermore, we have

$$\begin{aligned}
\max_k \omega_k \overline{F}_{L_k}(\ell) &\geq \omega_{s^\star} \overline{F}_{L_{s^\star}}(\ell) \\
&\overset{(a)}{=} \omega_{s^\star} \overline{F}_{\widetilde{L}_{s^\star}}(\ell) \\
&= \max_k \omega_k \overline{F}_{\widetilde{L}_k}(\ell) \\
&\overset{(b)}{\geq} \max_k \omega_k \overline{F}_{L_k}(\ell).
\end{aligned} \tag{2.82}$$

Here, (a) follows from (2.81), and (b) is due to the fact that $\omega_k \overline{F}_{\widetilde{L}_k}(\ell) \geq \omega_k \overline{F}_{L_k}(\ell)$ for every $k \in [K]$. Lastly, (2.82) concludes the proof of part (iv). □

Proof of Corollary 2.1. To prove (i), from the definition of $k^\star$, we get

$$\omega_{k^\star} \overline{F}_{\widetilde{L}_{k^\star}}(\ell) > \omega_{k^\star-1} \overline{F}_{\widetilde{L}_{k^\star-1}}(\ell).$$

This together with Lemma 2.3-(ii) for $k = k^\star$ arrives us at

$$\omega_{k^\star-1} \overline{F}_{\widetilde{L}_{k^\star-1}}(\ell) \geq \cdots \geq \omega_1 \overline{F}_{\widetilde{L}_1}(\ell).$$

The part (ii) can be directly derived from the definitions of $k^\star$ and $u^\star$.

In order to prove part (iii), from the definition of $u^\star$, we have

$$\omega_{u^\star}\overline{F}_{\widetilde{L}_{u^\star}}(\ell) > \omega_{u^\star+1}\overline{F}_{\widetilde{L}_{u^\star+1}}(\ell).$$

This combined with Lemma 2.3-(iii) for $k = u^\star + 1$ leads us to

$$\omega_{u^\star+1}\overline{F}_{\widetilde{L}_{u^\star+1}}(\ell) \geq \cdots \geq \omega_K \overline{F}_{\widetilde{L}_K}(\ell),$$

which completes the proof of the corollary □

Proof of Lemma 2.4. We first prove that

$$I\left(W_i^{(n)}; C_{\mathcal{S}}^{(n)}\right) = H\left(C_{\mathcal{S},i}^{(n)}\right).$$

To this end, we show $I\left(W_i^{(n)}; C_{\mathcal{S}}^{(n)}\right) \geq H\left(C_{\mathcal{S},i}^{(n)}\right)$ and $I\left(W_i^{(n)}; C_{\mathcal{S}}^{(n)}\right) \leq H\left(C_{\mathcal{S},i}^{(n)}\right)$. For the first inequality, we can write

$$\begin{aligned} I\left(W_i^{(n)}; C_{\mathcal{S}}^{(n)}\right) &= H\left(C_{\mathcal{S}}^{(n)}\right) - H\left(C_{\mathcal{S}}^{(n)} \middle| W_i^{(n)}\right) \\ &\geq H\left(C_{\mathcal{S}}^{(n)}\right) = H\left(C_{\mathcal{S},i}^{(n)}, C_{\mathcal{S},[N]\setminus\{i\}}^{(n)}\right) \geq H\left(C_{\mathcal{S},i}^{(n)}\right), \end{aligned} \tag{2.83}$$

where $C_{\mathcal{S},i}^{(n)} := \left(C_{k,i}^{(n)}\right)_{k\in\mathcal{S}}$ and

$$C_{\mathcal{S},[N]\setminus\{i\}}^{(n)} := \left(C_{k,1}^{(n)}, \ldots, C_{k,i-1}^{(n)}, C_{k,i+1}^{(n)}, \ldots, C_{k,N}^{(n)}\right)_{k\in\mathcal{S}}.$$

On the other hand, we have

$$\begin{aligned} I\left(W_i^{(n)}; C_{\mathcal{S}}^{(n)}\right) &= I\left(W_i^{(n)}; C_{\mathcal{S},i}^{(n)}, C_{\mathcal{S},[N]\setminus\{i\}}^{(n)}\right) \\ &= I\left(W_i^{(n)}; C_{\mathcal{S},[N]\setminus\{i\}}^{(n)}\right) + I\left(W_i^{(n)}; C_{\mathcal{S},i}^{(n)} \middle| C_{\mathcal{S},[N]\setminus\{i\}}^{(n)}\right) \\ &\leq I\left(W_i^{(n)}; W_{[N]\setminus\{i\}}^{(n)}\right) + I\left(W_i^{(n)}; C_{\mathcal{S},i}^{(n)} \middle| C_{\mathcal{S},[N]\setminus\{i\}}^{(n)}\right) \\ &\overset{(a)}{=} I\left(W_i^{(n)}; C_{\mathcal{S},i}^{(n)} \middle| C_{\mathcal{S},[N]\setminus\{i\}}^{(n)}\right) \\ &= H\left(C_{\mathcal{S},i}^{(n)} \middle| C_{\mathcal{S},[N]\setminus\{i\}}^{(n)}\right) - H\left(C_{\mathcal{S},i}^{(n)} \middle| C_{\mathcal{S},[N]\setminus\{i\}}^{(n)}, W_i^{(n)}\right) \\ &= H\left(C_{\mathcal{S},i}^{(n)} \middle| C_{\mathcal{S},[N]\setminus\{i\}}^{(n)}\right) \leq H\left(C_{\mathcal{S},i}^{(n)}\right), \end{aligned} \tag{2.84}$$

where (a) follows since files $W_1^{(n)}, \ldots, W_N^{(n)}$ are mutually independent. Hence, using (2.83), (2.84), and (2.4), we arrive at $I\left(W_i^{(n)}; C_{\mathcal{S}}^{(n)}\right) = H\left(C_{\mathcal{S},i}^{(n)}\right) \leq n\mu_{\mathcal{S}} f$. This completes the proof of the lemma. □

Proof of Lemma 2.5. Let $\boldsymbol{d} = (d_1, \ldots, d_K)$ be the demand vector. First, note that since user k is capable of decoding its requested file $W_{d_k}^{(n)}$ from its received signal and cache content $C_k^{(n)}$, there should exist some family of caching strategies, encoding, and decoding functions with block length n and decoding error probability ϵ_n where $\epsilon_n \to 0$ as $n \to \infty$. From Fano's inequality, we have

$$H\left(W_{d_k}^{(n)} \middle| Y_k^n, C_k^{(n)}\right) \le n\epsilon_n, \quad k \in [K].$$

Then, we can write

$$\begin{aligned}
nf(\mathfrak{C}, \boldsymbol{d}) - n\epsilon_n &\le H\left(W_{d_1}^{(n)}\right) - n\epsilon_n \\
&\le I\left(W_{d_1}^{(n)}; Y_1^n, C_1^{(n)}\right) \\
&= I\left(W_{d_1}^{(n)}; Y_1^n \middle| C_1^{(n)}\right) + I\left(W_{d_1}^{(n)}; C_1^{(n)}\right) \\
&= \sum_{i=1}^{n} I\left(W_{d_1}^{(n)}; Y_{1,i} \middle| Y_1^{i-1}, C_1^{(n)}\right) + I\left(W_{d_1}^{(n)}; C_1^{(n)}\right) \\
&\overset{(a)}{\le} \sum_{i=1}^{n} I\left(W_{d_1}^{(n)}, Y_1^{i-1}; Y_{1,i} \middle| C_1^{(n)}\right) + I\left(W_{d_1}^{(n)}; C_1^{(n)}\right) \\
&\overset{(b)}{=} \sum_{i=1}^{n} I\left(W_{d_1}^{(n)}, Y_1^{Q-1}; Y_{1,Q} \middle| C_1^{(n)}, Q = i\right) + I\left(W_{d_1}^{(n)}; C_1^{(n)}\right) \\
&= n \sum_{i=1}^{n} I\left(W_{d_1}^{(n)}, Y_1^{Q-1}; Y_{1,Q} \middle| C_1^{(n)}, Q = i\right) \mathbb{P}(Q = i) + I\left(W_{d_1}^{(n)}; C_1^{(n)}\right) \\
&= n \sum_{i=1}^{n} I\left(W_{d_1}^{(n)}, Y_1^{Q-1}; Y_{1,Q} \middle| C_1^{(n)}, Q\right) + I\left(W_{d_1}^{(n)}; C_1^{(n)}\right) \\
&\le nI\left(W_{d_1}^{(n)}, C_1^{(n)}, Y_1^{Q-1}, Q; Y_{1,Q}\right) + I\left(W_{d_1}^{(n)}; C_1^{(n)}\right) \\
&\overset{(c)}{=} nI(U_1; Y_{1,Q}) + I\left(W_{d_1}^{(n)}; C_1^{(n)}\right) \\
&= nI(U_1; Y_1) + I\left(W_{d_1}^{(n)}; C_1^{(n)}\right) \\
&\overset{(d)}{\le} nI(U_1; Y_1) + n\mu_{\{1\}} f(\mathfrak{C}, \boldsymbol{d}),
\end{aligned} \tag{2.85}$$

where (a) holds since

$$I\left(W_{d_1}^{(n)}, Y_1^{i-1}; Y_{1,i} \middle| C_1^{(n)}\right) = I\left(Y_1^{i-1}; Y_{1,i} \middle| C_1^{(n)}\right) + I\left(W_{d_1}^{(n)}; Y_{1,i} \middle| Y_1^{i-1}, C_1^{(n)}\right),$$

in (b) Q is a random variable independent of all other random variables which are uniformly distributed over $[n]$, in (c) we define $U_1 \coloneqq \left(W_{d_1}^{(n)}, C_1^{(n)}, Y_1^{Q-1}, Q\right)$, and in (d) we used (2.44). This implies inequality in (2.45).

We define the subset of indices $d_{[k]} = \{d_1, \ldots, d_k\}$ for every $k \in [K]$. Similarly, for $k \in [2 : K-1]$, we have

$$
\begin{aligned}
nf(\mathfrak{C}, \boldsymbol{d}) - n\epsilon_n &= H\left(W_{d_k}^{(n)}\right) - n\epsilon_n \\
&\le I\left(W_{d_k}^{(n)}; Y_k^n, C_k^{(n)}\right) \\
&= I\left(W_{d_k}^{(n)}; Y_k^n \middle| C_k^{(n)}\right) + I\left(W_{d_k}^{(n)}; C_k^{(n)}\right) \\
&\le I\left(W_{d_k}^{(n)}; Y_k^n, W_{d_{[k-1]}}^{(n)}, C_{[k-1]}^{(n)} \middle| C_k^{(n)}\right) + I\left(W_{d_k}^{(n)}; C_k^{(n)}\right) \\
&= I\left(W_{d_k}^{(n)}; C_{[k-1]}^{(n)} \middle| C_k^{(n)}\right) + I\left(W_{d_k}^{(n)}; W_{d_{[k-1]}}^{(n)} \middle| C_{[k]}^{(n)}\right) \\
&\quad + I\left(W_{d_k}^{(n)}; Y_k^n \middle| W_{d_{[k-1]}}^{(n)}, C_{[k]}^{(n)}\right) + I\left(W_{d_k}^{(n)}; C_k^{(n)}\right) \\
&\overset{(a)}{=} I\left(W_{d_k}^{(n)}; Y_k^n \middle| W_{d_{[k-1]}}^{(n)}, C_{[k]}^{(n)}\right) + I\left(W_{d_k}^{(n)}; C_{[k]}^{(n)}\right) \\
&= \sum_{i=1}^{n} I\left(W_{d_k}^{(n)}; Y_{k,i} \middle| W_{d_{[k-1]}}^{(n)}, C_{[k]}^{(n)}, Y_k^{i-1}\right) + I\left(W_{d_k}^{(n)}; C_{[k]}^{(n)}\right) \\
&\le \sum_{i=1}^{n} I\left(W_{d_k}^{(n)}; Y_{k,i}, Y_{[k-1]}^{i-1} \middle| W_{d_{[k-1]}}^{(n)}, C_{[k]}^{(n)}, Y_k^{i-1}\right) + I\left(W_{d_k}^{(n)}; C_{[k]}^{(n)}\right) \\
&\overset{(b)}{=} \sum_{i=1}^{n} I\left(W_{d_k}^{(n)}; Y_{k,i} \middle| \{W_{d_{[k-1]}}^{(n)}, C_{[k]}^{(n)}, Y_{[k]}^{i-1}\right) + I\left(W_{d_k}^{(n)}; C_{[k]}^{(n)}\right) \\
&\le \sum_{i=1}^{n} I\left(W_{d_k}^{(n)}, C_k^{(n)}, Y_k^{i-1}; Y_{k,i} \middle| W_{d_{[k-1]}}^{(n)}, C_{[k-1]}^{(n)}, Y_{[k-1]}^{i-1}\right) + I\left(W_{d_k}^{(n)}; C_{[k]}^{(n)}\right) \\
&\overset{(c)}{=} \sum_{i=1}^{n} I\left(W_{d_k}^{(n)}, C_k^{(n)}, Y_k^{Q-1}; Y_{k,Q} \middle| W_{d_{[k-1]}}^{(n)}, C_{[k-1]}^{(n)}, Y_{[k-1]}^{Q-1}, Q = i\right) + I\left(W_{d_k}^{(n)}; C_{[k]}^{(n)}\right) \\
&= nI\left(W_{d_k}^{(n)}, C_k^{(n)}, Y_k^{Q-1}; Y_{k,Q} \middle| W_{d_{[k-1]}}^{(n)}, C_{[k-1]}^{(n)}, Y_{[k-1]}^{Q-1}, Q\right) \\
&\quad + I\left(W_{d_k}^{(n)}; C_{[k]}^{(n)}\right) \qquad (2.86) \\
&\overset{(d)}{=} nI(U_k; Y_{k,Q} | U_{k-1}) + I\left(W_{d_k}^{(n)}; C_{[k]}^{(n)}\right) \\
&= nI(U_k; Y_k | U_{k-1}) + I\left(W_{d_k}^{(n)}; C_{[k]}^{(n)}\right) \overset{(e)}{\le} nI(U_k; Y_k | U_{k-1}) + n\mu_{[k]} f(\mathfrak{C}, \boldsymbol{d}), \qquad (2.87)
\end{aligned}
$$

where (a) holds since, for an uncoded caching strategy and mutually independent files,

we have

$$
\begin{aligned}
&I\left(W_{d_k}^{(n)};W_{d_{[k-1]}}^{(n)}\middle|C_{[k]}^{(n)}\right)\\
&=I\left(W_{d_k}^{(n)};W_{d_{[k-1]}}^{(n)}\middle|C_{[k],d_k}^{(n)},C_{[k],d_{[k-1]}}^{(n)},C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&=H\left(W_{d_k}^{(n)},C_{[k],d_k}^{(n)},C_{[k],d_{[k-1]}}^{(n)},C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&\quad+H\left(W_{d_{[k-1]}}^{(n)},C_{[k],d_k}^{(n)},C_{[k],d_{[k-1]}}^{(n)},C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&\quad-H\left(W_{d_k}^{(n)},W_{d_{[k-1]}}^{(n)},C_{[k],d_k}^{(n)},C_{[k],d_{[k-1]}}^{(n)},C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&\quad-H\left(C_{[k],d_k}^{(n)},C_{[k],d_{[k-1]}}^{(n)},C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&=H\left(W_{d_k}^{(n)},C_{[k],d_{[k-1]}}^{(n)},C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&\quad+H\left(W_{d_{[k-1]}}^{(n)},C_{[k],d_k}^{(n)},C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&\quad-H\left(W_{d_k}^{(n)},W_{d_{[k-1]}}^{(n)},C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&\quad-H\left(C_{[k],d_k}^{(n)},C_{[k],d_{[k-1]}}^{(n)},C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&=H\left(W_{d_k}^{(n)}\right)+H\left(C_{[k],d_{[k-1]}}^{(n)}\right)\\
&\quad+H\left(C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)+H\left(W_{d_{[k-1]}}^{(n)}\right)\\
&\quad+H\left(C_{[k],d_k}^{(n)}\right)+H\left(C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)\\
&\quad-H\left(W_{d_k}^{(n)}\right)-H\left(W_{d_{[k-1]}}^{(n)}\right)\\
&\quad-H\left(C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)-H\left(C_{[k],d_k}^{(n)}\right)\\
&\quad-H\left(C_{[k],[N]\setminus d_{[k]}}^{(n)}\right)=0.
\end{aligned}
$$

Moreover, (b) follows from the degradedness of the channel, which implies that for any time instance i, conditioned on $Y_{k,i}$, all channel outputs $\{Y_{u,i} : u < k\}$ are independent of the channel input and hence from the files and cache contents. More precisely, from

$\left(W_{[N]}^{(n)}, C_{[K]}^{(n)}\right) \leftrightarrow X_i \leftrightarrow Y_{K,i} \leftrightarrow \cdots \leftrightarrow Y_{1,i}$ we have

$$\begin{aligned}
&I\left(W_{d_k}^{(n)}; Y_{[k-1]}^{i-1} \middle| W_{d_{[k-1]}}^{(n)}, C_{[k]}^{(n)}, Y_k^{i-1}\right) \\
&= H\left(Y_{[k]}^{i-1} \middle| W_{d_{[k-1]}}^{(n)}, C_{[k]}^{(n)}, Y_k^{i-1}\right) - H\left(Y_{[k-1]}^{i-1} \middle| W_{d_{[k]}}^{(n)}, C_{[k]}^{(n)}, Y_k^{i-1}\right) \\
&= H\left(Y_{[k-1]}^{i-1} \middle| Y_k^{i-1}\right) - H\left(Y_{[k-1]}^{i-1} \middle| Y_k^{i-1}\right) = 0.
\end{aligned}$$

Furthermore, in the equality marked by (c), the random variable Q is independent of all other random variables and admits a uniform distribution over $[n]$. In the step (d) we have

$$U_k := \left(U_{k-1}, W_{d_k}^{(n)}, C_k^{(n)}, Y_k^{Q-1}\right) = \left(W_{d_{[k]}}^{(n)}, C_{[k]}^{(n)}, Y_{[k]}^{Q-1}, Q\right).$$

Finally, in the inequality (e) we used (2.44). Dividing both sides of (2.87) by n and letting $n \to \infty$, we arrive at (2.46), claimed in the lemma.

Finally, we can use a similar argument for the K-th and reach to (2.86). Continuing from there, we can write

$$\begin{aligned}
&nf(\mathfrak{C}, \boldsymbol{d}) - n\epsilon_n \\
&\le nI\left(W_{d_K}^{(n)}, C_K^{(n)}, Y_K^{Q-1}; Y_{K,Q} \middle| W_{d_{[K-1]}}^{(n)}, C_{[K-1]}^{(n)}, Y_{[K-1]}^{Q-1}, Q\right) + I\left(W_{d_K}^{(n)}; C_{[K]}^{(n)}\right) \\
&\overset{(a)}{=} nI\left(X_Q, W_{d_K}^{(n)}, C_K^{(n)}, Y_K^{Q-1}; Y_{K,Q} \middle| U_{K-1}\right) + I\left(W_{d_K}^{(n)}; C_{[K]}^{(n)}\right) \\
&\overset{(b)}{=} nI\left(X_Q; Y_{K,Q} | U_{K-1}\right) + I\left(W_{d_K}^{(n)}; C_{[K]}^{(n)}\right) \\
&= nI\left(X; Y_K | U_{K-1}\right) + I\left(W_{d_K}^{(n)}; C_{[K]}^{(n)}\right) \\
&\overset{(c)}{\le} nI\left(X; Y_K | U_{K-1}\right) + n\mu_{[K]} f(\mathfrak{C}, \boldsymbol{d}),
\end{aligned} \tag{2.88}$$

where (a) holds since the channel input X_Q is deterministically determined by the files and cache contents, (b) holds since condition on X_Q, the channel output $Y_{K,Q}$ is independent of all other variables, and (c) follows from (2.44). The last inequality in (2.47) can be obtained from (2.88).

It remains to show that the random variables $U_1, \ldots, U_{K-1}$ form a Markov chain. This is immediately implied by the recursive construction of U_k and the fact that U_{k-1} is deterministically known once U_k is given. This completes the proof of the lemma. □

Proof of Lemma 2.1. The proof of Lemma 2.1 is derived directly from the proof of Lemma 2.5. From (2.85) and (2.88), we have

$$nf(\mathfrak{C}, \boldsymbol{d}) - n\epsilon_n \leq nI(U_1; Y_1) + n\mu_{\{1\}} f(\mathfrak{C}, \boldsymbol{d}), \tag{2.89}$$

$$nf(\mathfrak{C}, \boldsymbol{d}) - n\epsilon_n \leq nI\left(X; Y_2 | U_1\right) + n\mu_{\{1,2\}} f(\mathfrak{C}, \boldsymbol{d}), \tag{2.90}$$

For the last term in (2.89), we can write

$$\mu_{\{1\}} = \Big| \bigcup_{\ell \in [N_1]} \mathcal{I}_{1,\ell} \Big| \leq \sum_{\ell \in [N_1]} |\mathcal{I}_{1,\ell}| = \mu. \tag{2.91}$$

Similarly, for the last term in (2.90) we get

$$\begin{aligned} \mu_{\{1,2\}} &= \Big| \bigcup_{u \in \{1,2\}} \bigcup_{\ell \in [N_u]} \mathcal{I}_{u,\ell} \Big| \\ &\leq \sum_{u \in \{1,2\}} \sum_{\ell \in [N_u]} |\mathcal{I}_{u,\ell}| \\ &= \sum_{\ell \in [N_1]} |\mathcal{I}_{1,\ell}| + \sum_{\ell \in [N_2]} |\mathcal{I}_{2,\ell}| = 2\mu. \end{aligned} \tag{2.92}$$

Plugging (2.91) and (2.92) into (2.89) and (2.90), respectively, we arrive at the desired inequalities. This completes the proof of the lemma. □

Part II

Algorithms for Distributed Machine Learning

Chapter 3

Adaptive Bit Allocation for Communication-Efficient Distributed Optimization

We propose an adaptive quantization method for two important distributed computation tasks: federated learning and distributed optimization. In both settings, we propose adaptive bit allocation schemes that allow nodes to trade their bandwidth with a minimal communication overhead. We show that the proposed schemes lead to an improvement in the speed of convergence of these methods compared to a uniform bit allocation method, especially when the data distribution among the nodes is skewed. Our theoretical results are corroborated by extensive simulations on various datasets.

3.1 Introduction

Large-scale data analytic and machine learning tasks on massive datasets require efficient frameworks for scalable stochastic optimization algorithms [51–53] on distributed computational architectures. Federated Learning and Distributed Optimization are two popular paradigms that are widely used to perform learning tasks in a decentralized fashion.

In federated learning, the network consists of a central node (server) that utilizes

a number of worker nodes capable of performing computation tasks at the edge of the network. This setting has been studied in various scenarios including the unbalanced data distribution and correlated and/or non-identical distribution of the data [54]. In federated learning, the goal is to fit a common estimate to the data generated and/or collected at individual computation nodes, without transferring the massive amount of (perhaps confidential) collected data from the edge to the back-end servers for processing [55]. A gradient-based optimization algorithm under a federated setting consists of several iterations, where in each iteration each node computes the gradient at the current optimal point estimate using its own data points, and then the estimate is updated by the server who collects all the gradients from the edge nodes. The frequent transmission of the local gradients and updated estimates between the worker nodes and the server yields a large communication overhead, which is a major bottleneck to achieve a fast convergence (see e.g., [56]).

In practice, the presence of a server in the learning phase may be costly or even infeasible. Distributed optimization is a paradigm to mitigate this barrier when there is no central computing node in the network. In this setting, each computing node has its own estimate, which will be updated throughout the algorithm based on the local data points. However, the goal is to achieve a consensus across all the nodes in the work, i.e., all the local estimate/models should converge to a common global model [51, 57–60]. Such a consensus can be achieved by exchanging the local estimates between the neighboring nodes, and a local update based on a (weighted) average of the estimates obtained from the neighbors. This data exchanged between all neighbors accounts for a communication overhead, which affects the overall delay and convergence rate of the algorithm.

While the mentioned frameworks provide solutions for learning tasks in a distributed fashion, their implementation over finite bandwidth communication channels is challenging. With the growth of the dimension of the optimization variable, the communication load of the distributed algorithms increases, and the satisfaction of a communication bottleneck becomes more challenging. Various compression approaches have been used in the literature to mitigate the communication constraint. Two commonly used gradient compression approaches are (i) quantization where the gradient vectors are represented with a finite number of bits and the quantized gradients are communicated

over the network [4,5], and (ii) sparsification of gradient vectors, i.e., only a number of most significant coordinates of the gradient vectors are transmitted over the network [61,62]. A method based on the combination of sparsification and quantization is studied in [9]. Other communication-efficient algorithms for distributed optimization are proposed in [63,64].

Contributions. The main communication load in federated learning is due to sending the local gradients from the computing nodes to the server. Similarly, in distributed optimization, the estimate of each node would be exchanged with its neighbors. The goal of a communication-efficient scheme studied is to provide the receiver of each communication with the best estimate of the transmitted data, while the communication cost (in terms of number of transmitted bits or delay) is subject to a constraint. The majority of the proposed schemes in the literature, including quantization and sparsification, treat all the data exchanges equally and allocate the same bandwidth to all communications taking place throughout the process. However, it is known that an adaptive allocation of the bandwidth among the transmitters can minimize the distortion and improve the overall quality of estimates at the receiver (see e.g., [65–67]). In [68] an adaptive bit-allocation algorithm is proposed for the mean-estimation task under the federated learning, which only guarantees satisfying a given *average* bandwidth constraint.

In this work, we focus on Quantized Stochastic Gradient Descent (QSGD) algorithms for federated and distributed optimization tasks. By minimizing the quantization error, we provide an adaptive bandwidth allocation algorithm (measured in terms of the number of bits used for quantization at each node) to achieve a faster convergence rate for QSGD. For the centralized setting (federated learning), the server can exactly determine this optimum solution without any communication overhead. The server broadcasts the achieved optimum solution to worker nodes that initiates a negligible communication overhead. We show when the data distribution over nodes is skewed the gain we can achieve is significant in comparison with the uniform bit assignment. In a decentralized framework (distributed optimization), a naive characterization of the optimum bandwidth allocation requires global network information. However, we propose a distributed algorithm for adaptive bit allocation that meets the decentralized nature of the network, with fairly low communication overhead. Our simulation results illustrate the improvement in the speed of converges offered by the proposed adaptive

bit allocation.

Throughout this chapter, we denote the set of integers $\{1, 2, \ldots, n\}$ by $[n]$, and for a vector $\mathbf{x} \in \mathbb{R}^d$ we use $\|\mathbf{x}\|$ to denote the ℓ_2-norm of $\mathbf{x}$.

3.2 Problem Formulation

In this section, we introduce the problem setup along with the underlying assumption.

3.2.1 Learning and Computation Model

This chapter is motivated by stochastic learning problems in which the goal is to solve

$$\min_{\mathbf{x}} L(\mathbf{x}) \coloneqq \min_{\mathbf{x}} \mathbb{E}_{\xi \sim \mathcal{P}} \left[\ell(\mathbf{x}, \xi) \right], \tag{3.1}$$

where $\ell : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ is a loss function, $\mathbf{x} \in \mathbb{R}^{1 \times d} = \mathbb{R}^d$ is the decision/optimization *row* vector, and ξ is a random vector taking values in $\mathbb{R}^p$ that is drawn from an unknown underlying distribution $\mathcal{P}$. One of the key practical considerations that renders (3.1) as a challenging task is that the underlying distribution $\mathcal{P}$ is often unknown. Instead, we have access to N independent realizations of ξ and focus on solving the corresponding empirical risk minimization (ERM) problem which is given by

$$\min_{\mathbf{x}} f(\mathbf{x}) \coloneqq \min_{\mathbf{x}} \frac{1}{N} \sum_{j=1}^{N} \ell(\mathbf{x}, \xi_j), \tag{3.2}$$

where $f(\mathbf{x})$ is the empirical risk with respect to the data points $\mathcal{D} = \{\xi_1, \ldots, \xi_N\}$. We assume that $\ell(\cdot, \cdot)$ is a non-convex loss function, which potentially results in a non-convex function $f(\cdot)$.

In distributed optimization, we have a network consisting of n computing nodes (agents, or workers), where each node i observes a non-overlapping subset of $m_i = r_i N$ data points, denoted by $\mathcal{D}_i = \{\xi_1^i, \ldots, \xi_{m_i}^i\}$, where $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_n$. Here, r_i represents the fraction of the data that is processed at node $i \in [n]$. Note that the vector $\mathbf{r} = (r_1, \ldots, r_n)$ is a strictly positive stochastic vector, i.e., $r_i > 0$ and $\sum_{i=1}^n r_i = 1$. Thus, the ERM problem in (3.2) can be written as the minimization of the weighted average of local empirical risk functions f_i for all nodes $i \in [n]$ in the network, i.e.,

$$\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\mathbf{x}} \sum_{i=1}^{n} r_i f_i(\mathbf{x}) = \min_{\mathbf{x}} \frac{1}{N} \sum_{i=1}^{n} \sum_{\xi \in \mathcal{D}_i} \ell(\mathbf{x}, \xi), \tag{3.3}$$

where $f_i(\mathbf{x}) := \frac{1}{m_i}\sum_{\xi\in\mathcal{D}_i}\ell(\mathbf{x},\xi) = \frac{1}{m_i}\sum_{j=1}^{m_i}\ell(\mathbf{x},\xi_j^i)$. Now, we provide the following assumptions on the loss function $\ell(\cdot,\cdot)$.

Assumption 3.1. We assume that the function $\ell(\cdot,\cdot)$ is K-smooth with respect to its first argument; i.e., for any $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ and $\xi\in\mathcal{D}$ we have $\|\nabla\ell(\mathbf{x},\xi)-\nabla\ell(\mathbf{y},\xi)\|\le K\|\mathbf{x}-\mathbf{y}\|$; this implies $\|\nabla f_i(\mathbf{x})-\nabla f_i(\mathbf{y})\|\le K\|\mathbf{x}-\mathbf{y}\|$ for any $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ and every $i\in[n]$. Moreover, since $\mathbf{r}$ is a stochastic vector, the function f is also K-smooth.

Assumption 3.2. Stochastic gradient $\nabla\ell(\mathbf{x},\xi)$ is unbiased and variance bounded, i.e.,

$$\mathbb{E}_\xi\left[\nabla\ell(\mathbf{x},\xi)\right]=\nabla L(\mathbf{x}),\quad \mathbb{E}_\xi\left[\|\nabla\ell(\mathbf{x},\xi)-\nabla L(\mathbf{x})\|^2\right]\le\sigma^2,$$

similarly, we have

$$\mathbb{E}_\xi\left[\nabla f_i(\mathbf{x})\right]=\nabla L(\mathbf{x}),\quad \mathbb{E}_\xi\left[\|\nabla f_i(\mathbf{x})-\nabla L(\mathbf{x})\|^2\right]\le\sigma^2/m_i.$$

3.2.2 Communication Model

To solve the optimization problem (3.2) distributively, we use iterative SGD-based algorithms. In these schemes, at each iteration, each computing node communicates its variables to its neighboring nodes (or to the central server). These variables are vectors in $\mathbb{R}^d$, where d is typically a large number, and hence, such communication schemes require infinite bandwidth. This motivates the use of quantization methods where the quantized versions of the vectors are communicated between the agents. Among various quantization methods, we use a quantization scheme, which we refer to it as *stochastic quantizer*[1] , and is widely used in the literature (see e.g. [4]).

Example 3.1. (Stochastic Quantizer). The stochastic quantizer with a number of quantization levels s maps a vector $\mathbf{x}\in\mathbb{R}^d$ to a random vector $Q_s^S(\mathbf{x})\in\mathbb{R}^d$, where its ℓ-th entry is given by

$$\left[Q_s^S(\mathbf{x})\right]_\ell := \|\mathbf{x}\|\cdot\operatorname{sgn}(x_\ell)\cdot\zeta\left(|x_\ell|/\|\mathbf{x}\|,s\right),\quad \ell\in[d], \tag{3.4}$$

[1] This quantizer is often referred to as the low-precision quantizer in the literature. However, in the original manuscript [4], low-precision quantizer is proposed as the combination of a stochastic quantizer and an efficient coding scheme for communicating the quantized values.

and $\zeta(x,s)$ is a random variable taking values

$$\zeta(x,s) = \begin{cases} \lceil sx \rceil / s & \text{w.p. } sx - \lfloor sx \rfloor \\ \lfloor sx \rfloor / s & \text{w.p. } \lceil sx \rceil - sx. \end{cases}$$

Note that, random variables $\{\zeta(\cdot,\cdot)\}$ are independent, across the coordinates, agents, and time steps. Thus, in this case, the relationship between $\tilde{\mathbf{x}}_j(t)$ and $\mathbf{x}_j(t)$ in Figure 4.1 would be $\tilde{\mathbf{x}}_j(t) = Q_s^S(\mathbf{x}_j(t))$. Furthermore, the noisy channel is perfect, and the decoder component is just an identity function, i.e., $\mathbf{y}_{i,j}(t) = \tilde{\mathbf{x}}_j(t)$ and $\hat{\mathbf{x}}_i(t) = \mathbf{y}_i(t)$. It is shown in [4] that the output of this quantizer for an input $\mathbf{x} \in \mathbb{R}^d$ satisfies $\mathbb{E}\left[Q_s^S(\mathbf{x})\right] = \mathbf{x}$ and $\mathbb{E}\left[\|Q_s^S(\mathbf{x}) - \mathbf{x}\|^2\right] \leq \|\mathbf{x}\|^2$ where $\gamma(s) = \min\left(\frac{\sqrt{d}}{s}, \frac{d}{s^2}\right)$.

Here, we set $s = 2^b$ for some number of bits b. In this work, we consider the case that d is sufficiently large, and thus, $\gamma(s) = \gamma(b) = \frac{\sqrt{d}}{2^b}$.

The key contribution of this work is to design algorithms that solve (3.3) under communication constraints using an *adaptive* (quantization) bit assignment, in which $b_i(t)$ bit is assigned to node i at time t, while respecting the capacity constraint

$$\sum_{i=1}^{n} b_i(t) \leq B, \quad b_i(t) \in \mathbb{Z}^+,\ i \in [n]. \tag{3.5}$$

Our goal is to find such an adaptive scheme to increase the convergence speed for various optimization/learning tasks.

Note that the random variable $\zeta(\cdot, b)$ will be independently applied to each coordinate of vector $\mathbf{x}$, and the stochastic quantizer uses db bits to communicate the outputs of $\zeta(\cdot, b)$ for all coordinates. Hence, even if b is not an integer, for a sufficiently large d, the communication bandwidth db can be closely approximated by an integer. As a result, we replace the condition $b_i(t) \in \mathbb{Z}^+$ by $b_i(t) \geq 0$ in (3.5) for the rest of the chapter.

3.3 Federated Learning Setup

In this section, we investigate the federated architecture where a central node (or server) aims at finding an optimum solution (estimate) for the optimization problem in (3.3) while distributing the computation tasks among a set of agents, which we refer to as workers.

In federated learning, at each iteration t of the algorithm, the server broadcasts its current estimate/model $\mathbf{x}(t)$ to all the nodes and each node $i \in [n]$ computes local gradient ∇f_i at point $\mathbf{x}(t)$ based on its local dataset $\mathcal{D}_i$. This gradient will be then quantized to $Q^S_{b_i}(\nabla f_i(\mathbf{x}(t)))$ using $b_i = b_i(t)$ bits per coordinate, and sent to the server. The server then aggregates the quantized gradients and updates its estimate according to

$$\mathbf{x}(t+1) = \mathbf{x}(t) - \eta(t) \sum_{i=1}^{n} r_i Q^S_{b_i}(\nabla f_i(\mathbf{x}(t))) = \mathbf{x}(t) - \eta(t)\boldsymbol{v}(t), \tag{3.6}$$

where $\eta(t)$ is a vanishing step size satisfying $\sum_t \eta(t) = \infty$ and $\sum_t \eta^2(t) < \infty$. Note that $\boldsymbol{v}(t)$ is a random vector, satisfying $\mathbb{E}[\boldsymbol{v}(t)] = \nabla f(\mathbf{x}(t))$. It is shown in [69] that under some regularity condition on the function $f(\cdot)$, the dynamics (3.6) converges to a critical point of $f(\mathbf{x})$.

Let us denote by $\mathcal{F}(t)$ the sigma-algebra that represents the history of the system up to iteration t. The expected loss function at iteration $t+1$ can be upper-bounded as

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}(t+1)) \mid \mathcal{F}(t)\right] \leq f(\mathbf{x}(t)) - \left(1 - \frac{K}{2}\eta(t)\right)\eta(t)\|\nabla f(\mathbf{x}(t))\|^2 \\ + \frac{K}{2}\eta^2(t)\sum_{i=1}^{n} r_i^2 \gamma(b_i(t))\|\nabla f_i(\mathbf{x}(t))\|^2.\end{aligned} \tag{3.7}$$

Therefore, a myopic strategy to accelerate the convergence of the algorithm would be to minimize the last term in (3.7). This minimization can be formulated as

$$\begin{aligned}&\min g(\boldsymbol{b}(t)),\\ &\text{s.t.} \sum_{i=1}^{n} b_i(t) \leq B,\\ &\quad 0 \leq b_i(t), \text{ for all } i \in [n],\end{aligned} \tag{3.8}$$

where $\boldsymbol{b}(t) = (b_1(t), \ldots, b_n(t))^T$ and $g(\boldsymbol{b}(t)) := \sum_{i=1}^n r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2 \gamma(b_i(t))$.

Algorithm 1 Centralized adaptive bits allocation for gradient quantization

Server sets clock $t = 0$ and initializes the estimate $\mathbf{x}(0)$.

repeat

Each node i computes the local gradient $\nabla f_i(\mathbf{x}(t))$.

Each node i sends the *norm* of the local gradient $\|\nabla f_i(\mathbf{x}(t))\|$.

Server calculates the vector $\boldsymbol{b}^\star(t)$ using equation (3.9) and broadcasts it.

Each node i sends the quantized gradient $Q^S_{b_i^\star}(\nabla f_i(\mathbf{x}(t)))$.

Server computes an SGD iteration according to (3.6).

until convergence or the number of iterations reaches the maximum setting.

Using KKT conditions, it can be shown that the solution to (3.8) is given by

$$b_i^\star(t) = \left(\log\left(r_i^2\|\nabla f_i(\mathbf{x}(t))\|^2\right) - \nu\right)^+, \tag{3.9}$$

where ν is a constant satisfying water-filling type relation

$$\sum_{i=1}^{n}\left(\log\left(r_i^2\|\nabla f_i(\mathbf{x}(t))\|^2\right) - \nu\right)^+ = B.$$

Therefore, our proposed adaptive bit-allocation algorithm is that at each iteration $t \geq 0$, the server evaluates the optimum solution in (3.9), and sends out $b_i^\star(t)$ to agent $i \in [n]$, where $\boldsymbol{b}^\star(t) = (b_1^\star(t), \ldots, b_n^\star(t))$. The overall procedure is summarized in Algorithm 1.

Note that the proposed adaptive scheme assigns a higher number of bits to the nodes with *larger* gradient vectors. This is very intuitive since the error in the estimate of the overall gradient is dominated by the quantization error associated with gradients with larger norms.

Communication Overhead. It is very crucial that solving for the optimum bit allocation in (3.8) only requires the knowledge of vector $\mathbf{r}$ (which is fixed throughout the algorithm) and the norm of the local gradients. Interestingly, those local gradient norms should be sent to the server, regardless of the bit application. Therefore, the server can evaluate the optimum bit allocation without any communication overhead. Lastly, the server sends $b_i^\star(t)$ to node i using at most $\log B$ bits. Therefore, the additional number of bits to communicate $\boldsymbol{b}^\star(t)$ is $n \log B$. Note that the communication cost of the original algorithm consists of $1 + b_i$ bits per coordinate and a fixed number of bits (typically 32 bits) to send the gradient vectors $\|\nabla f_i(\mathbf{x}(t))\|$ from each node i to the server, as

well as the cost of broadcasting $\mathbf{x}(t)$ from the server to all the nodes. This yields to a communication cost, lower bounded by $\sum_{i=1}^{n} 32 + d(1+b_i) = 32n + dn + dB$. Hence, the associated communication overhead is upper-bounded by $\frac{n \log B}{32n+dn+dB}$ which is negligible for practical values of $d \gg n$.

Analysis of Performance. Here, we study the gain offered by the adaptive bit allocation in comparison to the uniform bit assignment [6]. We denote by $\boldsymbol{b}^\star$ the optimum bit allocation in (3.9), and by $\boldsymbol{b}_u = (B/n, \ldots, B/n)$ the uniform bit allocation. Then, we have the following theorem, which is proved in Appendix 3.7.

Theorem 3.1. *If $b_i^\star(t) > 0$ for all $i \in [n]$, then the improvement in the descent in (3.7) offered by the adaptive bit allocation obtained in (3.9) over a uniform bit allocation is give by*

$$g(\boldsymbol{b}_u) - g(\boldsymbol{b}^\star(t)) = \frac{n\sqrt{d}}{2^{B/n}}(A - G), \tag{3.10}$$

where

$$A = \frac{1}{n}\sum_{i=1}^{n} r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2, \qquad G = \left(\prod_{i=1}^{n} r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2\right)^{\frac{1}{n}}$$

are the arithmetic and the geometric means of $\{r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2\}$'s, respectively.

It is well known that $A \geq G$, and an equality holds if and only if $r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2 = c$ for some constant c and for all $i \in [n]$. It can be immediately seen from Theorem 3.1 that the gain of the adaptive bit allocation is more significant when the weighted gradient norms are *highly skewed*.

3.4 Distributed Optimization Setup

In this section, we study solving the optimization problem in (3.3) in a fully distributed setting, namely, distributed optimization. This setting arises in a plethora of applications, where either the total number of samples N is massive and data cannot be stored or processed over a single node or the samples are available in parts at different nodes and, due to privacy or communication constraints, exchanging raw data points among

the nodes is not feasible. Moreover, the existence of a central server is costly or even infeasible in many applications.

Consider a network consisting of $n \geq 2$ nodes, which are partially connected via bidirectional links in a time-invariant topology. Such a network can be modeled as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = [n]$ represents the set of nodes, and $\mathcal{E} \subseteq \{\{i, j\} : i, j \in \mathcal{V}, i \neq j\}$ represents the set of links. Two nodes $i, j \in \mathcal{V}$ are called neighbors if and only if $\{i, j\} \in \mathcal{E}$. The set of neighbors of node i is denoted by $\mathcal{N}_i = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$. Each node can only communicate with its neighbors throughout the algorithm.

The optimization problem in (3.3) can be expressed as a classical decentralized optimization, given by

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}_i) \quad \text{s.t.} \qquad \mathbf{x}_1 = \cdots = \mathbf{x}_n. \tag{3.11}$$

A quantized version of SGD (called QSGD) is proposed by [6] to solve this optimization problem. In iteration t of this algorithm, each node i broadcasts a quantized version of its local model $Q_{b_i}^S(\mathbf{x}_i(t))$ to all its neighbors (b_i is the number of bits per coordinate allocated to node i and will be determined later). Consequently, it receives the quantized local models $Q_{b_j}^S(\mathbf{x}_j(t))$ from all its neighbors $j \in \mathcal{N}_i$. The model of node i will be then updated to a weighted average of its own and its neighbors' models and its local gradient. That is

$$\mathbf{x}_i(t+1) = (1 - \epsilon + \epsilon w_{ii})\mathbf{x}_i(t) + \epsilon \sum_{j \in \mathcal{N}_i} w_{ij} Q_{b_j}^S(\mathbf{x}_j(t)) - \alpha\epsilon \nabla f_i(\mathbf{x}_i(t)). \tag{3.12}$$

Here, α and ϵ are positive scalars set as the step-sizes and w_{ij} as the weight that node i assigns to the information that it receives from node j.

We can stack all such weights into a weight matrix $W = [w_{ij}]$, where $w_{ij} \geq 0$ and $w_{ij} = 0$ whenever $\{i, j\} \notin \mathcal{E}$. We make the following assumptions on the weight matrix W.

Assumption 3.3. We assume that $W \in [0, 1]^{n \times n}$ is a symmetric doubly-stochastic matrix, i.e., $W = W^T$, $W\mathbf{1} = \mathbf{1}$, $\mathbf{1}^T W = \mathbf{1}^T$, where $\mathbf{1} \in \mathbb{R}^n$ is the all-one column vector. Moreover, we assume the eigenvalues of W are $1 = |\lambda_1(W)| > |\lambda_2(W)| \geq \cdots \geq |\lambda_n(W)|$, and its spectral gap is $\beta := 1 - |\lambda_2(W)| \in (0, 1]$.

Let us denote the average (across nodes) estimates of (3.12) at time t by $\bar{\mathbf{x}}(t) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(t)$. In [6], it is shown that, under Assumptions 3.1–3.3 and boundedness of the trajectories of (3.12), for sufficiently large termination time T, the time-average (from the beginning to time T) of the expected gradients at $\{\bar{\mathbf{x}}(t)\}$ becomes arbitrarily small for a proper choice of α, and ϵ (depending on the termination time T).

To investigate the dynamics in (3.12), let us define

$$\begin{aligned} X(t) &= \Big[\mathbf{x}_1(t), \dots, \mathbf{x}_n(t)\Big]^T, \\ Q(X(t)) &= \Big[Q^S_{b_j}(\mathbf{x}_1(t)), \dots, Q^S_{b_n}(\mathbf{x}_n(t))\Big]^T, \\ \nabla f(X(t)) &= \Big[\nabla f(\mathbf{x}_1(t)), \dots, \nabla f(\mathbf{x}_n(t))\Big]^T. \end{aligned}$$

Using the matrix notation above, the dynamics in (3.12) can be rewritten as

$$X(t+1) = ((1-\epsilon)I + \epsilon W)X(t) + \epsilon(W - W_D)(Q(X(t)) - X(t)) - \alpha\epsilon\nabla f(X(t)),$$

where W_D is a diagonal matrix and its ith diagonal entry equals to w_{ii}, for $i \in [n]$. Moreover, we have $\bar{\mathbf{x}}(t) = \frac{1}{n}\mathbf{1}^T X(t)$, where $\mathbf{1}$ is an all-ones vector. Thus,

$$\bar{\mathbf{x}}(t+1) = \bar{\mathbf{x}}(t) + \frac{\epsilon}{n}\mathbf{1}^T(W - W_D)(Q(X(t)) - X(t)) - \frac{\alpha\epsilon}{n}\mathbf{1}^T\nabla f(X(t)).$$

Next, using Assumption 3.1, the expected loss at $\bar{\mathbf{x}}(t+1)$ can be upper bounded as

$$\begin{aligned} \mathbb{E}\left[f\left(\bar{\mathbf{x}}(t+1)\right)|\mathcal{F}(t)\right] \le f\left(\bar{\mathbf{x}}(t)\right) &- \alpha\epsilon\left\langle \nabla f\left(\bar{\mathbf{x}}(t)\right), \frac{\mathbf{1}^T}{n}\nabla f(X(t))\right\rangle \\ &+ \epsilon^2\alpha^2\frac{K}{2}\left\|\frac{\mathbf{1}^T}{n}\nabla f(X(t))\right\|^2 \\ &+ \epsilon^2\frac{K}{2n^2}\sum_{i=1}^{n}(1-w_{ii})^2\|\mathbf{x}_i(t)\|^2\gamma(b_i(t)). \end{aligned} \tag{3.13}$$

It is clear from (3.13) that minimizing the last term leads to greater progress in $\mathbb{E}\left[f\left(\bar{\mathbf{x}}(t+1)\right)|\mathcal{F}(t)\right]$, and eventually a faster convergence. This leads to the optimization problem

$$\begin{aligned} &\min_{\boldsymbol{b}} \sum_{i=1}^{n}(1-w_{ii})^2\|\mathbf{x}_i(t)\|^2\gamma(b_i(t)), \\ &\text{s.t.} \sum_{i=1}^{n} b_i(t) \le B, \qquad b_i(t) \ge 0, i \in [n]. \end{aligned} \tag{3.14}$$

Similar to the case of federated learning, the optimum solution of (3.14) is given by

$$b_i^\star(t) = \left(\log s_i(t) - \nu\right)^+, \tag{3.15}$$

where ν can be found from $\sum_{i=1}^n \left(\log s_i(t) - \nu\right)^+ = B$, and $s_i(t) = (1 - w_{ii})^2 \|\mathbf{x}_i(t)\|^2$. However, in the absence of a central node that has access to $\{s_i(t) : i \in [n]\}$, evaluation of (3.15) is challenging. A naïve approach to address this is to share all values of $s_i(t)$ across the network, and let each agent individually evaluate (3.15) which in addition to requiring coordination among the agents, and additional storage at each node, it introduces a flooding delay over the network.

Here, we seek to solve the optimization problem in (3.14) in a distributed way. An efficient mechanism to achieve this is to use gossip-based algorithms [70, 71]. One such algorithm is discussed in [72] where authors presented a gossip-based, distributed asynchronous algorithm that solves distributed constrained optimization problems over networks with a time-varying topology. The algorithm operates by forcing the nodes' estimates of an unknown minimizer (here, the vector $\boldsymbol{b}^\star$) to asymptotically achieve a consensus while satisfying a conservation condition derived from the Karush-Kuhn-Tucker conditions. The downside of this work is that *all* nodes are required to estimate *all* minimizers (here, $b_i^\star$ for $i \in [n]$). Hence, it imposes a large communication overhead on the system.

In the following, we provide a *Pairwise Equalizing* (PE) algorithm that results in a minimal communication overhead and a fast convergence rate to an optimal value. We need to run the gossip-based algorithm between any two iterations of the SGD. We use t and ℓ to refer to the iterations of SGD (outer) and gossip-based (inner) algorithms, respectively. For the sake of brevity, for a fixed SGD iteration $t \geq 0$, and every node $i \in [n]$, we denote $\mathbf{x}_i(t)$ and $s_i(t) = (1 - w_{ii})^2 \|\mathbf{x}_i(t)\|^2$ by $\mathbf{x}_i$ and s_i, respectively.

Algorithm 2 Gossip based bit allocation for model quantization

Input Row stochastic matrix P.

Initialize $\ell = 0$ and $\boldsymbol{b}(0) = (B/n, \ldots, B/n)$.

repeat

Increase ℓ; An edge $\{i, j\}$ is chosen w.p. $\frac{P_{ij}+P_{ji}}{n}$ (independent of the past).

Node i transmits $b_i(\ell-1)$ and s_i to node j and node j sends $b_j(\ell-1)$ and s_j to node i.

Nodes i and j update $b_i(\ell)$ and $b_j(\ell)$ according to (3.17) and (3.18).

until convergence or $\ell = T$.

Our algorithm (for each SGD iteration t) is as follows: We set $\boldsymbol{b}(0) = (B/n, \ldots, B/n)$ and $\ell = 1$. At each gossip (inner) iteration ℓ, we choose a node i uniformly at random, and then an edge $\{i, j\}$ for $j \in \mathcal{N}_i$ with probability P_{ij}, where P is a row-stochastic matrix. Then, node i transmits $b_i(\ell-1)$ and s_i to node j. Similarly, node j transmits $b_j(\ell-1)$ and s_j to node i. Finally, nodes i and j exchange their bit allocation variables to solve the following optimization problem:

$$\begin{aligned} \min_{b_i(\ell), b_j(\ell)} \; & s_i\gamma(b_i(\ell)) + s_j\gamma(b_j(\ell)), \\ \text{s.t.} \quad & b_i(\ell) + b_j(\ell) = b_i(\ell-1) + b_j(\ell-1). \end{aligned} \tag{3.16}$$

The explicit solution to this problem can be written as

$$b_i(\ell) = \frac{1}{2}\left(b_i(\ell-1) + b_j(\ell-1)\right) + \frac{1}{2}\log\frac{s_i}{s_j}, \tag{3.17}$$

$$b_j(\ell) = \frac{1}{2}\left(b_i(t-1) + b_j(t-1)\right) + \frac{1}{2}\log\frac{s_j}{s_i}. \tag{3.18}$$

Therefore, agents i and j update their estimates using (3.17) and (3.18), respectively, while the remaining agents' estimates would not change. We repeat similar steps for T iterations. The proposed method is summarized in Algorithm 2.

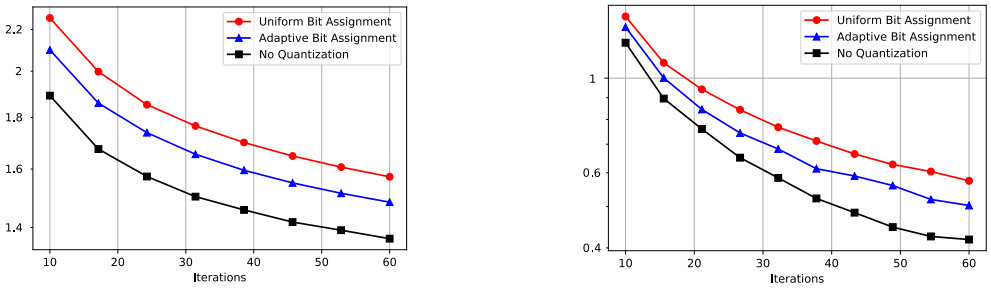

Figure 3.1: Training Loss vs. Iterations for Federated Learning with capacity constraint $B = 10$: Logistic Regression on CIFAR-10 (left) and Convolutional Neural Network on MNIST (right).

Next, we estimate the convergence time of $\boldsymbol{b}(\ell)$ to the limit $\boldsymbol{b}^\star$ (for each (outer) iteration t).

Definition 3.1. For any $0 < \kappa < 1$, we define the κ-convergence time of Algorithm 2 to be

$$T(\kappa, P) := \sup_{\boldsymbol{b}(0)} \inf \left\{ \ell : \Pr\left(\frac{\|\boldsymbol{b}(\ell) - \boldsymbol{b}^\star\|}{\|\boldsymbol{b}(0) - \boldsymbol{b}^\star\|} \geq \kappa \right) \leq \kappa \right\}.$$

Theorem 3.2. *Let $\boldsymbol{b}^\star$ be the optimal solution of (3.15). Then, if $b_i^\star > 0$ for all $i \in [n]$, then the κ-convergence time $T(\kappa, P)$ of Algorithm 2 is bounded as follows*

$$\frac{0.5 \log \kappa^{-1}}{\log \lambda_2^{-1}(U)} \leq T(\kappa, P) \leq \frac{3 \log \kappa^{-1}}{\log \lambda_2^{-1}(U)},$$

where

$$U = I - \frac{1}{2n} D + \frac{P + P^T}{2n},$$

and D is a diagonal matrix with entries

$$D_i = \sum_{j=1}^{n} [p_{ij} + p_{ji}]. \tag{3.19}$$

The proof of this theorem is provided in Appendix 3.8.

Communication Overhead. At each round in the gossip algorithm, two nodes send the norms of their models and their current number of bits. Note that communicating

the norms does not add an overhead, since it should be exchanged regardless of the bit allocation. Hence, the additional number of bits to communicate is at most $2\log B$ bits per round of the gossip algorithm. Thus, the overall communication overhead is upper-bounded by $\frac{2T\log B}{32n+dn+dB}$, which is negligible for large values of d and B.

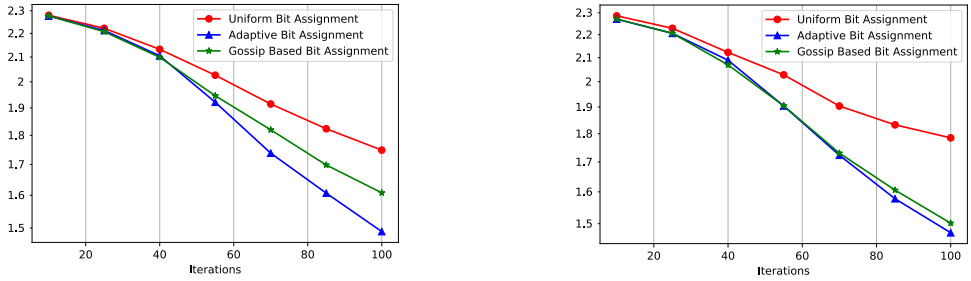

Figure 3.2: Distributed Optimization: Training Loss of Convolutional Neural Network on MNIST. The capacity constraint is $B = 20$, and the number of Gossip iterations is $T = 5$ (left) and $T = 20$ (right).

3.5 Experimental Results

Here, we provide experimental results supporting the effectiveness of the proposed algorithms. We use logistic regression and convolutional neural network (CNN) for the validation of our algorithms in convex and non-convex settings, respectively. Our adaptive bit allocation schemes show significant improvement over the uniform bit assignment approach. However, it should be mentioned that this comes at a cost of a negligible increase in communication cost.

Data and Experimental Setup. We implement two sets of experiments over CIFAR-10 and MNIST datasets. For both datasets, we set $n = 10$ as the number of worker nodes, and use $N = 1000$ data points for MNIST and $N = 5000$ data points for CIFAR-10. We distribute the data points across the nodes according to $r_i = p_i / \sum_{i=1}^{10} p_i$, where p_i is drawn uniformly at random from the interval $(0.02, 0.18)$. This results in a relatively skewed data distribution across the nodes. The loss function is the cross-entropy function. Each image is converted to a vector of length $28 \times 28 = 784$ for MNIST dataset and $3 \times 32 \times 32 = 3072$ for CIFAR-10 dataset. The CNN for MNIST dataset consists of two convolutional layers, two pooling layers, and two fully connected layers.

The mini-batch size is set to be 5 for CNN and 10 for logistic regression in both datasets. We implement the unbiased stochastic quantizer with various total bandwidth constraints. Note that the total communication costs are identical for uniform and adaptive bit assignments, and hence, we can plot the loss function versus iterations (as opposed to time).

Experiments in Federated Learning Setup. In Figure 3.1, the plot on the left represent the training time for a regularized logistic regression problem over CIFAR-10 dataset with the fine-tuned stepsize $\eta(t) = 0.1/\sqrt{t}$ and the plot on the right show the training time on CNN over MNIST dataset with $\eta(t) = 0.045/\sqrt{t}$. In each plot, 'No Quantization' refers to the performance of the FedAvg [54], and 'Uniform Bit Assignment' refers to the algorithm in [73]. Each curve shows the training loss versus the training time for the aggregated model at the server for each iteration. It can be observed that the proposed algorithm leads to a faster convergence, compared to the uniform bit assignment. This observation is consistent with the main idea behind our algorithm, which is to minimize gradient variance given a limited bandwidth budget.

Experiments in Distributed Optimization Setup We use a random Erdös-Renyi graph on $n = 10$ worker nodes and edge probability $p_c = 0.3$ for the connectivity network. The averaging weight matrix is chosen to be $W = I - \frac{L}{\mu}$ where L is the Laplacian matrix of the graph and μ is the maximum degree of the graph plus two. The parameters of the dynamics in (3.12) are finely tuned to $(\epsilon, \alpha) = (0.3, 0.015)$. Figure 3.2 demonstrates the training time of a convolutional neural network (CNN) for MNIST dataset over the network introduced above. The experiments are performed for 100 iterations of SGD, and $T = 5$ (left) and $T = 20$ (right) iterations of the gossip algorithm iterations. The row-stochastic matrix used for the random edge selection in each iteration of the Gossip algorithm is

$$P_{ij} = \begin{cases} \frac{1}{|\mathcal{N}_i|} & j \in \mathcal{N}_i \\ 0 & \text{otherwise} \end{cases},$$

where $|\mathcal{N}_i|$ is the number of neighbors of node i in the network. Moreover, a total bandwidth constraint is set to $B = 20$. The plots indicate that by increasing T, the performance of the (decentralized) gossip based bit assignment improves and approaches that of the centralized adaptive bit assignment scheme.

3.6 Concluding Remarks

We proposed two adaptive bit allocation schemes for the federated and distributed optimization settings to reduce the overall quantization error, and improve the convergence rate of QSGD. In the federated learning setting, the optimum bit allocation is obtained at the server, using the data received from the nodes. We also proposed a gossip-based algorithm for the distributed optimization setting to obtain the optimum bit allocation in a distributed manner. The communication overhead of the adaptive bit allocation is negligible compared to the communication cost of the main algorithm. Comprehensive experiments support the effectiveness of our algorithms in improving the convergence rate of QSGD.

3.7 Proof of Theorem 3.1

For $\boldsymbol{b}_u = (B/n, \ldots, B/n)$ we have

$$g(\boldsymbol{b}_u) = \sum_{i=1}^{n} \frac{\sqrt{d}}{2^{B/n}} r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2 = \frac{n\sqrt{d}A}{2^{B/n}}, \tag{3.20}$$

where $A = \frac{1}{n}\sum_{i=1}^{n} r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2$ is the arithmetic mean of $\{r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2\}_{i=1}^{n}$. Next, let $\nu^\star$ be the optimum value in (3.9). Then, since $b_i^\star(t) > 0$ for every $i \in [n]$, we get

$$b_i^\star(t) = \log r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2 - \nu^\star. \tag{3.21}$$

Summing up (3.21) for $i \in [n]$ we have

$$B = \sum_{i=1}^{n} \log\left(r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2\right) - n\nu^\star. \tag{3.22}$$

This implies

$$\nu^\star = \log\left(\prod_{i=1}^{n} r_i \|\nabla f_i(\mathbf{x}(t))\|^2\right)^{\frac{1}{n}} - \frac{B}{n} = \log G - \frac{B}{n}, \tag{3.23}$$

where G is the geometric mean of $\{r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2\}$'s. Plugging (3.23) into (3.21), we get

$$b_i^\star(t) = \log(r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2) - \log G + \frac{B}{n},$$

and

$$2^{b_i^\star(t)} = \frac{r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2}{G} 2^{B/n}.$$

Therefore, the $g(\boldsymbol{b}^\star(t))$ can be evaluated as

$$\begin{aligned} g(\boldsymbol{b}^\star(t)) &= \sum_{i=1}^{n} r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2 \frac{\sqrt{d}}{2^{b_i^\star(t)}} \\ &= \sum_{i=1}^{n} \frac{r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2 \sqrt{d} G}{r_i^2 \|\nabla f_i(\mathbf{x}(t))\|^2 2^{B/n}} = \frac{n\sqrt{d}G}{2^{B/n}}. \end{aligned} \tag{3.24}$$

Subtracting (3.24) from (3.20), we get the equality in (3.10). This completes the proof. ■

3.8 Proof of Theorem 3.2

Note that we use t to denote the iterations of the SGD, and ℓ to refer to the iterations of the Gossip algorithm, performed between the iterations of SGD to determine the bit allocation. Therefore, for the following analysis we fix the SGD iteration $t \geq 0$ and for the sake of brevity, we denote $\mathbf{x}_i(t)$ and $s_i(t) = (1 - w_{ii})^2 \|\mathbf{x}_i(t)\|^2$ by $\mathbf{x}_i$ and s_i, respectively, for all $i \in [n]$.

Let (neighboring) nodes i and j be the gossiping agents in the (inner) iteration ℓ, and exchange their variables $(b_i(\ell-1), s_i)$ and $(b_j(\ell-1), s_j)$. Then, they individually solve the local optimization problem

$$\begin{aligned} &\min_{b_i, b_j} \quad s_i \gamma(b_i) + s_j \gamma(b_j), \\ &\text{s.t.} \quad b_i + b_j = b_i(\ell-1) + b_j(\ell-1), \quad b_i, b_j \geq 0. \end{aligned} \tag{3.25}$$

It can be verified that if $b_i(\ell)$ and $b_j(\ell)$ in (3.17) and (3.18) are positive, then they form the optimal solution for (3.25). Then, the bit assignments of nodes i and j will be updated to $b_i(\ell)$ and $b_j(\ell)$, respectively. In matrix form, this can be written as

$$\boldsymbol{b}(\ell) = V(\ell)\boldsymbol{b}(\ell-1) + \frac{1}{2}\log\frac{s_i}{s_j}(e_i - e_j). \tag{3.26}$$

Here, $V(\ell) = V_{i,j}$ if neighboring nodes i and j exchange their information at iteration ℓ of the Gossip algorithm, and

$$V_{i,j} = I - \frac{(e_i - e_j)(e_i - e_j)^T}{2},$$

where e_i is the ith element of the standard basis in $\mathbb{R}^n$.

Let $\boldsymbol{b}^\star = (b_1^\star, \ldots, b_n^\star)$ be the optimum bit allocation at iteration t (of SGD) satisfying (3.15). It follows from (3.15) that $b_k^\star = \frac{B}{n} + \log s_k - \frac{1}{n}\sum_{p=1}^n \log s_p$ when $b_k^\star > 0$ for every $k \in [n]$. Therefore, we have

$$\boldsymbol{b}^\star = \sum_{k=1}^n b_k^\star e_k = \sum_{k=1}^n \left(\frac{B}{n} + \log s_k - \frac{1}{n}\sum_{p=1}^n \log s_p\right) e_k.$$

Replacing $\boldsymbol{b}^\star$ in (3.26), we get

$$\begin{aligned}
V(\ell)\boldsymbol{b}^\star + \frac{1}{2}\log\frac{s_i}{s_j}(e_i - e_j) &= \left(I - \frac{(e_i - e_j)(e_i - e_j)^T}{2}\right)\sum_{k=1}^n b_k^\star e_k + \frac{1}{2}\log\frac{s_i}{s_j}(e_i - e_j) \\
&= \boldsymbol{b}^\star - \frac{1}{2}\sum_{k=1}^n b_k^\star (e_i - e_j)(e_i - e_j)^T e_k + \frac{1}{2}\log\frac{s_i}{s_j}(e_i - e_j) \\
&= \boldsymbol{b}^\star - \frac{1}{2}\sum_{k\in\{i,j\}} b_k^\star (e_i - e_j)(e_i - e_j)^T e_k + \frac{1}{2}\log\frac{s_i}{s_j}(e_i - e_j) \\
&= \boldsymbol{b}^\star - \frac{1}{2}b_i^\star (e_i - e_j) + \frac{1}{2}b_j^\star (e_i - e_j) + \frac{1}{2}\log\frac{s_i}{s_j}(e_i - e_j) \\
&= \boldsymbol{b}^\star - \frac{1}{2}\log\frac{s_i}{s_j}(e_i - e_j) + \frac{1}{2}\log\frac{s_i}{s_j}(e_i - e_j) = \boldsymbol{b}^\star. \qquad (3.27)
\end{aligned}$$

This implies that for every choice of (i, j), the vector $\boldsymbol{b}^\star$ is a fixed point of the dynamics in (3.26). Moreover, denoting $\boldsymbol{h}(\ell) = \boldsymbol{b}(\ell) - \boldsymbol{b}^\star$ and subtracting (3.27) from (3.26), we get

$$\boldsymbol{h}(\ell) = \boldsymbol{b}(\ell) - \boldsymbol{b}^\star = V(\ell)(\boldsymbol{b}(\ell-1) - \boldsymbol{b}^\star) = V(\ell)\boldsymbol{h}(\ell-1). \qquad (3.28)$$

Recall that $V(\ell)$ takes value $V_{i,j}$ with probability $\frac{1}{n}(P_{ij} + P_{ji})$, where $\frac{1}{n}$ is the probability that node i is selected, and P_{ij} is the probability that node j is picked from the neighbors of i. Similarly, we may choose node j with probability $\frac{1}{n}$ and its neighbor i with probability P_{ji}. Hence, the expected averaging random matrix is given by

$$U = \mathbb{E}_{i,j}[V] = \sum_{i,j}\frac{1}{n}P_{ij}V_{i,j} = I - \frac{1}{2n}D + \frac{P + P^T}{2n},$$

where D is defined in The random dynamics in (3.28) is analyzed in [70, Theorem 3],

$$T(\kappa, P) = \sup_{\boldsymbol{h}(0)} \inf \left\{\ell : \Pr\left[\frac{\|\boldsymbol{h}(\ell) - \boldsymbol{h}^\star\|}{\|\boldsymbol{h}(0)\|} \geq \kappa\right] \leq \kappa\right\} \qquad (3.29)$$

satisfies

$$\frac{0.5\log\kappa^{-1}}{\log\lambda_2^{-1}(U)} \leq T(\kappa, P) \leq \frac{3\log\kappa^{-1}}{\log\lambda_2^{-1}(U)}.$$

Here, $\lambda_2(U)$ is the second largest eigenvalue of matrix U. Finally, using $\boldsymbol{h}^\star = \boldsymbol{b}^\star - \boldsymbol{b}^\star = 0$, $\boldsymbol{h}(\ell) = \boldsymbol{b}(\ell) - \boldsymbol{b}^\star$, and $\boldsymbol{h}(0) = \boldsymbol{b}(0) - \boldsymbol{b}^\star$ in (3.29), we get $T(\kappa, P)$ in the statement of the theorem. This completes the proof of the theorem. ■

Chapter 4

Distributed Optimization over Time-varying Graphs with Imperfect Sharing of Information

We study decentralized learning where a set of agents collaboratively solve a separable optimization problem that is distributed over a time-varying network. The existing methods to solve these problems rely on (at most) one time-scale algorithms, where each agent performs a diminishing or constant step-size gradient descent at the average estimate of the agents in the network. However, if possible at all, exchanging exact information, that is required to evaluate these average estimates, potentially introduces a massive communication overhead. Therefore, a reasonable practical assumption to be made is that agents only receive a rough approximation of the neighboring agents' information. To address this, we introduce and study a two time-scale decentralized algorithm with a broad class of lossy information sharing methods (that includes noisy, quantized, and/or compressed information sharing) over time-varying networks. In our method, one time-scale suppresses the (imperfect) incoming information from the neighboring agents, and one time-scale operates on local cost functions' gradients. For strongly convex loss functions, with a proper choice of step-sizes, we show that the agents' estimates converge to the global optimal state at a rate of $\mathcal{O}\left(T^{-1/2}\right)$. Moreover, we prove that with proper choices for the step-sizes' parameters, the algorithm achieves

a convergence rate of $\mathcal{O}(T^{-1/3+\epsilon})$ for non-convex distributed optimization problems over time-varying networks, for any $\epsilon > 0$. Further, we identify the sufficient conditions on the step-sizes sequences for the almost sure convergence of the agent's states to an optimal solution for the class of convex cost functions.

4.1 Introduction

Emergence of big data analytics, modern computer architectures, storage, and data collection have led to a growing interest in the study of multi-agent networked systems. These systems arises in various applications such as sensor networks [74, 75], network routing [76], large scale machine learning [77], power control [78], and distributed network resource allocations [79, 80], for which decentralized solutions offer promising results. In general and in the absence of a central entity, we are often dealing with a time-varying network of agents, each can perform local and on-device computation. The information can be shared throughout the network via local communication between neighboring agents. This communication among agents, especially when the dimension of the data is large, accounts for a significant delay in the overall running time of the algorithm. In this chapter, we study such a distributed optimization framework with lossy and imperfect information sharing and propose and analyze an gradient-based distributed algorithm that guarantees convergence to the optimum solution, in spite of a limitation on the communication load.

Related Works. Various methods have been proposed and studied to solve distributed optimization problems in convex settings [51, 52, 57, 70, 81–85], strongly convex settings [83, 86, 87], and non-convex settings [88, 89]. For the convex objective functions, a sub-gradient method with a fixed step-size is proposed over time-varying graphs in [90]. It is shown that the objective cost function reduces at rates of $\mathcal{O}\left(T^{-1}\right)$ until it reaches a neighbor of a minimizer of the original problem. To achieve exact convergence to a minimizer, various diminishing step-size sub-gradient methods have been proposed and studied [52, 57, 88, 89, 91, 92]. Considering convex loss functions that are Lipschitz continuous and have bounded gradients, a subgradient-push algorithm is proposed in [92]. There it is shown that the objective cost function convergences at the rate of $\mathcal{O}\left(T^{-1/2}\ln T\right)$ over uniformly strongly connected, directed time-varying graphs. Under

the same assumption and strong-convexity for loss functions, a better rate $\mathcal{O}\left(T^{-1}\ln T\right)$ for the objective loss function plus squared consensus residual is shown in [91].

Almost all the aforementioned works on this domain, consider distributed optimization with perfect sharing of information, i.e., the agents are allowed to communicate real-valued vectors perfectly over perfect communication channels. However, exchanging exact information among nodes initiates a massive communication overhead on the system that considerably slows down the convergence rate of these algorithms in real-world applications. Thus, it is reasonable to assume that each agent has access to a lossy version of neighboring agents' information.

To address lossy/noisy sharing of information, a (fixed steps-size) decentralized gradient descent method is proposed in [93]. Assuming fixed communication network and strongly convex local cost functions, it is shown that for a given iteration T, the algorithm's parameters (depending on T) can be chosen such that the local estimate of each agent at iteration T is (roughly) within $c\left(T^{-1/2+\epsilon}\right)$-distance of the global optimal solution for some $c>0$ and any $\epsilon>0$. Furthermore, the result holds for a termination time T which is required to satisfy $T\geq T_{\min}$, where $T_{\min}$ depends on ϵ as well as non-local parameters of the underlying fixed graph. Specifically, as ϵ goes to zero, $T_{\min}$ diverges to infinity. In a closely related recent work [94], a two time-scale gradient descent algorithm has been presented for strongly convex loss functions. Assuming a *fixed* topology for the underlying network, uniform weighting of the local cost functions, and a specific scheme for lossy sharing of information, it is shown that the expected objective loss function achieves a rate of $\mathcal{O}(T^{-1/2}(\ln T)^2)$. To compensate the quantization error, a decentralized (diminishing) gradient descent algorithm is proposed in in [95, 96] using error-feedback. The proposed algorithm achieves the convergence rate of $\mathcal{O}(T^{-1})$ and $\mathcal{O}(T^{-1/2})$ for strongly and non-convex objective functions, respectively. However, the nature of the algorithm restricts its use to time-invariant networks, and in addition, the feedback mechanism cannot compensate communication noise between the nodes. In another related work [97], a two-time-scale gradient descent algorithm was presented for distributed constrained and convex optimization problems over an i.i.d. communication graph with noisy communication links, and sub-gradient errors. It is shown that under certain conditions on the i.i.d. communication graph and proper choices of time-scale parameters the proposed dynamics result in almost sure convergence of local states to

the optimal point. Another interesting approach to address exact convergence for distributed optimization with fixed gradient step-sizes under a noiseless communication model is to use gradient tracking methods [98,99].

Contributions. In this work, we study distributed optimization problems for a *broad* class of lossy/noisy sharing of information over time-varying communication networks. The learning method relies only on local computations and received imperfect information from neighbor agents. We show that a two-time scale gradient descent algorithm with a proper choice of parameters reaches the global optima (in ℓ_2 and hence, in probability) for every agent with a rate of $\mathcal{O}\left(T^{-1/2}\right)$. Moreover, we prove that with a proper choice of the parameters for the two diminishing step-size sequences, the temporal average of the expected norm of the gradients decreases with the rate of $\mathcal{O}(T^{-1/3+\epsilon})$. Finally, we show that under certain conditions, the dynamics converge almost surely to an optimal point supported in the optimizer set of the loss function.

In addition, in the existing works on distributed optimization [51,57,83,87,88,90,91,93,94] (with perfect or imperfect sharing of information), either the underlying communication network is assumed to be fixed, or the non-zero elements of averaging weights are assumed to be uniformly bounded away from zero. In our proposed method, however, the weights are not uniformly bounded away from zero and they are evolving over an underlying time-varying communication network. One of the key contributions of this chapter is to develop tools and techniques to deal with diminishing averaging weights for distributed optimization over time-varying networks.

Notation. Throughout this chapter, we denote the set of integers $\{1,2,\ldots,n\}$ by $[n]$ and the set of non-negative real numbers by $\mathbb{R}^+$. In this chapter, we are dealing with n agents that are minimizing a function in $\mathbb{R}^d$. For notational convenience, throughout this chapter, we assume that the underlying functions are acting on **row** vectors, and hence, we view vectors in $\mathbb{R}^{1\times d} = \mathbb{R}^d$ as row vectors. The rest of the vectors, i.e., the vectors in $\mathbb{R}^{n\times 1} = \mathbb{R}^n$, are assumed to be column vectors. For a vector $\mathbf{x} \in \mathbb{R}^d$ we use $\|\mathbf{x}\|$ to denote the ℓ_2-norm of $\mathbf{x}$. A vector $\mathbf{r} \in \mathbb{R}^n$ is called stochastic if $r_i \geq 0$ and $\sum_{i=1}^n r_i = 1$. Similarly, a non-negative matrix $A \in \mathbb{R}^{n\times d}$ is called (row) stochastic if $\sum_{j=1}^d A_{ij} = 1$ for every $i \in [n]$. For a matrix $A \in \mathbb{R}^{n\times d}$, we denote its i-th row and j-th column by A_i and A^j, respectively. For an $n \times d$ matrix A and a strictly positive stochastic vector $\mathbf{r} \in \mathbb{R}^n$, we define the $\mathbf{r}$-norm of A by $\|A\|_{\mathbf{r}}^2 = \sum_{i=1}^n r_i \|A_i\|^2$. It can be verified that $\|\cdot\|_{\mathbf{r}}$ is

a norm on the space of $n \times d$ matrices. We denote the Frobenius norm of A by $\|A\|_F$, where $\|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} |A_{ij}|^2$. Moreover, $A \geq B$ indicates that all the entries of $A - B$ are non-negative.

4.2 Problem Setup and Main Result

In this section, first, we formulate distributed optimization problems over time-varying networks and introduce some standard assumptions on the underlying problem. After proposing our algorithm, we state our main result. Finally, we discuss the implications of our result in various important practical settings with imperfect information sharing.

4.2.1 Problem Setup

We can rewrite the ERM problem in (3.3) as a distributed consensus optimization problem, given by

$$\min_{\mathbf{x}_1,\dots,\mathbf{x}_n} \sum_{i=1}^{n} r_i f_i(\mathbf{x}_i) \quad \text{subject to} \quad \mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n. \tag{4.1}$$

Consider an $n \geq 2$ agents that are connected through a *time-varying* network. We represent this network at time $t \geq 1$ by the directed graph $\mathcal{G}(t) = ([n], \mathcal{E}(t))$, where the vertex set $[n]$ represents the set of agents and the edge set $\mathcal{E}(t) \subseteq \{(i,j) : i,j \in [n]\}$ represents the set of links at time t. At each discrete time $t \geq 1$, agent i can only send information to its (out-) neighbors in $\mathcal{E}(t)$, i.e., all $j \in [n]$ with $(i,j) \in \mathcal{E}(t)$.

To discuss our algorithm (`DIMIX`) for solving (4.1) distributively, let us first discuss its general structure and the required information at each node for its execution. In this algorithm, at each iteration $t \geq 1$, agent $i \in [n]$ updates its estimate $\mathbf{x}_i(t) \in \mathbb{R}^d$ of an optimizer of (3.3). To this end, it utilizes the gradient information of its own local cost function $f_i(\mathbf{x})$ as well as a *noisy/lossy* average of its current neighbor's estimates, denoted by $\hat{\mathbf{x}}_i(t) := \sum_{j=1}^{n} W_{ij}(t)\mathbf{x}_j(t) + \mathbf{e}_i(t)$. Here, $W(t)$ is a *row-stochastic* matrix that is consistent with the underlying connectivity network $\mathcal{G}(t)$ (i.e., $W_{ij}(t) > 0$ only if $(j,i) \in \mathcal{E}(t)$) and $\mathbf{e}_i(t) \in \mathbb{R}^d$ is a random noise vector. Later, in Section 4.2.4 we discuss several noisy and lossy information sharing architectures (quantization and noisy communication) that fit in this broad information structure.

Now we are ready to discuss the DIMIX algorithm. In this algorithm, using the information available to agent i at time t, agent i updates its current estimate by computing a *diminishing* weighted average of its own state and the noisy average of its neighbors' estimates, and moves along its local gradient. More formally, the update rule at node $i \in [n]$ is given by

$$\mathbf{x}_i(t+1) = (1-\beta(t))\mathbf{x}_i(t) + \beta(t)\hat{\mathbf{x}}_i(t) - \alpha(t)\beta(t)\nabla f_i(\mathbf{x}_i(t)), \tag{4.2}$$

where $\alpha(t) = \frac{\alpha_0}{(t+\tau)^\nu}$ and $\beta(t) = \frac{\beta_0}{(t+\tau)^\mu}$ for some $\mu, \nu \in (0,1)$ are the diminishing step-sizes of the algorithm, and $\tau \geq 0$ is an arbitrary shift, that is introduced to accelerate the finite-time performance of the algorithm. The description of DIMIX is summarized in Algorithm 3. For notational simplicity, let

$$X(t) := \begin{bmatrix} \mathbf{x}_1(t) \\ \vdots \\ \mathbf{x}_n(t) \end{bmatrix}, \quad E(t) := \begin{bmatrix} \mathbf{e}_1(t) \\ \vdots \\ \mathbf{e}_n(t) \end{bmatrix}, \quad \nabla f(X(t)) := \begin{bmatrix} \nabla f_1(\mathbf{x}_1(t)) \\ \vdots \\ \nabla f_n(\mathbf{x}_n(t)) \end{bmatrix}. \tag{4.3}$$

Since $\hat{\mathbf{x}}_i(t) = \sum_{j=1}^n W_{ij}(t)\mathbf{x}_j(t) + \mathbf{e}_i(t)$, we can rewrite the update rule in (4.2) in the matrix format

$$X(t+1) = ((1-\beta(t))I + \beta(t)W(t))X(t) + \beta(t)E(t) - \alpha(t)\beta(t)\nabla f(X(t)). \tag{4.4}$$

Algorithm 3 DIMIX at agent i

Input: Stochastic matrix sequence $\{W(t)\}$, Iteration T

Initialization: Set $\mathbf{x}_i(1) = 0$.

for $t = 1, \ldots, T-1$ **do**

Compute the local gradient $\nabla f_i(\mathbf{x}_i(t))$.

Obtain noisy average neighbors' estimate $\hat{\mathbf{x}}_i(t)$.

Update $\mathbf{x}_i(t+1) = (1-\beta(t))\mathbf{x}_i(t) + \beta(t)\hat{\mathbf{x}}_i(t) - \alpha(t)\beta(t)\nabla f_i(\mathbf{x}_i(t))$.

end for

Let us discuss some important aspects of the above update rule. Note that both $\alpha(t)$ and $\beta(t)$ are diminishing step-sizes. If $\beta(t) = \beta_0 < 1$ and $\alpha(t) = \alpha_0 < 1$ are both constants, then the dynamics in (4.4) reduces to the algorithm proposed in [6] for both convex and non-convex cost functions. Alternatively, if $\beta(t) = \beta_0 < 1$ is a constant sequence and

$E(t) = \mathbf{0}$ for all $t \geq 1$, (4.4) would be reduced to the averaging-based distributed optimization with diminishing steps-sizes (for gradients), which is introduced and studied in [52] for local convex cost functions $f_i(\mathbf{x})$. The newly introduced time-scale/step-size $\beta(t)$ suppresses the incoming noise $\mathbf{e}_i(t)$ from the neighboring agents. However, $\beta(t)$ also suppresses the incoming signal level $\sum_{j=1}^{n} W_{ij}(t)\mathbf{x}_j(t)$ at each node i. This casts a major technical challenge for establishing convergence-to-consensus guarantees for the algorithm over time-varying networks. On the other hand, the diminishing step-size for the gradient update is $\hat{\alpha}(t) = \alpha(t)\beta(t)$. We chose to represent our algorithm in this way to ensure that the local mixing (consensus) scheme is operated on a faster time-scale than the gradient descent.

4.2.2 Assumptions

To provide performance guarantees for `DIMIX`, we need to assume certain regularity conditions on (i) the statistics of the (neighbors' averaging) noise process $\{E(t)\}$, (ii) the mixing properties of the weight sequence $\{W(t)\}$, and (iii) the loss function $\ell(\cdot,\cdot)$.

First, we discuss our main assumption on the noise sequence $\{E(t)\}$.

Assumption 4.1. we suppose that $\{X(t)\}$ is adapted to a filtration $\{\mathcal{F}_t\}$ on the underlying probability space (see e.g. Section 5.2 in [100]). We assume that there exists some $\gamma > 0$ such that for all $i \in [n]$ and all $t \geq 1$, the noise sequence $\{\mathbf{e}_i(t)\}$ satisfies

$$\mathbb{E}\left[\mathbf{e}_i(t) \mid \mathcal{F}_t\right] = 0, \quad \text{and} \quad \mathbb{E}\left[\|\mathbf{e}_i(t)\|^2 \mid \mathcal{F}_t\right] \leq \gamma. \tag{4.5}$$

Note that the natural filtration of the random process $\{X(t)\}$ is one choice for $\{\mathcal{F}_t\}$. Thus, (4.5) reduces to $\mathbb{E}[\mathbf{e}_i(t)|X(1),\ldots,X(t)]=0$ and $\mathbb{E}[\|\mathbf{e}_i(t)\|^2|X(1),\ldots,X(t)]\leq\gamma$.

Next, we discuss the main assumption on the network connectivity which relates to information mixing over the time-varying network.

Assumption 4.2. We assume that the weight matrix sequence $\{W(t)\}$ in (4.4) satisfies the following properties.

(a) *Stochastic with Common Stationary Distribution*: for all $t \geq 1$, $W(t)$ is non-negative, $W(t)\mathbf{1} = \mathbf{1}$, and $\mathbf{r}^T W(t) = \mathbf{r}^T$, where $\mathbf{1} \in \mathbb{R}^n$ is the all-one vector, and $\mathbf{r} > 0$ is the weight vector.

(b) *Bounded Nonzero Elements*: There exists some $\eta > 0$ such that $W_{ij}(t) > 0$ implies $W_{ij}(t) \geq \eta$, for all $i, j \in [n]$ and $t \geq 1$.

(c) *B-Connected*: For a fixed integer $B \geq 1$, the graph $\left([n], \bigcup_{k=t+1}^{t+B} \mathcal{E}(k)\right)$ is strongly connected for all $t \geq 1$, where $\mathcal{E}(k) = \{(j,i) \mid W_{ij}(k) > 0\}$.

Assumption 4.3. The function $\ell(\cdot,\cdot)$ is ρ-strongly convex with respect to its first argument, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\xi \in \mathcal{D}$ we have that $\langle \nabla \ell(\mathbf{x}, \xi) - \nabla \ell(\mathbf{y}, \xi), \mathbf{x} - \mathbf{y} \rangle \geq \rho \|\mathbf{x} - \mathbf{y}\|^2$. Consequently, we get $\langle \nabla f_i(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \rho \|\mathbf{x} - \mathbf{y}\|^2$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

4.2.3 Main Result and Discussion

Here, we present the main results of this chapter.

Theorem 4.1. *Assume the conditions in Assumptions 3.1, 4.1, 4.2, 4.3 are satisfied and the step-sizes are set to* $\alpha(t) = \frac{\alpha_0}{t^\nu}$ *and* $\beta(t) = \frac{\beta_0}{t^\mu}$ *for* $\mu, \nu \in (0,1)$. *Then, if* $\mu + \nu < 1$, *the dynamics generated by Algorithm 3 satisfy*

$$\mathbb{E}\left[\|X(T) - \mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right] \leq \xi_1 T^{-\min(\mu, 2\nu)} + \xi_2 T^{-\min(\mu-\nu, 2\nu)}, \tag{4.6}$$

for any iteration $T \geq T_0 := \max(T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8)$, *where* T_1, T_2, T_3, T_4, T_5, T_6, T_7 *and* T_8 *are given in* (4.33), (4.40), (4.48), (4.62), (4.65), (4.67), (4.70), *and* (4.71), *respectively, and* $\mathbf{x}^\star := \arg\min f(\mathbf{x})$. *Furthermore, under the same assumptions, when* $\mu + \nu = 1$, *the dynamics generated by Algorithm 3 satisfy* (4.6), *for any iteration* $T \geq T_0$, *provided that* $\frac{\rho K}{\rho + K} \alpha_0 \beta_0 \geq 8 \min(\mu - \nu, 2\nu)$.

We refer to Section 4.5 for the proof of Theorem 4.1.

Remark 4.1. Theorem 4.1 guarantees the exact convergence (in ℓ_2 sense) of each local state to the global optimal with diminishing step-size even though the noises induced by random quantizations and gradients are non-vanishing with iterations. In order to maximize the exponents in the upper bound (4.6), it can be verified that the optimum choice is $(\mu, \nu) = (3/4, 1/4)$. Replacing this in (4.6), we conclude

$$\mathbb{E}\left[\|X(T) - \mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right] \leq \xi T^{-1/2},$$

for any $T \geq T_0$ and $\xi = \xi_1 + \xi_2$. Our algorithm and the main result are inspired by the fixed step-size variation of (4.4) that is proposed in [93] under the limited setting of

time-invariant networks, uniform weights $\mathbf{r}$, and a particular choice of lossy sharing of information. In that setting, it is shown that for any *given stopping time* $T \geq T_{\min}$ and any $\epsilon > 0$, the constant step-sizes $\alpha_0, \beta_0 > 0$ can be set such that

$$\mathbb{E}\left[\|X(T) - \mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right] \leq cT^{-1/2+\epsilon},$$

where $c, T_{\min}$ are positive constants depending on the problem's parameters (note that this is established for a fixed T). However, $T_{\min} \to \infty$ as $\epsilon \to 0$ [93]. Here, we provide a rigorous convergence rate analysis which reduces to $\mathcal{O}\left(T^{-1/2}\right)$ for *every* iteration T. In Theorem 4.1, for the case $\mu + \nu = 1$, the minimum number of required iterations is finite.

We also note that in a recent independent work [94], for a specific quantizer and the specific choice $\mathbf{r} = \frac{1}{n}\mathbf{1}$, the authors have shown the convergence rate of $\mathcal{O}(T^{-1/2}(\ln T)^2)$ for strongly convex loss functions over *fixed* underlying networks and a specific choice of lossy sharing of information. Note that the obtained rate, which is with respect to a weighted average of the previous iterates $\{X(t)\}_{t\leq T}$ instead of $X(T)$, is strictly *slower* than $\mathcal{O}(T^{-1/2})$. Furthermore, it is assumed that gradients are bounded in [94], while here we show that such a strong assumption is not needed and Lipschitz gradients result in expected bounded gradients for strongly convex functions.

One of the key distinctions of our work from the prior works in this domain is the introduction of $\hat{x}_i(t)$ that satisfies some general structural properties without being tied to any specific application. Before discussing the technical details of the main result's proof, let us provide some general practical settings for which our structural assumptions on $\hat{x}_i(t)$ hold.

Next, we characterize the convergence rates of our algorithm for the K-smooth non-convex loss functions. More precisely, we establish a rate for the temporal average of the expected norm of the gradients for various choices of the time-scale parameters ν, μ.

Theorem 4.2. *Suppose that Assumptions 3.1, 3.2, 4.1, and 4.2 hold and let* $\alpha(t) = \frac{\alpha_0}{(t+\tau)^\nu}$ *and* $\beta(t) = \frac{\beta_0}{(t+\tau)^\mu}$ *where* $\alpha_0, \beta_0 \in (0,1)$, $\tau \geq 0$, *and* $\nu, \mu \in (0,1)$ *are arbitrary constants with* $\mu \neq 1/2$ *and* $3\nu + \mu \neq 1$. *Then the weighted average estimates* $\bar{\mathbf{x}}(t) := \sum_{i=1}^n r_i \mathbf{x}_i(t)$ *generated by (4.2) satisfy*

$$M_\theta(\nu,\mu) := \left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]\right)^\theta\right]^{1/\theta} = \mathcal{O}\left(T^{-\min\{1-\nu-\mu,\mu-\nu,2\nu\}}\right), \tag{4.7}$$

where $\theta \in (0,1)$ is an arbitrary constant.

Furthermore, for $(\nu^\star, \mu^\star) = \left(\frac{1}{6}, \frac{1}{2}\right)$ we get the optimal rate of

$$M_\theta(\nu^\star, \mu^\star) = \mathcal{O}\left(T^{-1/3} \ln T\right). \tag{4.8}$$

Remark 4.2. Note that the expectation operator $\mathbb{E}[\cdot]$ is over the randomness of the dataset $\mathcal{D}$ and the compression/communication noise. Moreover, note that the theorem above shows that the gradient of $f(\cdot)$ (which depends on the choice of $\mathcal{D}$) at the average state of $\overline{\mathbf{x}}(t)$ (which also depends on $\mathcal{D}$) vanishes at a certain rate. It is worth mentioning that this is not the performance of the average trajectory for the average function.

Remark 4.3. From (4.7), one has to maximize $\min\{1-\nu-\mu, \mu-\nu, 2\nu\}$ over $\nu, \mu \in (0,1)$ to achieve the fastest convergence for M_θ. This leads to $(\nu^\star, \mu^\star) = (1/6, 1/2)$, which none of the conditions $\mu \neq 1/2$ and $3\nu + \mu \neq 1$ hold for. However, one can choose $(\nu, \mu) = (1/6+\epsilon/2, 1/2+\epsilon/2)$ and obtain $M_\theta = \mathcal{O}(T^{-1/3+\epsilon})$ for any $\epsilon > 0$. Nevertheless, note that (4.8) provides a faster convergence rate of $\mathcal{O}(T^{-1/3} \ln T)$ for $(\nu^\star, \mu^\star) = (1/6, 1/3)$.

Proposition 4.1. Under the conditions of Theorem 4.2, for the optimum choice of $(\nu^\star, \mu^\star) = (1/6, 1/3)$, we have

$$M_1(\nu^\star, \mu^\star) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \leq \mathcal{O}\left(T^{-1/3+\epsilon}\right), \tag{4.9}$$

for any $\epsilon > 0$. Furthermore, in this case, for each agent $i \in [n]$ the convergence rate to consensus is given by

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\|^2\right] \leq \mathcal{O}\left(T^{-1/3+\epsilon}\right). \tag{4.10}$$

As a result, combining (4.9), (4.10), and Assumptions 3.1 and 3.2, for all $i \in [n]$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\mathbf{x}_i(t))\|^2\right] \leq \mathcal{O}\left(T^{-1/3+\epsilon}\right).$$

Remark 4.4. We should comment that almost all the existing results and algorithms on distributed optimization algorithms for time-varying graphs assume a uniform positive lower bound on non-zero elements of the (effective) weight matrices [83, 88, 90, 91, 101–104]. Absence of such an assumption significantly increases the complexity of the

convergence analysis of the algorithm. In our work, even though the stochastic matrix sequence $\{W(t)\}$ is assumed to be B-connected, the **effective averaging** weight sequence, given by $\{(1-\beta(t))I+\beta(t)W(t)\}$, has vanishing weights. One of the major theoretical contributions of this work is to introduce tools and techniques to study distributed optimization with diminishing weight sequences.

Remark 4.5. In a related work [6] on distributed optimization with compressed information sharing among the nodes, authors considered a fixed step-size (zero time-scale) version of our dynamics (4.2) with a fixed averaging matrix W. It is shown that for a given **termination time** T, the algorithm's step-sizes can be chosen (depending on T) such that the temporal average (up to iteration T) of the expected norm of the gradient (i.e., M_1 defined in (4.9)) does not exceed $c(T^{-1/3})$ (where $c>0$ is a constant). However, the algorithm needs to be re-executed with re-evaluated step-sizes if one targets another termination time T'. In this work, we use vanishing step-sizes $\alpha(t)$ and $\beta(t)$ (which do not depend on the termination time) and show that the same temporal average vanishes at the rate of $\mathcal{O}\left(T^{-1/3+\epsilon}\right)$ for **every** iteration T and any arbitrarily small $\epsilon>0$.

4.2.4 Examples for Stochastic Noisy State Estimation

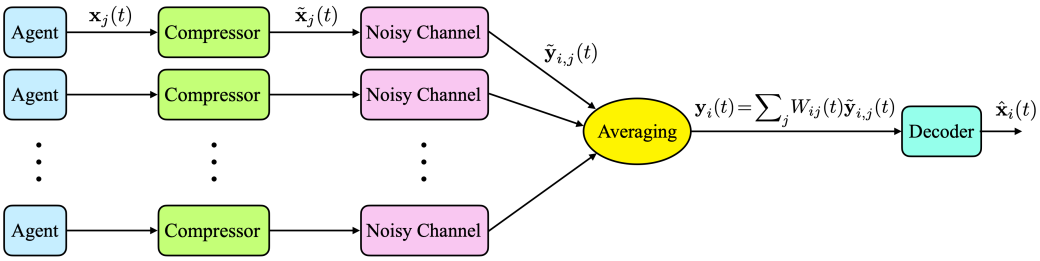

Figure 4.1: A general architecture for lossy information model.

The noisy information in (4.2) is very general and captures several models of imperfect data used in practice and/or theoretical studies. A rather general architecture that leads to such noisy/lossy estimates is demonstrated in Figure 4.1: Once the estimate $\mathbf{x}_j(t)$ of node j at iteration t is evaluated, node j may apply an operation (such as compression/sparsification or quantization) on its own model to generate $\tilde{\mathbf{x}}_j(t)$. This vector is sent over a potentially noisy communication channel, and a neighbor node

i receives a corrupted version of $\tilde{\mathbf{x}}_j(t)$, say $\tilde{\mathbf{y}}_{i,j}(t)$ from every neighbor node j. Upon collecting all channel outputs from its neighbors, node i computes their weighted average $\mathbf{y}_i(t) = \sum_{j=1}^{n} W_{ij}(t)\tilde{\mathbf{y}}_{i,j}(t)$, and decodes it to the approximate average model $\hat{\mathbf{x}}_i(t)$. In the following, we describe three popular frameworks, in which each node i can only use an imperfect neighbors' average $\hat{\mathbf{x}}_i(t)$ to update its estimate. It is worth emphasizing that these are just some examples that lie under the general model in (4.2).

Example 4.1. (Stochastic Quantizer with bounded trajectory). We consider the stochastic quantizer in Example 3.1 with a number of quantization levels s. Thus, in this case, the relationship between $\tilde{\mathbf{x}}_j(t)$ and $\mathbf{x}_j(t)$ in Figure 4.1 would be $\tilde{\mathbf{x}}_j(t) = Q_s^S(\mathbf{x}_j(t))$. Furthermore, the noisy channel is perfect, and the decoder component is just an identity function, i.e., $\mathbf{y}_{i,j}(t) = \tilde{\mathbf{x}}_j(t)$ and $\hat{\mathbf{x}}_i(t) = \mathbf{y}_i(t)$. From [4], the output of this quantizer for an input $\mathbf{x} \in \mathbb{R}^d$ with a bounded norm $\|\mathbf{x}\|^2 \leq D$ satisfies $\mathbb{E}\left[Q_s^S(\mathbf{x})\right] = \mathbf{x}$ and $\mathbb{E}\left[\|Q_s^S(\mathbf{x}) - \mathbf{x}\|^2\right] \leq \min\left(\frac{\sqrt{d}}{s}, \frac{d}{s^2}\right) D$. Therefore, the neighbors' estimate for node i will be

$$\hat{\mathbf{x}}_i(t) = \sum_{j=1}^{n} W_{ij}(t) Q_s^S(\mathbf{x}_j(t)) = \sum_{j=1}^{n} W_{ij}(t)\mathbf{x}_j(t) + \mathbf{e}_i(t),$$

where $\mathbf{e}_i(t) = \sum_{j=1}^{n} W_{ij}(t)\left(Q_b^S(\mathbf{x}_j(t)) - \mathbf{x}_j(t)\right)$ satisfies $\mathbb{E}\left[\mathbf{e}_i(t)|\mathcal{F}_t\right] = 0$ and

$$\mathbb{E}\left[\|\mathbf{e}_i(t)\|^2|\mathcal{F}_t\right] = \min\left(\sqrt{d}/s, d/s^2\right) D \sum_{j=1}^{n} W_{ij}^2(t) \leq \min\left(\sqrt{d}/s, d/s^2\right) D,$$

provided that $\|\mathbf{x}_j(t)\|^2 \leq D$ for every $j \in [n]$ and every $t \geq 1$. Therefore, the conditions of Assumption 4.1 are satisfied. Note that $\mathbf{e}_i(t)$ and $\mathbf{e}_j(t)$ might be correlated, especially when nodes i and j have common neighbor(s). However, this does not violate the conditions of Assumptions 3.1 and 3.2. The bounded variance noise sequence, which is ensured by bounded trajectory here, has been implicitly and explicitly assumed in many related works [6,94,105,106]. In addition, the above stochastic quantizer implicitly assumes a bounded trajectory as otherwise, the state norm ($\|\mathbf{x}\|$, whose communication cost is ignored) requires infinite bits to be transmitted.

Example 4.2. (Noisy Communication). The noisy neighbor estimate model may arise due to imperfect communication between the agents. Consider a wireless network, in which the computing nodes communicate with their neighbors over a Gaussian channel, i.e., when node j sends its state $\mathbf{x}_j(t)$ (without compression, i.e., $\tilde{\mathbf{x}}_j(t) = \mathbf{x}_j(t)$) to

its neighbor i, the signal received at node i is $\tilde{\mathbf{y}}_{i,j}(t) = \mathbf{x}_j(t) + \mathbf{z}_{i,j}(t)$ where $\mathbf{z}_{i,j}(t)$ is a zero-mean Gaussian noise with variance ζ^2, independent across (i,j), and t. Applying an identity map decoder at node i (i.e., $\hat{\mathbf{x}}_i(t) = \mathbf{y}_i(t)$) we have

$$\hat{\mathbf{x}}_i(t) = \sum_{j=1}^{n} W_{ij}(t)\left(\mathbf{x}_j(t) + \mathbf{z}_{i,j}(t)\right) = \sum_{j=1}^{n} W_{ij}(t)\mathbf{x}_j(t) + \sum_{j=1}^{n} W_{ij}(t)\mathbf{z}_{i,j}(t).$$

Therefore, we have $\mathbf{e}_i(t) = \sum_{j=1}^{n} W_{ij}(t)\mathbf{z}_{i,j}(t)$, from which we conclude $\mathbb{E}\left[\mathbf{e}_i(t)|\mathcal{F}_t\right] = 0$ and $\mathbb{E}\left[\|\mathbf{e}_i(t)\|^2|\mathcal{F}_t\right] = \zeta^2 \sum_{j=1}^{n} W_{ij}^2(t) \le \zeta^2$. Hence, the conditions of Assumption 4.1 are satisfied.

4.3 Experimental Results

In the following, we discuss two sets of simulations to verify our theoretical results. First, we perform Algorithm 3 for a *fixed* network to compare fixed step-sizes [6] versus diminishing step-sizes (this work). Then, we compare the performance of our algorithm on a fixed versus time-varying graph. Throughout this section, we use the unbiased stochastic quantizer in (3.4) with various total number of quantization levels s. For CNN experiments over CIFAR-10 and regularized logistic regression problem over MNIST dataset, the mini-batch size is set to be 10 and 20, respectively, the loss function is the cross-entropy function, we use $N = 10,000$ data points which are distributed across $n = 20$ agents. Moreover, for each set of experiments, we distribute the data points across the nodes according to $r_i = p_i / \sum_{i=1}^{20} p_i$, where p_i is drawn uniformly at random from the interval $(0.01, 0.9)$. Our codes are implemented using Python and tested on a 2017 MacBook Pro with 16 GB memory and a 2.5 GHz Intel Core i7 processor.

4.3.1 `DIMIX` vs. `QuanTimed-DSGD` over Fixed Network

We use regularized logistic regression to validate our algorithm on the benchmark dataset MNIST.

Data and Experimental Setup. In this experiment, we train a regularized logistic regression to classify the MNIST dataset over a fixed network. First, we generate a random (connected) undirected Erdös-Renyi graph with the edge probability $p_c = 0.3$, which is fixed throughout this experiment. We also generate a doubly stochastic weight matrix $A := I - (d_{\max} + 1)^{-1} L_G$ where L_G is the Laplacian matrix and $d_{\max}$ is

the maximum degree of the graph. Finally, we use the time-invariant weight sequence $W = I + c\hat{\mathbf{r}}_{\min}\text{diag}^{-1}(\mathbf{r})(A-I)$, where $\hat{\mathbf{r}}_{\min} = \min_{i\in[n]} r_i/(1-A_{ii})$, $c = 0.95$, and $\text{diag}(\mathbf{r})$ is a diagonal matrix with r_i as its ith diagonal element. It can be verified that W satisfies the conditions of Assumption 4.2, i.e., it is row-stochastic and satisfies $\mathbf{r}^T W = \mathbf{r}^T$. We implemented our algorithm with quantizer parameter $s = 3$ and tuned the step-size parameters $(\alpha_0, \nu^\star) = (0.005, 1/6)$, $(\beta_0, \mu^\star) = (0.6, 1/2)$, with $\tau = 2000$. For the fixed step-sizes, we used the termination time $T = 7500$, which results in $\alpha(t) = 0.001$ and $\beta(t) = 0.01$ for all $t \geq 1$.

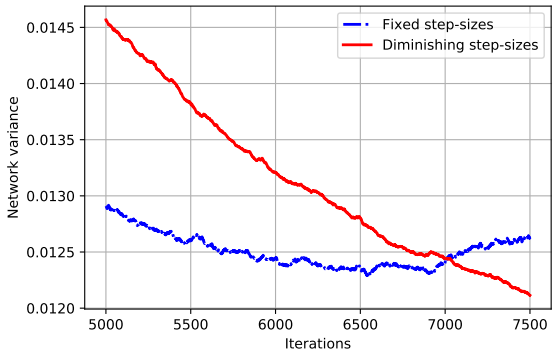

Figure 4.2: Logistic Regression on MNIST: network variance for fixed and vanishing step-sizes.

Results. The plot in Figure 4.2 shows *network variance* defined by $\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2$ for fixed and diminishing step-sizes. It can be observed that with a fixed step-sizes algorithm, we can only reach a *neighborhood* of the average model, but a *consensus* is not necessarily achieved. However, using diminishing step-sizes guarantees that each node's state converges to the average model.

4.3.2 Diminishing Step-sizes over Time-varying vs. Fixed Network

Here, we provide simulation results to demonstrate the performance of `DIMIX` on time-varying networks. For this, we conduct simulations on non-convex (CNN) and convex (linear regression) problems. For the sake of comparison, we apply our algorithm to a corresponding fixed network.

Data and Experimental Setup. For the non-convex setting, we study the problem of

learning the parameters of the LeNet-5 CNN [107] for CIFAR-10 dataset. The LeNet-5 CNN architecture consists of two convolutional layers, two sub-sampling (pooling) layers, and three fully connected layers, with the ReLU activation function. As before we have $n = 20$ nodes and $N = 10,000$ data points, which are distributed among the agents according to some stochastic vector $\mathbf{r}$. For the convex setting, we consider a 100-dimensional linear regression problem. We generate $N = 300$ data points $\{\xi_1, \ldots, \xi_{300}\}$, where $\xi_i = (\boldsymbol{u}_i, v_i)$, with $v_i = \boldsymbol{u}_i^T \tilde{\mathbf{x}} + \epsilon_i$, and $\boldsymbol{u}_i, \tilde{\mathbf{x}} \in \mathbb{R}^{100}$ for all $i \in [300]$. In order to generate the data, we uniformly and independently draw the entries of each $\boldsymbol{u}_i$ and each ϵ_i from $(0, 1)$ and $(0, 0.1)$, respectively. Similarly, entries of $\tilde{\mathbf{x}}$ are sampled uniformly and independently from $(-1, 1)$. The goal is to estimate the unknown coefficient vector $\tilde{\mathbf{x}}$, which leads to solving the minimum mean square error problem, i.e.,

$$f(\mathbf{x}) := \min_{\mathbf{x} \in \mathbb{R}^d} \frac{1}{2N} \sum_{i=1}^{300} \|v_i - \boldsymbol{u}_i^T \mathbf{x}\|^2$$

.

Experiments with Fixed Graph. We consider a fixed undirected cyclic graph $\mathcal{G}^C = ([n], \mathcal{E})$, where $\mathcal{E} = \{(\langle i \rangle, \langle i+1 \rangle) : i \in [n]\}$, and $\langle i \rangle := (i - 1 \bmod n) + 1$. For these experiments, the stochastic matrix sequence $W(t) = W$ for any t is given by

$$W_{ij} = \begin{cases} \frac{r_{\langle j \rangle}}{2(r_{\langle i \rangle} + r_{\langle j \rangle})} & j \in \{\langle i-1 \rangle, \langle i+1 \rangle\} \\ \frac{r_{\langle i \rangle}}{2(r_{\langle i \rangle} + r_{\langle i+1 \rangle})} + \frac{r_{\langle i \rangle}}{2(r_{\langle i \rangle} + r_{\langle i-1 \rangle})} & j = i \\ 0 & \text{otherwise.} \end{cases} \tag{4.11}$$

Experiments with Time-varying Graph. To evaluate Algorithm 3 for a time-varying network, we use cyclic gossip [70, 75, 84, 108, 109] iterations. Here, the algorithm goes through the cycle graph $\mathcal{G}^C$ (described above), and a pair of neighbors (on the cycle) exchange their information at a time. More precisely, at time t, the averaging graph $\mathcal{G}(t) = ([n], \mathcal{E}(t))$ only consists of a single edge $\mathcal{E}(t) = \{(\langle t \rangle, \langle t+1 \rangle)\}$. For $\mathcal{G}(t)$, we let

$$[W(t)]_{ij} = \begin{cases} \frac{r_{\langle j \rangle}}{r_{\langle t \rangle} + r_{\langle t+1 \rangle}} & i, j \in \{\langle t \rangle, \langle t+1 \rangle\} \\ 1 & i = j \notin \{\langle t \rangle, \langle t+1 \rangle\} \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

Note that each edge in $\mathcal{G}^C$ will be visited periodically, and hence, the sequence of stochastic matrices $\{W(t)\}$ in (4.12) is B-connected with $B = n$.

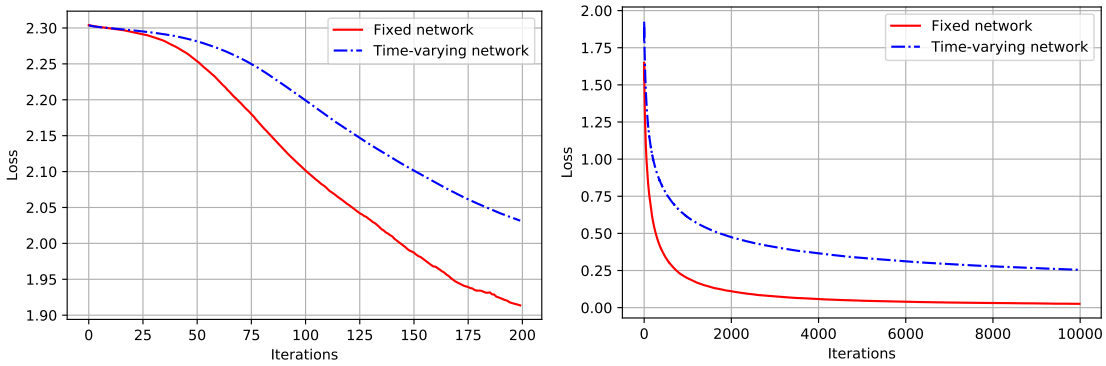

Figure 4.3: Training Loss vs. Iterations: LeNet-5 on CIFAR-10 (left), and Linear Regression (right).

Results. In Figure 4.3, the plot on the left demonstrates the training loss of LeNet-5 for CIFAR-10 dataset and the plot on the right shows the training loss of the linear regression. Here, 'Fixed network' refers to the full cycle with stochastic matrix in (4.11) and 'Time-varying network' refers to the network with stochastic matrix sequence in (4.12). For the CNN, the parameters of the dynamics in (4.2) are fine-tuned to $(\alpha_0, \nu^\star) = (0.02, 1/6)$, $(\beta_0, \mu^\star) = (0.05, 1/2)$, $\tau = 0$, and $s = 6$ is the parameter of the stochastic quantizer for both networks. For the linear regression model, the parameters are adjusted to $(\alpha_0, \nu) = (6, 1/4)$, $(\beta_0, \mu) = (16, 3/4)$, and $\tau = 1500$ with the quantizer parameter $s = 6$. It can be observed that the algorithm works for both fixed and time-varying networks. However, the performance of the algorithm over the time-varying network naturally suffers from a slower convergence, which is due to a slower mixing of information over the network.

4.4 Auxiliary Lemmas

In this section, we present auxiliary lemmas which play crucial roles in the proof of the main result, namely, Theorem 4.1 in Section 4.5. The proofs of Lemmas 4.1–4.6 are provided in Section 4.10. We refer to [110] for the proof of Lemma 4.9.

Lemma 4.1. Let $\{W(t)\}$ satisfy the connectivity Assumption 4.2 with parameters (B,η), and let $\{A(t)\}$ be given by $A(t) = (1-\beta(t))I + \beta(t)W(t)$ where $\beta(t) \in (0,1]$ for all t, and $\{\beta(t)\}$ is a non-increasing sequence. Then, for any matrix $U \in \mathbb{R}^{n\times d}$, and all $t > s \geq 1$, we have

$$\left\|\left(A(t-1)\cdots A(s+1) - \mathbf{1}\mathbf{r}^T\right)U\right\|_{\mathbf{r}}^2 \leq \|U\|_{\mathbf{r}}^2, \tag{4.13}$$

$$\left\|\left(A(t-1)\cdots A(s+1) - \mathbf{1}\mathbf{r}^T\right)U\right\|_{\mathbf{r}}^2 \leq \kappa \prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\,\|U\|_{\mathbf{r}}^2, \tag{4.14}$$

for every $t > s$. Furthermore, we have

$$\left\|\left(A(s+B)\cdots A(s+1) - \mathbf{1}\mathbf{r}^T\right)U\right\|_{\mathbf{r}}^2 \leq (1-\lambda B\beta(s+B))\,\|U\|_{\mathbf{r}}^2, \tag{4.15}$$

for all $s \geq 1$ where $\lambda := \frac{\eta \mathbf{r}_{\min}}{2Bn^2}$, $\kappa := (1-B\lambda\beta_0)^{-1}$ and $\beta_0 = \beta(1)$.

Lemma 4.2. For an $n \times m$ matrix A and $m \times q$ matrix B, we have

$$\|AB\|_{\mathbf{r}} \leq \|A\|_{\mathbf{r}}\,\|B\|_F\,.$$

Lemma 4.3. For any pair of vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and any scalar $\theta > 0$, we have

$$\|\boldsymbol{u}+\boldsymbol{v}\|^2 \leq (1+\theta)\|\boldsymbol{u}\|^2 + \left(1+\frac{1}{\theta}\right)\|\boldsymbol{v}\|^2.$$

Similarly, for matrices U and V and any scalar $\theta > 0$, we get

$$\|U+V\|_{\mathbf{r}}^2 \leq (1+\theta)\,\|U\|_{\mathbf{r}}^2 + \left(1+\frac{1}{\theta}\right)\|V\|_{\mathbf{r}}\,.$$

Lemma 4.4. For any $0 \leq \delta < 1$ and $0 < a < 1$ we have

$$\prod_{k=s}^{t-1}\left(1-\frac{a}{k^\delta}\right) \leq \exp\left(-\frac{a}{1-\delta}\left(t^{1-\delta}-s^{1-\delta}\right)\right).$$

For $\delta = 1$ and $0 \leq a < 1$ we have

$$\prod_{k=s}^{t-1}\left(1-\frac{a}{k}\right) \leq \left(\frac{t}{s}\right)^{-a}.$$

Lemma 4.5. Let $\{\beta(t)\}$ be a sequence in $\mathbb{R}$ and λ be a non-zero scalar. Then the following identities hold for all $t \geq 1$

$$\sum_{s=1}^{t-1}\beta(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k)) = \frac{1}{\lambda} - \frac{1}{\lambda}\prod_{k=1}^{t-1}(1-\lambda\beta(k)), \tag{4.16}$$

As a result, for any sequence $\{\beta(t)\}$ in $[0,1]$ and $\lambda > 0$, we get

$$\sum_{s=1}^{t-1}\beta(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k)) \leq \frac{1}{\lambda}.$$

Lemma 4.6. For any $0 \le \delta < \min(1, \sigma)$, $0 < a \le 1$, and every $t > \tau := \left(\frac{2(\sigma-\delta)}{a}\right)^{\frac{1}{1-\delta}}$, we have

$$\sum_{s=1}^{t-1}\left[\frac{1}{s^\sigma}\prod_{k=s+1}^{t-1}\left(1-\frac{a}{k^\delta}\right)\right] \le A(a,\sigma,\delta)t^{-(\sigma-\delta)},$$

where $A(a,\sigma,\delta)$ is given by

$$A(a,\sigma,\delta) := \begin{cases} 2^\sigma \max\left\{1+\frac{2}{a}, 1+\frac{1}{\sigma-1}\left(\frac{2(\sigma-\delta)}{a}\right)^{\frac{\sigma-\delta}{1-\delta}}\right\} & \text{if } \sigma > 1, \\ 2^\sigma \max\left\{1+\frac{2}{a}, 1+\frac{2}{a}\ln\left(\frac{2(1-\delta)}{a}\right)\right\} & \text{if } \sigma = 1, \\ 2^\sigma \max\left\{1+\frac{2}{a}, 1+\frac{2(\sigma-\delta)}{a(1-\sigma)}\right\} & \text{if } 0 < \sigma < 1. \end{cases} \tag{4.17}$$

Moreover, for $\delta = 1$ and $a - \sigma + 1 \ne 0$, we have

$$\sum_{s=1}^{t-1}\left[\frac{1}{s^\sigma}\prod_{k=s+1}^{t-1}\left(1-\frac{a}{k}\right)\right] \le A(a,\sigma,1)t^{-\min(\sigma-1,a)},$$

where $A(a,\sigma,1) = 2^\sigma\left(1+\frac{1}{|a-\sigma+1|}\right)$.

Lemma 4.7. For non-negative numbers $a, b, c > 0$, if $b \ne 1$ we have

$$(1-at^{-b})t^{-c} \le (t+1)^{-c} \tag{4.18}$$

for every $t \ge t_0 := \left(\frac{c}{a}\right)^{\frac{1}{1-b}}$. Moreover, if $b = 1$, the inequality in (4.18) holds for every $t \ge 1$ provided that $a \ge c$.

Lemma 4.8. For any $\delta \in \mathbb{R}$, $\tau \ge 0$, and $T \ge 1$, we have

$$\sum_{t=1}^{T}(t+\tau)^\delta \le \begin{cases} \frac{\tau^{1+\delta}}{|1+\delta|} & \text{if } \delta < -1, \\ \ln\left(\frac{T}{\tau}+1\right) & \text{if } \delta = -1, \\ \frac{2^{1+\delta}}{1+\delta}(T+\tau)^{1+\delta} & \text{if } \delta > -1. \end{cases} \tag{4.19}$$

Lemma 4.9. [110, Theorem 2.1.11] Suppose that ∇f is Lipschitz continuous with constant K and f is strongly convex with modulus ρ. Then, we have

$$\langle \mathbf{x}-\mathbf{x}^\star, \nabla f(\mathbf{x})\rangle \ge c_1\|\nabla f(\mathbf{x})\|^2 + c_2\|\mathbf{x}-\mathbf{x}^\star\|^2,$$

where $c_1 = \frac{1}{\rho+K}$, $c_2 = \frac{\rho K}{\rho+K}$, and $\nabla f(\mathbf{x}^\star) = 0$.

In this section, we present some results which will be used in the proof of the main theorem. The proof of Lemma 4.1 and Lemma 4.11 are provided in Appendix. We refer to the cited references for the proof of other preliminaries.

The first result is known as Robbins-Siegmund's Theorem [111] as described below.

Lemma 4.10. [111] Suppose that for non-negative random processes $\{w(t)\}, \{p(t)\}, \{u(t)\}$, and $\{q(t)\}$ that are adapted to a filtration $\{\mathcal{F}_t\}$, we have

$$\mathbb{E}\left[w(t+1) \mid \mathcal{F}_t\right] \leq (1+p(t))w(t) - u(t) + q(t), \tag{4.20}$$

almost surely, for all $t \geq 0$. Then, if $\sum_{t=1}^{\infty} q(t) < \infty$ and $\sum_{t=1}^{\infty} p(t) < \infty$ almost surely, we almost surely have $\sum_{t=1}^{\infty} u(t) < \infty$ and $w(t)$ converges almost surely to a (non-negative) random variable w.

The following theorem from [112] is a consequence of the Robbins-Siegmund's Theorem and plays a crucial role in the proof of Theorem 4.4.

Theorem 4.3. *[112, Lemma 3] Let the optimal set $\mathcal{X}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ be nonempty for a convex and continuous function $f : \mathbb{R}^d \to \mathbb{R}$. Moreover, assume $\{\mathbf{y}(t)\}$ is a sequence satisfying*

$$\mathbb{E}\left[\|\mathbf{y}(t+1) - \mathbf{x}^\star\|^2 | \mathcal{F}_t\right] \leq (1+c(t))\|\mathbf{y}(t) - \mathbf{x}^\star\|^2 - u(t)(f(\mathbf{y}(t)) - f(\mathbf{x}^\star)) + z(t),$$

for all $t \geq 1$ and for all $\mathbf{x}^\star \in \mathcal{X}^\star$ almost surely, where non-negative sequences $\zeta(t)$, $\xi(t)$, and $z(t)$ for satisfy $\sum_{t=1}^{\infty} \zeta(t) < \infty$, $\sum_{t=1}^{\infty} \xi(t) = \infty$, and $\sum_{t=1}^{\infty} z(t) < \infty$. Then, the sequence $\{\mathbf{y}(t)\}$ converges to some solution $\tilde{\mathbf{x}} \in \mathcal{X}^\star$ almost surely.

Lemma 4.11. Let $\{p(t)\}$ and $\{q(t)\}$ be two positive and non-increasing sequences for $t \geq 1$ and there exists for $0 < A < 1$ such that

$$-\Delta p(t) \leq A p(t) q(t), \tag{4.21}$$

for every $t \geq t_0$. Then, we have

$$\sum_{s=1}^{t-1} \left[p(s) \prod_{k=s+1}^{t-1} (1 - q(k)) \right] \leq S \frac{p(t)}{q(t)}. \tag{4.22}$$

for every $t \geq t_0$, and some positive S which is not a function of t (but may depend on t_0 and A).

4.5 Proof of Theorem 4.1

In this section, we provide the proof of the main result, namely, Theorem 4.1. The main steps of the proof are twofold: We first bound the deviation of the agents' states from their average, and then analyze the distance of the average state from the global optimal point. These together lead to the proof of the theorem.

4.5.1 State Deviation from the Average State

In this part, we study $C(t) := \mathbb{E}\left[\sum_{i=1}^{n} r_i \left\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\right\|^2\right]$, where $\bar{\mathbf{x}}(t) := \sum_{i=1}^{n} r_i \mathbf{x}_i(t)$ is the avarage of the states at time t. Note that the dynamics in (4.4) can be viewed as the linear time-varying system

$$X(t+1) = A(t)X(t) + U(t), \tag{4.23}$$

with

$$\begin{aligned} A(t) &= ((1-\beta(t))I + \beta(t)W(t)), \\ U(t) &= \beta(t)E(t) - \alpha(t)\beta(t)\nabla f(X(t)). \end{aligned}$$

Therefore, we have

$$X(t) = \sum_{s=1}^{t-1} \Phi(t:s)U(s) + \Phi(t:0)X(1), \tag{4.24}$$

where $\Phi(t:s) = A(t-1)\cdots A(s+1)$ with $\Phi(t:t-1) = I$ is the transition matrix of the linear system (4.23). For the notational simplicity, let us define

$$\begin{aligned} P(t:s) &:= \beta(s)(\Phi(t:s) - \mathbf{1}\mathbf{r}^T) \\ &= \beta(s)\left(A(t-1)\cdots A(s+1) - \mathbf{1}\mathbf{r}^T\right). \end{aligned}$$

As a result of Lemma 4.1, we have

$$\left\|P(t:s)U\right\|_{\mathbf{r}} \leq \pi(t:s)\left\|U\right\|_{\mathbf{r}},$$

where $\pi(t:s)$ is defined by

$$\pi(t:s) := \beta(s)\kappa^{\frac{1}{2}} \prod_{k=s+1}^{t-1} (1-\lambda\beta(k))^{\frac{1}{2}}. \tag{4.25}$$

Assuming $X(1) = \mathbf{0}$, the dynamic in (4.24) reduces to

$$X(t) = \sum_{s=1}^{t-1} \Phi(t:s)U(s). \tag{4.26}$$

Moreover, multiplying both sides of (4.26) from the left by $\mathbf{r}^T$ and using the fact that $\mathbf{r}^T A(t) = \mathbf{r}^T$, we get

$$\bar{\mathbf{x}}(t) := \mathbf{r}^T X(t) = \sum_{s=1}^{t-1} \mathbf{r}^T \Phi(t:s)U(s) = \sum_{s=1}^{t-1} \mathbf{r}^T U(s). \tag{4.27}$$

Then, subtracting (4.27) from (4.26), and plugging the definition of $U(s)$ we have

$$\begin{aligned} X(t) - \mathbf{1}\bar{\mathbf{x}}(t) &= \sum_{s=1}^{t-1} \Phi(t:s)U(s) - \sum_{s=1}^{t-1} \mathbf{1}\mathbf{r}^T U(s) \\ &= \sum_{s=1}^{t-1} (\Phi(t:s) - \mathbf{1}\mathbf{r}^T)U(s) \\ &= \sum_{s=1}^{t-1} \beta(s)(\Phi(t:s) - \mathbf{1}\mathbf{r}^T)\Big[E(s) - \alpha(s)\nabla f(X(s))\Big] \\ &= \sum_{s=1}^{t-1} P(t:s)E(s) - \sum_{s=1}^{t-1} \alpha(s)P(t:s)\nabla f(X(s)). \end{aligned}$$

Using Lemma 4.3 with $\theta = 1$, we get

$$\begin{aligned} \|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2 &\le 2\left\|\sum_{s=1}^{t-1} P(t:s)E(s)\right\|_{\mathbf{r}}^2 + 2\left\|\sum_{s=1}^{t-1} \alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2. \\ &= 2\sum_{s=1}^{t-1} \|P(t:s)E(s)\|_{\mathbf{r}}^2 + 2\sum_{s\neq q} \langle P(t:s)E(s), P(t:q)E(q)\rangle \\ &\quad + 2\left\|\sum_{s=1}^{t-1} \alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2. \end{aligned} \tag{4.28}$$

Using facts that $E(s)$ is measurable with respect to $\mathcal{F}_q$ for $q > s$ and $\mathbb{E}[E(q)|\mathcal{F}_q] = 0$, we have

$$\begin{aligned} \mathbb{E}[\langle P(t:s)E(s), P(t:q)E(q)\rangle] &= \mathbb{E}[\mathbb{E}[\langle P(t:s)E(s), P(t:q)E(q)\rangle|\mathcal{F}_q]] \\ &= \mathbb{E}[\langle P(t:s)E(s), P(t:q)\mathbb{E}[E(q)|\mathcal{F}_q]\rangle] = 0. \end{aligned} \tag{4.29}$$

Using a similar argument for $q < s$ and conditioning on $\mathcal{F}_s$, we conclude that (4.29) holds for all $q \neq s$. Recall that we defined $C(t) = \mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$. Taking expectations of

both sides of (4.28) and using the identity in (4.29), we get

$$C(t) = \mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$$
$$\leq 2\sum_{s=1}^{t-1}\mathbb{E}\left[\|P(t:s)E(s)\|_{\mathbf{r}}^2\right] + 2\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right]. \tag{4.30}$$

We continue with bounding the first term in (4.30). From Assumption 4.1, we have

$$\mathbb{E}\left[\|E(s)\|_{\mathbf{r}}^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|E(s)\|_{\mathbf{r}}^2|\mathcal{F}_s\right]\right] = \mathbb{E}\left[\sum_{i=1}^{n} r_i\mathbb{E}\left[\|\mathbf{e}_i(s)\|^2|\mathcal{F}_s\right]\right] \leq \mathbb{E}\left[\sum_{i=1}^{n} r_i\gamma\right] = \gamma.$$

This together with Lemma 4.1 lead to

$$\sum_{s=1}^{t-1}\mathbb{E}\left[\|P(t:s)E(s)\|_{\mathbf{r}}^2\right] \leq \sum_{s=1}^{t-1}\left[\beta^2(s)\kappa\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\mathbb{E}\left[\|E(s)\|_{\mathbf{r}}^2\right]\right]$$
$$\leq \gamma\kappa\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right] \tag{4.31}$$
$$= \gamma\kappa\sum_{s=1}^{t-1}\left[\frac{\beta_0^2}{s^{2\mu}}\prod_{k=s+1}^{t-1}\left(1-\lambda\frac{\beta_0}{k^{\mu}}\right)\right] \leq \epsilon_1 t^{-\mu}, \tag{4.32}$$

where the last inequality holds for $t \geq T_1$, where

$$T_1 := \left\lceil (2\mu/\lambda\beta_0)^{\frac{1}{1-\mu}}\right\rceil, \tag{4.33}$$

and follows from Lemma 4.6, with parameters $(\sigma, \delta, \tau) = (2\mu, \mu, T_1)$. Moreover, we have $\epsilon_1 := \gamma\kappa\beta_0^2 A(\lambda\beta_0, 2\mu, \mu)$.

Next, we bound the second term in (4.30). Note that $\|\cdot\|_{\mathbf{r}}$ is a norm. Hence, using the triangle inequality we have

$$\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right]$$
$$\leq \mathbb{E}\left[\left(\sum_{s=1}^{t-1}\|\alpha(s)P(t:s)\nabla f(X(s))\|_{\mathbf{r}}\right)^2\right]$$
$$= \sum_{1\leq s,q\leq t-1}\mathbb{E}\Big[\alpha(s)\,\|P(t:s)\nabla f(X(s))\|_{\mathbf{r}}\,\alpha(q)\,\|P(t:q)\nabla f(X(q))\|_{\mathbf{r}}\Big]. \tag{4.34}$$

Using Lemma 4.1 and the fact that $2ab \leq a^2 + b^2$, we can upper-bound this expression as

$$
\begin{aligned}
&\sum_{1\le s,q\le t-1}\mathbb{E}\Big[\alpha(s)\left\|P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}\cdot\alpha(q)\left\|P(t:q)\nabla f(X(q))\right\|_{\mathbf{r}}\Big]\\
&\qquad\le\sum_{1\le s,q\le t-1}\mathbb{E}\Big[\alpha(s)\pi(t:s)\left\|\nabla f(X(s))\right\|_{\mathbf{r}}\cdot\alpha(q)\pi(t:q)\left\|\nabla f(X(q))\right\|_{\mathbf{r}}\Big]\\
&\qquad=\sum_{1\le s,q\le t-1}\pi(t:s)\pi(t:q)\mathbb{E}\Big[\alpha(s)\left\|\nabla f(X(s))\right\|_{\mathbf{r}}\cdot\alpha(q)\left\|\nabla f(X(q))\right\|_{\mathbf{r}}\Big]\\
&\qquad\le\frac{1}{2}\sum_{1\le s,q\le t-1}\pi(t:s)\pi(t:q)\mathbb{E}\Big[\alpha^2(s)\left\|\nabla f(X(s))\right\|_{\mathbf{r}}^2+\alpha^2(q)\left\|\nabla f(X(q))\right\|_{\mathbf{r}}^2\Big]\\
&\qquad=\sum_{1\le s,q\le t-1}\pi(t:s)\pi(t:q)\mathbb{E}\left[\alpha^2(s)\left\|\nabla f(X(s))\right\|_{\mathbf{r}}^2\right]\\
&\qquad=\left(\sum_{q=1}^{t-1}\pi(t:q)\right)\cdot\left(\sum_{s=1}^{t-1}\alpha^2(s)\pi(t:s)\mathbb{E}\left[\left\|\nabla f(X(s))\right\|_{\mathbf{r}}^2\right]\right),
\end{aligned}
\tag{4.35}
$$

where $\pi(t:s)$ is given by (4.25). Then, using the fact that $\sqrt{1-x}\le 1-x/2$ and Lemma 4.5, we get

$$
\begin{aligned}
\sum_{q=1}^{t-1}\pi(t:q)&=\sum_{q=1}^{t-1}\left[\beta(q)\kappa^{\frac{1}{2}}\prod_{k=q+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}\right]\\
&\le\sum_{q=1}^{t-1}\beta(q)\kappa^{\frac{1}{2}}\prod_{k=q+1}^{t-1}\left(1-\frac{\lambda}{2}\beta(k)\right)\le\frac{2}{\lambda}\kappa^{\frac{1}{2}}.
\end{aligned}
\tag{4.36}
$$

Moreover, we can write

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\nabla f(X(s))\right\|_{\mathbf{r}}^2\right]\\
&=\mathbb{E}\left[\left\|\nabla f(X(s))-\nabla f(\mathbf{1}\bar{\mathbf{x}}(s))+\nabla f(\mathbf{1}\bar{\mathbf{x}}(s))-\nabla f(\mathbf{1}\mathbf{x}^\star)+\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2\right]\\
&\overset{(a)}{\le}\mathbb{E}\left[3\left\|\nabla f(X(s))-\nabla f(\mathbf{1}\bar{\mathbf{x}}(s))\right\|_{\mathbf{r}}^2+3\left\|\nabla f(\mathbf{1}\bar{\mathbf{x}}(s))-\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2+3\left\|\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2\right]\\
&=3\mathbb{E}\left[\left\|\nabla f(X(s))-\nabla f(\mathbf{1}\bar{\mathbf{x}}(s))\right\|_{\mathbf{r}}^2\right]+3\mathbb{E}\left[\left\|\nabla f(\mathbf{1}\bar{\mathbf{x}}(s))-\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2\right]+3\mathbb{E}\left[\left\|\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2\right]\\
&=3\mathbb{E}\left[\sum_{i=1}^n r_i\left\|\nabla f_i(\mathbf{x}_i(t))-\nabla f_i(\bar{\mathbf{x}}(s))\right\|^2\right]+3\mathbb{E}\left[\sum_{i=1}^n r_i\left\|\nabla f_i(\bar{\mathbf{x}}(s))-\nabla f_i(\mathbf{x}^\star)\right\|^2\right]+3\left\|\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2\\
&\overset{(b)}{\le}3\mathbb{E}\left[\sum_{i=1}^n r_iK^2\left\|\mathbf{x}_i(t)-\bar{\mathbf{x}}(s)\right\|^2\right]+3\mathbb{E}\left[\sum_{i=1}^n r_iK^2\left\|\bar{\mathbf{x}}(s)-\mathbf{x}^\star\right\|^2\right]+3\left\|\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2\\
&=3K^2\mathbb{E}\left[\left\|X(t)-\mathbf{1}\bar{\mathbf{x}}(s)\right\|_{\mathbf{r}}^2\right]+3K^2\mathbb{E}\left[\left\|\bar{\mathbf{x}}(s)-\mathbf{x}^\star\right\|^2\right]+3\left\|\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2\\
&=3K^2\left(C(s)+Q(s)\right)+3\left\|\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2,
\end{aligned}
\tag{4.37}
$$

where $Q(s) \coloneqq \mathbb{E}\left[\|\bar{\mathbf{x}}(s) - \mathbf{x}^\star\|^2\right]$, the inequality (a) follows from $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, and step (b) holds due to Assumption 3.1. Plugging (4.36) and (4.37) into (4.35) and (4.34), we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right] \\
&\le \frac{6}{\lambda}\kappa^{\frac{1}{2}}K^2\sum_{s=1}^{t-1}\alpha^2(s)\pi(t:s)\left(C(s)+Q(s)\right) + \frac{6}{\lambda}\kappa^{\frac{1}{2}}\left\|\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2\sum_{s=1}^{t-1}\alpha^2(s)\pi(t:s). \qquad (4.38)
\end{aligned}$$

To bound the second term in (4.38), we can write

$$\begin{aligned}
\sum_{s=1}^{t-1}\alpha^2(s)\pi(t:s) &\overset{(a)}{=} \sum_{s=1}^{t-1}\alpha^2(s)\left[\beta(s)\kappa^{\frac{1}{2}}\prod_{k=s+1}^{t-1}\left(1-\lambda\beta(k)\right)^{\frac{1}{2}}\right] \\
&\overset{(b)}{\le} \sum_{s=1}^{t-1}\alpha^2(s)\beta(s)\kappa^{\frac{1}{2}}\prod_{k=s+1}^{t-1}\left(1-\frac{\lambda}{2}\beta(k)\right) \\
&= \alpha_0^2\beta_0\kappa^{\frac{1}{2}}\sum_{s=1}^{t-1}\frac{1}{s^{2\nu+\mu}}\prod_{k=s+1}^{t-1}\left(1-\frac{\lambda\beta_0}{2}\frac{1}{k^\mu}\right) \\
&\overset{(c)}{\le} \alpha_0^2\beta_0\kappa^{\frac{1}{2}}A(\lambda\beta_0/2, 2\nu+\mu, \mu)t^{-2\nu}, \qquad (4.39)
\end{aligned}$$

for every $t \ge T_2$ where

$$T_2 \coloneqq \left\lceil (8\nu/\lambda\beta_0)^{\frac{1}{1-\mu}}\right\rceil. \qquad (4.40)$$

Note that in the chain of inequalities in (4.39), we used the definition of $\pi(t:s)$ from (4.25) in step (a), the inequality in (b) follows from $\sqrt{1-x} \le 1 - x/2$ for $x \le 1$, and we used Lemma 4.6 with $(\sigma, \delta, \tau) = (2\nu+\mu, \mu, T_2)$ in (c). Using (4.39) in (4.38), we get

$$\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right] \le \frac{6}{\lambda}\kappa^{\frac{1}{2}}K^2\sum_{s=1}^{t-1}\alpha^2(s)\pi(t:s)\left(C(s)+Q(s)\right) + \epsilon_2 t^{-2\nu}, \qquad (4.41)$$

where $\epsilon_2 \coloneqq \frac{6}{\lambda}\kappa\left\|\nabla f(\mathbf{1}\mathbf{x}^\star)\right\|_{\mathbf{r}}^2\alpha_0^2\beta_0 A(\lambda\beta_0/2, 2\nu+\mu, \mu)$ is a constant. Finally, using the bounds obtained in (4.32) and (4.41) in (4.30), we get

$$C(t) \le 2\epsilon_1 t^{-\mu} + \frac{12}{\lambda}\kappa^{\frac{1}{2}}K^2\sum_{s=1}^{t-1}\alpha^2(s)\pi(t:s)\left(C(s)+Q(s)\right) + 2\epsilon_2 t^{-2\nu}. \qquad (4.42)$$

4.5.2 Average State Distance to the Optimal Point

Now, we derive an upper bound for the average distance between the mean of the agents' states and the global optimal point, i.e., $Q(t) = \mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2\right]$, where $\mathbf{x}^\star$ is the minimizer of the function $f(\mathbf{x})$. Recall that $\bar{\mathbf{x}}(t) = \mathbf{r}^T X(t) = \sum_{i=1}^n r_i \mathbf{x}_i(t)$ and $\mathbf{r}^T W(t) = \mathbf{r}^T$. Hence, multiplying both sides of (4.4) by $\mathbf{r}^T$, we get

$$\bar{\mathbf{x}}(t+1) = \bar{\mathbf{x}}(t) + \beta(t)\mathbf{r}^T E(t) - \alpha(t)\beta(t)\mathbf{r}^T \nabla f(X(t)).$$

We define $g(t) := \mathbf{r}^T \nabla f(X(t)) = \sum_{i=1}^n r_i \nabla f_i(\mathbf{x}_i(t))$ and $\bar{g}(t) := \nabla f(\bar{\mathbf{x}}(t)) = \sum_{i=1}^n r_i \nabla f_i(\bar{\mathbf{x}}(t))$. Hence, we can write

$$\begin{aligned}\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1) - \mathbf{x}^\star\|^2 \middle| \mathcal{F}_t\right] &= \mathbb{E}\left[\|\bar{\mathbf{x}}(t) + \beta(t)\mathbf{r}^T E(t) - \alpha(t)\beta(t)g(t) - \mathbf{x}^\star\|^2 \middle| \mathcal{F}_t\right] \\ &= \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)g(t)\|^2 + \mathbb{E}\left[\|\beta(t)\mathbf{r}^T E(t)\|^2 \middle| \mathcal{F}_t\right], \end{aligned} \tag{4.43}$$

where the last equality follows from the fact that $X(t)$ is measurable with respect to $\mathcal{F}_t$ and Assumption 4.1 implying $\mathbb{E}\left[\beta(t)\mathbf{r}^T E(t) \middle| \mathcal{F}_t\right] = 0$, which leads to

$$2\left\langle \bar{\mathbf{x}}(t) - \alpha(t)\beta(t)g(t) - \mathbf{x}^\star, \mathbb{E}\left[\beta(t)\mathbf{r}^T E(t) \middle| \mathcal{F}_t\right]\right\rangle = 0.$$

Taking expectations of both sides of (4.43), and using the tower rule, we get

$$Q(t+1) = \mathbb{E}\left[\|\bar{\mathbf{x}}(t+1) - \mathbf{x}^\star\|^2\right] = \mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)g(t)\|^2\right] + \mathbb{E}\left[\|\beta(t)\mathbf{r}^T E(t)\|^2\right]. \tag{4.44}$$

In order to bound the first term in (4.43), we use Lemma 4.3 for vectors $\boldsymbol{u} = \bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)$ and $\boldsymbol{v} = \alpha(t)\beta(t)(\bar{g}(t) - g(t))$, and parameter $\theta = \frac{\rho K}{\rho + K}\alpha(t)\beta(t)$. Hence, we can write

$$\begin{aligned}\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)g(t)\|^2 &= \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t) + \alpha(t)\beta(t)\bar{g}(t) - \alpha(t)\beta(t)g(t)\|^2 \\ &\leq \left(1 + \frac{\rho K}{\rho + K}\alpha(t)\beta(t)\right)\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2 \\ &\quad + \alpha(t)\beta(t)\left(\alpha(t)\beta(t) + \frac{\rho + K}{\rho K}\right)\|\bar{g}(t) - g(t)\|^2. \end{aligned} \tag{4.45}$$

Next, we use Lemma 4.9 to bound the first term in (4.45). Note that Assumptions 3.1 and 3.2 guarantee that the conditions of Lemma 4.9 are satisfied. Thus, we have

$$\langle \bar{\mathbf{x}}(t) - \mathbf{x}^\star, \nabla f(\bar{\mathbf{x}}(t))\rangle \geq c_1\|\nabla f(\bar{\mathbf{x}}(t))\|^2 + c_2\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2,$$

or equivalently,

$$\langle \bar{\mathbf{x}}(t) - \mathbf{x}^\star, \bar{g}(t) \rangle \geq c_1 \|\bar{g}(t)\|^2 + c_2 \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2, \tag{4.46}$$

where $c_1 = \frac{1}{\rho+K}$ and $c_2 = \frac{\rho K}{\rho+K}$. Therefore, for the first term in (4.45), we can write

$$\begin{aligned}
&\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2 \\
&= \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2 + \alpha^2(t)\beta^2(t)\|\bar{g}(t)\|^2 - 2\alpha(t)\beta(t)\langle \bar{\mathbf{x}}(t) - \mathbf{x}^\star, \bar{g}(t)\rangle \\
&\leq (1 - 2c_2\alpha(t)\beta(t))\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2 + \alpha(t)\beta(t)(\alpha(t)\beta(t) - 2c_1)\|\bar{g}(t)\|^2.
\end{aligned} \tag{4.47}$$

Let us set

$$T_3 := \left\lceil \left(\frac{\alpha_0\beta_0}{2c_1} \right)^{\frac{1}{\mu+\nu}} \right\rceil = \left\lceil \left(\frac{\alpha_0\beta_0(\rho+K)}{2} \right)^{\frac{1}{\mu+\nu}} \right\rceil, \tag{4.48}$$

such that $\alpha(t)\beta(t) \leq 2c_1$ for any $t \geq T_3$. Hence, for $t \geq T_3$, the second term in (4.47) is non-positive, and thus

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2 \leq (1 - 2c_2\alpha(t)\beta(t))\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2.$$

Taking expectations of both sides, we get

$$\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2\right] \leq (1 - 2c_2\alpha(t)\beta(t))Q(t). \tag{4.49}$$

The average of the second term in (4.45) can be bounded as

$$\begin{aligned}
\mathbb{E}\big[\|\bar{g}(t) - g(t)\|^2\big] &= \mathbb{E}\left[\left\|\sum_{i=1}^n r_i \left((\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla f_i(\mathbf{x}_i(t))\right)\right\|^2\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\sum_{i=1}^n r_i \|\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla f_i(\mathbf{x}_i(t))\|^2\right] \\
&\overset{(b)}{\leq} K^2 \sum_{i=1}^n r_i \mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}_i(t)\|^2\right] \\
&= K^2 \mathbb{E}\left[\|\mathbf{1}\bar{\mathbf{x}}(t) - X(t)\|_{\mathbf{r}}^2\right] = K^2 C(t),
\end{aligned} \tag{4.50}$$

where (a) follows from the convexity of $\|\cdot\|^2$ and the inequality in (b) holds due to Assumption 3.1.

Taking expectations of both sides of (4.45) and recalling that $c_2 = \rho K/(\rho + K)$, we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)g(t)\|^2\right] \\
&\quad \le (1 + c_2\alpha(t)\beta(t))\,\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2\right] \\
&\qquad + \alpha(t)\beta(t)\left(\alpha(t)\beta(t) + 1/c_2\right)\mathbb{E}\left[\|\bar{g}(t) - g(t)\|^2\right] \\
&\quad \overset{(c)}{\le} (1 + c_2\alpha(t)\beta(t))(1 - 2c_2\alpha(t)\beta(t))Q(t) \\
&\qquad + \alpha(t)\beta(t)(\alpha(t)\beta(t) + 1/c_2)K^2C(t) \\
&\quad \overset{(d)}{\le} (1 - c_2\alpha(t)\beta(t))Q(t) + \alpha(t)\beta(t)(\alpha(t)\beta(t) + 1/c_2)K^2C(t),
\end{aligned} \tag{4.51}$$

where the inequality in (c) follows from (4.49) and (4.50), and (d) holds since

$$(1 + c_2\alpha(t)\beta(t))(1 - 2c_2\alpha(t)\beta(t)) \le 1 - c_2\alpha(t)\beta(t).$$

From Assumption 4.1, we get

$$\begin{aligned}
\left[\mathbb{E}\left[|E(t)E^T(t)|\big|\mathcal{F}_t\right]\right]_{ij} &= \mathbb{E}\left[|\mathbf{e}_i(t)\mathbf{e}_j^T(t)|\big|\mathcal{F}_t\right] \\
&\le \sqrt{\mathbb{E}\left[\|\mathbf{e}_i(t)\|^2|\mathcal{F}_t\right]\mathbb{E}\left[\|\mathbf{e}_j(t)\|^2|\mathcal{F}_t\right]} \le \gamma,
\end{aligned} \tag{4.52}$$

for all $1 \le i, j \le n$. Thus, for the second term in (4.43), we arrive at

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{r}^T E(t)\right\|^2\Big|\mathcal{F}_t\right] &= \mathbf{r}^T\mathbb{E}\left[E(t)E^T(t)\big|\mathcal{F}_t\right]\mathbf{r} \\
&\le \mathbf{r}^T\mathbb{E}\left[|E(t)E^T(t)|\big|\mathcal{F}_t\right]\mathbf{r} \\
&\le \mathbf{r}^T(\gamma\mathbf{1}\mathbf{1}^T)\mathbf{r} = \gamma.
\end{aligned} \tag{4.53}$$

Note that we used the fact that $\mathbf{r}^T\mathbf{1} = 1$. Taking expectations from both sides of (4.53), and using the tower rule, we arrive at

$$\mathbb{E}\left[\left\|\mathbf{r}^T E(t)\right\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{r}^T E(t)\right\|^2\Big|\mathcal{F}_t\right]\right] \le \gamma. \tag{4.54}$$

Using (4.51) and (4.54) in (4.44), we can write

$$Q(t+1) \le (1 - c_2\alpha(t)\beta(t))Q(t) + \alpha(t)\beta(t)(\alpha(t)\beta(t) + 1/c_2)K^2C(t) + \gamma\beta^2(t). \tag{4.55}$$

Now, we have two inequalities between $C(t)$ and $Q(t)$, namely (4.42) and (4.55). In the following, we use these inequalities to show that both $C(t)$ and $Q(t)$ vanish as

t grows and conclude that the state of all agents converges to $\mathbf{x}^\star$ in expectation. More precisely, we show that

$$Q(t) \le Dt^{-\min(\mu-\nu,2\nu)}, \quad \text{and} \quad C(t) \le \frac{c_2^2}{2K^2} Dt^{-\min(\mu,2\nu)}, \tag{4.56}$$

for every $t \ge T_0$, where

$$D := \max\left(\max_{1\le t\le T_0} Q(t)t^{\min(\mu-\nu,2\nu)}, \max_{1\le t\le T_0} \frac{2K^2}{c_2^2} C(t)t^{\min(\mu,2\nu)}, \frac{8\gamma\beta_0}{c_2\alpha_0}, \frac{8(\epsilon_1+\epsilon_2)K^2}{c_2^2} \right), \tag{4.57}$$

and $T_0 = \max(T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8)$.

Note that T_1, T_2, and T_3 are defined above in (4.33), (4.40), (4.48), and T_4, T_5, T_6, T_7 and T_8 will be determined as in (4.62), (4.65), (4.67), (4.70), and (4.71), respectively.

We use induction to prove (4.56). First, for the starting point T_0 the definition D in (4.57) implies $Q(T_0) \le Dt^{-\min(\mu-\nu,2\nu)}$ and $C(T_0) \le \frac{c_2^2}{2K^2} Dt^{-\min(\mu,2\nu)}$. Next, we assume that (4.56) holds for every $s \le t$, and prove that $t+1$, i.e, we show that $Q(t+1) \le Dt^{-\min(\mu-\nu,2\nu)}$ and $C(t+1) \le \frac{c_2^2}{2K^2} Dt^{-\min(\mu,2\nu)}$.

Using the facts that $\alpha(t) \le \alpha(T_0) \le \alpha(T_3)$ and $\beta(t) \le \beta(T_0) \le \beta(T_3)$ for any $t \ge T_0 \ge T_3$, we can write

$$c_2\alpha(t)\beta(t) \le c_2\alpha(T_3)\beta(T_3) = c_2\alpha_0\beta_0 \frac{1}{T_3^{\mu+\nu}} \overset{(a)}{\le} c_2\alpha_0\beta_0 \frac{2c_1}{\alpha_0\beta_0} = 2c_2c_1 = \frac{2\rho K}{(\rho+K)^2} \overset{(b)}{\le} \frac{1}{2}, \tag{4.58}$$

where in (a) we used the definition of T_3 in (4.48), and (b) follows from the fact that $(a+b)^2 \ge 4ab$. Then, from (4.55) we have

$$\begin{aligned}
Q(t+1) &\le (1 - c_2\alpha(t)\beta(t))Q(t) + \alpha(t)\beta(t)(c_2\alpha(t)\beta(t)+1)\frac{K^2}{c_2}C(t) + \gamma\beta^2(t) \\
&\overset{(c)}{\le} (1 - c_2\alpha(t)\beta(t))Q(t) + \frac{3}{2}\alpha(t)\beta(t)\frac{K^2}{c_2}C(t) + \gamma\beta^2(t) \\
&\overset{(d)}{\le} (1 - c_2\alpha(t)\beta(t))Dt^{-\min(\mu-\nu,2\nu)} + \frac{3}{2}\alpha(t)\beta(t)\frac{K^2}{c_2}\frac{c_2^2}{2K^2}Dt^{-\min(\mu,2\nu)} + \gamma\beta^2(t) \\
&= \left(1 - \frac{1}{4}c_2\alpha(t)\beta(t)\right) Dt^{-\min(\mu-\nu,2\nu)} + \gamma\beta^2(t) \\
&= \left(1 - \frac{1}{8}c_2\alpha(t)\beta(t)\right) Dt^{-\min(\mu-\nu,2\nu)} + \beta(t)\left(\gamma\beta(t) - \frac{1}{8}c_2 D\alpha(t)t^{-\min(\mu-\nu,2\nu)}\right),
\end{aligned} \tag{4.59}$$

where (c) follows from (4.58), for the inequality in (d) we used the induction assumption. Note that for the last term in (4.59) we have

$$\frac{1}{8}c_2 D\alpha(t)t^{-\min(\mu-\nu,2\nu)} = \frac{c_2 D\alpha_0}{8}t^{-(\nu+\min(\mu-\nu,2\nu))} = \frac{c_2 D\alpha_0}{8}t^{-\min(\mu,3\nu)} \geq \frac{c_2 D\alpha_0}{8}t^{-\mu} \geq \gamma\beta_0 t^{-\mu}, \tag{4.60}$$

where the last inequality holds provided that $D \geq \frac{8\gamma\beta_0}{c_2\alpha_0}$. Plugging (4.60) into (4.59), we arrive at

$$Q(t+1) \leq \left(1 - \frac{1}{8}c_2\alpha(t)\beta(t)\right) D \cdot t^{-\min(\mu-\nu,2\nu)} \overset{\text{(e)}}{\leq} D \cdot (t+1)^{-\min(\mu-\nu,2\nu)}, \tag{4.61}$$

where the inequality in (e) follows from Lemma 4.7 for $(a,b,c) = (c_2\alpha_0\beta_0/8, \mu+\nu, \min(\mu-\nu,2\nu))$, and $\mu+\nu \neq 1$, which holds for

$$t \geq T_4 := \begin{cases} \left\lceil \left(\frac{8\min(\mu-\nu,2\nu)}{c_2\alpha_0\beta_0}\right)^{\frac{1}{1-\mu-\nu}} \right\rceil & \text{if } \mu+\nu < 1, \\ 1 & \text{if } \mu+\nu = 1. \end{cases} \tag{4.62}$$

Note that when $\mu+\nu = 1$, Lemma 4.7 implies that the inequality in (4.61) holds for every $t \geq 1$, provided that $c_2\alpha_0\beta_0 \geq 8\min(\mu-\nu, 2\nu)$. This completes the proof of induction for $Q(t)$. Next, for $C(t)$ from (4.42), we can write

$$C(t+1) \leq 2\epsilon_1(t+1)^{-\mu} + \frac{12}{\lambda}\kappa^{\frac{1}{2}}K^2\sum_{s=1}^{t}\alpha^2(s)\pi(t+1:s)\left(C(s)+Q(s)\right) + 2\epsilon_2(t+1)^{-2\nu}. \tag{4.63}$$

Recall the induction assumption $C(s) \leq \frac{c_2^2}{2K^2}Ds^{-\min(\mu,2\nu)}$ and $Q(s) \leq Ds^{-\min(\mu-\nu,2\nu)}$ for every $s \leq t$. Therefore, we can bound the summations in (4.63) using a chain of inequalities similar to those used in (4.39) and Lemma 4.6 for $(a,\sigma,\delta) = (\lambda\beta_0/2, 2\nu+\mu+\min(\mu,2\nu), \mu)$. We have

$$\begin{aligned}
&\sum_{s=1}^{t}\alpha^2(s)\pi(t+1:s)C(s) \\
&\leq \sum_{s=1}^{t}\alpha^2(s)C(s)\left[\beta(s)\kappa^{\frac{1}{2}}\prod_{k=s+1}^{t}(1-\lambda\beta(k))^{\frac{1}{2}}\right] \\
&\leq \sum_{s=1}^{t}\alpha^2(s)\beta(s)C(s)\kappa^{\frac{1}{2}}\prod_{k=s+1}^{t}\left(1-\frac{\lambda}{2}\beta(k)\right) \\
&= \frac{c_2^2}{2K^2}D\alpha_0^2\beta_0\kappa^{\frac{1}{2}}\sum_{s=1}^{t}\frac{1}{s^{2\nu+\mu+\min(\mu,2\nu)}}\prod_{k=s+1}^{t}\left(1-\frac{\lambda\beta_0}{2}\frac{1}{k^{\mu}}\right) \\
&\leq \frac{c_2^2}{2K^2}D\alpha_0^2\beta_0\kappa^{\frac{1}{2}}A(\lambda\beta_0/2, 2\nu+\mu+\min(\mu,2\nu),\mu)(t+1)^{-2\nu-\min(\mu,2\nu)},
\end{aligned} \tag{4.64}$$

for every

$$t \geq T_5 := \left\lceil \left(\frac{8\nu + 4\min(\mu, 2\nu)}{\lambda\beta_0} \right)^{\frac{1}{1-\mu}} \right\rceil. \tag{4.65}$$

Similarly, we have

$$\begin{aligned}
&\sum_{s=1}^{t} \alpha^2(s)\pi(t+1:s)Q(s) \\
&\leq \sum_{s=1}^{t} \alpha^2(s)Q(s) \left[\beta(s)\kappa^{\frac{1}{2}} \prod_{k=s+1}^{t} (1-\lambda\beta(k))^{\frac{1}{2}} \right] \\
&\leq \sum_{s=1}^{t} \alpha^2(s)\beta(s)Q(s)\kappa^{\frac{1}{2}} \prod_{k=s+1}^{t} \left(1 - \frac{\lambda}{2}\beta(k)\right) \\
&= D\alpha_0^2\beta_0\kappa^{\frac{1}{2}} \sum_{s=1}^{t} \frac{1}{s^{2\nu+\mu+\min(\mu-\nu,2\nu)}} \prod_{k=s+1}^{t} \left(1 - \frac{\lambda\beta_0}{2}\frac{1}{k^{\mu}}\right) \\
&\leq D\alpha_0^2\beta_0\kappa^{\frac{1}{2}} A(\lambda\beta_0/2, 2\nu+\mu+\min(\mu-\nu,2\nu), \mu)(t+1)^{-2\nu-\min(\mu-\nu,2\nu)},
\end{aligned} \tag{4.66}$$

for every

$$t \geq T_6 := \left\lceil \left(\frac{8\nu + 4\min(\mu-\nu, 2\nu)}{\lambda\beta_0} \right)^{\frac{1}{1-\mu}} \right\rceil. \tag{4.67}$$

Plugging (4.64) and (4.66) in (4.63), we arrive at

$$\begin{aligned}
C(t+1) &\leq 2\epsilon_1(t+1)^{-\mu} + \epsilon_4 D(t+1)^{-2\nu-\min(\mu,2\nu)} + \epsilon_5 D(t+1)^{-2\nu-\min(\mu-\nu,2\nu)} + 2\epsilon_2(t+1)^{-2\nu} \\
&\leq \left(2\epsilon_1 + \epsilon_4 D(t+1)^{-2\nu} + \epsilon_5 D(t+1)^{-\nu} + 2\epsilon_2\right)(t+1)^{-\min(\mu,2\nu)},
\end{aligned} \tag{4.68}$$

where $\epsilon_4 := \frac{6}{\lambda}\kappa c_2^2\alpha_0^2\beta_0 A(\lambda\beta_0/2, 2\nu+\mu+\min(\mu,2\nu),\mu)$ and $\epsilon_5 := \frac{12}{\lambda}\kappa K^2\alpha_0^2\beta_0 A(\lambda\beta_0/2, 2\nu+\mu+\min(\mu-\nu,2\nu),\mu)$. Hence, in order to complete the induction for $C(t)$, it suffices to show that

$$2\epsilon_1 + \epsilon_4 D(t+1)^{-2\nu} + \epsilon_5 D(t+1)^{-\nu} + 2\epsilon_2 \leq \frac{c_2^2}{2K^2} D, \tag{4.69}$$

for $t \geq T_0$. Note for

$$t \geq T_0 \geq T_7 := \left\lceil \left(\frac{8K^2\epsilon_4}{c_2^2} \right)^{\frac{1}{2\nu}} \right\rceil, \tag{4.70}$$

we have $\epsilon_4(t+1)^{-2\nu} \leq \frac{c_2^2}{8K^2}$. Similarly, when

$$t \geq T_0 \geq T_8 := \left\lceil \left(\frac{8K^2\epsilon_5}{c_2^2} \right)^{\frac{1}{\nu}} \right\rceil, \tag{4.71}$$

we get $\epsilon_5(t+1)^{-\nu} \leq \frac{c_2^2}{8K^2}$. Therefore, for $t \geq \max(T_7, T_8)$, the inequality in (4.69) holds provided that

$$2\epsilon_1 + 2\epsilon_2 \leq \frac{c_2^2}{4K^2} D, \tag{4.72}$$

which clearly holds for $D \geq \frac{8(\epsilon_1+\epsilon_2)K^2}{c_2^2}$. This concludes the induction proof for $C(t)$.

4.5.3 Total State Deviation from the Optimum Solution

In the previous section, we identified a rate (and conditions) for which the deviation between the states and their average $C(t) = \mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$ vanishes as t grows. We established a similar result for the distance between the states' average and the optimum solution $Q(t) = \mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|_{\mathbf{r}}^2\right]$. Combining these bounds using Lemma 4.3 with $\theta = 1$, we can conclude the proof of Theorem 4.1. In particular, we have

$$\begin{aligned}
\mathbb{E}\left[\|X(T) - \mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right] &= \mathbb{E}\left[\|X(T) - \mathbf{1}\bar{\mathbf{x}}(T) + \mathbf{1}\bar{\mathbf{x}}(T) - \mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right] \\
&\leq 2\left(\mathbb{E}\left[\|X(T) - \mathbf{1}\bar{\mathbf{x}}(T)\|_{\mathbf{r}}^2\right] + \mathbb{E}\left[\|\mathbf{1}\bar{\mathbf{x}}(T) - \mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right]\right) \\
&= 2\left(\mathbb{E}\left[\|X(T) - \mathbf{1}\bar{\mathbf{x}}(T)\|_{\mathbf{r}}^2\right] + \mathbb{E}\left[\|\bar{\mathbf{x}}(T) - \mathbf{x}^\star\|^2\right]\right) \\
&\leq \frac{c_2^2}{K^2} D t^{-\min(\mu, 2\nu)} + 2D t^{\min(\mu-\nu, 2\nu)},
\end{aligned}$$

for every $T \geq T_0 = \max(T_1, T_2, T_0, T_4, T_5, T_6, T_7, T_8)$. This implies the claim of Theorem 4.1, where

$$\xi_1 := c_2^2/K^2, \qquad \xi_2 := 2D. \tag{4.73}$$

Note that when $\mu + \nu = 1$, the bound on $Q(t)$ only holds when $c_2\alpha_0\beta_0 \geq 8\min(\mu - \nu, 2\nu)$. This completes the proof Theorem 4.1.

4.6 Proof of Theorem 4.2

For our analysis, first we obtain an expression for the average reduction of the objective function $f(\cdot)$ at the average of the states, i.e., $\bar{\mathbf{x}}(t) = \sum_{i=1}^{n} r_i \mathbf{x}_i(t) = \mathbf{r}^T X(t)$. Recall that $\mathbf{r}^T W(t) = \mathbf{r}^T$ for all $t \geq 1$. Hence, multiplying both sides of (4.4) by $\mathbf{r}^T$, we get

$$\bar{\mathbf{x}}(t+1) = \bar{\mathbf{x}}(t) + \beta(t)\mathbf{r}^T E(t) - \alpha(t)\beta(t)\mathbf{r}^T \nabla f(X(t)).$$

From Assumption 3.1 and Lemma 3.4 in [113] we can conclude

$$f(\bar{\mathbf{x}}(t+1)) - f(\bar{\mathbf{x}}(t)) - \langle \nabla f(\bar{\mathbf{x}}(t)), \bar{\mathbf{x}}(t+1) - \bar{\mathbf{x}}(t) \rangle \leq \frac{K}{2} \left\| \bar{\mathbf{x}}(t+1) - \bar{\mathbf{x}}(t) \right\|^2,$$

or equivalently,

$$\begin{aligned} f(\bar{\mathbf{x}}(t+1)) \leq & f(\bar{\mathbf{x}}(t)) + \beta(t) \left\langle \nabla f(\bar{\mathbf{x}}(t)), \mathbf{r}^T E(t) \right\rangle - \alpha(t)\beta(t) \left\langle \nabla f(\bar{\mathbf{x}}(t)), \mathbf{r}^T \nabla f(X(t)) \right\rangle \\ & + \beta(t)^2 \frac{K}{2} \left\| \mathbf{r}^T E(t) - \alpha(t) \mathbf{r}^T \nabla f(X(t)) \right\|^2. \end{aligned}$$

Since $\{X(t)\}$ is adapted to the filtration $\{\mathcal{F}_t\}$ and using Assumption 4.1, we arrive at

$$\begin{aligned} \mathbb{E}\left[f\left(\bar{\mathbf{x}}(t+1)\right) | \mathcal{F}_t \right] \leq & f\left(\bar{\mathbf{x}}(t)\right) - \alpha(t)\beta(t) \left\langle \nabla f\left(\bar{\mathbf{x}}(t)\right), \mathbf{r}^T \nabla f(X(t)) \right\rangle \\ & + \beta^2(t) \frac{K}{2} \mathbb{E}\left[\left\| \mathbf{r}^T \left(E(t) - \alpha(t) \nabla f(X(t)) \right) \right\|^2 \middle| \mathcal{F}_t \right]. \end{aligned} \tag{4.74}$$

Using the identity $2\langle \mathbf{x} - \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{x}\|^2$, we can write

$$\begin{aligned} \left\langle \nabla f\left(\bar{\mathbf{x}}(t)\right), \mathbf{r}^T \nabla f(X(t)) \right\rangle = & \left\| \nabla f\left(\bar{\mathbf{x}}(t)\right) \right\|^2 + \left\| \mathbf{r}^T \nabla f(X(t)) \right\|^2 \\ & - 2 \| \mathbf{r}^T \nabla f(X(t)) - \nabla f\left(\bar{\mathbf{x}}(t)\right) \|^2. \end{aligned} \tag{4.75}$$

Moreover, we have

$$\begin{aligned} \| \mathbf{r}^T \nabla f(X(t)) - \nabla f(\bar{\mathbf{x}}(t)) \|^2 = & \left\| \sum_{i=1}^{n} r_i \left(\nabla f_i(\mathbf{x}_i(t)) - \nabla f_i(\bar{\mathbf{x}}(t)) \right) \right\|^2 \\ \leq & \sum_{i=1}^{n} r_i \left\| \nabla f_i(\mathbf{x}_i(t)) - \nabla f_i(\bar{\mathbf{x}}(t)) \right\|^2 \leq K^2 \sum_{i=1}^{n} r_i \| \mathbf{x}_i(t) - \bar{\mathbf{x}}(t) \|^2 = K^2 \left\| X(t) - \mathbf{1}\bar{\mathbf{x}}(t) \right\|_{\mathbf{r}}, \end{aligned} \tag{4.76}$$

where the first inequality holds as $\|\cdot\|^2$ is a convex function and $\mathbf{r}$ is a stochastic vector, and the second inequality follows from Assumption 3.1.

Next, we analyze the last term in (4.74). Note that Assumption 4.1 implies that $\mathbb{E}[E(t)|\mathcal{F}_t] = 0$, which leads to

$$\mathbb{E}\left[\left\|\mathbf{r}^T(E(t)-\alpha(t)\nabla f(X(t)))\right\|^2\middle|\mathcal{F}_t\right] = \mathbb{E}\left[\left\|\mathbf{r}^T E(t)\right\|^2\middle|\mathcal{F}_t\right] + \left\|\alpha(t)\mathbf{r}^T\nabla f(X(t))\right\|^2. \quad (4.77)$$

This together with (4.53) leads us to

$$\mathbb{E}\left[\left\|\mathbf{r}^T\left(E(t)-\alpha(t)\nabla f(X(t))\right)\right\|^2\middle|\mathcal{F}_t\right] \le \gamma + \alpha^2(t)\|\mathbf{r}^T\nabla f(X(t))\|^2. \quad (4.78)$$

Therefore, replacing (4.75), (4.76), and (4.78) in (4.74) we get

$$\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}(t+1))|\mathcal{F}_t] \\
&\le f(\bar{\mathbf{x}}(t)) - \alpha(t)\beta(t)\left\langle\nabla f(\bar{\mathbf{x}}(t)), \mathbf{r}^T\nabla f(X(t))\right\rangle \\
&\quad + \beta^2(t)\frac{K}{2}\mathbb{E}\left[\left\|\mathbf{r}^T[E(t)-\alpha(t)\nabla f(X(t))]\right\|^2\middle|\mathcal{F}_t\right] \\
&\le f(\bar{\mathbf{x}}(t)) - \frac{1}{2}\alpha(t)\beta(t)\left(\|\nabla f(\bar{\mathbf{x}}(t))\|^2 + \left\|\mathbf{r}^T\nabla f(X(t))\right\|^2 - 2K^2\sum_{i=1}^n r_i\|\bar{\mathbf{x}}(t)-\mathbf{x}_i(t)\|^2\right) \\
&\quad + \beta^2(t)\frac{K}{2}\left(\gamma + \alpha^2(t)\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\right) \\
&= f(\bar{\mathbf{x}}(t)) - \frac{1}{2}\alpha(t)\beta(t)\|\nabla f(\bar{\mathbf{x}}(t))\|^2 - \frac{1}{2}\alpha(t)\beta(t)(1-\alpha(t)\beta(t)K)\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2 \\
&\quad + \alpha(t)\beta(t)K^2\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2 + \beta^2(t)\frac{K}{2}\gamma.
\end{aligned}$$

Taking the expectation of both sides leads to

$$\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}(t+1))] \le \mathbb{E}[f(\bar{\mathbf{x}}(t))] &- \frac{1}{2}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
&- \frac{1}{2}\alpha(t)\beta(t)(1-\alpha(t)\beta(t)K)\,\mathbb{E}\left[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\right] \\
&+ \alpha(t)\beta(t)K^2\mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] + \beta^2(t)\frac{K}{2}\gamma. \quad (4.79)
\end{aligned}$$

4.6.1 State Deviations from the Average State: $\mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$

From (4.31), we have

$$\sum_{s=1}^{t-1}\mathbb{E}\left[\|P(t:s)E(s)\|_{\mathbf{r}}^2\right] \le \gamma\kappa\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]. \quad (4.80)$$

Using (4.36) in (4.35), we get

$$\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right] \leq \frac{2}{\lambda}\kappa^{\frac{1}{2}}\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right]\right]. \quad (4.81)$$

Plugging the bounds obtained in (4.80) and (4.81) into (4.30) we arrive at

$$\begin{aligned}\mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \leq{}& 2\gamma\kappa\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\\ &+\frac{4}{\lambda}\kappa^{\frac{1}{2}}\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right]\right]. \quad (4.82)\end{aligned}$$

4.6.2 Analysis of the overall deviation: $\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$

Our goal here is to bound the overall weighted deviation of the states from their average. First recall the bound for $\mathbb{E}\left[\|X(t)-\bar{\mathbf{x}}(t)\mathbf{1}\|_{\mathbf{r}}^2\right]$, derived in Section 4.6.1 for each t. Our goal here is to bound

$$\begin{aligned}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \leq{}& 2\gamma\kappa\sum_{t=1}^{T}\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\\ &+\frac{4}{\lambda}\kappa^{\frac{1}{2}}\sum_{t=1}^{T}\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right]\right]. \quad (4.83)\end{aligned}$$

Focusing on the first term in (4.83), we can write

$$\begin{aligned}\sum_{t=1}^{T}\left[\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\right] &= \sum_{s=1}^{T-1}\left[\beta^2(s)\sum_{t=s+1}^{T}\left[\alpha(t)\beta(t)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\right]\\ &\leq \sum_{s=1}^{T-1}\left[\alpha(s)\beta^2(s)\sum_{t=s+1}^{T}\left[\beta(t)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\right] \leq \frac{1}{\lambda}\sum_{s=1}^{T-1}\alpha(s)\beta^2(s), \quad (4.84)\end{aligned}$$

where the first inequality is due to the fact that $\alpha(t)\leq\alpha(s)$ for $t>s$, and the second one follows from Lemma 4.5. Similarly, using the fact that $\alpha(t)\leq\alpha(s)$ for $t>s$, for the second term in (4.83), we have

$$\begin{aligned}&\sum_{t=1}^{T}\left[\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right]\right]\right]\\ &=\sum_{s=1}^{T-1}\left[\alpha^2(s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right]\sum_{t=s+1}^{T}\alpha(t)\beta(t)\pi(t:s)\right]\\ &\leq\sum_{s=1}^{T-1}\left[\alpha^3(s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right]\sum_{t=s+1}^{T}\beta(t)\pi(t:s)\right]. \quad (4.85)\end{aligned}$$

Since $\pi(t:s) = \beta(s)\kappa^{\frac{1}{2}} \prod_{k=s+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}$, using $\sqrt{1-x} \le 1 - x/2$, we have

$$\sum_{t=s+1}^{T} \beta(t)\pi(t:s) = \sum_{t=s+1}^{T}\left[\beta(t)\beta(s)\kappa^{\frac{1}{2}} \prod_{k=s+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}\right]$$
$$\le \beta(s)\kappa^{\frac{1}{2}} \sum_{t=s+1}^{T}\left[\beta(t)\prod_{k=s+1}^{t-1}\left(1-\frac{\lambda}{2}\beta(k)\right)\right] \le \frac{2}{\lambda}\beta(s)\kappa^{\frac{1}{2}}, \tag{4.86}$$

where the last inequality follows from Lemma 4.5. Then, (4.85) and (4.86) imply

$$\sum_{t=1}^{T}\left[\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right]\right]\right] \le \frac{2\kappa^{\frac{1}{2}}}{\lambda}\sum_{s=1}^{T-1}\left[\alpha^3(s)\beta(s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right]\right]$$

Therefore, plugging this and (4.84) into (4.83), we can conclude

$$\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$$
$$\le \frac{2\gamma\kappa}{\lambda}\sum_{t=1}^{T}\alpha(t)\beta^2(t) + \frac{8\kappa}{\lambda^2}\sum_{t=1}^{T}\left[\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(X(t))\|_{\mathbf{r}}^2\right]\right]. \tag{4.87}$$

4.6.3 Bounding $\mathbb{E}\left[\|\nabla f(X(t))\|_{\mathbf{r}}^2\right]$

In this part, we study $\mathbb{E}\left[\|\nabla f(X(t)\|_{\mathbf{r}}\right]^2$, to provide an upper bound for it. Following [6], we can rewrite $\nabla f(X(t))$ as

$$\nabla f(X(t)) = 3\left[\frac{1}{3}(\nabla f(X(t)) - \nabla f(\mathbf{1}\bar{\mathbf{x}}(t)) + \frac{1}{3}(\nabla f(\mathbf{1}\bar{\mathbf{x}}(t)) - \mathbf{1}\nabla f(\bar{\mathbf{x}}(t))) + \frac{1}{3}\mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\right].$$

where $\nabla f(\bar{\mathbf{x}}(t)) := \sum_{i=1}^{n}\nabla f_i(\bar{\mathbf{x}}(t))$. Then, since $\|\cdot\|_{\mathbf{r}}^2$ is a convex function, we have

$$\mathbb{E}\left[\|\nabla f(X(t)\|_{\mathbf{r}}^2\right] \le 3\mathbb{E}\left[\|\nabla f(X(t)) - \nabla f(\mathbf{1}\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2\right]$$
$$+ 3\mathbb{E}\left[\|\nabla f(\mathbf{1}\bar{\mathbf{x}}(t)) - \mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2\right] + 3\mathbb{E}\left[\|\mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2\right]. \tag{4.88}$$

Next, we bound each term in (4.88). Using (4.76), we can write

$$\mathbb{E}\left[\|\nabla f(X(t)) - \nabla f(\mathbf{1}\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} r_i \|\nabla f_i(\mathbf{x}_i(t)) - \nabla f_i(\bar{\mathbf{x}}(t))\|^2\right]$$
$$\le K^2\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]. \tag{4.89}$$

Similarly, using the convexity of function $\|\cdot\|^2$, for the second term in (4.88) we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{1}\bar{\mathbf{x}}(t)) - \mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} r_i \|\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
&= \sum_{i=1}^{n} r_i \mathbb{E}\left[4\left\|\frac{1}{2}\Big(\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla K(\bar{\mathbf{x}}(t))\Big) - \frac{1}{2}\Big(\nabla f(\bar{\mathbf{x}}(t)) - \nabla K(\bar{\mathbf{x}}(t))\Big)\right\|^2\right] \\
&\le \sum_{i=1}^{n} 2r_i \mathbb{E}\left[\|\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla K(\bar{\mathbf{x}}(t))\|^2\right] + 2\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t)) - \nabla K(\bar{\mathbf{x}}(t))\|^2\right] \\
&\overset{(a)}{=} \sum_{i=1}^{n} 2r_i \mathbb{E}\left[\frac{1}{m_i^2}\left\|\sum_{j=1}^{m_i}\left[\nabla \ell(\bar{\mathbf{x}}(t), \xi_j^i) - \nabla K(\bar{\mathbf{x}}(t))\right]\right\|^2\right] + 2\mathbb{E}\left[\frac{1}{N^2}\left\|\sum_{j=1}^{N}\left[\nabla \ell(\bar{\mathbf{x}}(t), \xi_j) - \nabla K(\bar{\mathbf{x}}(t))\right]\right\|^2\right] \\
&\overset{(b)}{=} \sum_{i=1}^{n} \frac{2r_i}{m_i^2} \sum_{j=1}^{m_i} \mathbb{E}\left[\left\|\nabla \ell(\bar{\mathbf{x}}(t), \xi_j^i) - \nabla K(\bar{\mathbf{x}}(t))\right\|^2\right] + \frac{2}{N^2}\sum_{j=1}^{N} \mathbb{E}\left[\|\nabla \ell(\bar{\mathbf{x}}(t), \xi_j) - \nabla K(\bar{\mathbf{x}}(t))\|^2\right] \\
&\overset{(c)}{\le} \sum_{i=1}^{n} 2\frac{m_i}{N}\frac{1}{m_i^2} m_i \sigma^2 + \frac{2}{N^2} N\sigma^2 = \frac{2(n+1)}{N}\sigma^2,
\end{aligned} \tag{4.90}
$$

where in (a) we replaced the definitions of $f_i(\bar{\mathbf{x}}(t))$ and $f(\bar{\mathbf{x}}(t))$ from (3.3) and (3.2), respectively, the equality in (b) holds since ξ_js are independent samples from the underlying distribution, and (c) follows from Assumption 3.2 and the fact that $r_i = m_i/N$ for $i \in [n]$. Finally, for the third term in (4.88), we have

$$
\mathbb{E}\left[\|\mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} r_i \|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] = \mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]. \tag{4.91}
$$

Plugging (4.89)–(4.91) into (4.88), we get

$$
\mathbb{E}\left[\|\nabla f(X(t)\|_{\mathbf{r}}^2\right] \le 3K^2 \mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] + \frac{6(n+1)}{N}\sigma^2 + 3\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]. \tag{4.92}
$$

Next, replacing this bound in (4.87), we arrive at

$$
\begin{aligned}
\sum_{t=1}^{T} \alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] &\le \frac{2\gamma\kappa}{\lambda}\sum_{t=1}^{T}\alpha(t)\beta^2(t) + \frac{8\kappa}{\lambda^2}\sum_{t=1}^{T}\left[\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(X(t))\|_{\mathbf{r}}^2\right]\right] \\
&\le \frac{2\gamma\kappa}{\lambda}\sum_{t=1}^{T}\alpha(t)\beta^2(t) + \frac{24\kappa K^2}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \\
&\quad + \frac{48\kappa(n+1)\sigma^2}{N\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t) + \frac{24\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right].
\end{aligned} \tag{4.93}
$$

Now, define $\phi_{i,j}(T) := \sum_{t=1}^{T}\alpha^i(t)\beta^j(t)$. Then, $\frac{2\gamma\kappa}{\lambda}\sum_{t=1}^{T}\alpha(t)\beta^2(t) = \epsilon_1\phi_{1,2}(T)$ and $\frac{48\kappa(n+1)\sigma^2}{N\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t) = \epsilon_2\phi_{3,1}(T)$, where $\epsilon_1 := \frac{2\gamma\kappa}{\lambda}$ and $\epsilon_2 := \frac{48\kappa(n+1)\sigma^2}{N\lambda^2}$. Furthermore,

we set $T_0 \coloneqq \left\lceil \left(\frac{14\alpha_0 \kappa^{\frac{1}{2}} K}{\lambda} \right)^{\frac{1}{\nu}} \right\rceil$ such that $\frac{24\kappa K^2}{\lambda^2}\alpha^2(T_0) \leq \frac{24}{196} < \frac{1}{2}$. Then, for $T \geq T_0$ we can rewrite (4.93) as

$$
\begin{aligned}
&\sum_{t=1}^{T_0} \alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] + \sum_{t=T_0+1}^{T} \alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \\
&\leq \epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \frac{24\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
&\quad + \frac{24\kappa K^2}{\lambda^2}\left[\sum_{t=1}^{T_0}\alpha^3(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] + \sum_{t=T_0+1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]\right] \\
&\leq \epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \frac{24\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
&\quad + \frac{24\kappa K^2}{\lambda^2}\left[\sum_{t=1}^{T_0}\alpha^3(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] + \alpha^2(T_0)\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]\right] \\
&\leq \epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \frac{24\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
&\quad + \epsilon_3 + \frac{1}{2}\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right],
\end{aligned}
$$

where $\epsilon_3 \coloneqq \frac{24\kappa K^2}{\lambda^2}\sum_{t=1}^{T_0}\alpha^3(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$ does not grow with T, and the second inequality holds since $\alpha(t)$ is a non-increasing sequence. Therefore, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \\
&\qquad \leq 2\sum_{t=1}^{T_0}\alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] + \sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \\
&\qquad \leq 2\epsilon_1\phi_{1,2}(T) + 2\epsilon_2\phi_{3,1}(T) + 2\epsilon_3 + \frac{48\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]. \quad (4.94)
\end{aligned}
$$

4.6.4 Back to the Main Dynamics

Recall the dynamics in (4.79), that is,

$$
\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}(t+1))] \leq \mathbb{E}\left[f(\bar{\mathbf{x}}(t))\right] - \frac{1}{2}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] + \frac{K\gamma}{2}\beta^2(t) \\
&- \frac{1}{2}\alpha(t)\beta(t)\left(1 - \alpha(t)\beta(t)K\right)\mathbb{E}\left[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\right] + \alpha(t)\beta(t)K^2\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right].
\end{aligned}
$$

Summing (4.79) for $t = 1, 2, \ldots, T$, and using (4.94) we get

$$\begin{aligned}
&\mathbb{E}\left[f\left(\bar{\mathbf{x}}(T+1)\right)\right] \leq \mathbb{E}\left[f\left(\bar{\mathbf{x}}(1)\right)\right] - \frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right] + \frac{K\gamma}{2}\sum_{t=1}^{T}\beta^2(t) \\
&- \frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}\left[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\right] + K^2\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\left\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\right\|_{\mathbf{r}}^2\right] \\
&\leq \mathbb{E}\left[f\left(\bar{\mathbf{x}}(1)\right)\right] - \frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right] + \frac{K}{2}\gamma\phi_{0,2}(T) \\
&- \frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}\left[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\right] \\
&+ K^2\left[2\epsilon_1\phi_{1,2}(T) + 2\epsilon_2\phi_{3,1}(T) + 2\epsilon_3 + \frac{48\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right]\right], \qquad (4.95)
\end{aligned}$$

where $\phi_{0,2}(T) = \sum_{t=1}^{T}\beta^2(t)$. Next, note that for $t \geq T_0 = \left\lceil\left(\frac{14\alpha_0\kappa^{\frac{1}{2}}K}{\lambda}\right)^{\frac{1}{\nu}}\right\rceil$ we have $\alpha(t)\beta(t)K \leq \alpha(T_0)\beta(T_0)K \leq \alpha(T_0)K \leq \frac{\lambda}{14\kappa^{\frac{1}{2}}} < 1$, where the last inequality holds since $\lambda \leq 1$ and $\kappa > 1$ (see Lemma 4.1). Therefore, the coefficient $1 - \alpha(t)\beta(t)K$ is non-negative for $t \geq T_0$. Thus, for $T \geq T_0$ we have

$$\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}\left[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\right] \\
&\quad\geq \frac{1}{2}\sum_{t=1}^{T_0}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}\left[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\right] := \epsilon_4, \qquad (4.96)
\end{aligned}$$

where ϵ_4 does not grow with T. Similarly, $\frac{48\kappa K^2}{\lambda^2}\alpha^2(T_0) \leq \frac{48}{196}\alpha_0 \leq \frac{1}{4}$. Therefore, for $T > T_0$, the last summation in (4.95) can be upper bounded by

$$\begin{aligned}
&\frac{48\kappa K^2}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right] \\
&\quad= \frac{48\kappa K^2}{\lambda^2}\left[\sum_{t=1}^{T_0}\alpha^3(t)\beta(t)\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right] + \sum_{t=T_0+1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right]\right], \\
&\quad\leq \epsilon_5 + \frac{48\kappa K^2}{\lambda^2}\alpha^2(T_0)\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right] \\
&\quad\leq \epsilon_5 + \frac{1}{4}\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right], \qquad (4.97)
\end{aligned}$$

where $\epsilon_5 := \frac{48\kappa K^2}{\lambda}\sum_{t=1}^{T_0}\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]$ does not depend on T. Next, plugging (4.96) and (4.97) into (4.95), for $T > T_0$, we get

$$\begin{aligned}
\mathbb{E}\left[f\left(\bar{\mathbf{x}}(T+1)\right)\right] \leq\ & \mathbb{E}\left[f\left(\bar{\mathbf{x}}(1)\right)\right] + \frac{K}{2}\phi_{0,2}(T) + 2K^2\left(\epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \epsilon_3\right) \\
& - \frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}\left[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\right] \\
& - \frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] + \frac{48\kappa K^2}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
\leq\ & \mathbb{E}\left[f\left(\bar{\mathbf{x}}(1)\right)\right] + \frac{K}{2}\phi_{0,2}(T) + 2K^2\left(\epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \epsilon_3\right) - \epsilon_4 \\
& - \frac{1}{2}\sum_{t=1}^{T_0}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] - \frac{1}{2}\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
& + \epsilon_5 + \frac{1}{4}\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
\leq\ & \mathbb{E}\left[f\left(\bar{\mathbf{x}}(1)\right)\right] + \frac{K}{2}\phi_{0,2}(T) + 2K^2\left(\epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \epsilon_3\right) - \epsilon_4 \\
& + \epsilon_5 - \frac{1}{4}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right].
\end{aligned}$$

By re-arrangement of the terms above, we get

$$\begin{aligned}
& \sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
& \leq 4\mathbb{E}\left[f(\bar{\mathbf{x}}(1))\right] - 4\mathbb{E}\left[f(\bar{\mathbf{x}}(T+1))\right] + 2K\phi_{0,2}(T) + 8K^2\left(\epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \epsilon_3\right) - 4\epsilon_4 + 4\epsilon_5 \\
& \leq 4\mathbb{E}\left[f(\bar{\mathbf{x}}(1))\right] - 4\mathbb{E}\left[f(\mathbf{x}^\star)\right] + 2K\phi_{0,2}(T) + 8K^2\left(\epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \epsilon_3\right) - 4\epsilon_4 + 4\epsilon_5 \\
& = \epsilon_6 + 2K\phi_{0,2}(T) + 8K^2\epsilon_1\phi_{1,2}(T) + 8K^2\epsilon_2\phi_{3,1}(T) \\
& \leq \epsilon_6 + (2K + 8K^2\epsilon_1\alpha_0)\phi_{0,2}(T) + 8K^2\epsilon_2\phi_{3,1}(T),
\end{aligned} \tag{4.98}$$

where $\epsilon_6 := 8K^2\epsilon_3 - 4\epsilon_4 + 4\epsilon_5 + 4\mathbb{E}\left[f\left(\bar{\mathbf{x}}(1)\right)\right] - 4\mathbb{E}\left[f(\mathbf{x}^\star)\right]$ is a constant (does not depend on T), and the last inequality in (4.98) follows from the fact that

$$\phi_{1,2}(T) = \sum_{t=1}^{T}\alpha(t)\beta^2(t) = \alpha_0\sum_{t=1}^{T}\frac{1}{(t+\tau)^\nu}\beta^2(t) \leq \alpha_0\sum_{t=1}^{T}\beta^2(t) = \alpha_0\phi_{0,2}(T). \tag{4.99}$$

4.6.5 Bound on the Moments of $\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]$ and $\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$

The inequality in (4.98) provides us with an upper bound on the temporal average of $\left\{\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right\}$. However, our goal is to derive a bound on the temporal

average of $\left\{\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right\}$. To this end, define the convergence measure

$$M_\theta(\nu,\mu) := \left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]\right)^\theta\right]^{\frac{1}{\theta}},$$

for a given $\theta \in (0,1)$. Note that by Hölder's inequality [114, Theorem 6.2] for any $p,q>1$ with $\frac{1}{p}+\frac{1}{q}=1$, and non-negative sequences $\{a_t\}_{t=1}^T$ and $\{b_t\}_{t=1}^T$, we have

$$\left(\sum_{t=1}^{T} a_t b_t\right)^q \le \left(\sum_{t=1}^{T} a_t^p\right)^{\frac{q}{p}}\left(\sum_{t=1}^{T} b_t^q\right). \tag{4.100}$$

Let $(p,q) := \left(\frac{1}{1-\theta},\frac{1}{\theta}\right)$ so that $\frac{1}{p}+\frac{1}{q}=1$. Furthermore, define

$$a_t := \left(\frac{1}{\alpha(t)\beta(t)}\right)^\theta = \frac{(t+\tau)^{(\mu+\nu)\theta}}{(\alpha_0\beta_0)^\theta} \quad \text{and} \quad b_t := \left(\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]\right)^\theta. \tag{4.101}$$

Then, applying Hölder's inequality (4.100), we arrive at

$$M_\theta(\nu,\mu) = \left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]\right)^\theta\right]^{\frac{1}{\theta}} = \left(\frac{1}{T}\sum_{t=1}^{T} a_t b_t\right)^q \le \frac{1}{T^q}\left(\sum_{t=1}^{T} a_t^p\right)^{\frac{q}{p}}\left(\sum_{t=1}^{T} b_t^q\right). \tag{4.102}$$

It remains to upper bound the terms in the RHS of (4.102). First, using Lemma 4.8 we get

$$\sum_{t=1}^{T} a_t^p = \frac{1}{(\alpha_0\beta_0)^{\frac{\theta}{1-\theta}}}\sum_{t=1}^{T}(t+\tau)^{\frac{(\nu+\mu)\theta}{1-\theta}} \le \frac{2^{1+\frac{(\nu+\mu)\theta}{1-\theta}}}{(\alpha_0\beta_0)^{\frac{\theta}{1-\theta}}\left(1+\frac{(\nu+\mu)\theta}{1-\theta}\right)}(T+\tau)^{1+\frac{(\nu+\mu)\theta}{1-\theta}}.$$

Therefore,

$$\left(\sum_{t=1}^{T} a_t^p\right)^{\frac{q}{p}} \le \frac{2^{\frac{1-\theta}{\theta}+(\nu+\mu)}}{\alpha_0\beta_0\left(1+\frac{(\nu+\mu)\theta}{1-\theta}\right)^{\frac{1-\theta}{\theta}}}(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu} := \epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}. \tag{4.103}$$

Next, continuing from (4.98), for $T \ge T_0$ we can write

$$\begin{aligned}\sum_{t=1}^{T} b_t^q = \sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] &\le \epsilon_6 + (2K+8K^2\epsilon_1\alpha_0)\phi_{0,2}(T) + 8K^2\epsilon_2\phi_{3,1}(T)\\ &\le \max\left\{3\epsilon_6, (6K+24K^2\epsilon_1\alpha_0)\phi_{0,2}(T), 24K^2\epsilon_2\phi_{3,1}(T)\right\},\end{aligned} \tag{4.104}$$

where the last inequality follows from the fact that $a+b+c \le \max\{3a, 3b, 3c\}$. Then, we can use Lemma 4.8 and the fact that $2^{1-2\mu} \le 2$ for $0 \le \mu < 1$, to bound $\phi_{0,2}(T)$, as

$$\phi_{0,2}(T) = \sum_{t=1}^{T} \beta^2(t) = \frac{1}{\beta_0^2} \sum_{t=1}^{T} (t+\tau)^{-2\mu} \le \begin{cases} \frac{\tau^{1-2\mu}}{\beta_0^2 |1-2\mu|} & \text{if } \mu > 1/2, \\ \frac{1}{\beta_0^2} \ln\left(\frac{T}{\tau}+1\right) & \text{if } \mu = 1/2, \\ \frac{2(T+\tau)^{1-2\mu}}{\beta_0^2(1-2\mu)} & \text{if } 0 \le \mu < 1/2. \end{cases} \tag{4.105}$$

Similarly, applying Lemma 4.8 on $\phi_{3,1}(T)$, we get

$$\phi_{3,1}(T) = \sum_{t=1}^{T} \alpha^3(t)\beta(t) = \frac{1}{\alpha_0^3 \beta_0} \sum_{t=1}^{T} (t+\tau)^{-3\nu-\mu} \le \begin{cases} \frac{\tau^{1-3\nu-\mu}}{\alpha_0^3 \beta_0 |1-3\nu-\mu|} & \text{if } 3\nu+\mu > 1, \\ \frac{1}{\alpha_0^3 \beta_0} \ln\left(\frac{T}{\tau}+1\right) & \text{if } 3\nu+\mu = 1, \\ \frac{2(T+\tau)^{1-3\nu-\mu}}{\alpha_0^3 \beta_0 (1-3\nu-\mu)} & \text{if } 0 \le 3\nu+\mu < 1, \end{cases} \tag{4.106}$$

in which $2^{1-3\nu-\mu}$ is bounded by 2, for $0 \le 3\nu+\mu < 1$. Note that the upper bound in (4.104) is the maximum of a constant, and two $\phi_{\cdot,\cdot}(T)$ functions. Moreover, depending on the values of μ and ν, $\phi_{\cdot,\cdot}(T)$ functions can be upper bounded as in (4.105) and (4.106). Figure 4.4 illustrates the four regions of (μ, ν). In the following, we first analyze the interior of the four regions, and then study the boundary cases.

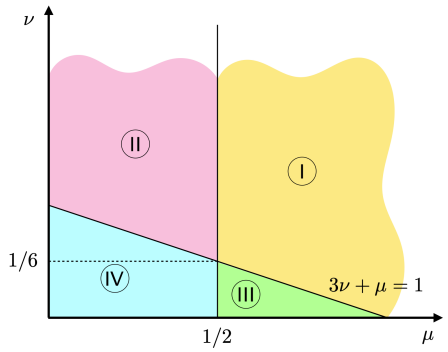

Figure 4.4: Regions of (μ, ν).

(**Region I**) $\mu > 1/2$ **and** $3\nu + \mu > 1$: Recall from (4.105) and (4.106) that $\phi_{0,2}(T)$ and

$\phi_{3,1}(T)$ are both upper bounded by constants. Hence, in this regime, (4.104) leads to

$$\sum_{t=1}^{T} b_t^q \le \max\left\{3\epsilon_6, (6K + 24K^2\epsilon_1\alpha_0)\phi_{0,2}(T), 24K^2\epsilon_2\phi_{3,1}(T)\right\}$$
$$\le \max\left\{3\epsilon_6, (6K + 24K^2\epsilon_1\alpha_0)\frac{\tau^{1-2\mu}}{\beta_0^2|1-2\mu|}, 24K^2\epsilon_2\frac{\tau^{1-3\nu-\mu}}{\alpha_0^3\beta_0|1-3\nu-\mu|}\right\} := \epsilon_8. \quad (4.107)$$

Note that ϵ_8 is a constant. Plugging (4.103) and (4.107) into (4.102), we arrive at

$$M_\theta(\nu,\mu) \le \frac{1}{T^{1/\theta}}\epsilon_7(\theta)\cdot(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}\cdot\epsilon_8 = \mathcal{O}\left(T^{-(1-\nu-\mu)}\right).$$

(**Region II)** $\mu < 1/2$ **and** $3\nu+\mu > 1$: Note that (4.105) and (4.106) imply that $\phi_{0,2}(T)$ and $\phi_{3,1}(T)$ are upper bounded by a polynomial (in T) and a constant. Since for sufficiently large T, a polynomial function beats any constant, we can write

$$\sum_{t=1}^{T} b_t^q \le \frac{12K + 48K^2\epsilon_1\alpha_0}{\beta_0^2(1-2\mu)}(T+\tau)^{1-2\mu} := \epsilon_9\cdot(T+\tau)^{1-2\mu}. \quad (4.108)$$

This together with (4.102) and (4.103) implies

$$M_\theta(\nu,\mu) \le \frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}\cdot\epsilon_9\cdot(T+\tau)^{1-2\mu} = \mathcal{O}\left(T^{-(\mu-\nu)}\right).$$

(**Region III)** $\mu > 1/2$ **and** $3\nu+\mu < 1$: Recall from (4.105) and (4.106) that $\phi_{0,2}(T)$ and $\phi_{3,1}(T)$ are upper bounded by a constant and a polynomial function of T, respectively. Therefore, for sufficiently large T, we get

$$\sum_{t=1}^{T} b_t^q \le \left(\frac{48K^2\epsilon_2}{\alpha_0^3\beta_0(1-3\nu-\mu)}\right)(T+\tau)^{1-3\nu-\mu} := \epsilon_{10}\cdot(T+\tau)^{1-3\nu-\mu}. \quad (4.109)$$

Therefore, plugging (4.109) and (4.103) to (4.102), we get

$$M_\theta(\nu,\mu) \le \frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}\cdot\epsilon_{10}\cdot(T+\tau)^{1-3\nu-\mu} = \mathcal{O}\left(T^{-2\nu}\right).$$

(**Region IV)** $\mu < 1/2$ **and** $3\nu+\mu < 1$: In this region, we can use (4.105) and (4.106) to upper bound both $\phi_{0,2}(T)$ and $\phi_{3,1}(T)$ by polynomial functions. Thus, for sufficiently large T, we get

$$\sum_{t=1}^{T} b_t^q \le \max\left\{\frac{12K + 48K^2\epsilon_1\alpha_0}{\beta_0^2(1-2\mu)}, \frac{48K^2\epsilon_2}{\alpha_0^3\beta_0(1-3\nu-\mu)}\right\}(T+\tau)^{\max\{1-2\mu,1-3\nu-\mu\}}$$
$$:= \epsilon_{11}\cdot(T+\tau)^{\max\{1-2\mu,1-3\nu-\mu\}}. \quad (4.110)$$

Then, we can plug (4.110) and (4.103) into (4.102), to conclude

$$M_\theta(\nu,\mu) \le \frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}\cdot\epsilon_{11}\cdot(T+\tau)^{\max\{1-2\mu,1-3\nu-\mu\}} = \mathcal{O}\left(T^{-\min\{\mu-\nu,2\nu\}}\right).$$

The result of the four cases above concludes the first claim of Theorem 4.2.

Recall that our goal is to find (ν,μ) that achieves the best convergence rate. This is equivalent to optimizing the exponent of $1/T$ in each of the four (open) regions I – IV (shown in Figure 4.4). Interestingly, it turns out that the respective supremum value in all four regions is $1/3$, which corresponds to the boundary point $(\nu^\star,\mu^\star) = (1/6,1/2)$. However, this point does not belong to any of the corresponding open sets, which motivates the convergence analysis of M_θ for $(\nu^\star,\mu^\star) = (1/6,1/2)$.

(**Boundary Case) $\mu = 1/2$ and $3\nu+\mu = 1$**: First note that the two lines of interest intersect at $(\nu^\star,\mu^\star) = (1/6,1/2)$, as shown in Figure 4.4. Applying Lemma 4.8 on $\phi_{0,2}(T)$ and $\phi_{3,1}(T)$ for $(\nu^\star,\mu^\star) = (1/6,1/2)$ we get $\phi_{0,2}(T) \le \frac{1}{\beta_0^2}\ln\left(\frac{T}{\tau}+1\right)$ and $\phi_{3,1}(T) \le \frac{1}{\alpha_0^3\beta_0}\ln\left(\frac{T}{\tau}+1\right)$. Therefore, (4.104) reduces to

$$\sum_{t=1}^{T} b_t^q \le \max\left\{\frac{6K+24K^2\epsilon_1\alpha_0}{\beta_0^2}, \frac{24K^2\epsilon_2}{\alpha_0^3\beta_0}\right\}\ln\left(\frac{T}{\tau}+1\right) := \epsilon_{12}\cdot\ln\left(\frac{T}{\tau}+1\right). \tag{4.111}$$

Plugging (4.111) and (4.103) into (4.102), we arrive at

$$M_\theta\left(\frac{1}{6},\frac{1}{2}\right) \le \frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\frac{2}{3}}\cdot\epsilon_{12}\cdot\ln\left(\frac{T}{\tau}+1\right) = \mathcal{O}\left(T^{-1/3}\ln T\right), \tag{4.112}$$

which is the second claim of the theorem.

4.7 Proof of Proposition 4.1

First note that we cannot directly conclude the proposition from Theorem 4.2, since the theorem only holds for $\theta \in (0,1)$, and not $\theta = 1$. In order to show the claim, for a vector $\mathbf{y} \in \mathbb{R}^T$ and some $\theta \in (0,1)$ we define[1] $\|\mathbf{y}\|_\theta := \left(|y_1|^\theta + |y_2|^\theta + \ldots + |y_T|^\theta\right)^{1/\theta}$. Then we have $\|\mathbf{y}\|_1 \le \|\mathbf{y}\|_\theta$ since

$$\left(\frac{\|\mathbf{y}\|_\theta}{\|\mathbf{y}\|_1}\right)^\theta = \frac{|y_1|^\theta+\ldots+|y_T|^\theta}{(|y_1|+\ldots+|y_T|)^\theta} = \sum_{t=1}^{T}\left(\frac{|y_t|}{|y_1|+\ldots+|y_T|}\right)^\theta \ge \sum_{t=1}^{T}\frac{|y_t|}{|y_1|+\ldots+|y_T|} = 1,$$

[1] Note that for $\|\mathbf{y}\|_\theta$ is not a norm, since it is not a subadditive function for $\theta < 1$.

where the inequality holds since we have $0 \le |y_t| / \sum_{i=1}^{T} |y_i| \le 1$, and $\theta < 1$.

Now, for the vector $\mathbf{y}$ with $y_t = \mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]$ we have

$$M_1 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] = \frac{1}{T}\|\mathbf{y}\|_1 \le \frac{1}{T}\|\mathbf{y}\|_\theta = \frac{1}{T}\left(\sum_{t=1}^{T}\left(\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]\right)^\theta\right)^{1/\theta}$$
$$= \frac{1}{T^{1-1/\theta}}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right]\right)^\theta\right)^{1/\theta} = \frac{1}{T^{1-1/\theta}} M_\theta.$$

Then, from Theorem 4.2 for $(\nu^\star, \mu^\star) = (1/6, 1/2)$ and $\theta = \frac{2}{2+\epsilon}$, we get

$$M_1 \le \frac{1}{T^{1-1/\theta}} M_\theta \le \frac{1}{T^{1-1/\theta}} \mathcal{O}\left(T^{-1/3}\ln T\right) = T^{\epsilon/2}\mathcal{O}\left(T^{-1/3}\ln T\right) = \mathcal{O}\left(T^{-1/3+\epsilon}\right), \tag{4.113}$$

where the last equality holds since $\ln T = \mathcal{O}(T^{\epsilon/2})$ for any $\epsilon > 0$. Similarly, for the vector $\mathbf{z} \in \mathbb{R}^T$ with $z_t := \mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$ and any $\theta \in (0,1)$ we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] = \frac{1}{T}\|\mathbf{z}\|_1 \le \frac{1}{T}\|\mathbf{z}\|_\theta = \frac{1}{T}\left[\sum_{t=1}^{T}\left(\mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]\right)^\theta\right]^{\frac{1}{\theta}}$$
$$= \frac{1}{T^{1-1/\theta}}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]\right)^\theta\right]^{\frac{1}{\theta}}. \tag{4.114}$$

We need to bound the RHS of (4.114). Let $a_t := (\alpha(t)\beta(t))^{-\theta} = (t+\tau)^{2\theta/3}/(\alpha_0\beta_0)^\theta$ and $c_t := \left(\alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]\right)^\theta$. Then, using the Hölder's inequality in (4.100) for $(p,q) = \left(\frac{1}{1-\theta}, \frac{1}{\theta}\right)$ we can write

$$\left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]\right)^\theta\right]^{\frac{1}{\theta}} = \left(\frac{1}{T}\sum_{t=1}^{T} a_t c_t\right)^q \le \frac{1}{T^q}\left(\sum_{t=1}^{T} a_t^p\right)^{\frac{q}{p}}\left(\sum_{t=1}^{T} c_t^q\right). \tag{4.115}$$

Note that $\left(\sum_{t=1}^{T} a_t^p\right)^{\frac{q}{p}}$ is bounded in (4.103). Moreover, for $\sum_{t=1}^{T} c_t^q$ we can write

$$\begin{aligned}
\sum_{t=1}^{T} c_t^q &= \sum_{t=1}^{T} \alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \\
&\overset{(a)}{\leq} 2\epsilon_1\phi_{1,2}(T) + 2\epsilon_2\phi_{3,1}(T) + 2\epsilon_3 + \frac{48\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
&\overset{(b)}{\leq} 2\epsilon_1\phi_{1,2}(T) + 2\epsilon_2\phi_{3,1}(T) + 2\epsilon_3 + \frac{48\kappa\alpha_0^2}{\lambda^2}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\right] \\
&\overset{(c)}{\leq} 2\alpha_0\epsilon_1\phi_{0,2}(T) + 2\epsilon_2\phi_{3,1}(T) + 2\epsilon_3 + \frac{48\kappa\alpha_0^2}{\lambda^2}\left(\epsilon_6 + K(2+8K\alpha_0\epsilon_1)\phi_{0,2}(T) + 8K^2\epsilon_2\phi_{3,1}(T)\right) \\
&\leq 2\epsilon_3 + \frac{48\kappa\alpha_0^2\epsilon_6}{\lambda^2} + \left(2\alpha_0\epsilon_1 + \frac{96\kappa\alpha_0^2K(1+4K\alpha_0\epsilon_1)}{\lambda^2}\right)\phi_{0,2}(T) + \left(2\epsilon_2 + \frac{384\kappa\alpha_0^2K^2\epsilon_2}{\lambda^2}\right)\phi_{3,1}(T) \\
&\overset{(d)}{\leq} 2\epsilon_3 + \frac{48\kappa\alpha_0^2\epsilon_6}{\lambda^2} + \left(\frac{2\alpha_0\epsilon_1}{\beta_0^2} + \frac{96\kappa\alpha_0^2K(1+4K\alpha_0\epsilon_1)}{\lambda^2\beta_0^2}\right)\ln\left(\frac{T}{\tau}+1\right) \\
&\quad + \left(\frac{2\epsilon_2}{\alpha_0^3\beta_0} + \frac{384\kappa K^2\epsilon_2}{\lambda^2\alpha_0\beta_0}\right)\ln\left(\frac{T}{\tau}+1\right) \overset{(e)}{\leq} \epsilon_{13}\cdot\ln\left(\frac{T}{\tau}+1\right), \qquad (4.116)
\end{aligned}$$

where (a) follows from (4.94), (b) holds since $\alpha^2(t) = \frac{\alpha_0^2}{(t+\tau)^{1/3}} \leq \alpha_0^2$, the inequality in (c) follows from (4.98) and (4.99), we have used a bounds in (4.105) and (4.106) for $(\nu^\star, \mu^\star) = (1/6, 1/2)$ in (d), and (e) holds since the constant term $2\epsilon_3 + 48\kappa\alpha_0^2\epsilon_6/\lambda^2$ is upper bounded by $\ln(T/\tau+1)$ for large enough T. Next, plugging (4.103) for $\nu^\star+\mu^\star = 2/3$ and (4.116) into (4.115) and (4.114), and setting $\theta = \frac{2}{2+\epsilon}$ we arrive at

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\|^2\Big] &\leq \frac{1}{\mathbf{r}_{\min}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]\right] \\
&\leq \frac{1}{\mathbf{r}_{\min}}\frac{1}{T^{1-1/\theta}}\frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\frac{2}{3}}\cdot\epsilon_{13}\cdot\ln\left(\frac{T}{\tau}+1\right) = \mathcal{O}\left(T^{-1/3+\epsilon}\right), \qquad (4.117)
\end{aligned}$$

where the last equality holds since $\ln T = \mathcal{O}(T^{\epsilon/2})$ for any $\epsilon > 0$. Finally, combining (4.9), (4.10), and using Assumption 3.1 and Lemma 4.3 (for $\omega = 1$) we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla f(\mathbf{x}_i(t))\|^2\Big] &\leq \frac{2}{T}\sum_{t=1}^{T}\Big\{\mathbb{E}\Big[\|\nabla f(\mathbf{x}_i(t)) - \nabla f(\bar{\mathbf{x}}(t))\|^2\Big] + \mathbb{E}\Big[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\Big]\Big\} \\
&\leq \frac{2}{T}\sum_{t=1}^{T}\Big\{K^2\mathbb{E}\Big[\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\|^2\Big] + \mathbb{E}\Big[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\Big]\Big\} \leq \mathcal{O}(T^{-1/3+\epsilon}).
\end{aligned}$$

for every $i \in [n]$. This concludes the proof of Proposition 4.1. ∎

4.8 Extension to Almost Sure Sense

In this section, we study (4.2) to identify the sufficient conditions on the step-sizes sequences for the almost sure convergence of the agent's states to an optimal solution for the class of convex cost functions.

Regarding the local cost function, we assume that agent $i \in [n]$ has access to a subgradient $\mathbf{g}_i(\mathbf{x}_i(t))$ of the local cost function $f_i(\cdot)$ at each local decision variable $\mathbf{x}_i(t)$ at time t. Thus, we can rewrite (4.2) as

$$\mathbf{x}_i(t+1) = (1-\beta(t))\mathbf{x}_i(t) + \beta(t)\hat{\mathbf{x}}_i(t) - \hat{\alpha}(t)\mathbf{g}_i(\mathbf{x}_i(t)), \tag{4.118}$$

where $\{\beta(t)\}$ and $\{\hat{\alpha}(t)\}$ are the sequences of step-sizes of the algorithm. This section identifies *sufficient* conditions on the sequences of step-sizes for almost sure convergence of the dynamics (4.118) to an optimal point $\mathbf{x}^\star \in \mathcal{X}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^n f_i(\mathbf{x})$. For simplicity of notation, let

$$G(t) := \begin{bmatrix} \mathbf{g}_1(\mathbf{x}_1(t)) \\ \vdots \\ \mathbf{g}_n(\mathbf{x}_n(t)) \end{bmatrix}.$$

Using these matrices, we can write the update rule (4.118) in the form of a linear time-varying system given by

$$X(t+1) = A(t)X(t) + V(t), \tag{4.119}$$

where $V(t) := \beta(t)E(t) - \hat{\alpha}(t)G(t)$. We also define $\Phi(t:s) := A(t-1)\cdots A(s+1)$ for $t > s$, with $\Phi(t:t-1) = I$.

4.8.1 Assumptions

To proceed with our main result, we need to make certain assumptions in addition to Assumptions 4.1 and 4.2, including those regarding the local cost functions and the sequences of step-sizes.

Assumption 4.4 (Objective Function Assumptions)**.** We assume that objective functions f_i satisfy the following properties.

(a) f_i is convex for all $i \in [n]$.

(b) The optimizer set $\mathcal{X}^\star := \arg\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^n r_i f_i(\mathbf{x})$ is non-empty.

(c) Each f_i has bounded subgradients, i.e., there exists $L > 0$ such that $\|\mathbf{g}_i\| < L$ for all subgradients $\mathbf{g}_i$ of $f_i(\mathbf{x})$ at every $\mathbf{x} \in \mathbb{R}^d$. This also implies that each $f_i(\cdot)$ is L-Lipschitz continuous, i.e.,

$$|f_i(\mathbf{x}) - f_i(\mathbf{y})| < L\,\|\mathbf{x} - \mathbf{y}\|\,,$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Assumption 4.5 (Step-size Sequences Assumptions)**.** For the non-increasing step-size sequences $\{\hat{\alpha}(t)\}$ and $\{\beta(t)\}$ where $\{\beta(t)\}$ take values in $[0, 1]$, we assume that

(a) $\sum_{t=1}^\infty \hat{\alpha}(t) = \infty$,

(b) $\sum_{t=1}^\infty \hat{\alpha}^2(t) < \infty$, $\sum_{t=1}^\infty \beta^2(t) < \infty$, and

(c) $\sum_{t=1}^\infty \frac{\hat{\alpha}^2(t)}{\beta(t)} < \infty$.

Also, there exists some $t_0 \geq 1$ such that for every $t \geq t_0$

(d) $-\Delta\beta(t) \leq c_1\beta^2(t)$,

(e) $-\Delta\hat{\alpha}(t) \leq c_2\hat{\alpha}(t)\beta(t)$,

for some positive constants $c_1 < \frac{\lambda}{2}$ and $c_2 < \frac{\lambda}{4}$, where $\lambda := \frac{\eta \mathbf{r}_{\min}}{2Bn^2} < 1$ and $\mathbf{r}_{\min} := \min_{i\in[n]}\{r_i\} \leq 1$.

Remark 4.6. Note that Assumptions 4.5-(a), (b), and (c) imply $\sum_{t=1}^\infty \beta(t) = \infty$ as if $\sum_{t=1}^\infty \beta(t) < \infty$, using the Cauchy-Schwarz inequality we get

$$\sum_{t=1}^\infty \hat{\alpha}(t) \leq \left(\sum_{t=1}^\infty \frac{\hat{\alpha}^2(t)}{\beta(t)}\right)^{\frac{1}{2}} \left(\sum_{t=1}^\infty \beta(t)\right)^{\frac{1}{2}} < \infty,$$

which is a contradiction with Assumption 4.5-(a). Similarly, we can write

$$\sum_{t=1}^\infty \hat{\alpha}(t)\beta^{\frac{1}{2}}(t) \leq \left(\sum_{t=1}^\infty \frac{\hat{\alpha}^2(t)}{\beta(t)}\right)^{\frac{1}{2}} \left(\sum_{t=1}^\infty \beta^2(t)\right)^{\frac{1}{2}} < \infty, \tag{4.120}$$

where the second inequality follows from Assumptions 4.5-(b) and (c).

Remark 4.7. Note that unlike Assumption 4.5-(a)-(c) that do not depend on the dynamics parameters, Assumption 4.5-(d)-(e) depend on those parameters. However, we will show in Section 4.8.5 that Assumption 4.5-(d)-(e) will be satisfied for sufficiently large t, regardless of the dynamic parameters.

4.8.2 Main Results

Now, we present the main results. First, we present sufficient conditions for the sequences $\{\beta(t)\}$ and $\{\hat{\alpha}(t)\}$ for the almost sure convergence to an optimal point for the agents acting under the dynamics (4.118). Then, for step-sizes of the form $\hat{\alpha}(t) = \frac{\alpha_0}{t^\nu}$ and $\beta(t) = \frac{\beta_0}{t^\mu}$, we provide the region of (μ, ν) for which the almost sure convergence is guaranteed.

Theorem 4.4. *If Assumptions 4.1-4.5 are satisfied, then, for the dynamics (4.118), for all $i \in [n]$ we have $\lim_{t\to\infty} \mathbf{x}_i(t) = \tilde{\mathbf{x}}$ almost surely, where $\tilde{\mathbf{x}}$ is an optimal point in the set of optimal solutions $\mathcal{X}^\star$.*

The proof of Theorem 4.4 is provided in Section 4.8.4. The implication of the above result for the practical step-sizes $\hat{\alpha}(t) = \frac{\alpha_0}{t^\nu}$ and $\beta(t) = \frac{\beta_0}{t^\mu}$ as as follows.

Proposition 4.1. Let Assumptions 4.1-4.4 hold. Then, for every $i \in [n]$, the dynamics (4.118) with step-sizes $\hat{\alpha}(t) = \frac{\alpha_0}{t^\nu}$ and $\beta(t) = \frac{\beta_0}{t^\mu}$ converges almost surely, i.e., we have $\lim_{t\to\infty} \mathbf{x}_i(t) = \tilde{\mathbf{x}}$ almost surely for some optimal point $\tilde{\mathbf{x}} \in \mathcal{X}^\star$, provided that $\beta_0 \leq 1$, $\frac{1}{2} < \mu \leq 1$, and $\frac{1}{2}(1+\mu) < \nu \leq 1$.

The proof of Proposition 4.1 is provided in Section 4.8.5.

Remark 4.8. Proposition 4.1 identifies sufficient conditions for (μ, ν) such that the dynamics (4.118) converges almost surely to the optimal set of a *convex* objective function over *time-varying* networks, when utilizing step-sizes of the form $\hat{\alpha}(t) = \frac{\alpha_0}{t^\nu}$ and $\beta(t) = \frac{\beta_0}{t^\mu}$. The same dynamic is studied for *constrained* optimization problems with *i.i.d.* weight matrices with symmetric expected weight, and *independent* noisy communication links in [97], where interestingly, the same (μ, ν)-region for the convergence of the dynamic is obtained.

Remark 4.9. Figure 4.5 compares the region of (μ, ν) to guarantee the ℓ_2-convergence which is $\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3$ and the region for the almost sure convergence, i.e., $\mathcal{R}_1$. Interestingly, the optimal parameters $\mu^\star = 0.75$ and $\nu^\star = 1$ that lead to the fastest convergence in ℓ_2 sense for strongly convex loss functions also guarantee the almost sure convergence.

Note that both regions only characterize *sufficient* conditions for two types of convergence criteria under different function properties, and hence, not comparable. For example, if for ℓ_2 convergence, we relax the class of strongly convex functions to general convex functions, we can show that it is *necessary* to have $\mu > \frac{1}{2}$. In other words, we cannot have ℓ_2 convergence in the region $\mathcal{R}_3$ for the general class of (not necessarily strong) convex functions. To see this, consider the convex functions $f_i(\mathbf{x}) = 0$ for $i \in [n]$ with the set of optimizers $\mathcal{X}^\star = \mathbb{R}^d$, and zero-mean *i.i.d.* noise sequences with variance γ, which satisfy Assumption 4.1. Then, multiplying both sides of (4.119) by $\mathbf{r}^T$, we get $\bar{\mathbf{x}}(k+1) = \bar{\mathbf{x}}(k) + \beta(k)\mathbf{r}^T E(k)$. Therefore, we can write

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\mathbf{x}}(k+1)\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\bar{\mathbf{x}}(k+1)\|^2\right]\middle|\mathcal{F}_k\right] \\
&= \mathbb{E}\left[\|\bar{\mathbf{x}}(k)\|^2\right] + \beta(k)\langle \mathbf{r}^T \mathbb{E}\left[E(k)|\mathcal{F}_k\right], \bar{\mathbf{x}}(k)\rangle + \beta^2(k)\mathbb{E}\left[\left\|\mathbf{r}^T E(k)\right\|^2\middle|\mathcal{F}_k\right] \\
&= \mathbb{E}\left[\|\bar{\mathbf{x}}(k)\|^2\right] + \gamma\beta^2(k), \qquad (4.121)
\end{aligned}$$

where the last equality is due to Assumption 4.1 (see (4.53) for more details). Summing up (4.121) over k, we arrive at

$$\lim_{t\to\infty} \mathbb{E}\left[\|\bar{\mathbf{x}}(t)\|^2\right] = \|\bar{\mathbf{x}}(1)\|^2 + \gamma \sum_{k=1}^{\infty} \beta^2(k).$$

If $\mathbf{x}_i(t)$ converges to some $\tilde{\mathbf{x}}$ in ℓ_2 for all $i \in [n]$, then we have $\lim_{t\to\infty} \mathbb{E}\left[\|\bar{\mathbf{x}}(t)\|^2\right] = \mathbb{E}\left[\|\tilde{\mathbf{x}}\|^2\right] < \infty$. This implies that $\sum_{k=1}^{\infty} \beta^2(k) < \infty$, which means that we need to have $\mu > \frac{1}{2}$. In other words, the condition $\mu > \frac{1}{2}$ is necessary for the class of convex functions if ℓ_2-convergence is desired. Further investigation is required to determine whether region $\mathcal{R}_2$ leads to a convergence.

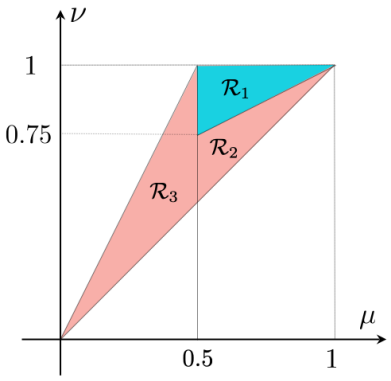

Figure 4.5: $\mathcal{R}_1$ is the (μ, ν)-region for the almost sure convergence of the dynamic when applied on strongly convex functions. The dynamic converges in the ℓ_2-sense in $\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3$, when the objective functions are convex.

4.8.3 Experimental Results

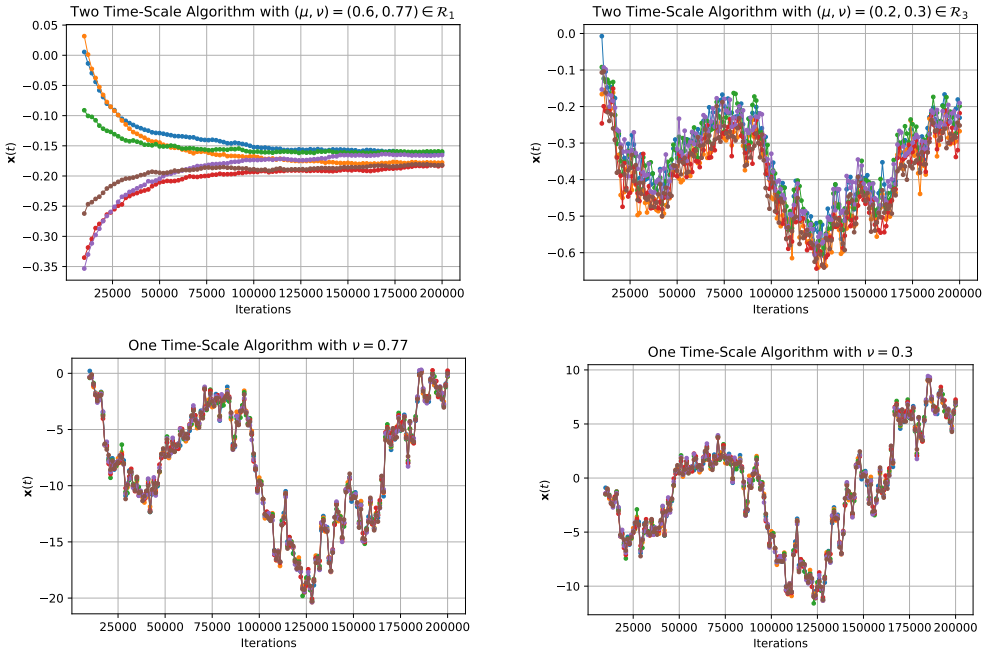

Figure 4.6: Trajectory vs. Iterations: Two Time-Scale Algorithm with $(\mu, \nu) = (0.6, 0.77)$ (top-left), $(0.2, 0.3)$ (top-right), and One Time-Scale Algorithm with $\nu = 0.77$ (bottom-left), 0.3 (bottom-right).

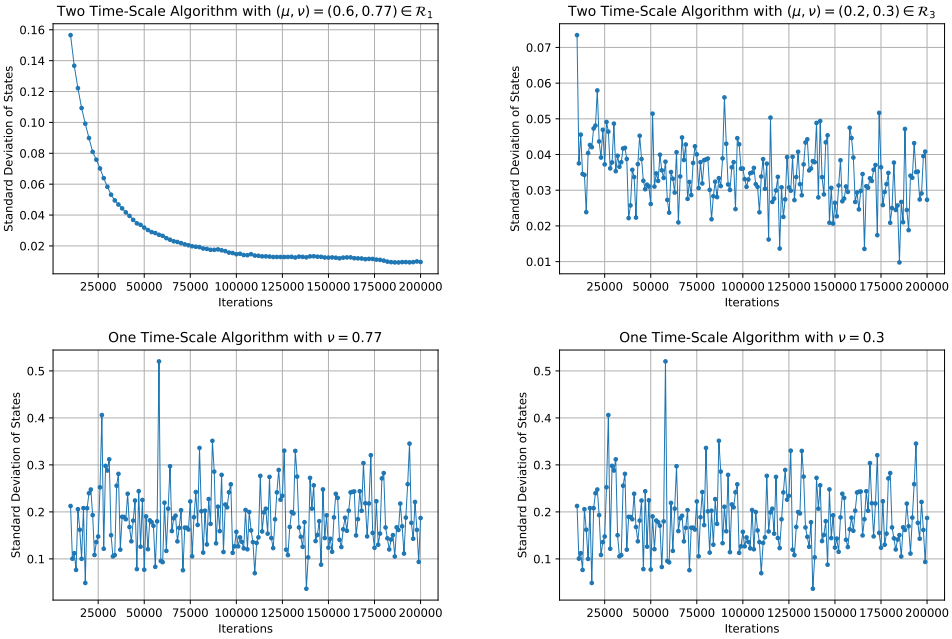

Figure 4.7: Standard Deviation of States vs. Iterations: Two Time-Scale Algorithm with $(\mu, \nu) = (0.6, 0.77)$ (top-left), $(0.2, 0.3)$ (top-right), and One Time-Scale Algorithm with $\nu = 0.77$ (bottom-left), 0.3 (bottom-right).

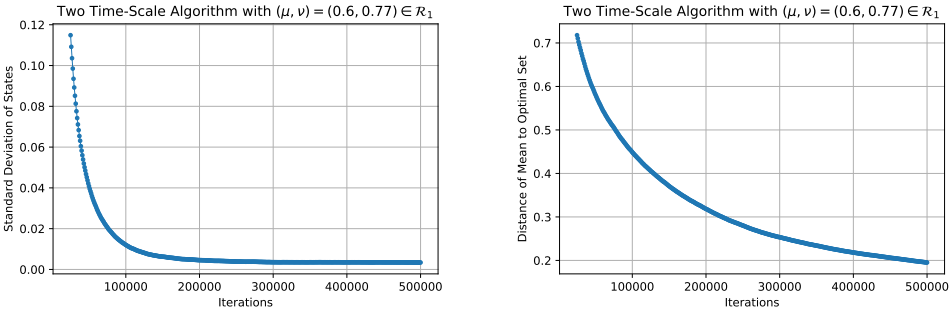

Figure 4.8: Standard Deviation of States and Distance of Mean State to Optimal Set vs. Iterations: Two Time-Scale Algorithm with $(\mu, \nu) = (0.6, 0.77)$.

We provide some numerical results to corroborate the derived theoretical analysis.

Data and Experimental Setup. We consider a time-varying network with $n = 6$ agents, with lost functions given by $f_i(x) = |x - v_i|$, where $x \in \mathbb{R}$ $(d = 1)$, and $v_i =$

$2 \times (i \bmod 2) - 1$. Note that the cost functions are convex (but not strongly convex) and non-differentiable at some points. We exploit the used time-varying graph in [18] where the mixing weight matrices are given by

$$[W(t)]_{ij} = \begin{cases} \frac{r_{\langle j \rangle}}{r_{\langle t \rangle} + r_{\langle t+1 \rangle}} & i, j \in \{\langle t \rangle, \langle t+1 \rangle\} \\ 1 & i = j \notin \{\langle t \rangle, \langle t+1 \rangle\} \\ 0 & \text{otherwise.} \end{cases}$$

Here $\langle i \rangle = (i - 1 \bmod n) + 1$. We assume the elements of the noise vector $E(t)$ to be i.i.d. $\mathcal{N}(0, 0.1)$ Gaussian random variables. The parameters of the dynamics in (1) are fine-tuned to $\alpha_0 = 0.0055$ and $\beta_0 = 0.21$. To complete the validation, we implement the one time-scale without any damping mechanism for the noise vectors, i.e., $\beta(t) = 1$ for every t.

Results. In Figure 4.6, top-left and top-right plots demonstrate the trajectories vs. training time for the dynamics (1) with choices of $(\mu, \nu) = (0.6, 0.77) \in \mathcal{R}_1$, and $(\mu, \nu) = (0.2, 0.3) \in \mathcal{R}_3$, respectively. It can be verified that the state of each node converges to an optimal point for the *first* pair (in $\mathcal{R}_1$) of (μ, ν) while the trajectories follow a random walk and do not converge for the *second* pair (in $\mathcal{R}_3$) of (μ, ν). In Figure 4.6, the bottom-left and bottom-right plots show the trajectories vs. training time for the one time-scale method with $\nu = 0.77$ and 0.3, respectively. It can be observed that the states of the node form a random walk, and it shows the privilege of exploiting the two-time scale in the presence of noise.

Furthermore, Figure 4.7 demonstrates the deviation of nodes' states from the mean state for every iteration. It can be verified that one-time scale and two-time scale algorithms with (μ, ν) in $\mathcal{R}_3$ both exhibit a random walk behavior for the states of the node. On the other hand, the two-time scale algorithm with (μ, ν) in $\mathcal{R}_1$ results in consensus among the states of all nodes.

To further demonstrate the effectiveness of the two-time scale algorithm, we conducted an experiment on a time-varying network consisting of $n = 6$ agents with each node's cost function defined as $f_i(\mathbf{x}) = \|\mathbf{x} - \boldsymbol{v}_i\|_1$, where $\boldsymbol{v}_i$ is a constant vector in $\mathbb{R}^{10}$. For our experiment, we set $\boldsymbol{v}_i = \mathbf{w}_1$ for nodes $i = 1, 3, 5$, and $\boldsymbol{v}_i = \mathbf{w}_2$ for nodes $i = 2, 4, 6$ where the elements of $\mathbf{w}_1$ and $\mathbf{w}_1$ are generated randomly from Gaussian distribution with $\mathcal{N}(0, 0.1)$. It is worth noting that these cost functions are convex, but not

strongly convex, and are non-differentiable at certain points. For the noise vectors, we sample points from $\mathcal{N}(0, 0.1)$. The parameters of the used dynamics are fine-tuned to $\alpha_0 = 0.0075$ and $\beta_0 = 0.12$. In Figure 4.8, the left and right plots demonstrate the deviation of states' nodes from the average state and the distance of the mean state from the optimal set for every iteration. By analyzing the plots, we can observe that each node's state gradually converges to an optimal point, as the distance from the average state and the optimal set decreases over time.

4.8.4 Proof of Theorem 4.4

Here, we provide the proof of Theorem 4.4. We first show that the deviation of the agents' states from their *average* converges to a random variable, which is later shown to be zero with probability 1. Next, we analyze the distance of the average state from an arbitrary point $\mathbf{x}^\star$ in the optimal set $\mathcal{X}^\star$ of the function $f(\cdot)$. These together lead to the proof of the theorem.

State Deviation from the Average State

We first define $\delta(t) := \|D(t)\|_{\mathbf{r}}$ where $D(t) := X(t) - \mathbf{1}\bar{\mathbf{x}}(t)$ and $\bar{\mathbf{x}}(t) := \mathbf{r}^T X(t) = \sum_{i=1}^n r_i \mathbf{x}_i(t)$. Our ultimate goal is to show that $\delta(t)$ vanishes almost surely. To that end, we first show its convergence in this section, and then show that it converges to 0. Since we are dealing with time-varying graphs, we cannot guarantee any decent for $\delta(t)$ in every iteration. However, since the graph is B-connected (see Assumption 4.2), such a claim can be made from $\delta(t)$ and $\delta(t+B)$. As a result, we can show that for any $1 \le \tau \le B$ the sequence $\{\delta(\tau + kB)\}_{k=0}^{\infty}$ converges almost surely to a (non-negative) random variable v_τ.

Starting from (4.119), we can write

$$X(t) = \sum_{s=1}^{t-1} \Phi(t:s)U(s) + \Phi(t:0)X(1). \tag{4.122}$$

Assuming $X(1) = \mathbf{0}$, the dynamics in (4.122) reduces to

$$X(t) = \sum_{s=1}^{t-1} \Phi(t:s)U(s). \tag{4.123}$$

By multiplying both sides of (4.123) from the left by $\mathbf{r}^T$ and using the fact $\mathbf{r}^T A(t) = \mathbf{r}^T$ for every $t \ge 1$, we get $\bar{\mathbf{x}} = \mathbf{r}^T X(t) = \sum_{s=1}^{t-1} \mathbf{r}^T \Phi(t:s)U(s) = \sum_{s=1}^{t-1} \mathbf{r}^T U(s)$. Subtracting

$\mathbf{1}\bar{\mathbf{x}}(t)$ from (4.123), we get

$$D(t) = \sum_{s=1}^{t-1}(\Phi(t:s) - \mathbf{1}\mathbf{r}^T)U(s).$$

Writing this equation for iteration $t+B$ we get

$$\begin{aligned}
D(t+B) &= \sum_{s=1}^{t+B-1}(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)U(s) \\
&= \sum_{s=1}^{t-1}(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)U(s) + \sum_{s=t}^{t+B-1}(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)U(s) \\
&\overset{(a)}{=} \Phi(t+B:t-1)\sum_{s=1}^{t-1}(\Phi(t:s) - \mathbf{1}\mathbf{r}^T)U(s) + \sum_{s=t}^{t+B-1}(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)U(s) \\
&= \Phi(t+B:t-1)D(t) + \sum_{s=t}^{t+B-1}(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)U(s), \qquad (4.124)
\end{aligned}$$

where in (a) we used the fact $A(s)\mathbf{1} = \mathbf{1}$ for every $s \geq 1$. This leads to

$$\begin{aligned}
&\mathbb{E}\left[\delta^2(t+B)\middle|\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\|D(t+B)\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|\Phi(t+B:t-1)D(t) + \sum_{s=t}^{t+B-1}(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)U(s)\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\|\Phi(t+B:t-1)D(t)\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right] + \mathbb{E}\left[\left\|\sum_{s=t}^{t+B-1}(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)U(s)\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right] \\
&\quad + 2\sum_{i=1}^{n} r_i \mathbb{E}\left[\left\langle \sum_{s=t}^{t+B-1}\left[(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)U(s)\right]_i, \left[\Phi(t+B:t-1)D(t)\right]_i \right\rangle \middle| \mathcal{F}_t\right]. \qquad (4.125)
\end{aligned}$$

Next, we bound each term in (4.125), separately. Note that $\mathbf{r}^T X(t) = \bar{\mathbf{x}}(t)$ and $\mathbf{r}^T\mathbf{1} = 1$. Hence

$$\mathbf{1}\mathbf{r}^T D(t) = \mathbf{1}\mathbf{r}^T(X(t) - \mathbf{1}\bar{\mathbf{x}}(t)) = \mathbf{1}\bar{\mathbf{x}}(t) - \mathbf{1}\bar{\mathbf{x}}(t) = 0.$$

Therefore, using Lemma 4.1, the first term can be bounded as

$$\begin{aligned}
\mathbb{E}\left[\|\Phi(t+B:t-1)D(t)\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right] &= \mathbb{E}\left[\left\|\left(\Phi(t+B:t-1) - \mathbf{1}\mathbf{r}^T\right)D(t)\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right] \\
&= \left\|\left(\Phi(t+B:t-1) - \mathbf{1}\mathbf{r}^T\right)D(t)\right\|_{\mathbf{r}}^2 \\
&\leq \left(1 - \lambda B\beta(t+B-1)\right)\delta^2(t). \qquad (4.126)
\end{aligned}$$

Next, in order to bound the second term, we can use the convexity of the $\|\cdot\|_{\mathbf{r}}$ to write

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{s=t}^{t+B-1}(\Phi(k+B:s)-\mathbf{1}\mathbf{r}^T)U(s)\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right]\\
&\leq \mathbb{E}\left[B\sum_{s=t}^{t+B-1}\left\|(\Phi(t+B:s)-\mathbf{1}\mathbf{r}^T)U(s)\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right]\\
&\leq 2B\sum_{s=t}^{t+B-1}\beta^2(s)\mathbb{E}\left[\left\|(\Phi(t+B:s)-\mathbf{1}\mathbf{r}^T)E(s)\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right]\\
&\quad+2B\sum_{s=t}^{t+B-1}\hat{\alpha}^2(s)\mathbb{E}\left[\left\|(\Phi(t+B:s)-\mathbf{1}\mathbf{r}^T)G(s)\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right],
\end{aligned}\tag{4.127}$$

where the second inequality follows from the fact that $U(s)=\beta(s)E(s)\hat{\alpha}(s)G(s)$ and Lemma 4.3 with $\theta=1$. For the first term in (4.127), from Lemma 4.1 we have

$$\begin{aligned}
&\sum_{s=t}^{t+B-1}\beta^2(s)\mathbb{E}\left[\left\|(\Phi(t+B:s)-\mathbf{1}\mathbf{r}^T)E(s))\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right]\\
&\leq \sum_{s=t}^{t+B-1}\beta^2(s)\mathbb{E}\left[\|E(s)\|_{\mathbf{r}}^2|\mathcal{F}_t\right]\leq\gamma\sum_{s=t}^{t+B-1}\beta^2(s),
\end{aligned}\tag{4.128}$$

where the last inequality follows from Assumption 4.1 for every $s\geq t$ we can write

$$\begin{aligned}
\mathbb{E}\left[\|E(s)\|_{\mathbf{r}}^2|\mathcal{F}_t\right]&=\mathbb{E}\left[\mathbb{E}\left[\|E(s)\|_{\mathbf{r}}^2|\mathcal{F}_s\right]|\mathcal{F}_t\right]\\
&=\sum_{i=1}^{n}r_i\mathbb{E}\left[\mathbb{E}\left[\|\mathbf{e}_i(s)\|^2|\mathcal{F}_s\right]|\mathcal{F}_t\right]\leq\sum_{i=1}^{n}r_i\mathbb{E}\left[\gamma|\mathcal{F}_t\right]=\gamma.
\end{aligned}$$

Similarly, for the second term in (4.127), we can apply Lemma 4.1 and write

$$\begin{aligned}
\sum_{s=t}^{t+B-1}\hat{\alpha}^2(s)\mathbb{E}\left[\left\|(\Phi(t+B:s)-\mathbf{1}\mathbf{r}^T)G(s)\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right]&\leq\sum_{s=t}^{t+B-1}\hat{\alpha}^2(s)\mathbb{E}\left[\|G(s)\|_{\mathbf{r}}^2|\mathcal{F}_t\right]\\
&\leq L^2\sum_{s=t}^{t+B-1}\hat{\alpha}^2(s),
\end{aligned}\tag{4.129}$$

where in the last inequality follows from Assumption 4.4-(c). Plugging (4.128) and (4.129) into (4.127), we arrive at

$$\mathbb{E}\left[\left\|\sum_{s=t}^{t+B-1}(\Phi(t+B:s)-\mathbf{1}\mathbf{r}^T)U(s)\right\|_{\mathbf{r}}^2\middle|\mathcal{F}_t\right]\leq 2B\sum_{s=t}^{t+B-1}\left(\gamma\hat{\alpha}^2(s)+L^2\beta^2(s)\right).\tag{4.130}$$

Next, we focus on each term of the last summation in (4.125). Since $U(s) = \beta(s)E(s) - \hat{\alpha}(s)G(s)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \sum_{s=t}^{t+B-1} \left[(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)U(s)\right]_i, \left[\Phi(t+B:t-1)D(t)\right]_i \right\rangle \middle| \mathcal{F}_t \right] \\
&= \mathbb{E}\left[\left\langle \sum_{s=t}^{t+B-1} \left[(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)\beta(s)E(s)\right]_i, \left[\Phi(t+B:t-1)D(t)\right]_i \right\rangle \middle| \mathcal{F}_t \right] \\
&\quad + \mathbb{E}\left[\left\langle \sum_{s=t}^{t+B-1} \left[(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)(-\hat{\alpha}(s)G(s))\right]_i, \left[\Phi(t+B:t-1)D(t)\right]_i \right\rangle \middle| \mathcal{F}_t \right]. \qquad (4.131)
\end{aligned}$$

For the first term in (4.131), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \sum_{s=t}^{t+B-1} \left[(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)\beta(s)E(s)\right]_i, \left[\Phi(t+B:t-1)D(t)\right]_i \right\rangle \middle| \mathcal{F}_t \right] \\
&= \mathbb{E}\left[\left\langle \sum_{s=t}^{t+B-1} \beta(s)\left[(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)\mathbb{E}\left[E(s)|\mathcal{F}_t\right]\right]_i, \left[\Phi(t+B:t-1)D(t)\right]_i \right\rangle \right] = 0, \qquad (4.132)
\end{aligned}$$

where the last inequality holds since from Assumption 4.1 for every $s \geq t$ we get

$$\mathbb{E}\left[E(s)|\mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{E}\left[E(s)|\mathcal{F}_s\right]|\mathcal{F}_t\right] = 0.$$

For the second term in (4.131), using the Cauchy-Schwarz inequality and the fact that $2ab \leq a^2 + b^2$, we can write

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \sum_{s=t}^{t+B-1} \left[(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)(-\hat{\alpha}(s)G(s))\right]_i, \left[\Phi(t+B:t-1)D(t)\right]_i \right\rangle \middle| \mathcal{F}_t \right] \\
&\leq \mathbb{E}\left[\left\| \sum_{s=t}^{t+B-1} \hat{\alpha}(s)\left[(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)G(s)\right]_i \right\| \times \left\|\left[\Phi(t+B:t-1)D(k)\right]_i\right\| \middle| \mathcal{F}_t \right] \\
&= \frac{1}{2}\mathbb{E}\left[\frac{1}{\omega(t)}\left\| \sum_{s=t}^{t+B-1} \hat{\alpha}(s)\left[(\Phi(t+B:s) - \mathbf{1}\mathbf{r}^T)G(s)\right]_i \right\|^2 + \omega(t)\left\|\left[\Phi(t+B:t-1)D(t)\right]_i\right\|^2 \middle| \mathcal{F}_t \right], \\
&\qquad (4.133)
\end{aligned}$$

for any $\omega(t) > 0$, which will be determined later. Hence, using (4.132) and (4.133)

in (4.131) we get

$$
\begin{aligned}
&2\sum_{i=1}^{n} r_i \mathbb{E}\left[\left\langle \sum_{s=t}^{t+B-1} \left[(\Phi(t+B:s)-\mathbf{1}\mathbf{r}^T)U(s)\right]_i, [\Phi(t+B:t-1)D(t)]_i \right\rangle \middle| \mathcal{F}_t\right] \\
&\leq \frac{1}{\omega(t)} \sum_{i=1}^{n} r_i \mathbb{E}\left[\left\| \sum_{s=t}^{t+B-1} \hat{\alpha}(s) \left[(\Phi(t+B:s)-\mathbf{1}\mathbf{r}^T)G(s)\right]_i \right\|^2 \middle| \mathcal{F}_t\right] \\
&\quad + \omega(t) \sum_{i=1}^{n} r_i \left\|[\Phi(t+B:t-1)D(t)]_i\right\|^2 \\
&\overset{(a)}{\leq} \frac{B}{\omega(t)} \sum_{s=t}^{t+B-1} \hat{\alpha}^2(s) \mathbb{E}\left[\left\|(\Phi(t+B:s)-\mathbf{1}\mathbf{r}^T)G(s)\right\|_{\mathbf{r}}^2 \middle| \mathcal{F}_t\right] + \omega(t) \left\|\Phi(t+B:t-1)D(t)\right\|_{\mathbf{r}}^2 \\
&\overset{(b)}{\leq} \frac{BL^2}{\omega(t)} \sum_{s=t}^{t+B-1} \hat{\alpha}^2(s) + \omega(t)(1-\lambda B\beta(t+B-1))\,\delta^2(t),
\end{aligned} \tag{4.134}
$$

where step (a) follows from the convexity of $\|\cdot\|_{\mathbf{r}}$ and the inequality in (b) follows from (4.129) and (4.126). Finally, plugging (4.126), (4.130), and (4.134) into (4.125) we have

$$
\begin{aligned}
\mathbb{E}\left[\delta^2(t+B)\middle|\mathcal{F}_t\right] &\leq (1+\omega(t))\,(1-\lambda B\beta(t+B-1))\,\delta^2(t) \\
&\quad + 2B \sum_{s=t}^{t+B-1} \left(\gamma\hat{\alpha}^2(s) + L^2\beta^2(s)\right) + \frac{BL^2}{\omega(t)} \sum_{s=t}^{t+B-1} \hat{\alpha}^2(s) \\
&\overset{(a)}{\leq} (1+\omega(t))\,(1-\lambda B\beta(t+B-1))\,\delta^2(t) \\
&\quad + 2B^2 \left(\gamma\hat{\alpha}^2(t) + L^2\beta^2(t)\right) + \frac{B^2L^2}{\omega(t)}\hat{\alpha}^2(t) \\
&\overset{(b)}{=} (1+\lambda B\beta(t))\,(1-\lambda B\beta(t+B-1))\,\delta^2(t) \\
&\quad + 2B^2 \left(\gamma\hat{\alpha}^2(t) + L^2\beta^2(t)\right) + \frac{BL^2}{\lambda}\frac{\hat{\alpha}^2(t)}{\beta(t)} \\
&\leq (1+\lambda B(\beta(t)-\beta(t+B-1)))\,\delta^2(t) \\
&\quad + 2B^2 \left(\gamma\hat{\alpha}^2(t) + L^2\beta^2(t)\right) + \frac{BL^2}{\lambda}\frac{\hat{\alpha}^2(t)}{\beta(t)},
\end{aligned} \tag{4.135}
$$

where the inequality in (a) follows from this assumption that $\{\hat{\alpha}(t)\}$ and $\{\beta(t)\}$ are non-increasing step-size sequences and in the step (b) we set $\omega(t) = \lambda B\beta(t)$.

Now, for $\tau = 1, 2, \ldots, B$, consider B random processes $\left\{\delta^2(\tau + kB)\right\}_{k=0}^{\infty}$. Due to (4.135),

each of these random processes satisfy the inequality (4.20) in Theorem 4.10, with

$$\zeta(k) := \lambda B(\beta(\tau + kB) - \beta(\tau + (k+1)B - 1)),$$
$$u(k) := 0,$$
$$z(k) := 2B^2\left(\gamma\hat{\alpha}^2(\tau + kB) + L^2\beta^2(\tau + kB)\right) + \frac{BL^2}{\lambda}\frac{\hat{\alpha}^2(\tau + kB)}{\beta(\tau + kB)}.$$

Since $\{\beta(t)\}$ is a non-increasing sequence, we have

$$\begin{aligned}\sum_{k=0}^{\infty}\zeta(k) &= B\lambda\sum_{k=0}^{\infty}\beta(\tau + kB) - \beta(\tau + (k+1)B) - 1)\\ &\leq B\lambda\sum_{k=0}^{\infty}\beta(\tau + kB) - \beta(\tau + (k+1)B)\\ &\leq B\lambda\beta(\tau) < \infty.\end{aligned}$$

Moreover, we have $\sum_{k=0}^{\infty}u(k) = 0$ and Assumption 4.5-(b) and (c) imply that $\sum_{k=0}^{\infty}z(k) < \infty$. Thus, all the conditions of Robbin-Sigmund Theorem are satisfied, and hence, for any $\tau = 1, \ldots, B$, the random process $\left\{\delta^2(\tau + kB)\right\}_{k=0}^{\infty}$ converges, almost surely. Consequently, there exist random variables v_τ such that $\lim_{k\to\infty}\delta^2(\tau + kB) = v_\tau$ almost surely, for $\tau = 1, \ldots, B$.

Accumulative Variance from the States

Here, we study the summation $\sum_{t=1}^{\infty}\hat{\alpha}(t)\delta(t)$, and show that it converges. This will imply that $\delta(t)$ converges to zero. Assuming 4.1, 4.2, 4.4-(c), and for non-increasing sequence $\{\beta(t)\}$ with $0 < \beta(t) \leq 1$ for all $t \geq 1$, we have

$$\begin{aligned}\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \leq\ & c_3\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\\ &+ c_4\sum_{s=1}^{t-1}\left[\frac{\hat{\alpha}^2(s)}{\beta(s)}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}\right],\end{aligned}$$

for some constants $c_3, c_4 > 0$. This is shown as part of Theorem 4.1. This together with the fact that

$$\sqrt{a+b} \leq \sqrt{a} + \sqrt{b},$$

leads us to

$$\begin{aligned}\mathbb{E}\left[\delta(t)\right] &= \mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}\right] \\ &\leq \left(c_3 \sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\right)^{\frac{1}{2}} \\ &\quad + \left(c_4 \sum_{s=1}^{t-1}\left[\frac{\hat{\alpha}^2(s)}{\beta(s)}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}\right]\right)^{\frac{1}{2}}.\end{aligned} \tag{4.136}$$

Next, we bound each summation in (4.136). To this end, we use Lemma 4.11 with $p_1(t) = \beta^2(t)$, $q_1(t) = \lambda\beta(t)$, and $A_1 = \frac{2c_1}{\lambda} < 1$ for the first summation. Using the fact that $\{\beta(t)\}$ is a non-increasing sequence and Assumption 4.5-(d), for every $t \geq t_0$ we have

$$\begin{aligned}-\Delta p_1(t) &= \beta^2(t) - \beta^2(t+1) \\ &= -\Delta\beta(t)(\beta(t) + \beta(t+1)) \\ &\leq (c_1\beta^2(t)) \cdot (2\beta(t)) = 2c_1\beta^3(t) = A_1 p_1(t) q_1(t).\end{aligned}$$

Thus, Lemma 4.11 leads to

$$\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right] \leq \frac{S_1}{\lambda}\beta(t), \tag{4.137}$$

for some constant S_1 and all $t \geq t_0$.

Similarly, we use Lemma 4.11 with $p_2(t) = \frac{\hat{\alpha}^2(t)}{\beta(t)}$, $q_2(t) = \frac{\lambda}{2}\beta(t)$, and $A_2 = \frac{4c_2}{\lambda} < 1$ to bound the second summation in (4.136). Using the fact that $\{\hat{\alpha}(t)\}$ and $\{\beta(t)\}$ are non-increasing sequences and Assumption 4.5-(e), we can write

$$\begin{aligned}-\Delta p_2(t) &= \frac{\hat{\alpha}^2(t)}{\beta(t)} - \frac{\hat{\alpha}^2(t+1)}{\beta(t+1)} \\ &\leq \frac{\hat{\alpha}^2(t)}{\beta(t)} - \frac{\hat{\alpha}^2(t+1)}{\beta(t)} \\ &= \frac{(-\Delta\hat{\alpha}(t))(\hat{\alpha}(t+1) + \hat{\alpha}(t))}{\beta(t)} \\ &= \frac{(-\Delta\hat{\alpha}(t))(\hat{\alpha}(t+1) + \hat{\alpha}(t))}{\beta(t)} \\ &\leq \frac{(c_2\hat{\alpha}(t)\beta(t))(2\hat{\alpha}(t))}{\beta(t)} \\ &\leq 2c_2\hat{\alpha}^2(t) = A_2 p_2(t) q_2(t),\end{aligned}$$

for $t \geq t_0$. Thus, Lemma 4.11 together with the fact that $\sqrt{1-x} \leq 1-x/2$ imply

$$\sum_{s=1}^{t-1}\left[\frac{\hat{\alpha}^2(s)}{\beta(s)}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}\right] \leq \sum_{s=1}^{t-1}\left[\frac{\hat{\alpha}^2(s)}{\beta(s)}\prod_{k=s+1}^{t-1}\left(1-\frac{\lambda}{2}\beta(k)\right)\right] \leq \frac{2S_2}{\lambda}\frac{\hat{\alpha}^2(t)}{\beta^2(t)}, \tag{4.138}$$

for some constant S_2 and every $t \geq t_0$. Plugging (4.137) and (4.138) into (4.136), we get

$$\mathbb{E}\left[\delta(t)\right] \leq \left(\frac{c_3 S_1}{\lambda}\beta(t)\right)^{\frac{1}{2}} + \left(\frac{2c_4 S_2}{\lambda}\frac{\hat{\alpha}^2(t)}{\beta^2(t)}\right)^{\frac{1}{2}}. \tag{4.139}$$

Therefore, from Equation (4.120) in Remark 4.6 and Assumption 4.5-(c) we can conclude

$$\begin{aligned}
&\lim_{T\to\infty}\mathbb{E}\left[\sum_{t=1}^{T}\hat{\alpha}(t)\delta(t)\right] \\
&= \lim_{T\to\infty}\sum_{t=1}^{T}\hat{\alpha}(t)\mathbb{E}\left[\delta(t)\right] \\
&\leq \sqrt{\frac{c_3 S_1}{\lambda}}\sum_{t=1}^{\infty}\hat{\alpha}(t)\beta^{\frac{1}{2}}(t) + \sqrt{\frac{2c_4 S_2}{\lambda}}\sum_{t=1}^{\infty}\frac{\hat{\alpha}^2(t)}{\beta(t)} < \infty.
\end{aligned} \tag{4.140}$$

Using Monotone Convergence Theorem, we have

$$\mathbb{E}\left[\sum_{t=1}^{\infty}\hat{\alpha}(t)\delta(t)\right] = \lim_{T\to\infty}\mathbb{E}\left[\sum_{t=1}^{T}\hat{\alpha}(t)\delta(t)\right] < \infty,$$

which implies

$$\sum_{t=1}^{\infty}\hat{\alpha}(t)\delta(t) < \infty, \tag{4.141}$$

almost surely.

Now, we aim to show that random variables $v_1, v_2, \ldots, v_B$ are all zero, almost surely. We prove this claim by contradiction. Assume there exists some $\tau \in \{1, 2, \ldots, B\}$ such that $p := \Pr[v_\tau > 0] > 0$. Hence, by the continuity of measure, there exists some (deterministic) $\epsilon > 0$ such that $\Pr[v_\tau > \epsilon] > p/2 > 0$. Consider the event $\mathcal{A} = \{v_\tau > \epsilon\}$. Then $\lim_{k\to\infty}\delta^2(\tau + kB) = v_\tau$ and $\delta(t) \geq 0$ imply that for all $\omega \in \mathcal{A}$, there exists some k_0 (possibly depending on ω) such that $\delta(\tau + kB) \geq \sqrt{\epsilon/2}$ for $k \geq k_0$. Then, since $\{\hat{\alpha}(t)\}$ is

a non-increasing sequence, we have

$$\begin{aligned}\sum_{t=1}^{\infty}\hat{\alpha}(t)\delta(t) &\geq \sum_{k=0}^{\infty}\hat{\alpha}(\tau+kB)\delta(\tau+kB)\\ &\geq \sum_{k=k_0}^{\infty}\hat{\alpha}(\tau+kB)\delta(\tau+kB)\\ &\geq \sqrt{\frac{\epsilon}{2}}\sum_{k=k_0}^{\infty}\hat{\alpha}(\tau+kB)\\ &\geq \sqrt{\frac{\epsilon}{2}}\sum_{k=k_0}^{\infty}\frac{1}{B}\sum_{j=0}^{B-1}\hat{\alpha}(\tau+kB+j)\\ &= \frac{1}{B}\sqrt{\frac{\epsilon}{2}}\sum_{\ell=\tau+k_0B}^{\infty}\hat{\alpha}(\ell)=\infty,\end{aligned} \tag{4.142}$$

where the last equality follows from Assumption 4.5-(a). This implies that

$$\Pr\left[\sum_{t=1}^{\infty}\hat{\alpha}(t)\delta(t)=\infty\right]\geq \Pr(\mathcal{A})>\frac{p}{2}>0,$$

which is in contradiction with (4.141). Therefore, we have $v_1=v_2=\cdots=v_B=0$, with probability 1.

Average State Distance to an Optimal Point

Now, we derive an upper bound for the expected distance between the (weighted) average of the agents' states, i.e., $\bar{\mathbf{x}}(t)=\mathbf{r}^TX(t)$ and an arbitrary minimizer of $\mathbf{x}^\star\in\mathcal{X}^\star$ of the function $f(\mathbf{x})$. Recall that $\mathbf{r}^TA(t)=\mathbf{r}^T$. Hence, multiplying both sides of (4.119) by $\mathbf{r}^T$, subtracting $\mathbf{x}^\star$, and taking expectation, we arrive at

$$\begin{aligned}\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1)-\mathbf{x}^\star\|^2\middle|\mathcal{F}_t\right]&=\mathbb{E}\left[\left\|\bar{\mathbf{x}}(t)+\mathbf{r}^TU(t)-\mathbf{x}^\star\right\|^2\middle|\mathcal{F}_t\right]\\ &=\|\bar{\mathbf{x}}(t)-\mathbf{x}^\star\|^2+\mathbb{E}\left[\left\|\mathbf{r}^TU(t)\right\|^2\middle|\mathcal{F}_t\right]+2\left\langle\mathbb{E}\left[\mathbf{r}^TU(t)\middle|\mathcal{F}_t\right],\bar{\mathbf{x}}(t)-\mathbf{x}^\star\right\rangle.\end{aligned} \tag{4.143}$$

Using Lemma 4.3 with $\theta=1$, we can bound the second term in (4.143) as

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{r}^TU(t)\right\|^2\middle|\mathcal{F}_t\right]&=\mathbb{E}\left[\left\|\beta(t)\mathbf{r}^TE(t)-\hat{\alpha}(t)\mathbf{r}^TG(t)\right\|^2\middle|\mathcal{F}_t\right]\\ &\leq 2\beta^2(t)\mathbb{E}\left[\left\|\mathbf{r}^TE(t)\right\|^2\middle|\mathcal{F}_t\right]+2\hat{\alpha}^2(t)\left\|\mathbf{r}^TG(t)\right\|^2.\end{aligned} \tag{4.144}$$

From Assumption 4.4-(c), we arrive at

$$\begin{aligned}\left\|\mathbf{r}^TG(t)\right\|^2&=\mathbf{r}^TG(t)\left[G(t)\right]^T\mathbf{r}\\ &\leq \mathbf{r}^T(L^2\mathbf{1}\mathbf{1}^T)\mathbf{r}=L^2.\end{aligned} \tag{4.145}$$

Plugging (4.53) and (4.145) into (4.144), we can write

$$\mathbb{E}\left[\left\|\mathbf{r}^T U(t)\right\|_2^2 \middle| \mathcal{F}_t\right] \leq 2\beta^2(t)\gamma + 2\hat{\alpha}^2(t)L^2. \tag{4.146}$$

Recall that Assumption 4.1 implies $\mathbb{E}\left[\beta(t)\mathbf{r}^T E(t)\middle|\mathcal{F}_t\right] = 0$. Using this fact and linearity of inner product, we can bound the last term in (4.143) as

$$\begin{aligned}
&\langle \mathbb{E}\left[\mathbf{r}^T U(t)\middle|\mathcal{F}_t\right], \bar{\mathbf{x}}(t) - \mathbf{x}^\star\rangle \\
&= \langle \mathbb{E}\left[\beta(t)\mathbf{r}^T E(t)\middle|\mathcal{F}_t\right] - \mathbb{E}\left[\hat{\alpha}(t)\mathbf{r}^T G(t)\middle|\mathcal{F}_t\right], \bar{\mathbf{x}}(t) - \mathbf{x}^\star\rangle \\
&= -\hat{\alpha}(t)\langle \mathbf{r}^T G(t), \bar{\mathbf{x}}(t) - \mathbf{x}^\star\rangle \\
&= -\hat{\alpha}(t)\left\langle \textstyle\sum_{i=1}^n r_i \mathbf{g}_i(\mathbf{x}_i(t)), \bar{\mathbf{x}}(t) - \mathbf{x}^\star\right\rangle \\
&= -\hat{\alpha}(t)\textstyle\sum_{i=1}^n r_i \langle \mathbf{g}_i(\mathbf{x}_i(t)), \bar{\mathbf{x}}(t) - \mathbf{x}^\star\rangle.
\end{aligned} \tag{4.147}$$

Let us consider each summand in (4.147), where we can write

$$\langle \mathbf{g}_i(\mathbf{x}_i(t)), \bar{\mathbf{x}}(t) - \mathbf{x}^\star\rangle = \langle \mathbf{g}_i(\mathbf{x}_i(t)), \bar{\mathbf{x}}(t) - \mathbf{x}_i(t)\rangle + \langle \mathbf{g}_i(\mathbf{x}_i(t)), \mathbf{x}_i(t) - \mathbf{x}^\star\rangle. \tag{4.148}$$

Using the Cauchy-Schwarz inequality and Assumption 4.4-(c), the first term in (4.148) can be lower bounded as

$$\begin{aligned}
\langle \mathbf{g}_i(\mathbf{x}_i(t)), \bar{\mathbf{x}}(t) - \mathbf{x}_i(t)\rangle &\geq -\|\mathbf{g}_i(\mathbf{x}_i(t))\| \, \|\bar{\mathbf{x}}(t) - \mathbf{x}_i(t)\| \\
&\geq -L \, \|\bar{\mathbf{x}}(t) - \mathbf{x}_i(t)\| \, .
\end{aligned} \tag{4.149}$$

From the convexity of $f_i(\cdot)$ in Assumption 4.4-(a), for the second term in (4.148) we have

$$\begin{aligned}
\langle \mathbf{g}_i(\mathbf{x}_i(t)), \mathbf{x}_i(t) - \mathbf{x}^\star\rangle &\geq f_i(\mathbf{x}_i(t)) - f_i(\mathbf{x}^\star) \\
&= f_i(\bar{\mathbf{x}}(t)) - f_i(\mathbf{x}^\star) + f_i(\mathbf{x}_i(t)) - f_i(\bar{\mathbf{x}}(t)) \\
&\geq f_i(\bar{\mathbf{x}}(t)) - f_i(\mathbf{x}^\star) \ - K \, \|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\| \, ,
\end{aligned} \tag{4.150}$$

where the last inequality follows from Assumption 4.4-(c). Therefore, substituting (4.149) and (4.150) into (4.148), we get

$$\langle \mathbf{g}_i(\mathbf{x}_i(t)), \bar{\mathbf{x}}(t) - \mathbf{x}^\star\rangle \geq -2K \, \|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\| + f_i(\bar{\mathbf{x}}(t)) - f_i(\mathbf{x}^\star). \tag{4.151}$$

Replacing (4.151) in (4.147), and using the Cauchy-Schwarz inequality and the fact that $\sum_{i=1}^{n} r_i = 1$, we have

$$\begin{aligned}
&\langle \mathbb{E}\left[\mathbf{r}^T U(t) \middle| \mathcal{F}_t\right], \bar{\mathbf{x}}(t) - \mathbf{x}^\star \rangle \\
&\leq 2\hat{\alpha}(t) K \sum_{i=1}^{n} r_i \left\| \mathbf{x}_i(t) - \bar{\mathbf{x}}(t) \right\| - \hat{\alpha}(t) \sum_{i=1}^{n} r_i (f_i(\bar{\mathbf{x}}(t)) - f_i(\mathbf{x}^\star)) \\
&\leq 2\hat{\alpha}(t) K \sqrt{\sum_{i=1}^{n} r_i \cdot \sum_{i=1}^{n} r_i \left\| \mathbf{x}_i(t) - \bar{\mathbf{x}}(t) \right\|^2} \\
&\quad - \hat{\alpha}(t) \sum_{i=1}^{n} r_i (f_i(\bar{\mathbf{x}}(t)) - f_i(\mathbf{x}^\star)) \\
&= 2\hat{\alpha}(t) K \delta(t) - \hat{\alpha}(t) \left(f(\bar{\mathbf{x}}(t)) - f(\mathbf{x}^\star) \right). \qquad (4.152)
\end{aligned}$$

Plugging (4.146) and (4.152) into (4.143), we get

$$\begin{aligned}
\mathbb{E}\left[\left\| \bar{\mathbf{x}}(t+1) - \mathbf{x}^\star \right\|^2 \middle| \mathcal{F}_t \right] \leq \left\| \bar{\mathbf{x}}(t) - \mathbf{x}^\star \right\|^2 + 2(\beta^2(t)\gamma + \hat{\alpha}^2(t) K^2) \\
+ 4\hat{\alpha}(t) K \delta(t) - 2\hat{\alpha}(t) (f(\bar{\mathbf{x}}(t)) - f(\mathbf{x}^\star)),
\end{aligned}$$

which is identical to the inequality in Theorem 4.3, with $\mathbf{y}(t) = \bar{\mathbf{x}}(t)$, $\zeta(t) = 0$, $\xi(t) = 2\hat{\alpha}(t)$, and

$$z(t) = 4L\hat{\alpha}(t)\delta(t) + 2\gamma\beta^2(t) + 2L^2\hat{\alpha}^2(t).$$

Note that $\zeta(t)$, $\xi(t)$, and and $z(t)$ are all none-negative for $t \geq 1$, and $\sum_{t=1}^{\infty} \zeta(t) = 0 < \infty$. Moreover, Assumption 4.5-(a) implies $\sum_{t=1}^{\infty} \xi(t) = 2\sum_{t=1}^{\infty} \hat{\alpha}(t) = \infty$. Finally, from (4.141) and Assumption 4.5-b we have $\sum_{t=1}^{\infty} z(t) < \infty$. Therefore, we can apply Theorem 4.3, and conclude that $\{\bar{\mathbf{x}}(t)\}$ converges to some $\tilde{\mathbf{x}} \in \mathcal{X}^\star$, almost surely.

Almost Sure Convergence of State to an Optimal Point

We showed that under the Assumptions 4.1-4.5, we have $\lim_{t\to\infty} \left\| X(t) - \mathbf{1}\bar{\mathbf{x}}(t) \right\|_{\mathbf{r}} = 0$. This implies $\lim_{t\to\infty} \left\| \mathbf{x}_i(t) - \bar{\mathbf{x}}(t) \right\|_{\mathbf{r}} = 0$ for every $i \in [n]$. Moreover, we proved that $\bar{\mathbf{x}}(t)$ converges to some $\tilde{\mathbf{x}} \in \mathcal{X}^\star$, almost surely. Combining these two results, immediately conclude the claim of Theorem 4.4.

4.8.5 Proof of Proposition 4.1

Now, we prove Proposition 4.1. We only need to show that step-sizes $\beta(t) = \frac{\beta_0}{t^\mu}$ and $\hat{\alpha}(t) = \frac{\alpha_0}{t^\nu}$ with $\frac{1}{2} < \mu \leq 1$ and $\frac{1}{2}(1+\mu) < \nu \leq 1$ satisfy all conditions in Assumption 4.5.

First, from $\beta_0 \le 1$, we get $\beta(t) = \frac{\beta_0}{t^\mu} \le 1$ for all $t \ge 1$. Using Lemma 4.8 with $\nu \le 1$, we have $\sum_{t=1}^{\infty} \hat{\alpha}(t) = \sum_{t=1}^{\infty} \frac{\alpha_0}{t^\nu} = \infty$. Moreover, from Lemma 4.8 with $\mu > \frac{1}{2}$, $\nu > \frac{1}{2}$, and $2\nu - \mu > 1$ we arrive at

$$\sum_{t=1}^{\infty} \beta^2(t) = \sum_{t=1}^{\infty} \frac{\beta_0^2}{t^{2\mu}} < \infty,$$
$$\sum_{t=1}^{\infty} \hat{\alpha}^2(t) = \sum_{t=1}^{\infty} \frac{\alpha_0^2}{t^{2\nu}} < \infty,$$
$$\sum_{t=1}^{\infty} \frac{\hat{\alpha}^2(t)}{\beta(t)} = \sum_{t=1}^{\infty} \frac{\alpha_0^2}{\beta_0} \frac{1}{t^{2\nu-\mu}} < \infty.$$

Now, we need to show that for some positive constants $c_1 < \frac{\lambda}{2}$, $c_2 < \frac{\lambda}{4}$, and $t_0 \ge 1$ we have $-\Delta\beta(t) \le c_1\beta^2(t)$ and $-\Delta\hat{\alpha}(t) \le c_2\hat{\alpha}(t)\beta(t)$ for all $t \ge t_0$.

Using the mean value theorem for the function $\beta(t) = \frac{\beta_0}{t^\mu}$ we have $\beta(t+1) - \beta(t) = \beta'(\zeta_1)$ for some $\zeta_1 \in [t, t+1]$. Therefore, we arrive at

$$\begin{aligned} -\Delta\beta(t) &= \beta(t) - \beta(t+1) \\ &= -\beta'(\zeta_1) = \mu \frac{\beta_0}{\zeta_1^{\mu+1}} \le \mu \frac{\beta_0}{t^{\mu+1}} \le c_1 \frac{\beta_0^2}{t^{2\mu}} = c_1\beta^2(t), \end{aligned}$$

where the latter holds for $t \ge t_1 := \left(\frac{\mu}{\beta_0 c_1}\right)^{\frac{1}{1-\mu}}$ provided that $\mu < 1$. Similarly, for the the function $\hat{\alpha}(t) = \frac{\alpha_0}{t^\nu}$ we get $\hat{\alpha}(t+1) - \hat{\alpha}(t) = \hat{\alpha}'(\zeta_2)$ for some $\zeta_2 \in [t, t+1]$. Hence, we can write

$$\begin{aligned} -\Delta\hat{\alpha}(t) &= \hat{\alpha}(t) - \hat{\alpha}(t+1) \\ &= -\hat{\alpha}'(\zeta_2) = \nu \frac{\alpha_0}{\zeta_2^{\nu+1}} \le \nu \frac{\alpha_0}{t^{\nu+1}} \le c_2 \frac{\alpha_0\beta_0}{t^{\nu+\mu}} = c_2\hat{\alpha}(t)\beta(t), \end{aligned}$$

for every $t \ge t_2$ where $t_2 := \left(\frac{\nu}{\beta_0 c_2}\right)^{\frac{1}{1-\mu}}$. Therefore, for any pair of (fixed) positive constants $c_1 < \frac{\lambda}{2}$ and $c_2 < \frac{\lambda}{4}$ we have $-\Delta\beta(t) \le c_1\beta^2(t)$ and $-\Delta\hat{\alpha}(t) \le c_2\hat{\alpha}(t)\beta(t)$ for all $t \ge t_0 = \max(t_1, t_2)$. This shows that Assumption 4.5-(d)-(e) are satisfied for sufficiently large t, regardless of the dynamic parameters. This completes the proof of Proposition 4.1.

4.9 Concluding Remarks

We have studied distributed optimization over time-varying networks suffering from noisy and imperfect sharing of information. For the proposed algorithm, we obtain a convergence rate as a function of the damping and diminishing step-size parameters. By optimizing the achieved rate over all feasible choices for parameters, the algorithm obtains a convergence rate of $\mathcal{O}(T^{-1/2})$ and $\mathcal{O}(T^{-1/3+\epsilon})$ for strongly convex and non-convex settings for any $\epsilon > 0$, respectively. Further, we identified sufficient conditions for general step-sizes sequences for the two-time-scales to guarantee the algorithm's almost sure convergence for convex cost functions. Moreover, we used this result to characterize conditions on practical step-size sequences that enable almost sure convergence in this setting.

4.10 Proof of The Auxiliary Lemmas

In this section, we provide the proofs of auxiliary lemmas.

Proof of Lemma 4.1: Due to the separable nature of $\|\cdot\|_{\mathbf{r}}$, i.e., $\|U\|_{\mathbf{r}}^2 = \sum_{j=1}^{d} \|U^j\|_{\mathbf{r}}^2$, without loss of generality, we may assume that $d = 1$. Thus, let $U = \boldsymbol{u} \in \mathbb{R}^n$. Define $V_{\mathbf{r}} : \mathbb{R}^n \to \mathbb{R}^+$ by

$$V_{\mathbf{r}}(\boldsymbol{u}) := \left\|\boldsymbol{u} - \mathbf{1}\mathbf{r}^T\boldsymbol{u}\right\|_{\mathbf{r}}^2 = \sum_{i=1}^{n} r_i(u_i - \mathbf{r}^T\boldsymbol{u})^2. \tag{4.153}$$

For notational simplicity, let $\boldsymbol{u}(s) = \boldsymbol{u} = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix}$, and $\boldsymbol{u}(k+1) = A(k+1)\boldsymbol{u}(k)$. In addition with a slight abuse of notation, we denote $V_{\mathbf{r}}(\boldsymbol{u}(k))$ by $V_{\mathbf{r}}(k)$ for $k = s, \dots, t$.

Using Theorem 1 in [115], we have

$$V_{\mathbf{r}}(t) = V_{\mathbf{r}}(s) - \sum_{k=s+1}^{t} \sum_{i<j} H_{ij}(k)(u_i(k) - u_j(k))^2, \tag{4.154}$$

where $H(k) = A^T(k)\,\mathrm{diag}(\mathbf{r})A(k)$. Then, for $\boldsymbol{u}(s) = \boldsymbol{u}$ and $\boldsymbol{u}(t-1) = \Phi(t:s)\boldsymbol{u}(s)$ we have

$$\begin{aligned}
&\left\|(\Phi(t:s) - \mathbf{1}\mathbf{r}^T)\boldsymbol{u}\right\|_{\mathbf{r}}^2 \overset{(a)}{=} \left\|(I - \mathbf{1}\mathbf{r}^T)\Phi(t:s)\boldsymbol{u}\right\|_{\mathbf{r}}^2 \\
&= \left\|(I - \mathbf{1}\mathbf{r}^T)\boldsymbol{u}(t-1)\right\|_{\mathbf{r}}^2 = V_{\mathbf{r}}(t-1) \le V_{\mathbf{r}}(s) \\
&= \left\|\boldsymbol{u} - \mathbf{1}\mathbf{r}^T\boldsymbol{u}\right\|_{\mathbf{r}}^2 = \|\boldsymbol{u}\|_{\mathbf{r}}^2 - \left\|\mathbf{1}\mathbf{r}^T\boldsymbol{u}\right\|_{\mathbf{r}}^2 \le \|\boldsymbol{u}\|_{\mathbf{r}}^2,
\end{aligned} \tag{4.155}$$

where (a) follows from Assumption 4.2-(a) and the fact that $A(k) = (1-\beta(k))I + \beta(k)W(k)$, which imply $\mathbf{r}^T\Phi(t:s) = \mathbf{r}^T$. This shows the claim in (4.13).

Note that $A(k)$ is a non-negative matrix, then $H(k) \geq \mathbf{r}_{\min} A^T(k)A(k)$, for $k = s+1, \ldots, t$. Also, since $A(k) = (1-\beta(k))I + \beta(k)W(k)$, then Assumption 4.2-(b) implies that the minimum non-zero elements of $A(k)$ are bounded bellow by $\eta\beta(k)$. Therefore, since $\beta(k)$ is non-increasing, on the window $k = s+1, \ldots, s+B$, the minimum non-zero elements of $A(k)$ for k in this window are lower bounded by $\eta\beta(s+B)$. Without loss of generality, assume that the entries of $\boldsymbol{u}$ are sorted, i.e., $u_1 \leq \ldots \leq u_n$, otherwise, we can relabel the agents (rows and columns of $A(k)$s and $\boldsymbol{u}$ to achieve this). Therefore, by Lemma 8 in [51], for (4.154), we have

$$\begin{aligned} V_{\mathbf{r}}(s+B) &\leq V_{\mathbf{r}}(s) - \mathbf{r}_{\min} \sum_{k=s+1}^{s+B} \sum_{i<j} [A^T(k)A(k)]_{ij} (u_i(k) - u_j(k))^2 \\ &\leq V_{\mathbf{r}}(s) - \frac{\eta \mathbf{r}_{\min}}{2} \beta(s+B) \sum_{\ell=1}^{n-1} (u_{\ell+1} - u_\ell)^2. \end{aligned} \tag{4.156}$$

We may comment here that although Lemma 8 in [51] is written for doubly stochastic matrices, and its statement is about the decrease of $V_{\mathbf{r}}(\mathbf{x})$ for the special case of $\mathbf{r} = \frac{1}{n}\mathbf{1}$, but in fact, it is a result on bounding $\sum_{k=s+1}^{s+B} \sum_{i<j} [A^T(k)A(k)]_{ij}(u_i(k) - u_j(k))^2$ for a sequence of B-connected stochastic matrices $A(k)$ in terms of the minimum non-zero entries of stochastic matrices $A(s+1), \ldots, A(s+B)$.

Next, we will show that $\sum_{\ell=1}^{n-1}(u_{\ell+1} - u_\ell)^2 \geq n^{-2} V_{\mathbf{r}}(\boldsymbol{u})$. This argument adapts a similar argument used in the proof of Theorem 18 in [51] to the general $V_{\mathbf{r}}(\cdot)$.

For a $\boldsymbol{v} \in \mathbb{R}^n$ with $V_{\mathbf{r}}(\boldsymbol{v}) > 0$, define the quotient

$$h(\boldsymbol{v}) = \frac{\sum_{\ell=1}^{n-1}(v_{\ell+1} - v_\ell)^2}{\sum_{i=1}^{n} r_i (v_i - \mathbf{r}^T\boldsymbol{v})^2} = \frac{\sum_{\ell=1}^{n-1}(v_{\ell+1} - v_\ell)^2}{V_{\mathbf{r}}(\boldsymbol{v})}. \tag{4.157}$$

Note that $h(\boldsymbol{v})$ is invariant under scaling and translations by all-one vector, i.e., $h(\omega\boldsymbol{v}) = h(\boldsymbol{v})$ for all non-zero $\omega \in \mathbb{R}$ and $h(\boldsymbol{v} + \omega\mathbf{1}) = h(\boldsymbol{v})$ for all $\omega \in \mathbb{R}$. Therefore,

$$\begin{aligned} \min_{\substack{v_1 \leq v_2 \leq \cdots \leq v_n \\ V_{\mathbf{r}}(\boldsymbol{v}) \neq 0}} h(\boldsymbol{v}) &= \min_{\substack{v_1 \leq v_2 \leq \cdots \leq v_n \\ \mathbf{r}^T\boldsymbol{v} = 0, V_{\mathbf{r}}(\boldsymbol{v}) = 1}} h(\boldsymbol{v}) \\ &= \min_{\substack{v_1 \leq v_2 \leq \cdots \leq v_n \\ \mathbf{r}^T\boldsymbol{v} = 0, V_{\mathbf{r}}(\boldsymbol{v}) = 1}} \sum_{\ell=1}^{n-1} (v_{\ell+1} - v_\ell)^2. \end{aligned} \tag{4.158}$$

Since $\mathbf{r}$ is a stochastic vector, then for a vector $\boldsymbol{v}$ with $v_1 \le \ldots \le v_n$ and $\mathbf{r}^T\boldsymbol{v} = 0$, we would have $v_1 \le \mathbf{r}^T\boldsymbol{v} = 0 \le v_n$. On the other hand, the fact that $V_{\mathbf{r}}(\boldsymbol{v}) = \sum_{i=1}^n r_i v_i^2 = 1$ would imply $\max(|v_1|, |v_n|) \ge \frac{1}{\sqrt{n}}$. Let us consider the difference sequence $\hat{v}_\ell = v_{\ell+1} - v_\ell$ for $\ell = 1, \ldots, n-1$, for which we have $\sum_{i=1}^{n-1} \hat{v}_\ell = v_n - v_1 \ge v_n \ge \frac{1}{\sqrt{n}}$. Therefore, the optimization problem (4.158) can be rewritten as

$$\begin{aligned}\min_{\substack{v_1 \le v_2 \le \cdots \le v_n \\ V_{\mathbf{r}}(\boldsymbol{v}) \neq 0}} h(\boldsymbol{v}) &= \min_{\substack{v_1 \le v_2 \le \cdots \le v_n \\ \mathbf{r}^T\boldsymbol{v} = 0, V_{\mathbf{r}}(\boldsymbol{v}) = 1}} \sum_{\ell=1}^{n-1} (v_{\ell+1} - v_\ell)^2 \\ &\ge \min_{\substack{\hat{v}_1, \ldots, \hat{v}_{n-1} \ge 0 \\ \sum_{i=1}^{n-1} \hat{v}_i \ge \frac{1}{\sqrt{n}}}} \sum_{\ell=1}^{n-1} \hat{v}_\ell^2. \end{aligned} \tag{4.159}$$

Using the Cauchy-Schwarz inequality, we get

$$\left(\sum_{\ell=1}^{n-1} \hat{v}_\ell^2\right)\left(\sum_{\ell=1}^{n-1} 1^2\right) \ge \left(\sum_{\ell=1}^{n-1} \hat{v}_\ell\right)^2 \ge \left(\frac{1}{\sqrt{n}}\right)^2 = \frac{1}{n}.$$

Hence, we arrive at

$$\min_{\substack{v_1 \le v_2 \le \ldots \le v_n \\ V_{\mathbf{r}}(\boldsymbol{v}) \neq 0}} h(\boldsymbol{v}) \ge \frac{1}{n(n-1)} \ge \frac{1}{n^2}. \tag{4.160}$$

Thus, for $v_1 \le \ldots \le v_n$, we have $\sum_{\ell=1}^{n-1} (v_{\ell+1} - v_\ell)^2 \ge n^{-2} V_{\mathbf{r}}(\boldsymbol{v})$ (note that this inequality also holds for $\boldsymbol{v} \in \mathbb{R}^n$ with $V_{\mathbf{r}}(\boldsymbol{v}) = 0$). Using this fact in (4.156) implies

$$V_{\mathbf{r}}(s+B) \le \left(1 - \frac{\eta \mathbf{r}_{\min}}{2n^2} \beta(s+B)\right) V_{\mathbf{r}}(s). \tag{4.161}$$

Applying (4.161) for $\Delta := \lfloor \frac{t-1-s}{B} \rfloor$ steps recursively, we get

$$V_{\mathbf{r}}(s + \Delta B) \le \prod_{j=1}^{\Delta} \left(1 - \frac{\eta \mathbf{r}_{\min}}{2n^2} \beta(s + jB)\right) V_{\mathbf{r}}(s)$$

Using the fact that $(1-x)^{1/B} \le 1 - x/B$ and since $\{\beta(k)\}$ is a non-increasing sequence, we have

$$\begin{aligned} 1 - \frac{\eta \mathbf{r}_{\min}}{2n^2} \beta(s + jB) &= \prod_{\ell=1}^{B} \left(1 - \frac{\eta \mathbf{r}_{\min}}{2n^2} \beta(s + jB)\right)^{1/B} \\ &\le \prod_{\ell=1}^{B} \left(1 - \frac{\eta \mathbf{r}_{\min}}{2Bn^2} \beta(s + jB)\right) \\ &\le \prod_{\ell=1}^{B} \left(1 - \lambda \beta(s + jB + \ell)\right). \end{aligned}$$

From (4.154), $V_{\mathbf{r}}(t)$ is a non-increasing function of t. Thus, for $s+\Delta B \le t-1 < s+(\Delta+1)B$ we have

$$\begin{aligned} V_{\mathbf{r}}(t-1) &\le V_{\mathbf{r}}(s+\Delta B) \\ &\le \prod_{j=1}^{\Delta}\left(1-\frac{\eta \mathbf{r}_{\min}}{2n^2}\beta(s+jB)\right) V_{\mathbf{r}}(s) \\ &\le \prod_{j=1}^{\Delta}\prod_{\ell=1}^{B}\left(1-\lambda\beta(s+jB+\ell)\right) V_{\mathbf{r}}(s) \\ &= \prod_{k=s+B+1}^{s+(\Delta+1)B}\left(1-\lambda\beta(k)\right) V_{\mathbf{r}}(s) \\ &\le \prod_{k=s+B+1}^{t-1}\left(1-\lambda\beta(k)\right) V_{\mathbf{r}}(s). \end{aligned} \tag{4.162}$$

Next, noting that $\{\beta(k)\}$ is a non-increasing sequence, we have $\beta(k) \le \beta(1) = \beta_0$. Thus,

$$\begin{aligned} \prod_{k=s+1}^{s+B}\left(1-\lambda\beta(k)\right) &\ge \prod_{k=s+1}^{s+B}\left(1-\lambda\beta_0\right) \\ &= \left(1-\lambda\beta_0\right)^B \ge 1-B\lambda\beta_0. \end{aligned} \tag{4.163}$$

Therefore, combining (4.162) and (4.163), we get

$$\begin{aligned} V_{\mathbf{r}}(t-1) &\le \prod_{k=s+B+1}^{t-1}\left(1-\lambda\beta(k)\right) V_{\mathbf{r}}(s) \\ &\le \frac{\prod_{k=s+1}^{s+B}\left(1-\lambda\beta(k)\right)}{1-B\lambda\beta_0}\prod_{k=s+B+1}^{t-1}\left(1-\lambda\beta(k)\right) V_{\mathbf{r}}(s) \\ &= \frac{1}{1-B\lambda\beta_0}\prod_{k=s+1}^{t-1}\left(1-\lambda\beta(k)\right) V_{\mathbf{r}}(s). \end{aligned} \tag{4.164}$$

Now, we define

$$\Phi(t:s) = A(t-1)\cdots A(s+1)$$

for $t \ge s$ with $\Phi(t:t-1) = I$.

Note that Assumption 4.2-(a) and the fact that $A(k) = (1-\beta(k))I + \beta(k)W(k)$ imply $\mathbf{r}^T\Phi(t:s) = \mathbf{r}^T$. Then, setting $\boldsymbol{u}(s) = \boldsymbol{u} = U$ and $\boldsymbol{u}(t-1) = \Phi(t:s)\boldsymbol{u}(s) = \Phi(t:s)U$,

we can write

$$\begin{aligned}\left(\Phi(t:s)-\mathbf{1}\mathbf{r}^T\right)U &= \Phi(t:s)U-\mathbf{1}\mathbf{r}^TU\\ &= \Phi(t:s)U-\mathbf{1}\mathbf{r}^T\Phi(t:s)U\\ &= \boldsymbol{u}(t-1)-\mathbf{1}\mathbf{r}^T\boldsymbol{u}(t-1).\end{aligned}$$

Therefore, using (4.164) we have

$$\begin{aligned}\left\|\left(\Phi(t:s)-\mathbf{1}\mathbf{r}^T\right)U\right\|_{\mathbf{r}}^2 &= \left\|\boldsymbol{u}(t-1)-\mathbf{1}\mathbf{r}^T\boldsymbol{u}(t-1)\right\|_{\mathbf{r}}^2 = V_{\mathbf{r}}(t-1)\\ &\le \frac{1}{1-B\lambda\beta_0}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))V_{\mathbf{r}}(s)\\ &= \frac{1}{1-B\lambda\beta_0}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\left\|\boldsymbol{u}-\mathbf{1}\mathbf{r}^T\boldsymbol{u}\right\|_{\mathbf{r}}^2\\ &\le \frac{1}{1-B\lambda\beta_0}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\left\|\boldsymbol{u}\right\|_{\mathbf{r}}^2,\end{aligned}\tag{4.165}$$

where the second inequality follows from the fact that $\left\|\boldsymbol{u}-\mathbf{1}\mathbf{r}^T\boldsymbol{u}\right\|_{\mathbf{r}}^2+\left\|\mathbf{1}\mathbf{r}^T\boldsymbol{u}\right\|_{\mathbf{r}}^2=\left\|\boldsymbol{u}\right\|_{\mathbf{r}}^2$. Applying the inequality in (4.165) on each column of a matrix U, we can conclude the same result for matrices.

To show (4.13), we can continue from (4.161) and write

$$\begin{aligned}\left\|\left(\Phi(s+B+1:s)-\mathbf{1}\mathbf{r}^T\right)\boldsymbol{u}\right\|_{\mathbf{r}}^2 &= \left\|\left(I-\mathbf{1}\mathbf{r}^T\right)\Phi(s+B+1:s)\boldsymbol{u}\right\|_{\mathbf{r}}^2\\ &= \left\|(I-\mathbf{1}\mathbf{r}^T)\boldsymbol{u}(s+B)\right\|_{\mathbf{r}}^2\\ &= V_{\mathbf{r}}(s+B)\\ &\le (1-\lambda B\beta(s+B))\,V_{\mathbf{r}}(s)\\ &= (1-\lambda B\beta(s+B))\left\|\boldsymbol{u}-\mathbf{1}\mathbf{r}^T\boldsymbol{u}\right\|_{\mathbf{r}}^2\\ &\le (1-\lambda B\beta(s+B))\left\|\boldsymbol{u}\right\|_{\mathbf{r}}^2.\end{aligned}$$

Applying this inequality on each column of a matrix U, we can conclude the same result for matrices. This completes the proof of the lemma. ■ ■

Proof of Lemma 4.2: Recall that we denote the ith row of A by A_i, and the jth column of B by B^j. Then, applying the Cauchy-Schwartz inequality to vectors A_i and

B^j, we have $|[AB]_{ij}| = |\langle A_i, B^j \rangle| \le \|A_i\| \|B^j\|$. Therefore,

$$\begin{aligned}\|[AB]_i\|^2 &= \sum_{j=1}^{m} |[AB]_{ij}| = \sum_{j=1}^{m} |\langle A_i, B^j \rangle|^2 \\ &\le \|A_i\|^2 \sum_{j=1}^{m} \|B^j\|^2 \le \|A_i\|^2 \|B\|_F^2 .\end{aligned}$$

Using this inequality and the definition of $\mathbf{r}$-norm, we get

$$\|AB\|_{\mathbf{r}}^2 = \sum_{i=1}^{n} r_i \|[AB]_i\|^2 \le \sum_{i=1}^{n} r_i \|A_i\|^2 \|B\|_F^2 = \|A\|_{\mathbf{r}}^2 \|B\|_F^2 ,$$

as claimed in the lemma. ∎

Proof of Lemma 4.3: For two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ and any scalar $\theta > 0$, we have

$$\begin{aligned}\|\boldsymbol{u}+\boldsymbol{v}\|^2 &= \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 + 2\langle \boldsymbol{u}, \boldsymbol{v} \rangle \\ &\overset{(a)}{\le} \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 + 2\|\boldsymbol{u}\|\|\boldsymbol{v}\| \\ &= \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 + 2\left(\sqrt{\theta}\|\boldsymbol{u}\| \cdot \frac{1}{\sqrt{\theta}}\|\boldsymbol{v}\|\right) \\ &\overset{(b)}{\le} \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 + \theta\|\boldsymbol{u}\|^2 + \frac{1}{\theta}\|\boldsymbol{v}\|^2 \\ &= (1+\theta)\|\boldsymbol{u}\|^2 + \left(1+\frac{1}{\theta}\right)\|\boldsymbol{v}\|^2,\end{aligned}$$

where (a) follows from the Cauchy–Schwarz inequality and (b) is concluded from the geometric-arithmetic inequality. Similarly, recalling $\mathbf{r}$-norm, for matrices U and V, we have

$$\begin{aligned}\|U+V\|_{\mathbf{r}}^2 &= \sum_{i=1}^{n} r_i \|U_i + V_i\|^2 \\ &\le \sum_{i=1}^{n} r_i \left[(1+\theta)\|U_i\|^2 + \left(1+\frac{1}{\theta}\right)\|V_i\|^2\right] \\ &= (1+\theta)\sum_{i=1}^{n} r_i \|U_i\|^2 + \left(1+\frac{1}{\theta}\right)\sum_{i=1}^{n} r_i \|V_i\|^2 \\ &= (1+\theta)\|U\|_{\mathbf{r}}^2 + \left(1+\frac{1}{\theta}\right)\|V\|_{\mathbf{r}}^2.\end{aligned}$$

This completes the proof of the lemma. ∎

Proof of Lemma 4.4: Since $\log(1-x) \le -x$ and $\frac{a}{\tau^\delta}$ is a decreasing function of τ, for $0 < \delta < 1$ we have

$$\begin{aligned}\log \prod_{k=s}^{t-1}\left(1-\frac{a}{k^\delta}\right) &= \sum_{k=s}^{t-1}\log\left(1-\frac{a}{k^\delta}\right)\\ &\le -\sum_{k=s}^{t-1}\frac{a}{k^\delta} \le -\int_s^t \frac{a}{\tau^\delta}d\tau\\ &= -\frac{a}{1-\delta}\left(t^{1-\delta}-s^{1-\delta}\right).\end{aligned}$$

Thus, we have $\prod_{k=s}^{t-1}\left(1-\frac{a}{k^\delta}\right) \le \exp\left(-\frac{a}{1-\delta}\left(t^{1-\delta}-s^{1-\delta}\right)\right)$. Using a similar argument for $\delta = 1$, we can write

$$\begin{aligned}\log \prod_{k=s}^{t-1}\left(1-\frac{a}{k}\right) &= \sum_{k=s}^{t-1}\log\left(1-\frac{a}{k}\right)\\ &\le -\sum_{k=s}^{t-1}\frac{a}{k} \le -\int_s^t \frac{a}{\tau}d\tau\\ &= -a\ln\left(\frac{t}{s}\right).\end{aligned}$$

This implies that $\prod_{k=s}^{t-1}\left(1-\frac{a}{k}\right) \le \exp\left(-a\ln\left(\frac{t}{s}\right)\right) = \left(\frac{t}{s}\right)^{-a}$. This completes the proof of the lemma. ∎

Proof of Lemma 4.5: In order to prove (4.16), we define $p = \sum_{s=1}^{t-1}\beta(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))$. Then, we have

$$\begin{aligned}\lambda p &= \sum_{s=1}^{t-1}\lambda\beta(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\\ &= \sum_{s=1}^{t-1}(1-(1-\lambda\beta(s)))\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\\ &= \sum_{s=1}^{t-1}\left[\prod_{k=s+1}^{t-1}(1-\lambda\beta(k)) - \prod_{k=s}^{t-1}(1-\lambda\beta(k))\right].\end{aligned}$$

Noticing that the last sum is a telescopic sum implies

$$\begin{aligned}\lambda p &= \prod_{k=t}^{t-1}(1-\lambda\beta(k)) - \prod_{k=1}^{t-1}(1-\lambda\beta(k))\\ &= 1-\prod_{k=1}^{t-1}(1-\lambda\beta(k)).\end{aligned}$$

Dividing both sides by $\lambda \neq 0$ arrives at (4.16). ∎

Proof of Lemma 4.6: First, note from Lemma 4.4 that

$$\prod_{k=s}^{t-1}\left(1-\frac{a}{k^{\delta}}\right)\leq \exp\left(-\frac{a}{1-\delta}\left(t^{1-\delta}-s^{1-\delta}\right)\right). \tag{4.166}$$

Therefore, we get

$$\begin{aligned}
&\sum_{s=1}^{t-1}\left[\frac{1}{s^{\sigma}}\prod_{k=s+1}^{t-1}\left(1-\frac{a}{k^{\delta}}\right)\right]\\
&=\sum_{s=2}^{t}\left[\left(\frac{s}{s-1}\right)^{\sigma}\frac{1}{s^{\sigma}}\prod_{k=s}^{t-1}\left(1-\frac{a}{k^{\delta}}\right)\right]\\
&\leq 2^{\sigma}\left[\sum_{s=2}^{t-1}\left[\frac{1}{s^{\sigma}}\prod_{k=s}^{t-1}\left(1-\frac{a}{k^{\delta}}\right)\right]+\frac{1}{t^{\sigma}}\right]\\
&\leq 2^{\sigma}\left[\sum_{s=2}^{t-1}s^{-\sigma}\exp\left(-\frac{a}{1-\delta}\left(t^{1-\delta}-s^{1-\delta}\right)\right)+t^{-\sigma}\right]\\
&=2^{\sigma}\left[e^{-\frac{a}{1-\delta}t^{1-\delta}}\sum_{s=2}^{t-1}s^{-\sigma}\exp\left(\frac{a}{1-\delta}s^{1-\delta}\right)+t^{-\sigma}\right].
\end{aligned} \tag{4.167}$$

Now, consider function $h(\tau)=\tau^{-\sigma}\exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right)$, with

$$\frac{dh(\tau)}{d\tau}=\left(a\tau^{-\sigma}\tau^{-\delta}-\sigma\tau^{-\sigma-1}\right)\exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right). \tag{4.168}$$

Let $t_0=\lceil(\sigma/a)^{\frac{1}{1-\delta}}\rceil$. Then the function $h(\tau)$ is a decreasing function for $\tau\leq t_0-1$ and an increasing function for $\tau\geq t_0$. Therefore, we can write

$$\sum_{s=2}^{t_0-1}s^{-\sigma}\exp\left(\frac{a}{1-\delta}s^{1-\delta}\right)\leq\int_{1}^{t_0-1}\tau^{-\sigma}e^{\frac{a}{1-\delta}\tau^{1-\delta}}d\tau, \tag{4.169}$$

$$\sum_{s=t_0}^{t-1}s^{-\sigma}\exp\left(\frac{a}{1-\delta}s^{1-\delta}\right)\leq\int_{t_0}^{t}\tau^{-\sigma}e^{\frac{a}{1-\delta}\tau^{1-\delta}}d\tau. \tag{4.170}$$

Summing up inequalities in (4.169) and (4.170), we arrive at

$$\sum_{s=2}^{t-1}s^{-\sigma}\exp\left(\frac{a}{1-\delta}s^{1-\delta}\right)\leq\int_{1}^{t}\tau^{-\sigma}e^{\frac{a}{1-\delta}\tau^{1-\delta}}d\tau. \tag{4.171}$$

Substituting this into (4.167), we get

$$\sum_{s=2}^{t-1}\left[\frac{1}{s^{\sigma}}\prod_{k=s}^{t-1}\left(1-\frac{a}{k^{\delta}}\right)\right]\leq\frac{\int_{1}^{t}\tau^{-\sigma}\exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right)d\tau}{\exp\left(\frac{a}{1-\delta}t^{1-\delta}\right)}. \tag{4.172}$$

Now, let us define $p(t) \coloneqq \int_1^t \tau^{-\sigma} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) d\tau$ and $q(t) \coloneqq Ct^{-(\sigma-\delta)} \exp\left(\frac{a}{1-\delta}t^{1-\delta}\right)$ for some constant $C > 0$ (independent of t). In the following, we will show that $p(t) \leq q(t)$ for $t \geq \tau \coloneqq \max\left\{1, \left(\frac{2(\sigma-\delta)}{a}\right)^{\frac{1}{1-\delta}}\right\}$ and a proper choice of C. To this end, we show that $p'(\tau) \leq q'(\tau)$ and $p(\tau) \leq q(\tau)$. First note that

$$\begin{aligned}
p'(t) &= t^{-\sigma} \exp\left(\frac{a}{1-\delta}t^{1-\delta}\right), \\
q'(t) &= C\left(at^{-\sigma} - (\sigma-\delta)t^{-(\sigma-\delta)-1}\right) \exp\left(\frac{a}{1-\delta}t^{1-\delta}\right) \\
&= aC\left(1 - \frac{(\sigma-\delta)}{a}t^{-(1-\delta)}\right) t^{-\sigma} \exp\left(\frac{a}{1-\delta}t^{1-\delta}\right).
\end{aligned}$$

Hence, for $t \geq \tau$ we have

$$\begin{aligned}
q'(t) &= aC\left(1 - \frac{\sigma-\delta}{a}t^{-(1-\delta)}\right) t^{-\sigma} \exp\left(\frac{a}{1-\delta}t^{1-\delta}\right) \\
&\geq aC\left(1 - \frac{\sigma-\delta}{a}\left(\frac{a}{2(\sigma-\delta)}\right)\right) t^{-\sigma} \exp\left(\frac{a}{1-\delta}t^{1-\delta}\right) \\
&= \frac{a}{2}Ct^{-\sigma} \exp\left(\frac{a}{1-\delta}t^{1-\delta}\right),
\end{aligned}$$

which is greater than or equal to $p'(t)$ provided that $C \geq \frac{2}{a}$. It only remains to determine C such that show that $p(\tau) \leq q(\tau)$. First note that if $\tau = 1$, then $p(\tau) = 0 \leq q(\tau)$ for any $C \geq 0$. We will prove the claim for $\sigma > 1$, $\sigma = 1$, and $0 < \sigma < 1$, separately, for the case of $\tau \coloneqq \left(\frac{2(\sigma-\delta)}{a}\right)^{\frac{1}{1-\delta}}$. When $\sigma > 1$, we have

$$\begin{aligned}
p(\tau) &= \int_1^\tau \tau^{-\sigma} e^{\frac{a}{1-\delta}\tau^{1-\delta}} d\tau \leq \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \int_1^\tau \tau^{-\sigma} d\tau \\
&= \frac{1-\tau^{1-\sigma}}{\sigma-1} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \\
&< \frac{1}{\sigma-1} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \\
&= \frac{1}{\sigma-1}\tau^{\sigma-\delta}\tau^{-(\sigma-\delta)} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \\
&= \frac{1}{\sigma-1}\left(\frac{2(\sigma-\delta)}{a}\right)^{\frac{\sigma-\delta}{1-\delta}} \tau^{-(\sigma-\delta)} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \\
&\leq q(\tau),
\end{aligned}$$

where the last inequality holds for $C \geq \frac{1}{\sigma-1}\left(\frac{2(\sigma-\delta)}{a}\right)^{\frac{\sigma-\delta}{1-\delta}}$. For the case of $\sigma = 1$, we have

$$\begin{aligned}
p(\tau) &= \int_1^\tau \tau^{-1} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) d\tau \\
&\leq \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \int_1^\tau \tau^{-1} d\tau \\
&= \ln(\tau) \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \\
&= \ln(\tau)\tau^{1-\delta}\tau^{-(1-\delta)} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \\
&= \frac{2}{a}\ln\left(\frac{2(1-\delta)}{a}\right)\tau^{-(1-\delta)} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \leq q(\tau),
\end{aligned}$$

where the last inequality holds provided that $C \geq \frac{2}{a}\ln\left(\frac{2(1-\delta)}{a}\right)$. Lastly, for the case of $0 < \sigma < 1$, we can write

$$\begin{aligned}
p(\tau) &= \int_1^\tau \tau^{-\sigma} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) d\tau \\
&\leq \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \int_1^\tau \tau^{-\sigma} d\tau \\
&= \frac{\tau^{1-\sigma}-1}{1-\sigma} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \\
&< \frac{\tau^{1-\delta}}{1-\sigma}\tau^{-(\sigma-\delta)} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \\
&= \frac{2(\sigma-\delta)}{a(1-\sigma)}\tau^{-(\sigma-\delta)} \exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right) \leq q(\tau),
\end{aligned}$$

where the last inequality holds for $C \geq \frac{2(\sigma-\delta)}{a(1-\sigma)}$. Therefore, we have $p(t) \leq q(t)$ for $t \geq \tau$ where C is given by

$$C = \begin{cases} \max\left\{\frac{2}{a}, \frac{1}{\sigma-1}\left(\frac{2(\sigma-\delta)}{a}\right)^{\frac{\sigma-\delta}{1-\delta}}\right\} & \text{if } \sigma > 1, \\ \max\left\{\frac{2}{a}, \frac{2}{a}\ln\left(\frac{2(1-\delta)}{a}\right)\right\} & \text{if } \sigma = 1, \\ \max\left\{\frac{2}{a}, \frac{2(\sigma-\delta)}{a(1-\sigma)}\right\} & \text{if } 0 < \sigma < 1. \end{cases}$$

Plugging this into (4.172), we get

$$\begin{aligned}\sum_{s=2}^{t-1}\left[\frac{1}{s^\sigma}\prod_{k=s}^{t-1}\left(1-\frac{a}{k^\delta}\right)\right] &\le \frac{\int_1^t \tau^{-\sigma}\exp\left(\frac{a}{1-\delta}\tau^{1-\delta}\right)d\tau}{\exp\left(\frac{a}{1-\delta}t^{1-\delta}\right)}\\ &= \frac{p(t)}{\exp\left(\frac{a}{1-\delta}t^{1-\delta}\right)}\\ &\le \frac{q(t)}{\exp\left(\frac{a}{1-\delta}t^{1-\delta}\right)} = Ct^{-(\sigma-\delta)},\end{aligned}$$

for $t \ge \tau$. Then, continuing from (4.167), we get

$$\begin{aligned}&\sum_{s=1}^{t-1}\left[\frac{1}{s^\sigma}\prod_{k=s+1}^{t-1}\left(1-\frac{a}{k^\delta}\right)\right]\\ &\le 2^\sigma\left[\sum_{s=2}^{t-1}\left[\frac{1}{s^\sigma}\prod_{k=s}^{t-1}\left(1-\frac{a}{k^\delta}\right)\right]+\frac{1}{t^\sigma}\right]\\ &\le 2^\sigma\left[Ct^{-(\sigma-\delta)}+t^{-\sigma}\right]\\ &\le 2^\sigma\left[C+t^{-\delta}\right]t^{-(\sigma-\delta)}\\ &\le 2^\sigma[C+1]t^{-(\sigma-\delta)}.\end{aligned} \tag{4.173}$$

Therefore, $A(a,\sigma,\delta)$ in the statement of the lemma is determined to be $2^\sigma(C+1)$. Finally, let us consider the case of $\delta = 1$. Similar to (4.167), and using Lemma 4.4 for $a-\sigma+1 \ne 0$, we have

$$\begin{aligned}\sum_{s=1}^{t-1}\left[\frac{1}{s^\sigma}\prod_{k=s+1}^{t-1}\left(1-\frac{a}{k}\right)\right] &\le 2^\sigma\left[\sum_{s=2}^{t-1}\left[\frac{1}{s^\sigma}\prod_{k=s}^{t-1}\left(1-\frac{a}{k}\right)\right]+\frac{1}{t^\sigma}\right]\\ &\le 2^\sigma\left[\sum_{s=2}^{t-1}\frac{1}{s^\sigma}\left(\frac{t}{s}\right)^{-a}+\frac{1}{t^\sigma}\right]\\ &= 2^\sigma t^{-a}\sum_{s=2}^{t-1}s^{a-\sigma}+2^\sigma t^{-\sigma}\\ &\le 2^\sigma t^{-a}\int_1^t \tau^{a-\sigma}d\tau+2^\sigma t^{-\sigma}\\ &= 2^\sigma t^{-a}\left|\frac{t^{a-\sigma+1}-1}{a-\sigma+1}\right|+2^\sigma t^{-\sigma}\\ &= 2^\sigma\left|\frac{t^{-(\sigma-1)}-t^{-a}}{a-\sigma+1}\right|+2^\sigma t^{-\sigma}\\ &\le 2^\sigma\left(\frac{1}{|a-\sigma+1|}+1\right)t^{-\min(\sigma-1,a)}.\end{aligned}$$

This completes the proof of the lemma. ∎

Proof of Lemma 4.7: First let $b \neq 1$. Then, the claimed inequality is equivalent to

$$\frac{a}{t^{c+b}} \geq \frac{1}{t^c} - \frac{1}{(t+1)^c}.$$

The mean value theorem for $h(t) = 1/t^c$ implies that

$$\frac{h(t+1) - h(t)}{(t+1) - t} = h'(z) = -\frac{c}{z^{c+1}}$$

for some $z \in (t, t+1)$. Therefore, we have

$$\frac{1}{t^c} - \frac{1}{(t+1)^c} = \frac{c}{z^{c+1}} \leq \frac{c}{t^{c+1}} = \left(\frac{c/a}{t^{1-b}}\right)\frac{a}{t^{c+b}} \leq \frac{a}{t^{c+b}}, \tag{4.174}$$

where the first inequality follows from $z \geq t$ and the second inequality holds for $t \geq \left(\frac{c}{a}\right)^{\frac{1}{1-b}}$.

Finally, if $b = 1$ and $a \geq c$, then we have $c/t^{c+1} \leq a/t^{c+b}$, and the inequality in (4.174) holds for all values of $t \geq 1$. This completes the proof of the lemma. ∎

Proof of Lemma 4.11: We first define a sequence $g(t)$ via

$$g(t+1) = (1 - q(t))g(t) + p(t), \tag{4.175}$$

for $t \geq 1$ and $g(1) = 0$. Then, we can verify that

$$g(t) = \sum_{s=1}^{t-1}\left[p(s)\prod_{k=s+1}^{t-1}(1 - q(k))\right].$$

for all values of $t \geq 1$. We set $S = \max\left\{\frac{g(t_0)q(t_0)}{p(t_0)}, \frac{1}{1-A}\right\}$, and we aim to show that $g(t) \leq Sp(t)/q(t)$ for every $t \geq t_0$.

We use induction to prove the claim. First note that for $t = t_0$, we have $g(t_0) \leq Sp(t_0)/q(t_0)$. Assume the claim holds for t. Then, for $t+1$ we have

$$\begin{aligned}
\sum_{s=1}^{t}\left[p(s)\prod_{k=s+1}^{t}(1 - q(k))\right] &= g(t+1) \\
&= (1 - q(t))g(t) + p(t) \\
&\leq (1 - q(t))S\frac{p(t)}{q(t)} + p(t) \\
&= S\frac{p(t)}{q(t)} - (S-1)p(t).
\end{aligned}$$

Thus, in order to show that $g(t+1) \leq S\frac{p(t+1)}{q(t+1)}$, it suffice to show that

$$\frac{p(t)}{q(t)} - \frac{p(t+1)}{q(t+1)} \leq \frac{S-1}{S}p(t).$$

To this end, we can write

$$\begin{aligned}
\frac{p(t)}{q(t)} - \frac{p(t+1)}{q(t+1)} &= \frac{p(t)}{q(t)} - \frac{p(t+1)}{q(t)} + \frac{p(t+1)}{q(t)} - \frac{p(t+1)}{q(t+1)} \\
&= \frac{-\Delta p(t)}{q(t)} + p(t+1)\frac{q(t+1)-q(t)}{q(t)q(t+1)} \\
&\leq \frac{-\Delta p(t)}{q(t)} \\
&\leq \frac{Ap(t)q(t)}{q(t)} \\
&\leq \frac{S-1}{S}p(t),
\end{aligned}$$

where the first inequality holds since $\{q(t)\}$ is a non-increasing sequence, the second inequality follows from (4.21), and the last inequality holds since $S \geq \frac{1}{1-A}$. This completes the proof of the lemma. ■

Chapter 5

DNA-DP: Decentralized Nesterov Acceleration with Differential Privacy

We study a decentralized learning scenario involving a group of computing agents aiming to collaboratively solve a convex optimization problem. Our contribution lies in proposing an accelerated distributed Nesterov gradient descent method integrated with differential privacy. In our algorithm, the local variables are perturbed by a Laplace noise before being shared with neighboring agents. We choose the parameter of the injected noises to (1) guarantee differential privacy against an adversary who observes all the exchanged data and (2) guarantee convergence to the minimizer of the objective function. Subsequently, we conduct an analysis of the convergence behavior of the algorithm and show that the proposed algorithm converges to the optimal solution at the rate of $\mathcal{O}(T^{-1/2+\eta})$ for any $\eta > 0$. The proposed algorithm outperforms the benchmarks of the decentralized learning with differential privacy under the same privacy budget. Our simulation results corroborate the findings of this chapter.

5.1 Introduction

Decentralized learning serves as a learning paradigm in which a group of computing agents collaboratively engages in solving an optimization problem. In the absence of a central server, the learning process relies entirely on on-device computation and local communication among neighboring agents. Decentralized learning is emerging as a core in various applications due to its ability to scale to larger datasets and systems, data locality, ownership, and privacy. Decentralized learning arises in various applications such as sensor networks [74, 75], network routing [76], large scale machine learning [77], power control [78], and distributed network resource allocations [79, 80], for which decentralized solutions offer promising result.

In decentralized learning, safeguarding sensitive information from disclosure may involve protecting data sample points, gradients, or model vectors. We note that in this setup, data points are not directly shared, and each agent sends information about its local model variables. In machine learning, it has been shown that shared gradients can be used by an adversary to recover the data sample points used for training reversely (pixel-wise accurate for images and token-wise matching for texts) [116]. Thus, the models could carry sensitive information abstracted from raw data. However, note that the objective of decentralized optimization is for individual agents to learn the same optimal model, and hence, the final consensual optimization variable should be disclosed to all agents and not be a target of privacy protection.

Differential privacy (DP) plays a crucial role in decentralized learning, ensuring the protection of individual privacy while extracting valuable insights from distributed datasets. This approach allows computing agents to collaborate in training learning models without compromising the confidentiality of their data. DP stochastic gradient descent [15] is particularly important for enabling private empirical risk minimization (ERM) of machine learning models. Various works have analyzed the convergence behavior of DP ERM methods, including DP-SGD [117–120].

For strongly convex and smooth objection functions, both accelerated and non-accelerated gradient-based methods achieve a linear convergence rate, i.e., $\mathcal{O}(\exp(-\alpha T))$ for some $\alpha > 0$ [121].

Such a fascinating convergence rate disappears when the objective function is only

convex. In particular, the gradient descent method converges to the optimum solution at the rate of $\mathcal{O}(T^{-1}\ln T)$ [57]. However, this rate can be significantly improved using the accelerated Nesterov gradient descent method, as the rate of $\mathcal{O}(T^{-1.4+\eta})$ for any $\eta \in (0, 1.4)$ is shown to be achievable [121]. When privacy constraints are imposed, we have to perturb the exchanged information between the workers in various stages of the algorithm to avoid leakage of the information. As a consequence, convergence to the optimum solution is more challenging when the algorithm is performed over noisy data. A differentially-private decentralized learning algorithm is proposed and studied in [122], where it is shown to converge to the optimal solution at the rate of $\mathcal{O}(T^{-1/3})$ for the convex setting.

In this work, we consider the decentralized learning paradigm where data is distributed among various computing nodes and develop a robust distributed optimization method against adversaries. Our main contributions are:

- We propose a novel accelerated distributed Nesterov gradient descent method with differential privacy. Each agent adds a well-designed level of Laplace noise to information before exchanging with neighboring agents to ensure DP.

- We provide privacy guarantees where the parameters for the Laplace mechanism[1] are characterized for a pre-defined level of privacy budget. Then, we analyze the convergence rate for the proposed method for the convex setting. We show that our algorithm converges to the optimal loss at the rate of $\mathcal{O}(T^{-1/2+\eta})$ for any $\eta > 0$.

- We conduct numerical experiments on widely used datasets such as MNIST and Fashion-MNIST for various privacy budgets. Moreover, we compare our algorithm with two benchmarks under the same privacy leakage. These experimental results support our theoretical findings.

Related Work.

Decentralized learning and optimization. The literature has extensively studied the decentralized/distributed optimization framework for various classes of objective functions, including non-convex, convex, and strongly-convex settings. Most

[1] We note that the analysis can be easily extended to the Gaussian mechanism.

popular approaches for the convex setting are distributed gradient descent-type methods [57, 78, 83, 92, 123–127], distributed variants of the alternating direction method of multipliers (ADMM) [81, 82, 128–131], augmented Lagrangian algorithms [132–134], dual averaging [77, 85].

Differential privacy for distributed learning. Decentralized learning, by nature, consists of the exchange of messages between various parts of the networks, which can potentially lead to information leakage to an external observer. To protect the privacy of data, DP essentially proposes to apply a perturbation on the data before revealing it to the adversary. DP is widely studied in the presence of a central server, referred to as federated learning, [135–139]. Focusing on the client's perspective, an algorithm is proposed in [140] to hide clients' contributions during training, balancing the trade-off between privacy loss and model performance. Moreover, it is shown how noise injection can be tailored at the client level to achieve differential privacy guarantees.

Further, DP has also been exploited in decentralized learning by introducing independent additive noises [141–143], for time-varying objective functions [144–146]. To improve the accuracy of the model, a gradient tracking-based approach is developed in [147]. However, to achieve privacy, a very strong assumption is considered on the gradient function, that is, all the functions in the adjacent class of functions to be kept private (see Definition 5.1) have identical gradients!

Gradient methods for differentially-private distributed optimization are studied for strongly convex setting in [148] and convex setting in [149]. Using a *geometrically* vanishing step-size sequence, the algorithm in [148] converges to a *neighborhood* of the optimal solution. In an almost sure sense, the algorithm proposed in [149] converges to an optimal solution. However, for the privacy analysis, a very restrictive assumption about the behavior of the function on a neighborhood of the optimizers is considered: the gradient of the true function and all its adjacent functions (see Definition 5.1) in a neighborhood of the optimum solution are constant.

5.2 Problem Setup

Notation. We denote the set of integers $\{1, 2, \ldots, n\}$ by $[n]$. In this chapter, we are considering n agents that collaborate to minimize a function $f : \mathbb{R}^{1\times d} \to \mathbb{R}$. Hence, we

view vectors in $\mathbb{R}^{1\times d} = \mathbb{R}^d$ as row vectors. The remaining vectors, i.e., the vectors in $\mathbb{R}^{n\times 1} = \mathbb{R}^n$, are assumed to be column vectors. We denote the ℓ_2 norm of a vector $\mathbf{x} \in \mathbb{R}^d$ by $\|\mathbf{x}\| = \left(\sum_{i=1}^d |x_i|^2\right)^{1/2}$ and its ℓ_1 norm by $\|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i|$. For a matrix $A \in \mathbb{R}^{n\times d}$, we use $\|A\| = \left(\sum_{i=1}^n \sum_{j=1}^d |A_{ij}|^2\right)^{1/2}$ to refer to its Frobenius norm, and $\|A\|_2 = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$ shows the spectral norm for A.

This chapter is motivated by stochastic learning problems in which the goal is to solve

$$\min_{\mathbf{x}} L(\mathbf{x}) := \min_{\mathbf{x}} \mathbb{E}_{\xi\sim\mathcal{P}}[\ell(\mathbf{x},\xi)], \tag{5.1}$$

where $\ell : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ is a loss function, $\mathbf{x} \in \mathbb{R}^{1\times d} = \mathbb{R}^d$ is the model row vector, and ξ is a random vector taking values in $\mathbb{R}^p$ that is drawn from an unknown underlying distribution $\mathcal{P}$. The inherent challenge in (5.1) arises from the practical scenario where the underlying distribution $\mathcal{P}$ is often unknown. Instead, we have access to $N = mn$ independent realizations of ξ and focus on solving the corresponding ERM problem, defined as

$$\min_{\mathbf{x}} f(\mathbf{x}) := \min_{\mathbf{x}} \frac{1}{N} \sum_{j=1}^{N} \ell\left(\mathbf{x}, \xi_j\right), \tag{5.2}$$

where $f(\mathbf{x})$ is the empirical risk with respect to the data points $\mathcal{D} = \{\xi_1, \ldots, \xi_N\}$.

In decentralized learning, we consider a set of $n \geq 2$ agents that are connected through a fixed network. Each agent $i \in [n]$ observes a subset of m data points, denoted by $\mathcal{D}_i = \{\xi_1^i, \ldots, \xi_m^i\}$, where $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_n$. Thus, the ERM problem in (5.2) can be written as the minimization of the average of local empirical risk functions f_i for all nodes $i \in [n]$ in the network, i.e.,

$$\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\mathbf{x}} \frac{1}{N} \sum_{i=1}^{n} \sum_{\xi\in\mathcal{D}_i} \ell(\mathbf{x},\xi) = \min_{\mathbf{x}} \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}), \tag{5.3}$$

where $f_i(\mathbf{x}) := \frac{1}{m}\sum_{\xi\in\mathcal{D}_i} \ell(\mathbf{x},\xi) = \frac{1}{m}\sum_{j=1}^m \ell(\mathbf{x},\xi_j^i)$. We can rewrite the ERM problem in (5.3) as a distributed consensus optimization problem, given by

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_n\in\mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}_i) \quad \text{s.t.} \quad \mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n. \tag{5.4}$$

Information Exchange. In decentralized learning, each agent i is trying to minimize its own objective function f_i as in (5.4). However, the constraint $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n$ needs

to be satisfied, i.e., the agents should reach a consensus and agree on a global minimizer rather than individual minimizers. The key challenge here is to achieve a consensus by only exchanging information with neighboring nodes on the underlying network.

An algorithm can solve the problem in (5.4) if it ensures that the node's models (i.e., $\mathbf{x}_i(t)$) remain close to each other and eventually converge the same value, which is the minimizer of $f(\mathbf{x})$. Here, we assume a fixed network topology among the nodes, which is represented by an undirected connected graph $\mathcal{G} := ([n], \mathcal{E})$. The vertex set $[n]$ represents the set of agents, and the edge set $\mathcal{E} \subseteq \{(i,j) : i,j \in [n]\}$ represents the set of links between the agents. We denote the set of neighbors of node i by $\mathcal{N}_i := \{j \in [n] : (i,j) \in \mathcal{E}\}$. Each agent i can only communicate with its neighbors, i.e., $j \in \mathcal{N}_i$. We consider a symmetric doubly-stochastic matrix $W = [w_{ij}]$, which is consistent with the underlying network $\mathcal{G}$, that is, $w_{ij} > 0$ if only if $(j,i) \in \mathcal{E}$. The coefficient $w_{i,j}$ is used to denote the weight associated to the information of agent j when averaged at agent i. As mentioned before, at each iteration t, each node j broadcasts its local data, say some $\boldsymbol{d}_j(t)$ to all its neighbors, i.e., i satisfying $(j,i) \in \mathcal{E}$. Then, node i can evaluate $\sum_{j=1}^{n} w_{ij}\boldsymbol{d}_j = \sum_{j\in\mathcal{N}_i} w_{ij}\boldsymbol{d}_j(t)$, as the weighted average of the data available at its neighbors.

Privacy Consideration. In decentralized learning, each agent shares its local model variables with neighboring agents to achieve a consensus. It is shown that the exchanged intermediate variables may reveal information about the agents' private data through model inversion attacks [150] and deep leakage from gradient [151]. Hence, we need to apply privacy-preserving methods to protect the privacy of data. We consider two potential attacks in decentralized learning, which are the two most common attacks in privacy research [135, 143, 152, 153]:

- *Honest-but-curious attacks* involve each computing node following all protocols for the training phase but being curious and collecting all transmitted intermediate data from the neighboring nodes to learn the sensitive information about the private data of other agents;

- *Eavesdropping attacks* in which an external eavesdropper has access to all communication links to intercept exchanged messages to learn sensitive information about the private data of agents.

Note that each computing agent has access to some local model variables which are unavailable to external eavesdroppers. In contrast, an eavesdropper has access to all shared information in the network, whereas each agent gets the shared information only from its neighboring nodes. There is no trusted third party. Hence, to guarantee privacy against both types of attacks, a privacy-preserving mechanism should be applied locally by each agent.

To this end, we formally define our privacy model using the notion of differential privacy. We represent the distributed optimization problem in (5.4) by a tuple $\mathcal{P} = (\mathbb{R}^d, \mathcal{D}, \mathcal{G})$, where $\mathbb{R}^d$ is the domain of the function f which we aim to minimize, $\mathcal{D}$ is the dataset that determines the objective function $f(\mathbf{x}) = \frac{1}{|\mathcal{D}|}\sum_{\xi\in\mathcal{D}} \ell(\mathbf{x}, \xi)$, and $\mathcal{G}$ is the underlying graph. Then, we adopt the following definition from [143].

Definition 5.1. Two distributed optimization problems $\mathcal{P} = (\mathbb{R}^d, \mathcal{D}, \mathcal{G})$ and $\mathcal{P}' = (\mathbb{R}^d, \mathcal{D}', \mathcal{G}')$ are called adjacent, if $\mathcal{G} = \mathcal{G}'$ and, there exists an $i \in [n]$ such that $\mathcal{D}_i$ and $\mathcal{D}'_i$ differ in exactly one data point, i.e., $\xi^i_m \neq \xi'^i_m$, but $\xi^i_k = \xi'^i_k$ for $k \neq m$. Moreover, we have $\mathcal{D}_j = \mathcal{D}'_j$ for every $j \in [n] \setminus \{i\}$.

In other words, two optimization problems are adjacent if and only if for one agent i, one of the data points in $\mathcal{P}$ and $\mathcal{P}'$ are different, but all other settings and parameters are identical. Note that this yields $f_i \neq f'_i$, but $f_j = f'_j$ for every $j \in [n] \setminus \{i\}$. Assume that we apply a decentralized algorithm $\mathcal{A}$ on two adjacent optimization problems with $\mathcal{D}$ and $\mathcal{D}'$. We denote the set of all variables that are generated and exchanged over the network throughout the course of execution of $\mathcal{A}$ by $\mathcal{A}(\mathcal{D})$ and $\mathcal{A}(\mathcal{D}')$, respectively. Next, we provide the definition of a private mechanism.

Definition 5.2. A randomized mechanism $\mathcal{A}$ is called ϵ-differentially private if, for any two adjacent problems with datasets $\mathcal{D}$ and $\mathcal{D}'$ we have

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{O}] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{O}]$$

for any measurable set $\mathcal{O}$ of possible generated variables.

5.3 The Proposed DNA-DP Algorithm

An accelerated distributed Nesterov gradient descent algorithm is proposed in [121] for minimizing a convex function without any privacy constraint. It is shown that the

proposed algorithm can converge to the optimum loss as fast as $\mathcal{O}(T^{-1.4+\eta})$ for any $\eta > 0$. However, an external adversary can potentially learn the local function and the private data by observing the exact information exchanged between the agents. Inspired by this algorithm, we propose an accelerated distributed Nesterov gradient descent method with differential privacy, which we refer to as *D*ecentralized *N*esterov *A*cceleration with *D*ifferential *P*rivacy (DNA-DP).

The DNA-DP consists of iterative updates of four local variables, namely, $\mathbf{y}_i(t)$, $\mathbf{g}_i(t)$, $\mathbf{v}_i(t)$, and $\mathbf{x}_i(t)$. Each agent updates its local variables over the iterations of the algorithm. At each iteration t, each node exchanges $\mathbf{y}_i(t)$, $\mathbf{g}_i(t)$, and $\boldsymbol{v}_i(t)$ with its neighbors. However, before transmission, node i perturbs its local variables and sends

$$\hat{\mathbf{y}}_i(t) = \mathbf{y}_i(t) + \mathbf{e}_{i,y}(t), \tag{5.5a}$$

$$\hat{\mathbf{g}}_i(t) = \mathbf{g}_i(t) + \mathbf{e}_{i,g}(t), \tag{5.5b}$$

$$\hat{\boldsymbol{v}}_i(t) = \boldsymbol{v}_i(t) + \mathbf{e}_{i,v}(t) \tag{5.5c}$$

to its neighbors. Here, $\mathbf{e}_{i,y}(t)$, $\mathbf{e}_{i,g}(t)$, and $\mathbf{e}_{i,v}(t)$ are random noise vectors in $\mathbb{R}^d$. In particular, we assume zero-mean Laplace noises with proper variances for each noise sequence which are specified in Assumption 5.1.

The algorithm starts with initial conditions $\mathbf{y}_i(0) = \mathbf{g}_i(0) = \boldsymbol{v}_i(0) = \mathbf{x}_i(0) = \mathbf{0}$ for every $i \in [n]$. At iteration t, given the noisy information received from the neighbors, each node i updates its local variables via

$$\mathbf{g}_i(t+1) = \hat{\mathbf{g}}_i(t) + \gamma \sum_{j \in \mathcal{N}_i} w_{ij} \left(\hat{\mathbf{g}}_j(t) - \hat{\mathbf{g}}_i(t)\right) + \nu(t) \nabla f_i(\mathbf{y}_i(t)), \tag{5.6a}$$

$$\mathbf{x}_i(t+1) = \hat{\mathbf{y}}_i(t) + \gamma \sum_{j \in \mathcal{N}_i} w_{ij} (\hat{\mathbf{y}}_j(t) - \hat{\mathbf{y}}_i(t)) - (\mathbf{g}_i(t+1) - \mathbf{g}_i(t)), \tag{5.6b}$$

$$\boldsymbol{v}_i(t+1) = \hat{\boldsymbol{v}}_i(t) + \gamma \sum_{j \in \mathcal{N}_i} w_{ij} (\hat{\boldsymbol{v}}_j(t) - \hat{\boldsymbol{v}}_i(t)) - \frac{1}{\alpha(t)} (\mathbf{g}_i(t+1) - \mathbf{g}_i(t)), \tag{5.6c}$$

$$\mathbf{y}_i(t+1) = (1 - \alpha(t+1)) \mathbf{x}_i(t+1) + \alpha(t+1) \boldsymbol{v}_i(t+1). \tag{5.6d}$$

Here, $\nu(t) = \frac{\nu_0}{(t+\tau)^\sigma}$ is a vanishing step-size for arbitrary exponent of $\sigma \in (3/2, 2)$ and shift of $\tau \geq 0$, which will be determined later. Moreover, $\alpha(t)$ is another step-size that can be recursively determined from $\nu(t)$ via

$$\alpha^2(t+1) = \frac{\nu(t+1)}{\nu(t)} (1 - \alpha(t+1)) \alpha^2(t), \tag{5.7}$$

for every $t \geq 0$ with initial condition $\alpha(0) = \alpha_0 := \sqrt{\nu_0 K}$. Finally, $\gamma \in (0, 1]$ is an arbitrary mixing parameter.

The description of DNA-DP is summarized in Algorithm 4.

Algorithm 4 DNA-DP at agent i

Input: Stochastic matrix W, Iteration T

Set $\mathbf{x}_i(0) = \boldsymbol{v}_i(0) = \mathbf{y}_i(0) = \mathbf{g}_i(0) = \mathbf{0}$.

for $t = 1$ **to** T **do**

  for $i = 1$ **to** n **do**

    Perturb $\boldsymbol{v}_i(t)$, $\mathbf{y}_i(t)$, and $\mathbf{g}_i(t)$ according to (5.5).

    Send $\hat{\boldsymbol{v}}_i(t)$, $\hat{\mathbf{y}}_i(t)$, $\hat{\mathbf{g}}_i(t)$ to $j \in \mathcal{N}_i$ and receive $\hat{\boldsymbol{v}}_j(t)$, $\hat{\mathbf{y}}_j(t)$, $\hat{\mathbf{g}}_j(t)$ from $j \in \mathcal{N}_i$.

    Compute the local gradient $\nabla f_i(\mathbf{y}_i(t))$.

    Update $\mathbf{g}_i(t+1)$, $\mathbf{x}_i(t+1)$, $\boldsymbol{v}_i(t+1)$, and $\mathbf{y}_i(t+1)$ according to (5.6).

  end for

end for

5.4 Theoretical Results and Analysis

In this section, we provide the theoretical guarantees of the DNA-DP algorithm presented in Section 5.3. To this end, we first present the assumptions on the noise sequences, the stochastic weight matrix, and loss functions, which are needed to guarantee the convergence and privacy of the algorithm. Then, we present the theoretical findings of this work for the privacy guarantees and convergence analysis.

Assumption 5.1. We assume that the noise sequences are drawn independently across the nodes and iterations. Moreover, we assume they are zero mean, i.e.,

$$\mathbb{E}\left[\mathbf{e}_{i,y}(t)|\mathcal{F}_t\right] = \mathbb{E}\left[\mathbf{e}_{i,g}(t)|\mathcal{F}_t\right] = \mathbb{E}\left[\mathbf{e}_{i,v}(t)|\mathcal{F}_t\right] = 0,$$

and their variances are bounded by

$$\mathbb{E}\left[\|\mathbf{e}_{i,y}(t)\|^2|\mathcal{F}_t\right] \leq \zeta_1 \nu^2(t), \quad \mathbb{E}\left[\|\mathbf{e}_{i,g}(t)\|^2|\mathcal{F}_t\right] \leq \zeta_1 \nu^2(t), \quad \mathbb{E}\left[\|\mathbf{e}_{i,v}(t)\|^2|\mathcal{F}_t\right] \leq \zeta_1 \frac{\nu^2(t)}{\alpha^2(t)},$$

for some constant $\zeta_1 > 0$, for every node $i \in [n]$ and every iteration $t \geq 1$. Here $\{\mathcal{F}_t\}$ is the natural filtration for the processes $\{\mathbf{y}(t), \mathbf{g}(t), \boldsymbol{v}(t)\}_{i=1}^n$.

Later, in Theorem 5.1, we show that the noise sequences satisfying Assumption 5.1 are capable of protecting the data points in the differential privacy sense.

Assumption 5.2. We assume the following properties on $\ell(\cdot,\cdot)$ and $f(\cdot)$:

(a) The function $\ell(\cdot,\cdot)$ is convex with respect to its first argument; Consequently, function $f_i(\cdot)$ is convex for every $i \in [n]$;

(b) The function $\ell(\cdot,\cdot)$ has bounded gradient, i.e., there exists a constant ζ_2 so that $\|\nabla \ell(\mathbf{x},\xi)\| \leq \zeta_2$; similarly, we have $\|\nabla f_i(\mathbf{x})\| \leq \zeta_2$ for all $i \in [n]$;

(c) The set of minimizers of $f(\cdot)$ is compact.

In the following, we present the main results of this work. In particular, in Theorem 5.1, we show that the algorithm in (5.6) is differentially private. The proof of Theorem 5.1 is presented in Section 5.8. Next, in Theorem 5.2, we show the convergence of the algorithm to the optimum value. The detailed proof of Theorem 5.1 is presented in Section 5.9.

5.4.1 Differential Privacy

We prove that each iteration of update rules in (5.6) guarantees ϵ-differential privacy. To ensure the differential privacy, we consider the zero-mean Laplace distribution with scale parameter b for the noise sequences, where its probability density function is given by $\frac{1}{2b}e^{-|x|/b}$.

Theorem 5.1. *Given $\epsilon > 0$, assume $\mathbf{e}_{i,y}(t)$, $\mathbf{e}_{i,g}(t)$, and $\mathbf{e}_{i,v}(t)$ be noise sequences sampled from a zero-mean Laplace distribution with scale parameter $b_y(t)$, $b_g(t)$, and $b_v(t)$, respectively where*

$$b_y(t) = \frac{8\zeta_2\sqrt{d}}{m\epsilon}\nu(t-2), \quad b_g(t) = \frac{2\zeta_2\sqrt{d}}{m\epsilon}\nu(t-1), \quad b_v(t) = \frac{4\zeta_2\sqrt{d}}{m\epsilon}\frac{\nu(t-2)}{\alpha(t-1)}. \tag{5.8}$$

Then, Assumption 5.1 holds. Moreover, each iteration of update rules in (5.6) satisfies ϵ-differential privacy. Specifically, for any adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$, any iteration t, any agent i, and any outputs $\hat{\mathbf{g}}_i(t)$, $\hat{\boldsymbol{v}}_i(t)$, and $\hat{\mathbf{y}}_i(t)$, the following inequalities hold

$$\Pr[\hat{\mathbf{y}}_i(t)|\mathcal{D}] \leq e^{\epsilon}\Pr[\hat{\mathbf{y}}_i(t)|\mathcal{D}'], \ \Pr[\hat{\mathbf{g}}_i(t)|\mathcal{D}] \leq e^{\epsilon}\Pr[\hat{\mathbf{g}}_i(t)|\mathcal{D}'], \ \Pr[\hat{\boldsymbol{v}}_i(t)|\mathcal{D}] \leq e^{\epsilon}\Pr[\hat{\boldsymbol{v}}_i(t)|\mathcal{D}'].$$

Remark 5.1. Theorem 5.1 shows that each iteration of the proposed algorithm DNA-DP is ϵ-differentially private. Hence, the total privacy leakage grows at most linearly with the number of iterations T and can be kept limited below any tolerable privacy level. Such analysis is rather standard in the literature [15, 135], where a Gaussian mechanism for differential privacy and overall leakage in T iterations scales as $\mathcal{O}(\sqrt{T})$. In Section 5.5, for a fair comparison with other schemes, we normalize the privacy parameter of the proposed scheme to ensure that the overall leakages of all the algorithms are the same.

5.4.2 Convergence

The next theorem shows that the dynamic in (5.6) converges to the optimum solution of the optimization problem.

Theorem 5.2. *Suppose that Assumptions 3.1, 3.3, 5.1, 5.2 hold. Then, the average model* $\bar{\mathbf{x}}(t) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(t)$ *generated by Algorithm 4 with step-size* $\nu(t) = \frac{\nu_0}{(t+\tau)^\sigma}$ *satisfies*

$$\mathbb{E}\left[f(\bar{\mathbf{x}}(t))\right] - f^\star \leq \frac{4\tau^2\zeta/\nu_0}{(t+\tau)^{2-\sigma}}, \tag{5.9}$$

for any $\sigma \in (\frac{3}{2}, 2)$*, where* ζ *is a constant, which will be determined later.*

Remark 5.2. Theorem 5.2 guarantees the convergence of the loss function for the average model to the optimal value with diminishing step-size. To maximize the convergence rate, we need to choose σ such that the exponent of the dominating term in the upper bound (5.9) is minimized. It can be verified that the optimum choice is $\sigma = 3/2 + \eta$ for any $\eta > 0$, which leads to

$$\mathbb{E}\left[f(\bar{\mathbf{x}}(t))\right] - f^\star \leq \frac{4\tau^2\zeta}{\nu_0(t+\tau)^{1/2-\eta}}, \tag{5.10}$$

This achieves a convergence rate of $\mathcal{O}(T^{-1/2+\eta})$ for any $\eta > 0$.

Note that Theorem 5.2 demonstrates an explicit utility-privacy trade-off for our algorithm; when the privacy constraint is weaker (corresponding to a larger ϵ), we can satisfy privacy by a small choice of ζ (see (5.17) and (5.154), which implies $\zeta \propto 1/\epsilon^2$). Consequently, the gap to optimality in (5.10) gets smaller. We first prove that the proposed algorithm is differentially private.

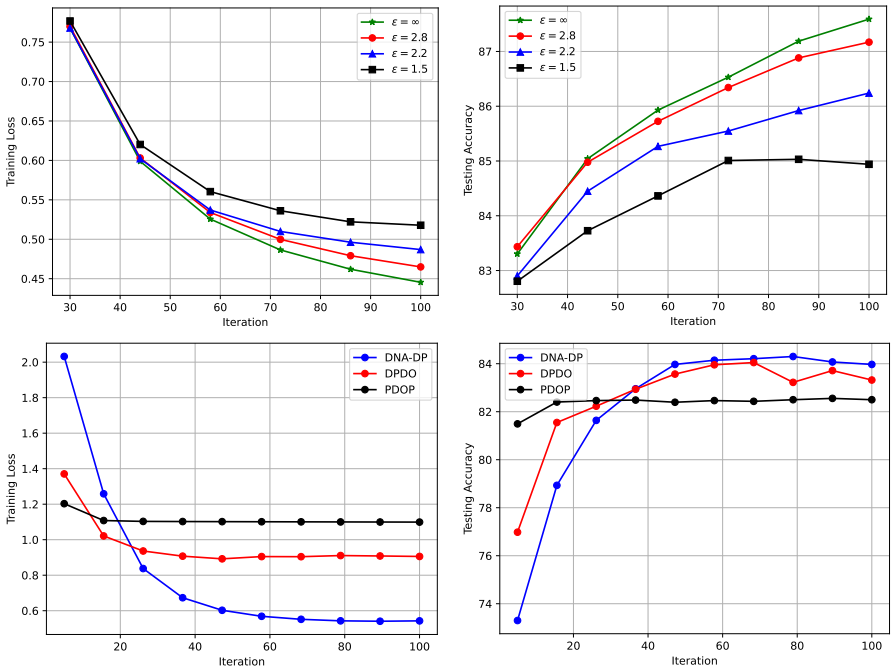

Figure 5.1: Logistic Regression on MNIST

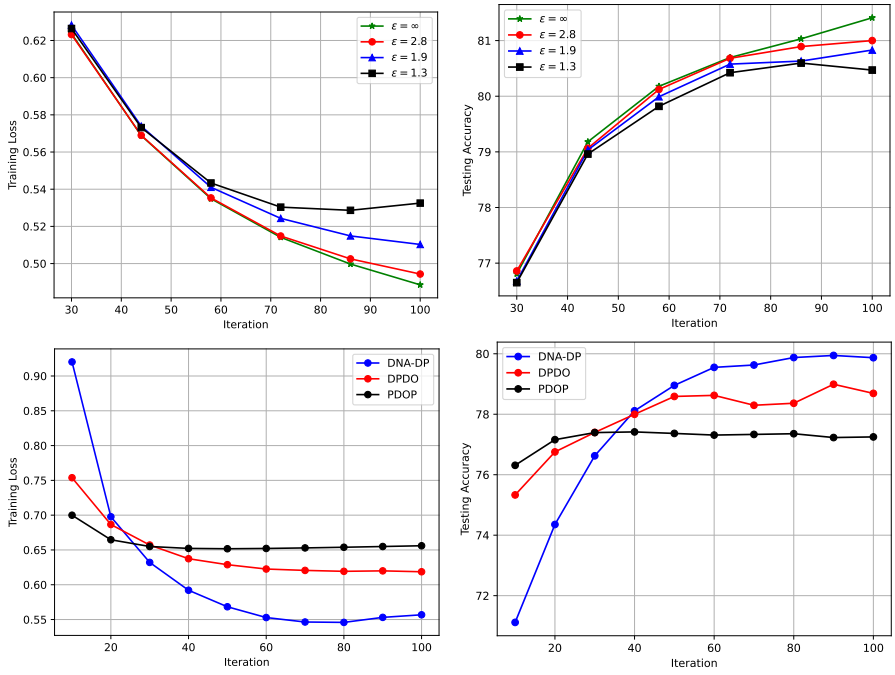

Figure 5.2: Logistic Regression on Fashion-MNIST

5.5 Experimental Results

In this section, we numerically evaluate the performance of the proposed DNA-DP method described in Algorithm 4 for various privacy parameters ϵ. Further, we compare the performance of DNA-DP with two benchmarks under the same privacy budget, briefly described below.

- **Private Distributed Optimization (PDOP)** [148]: Each worker updates its model variables as

$$\begin{aligned}\mathbf{y}_i(t) &= \mathbf{x}_i(t-1) + \mathbf{e}_i(t),\\ \mathbf{z}_i(t) &= \sum_{j=1}^{n} w_{ij}\mathbf{y}_j(t),\\ \mathbf{x}_i(t) &= \mathbf{z}_i(t) - \gamma(t)\left(\nabla f_i\left(\mathbf{z}_i(t)\right)\right),\end{aligned}$$

  where $\mathbf{e}_i(t)$ are random noise vectors drawn from a zero-mean Laplace distribution with the scale parameter $b(t) = c_0 p^{t-1}$, and $\gamma(t) = c_1 q^{t-1}$ for some constants p, q, c_0, and c_1.

- **Differentially Private Distributed Optimization (DPDO)** [149]: Each node updates its model as

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \sum_{j=1}^{n} \gamma(t) w_{ij}\left(\mathbf{x}_j(t) + \mathbf{e}_j(t) - \mathbf{x}_i(t)\right) - \lambda(t)\nabla f_i\left(\mathbf{x}_i(t)\right),$$

  where $\mathbf{e}_i(t)$ are noise sequences sampled from a zero-mean Laplace distribution with some diminishing scale parameter $b(t)$, and step-size learning $\gamma(t)$ and $\lambda(t)$.

Remark 5.3. The algorithm DPOP [148] exploits geometrically decaying step-size parameters to ensure a finite privacy budget. However, such fast-decaying steps lead to poor performance in the training phase. We use polynomial decay rates to ensure the convergence of the local models to the minimizer $\mathbf{x}^\star$ as shown in Theorem 5.2.

Remark 5.4. To show privacy in [149], it is assumed that for two adjacent problems $\mathcal{P}$ and $\mathcal{P}'$, functions f_i and f_i' have similar behaviors around $\mathbf{x}^\star = \arg\min_{\mathbf{x}}\left[f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})\right]$, the minimizer of problem $\mathcal{P}$, i.e., there exits some $\delta > 0$ such that for every $\mathbf{x}, \mathbf{y} \in B_\delta(\mathbf{x}^\star)$ where $B_\delta(\mathbf{x}^\star) := \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^\star\| < \delta\}$, we have $\nabla f_i(\mathbf{x}) = \nabla f_i'(\mathbf{y})$. This assumption could

easily fail in practice. In our work, we prove privacy regardless of the behaviors of f_i and f_i' around the minimizers.

Remark 5.5. As Theorem 5.1 shows parameters $b_g(t)$, $b_v(t)$, and $b_y(t)$ scale with $\zeta_2\sqrt{d}/m$. Hence, in our experiments, we consider the level of privacy ϵ, which is *normalized* to $m/\zeta_2\sqrt{d}$. Moreover, under the same privacy budget, to provide a fair comparison, we normalized the derived scale parameters for the proposed algorithms PDOP and DPDO.

Data and Experimental Setup. We run two sets of experiments over MNIST[2] and Fashion-MNIST[3] datasets, where each worker is assigned a sample set of size $m = 6,000$ for both datasets. The connectivity network is a random Erös-Rènyi graph with edge connectivity $p_c = 0.8$ and $n = 10$ nodes. The matrix is designed as $W = I - 2G/k$ where G is the Laplacian matrix of the graph and $k = 1.2$. We exploit the logistic regression model for the learning with the total number of iterations $T = 100$. For MNIST, we set mixing parameter $\gamma = 0.04$, batch size $B = 100$, fine-tuned step-size parameters $(\nu_0, \alpha_0) = (0.25, 0.05)$, and $\tau = 100$. For Fashion-MNIST, we tune $\gamma = 0.01$, $B = 100$, $(\nu_0, \alpha_0) = (1.1, 0.1)$, and $\tau = 300$. Our codes are implemented using Python programming language and Pytorch package and tested on a 2017 MacBook Pro with 16 GB of memory and a 2.5 GHz Intel Core i7 processor.

Results. In Figure 5.1, the top-left and top-right plots demonstrate the training loss and testing accuracy, respectively, of the average model vs. iteration for Algorithm 4 of the logistic regression model on the MNIST dataset. Moreover, the bottom-left and bottom-right plots compare the performance of our algorithm DNA-DP to two benchmarks, PDOP and DPDO, with $\epsilon = 1.6$, for the training loss and testing accuracy of the average model vs. iteration, respectively.

Similarly, in Figure 5.2, the top plots show the performance of DNA-DP for various values of (normalized) ϵ, and the bottom plots demonstrate the performance of DNA-DP in comparison to PDOP and DPDO with $\epsilon = 1$ for the logistic regression model on the Fashion-MNIST dataset.

Top plots in Figure 5.1 and5.2 show that our approach with larger ϵ has better convergence, consistent with Theorem 5.2. Moreover, we can observe the utility-privacy

[2] http://yann.lecun.com/exdb/mnist/

[3] The MIT License (MIT) Copyright © [2017] Zalando SE, https://tech.zalando.com

trade-off of our scheme. Larger ϵ indicating a weaker privacy guarantee would result in better utility. Finally, the bottom plots in Figure 5.1 and5.2 clearly show that our algorithm outperforms the previous methods for decentralized learning under the same budget of privacy.

5.6 Concluding Remarks

In our study, we introduced an accelerated distributed Nesterov gradient-based method with differential privacy. To ensure privacy, we inject Laplace noise to perturb the exchanged information. For any given tolerance of information leakage, we characterized the parameters of the added noise for each exchanged model variable. Subsequently, we derived the convergence rate of our proposed algorithm for the convex setting.

The rise of big data analytics, modern decentralized computer architectures, and extensive data collection has increased interest in exploring multi-agent networked systems. Such systems preserve data privacy and ownership. However, implementing distributed algorithms in such cases involves multiple rounds of exchanging local variables among agents, which poses a risk of private information leakage. This work significantly enhances the practicality of distributed learning and optimization in data-sensitive applications. A notable societal implication of our work is the feasibility of establishing a united healthcare system, where patients and researchers worldwide could allow the use of their private medical records for scientific research without concerns about individual privacy or relinquishing ownership of their valuable collected data.

Several avenues are left for future research, including the study of the proposed algorithm over time-varying networks and in non-convex settings.

5.7 The Preliminaries

In this section, we present some auxiliary results, which will be used in the proof of the main theorems. The proof of Lemma 5.5-5.9, 5.11, and Corollary 5.1-5.7 are provided in Section 5.10. We refer to the cited references for proof of other preliminaries.

Lemma 5.1. For any vector $\boldsymbol{u} \in \mathbb{R}^d$, we have

$$\|\boldsymbol{u}\| \leq \|\boldsymbol{u}\|_1 \leq \sqrt{d}\|\boldsymbol{u}\|.$$

Lemma 5.2 (Lemma 2 in [59]). For any pair of vectors $\boldsymbol{u}$, $\mathbf{r}$, and any scalar $\theta > 0$, we have

$$\|\boldsymbol{u}+\mathbf{r}\|^2 \le (1+\theta)\|\boldsymbol{u}\|^2 + (1+\theta^{-1})\|\mathbf{r}\|^2.$$

Corollary 5.1. For any vectors $\{\boldsymbol{u}_i\}_{i=1}^n \in \mathbb{R}^d$, we get

$$\left\|\sum_{i=1}^n \boldsymbol{u}_i\right\|^2 \le n\sum_{i=1}^n \|\boldsymbol{u}_i\|^2. \tag{5.11}$$

Corollary 5.2. For any pair of vectors $\boldsymbol{u}$, $\mathbf{r}$, and any scalar $\theta > 0$, we have

$$\|\boldsymbol{u}+\mathbf{r}\|^2 \ge \frac{1}{1+\theta}\|\boldsymbol{u}\|^2 - \frac{1}{\theta}\|\mathbf{r}\|^2.$$

Lemma 5.3 (Remark 6 in [95]). For $A \in \mathbb{R}^{d\times n}, B \in \mathbb{R}^{n\times n}$

$$\|AB\| \le \|A\|\|B\|_2.$$

Lemma 5.4 (Lemma 16 in [95]). For W satisfying Assumption 3.3, we get

$$\left\|W - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right\|_2 \le 1-\beta.$$

Corollary 5.3. For any $A \in \mathbb{R}^{n\times d}$, we can write

$$\left\|\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)A\right\| \le \|A\|.$$

Lemma 5.5. For any $A \in \mathbb{R}^{n\times d}$, matrix W satisfying Assumption 3.3, and any $\gamma \in [0,1]$, we have

$$\left\|((1-\gamma)I + \gamma W)\left(A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right)\right\| \le \rho\left\|A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right\|,$$

where $\rho := 1-\gamma\beta \in (0,1]$ and $\beta = 1-|\lambda_2(W)| \in (0,1]$ as defined in Assumption 3.3.

Lemma 5.6. Let c, ρ, L be positive constants such that $\rho < c\rho < 1$, and let

$$\delta_0 := \frac{6K^2}{\sqrt{\rho}(1-\sqrt{\rho})^3}. \tag{5.12}$$

For ν satisfying $\nu^2 < (c-1)^2\rho^2/5\delta_0$, define

$$A(\nu) := \begin{bmatrix} \rho & 0 & \frac{\nu^2}{1-\sqrt{\rho}} \\ \frac{1}{1-\sqrt{\rho}} & \rho & \frac{\nu^2}{\sqrt{\rho}(1-\sqrt{\rho})^2} \\ \frac{6K^2}{(1-\sqrt{\rho})^2} & \frac{12K^2}{1-\sqrt{\rho}} & c\rho + \frac{6K^2\nu^2}{\sqrt{\rho}(1-\sqrt{\rho})^3}. \end{bmatrix},$$

and let $\mu(\nu)$ be the largest eigenvalue of the matrix $A(\nu)$, and $\boldsymbol{\Theta}(\nu) := [\Theta_1(\nu), \Theta_2(\nu), 1]^T$ be its corresponding (normalized) eigenvector. Then, we have

(a) $c\rho + \delta_0\nu^2 < \mu(\nu) < c\rho + 4\delta_0^{1/3}\nu^{2/3}$,

(b) $\frac{\nu^2}{4(1-\sqrt{\rho})} < \Theta_1(\nu) < \frac{\nu^2}{(c-1)\rho(1-\sqrt{\rho})}$,

(c) $\Theta_2(\nu) < \frac{\nu^2}{(c-1)\rho(1-\sqrt{\rho})^2}\left(\frac{1}{(c-1)\rho} + \frac{1}{\sqrt{\rho}}\right)$.

Corollary 5.4. For any $\nu > 0$ with $\nu^2 < (c\rho)^2\left(1-(c\rho)^{1/3}\right)^3/64\delta_0$, we arrive at $\mu(\nu) < (c\rho)^{2/3}$.

Lemma 5.7. For any $\nu > 0$ with $\nu^2 < \sqrt{\rho}(c-1)^3\rho^3/64\delta_0$, we obtain

(a) $c\rho + \frac{3\delta_0\nu^2}{(c-1)^{3/2}\rho^{3/2}} < \mu(\nu)$,

(b) $\frac{\nu^2}{4(1-\sqrt{\rho})^2(c-1)^2\rho^2} < \Theta_2(\nu)$,

(c) $\boldsymbol{\Theta}(\nu) > \frac{1}{4}[\nu^2, \nu^2, 1]^T$.

Lemma 5.8. Consider $0 < \nu_1 < \nu_2$ such that $\nu_2^2 < \sqrt{\rho}(c-1)^3\rho^3/64\delta_0$. Then, the entries of the eigenvectors of the matrix $A(\nu)$ defined in (5.61) for ν_1 and ν_2 satisfy

$$\frac{\Theta_1(\nu_2)}{\Theta_1(\nu_1)} < \left(\frac{\nu_2}{\nu_1}\right)^{\delta_1},$$
$$\frac{\Theta_2(\nu_2)}{\Theta_2(\nu_1)} < \left(\frac{\nu_2}{\nu_1}\right)^{\delta_2}.$$

Here, we have $\delta_1 := \frac{6}{(c-1)^{1/2}\rho^{1/2}}$ and $\delta_2 := \frac{32c+16}{(c-1)^{1/2}\rho^{1/2}}$.

Corollary 5.5. Note that for $c \geq 0$, parameters δ_1, δ_2 defined in Lemma 5.8 satisfy $\delta_2 > \delta_1 > 1$. This implies that under the assumptions of Lemma 5.8 we have

$$\boldsymbol{\Theta}(\nu_2) < \left(\frac{\nu_2}{\nu_1}\right)^{\delta_2}\boldsymbol{\Theta}(\nu_1).$$

Lemma 5.9. Consider the function $h(x) := x^{2+\frac{16x+8}{(x-1)^{1/2}\rho^{1/2}}}$, defined for $x > 1$, where $\rho > 0$ is a constant. Then, for any $\delta_5 > 0$, there exists some $c_0 > 1$ such that $h(x) \leq 1 + \delta_5$ for every $1 < x \leq c_0$.

Lemma 5.10 (Lemma 9 in [121])**.** Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be a convex function for $i \in [n]$. Then, for function $\hat{f}(t) := \frac{1}{n}\sum_{i=1}^n \left(f_i(\mathbf{y}_i(t)) + \langle \nabla f_i(\mathbf{y}_i(t)), \bar{\mathbf{y}}(t) - \bar{\mathbf{y}}_i(t)\rangle\right)$, variable $\bar{\mathbf{z}}(t) := \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{y}_i(t))$, and any $\boldsymbol{\omega} \in \mathbb{R}^d$, we have

$$f(\boldsymbol{\omega}) \geq \hat{f}(t) + \langle \boldsymbol{\omega} - \bar{\mathbf{y}}(t), \bar{\mathbf{z}}(t)\rangle. \tag{5.13}$$

Moreover, under Assumption 3.1 we get

$$f(\boldsymbol{\omega}) \le \hat{f}(t) + \langle \boldsymbol{\omega} - \bar{\mathbf{y}}(t), \bar{\mathbf{z}}(t) \rangle + K\|\boldsymbol{\omega} - \bar{\mathbf{y}}(t)\|^2 + \frac{K}{n}\|Y(t) - \mathbf{1}\bar{\mathbf{y}}(t)\|^2. \tag{5.14}$$

Lemma 5.11. Let $\sigma \in (0,2)$ and $\nu_0 > 0$ be two constants, and pick some $\tau \ge \left(\frac{4(2-\sigma)}{\alpha_0(4+\alpha_0)}\right)^{\frac{1}{1-\sigma/2}}$. Define $\nu(t) = \frac{\nu_0}{(t+\tau)^\sigma}$ with $\nu_0 < \frac{1}{4K}$ and $\alpha(t)$ as introduced in (5.7). Then, we have

(a) $0 \le \alpha(t) \le \alpha(t+1) \le 1$,

(b) $0 \le -\Delta\nu(t) = \nu(t) - \nu(t+1) \le \sigma\frac{\nu(t)}{t+\tau}$,

(c) $\frac{4(2-\sigma)}{(4+\alpha_0)(t+\tau)} \le \alpha(t) \le \frac{2}{t+1}$,

for every $t \ge 0$.

5.8 Proof of Theorem 5.1

Using the fact that $\mathbf{e}_{i,y}(t)$ is sampled from a zero-mean Laplace distribution with scale parameter $b_y(t)$, we have $\mathbb{E}[\mathbf{e}_{i,y}(t)|\mathcal{F}_t] = 0$. Moreover, we can write

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{e}_{i,y}(t)\|^2|\mathcal{F}_t\right] &= 2b_y^2(t) \\
&= 2\left(\frac{8\zeta_2\sqrt{d}}{m\epsilon}\nu(t-2)\right)^2 \\
&= \frac{128\zeta_2^2 d}{m^2\epsilon^2}\left(\frac{\nu(t-2)}{\nu(t)}\right)^2\nu^2(t) \\
&= \frac{128\zeta_2^2 d}{m^2\epsilon^2}\left(\frac{t+\tau}{t+\tau-2}\right)^2\nu^2(t) \\
&\le \frac{1152\zeta_2^2 d}{m^2\epsilon^2}\nu^2(t),
\end{aligned} \tag{5.15}$$

where the last step follows from $\frac{t+\tau}{t+\tau-2} \le 3$ for every $t \ge 3$. Similarly, we get $\mathbb{E}[\mathbf{e}_{i,g}(t)|\mathcal{F}_t] = \mathbb{E}[\mathbf{e}_{i,v}(t)|\mathcal{F}_t] = 0$ and

$$\mathbb{E}\left[\|\mathbf{e}_{i,g}(t)\|^2|\mathcal{F}_t\right] \le \frac{32\zeta_2^2 d}{m^2\epsilon^2}\nu^2(t), \quad \mathbb{E}\left[\|\mathbf{e}_{i,v}(t)\|^2|\mathcal{F}_t\right] \le \frac{288\zeta_2^2 d}{m^2\epsilon^2}\frac{\nu^2(t)}{\alpha^2(t)}. \tag{5.16}$$

Hence, Assumption 5.1 holds for

$$\zeta_1 := \frac{1152\zeta_2^2 d}{m^2\epsilon^2}. \tag{5.17}$$

Now, we show that each iteration of update rules in (5.6) provides ϵ-differential privacy.

Note that $\mathcal{D}$ and $\mathcal{D}'$ are two adjacent datasets, in which $\mathcal{D}_j = \mathcal{D}'_j$ for every $j \neq i$. Therefore, all the variables sent from node $j \in [n] \setminus \{i\}$ will be identical under both problems, provided that node i sends identical variables. Hence, the two probabilities on the two sides of the inequalities are identical for $j \in [n] \setminus \{i\}$.

Now, consider node i, for which $\mathcal{D}_i \neq \mathcal{D}'_i$. The privacy loss for $\hat{\mathbf{g}}_i(t)$ is given by

$$\ln \frac{\Pr[\hat{\mathbf{g}}_i(t)|\mathcal{D}]}{\Pr[\hat{\mathbf{g}}_i(t)|\mathcal{D}']} = \ln \frac{\Pr[\mathbf{e}_{i,g}(t) = \hat{\mathbf{g}}_i(t) - \mathbf{g}_i(t)]}{\Pr[\mathbf{e}'_{i,g}(t) = \hat{\mathbf{g}}_i(t) - \mathbf{g}'_i(t)]}.$$

Note that each entry of $\mathbf{e}_{i,g}(t)$ and $\mathbf{e}'_{i,g}(t)$ is sampled from zero-mean Laplace distribution with the scale parameter $b_g(t)$, thus we can write

$$\ln \frac{\Pr[\hat{\mathbf{g}}_i(t)|\mathcal{D}]}{\Pr[\hat{\mathbf{g}}_i(t)|\mathcal{D}']} = \frac{1}{b_g(t)} \left[\|\hat{\mathbf{g}}_i(t) - \mathbf{g}'_i(t)\|_1 - \|\hat{\mathbf{g}}_i(t) - \mathbf{g}_i(t)\|_1 \right] \leq \frac{1}{b_g(t)} \|\mathbf{g}_i(t) - \mathbf{g}'_i(t)\|_1,$$

where the last step follows from the triangle inequality. Similarly, we get

$$\ln \frac{\Pr[\hat{\boldsymbol{v}}_i(t)|\mathcal{D}]}{\Pr[\hat{\boldsymbol{v}}_i(t)|\mathcal{D}']} \leq \frac{1}{b_v(t)} \|\boldsymbol{v}_i(t) - \boldsymbol{v}'_i(t)\|_1, \quad \ln \frac{\Pr[\hat{\mathbf{y}}_i(t)|\mathcal{D}]}{\Pr[\hat{\mathbf{y}}_i(t)|\mathcal{D}']} \leq \frac{1}{b_y(t)} \|\mathbf{y}_i(t) - \mathbf{y}'_i(t)\|_1.$$

We assume that two adjacent datasets $\mathcal{D}_i$ and $\mathcal{D}'_i$ only differ in the mth data point. Starting from the dynamics in (5.6a), we obtain

$$\begin{aligned} \mathbf{g}_i(t) - \mathbf{g}'_i(t) &= \nu(t-1) \left(\nabla f_i(\mathbf{y}_i(t-1)) - \nabla f'_i(\mathbf{y}'_i(t-1)) \right) \\ &= \frac{\nu(t-1)}{m} \left(\nabla \ell(\mathbf{y}_i(t-1), \xi^i_m) - \nabla \ell(\mathbf{y}'_i(t-1), \xi'^i_m) \right). \end{aligned}$$

This implies

$$\begin{aligned} \|\mathbf{g}_i(t) - \mathbf{g}'_i(t)\|_1 &= \frac{\nu(t-1)}{m} \|\nabla \ell(\mathbf{y}_i(t-1), \xi^i_m) - \nabla \ell(\mathbf{y}'_i(t-1), \xi'^i_m)\|_1 \\ &\leq \frac{\nu(t-1)}{m} \left(\|\nabla \ell(\mathbf{y}_i(t-1), \xi^i_m)\|_1 + \|\nabla \ell(\mathbf{y}'_i(t-1), \xi'^i_m)\|_1 \right) \leq \frac{2\zeta_2 \sqrt{d}}{m} \nu(t-1), \end{aligned} \tag{5.18}$$

where in the last inequality we used Lemma 5.1 and Assumption 5.2-(b). Using the update rule in (5.6b), we have $\mathbf{x}_i(t) - \mathbf{x}'_i(t) = \mathbf{g}'_i(t) - \mathbf{g}_i(t) + \mathbf{g}_i(t-1) - \mathbf{g}'_i(t-1)$. This combined with (5.18) leads to

$$\begin{aligned} \|\mathbf{x}_i(t) - \mathbf{x}'_i(t)\|_1 &= \|\mathbf{g}'_i(t) - \mathbf{g}_i(t) + \mathbf{g}_i(t-1) - \mathbf{g}'_i(t-1)\|_1 \\ &\leq \left(\|\mathbf{g}'_i(t) - \mathbf{g}_i(t)\|_1 + \|\mathbf{g}_i(t-1) - \mathbf{g}'_i(t-1)\|_1 \right) \\ &\leq \frac{2\zeta_2 \sqrt{d}}{m} (\nu(t-2) + \nu(t-1)) \leq \frac{4\zeta_2 \sqrt{d}}{m} \nu(t-2), \end{aligned}$$

where the last inequality holds since $\{\nu(t)\}$ is a non-increasing sequence. Moreover, from (5.6c) we have

$$\begin{aligned}\|\mathbf{v}_i(t)-\mathbf{v}'_i(t)\|_1 &= \frac{1}{\alpha(t-1)}\|\mathbf{g}'_i(t)-\mathbf{g}_i(t)+\mathbf{g}_i(t-1)-\mathbf{g}'_i(t-1)\|_1\\ &\leq \frac{2\zeta_2\sqrt{d}}{m\cdot\alpha(t-1)}(\nu(t-2)+\nu(t-1))\leq\frac{4\zeta_2\sqrt{d}}{m\cdot\alpha(t-1)}\nu(t-2).\end{aligned}$$

Finally, from (5.6d), we can write

$$\begin{aligned}\|\mathbf{y}_i(t)-\mathbf{y}'_i(t)\|_1 &= \|(1-\alpha(t))(\mathbf{x}_i(t)-\mathbf{x}'_i(t))+\alpha(t)(\mathbf{v}_i(t)-\mathbf{v}'_i(t))\|_1\\ &\leq (1-\alpha(t))\|\mathbf{x}_i(t)-\mathbf{x}'_i(t)\|_1+\alpha(t)\|\mathbf{v}_i(t)-\mathbf{v}'_i(t)\|_1\\ &\leq \frac{4\zeta_2\sqrt{d}}{m}(1-\alpha(t))\nu(t-2)+\frac{4\zeta_2\sqrt{d}}{m\cdot\alpha(t-1)}\alpha(t)\nu(t-2)\leq\frac{8\zeta_2\sqrt{d}}{m}\nu(t-2).\end{aligned}$$

Thus, we can write

$$\ln\frac{\Pr[\hat{\mathbf{g}}_i(t)|\mathcal{D}]}{\Pr[\hat{\mathbf{g}}_i(t)|\mathcal{D}']}\leq\frac{1}{b_g(t)}\frac{2\zeta_2\sqrt{d}}{m}\nu(t-1)=\epsilon,\quad \ln\frac{\Pr[\hat{\boldsymbol{v}}_i(t)|\mathcal{D}]}{\Pr[\hat{\boldsymbol{v}}_i(t)|\mathcal{D}']}\leq\frac{1}{b_v(t)}\frac{4\zeta_2\sqrt{d}}{m}\frac{\nu(t-2)}{\alpha(t-1)}=\epsilon,$$
$$\ln\frac{\Pr[\hat{\mathbf{y}}_i(t)|\mathcal{D}]}{\Pr[\hat{\mathbf{y}}_i(t)|\mathcal{D}']}\leq\frac{1}{b_y(t)}\frac{8\zeta_2\sqrt{d}}{m}\nu(t-2)=\epsilon,$$

which completes the proof of Theorem 5.1. □

5.9 Proof of Theorem 5.2

In this section, we provide the proof of the result on the convergence rate of the proposed algorithm, namely Theorem 5.2. The key steps of the proof are twofold: We first establish an upper bound on the deviation of the agents' models from their average and then analyze the distance of the average model's loss from the minimum loss value. These together lead to the proof of the theorem. We consider that the parameters τ, ρ, and σ should be chosen such that

(a) The parameters τ and σ satisfy

(1) $\zeta_3 := \sup_{t\geq 0}\frac{\nu(t)}{\nu(t+1)}<\rho^{-\frac{1}{6(\delta_2+4)}}$,

(2) $\tau\geq\left(\frac{4(2-\sigma)}{\alpha_0(4+\alpha_0)}\right)^{\frac{1}{1-\sigma/2}}$.

(b) The initial step-size ν_0 is small enough such that

(1) $\nu_0^2 < \left(\zeta_3^2\rho\right)^2 \left(1-\left(\zeta_3^2\rho\right)^{1/3}\right)^3 /64\delta_0$,

(2) $\nu_0^2 < \sqrt{\rho}(\zeta_3^2-1)^3\rho^3/64\delta_0$,

(3) $\nu_0 < \frac{1}{8K}$,

(4) $K^2\zeta_4\left(\zeta_6\zeta_9\nu_0^2+48\right)\nu_0 \leq \frac{\zeta_{11}}{\zeta_3^2}(\zeta_3-1)$,

(5) $\nu_0 < 2\zeta_4\zeta_7 - \zeta_3$,

(6) $\nu_0 < \rho^{-1/3} - \zeta_3$,

(7) $\nu_0 < \frac{1}{K}\left(1-\rho^{1/9}\right)^2$,

(8) $K^2\zeta_1\zeta_4\zeta_{10}\nu_0^3 \leq 1 - \zeta_3^{-2}\rho^{1/3}$,

(9) $\nu_0^3 < \frac{1}{8}\left(\frac{96K^3\zeta_4\zeta_6}{(\zeta_3^2-1)\rho(1-\sqrt{\rho})^3}\left(\frac{1}{(\zeta_3^2-1)\rho}+\frac{1}{\sqrt{\rho}}\right)\right)^{-1}$,

(10) $\nu_0^2 < \left(\frac{24\zeta_5\zeta_6}{\zeta_1(\zeta_3^2-1)\rho(1-\sqrt{\rho})^3}\left(\frac{1}{(\zeta_3^2-1)\rho}+\frac{1}{\sqrt{\rho}}\right)\right)^{-1}$,

where δ_0, δ_2, ζ_4, ζ_5, ζ_6, ζ_7, ζ_{10}, ζ_{11} are constants are determined later and $\rho := 1-\gamma\beta$. in (5.12), (5.65),(5.86), (5.59), (5.87), (5.88), (5.79), (5.81), respectively.

5.9.1 Models Deviation from the Average Model

For simplicity of notation, let

$$G(t) := \begin{bmatrix} \mathbf{g}_1(t) \\ \vdots \\ \mathbf{g}_n(t) \end{bmatrix}, \quad X(t) := \begin{bmatrix} \mathbf{x}_1(t) \\ \vdots \\ \mathbf{x}_n(t) \end{bmatrix}, \quad Y(t) := \begin{bmatrix} \mathbf{y}_1(t) \\ \vdots \\ \mathbf{y}_n(t) \end{bmatrix}, \quad V(t) := \begin{bmatrix} \boldsymbol{v}_1(t) \\ \vdots \\ \boldsymbol{v}_n(t) \end{bmatrix},$$

$$\nabla f(Y(t)) := \begin{bmatrix} \nabla f_1(\mathbf{y}_1(t)) \\ \vdots \\ \nabla f_n(\mathbf{y}_n(t)), \end{bmatrix}, \quad E_g(t) := \begin{bmatrix} \mathbf{e}_{1,g}(t) \\ \vdots \\ \mathbf{e}_{n,g}(t) \end{bmatrix}, \quad E_y(t) := \begin{bmatrix} \mathbf{e}_{1,y}(t) \\ \vdots \\ \mathbf{e}_{n,y}(t) \end{bmatrix}, \quad E_v(t) := \begin{bmatrix} \mathbf{e}_{1,v}(t) \\ \vdots \\ \mathbf{e}_{n,v}(t) \end{bmatrix},$$

and $\hat{G}(t)=G(t)+E_g(t)$, and $\hat{Y}(t)=Y(t)+E_y(t)$, and $\hat{V}(t)=V(t)+E_v(t)$. Therefore, we can rewrite the update algorithm (5.6) in the matrix format as

$$G(t+1) = ((1-\gamma)I+\gamma W)\hat{G}(t) + \nu(t)\nabla f(Y(t)), \tag{5.19a}$$
$$X(t+1) = ((1-\gamma)I+\gamma W)\hat{Y}(t) - (G(t+1)-G(t)), \tag{5.19b}$$
$$V(t+1) = ((1-\gamma)I+\gamma W)\hat{V}(t) - \frac{1}{\alpha(t)}(G(t+1)-G(t)), \tag{5.19c}$$
$$Y(t+1) = (1-\alpha(t+1))X(t+1) + \alpha(t+1)V(t+1). \tag{5.19d}$$

We define $\widetilde{\nabla} f(Y(t)) := \nu(t)\nabla f(Y(t)) + ((1-\gamma)I+\gamma W)E_g(t)$ and $S(t) := G(t+1)-G(t)$ where $S(0)=G(1)=\widetilde{\nabla} f(Y(0))$. Then, the updating rules in (5.19), can be rephrased as

$$X(t+1) = ((1-\gamma)I+\gamma W)\hat{Y}(t) - S(t), \tag{5.20a}$$
$$V(t+1) = ((1-\gamma)I+\gamma W)\hat{V}(t) - \frac{1}{\alpha(t)}S(t), \tag{5.20b}$$
$$Y(t+1) = (1-\alpha(t+1))X(t+1) + \alpha(t+1)V(t+1). \tag{5.20c}$$

Moreover, we can rewrite the update rule in (5.19a) as

$$\begin{aligned}
&S(t+1)\\
&= G(t+2) - G(t+1)\\
&= ((1-\gamma)I+\gamma W)(\hat{G}(t+1)-\hat{G}(t)) + \nu(t+1)\nabla f(Y(t+1)) - \nu(t)\nabla f(Y(t))\\
&= ((1-\gamma)I+\gamma W)(G(t+1)+E_g(t+1)-G(t)-E_g(t)) + \nu(t+1)\nabla f(Y(t+1)) - \nu(t)\nabla f(Y(t))\\
&= ((1-\gamma)I+\gamma W)S(t) + \widetilde{\nabla} f(Y(t+1)) - \widetilde{\nabla} f(Y(t)),
\end{aligned} \tag{5.20d}$$

Hence, we can study $(X(t),V(t),Y(t),S(t))$ instead of $(X(t),V(t),Y(t),G(t))$ in the remainder of this work. Further, we define

$$\Delta X(t) := X(t)-\mathbf{1}\bar{\mathbf{x}}(t),\ \Delta V(t) := V(t)-\mathbf{1}\bar{\boldsymbol{v}}(t),\ \Delta Y(t) := Y(t)-\mathbf{1}\bar{\mathbf{y}}(t),\ \Delta S(t) := S(t)-\mathbf{1}\bar{\boldsymbol{\sigma}}(t),$$
$$\bar{\mathbf{x}}(t) := \frac{1}{n}\mathbf{1}^T X(t),\quad \bar{\boldsymbol{v}}(t) := \frac{1}{n}\mathbf{1}^T V(t),\quad \bar{\mathbf{y}}(t) := \frac{1}{n}\mathbf{1}^T Y(t),\quad \bar{\boldsymbol{\sigma}}(t) := \frac{1}{n}\mathbf{1}^T S(t),$$

and the following notations for the noise sequences

$$\Delta E_v(t) := E_v(t) - \mathbf{1}\bar{\mathbf{e}}_v(t),\qquad \Delta E_y(t) := E_y(t) - \mathbf{1}\bar{\mathbf{e}}_y(t),\qquad \Delta E_g(t) := E_g(t) - \mathbf{1}\bar{\mathbf{e}}_g(t),$$
$$\bar{\mathbf{e}}_v(t) := \frac{1}{n}\mathbf{1}^T E_v(t),\quad \bar{\mathbf{e}}_y(t) := \frac{1}{n}\mathbf{1}^T E_y(t),\quad \bar{\mathbf{e}}_g(t) := \frac{1}{n}\mathbf{1}^T E_g(t).$$

From (5.20a)-(5.20d) and Assumption 3.3, we get recursive equations for the average variables, as

$$\bar{\mathbf{x}}(t+1) = \bar{\mathbf{y}}(t) + \bar{\mathbf{e}}_y(t) - \bar{\boldsymbol{\sigma}}(t), \tag{5.21}$$

$$\bar{\boldsymbol{v}}(t+1) = \bar{\boldsymbol{v}}(t) + \bar{\mathbf{e}}_v(t) - \frac{1}{\alpha(t)}\bar{\boldsymbol{\sigma}}(t), \tag{5.22}$$

$$\bar{\mathbf{y}}(t+1) = (1-\alpha(t+1))\bar{\mathbf{x}}(t+1) + \alpha(t+1)\bar{\boldsymbol{v}}(t+1), \tag{5.23}$$

$$\bar{\boldsymbol{\sigma}}(t+1) = \bar{\boldsymbol{\sigma}}(t) + \frac{1}{n}\mathbf{1}^T\widetilde{\nabla} f(Y(t+1)) - \frac{1}{n}\mathbf{1}^T\widetilde{\nabla} f(Y(t)), \tag{5.24}$$

Further, from (5.24), we have

$$\sum_{\ell=0}^{t-1}\bar{\boldsymbol{\sigma}}(\ell+1) = \sum_{\ell=0}^{t-1}\bar{\boldsymbol{\sigma}}(\ell) + \sum_{\ell=0}^{t-1}\frac{1}{n}\mathbf{1}^T\widetilde{\nabla} f(Y(\ell+1)) - \sum_{\ell=0}^{t-1}\frac{1}{n}\mathbf{1}^T\widetilde{\nabla} f(Y(\ell)).$$

This together with $\bar{\boldsymbol{\sigma}}(0) = \frac{1}{n}\mathbf{1}^T\widetilde{\nabla} f(Y(0))$ leads us to

$$\bar{\boldsymbol{\sigma}}(t) = \bar{\boldsymbol{\sigma}}(0) + \frac{1}{n}\mathbf{1}^T\widetilde{\nabla} f(Y(t)) - \frac{1}{n}\mathbf{1}^T\widetilde{\nabla} f(Y(0)) = \frac{1}{n}\mathbf{1}^T\widetilde{\nabla} f(Y(t)). \tag{5.25}$$

We start from from (5.20a) and use (5.21), to arrive at

$$\begin{aligned}
\Delta X(t+1) &= X(t+1) - \mathbf{1}\bar{\mathbf{x}}(t+1) \\
&= \left(((1-\gamma)I + \gamma W)\hat{Y}(t) - S(t)\right) - \mathbf{1}\left(\bar{\mathbf{y}}(t) + \bar{\mathbf{e}}_y(t) - \bar{\boldsymbol{\sigma}}(t)\right) \\
&= ((1-\gamma)I + \gamma W)(Y(t) + E_y(t)) - \mathbf{1}\bar{\mathbf{y}}(t) - \mathbf{1}\bar{\mathbf{e}}_y(t) - (S(t) - \mathbf{1}\bar{\boldsymbol{\sigma}}(t)) \\
&\overset{(a)}{=} ((1-\gamma)I + \gamma W)\left(Y(t) - \mathbf{1}\bar{\mathbf{y}}(t) + E_y(t) - \mathbf{1}\bar{\mathbf{e}}_y(t)\right) - (S(t) - \mathbf{1}\bar{\boldsymbol{\sigma}}(t)) \\
&= ((1-\gamma)I + \gamma W)\left(\Delta Y(t) + \Delta E_y(t)\right) - \Delta S(t),
\end{aligned} \tag{5.26}$$

where (a) follows from Assumption 3.3. Similarly, from (5.20b) and (5.22), we have

$$\begin{aligned}
\Delta V(t+1) &= V(t+1) - \mathbf{1}\bar{\boldsymbol{v}}(t+1) \\
&= \left(((1-\gamma)I + \gamma W)\hat{V}(t) - \frac{1}{\alpha(t)}S(t)\right) - \mathbf{1}\left(\bar{\boldsymbol{v}}(t) + \bar{\mathbf{e}}_v(t) + \frac{1}{\alpha(t)}\bar{\boldsymbol{\sigma}}(t)\right) \\
&= ((1-\gamma)I + \gamma W)(V(t) + E_v(t)) - \mathbf{1}\bar{\boldsymbol{v}}(t) - \mathbf{1}\bar{\mathbf{e}}_v(t) - \frac{1}{\alpha(t)}\Delta S(t) \\
&= ((1-\gamma)I + \gamma W)\left(\Delta V(t) + \Delta E_v(t)\right) - \frac{1}{\alpha(t)}\Delta S(t).
\end{aligned} \tag{5.27}$$

Moreover, using (5.20c) and (5.23), we arrive at

$$\begin{aligned}
&\Delta Y(t+1)\\
&= Y(t+1) - \mathbf{1}\bar{\mathbf{y}}(t+1)\\
&= ((1-\alpha(t+1))X(t+1) + \alpha(t+1)V(t+1)) - ((1-\alpha(t+1))\mathbf{1}\bar{\mathbf{x}}(t+1) + \alpha(t+1)\mathbf{1}\bar{\boldsymbol{v}}(t+1))\\
&= (1-\alpha(t+1))\Delta X(t+1) + \alpha(t+1)\Delta V(t+1). \qquad (5.28)
\end{aligned}$$

Finally, from (5.20d), we can write

$$\begin{aligned}
&\Delta S(t+1)\\
&= S(t+1) - \mathbf{1}\bar{\boldsymbol{\sigma}}(t+1)\\
&= \left(((1-\gamma)I + \gamma W)S(t) + \widetilde{\nabla} f(Y(t+1)) - \widetilde{\nabla} f(Y(t))\right) - \mathbf{1}\bar{\boldsymbol{\sigma}}(t+1)\\
&= ((1-\gamma)I+\gamma W)(S(t) - \mathbf{1}\bar{\boldsymbol{\sigma}}(t)) + ((1-\gamma)I+\gamma W)\mathbf{1}\bar{\boldsymbol{\sigma}}(t)\\
&\quad + \widetilde{\nabla} f(Y(t+1)) - \widetilde{\nabla} f(Y(t)) - \mathbf{1}\bar{\boldsymbol{\sigma}}(t+1)\\
&\overset{(a)}{=} ((1-\gamma)I + \gamma W)\Delta S(t) + \mathbf{1}\bar{\boldsymbol{\sigma}}(t) + \widetilde{\nabla} f(Y(t+1)) - \widetilde{\nabla} f(Y(t)) - \mathbf{1}\bar{\boldsymbol{\sigma}}(t+1)\\
&\overset{(b)}{=} ((1-\gamma)I + \gamma W)\Delta S(t) + \frac{1}{n}\mathbf{1}\mathbf{1}^T\nu(t)\nabla f(Y(t)) + \mathbf{1}\bar{\mathbf{e}}_g(t)\\
&\quad + \nu(t+1)\nabla f(Y(t+1)) + ((1-\gamma)I+\gamma W)E_g(t+1) - \nu(t)\nabla f(Y(t)) - ((1-\gamma)I+\gamma W)E_g(t)\\
&\quad - \frac{1}{n}\mathbf{1}\mathbf{1}^T\nu(t+1)\nabla f(Y(t+1)) - \mathbf{1}\bar{\mathbf{e}}_g(t+1)\\
&\overset{(c)}{=} ((1-\gamma)I + \gamma W)\Delta S(t) + ((1-\gamma)I + \gamma W)\left(\Delta E_g(t+1) - \Delta E_g(t)\right)\\
&\quad + \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left[\nu(t+1)\nabla f(Y(t+1)) - \nu(t)\nabla f(Y(t))\right]\\
&= ((1-\gamma)I + \gamma W)\left(\Delta S(t) + \Delta E_g(t+1) - \Delta E_g(t)\right)\\
&\quad + \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left[\nu(t+1)\nabla f(Y(t+1)) - \nu(t)\nabla f(Y(t))\right], \qquad (5.29)
\end{aligned}$$

where (a), (c) follow from Assumption 3.3 and (b) holds due to (5.25).

Next, applying Lemma 5.2 with $\theta = \frac{1}{\sqrt{\rho}} - 1$ for (5.26), we get

$$\mathbb{E}\left[\|\Delta X(t+1)\|^2\right] \le \frac{1}{\sqrt{\rho}}\mathbb{E}\left[\|((1-\gamma)I+\gamma W)\left(\Delta Y(t) + \Delta E_y(t)\right)\|^2\right] + \frac{1}{1-\sqrt{\rho}}\mathbb{E}\left[\|\Delta S(t)\|^2\right]. \qquad (5.30)$$

We continue with bounding the first term in (5.30). Using Lemma 5.5 for $\Delta Y(t) =$

$Y(t) - \frac{1}{n}\mathbf{1}\mathbf{1}^T Y(t)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|((1-\gamma)I+\gamma W)\left(\Delta Y(t)+\Delta E_y(t)\right)\|^2\right] \\
&\leq \rho^2 \mathbb{E}\left[\|\Delta Y(t)+\Delta E_y(t)\|^2\right] \\
&= \rho^2 \mathbb{E}\left[\|\Delta Y(t)\|^2\right] + \rho^2 \mathbb{E}\left[\|\Delta E_y(t)\|^2\right] + 2\rho^2 \mathbb{E}\left[\langle \Delta Y(t), \Delta E_y(t)\rangle\right] \\
&= \rho^2 \mathbb{E}\left[\|\Delta Y(t)\|^2\right] + \rho^2 \mathbb{E}\left[\left\|\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_y(t)\right\|^2\right] + 2\rho^2 \mathbb{E}\left[\left\langle \Delta Y(t), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_y(t)\right\rangle\right] \\
&\overset{(a)}{\leq} \rho^2 \mathbb{E}\left[\|\Delta Y(t)\|^2\right] + \rho^2 \mathbb{E}\left[\|E_y(t)\|^2\right] + 2\rho^2 \mathbb{E}\left[\left\langle \Delta Y(t), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_y(t)\right\rangle\right] \\
&\overset{(b)}{=} \rho^2 \mathbb{E}\left[\|\Delta Y(t)\|^2\right] + \rho^2 \mathbb{E}\left[\|E_y(t)\|^2\right], \qquad (5.31)
\end{aligned}$$

where (a) follows from Corollary 5.3. The step (b) relies on Assumption 5.1 and

$$\begin{aligned}
\mathbb{E}\left[\left\langle \Delta Y(t), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_y(t)\right\rangle\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\langle \Delta Y(t), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_y(t)\right\rangle \middle| F_t\right]\right] \\
&= \mathbb{E}\left[\left\langle \Delta Y(t), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbb{E}\left[E_y(t)|\mathcal{F}_t\right]\right\rangle\right] = 0.
\end{aligned}$$

To bound the second term in (5.31), from Assumption 5.1, we have

$$\mathbb{E}\left[\|E_y(t)\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|E_y(t)\|^2|\mathcal{F}_t\right]\right] = \mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{e}_{i,y}(t)\|^2|\mathcal{F}_t\right]\right] \leq n\zeta_1\nu^2(t). \qquad (5.32)$$

Plugging (5.31) and (5.32) into (5.30), we have

$$\mathbb{E}\left[\|\Delta X(t+1)\|^2\right] \leq \rho^{3/2}\left(\mathbb{E}\left[\|\Delta Y(t)\|^2\right] + n\zeta_1\nu^2(t)\right) + \frac{1}{1-\sqrt{\rho}}\mathbb{E}\left[\|\Delta S(t)\|^2\right]. \qquad (5.33)$$

This establishes the recursive bound for deviation $\mathbb{E}\left[\|\Delta X(t+1)\|^2\right]$. Similarly, from (5.27),

we can write

$$
\begin{aligned}
&\mathbb{E}\left[\|\Delta V(t+1)\|^2\right] \\
&\leq \frac{1}{\sqrt{\rho}}\mathbb{E}\left[\|((1-\gamma)I+\gamma W)\left(\Delta V(t)+\Delta E_v(t)\right)\|^2\right]+\frac{1}{(1-\sqrt{\rho})\alpha^2(t)}\mathbb{E}\left[\|\Delta S(t)\|^2\right] \\
&\leq \rho^{3/2}\mathbb{E}\left[\|\Delta V(t)\|^2\right]+\rho^{3/2}\mathbb{E}\left[\|E_v(t)\|^2\right]+2\rho^{3/2}\mathbb{E}\left[\left\langle \Delta V(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_v(t)\right\rangle\right] \\
&\quad+\frac{1}{(1-\sqrt{\rho})\alpha^2(t)}\mathbb{E}\left[\|\Delta S(t)\|^2\right] \\
&\overset{(a)}{=} \rho^{3/2}\mathbb{E}\left[\|\Delta V(t)\|^2\right]+\rho^{3/2}\mathbb{E}\left[\|E_v(t)\|^2\right]+\frac{1}{(1-\sqrt{\rho})\alpha^2(t)}\mathbb{E}\left[\|\Delta S(t)\|^2\right] \\
&\overset{(b)}{\leq} \rho^{3/2}\left(\mathbb{E}\left[\|\Delta V(t)\|^2\right]+n\zeta_1\frac{\nu^2(t)}{\alpha^2(t)}\right)+\frac{1}{(1-\sqrt{\rho})\alpha^2(t)}\mathbb{E}\left[\|\Delta S(t)\|^2\right] \\
&\leq \rho\left(\mathbb{E}\left[\|\Delta V(t)\|^2\right]+n\zeta_1\frac{\nu^2(t)}{\alpha^2(t)}\right)+\frac{1}{(1-\sqrt{\rho})\alpha^2(t)}\mathbb{E}\left[\|\Delta S(t)\|^2\right],
\end{aligned}
\tag{5.34}
$$

where (a) and (b) follow from Assumption 5.1 and can be explained by

$$
\begin{aligned}
&\mathbb{E}\left[\left\langle \Delta V(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_v(t)\right\rangle\right]=\mathbb{E}\left[\left\langle \Delta V(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbb{E}\left[E_v(t)|\mathcal{F}_t\right]\right\rangle\right]=0, \\
&\mathbb{E}\left[\|E_v(t)\|^2\right]=\mathbb{E}\left[\mathbb{E}\left[\|E_v(t)\|^2|\mathcal{F}_t\right]\right]=\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{e}_{i,v}(t)\|^2|\mathcal{F}_t\right]\right]\leq n\zeta_1\frac{\nu^2(t)}{\alpha^2(t)},
\end{aligned}
$$

and the last inequality holds due to $0<\rho^{3/2}<\rho<1$. From (5.34) and Lemma 5.11-(a), we can write

$$
\begin{aligned}
\alpha^2(t+1)\mathbb{E}\left[\|\Delta V(t+1)\|^2\right] &\leq \alpha^2(t)\mathbb{E}\left[\|\Delta V(t+1)\|^2\right] \\
&\leq \rho\alpha^2(t)\left(\mathbb{E}\left[\|\Delta V(t)\|^2\right]+n\zeta_1\frac{\nu^2(t)}{\alpha^2(t)}\right)+\frac{1}{1-\sqrt{\rho}}\mathbb{E}\left[\|\Delta S(t)\|^2\right] \\
&= \rho\left(\alpha^2(t)\mathbb{E}\left[\|\Delta V(t)\|^2\right]+n\zeta_1\nu^2(t)\right)+\frac{1}{1-\sqrt{\rho}}\mathbb{E}\left[\|\Delta S(t)\|^2\right].
\end{aligned}
\tag{5.35}
$$

Thus, we got a recursive upper bound for $\alpha^2(t+1)\mathbb{E}\left[\|\Delta V(t+1)\|^2\right]$.

Similarly, from (5.28) and using Lemma 5.2 with $\theta = \frac{1}{\sqrt{\rho}} - 1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\Delta Y(t+1)\|^2\right] \\
&= \mathbb{E}\left[\|(1-\alpha(t+1))\Delta X(t+1) + \alpha(t+1)\Delta V(t+1)\|^2\right] \\
&\leq \frac{1}{\sqrt{\rho}}\mathbb{E}\left[\|(1-\alpha(t+1))\Delta X(t+1)\|^2\right] + \frac{1}{1-\sqrt{\rho}}\mathbb{E}\left[\|\alpha(t+1)\Delta V(t+1)\|^2\right] \\
&\overset{(a)}{\leq} \frac{1}{\sqrt{\rho}}\mathbb{E}\left[\|\Delta X(t+1)\|^2\right] + \frac{1}{1-\sqrt{\rho}}\alpha^2(t)\mathbb{E}\left[\|\Delta V(t+1)\|^2\right] \\
&\overset{(b)}{\leq} \rho(\mathbb{E}\left[\|\Delta Y(t)\|^2\right] + n\zeta_1\nu^2(t)) + \frac{\rho}{1-\sqrt{\rho}}\left(\alpha^2(t)\mathbb{E}\left[\|\Delta V(t)\|^2\right] + n\zeta_1\nu^2(t)\right) \\
&\quad + \frac{1}{\sqrt{\rho}(1-\sqrt{\rho})^2}\mathbb{E}\left[\|\Delta S(t)\|^2\right] \\
&\leq \rho(\mathbb{E}\left[\|\Delta Y(t)\|^2\right] + n\zeta_1\nu^2(t)) + \frac{1}{1-\sqrt{\rho}}\left(\alpha^2(t)\mathbb{E}\left[\|\Delta V(t)\|^2\right] + n\zeta_1\nu^2(t)\right) \\
&\quad + \frac{1}{\sqrt{\rho}(1-\sqrt{\rho})^2}\mathbb{E}\left[\|\Delta S(t)\|^2\right],
\end{aligned} \tag{5.36}$$

where (a) follows from Lemma 5.11-(a), we used (5.33) and (5.34) in (b), and the last inequality holds due to $0 < \rho < 1$. This establishes the desired recursive bound for $\mathbb{E}\left[\|\Delta Y(t+1)\|^2\right]$.

Finally, from (5.29) and using Lemma (5.2) with $\theta = \frac{1}{\rho} - 1$, we can write

$$\begin{aligned}
\mathbb{E}\left[\|\Delta S(t+1)\|^2\right] \leq \frac{1}{\rho}\mathbb{E}\left[\|((1-\gamma)I + \gamma W)(\Delta S(t) + \Delta E_g(t+1) - \Delta E_g(t))\|^2\right] \\
+ \frac{1}{1-\rho}\mathbb{E}\left[\left\|\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)(\nu(t+1)\nabla f(Y(t+1)) - \nu(t)\nabla f(Y(t)))\right\|^2\right].
\end{aligned} \tag{5.37}$$

To bound the first term in (5.37), from Lemma 5.5, we get

$$\begin{aligned}
&\mathbb{E}\left[\|((1-\gamma)I+\gamma W)(\Delta S(t)+\Delta E_g(t+1)-\Delta E_g(t))\|^2\right] \\
&\leq \rho^2\mathbb{E}\left[\|\Delta S(t)+\Delta E_g(t+1)-\Delta E_g(t)\|^2\right] \\
&= \rho^2\mathbb{E}\left[\|\Delta S(t)\|^2\right]+\rho^2\mathbb{E}\left[\left\|\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)(E_g(t+1)-E_g(t))\right\|^2\right] \\
&\quad +2\rho^2\mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)(E_g(t+1)-E_g(t))\right\rangle\right] \\
&\overset{(b)}{\leq} \rho^2\mathbb{E}\left[\|\Delta S(t)\|^2\right]+\rho^2\mathbb{E}\left[\|E_g(t+1)-E_g(t)\|^2\right] \\
&\quad +2\rho^2\mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)(E_g(t+1)-E_g(t))\right\rangle\right]. \qquad (5.38)
\end{aligned}$$

where in (b) we used Corollary 5.3. Note that from Assumption 5.1 we arrive at

$$\mathbb{E}\left[\|E_g(t)\|^2\right]=\mathbb{E}\left[\mathbb{E}\left[\|E_g(t)\|^2|\mathcal{F}_t\right]\right]=\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{e}_{i,g}(t)\|^2|\mathcal{F}_t\right]\right]\leq n\zeta_1\nu^2(t). \qquad (5.39)$$

Now, we continue with bounding the second term in (5.38) as

$$\begin{aligned}
\mathbb{E}\left[\|E_g(t+1)-E_g(t)\|^2\right] &= \mathbb{E}\left[\|E_g(t+1)\|^2\right]+\mathbb{E}\left[\|E_g(t)\|^2\right]+2\mathbb{E}\left[\langle E_g(t+1),E_g(t)\rangle\right] \\
&= \mathbb{E}\left[\|E_g(t+1)\|^2\right]+\mathbb{E}\left[\|E_g(t)\|^2\right]+2\mathbb{E}\left[\mathbb{E}\left[\langle E_g(t+1),E_g(t)\rangle|\mathcal{F}_{t+1}\right]\right] \\
&= \mathbb{E}\left[\|E_g(t+1)\|^2\right]+\mathbb{E}\left[\|E_g(t)\|^2\right] \\
&\overset{(a)}{\leq} n\zeta_1\nu^2(t+1)+n\zeta_1\nu^2(t)\leq 2n\zeta_1\nu^2(t), \qquad (5.40)
\end{aligned}$$

where (a) follows from (5.39) and the last inequality holds since $\{\nu(t)\}$ is a non-increasing sequence. For the last term in (5.38), we can write

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)(E_g(t+1)-E_g(t))\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_g(t+1)\right\rangle\right]-\mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_g(t)\right\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_g(t+1)\right\rangle\middle|\mathcal{F}_{t+1}\right]\right]-\mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_g(t)\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbb{E}\left[E_g(t+1)|\mathcal{F}_{t+1}\right]\right\rangle\right]-\mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_g(t)\right\rangle\right] \\
&= -\mathbb{E}\left[\left\langle\Delta S(t),\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E_g(t)\right\rangle\right]. \qquad (5.41)
\end{aligned}$$

From (5.29), we get

$$
\begin{aligned}
&\mathbb{E}\left[\left\langle \Delta S(t), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t)\right\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle \Delta S(t), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t)\right\rangle \middle| \mathcal{F}_t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle ((1-\gamma)I+\gamma W)\left(\Delta S(t-1)+\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)(E_g(t)-E_g(t-1))\right), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t)\right\rangle \middle| \mathcal{F}_t\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\left\langle \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)[\nu(t)\nabla f(Y(t)) - \nu(t-1)\nabla f(Y(t-1))], \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t)\right\rangle \middle| \mathcal{F}_t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle ((1-\gamma)I+\gamma W)\Delta S(t-1), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t)\right\rangle \middle| \mathcal{F}_t\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\left\langle \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t)\right\rangle \middle| \mathcal{F}_t\right]\right] \\
&\quad - \mathbb{E}\left[\mathbb{E}\left[\left\langle \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t-1), \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t)\right\rangle \middle| \mathcal{F}_t\right]\right] \\
&\quad + \mathbb{E}\left[\left\langle \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)[\nu(t)\nabla f(Y(t)) - \nu(t-1)\nabla f(Y(t-1))], \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbb{E}[E_g(t)|\mathcal{F}_t]\right\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\| \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) E_g(t)\right\|^2\right] \middle| \mathcal{F}_t\right] \geq 0. \qquad (5.42)
\end{aligned}
$$

Plugging (5.42) and (5.40) into (5.41) and (5.38), we get

$$
\mathbb{E}\left[\|((1-\gamma)I+\gamma W)(\Delta S(t)+\Delta E_g(t+1) - \Delta E_g(t))\|^2\right] \leq \rho^2 \mathbb{E}\left[\|\Delta S(t)\|^2\right] + 2\rho^2 n\zeta_1 \nu^2(t). \qquad (5.43)
$$

To bound the second term in (5.37), from Corollary 5.3 we can write

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)(\nu(t+1)\nabla f(Y(t+1)) - \nu(t)\nabla f(Y(t)))\right\|^2\right] \\
&\leq \mathbb{E}\left[\|\nu(t+1)\nabla f(Y(t+1)) - \nu(t)\nabla f(Y(t))\|^2\right] \\
&= \mathbb{E}\left[\|\nu(t+1)\nabla f(Y(t+1)) - \nu(t)\nabla f(Y(t)) + \nu(t+1)\nabla f(Y(t)) - \nu(t+1)\nabla f(Y(t))\|^2\right] \\
&\overset{(a)}{\leq} 2\nu^2(t+1)\mathbb{E}\left[\|\nabla f(Y(t+1)) - \nabla f(Y(t))\|^2\right] + 2(\nu(t) - \nu(t+1))^2 \mathbb{E}\left[\|\nabla f(Y(t))\|^2\right] \\
&\leq 2\nu^2(t+1)\mathbb{E}\left[\|\nabla f(Y(t+1)) - \nabla f(Y(t))\|^2\right] + 2\sigma^2 \frac{\nu^2(t)}{(t+\tau)^2}\mathbb{E}\left[\|\nabla f(Y(t))\|^2\right], \qquad (5.44)
\end{aligned}
$$

where (a) follows from Lemma 5.2 with $\theta = 1$ and the last step holds due to to

Lemma 5.11-(b) for every $t \geq 0$. We further bound each term in (5.44). Using Assumption 3.1 and the Cauchy-Schwarz inequality, we can bound the first term as

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(Y(t+1)) - \nabla f(Y(t))\|^2\right] \\
&\leq K^2 \mathbb{E}\left[\|Y(t+1) - Y(t)\|^2\right] \\
&\leq 3K^2 \left(\mathbb{E}\left[\|Y(t+1) - \mathbf{1}\bar{\mathbf{y}}(t+1)\|^2\right] + \mathbb{E}\left[\|Y(t) - \mathbf{1}\bar{\mathbf{y}}(t)\|^2\right] + \mathbb{E}\left[\|\mathbf{1}(\bar{\mathbf{y}}(t+1) - \mathbf{1}\bar{\mathbf{y}}(t))\|^2\right]\right) \\
&= 3K^2 \left(\mathbb{E}\left[\|\Delta Y(t+1)\|^2\right] + \mathbb{E}\left[\|\Delta Y(t)\|^2\right] + n\mathbb{E}\left[\|\bar{\mathbf{y}}(t+1) - \bar{\mathbf{y}}(t)\|^2\right]\right). \qquad (5.45)
\end{aligned}$$

Starting from (5.22) and using (5.25), we have

$$\begin{aligned}
\bar{\boldsymbol{v}}(t+1) &= \bar{\boldsymbol{v}}(t) + \bar{\mathbf{e}}_v(t) - \frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t) - \frac{1}{\alpha(t)}\left(\frac{1}{n}\mathbf{1}^T((1-\gamma)I + \gamma W)\right) E_g(t) \\
&= \bar{\boldsymbol{v}}(t) + \bar{\mathbf{e}}_v(t) - \frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t) - \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t), \qquad (5.46)
\end{aligned}$$

where the last equality holds due to Assumption 3.3. Similarly, from (5.21), we can write

$$\begin{aligned}
\bar{\mathbf{x}}(t+1) &= \bar{\mathbf{y}}(t) + \bar{\mathbf{e}}_y(t) - \nu(t)\bar{\mathbf{z}}(t) - \left(\frac{1}{n}\mathbf{1}^T((1-\gamma)I + \gamma W)\right) E_g(t) \\
&= \bar{\mathbf{y}}(t) + \bar{\mathbf{e}}_y(t) - \nu(t)\bar{\mathbf{z}}(t) - \bar{\mathbf{e}}_g(t). \qquad (5.47)
\end{aligned}$$

Using (5.23), (5.46), and (5.47), we get

$$\begin{aligned}
&\bar{\mathbf{y}}(t+1) \\
&= (1-\alpha(t+1))\bar{\mathbf{x}}(t+1) + \alpha(t+1)\bar{\boldsymbol{v}}(t+1) \\
&= (1-\alpha(t+1))(\bar{\mathbf{y}}(t) + \bar{\mathbf{e}}_y(t) - \nu(t)\bar{\mathbf{z}}(t) - \bar{\mathbf{e}}_g(t)) + \alpha(t+1)\left(\bar{\boldsymbol{v}}(t) - \frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t) + \bar{\mathbf{e}}_v(t) - \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right) \\
&= (1-\alpha(t+1))\bar{\mathbf{y}}(t) + \alpha(t+1)\bar{\boldsymbol{v}}(t) - \left(\frac{\alpha(t+1)}{\alpha(t)} + 1 - \alpha(t+1)\right)\nu(t)\bar{\mathbf{z}}(t) \\
&\quad + (1-\alpha(t+1))\bar{\mathbf{e}}_y(t) + \alpha(t+1)\bar{\mathbf{e}}_v(t) - \left(\frac{\alpha(t+1)}{\alpha(t)} + 1 - \alpha(t+1)\right)\bar{\mathbf{e}}_g(t). \qquad (5.48)
\end{aligned}$$

or equivalently,

$$\begin{aligned}
\bar{\mathbf{y}}(t+1) - \bar{\mathbf{y}}(t) &= \alpha(t+1)(\bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t)) + (1-\alpha(t+1))\bar{\mathbf{e}}_y(t) + \alpha(t+1)\bar{\mathbf{e}}_v(t) \\
&\quad - \left(\frac{\alpha(t+1)}{\alpha(t)} + 1 - \alpha(t+1)\right)\nu(t)\bar{\mathbf{z}}(t) - \left(\frac{\alpha(t+1)}{\alpha(t)} + 1 - \alpha(t+1)\right)\bar{\mathbf{e}}_g(t). \qquad (5.49)
\end{aligned}$$

Thus, using Lemma 5.2 with $\theta = 1$, we can write

$$
\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{y}}(t+1)-\bar{\mathbf{y}}(t)\|^2\right] \\
&\le 4\alpha^2(t+1)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+4(1-\alpha(t+1))^2\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)\|^2\right]+4\alpha^2(t+1)\mathbb{E}\left[\|\bar{\mathbf{e}}_v(t)\|^2\right] \\
&\quad+4\left(\frac{\alpha(t+1)}{\alpha(t)}+1-\alpha(t+1)\right)^2\mathbb{E}\left[\|\nu(t)\bar{\mathbf{z}}(t)+\bar{\mathbf{e}}_g(t)\|^2\right] \\
&\overset{(a)}{\le} 4\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+4\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)\|^2\right]+4\alpha^2(t)\mathbb{E}\left[\|\bar{\mathbf{e}}_v(t)\|^2\right]+16\mathbb{E}\left[\|\nu(t)\bar{\mathbf{z}}(t)+\bar{\mathbf{e}}_g(t)\|^2\right] \\
&= 4\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+4\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)\|^2\right]+4\alpha^2(t)\mathbb{E}\left[\|\bar{\mathbf{e}}_v(t)\|^2\right] \\
&\quad+16\mathbb{E}\left[\|\bar{\mathbf{e}}_g(t)\|^2\right]+16\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]+32\nu(t)\mathbb{E}\left[\langle\bar{\mathbf{e}}_g(t),\bar{\mathbf{z}}(t)\rangle\right] \\
&= 4\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+4\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)\|^2\right]+4\alpha^2(t)\mathbb{E}\left[\|\bar{\mathbf{e}}_v(t)\|^2\right] \\
&\quad+16\mathbb{E}\left[\|\bar{\mathbf{e}}_g(t)\|^2\right]+16\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right], \qquad (5.50)
\end{aligned}
$$

where (a) holds since $\alpha(t+1) \le \alpha(t)$, which also implies $|\frac{\alpha(t+1)}{\alpha(t)}+1-\alpha(t+1)| < 2$ and the last step follows from

$$
\mathbb{E}\left[\langle\bar{\mathbf{e}}_g(t),\bar{\mathbf{z}}(t)\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle\bar{\mathbf{e}}_g(t),\bar{\mathbf{z}}(t)\rangle\right]|\mathcal{F}_t\right] = \mathbb{E}\left[\langle\mathbb{E}\left[\bar{\mathbf{e}}_g(t)|\mathcal{F}_t\right],\bar{\mathbf{z}}(t)\rangle\right] = 0.
$$

From Assumption 5.1, we have

$$
\begin{aligned}
\left[\mathbb{E}\left[E_y(t)E_y(t)^T|\mathcal{F}_t\right]\right]_{ij} &= \mathbb{E}\left[\mathbf{e}_{i,y}(t)\mathbf{e}_{j,y}^T(t)|\mathcal{F}_t\right] \\
&\le \frac{1}{2}\mathbb{E}\left[\|\mathbf{e}_{i,y}(t)\|^2|\mathcal{F}_t\right]+\frac{1}{2}\mathbb{E}\left[\|\mathbf{e}_{j,y}(t)\|^2|\mathcal{F}_t\right] \le \zeta_1\nu^2(t),
\end{aligned}
$$

for all $1 \le i,j \le n$. Hence, we get

$$
\begin{aligned}
\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)\|^2|\mathcal{F}_t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{1}^T E_y(t)\right\|^2\middle|\mathcal{F}_t\right]\right] \\
&= \frac{1}{n^2}\mathbf{1}^T\mathbb{E}\left[\mathbb{E}\left[E_y(t)E_y(t)^T|\mathcal{F}_t\right]\right]\mathbf{1} \le \frac{1}{n^2}\mathbf{1}^T(\zeta_1\nu^2(t)\mathbf{1}\mathbf{1}^T)\mathbf{1} \le \zeta_1\nu^2(t). \qquad (5.51)
\end{aligned}
$$

Similarly, from Assumption 5.1, we have

$$
\mathbb{E}\left[\|\bar{\mathbf{e}}_v(t)\|^2\right] \le \zeta_1\frac{\nu^2(t)}{\alpha^2(t)}, \quad \mathbb{E}\left[\|\bar{\mathbf{e}}_g(t)\|^2\right] \le \zeta_1\nu^2(t). \qquad (5.52)
$$

Plugging (5.51) and (5.52) into (5.50), we get

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{y}}(t+1)-\bar{\mathbf{y}}(t)\|^2\right] \\
&\leq 4\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+4\zeta_1\nu^2(t)+4\zeta_1\nu^2(t)+16\zeta_1\nu^2(t)+16\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right] \\
&= 4\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+24\zeta_1\nu^2(t)+16\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]. \qquad (5.53)
\end{aligned}$$

This together with (5.45) leads us to

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(Y(t+1))-\nabla f(Y(t))\|^2\right] \\
&\leq 3K^2\left(\mathbb{E}\left[\|\Delta Y(t+1)\|^2\right]+\mathbb{E}\left[\|\Delta Y(t)\|^2\right]\right) \\
&\quad +12K^2n\left(\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+6\zeta_1\nu^2(t)+4\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\right). \qquad (5.54)
\end{aligned}$$

Lastly, we can bound $\mathbb{E}\left[\|\Delta Y(t+1)\|^2\right]$ in (5.54) using the inequality in (5.36), and get

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(Y(t+1))-\nabla f(Y(t))\|^2\right] \\
&\leq 3K^2\left(2\mathbb{E}\left[\|\Delta Y(t)\|^2\right]+n\rho\zeta_1\nu^2(t)+\frac{1}{1-\sqrt{\rho}}\left(\alpha^2(t)\mathbb{E}\left[\|\Delta V(t)\|^2\right]+n\zeta_1\nu^2(t)\right)\right) \\
&\quad +\frac{3K^2\mathbb{E}\left[\|\Delta S(t)\|^2\right]}{\sqrt{\rho}(1-\sqrt{\rho})^2}+12K^2n\left(\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+6\zeta_1\nu^2(t)+4\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\right). \qquad (5.55)
\end{aligned}$$

where the last step holds for $\rho \leq 1$. This provides an upper for the first term in (5.44). Next, we can use Assumption 5.2-(b) to upper bound the second term in (5.44) as

$$\mathbb{E}\left[\|\nabla f(Y(t))\|^2\right]=\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla f_i(\mathbf{y}_i(t))\|^2\right]\leq n\zeta_2^2. \qquad (5.56)$$

Plugging (5.55) and (5.56) into (5.44), we get

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(\nu(t+1)\nabla f(Y(t+1))-\nu(t)\nabla f(Y(t))\right)\right\|^2\right] \\
&\leq 6K^2\nu^2(t+1)\left(2\mathbb{E}\left[\|\Delta Y(t)\|^2\right]+\left(\rho+\frac{1}{1-\sqrt{\rho}}\right)n\zeta_1\nu^2(t)+\frac{1}{1-\sqrt{\rho}}\alpha^2(t)\mathbb{E}\left[\|\Delta V(t)\|^2\right]\right) \\
&\quad +\frac{6K^2\nu^2(t+1)}{\sqrt{\rho}(1-\sqrt{\rho})^2}\mathbb{E}\left[\|\Delta S(t)\|^2\right]+2n\zeta_2^2\sigma^2\frac{\nu^2(t)}{(t+\tau)^2} \\
&\quad +24K^2n\nu^2(t+1)\left(\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+6\zeta_1\nu^2(t)+4\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\right). \qquad (5.57)
\end{aligned}$$

Plugging (5.43) and (5.57) into (5.37), we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\|\Delta S(t+1)\|^2\right] \\
&\leq \rho\mathbb{E}\left[\|\Delta S(t)\|^2\right] + \frac{6K^2}{1-\rho}\nu^2(t+1)\left(2\mathbb{E}\left[\|\Delta Y(t)\|^2\right] + \frac{\alpha^2(t)}{1-\sqrt{\rho}}\mathbb{E}\left[\|\Delta V(t)\|^2\right] + \frac{\mathbb{E}\left[\|\Delta S(t)\|^2\right]}{\sqrt{\rho}(1-\sqrt{\rho})^2}\right) \\
&\quad + 2n\rho\zeta_1\nu^2(t) + \frac{6K^2 n}{1-\rho}\nu^2(t+1)\left(\rho + \frac{1}{1-\sqrt{\rho}}\right)\zeta_1\nu^2(t) + \frac{144K^2}{1-\rho}n\nu^2(t+1)\zeta_1\nu^2(t) \\
&\quad + \frac{24K^2}{1-\rho}n\nu^2(t+1)\left(\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t)\|^2\right] + 4\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\right) + \frac{2}{1-\rho}n\zeta_2^2\sigma^2\frac{\nu^2(t)}{(t+\tau)^2}.
\end{aligned}$$

Recall that $\zeta_3 = \sup_{t\geq 0}\frac{\nu(t)}{\nu(t+1)}$. Then, we have $\frac{1}{\nu(t+1)} \leq \zeta_3\frac{1}{\nu(t)}$ for every $t \geq 0$. Therefore, we have

$$\begin{aligned}
&\frac{1}{\nu^2(t+1)}\mathbb{E}\left[\|\Delta S(t+1)\|^2\right] \\
&\leq \frac{\rho\zeta_3^2}{\nu^2(t)}\mathbb{E}\left[\|\Delta S(t)\|^2\right] + \frac{2n}{1-\sqrt{\rho}}\left(3K^2\left(25 + \frac{1}{1-\sqrt{\rho}}\right)\zeta_1\nu^2(t) + (\zeta_2\sigma\zeta_3)^2\frac{1}{(t+\tau)^2}\right) \\
&\quad + 2n\rho\zeta_1\zeta_3^2 + \frac{6K^2}{1-\sqrt{\rho}}\left(2\mathbb{E}\left[\|\Delta Y(t)\|^2\right] + \frac{\alpha^2(t)}{1-\sqrt{\rho}}\left(\mathbb{E}\left[\|\Delta V(t)\|^2\right] + \frac{1}{\sqrt{\rho}(1-\sqrt{\rho})^2}\mathbb{E}\left[\|\Delta S(t)\|^2\right]\right)\right. \\
&\quad + \frac{24K^2}{1-\sqrt{\rho}}n\left(\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t)\|^2\right] + 4\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\right),
\end{aligned} \tag{5.58}$$

where have used the fact that $\rho < 1$ to bound the fourth term. This provides the desired recursive bound for (a scaled version of) $\frac{1}{\nu^2(t+1)}\mathbb{E}\left[\|\Delta S(t+1)\|^2\right]$.

Now, having all the desired bounds available, we can rephrase them in the matrix form. To this end, we define

$$\begin{aligned}
&\boldsymbol{\Phi}(t) := \left[\alpha^2(t)\mathbb{E}\left[\|\Delta V(t)\|^2\right], \mathbb{E}\left[\|\Delta Y(t)\|^2\right], \frac{1}{\nu^2(t)}\mathbb{E}\left[\|\Delta S(t)\|^2\right]\right]^T, \\
&\mathbf{r}(t) := [0, 0, \kappa(t)]^T, \\
&\kappa(t) := K^2\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t)\|^2\right] + 4K^2\zeta_4\nu^2(t)\ \mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right] + \nu^2(t)\zeta_5 + \frac{1}{12}\frac{(\zeta_2\sigma\zeta_3)^2}{(t+\tau)^2}, \\
&\zeta_5 := \frac{1}{12}\left(3K^2\left(25 + \frac{1}{1-\sqrt{\rho}}\right)\zeta_1\right),
\end{aligned} \tag{5.59}$$

where $\zeta_4 > 1$ is given in (5.86). Thus, from (5.35), (5.36), and (5.58), we have the following linear time-varying system for every $t \geq t_0$

$$\boldsymbol{\Phi}(t+1) \leq A(\nu(t))\boldsymbol{\Phi}(t) + n\zeta_1\boldsymbol{b}(t) + \frac{24n}{1-\sqrt{\rho}}\mathbf{r}(t), \tag{5.60}$$

where $A(\nu)$ is defined as

$$A(\nu) := \begin{bmatrix} \rho & 0 & \frac{\nu^2}{1-\sqrt{\rho}} \\ \frac{1}{1-\sqrt{\rho}} & \rho & \frac{\nu^2}{\sqrt{\rho}(1-\sqrt{\rho})^2} \\ \frac{6K^2}{(1-\sqrt{\rho})^2} & \frac{12K^2}{1-\sqrt{\rho}} & c\rho + \frac{6K^2\nu^2}{\sqrt{\rho}(1-\sqrt{\rho})^3}. \end{bmatrix}, \tag{5.61}$$

with $c = \zeta_3^2$ and $\boldsymbol{b}(t) := \left[\rho\nu^2(t), \frac{2}{1-\sqrt{\rho}}\nu^2(t), 2\rho\zeta_3^2\right]^T$. We denote by $\mu(\nu)$ and $\boldsymbol{\Theta}(\nu) = [\Theta_1(\nu), \Theta_2(\nu), 1]^T$ the largest eigenvalue and the corresponding (normalized) eigenvector of the matrix $A(\nu)$, respectively. Lemma 5.6 presents some properties of $\mu(\nu)$ and $\boldsymbol{\Theta}(\nu)$. Next, having the matrix inequality in (5.60), we aim to show that

$$\boldsymbol{\Phi}(t) \le \frac{24n}{1-\sqrt{\rho}}\zeta_6\left(\kappa(t) + \zeta_1\right)\boldsymbol{\Theta}(\nu(t)), \tag{5.62}$$

for every $t \ge 0$, and some constant ζ_6 which will be determined later. We use induction to prove this claim. First note that for the induction base $t = 0$, we have $\boldsymbol{\Phi}(0) = 0$ and $\mathbf{r}(t) \ge 0$. Moreover, from Lemma 5.6, we have $\boldsymbol{\Theta}(\nu(0)) \ge 0$. Hence, (5.62) holds for $t = 0$.

Next, assume that (5.62) holds for some non-negative iteration t. For iteration $t+1$, starting from (5.60) and bounding $\boldsymbol{\Phi}(t)$ as in (5.62), we arrive at

$$\begin{aligned} \boldsymbol{\Phi}(t+1) &\le A(\nu(t))\frac{24n}{1-\sqrt{\rho}}\zeta_6\left(\kappa(t)+\zeta_1\right)\boldsymbol{\Theta}(\nu(t)) + n\zeta_1\boldsymbol{b}(t) + \frac{24n}{1-\sqrt{\rho}}\mathbf{r}(t) \\ &\overset{(a)}{=} \mu(\nu(t))\frac{24n}{1-\sqrt{\rho}}\zeta_6\left(\kappa(t)+\zeta_1\right)\boldsymbol{\Theta}(\nu(t)) + n\zeta_1\boldsymbol{b}(t) + \frac{24n}{1-\sqrt{\rho}}\mathbf{r}(t) \\ &\le \mu(\nu(t))\frac{24n}{1-\sqrt{\rho}}\zeta_6\left(\kappa(t)+\zeta_1\right)\boldsymbol{\Theta}(\nu(t)) + n\zeta_1\boldsymbol{b}(t) + \frac{24n}{1-\sqrt{\rho}}\kappa(t)\boldsymbol{\Theta}(\nu(t)). \end{aligned} \tag{5.63}$$

Here, (a) holds since $(\mu(\nu(t), \boldsymbol{\Theta}(\nu(t))))$ is an eigenpair of $A(\nu(t))$, and the last step follows from the fact that $\Theta_1(\nu(t)), \Theta_2(\nu(t)) \ge 0$ and $\Theta_3(\nu(t)) = 1$.

To bound the second term in (5.63), using $0 < \rho < \zeta_3^2\rho < 1$ and Lemma 5.7-(c), we get

$$\boldsymbol{b}(t) \le \frac{2}{1-\sqrt{\rho}}\begin{bmatrix}\nu^2(t)\\ \nu^2(t)\\ 1\end{bmatrix} < \frac{8}{1-\sqrt{\rho}}\boldsymbol{\Theta}(\nu(t)).$$

This combined with (5.63) and Corollary 5.5 with $\nu_1 = \nu(t+1)$, $\nu_2 = \nu(t)$, and $c := \zeta_3^2$ leads us to

$$\begin{aligned}\boldsymbol{\Phi}(t+1) &< \mu(\nu(t))\frac{24n}{1-\sqrt{\rho}}\zeta_6\left(\kappa(t)+\zeta_1\right)\boldsymbol{\Theta}(\nu(t)) + \frac{8n}{1-\sqrt{\rho}}\zeta_1\boldsymbol{\Theta}(\nu(t)) + \frac{24n}{1-\sqrt{\rho}}\kappa(t)\boldsymbol{\Theta}(\nu(t)) \\ &= \left(\mu(\nu(t))\frac{24n}{1-\sqrt{\rho}}\zeta_6\left(\kappa(t)+\zeta_1\right) + \frac{8n}{1-\sqrt{\rho}}\zeta_1 + \frac{24n}{1-\sqrt{\rho}}\kappa(t)\right)\boldsymbol{\Theta}(\nu(t)) \\ &< \frac{24n}{1-\sqrt{\rho}}\zeta_3^{\delta_2}\left(\left(\mu(\nu(t))\zeta_6+1\right)\kappa(t) + \left(\mu(\nu(t))\zeta_6+\frac{1}{3}\right)\zeta_1\right)\boldsymbol{\Theta}(\nu(t+1)), \end{aligned} \tag{5.64}$$

where

$$\delta_2 = \frac{32\zeta_3^2+16}{(\zeta_3^2-1)^{1/2}\rho^{1/2}}. \tag{5.65}$$

Now, we prove that $\kappa(t) \le \frac{\zeta_3^2}{\zeta_{11}}\kappa(t+1) + \frac{\zeta_3^2}{\zeta_{11}}K^2\zeta_1\zeta_4\zeta_{10}\nu_0^3$ for every $t \ge 0$ in which ζ_{11} will be determined later. From the definition $\kappa(t)$, we have

$$\begin{aligned}\kappa(t+1) = K^2\alpha^2(t+1)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t+1)-\bar{\mathbf{y}}(t+1)\|^2\right] + 4K^2\zeta_4\nu^2(t+1)\mathbb{E}\left[\|\bar{\mathbf{z}}(t+1)\|^2\right] + \nu^2(t+1)\zeta_5 \\ + \frac{1}{12}\frac{(\zeta_2\sigma\zeta_3)^2}{(t+\tau+1)^2}. \end{aligned} \tag{5.66}$$

To bound the second term in (5.66), subtracting (5.46) from (5.48), then taking expectations of both sides, we arrive at

$$\begin{aligned}&\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t+1)-\bar{\mathbf{y}}(t+1)\|^2\right] \\ &= \mathbb{E}\left[\left\|(1-\alpha(t+1))(\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)) + (1-\alpha(t+1))\left(1-\frac{1}{\alpha(t)}\right)\nu(t)\bar{\mathbf{z}}(t)\right\|^2\right] \\ &\quad + \mathbb{E}\left[\left\|(1-\alpha(t+1))(\bar{\mathbf{e}}_v(t)-\bar{\mathbf{e}}_y(t)) + (1-\alpha(t+1))\left(1-\frac{1}{\alpha(t)}\right)\bar{\mathbf{e}}_g(t)\right\|^2\right] \\ &\quad + (1-\alpha(t+1))^2\left(\mathbb{E}\left[\langle\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t), \bar{\mathbf{e}}_v(t)-\bar{\mathbf{e}}_y(t)\rangle\right] + \left(1-\frac{1}{\alpha(t)}\right)\mathbb{E}\left[\langle\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t), \bar{\mathbf{e}}_g(t)\rangle\right]\right) \\ &\quad + (1-\alpha(t+1))^2\left(1-\frac{1}{\alpha(t)}\right)\nu(t)\left(\mathbb{E}\left[\langle\bar{\mathbf{z}}(t), \bar{\mathbf{e}}_v(t)-\bar{\mathbf{e}}_y(t)\rangle\right] + \left(1-\frac{1}{\alpha(t)}\right)\mathbb{E}\left[\langle\bar{\mathbf{z}}(t), \bar{\mathbf{e}}_g(t)\rangle\right]\right). \end{aligned} \tag{5.67}$$

Using Assumption 5.1, we get

$$\begin{aligned}
&\mathbb{E}\left[\langle \bar{\boldsymbol{v}}(t), \bar{\mathbf{e}}_v(t)\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \bar{\boldsymbol{v}}(t), \bar{\mathbf{e}}_v(t)\rangle \,\middle|\, \mathcal{F}_t\right]\right] = \mathbb{E}\left[\langle \bar{\boldsymbol{v}}(t), \mathbb{E}\left[\bar{\mathbf{e}}_v(t)|\mathcal{F}_t\right]\rangle\right] = 0,\\
&\mathbb{E}\left[\langle \bar{\boldsymbol{v}}(t), \bar{\mathbf{e}}_y(t)\rangle\right] = 0,\\
&\mathbb{E}\left[\langle \bar{\boldsymbol{v}}(t), \bar{\mathbf{e}}_g(t)\rangle\right] = 0.
\end{aligned} \tag{5.68}$$

Similarly, we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\langle \bar{\mathbf{y}}(t), \bar{\mathbf{e}}_v(t)\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \bar{\mathbf{y}}(t), \bar{\mathbf{e}}_v(t)\rangle \,\middle|\, \mathcal{F}_t\right]\right] = \mathbb{E}\left[\langle \bar{\mathbf{y}}(t), \mathbb{E}\left[\bar{\mathbf{e}}_v(t)|\mathcal{F}_t\right]\rangle\right] = 0,\\
&\mathbb{E}\left[\langle \bar{\mathbf{y}}(t), \bar{\mathbf{e}}_y(t)\rangle\right] = 0,\\
&\mathbb{E}\left[\langle \bar{\mathbf{y}}(t), \bar{\mathbf{e}}_g(t)\rangle\right] = 0.
\end{aligned} \tag{5.69}$$

Finally, we have

$$\begin{aligned}
&\mathbb{E}\left[\langle \bar{\mathbf{z}}(t), \bar{\mathbf{e}}_v(t)\rangle\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\langle \nabla f_i(\mathbf{y}_i(t)), \bar{\mathbf{e}}_v(t)\rangle \,\middle|\, \mathcal{F}_t\right]\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle \nabla f_i(\mathbf{y}_i(t)), \mathbb{E}\left[\bar{\mathbf{e}}_v(t)\right]\rangle\right] = 0,\\
&\mathbb{E}\left[\langle \bar{\mathbf{z}}(t), \bar{\mathbf{e}}_y(t)\rangle\right] = 0\\
&\mathbb{E}\left[\langle \bar{\mathbf{z}}(t), \bar{\mathbf{e}}_g(t)\rangle\right] = 0.
\end{aligned} \tag{5.70}$$

Plugging (5.68), (5.69), and (5.70) into (5.67), we get

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t+1) - \bar{\mathbf{y}}(t+1)\|^2\right]\\
&= (1-\alpha(t+1))^2\mathbb{E}\left[\left\|\bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t) + \left(1 - \frac{1}{\alpha(t)}\right)\nu(t)\bar{\mathbf{z}}(t)\right\|^2\right]\\
&\quad + \mathbb{E}\left[\left\|(1-\alpha(t+1))(\bar{\mathbf{e}}_v(t) - \bar{\mathbf{e}}_y(t)) + (1-\alpha(t+1))\left(1 - \frac{1}{\alpha(t)}\right)\bar{\mathbf{e}}_g(t)\right\|^2\right]\\
&\geq (1-\alpha(t+1))^2\mathbb{E}\left[\left\|\bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t) + \left(1 - \frac{1}{\alpha(t)}\right)\nu(t)\bar{\mathbf{z}}(t)\right\|^2\right]\\
&\geq (1-\alpha(t+1))^2\left(\frac{1}{1+\zeta_7}\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t)\|^2\right] - \frac{1}{\zeta_7}\left(1-\alpha(t)\right)^2\frac{\nu^2(t)}{\alpha^2(t)}\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\right),
\end{aligned}$$

where the last step follows from Corollary 5.2 with $\theta = \zeta_7 > 0$ that will be determined

later. Hence, we can write

$$\begin{aligned}
&\alpha^2(t+1)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t+1)-\bar{\mathbf{y}}(t+1)\|^2\right]\\
&\geq \frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2\left(\frac{1}{1+\zeta_7}\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]-\frac{1}{\zeta_7}(1-\alpha(t))^2\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\right).
\end{aligned} \tag{5.71}$$

Moreover, from Corollary 5.2 with $\theta = \nu(t+1)$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\mathbf{z}}(t+1)\|^2\right] &= \mathbb{E}\left[\|\bar{\mathbf{z}}(t)+\bar{\mathbf{z}}(t+1)-\bar{\mathbf{z}}(t)\|^2\right]\\
&\geq \frac{1}{1+\nu(t+1)}\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]-\frac{1}{\nu(t+1)}\mathbb{E}\left[\|\bar{\mathbf{z}}(t+1)-\bar{\mathbf{z}}(t)\|^2\right].
\end{aligned} \tag{5.72}$$

Plugging (5.71) and (5.72) into (5.66), we arrive at

$$\begin{aligned}
\kappa(t+1) \geq{}& K^2\frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2\left(\frac{\alpha^2(t)}{1+\zeta_7}\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]-(1-\alpha(t))^2\frac{\nu^2(t)}{\zeta_7}\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\right)\\
&+\frac{4}{1+\nu(t+1)}K^2\zeta_4\nu^2(t+1)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]-4K^2\zeta_4\nu(t+1)\mathbb{E}\left[\|\bar{\mathbf{z}}(t+1)-\bar{\mathbf{z}}(t)\|^2\right]\\
&+\nu^2(t+1)\zeta_5+\frac{1}{12}\frac{(\zeta_2\sigma\zeta_3)^2}{(t+\tau+1)^2}.
\end{aligned} \tag{5.73}$$

Hence, we can write

$$\begin{aligned}
&\kappa(t)-\kappa(t+1)\\
&\leq\left(1-\frac{1}{1+\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2\right)K^2\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]\\
&+4\zeta_4\left(1+\frac{1}{4\zeta_4\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2(1-\alpha(t))^2-\frac{1}{1+\nu(t+1)}\frac{\nu^2(t+1)}{\nu^2(t)}\right)K^2\nu^2(t)\ \mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\\
&+4K^2\zeta_4\nu(t+1)\mathbb{E}\left[\|\bar{\mathbf{z}}(t+1)-\bar{\mathbf{z}}(t)\|^2\right]\\
&+\left(1-\frac{\nu^2(t+1)}{\nu^2(t)}\right)\nu^2(t)\zeta_5+\frac{1}{12}\frac{(\zeta_2\sigma\zeta_3)^2}{(t+\tau)^2}\left(1-\left(\frac{t+\tau}{t+\tau+1}\right)^2\right).
\end{aligned} \tag{5.74}$$

To bound the third term in (5.74), we get

$$
\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{z}}(t+1)-\bar{\mathbf{z}}(t)\|^2\right] \\
&\le 3\mathbb{E}\left[\|\bar{\mathbf{z}}(t+1)-\nabla f(\bar{\mathbf{y}}(t+1))\|^2\right]+3\mathbb{E}\left[\|\bar{\mathbf{z}}(t)-\nabla f(\bar{\mathbf{y}}(t))\|^2\right]+3\mathbb{E}\left[\|\nabla f(\bar{\mathbf{y}}(t+1))-\nabla f(\bar{\mathbf{y}}(t))\|^2\right] \\
&\overset{(a)}{\le} \frac{3K^2}{n}\left(\mathbb{E}\left[\|\Delta Y(t+1)\|^2\right]+\mathbb{E}\left[\|\Delta Y(t)\|^2\right]\right)+3K^2\mathbb{E}\left[\|\bar{\mathbf{y}}(t+1)-\bar{\mathbf{y}}(t)\|^2\right] \\
&\overset{(b)}{\le} \frac{3K^2}{n}\left(2\mathbb{E}\left[\|\Delta Y(t)\|^2\right]+n\rho\zeta_1\nu^2(t)+\frac{1}{1-\sqrt{\rho}}\alpha^2(t)\mathbb{E}\left[\|\Delta V(t)\|^2\right]+\frac{n}{1-\sqrt{\rho}}\zeta_1\nu^2(t)\right) \\
&\quad+\frac{3K^2}{n}\frac{\nu^2(t)}{\sqrt{\rho}(1-\sqrt{\rho})^2}\frac{1}{\nu^2(t)}\mathbb{E}\left[\|\Delta S(t)\|^2\right]+3K^2\mathbb{E}\left[\|\bar{\mathbf{y}}(t+1)-\bar{\mathbf{y}}(t)\|^2\right] \\
&\overset{(c)}{\le} \frac{72}{1-\sqrt{\rho}}\zeta_6\kappa(t)K^2\left(2\Theta_2(\nu(t))+\frac{1}{1-\sqrt{\rho}}\Theta_1(\nu(t))+\frac{\nu^2(t)}{\sqrt{\rho}(1-\sqrt{\rho})^2}\right) \\
&\quad+\frac{72}{1-\sqrt{\rho}}\zeta_6\zeta_1K^2\left(2\Theta_2(\nu(t))+\frac{1}{1-\sqrt{\rho}}\Theta_1(\nu(t))+\frac{\nu^2(t)}{\sqrt{\rho}(1-\sqrt{\rho})^2}\right) \\
&\quad+3K^2\left(\rho+\frac{1}{1-\sqrt{\rho}}\right)\zeta_1\nu^2(t)+3K^2\mathbb{E}\left[\|\bar{\mathbf{y}}(t+1)-\bar{\mathbf{y}}(t)\|^2\right], \qquad (5.75)
\end{aligned}
$$

where (a) follows from

$$
\begin{aligned}
\mathbb{E}\left[\|\bar{\mathbf{z}}(t)-\nabla f(\bar{\mathbf{y}}(t))\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\mathbf{y}_i(t))-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\bar{\mathbf{y}}(t))\right\|^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n\left(\nabla f_i(\mathbf{y}_i(t))-\nabla f_i(\bar{\mathbf{y}}(t))\right)\right\|^2\right] \\
&\le \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\|\nabla f_i(\mathbf{y}_i(t))-\nabla f_i(\bar{\mathbf{y}}(t))\|^2\right] \\
&\le \frac{K^2}{n}\sum_{i=1}^n\mathbb{E}\left[\|\mathbf{y}_i(t)-\bar{\mathbf{y}}(t)\|^2\right]=\frac{K^2}{n}\mathbb{E}\left[\|\Delta Y(t)\|^2\right].
\end{aligned}
$$

The step (b) holds due to (5.36) and in (c) we used the induction assumption (5.62). For the first term in (5.75), from Lemma 5.6 we have

$$
\begin{aligned}
&K^2\left(2\Theta_2(\nu(t))+\frac{1}{1-\sqrt{\rho}}\Theta_1(\nu(t))+\frac{\nu^2(t)}{\sqrt{\rho}(1-\sqrt{\rho})^2}\right) \\
&\le \frac{2K^2}{(\zeta_3^2-1)\rho(1-\sqrt{\rho})^2}\left(\frac{1}{(\zeta_3^2-1)\rho}+\frac{1}{\sqrt{\rho}}\right)\nu^2(t)+\frac{K^2}{(\zeta_3^2-1)\rho(1-\sqrt{\rho})^2}\nu^2(t)+\frac{K^2\nu^2(t)}{\sqrt{\rho}(1-\sqrt{\rho})^2} \\
&= \left(\frac{2}{(\zeta_3^2-1)\rho(1-\sqrt{\rho})^2}\left(\frac{1}{(\zeta_3^2-1)\rho}+\frac{1}{\sqrt{\rho}}\right)+\frac{1}{(\zeta_3^2-1)\rho(1-\sqrt{\rho})^2}+\frac{1}{\sqrt{\rho}(1-\sqrt{\rho})^2}\right)K^2\nu^2(t) \\
&:= \zeta_8\nu^2(t). \qquad (5.76)
\end{aligned}
$$

To bound the last term in (5.75) from (5.53), we can write

$$\mathbb{E}\left[\|\bar{\mathbf{y}}(t+1)-\bar{\mathbf{y}}(t)\|^2\right] \le 4\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]+24\zeta_1\nu^2(t)+16\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]$$
$$= \frac{4}{K^2}\kappa(t) - \frac{4}{K^2}\nu^2(t)\zeta_5+24\zeta_1\nu^2(t). \tag{5.77}$$

Plugging (5.76) and (5.77) into (5.75), we get

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{z}}(t+1)-\bar{\mathbf{z}}(t)\|^2\right]\\
&\le \frac{72}{1-\sqrt{\rho}}\zeta_6\zeta_8\nu^2(t)\kappa(t)+\frac{72}{1-\sqrt{\rho}}\zeta_1\zeta_6\zeta_8\nu^2(t)+3K^2\left(\rho+\frac{1}{1-\sqrt{\rho}}\right)\zeta_1\nu^2(t)\\
&\quad+12\kappa(t)-12\nu^2(t)\zeta_5+72K^2\zeta_1\nu^2(t)\\
&\le \frac{72}{1-\sqrt{\rho}}\zeta_6\zeta_8\nu_0^2\kappa(t)+\frac{72}{1-\sqrt{\rho}}\zeta_1\zeta_6\zeta_8\nu^2(t)+3K^2\left(\rho+\frac{1}{1-\sqrt{\rho}}\right)\zeta_1\nu^2(t)\\
&\quad+12\kappa(t)-12\nu^2(t)\zeta_5+72K^2\zeta_1\nu^2(t).
\end{aligned}$$

This together with (5.74) leads us to

$$\begin{aligned}
&\kappa(t)-\kappa(t+1)\\
&\le\left(1-\frac{1}{1+\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2\right)K^2\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]\\
&\quad+4\zeta_4\left(1+\frac{1}{2\zeta_4\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2(1-\alpha(t))^2-\frac{1}{1+\nu(t+1)}\frac{\nu^2(t+1)}{\nu^2(t)}\right)K^2\nu^2(t)\ \mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\\
&\quad+\left(1-\frac{\nu^2(t+1)}{\nu^2(t)}-48K^2\zeta_4\nu(t+1)\right)\nu^2(t)\zeta_5+\frac{1}{12}\frac{(\zeta_2\sigma\zeta_3)^2}{(t+\tau)^2}\left(1-\left(\frac{t+\tau}{t+\tau+1}\right)^2\right)\\
&\quad+K^2\zeta_4\left(\zeta_6\zeta_9\nu_0^2+48\right)\nu(t+1)\kappa(t)+K^2\zeta_1\zeta_4\zeta_{10}\nu^2(t)\nu(t+1),
\end{aligned}\tag{5.78}$$

where $\zeta_9 \coloneqq \frac{288}{1-\sqrt{\rho}}\zeta_8$ and

$$\zeta_{10} \coloneqq \zeta_6\zeta_9+288K^2+12K^2\left(\rho+\frac{1}{1-\sqrt{\rho}}\right). \tag{5.79}$$

For the first term in (5.78), from (5.7), we can write

$$\begin{aligned}
1-\frac{1}{1+\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2 &= 1-\frac{1}{1+\zeta_7}\frac{\nu(t+1)}{\nu(t)}(1-\alpha(t+1))^3\\
&\overset{(a)}{\le} 1-\frac{\nu(t+1)}{\nu(t)}\frac{(1-\alpha_0)^3}{1+\zeta_7}\\
&\le 1-\frac{\nu(t+1)}{\nu(t)}\zeta_{11},
\end{aligned}\tag{5.80}$$

where

$$\zeta_{11} \coloneqq \min\left(\frac{(1-\alpha_0)^3}{1+\zeta_7}, \frac{1}{\zeta_3+\nu_0} - \frac{1}{2\zeta_4\zeta_7}\right), \tag{5.81}$$

and (a) follows from Lemma 5.11-(a). Moreover, for the second term in (5.78), we get

$$\begin{aligned}
&1 + \frac{1}{2\zeta_4\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2(1-\alpha(t))^2 - \frac{1}{1+\nu(t+1)}\frac{\nu^2(t+1)}{\nu^2(t)} \\
&\le 1 + \frac{1}{2\zeta_4\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)(1-\alpha(t+1))} - \frac{1}{1+\nu(t+1)}\frac{\nu^2(t+1)}{\nu^2(t)} \\
&= 1 + \frac{1}{2\zeta_4\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)(1-\alpha(t+1))} - \frac{1}{\frac{\nu(t)}{\nu(t+1)}+\nu(t)}\frac{\nu(t+1)}{\nu(t)} \\
&\overset{(a)}{\le} 1 + \frac{1}{2\zeta_4\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)(1-\alpha(t+1))} - \frac{1}{\zeta_3+\nu_0}\frac{\nu(t+1)}{\nu(t)} \\
&= 1 - \frac{\nu(t+1)}{\nu(t)}\left(\frac{1}{\zeta_3+\nu_0} - \frac{1}{2\zeta_4\zeta_7}\right) \le 1 - \frac{\nu(t+1)}{\nu(t)}\zeta_{11},
\end{aligned} \tag{5.82}$$

where (a) follows from the fact that $\frac{\nu(t)}{\nu(t+1)} \le \zeta_3$ for all $t \ge 0$. Note that for the third term in (5.78), we have

$$\begin{aligned}
&1 - \frac{\nu^2(t+1)}{\nu^2(t)} - 48K^2\zeta_4\nu(t+1) \\
&\le 1 - \frac{1}{1+\nu(t+1)}\frac{\nu^2(t+1)}{\nu^2(t)} + \frac{1}{2\zeta_4\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2(1-\alpha(t))^2 \le 1 - \frac{\nu(t+1)}{\nu(t)}\zeta_{11}.
\end{aligned} \tag{5.83}$$

Finally, for the fourth term in (5.78) and $\sigma \in (\frac{1}{2}, 2)$, we can write

$$\begin{aligned}
1 - \left(\frac{t+\tau}{t+\tau+1}\right)^2 &\le 1 - \left(\frac{t+\tau}{t+\tau+1}\right)^{2\sigma} \\
&= 1 - \frac{\nu^2(t+1)}{\nu^2(t)} \\
&\le 1 - \frac{1}{1+\nu(t+1)}\frac{\nu^2(t+1)}{\nu^2(t)} + \frac{1}{2\zeta_4\zeta_7}\frac{\alpha^2(t+1)}{\alpha^2(t)}(1-\alpha(t+1))^2(1-\alpha(t))^2 \\
&\le 1 - \frac{\nu(t+1)}{\nu(t)}\zeta_{11}.
\end{aligned} \tag{5.84}$$

Using (5.78) and (5.80)-(5.84), we arrive at

$$\begin{aligned}
&\kappa(t)-\kappa(t+1)\\
&\leq \left(1-\frac{\nu(t+1)}{\nu(t)}\zeta_{11}+K^2\zeta_4\left(\zeta_6\zeta_9\nu_0^2+48\right)\nu(t+1)\right)\kappa(t)+K^2\zeta_1\zeta_4\zeta_{10}\nu^2(t)\nu(t+1),\\
&\overset{(a)}{\leq} \left(1-\frac{\zeta_{11}}{\zeta_3}+K^2\zeta_4\left(\zeta_6\zeta_9\nu_0^2+48\right)\nu_0\right)\kappa(t)+K^2\zeta_1\zeta_4\zeta_{10}\nu_0^3,\\
&\overset{(b)}{\leq} \left(1-\frac{\zeta_{11}}{\zeta_3^2}\right)\kappa(t)+K^2\zeta_1\zeta_4\zeta_{10}\nu_0^3,
\end{aligned}$$

where step (a) follows from $\frac{\nu(t)}{\nu(t+1)} \leq \zeta_3$ for all $t \geq 0$, and (b) holds due to condition (b)-(4). Thus, we can write

$$\kappa(t) \leq \frac{\zeta_3^2}{\zeta_{11}}\kappa(t+1)+\frac{\zeta_3^2}{\zeta_{11}}K^2\zeta_1\zeta_4\zeta_{10}\nu_0^3.$$

This combined with (5.64) arrives us at

$$\begin{aligned}
\mathbf{\Phi}(t+1) \leq{}& \frac{\zeta_3^{\delta_2+2}}{\zeta_{11}}\left(\mu(\nu(t))+\frac{1}{\zeta_6}\right)\frac{24n}{1-\sqrt{\rho}}\zeta_6\kappa(t+1)\mathbf{\Theta}(\nu(t+1))\\
&+\zeta_3^{\delta_2}\left(\frac{\zeta_3^2}{\zeta_{11}}\left(\mu(\nu(t))+\frac{1}{\zeta_6}\right)K^2\zeta_1\zeta_4\zeta_{10}\nu_0^3\right)+\mu(\nu(t))+\frac{1}{3\zeta_6}\Bigg)\frac{24n}{1-\sqrt{\rho}}\zeta_6\zeta_1\mathbf{\Theta}(\nu(t+1)).
\end{aligned} \tag{5.85}$$

Now, we choose

$$\zeta_4 = \frac{1}{2}\left(\frac{1}{\zeta_3+\nu_0}-\rho^{1/3}\right)^{-1}\left((1-\alpha_0)^3\rho^{-1/3}-1\right)^{-1}, \tag{5.86}$$

$$\zeta_6 = \rho^{-2/3}\left(\rho^{-1/6}-1\right)^{-1}, \tag{5.87}$$

$$\zeta_7 = (1-\alpha_0)^3\rho^{-1/3}-1. \tag{5.88}$$

From Corollary 5.4, we get

$$\begin{aligned}
\frac{\zeta_3^{\delta_2+2}}{\zeta_{11}}\left(\mu(\nu(t))+\frac{1}{\zeta_6}\right) &\leq \frac{\zeta_3^{\delta_2+2}}{\zeta_{11}}\left(\left(\zeta_3^2\rho\right)^{2/3}+\frac{1}{\zeta_6}\right)\\
&\overset{(a)}{\leq} \frac{\zeta_3^{\delta_2+4}}{\zeta_{11}}\left(\rho^{2/3}+\frac{1}{\zeta_6}\right)\overset{(b)}{\leq}\frac{\zeta_3^{\delta_2+4}}{\zeta_{11}}\rho^{1/2}\overset{(c)}{\leq}\frac{\rho^{1/3}}{\zeta_{11}}\overset{(d)}{=}1,
\end{aligned} \tag{5.89}$$

where (a) follows from $\zeta_3 \geq 1$, (b) holds due to (5.87), in step (c), we used Lemma 5.9 with $\delta_5 = \rho^{-1/6}-1$, and (d) is true from (5.86) and (5.88). Further, from (5.89), we can

write

$$\begin{aligned}
&\frac{\zeta_3^{\delta_2+2}}{\zeta_{11}}\left(\mu(\nu(t))+\frac{1}{\zeta_6}\right)K^2\zeta_1\zeta_4\zeta_{10}\nu^3(t_0)+\zeta_3^{\delta_2}\left(\mu(\nu(t))+\frac{1}{3\zeta_6}\right)\\
&\leq K^2\zeta_1\zeta_4\zeta_{10}\nu^3(t_0)+\zeta_3^{\delta_2}\left(\mu(\nu(t))+\frac{1}{3\zeta_6}\right)\\
&\overset{(a)}{\leq} K^2\zeta_1\zeta_4\zeta_{10}\nu^3(t_0)+\zeta_3^{\delta_2+2}\rho^{1/2}\\
&\overset{(b)}{\leq} K^2\zeta_1\zeta_4\zeta_{10}\nu^3(t_0)+\zeta_3^{-2}\rho^{1/3}\overset{(c)}{\leq}1,
\end{aligned}\tag{5.90}$$

where (a) follows from Corollary 5.4 and (5.87)

$$\mu(\nu(t))+\frac{1}{3\zeta_6}\leq\left(\zeta_3^2\rho\right)^{2/3}+\frac{1}{3\zeta_6}\leq\zeta_3^2\left(\rho^{2/3}+\frac{1}{3\zeta_6}\right)\leq\zeta_3^2\rho^{1/2},$$

step (b) holds due to Lemma 5.9 with $\delta_5=\rho^{-1/6}-1$, and (c) is true from condition (b)-(8). Plugging (5.89) and (5.90) into (5.85), we arrive at

$$\boldsymbol{\Phi}(t+1)\leq\frac{24n}{1-\sqrt{\rho}}\zeta_6\left(\kappa(t+1)+\zeta_1\right)\boldsymbol{\Theta}(\nu(t+1)).$$

Now, (5.62) is proven for $t+1$, and hence is true for all $t\geq 0$. Therefore, we have

$$\mathbb{E}\left[\|\Delta Y(t)\|^2\right]\leq\frac{24n}{1-\sqrt{\rho}}\zeta_6\left(\kappa(t)+\zeta_1\right)\Theta_2(\nu(t)).\tag{5.91}$$

5.9.2 Average Model's Loss Distance to the Optimal Loss

We define a series of Lyapunov functions $\Psi_t:\mathbb{R}^d\to\mathbb{R}$ for every $t\geq 0$ with $\Psi_0(\boldsymbol{\omega})=f(\bar{\mathbf{x}}(0))+\frac{\lambda_0}{2}\|\boldsymbol{\omega}-\bar{\boldsymbol{v}}(0)\|^2$ and

$$\Psi_{t+1}(\boldsymbol{\omega})=(1-\alpha(t))\Psi_t(\boldsymbol{\omega})+\alpha(t)\left(\hat{f}(t)+\langle\bar{\mathbf{z}}(t),\boldsymbol{\omega}-\bar{\mathbf{y}}(t)\rangle\right)+\frac{\alpha(t)}{\nu(t)}\langle\bar{\mathbf{e}}_g(t)-\alpha(t)\bar{\mathbf{e}}_v(t),\boldsymbol{\omega}-\bar{\mathbf{y}}(t)\rangle,\tag{5.92}$$

where $\lambda_0=\frac{\alpha_0^2}{\nu_0(1-\alpha_0)}$. Now, we consider a point $\mathbf{x}^\star$ in $\arg\min f(\mathbf{x})$, i.e., $f(\mathbf{x}^\star)=f^\star$. Taking expectations of both sides of (5.92) with $\boldsymbol{\omega}=\mathbf{x}^\star$, we have

$$\begin{aligned}
\mathbb{E}\left[\Psi_{t+1}(\mathbf{x}^\star)\right]&=(1-\alpha(t))\mathbb{E}\left[\Psi_t(\mathbf{x}^\star)\right]+\alpha(t)\mathbb{E}\left[\hat{f}(t)+\langle\bar{\mathbf{z}}(t),\mathbf{x}^\star-\bar{\mathbf{y}}(t)\rangle\right]\\
&\quad+\frac{\alpha(t)}{\nu(t)}\mathbb{E}\left[\langle\bar{\mathbf{e}}_g(t)-\alpha(t)\bar{\mathbf{e}}_v(t),\mathbf{x}^\star-\bar{\mathbf{y}}(t)\rangle\right].
\end{aligned}\tag{5.93}$$

We continue with bounding the first term in (5.93). Using Assumption 5.1, we have

$$\begin{aligned}
&\mathbb{E}\left[\langle \bar{\mathbf{e}}_g(t) - \alpha(t)\bar{\mathbf{e}}_v(t), \mathbf{x}^\star - \bar{\mathbf{y}}(t)\rangle\right] \\
&= \mathbb{E}\left[\langle \bar{\mathbf{e}}_g(t), \mathbf{x}^\star - \bar{\mathbf{y}}(t)\rangle\right] - \alpha(t)\mathbb{E}\left[\langle \bar{\mathbf{e}}_v(t), \mathbf{x}^\star - \bar{\mathbf{y}}(t)\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\langle \bar{\mathbf{e}}_g(t), \mathbf{x}^\star - \bar{\mathbf{y}}(t)\rangle \,|\mathcal{F}_t\right]\right] - \alpha(t)\mathbb{E}\left[\mathbb{E}\left[\langle \bar{\mathbf{e}}_v(t), \mathbf{x}^\star - \bar{\mathbf{y}}(t)\rangle \,|\mathcal{F}_t\right]\right] \\
&= \mathbb{E}\left[\langle \mathbb{E}\left[\bar{\mathbf{e}}_g(t)|\mathcal{F}_t\right], \mathbf{x}^\star - \bar{\mathbf{y}}(t)\rangle\right] - \alpha(t)\mathbb{E}\left[\langle \mathbb{E}\left[\bar{\mathbf{e}}_v(t)|\mathcal{F}_t\right], \mathbf{x}^\star - \bar{\mathbf{y}}(t)\rangle\right] = 0,
\end{aligned}$$

This together with (5.93) leads us to

$$\begin{aligned}
\mathbb{E}\left[\Psi_{t+1}(\mathbf{x}^\star)\right] &= (1-\alpha(t))\mathbb{E}\left[\Psi_t(\mathbf{x}^\star)\right] + \alpha(t)\mathbb{E}\left[\hat{f}(t) + \langle \bar{\mathbf{z}}(t), \mathbf{x}^\star - \bar{\mathbf{y}}(t)\rangle\right] \\
&\overset{(a)}{\leq} (1-\alpha(t))\mathbb{E}\left[\Psi_t(\mathbf{x}^\star)\right] + \alpha(t)\mathbb{E}\left[f^\star\right], \\
&= (1-\alpha(t))\mathbb{E}\left[\Psi_t(\mathbf{x}^\star)\right] + \alpha(t) f^\star, \qquad (5.94)
\end{aligned}$$

where (a) follows from (5.13). Next, we prove that

$$\mathbb{E}\left[\Psi_t(\mathbf{x}^\star)\right] \leq f^\star + \sigma(t)(\Psi_0(\mathbf{x}^\star) - f^\star), \qquad (5.95)$$

for every $t \geq 0$ where $\sigma(0) = 1$ and

$$\sigma(t+1) = (1-\alpha(t))\sigma(t). \qquad (5.96)$$

It is easy to verify that the inequality holds for $t = 0$. Assume it is true for t. For $t+1$, from (5.94), we have

$$\begin{aligned}
\mathbb{E}\left[\Psi_{t+1}(\mathbf{x}^\star)\right] &\leq (1-\alpha(t))f^\star + (1-\alpha(t))\sigma(t)(\Psi_0(\mathbf{x}^\star) - f^\star) + \alpha(t) f^\star \\
&= f^\star + \sigma(t+1)(\Psi_0(\mathbf{x}^\star) - f^\star).
\end{aligned}$$

Starting from (5.92), we note that

$$\nabla\Psi_{t+1}(\boldsymbol{\omega}) = (1-\alpha(t))\nabla\Psi_t(\boldsymbol{\omega}) + \alpha(t)\bar{\mathbf{z}}(t) + \frac{\alpha(t)}{\nu(t)}\left(\bar{\mathbf{e}}_g(t) - \alpha(t)\bar{\mathbf{e}}_v(t)\right), \qquad (5.97)$$

implying $\nabla^2\Psi_{t+1}(\boldsymbol{\omega}) = (1-\alpha(t))\nabla^2\Psi_t(\boldsymbol{\omega})$. This combined with $\nabla^2\Psi_0(\boldsymbol{\omega}) = \lambda_0 I$ leads us to $\nabla^2\Psi_t(\boldsymbol{\omega}) = \lambda(t) I$ for all $t \geq 0$ where

$$\lambda(t+1) = (1-\alpha(t))\lambda(t), \qquad (5.98)$$

and $\lambda(0) = \lambda_0$. Hence, $\Psi_t(\boldsymbol{\omega})$ is a quadratic function of $\boldsymbol{\omega}$. Now, by induction, we prove that

$$\nabla\Psi_t(\boldsymbol{\omega}) = \lambda(t)(\boldsymbol{\omega} - \bar{\boldsymbol{v}}(t)), \tag{5.99}$$

for every $t \geq 0$. First, from the definition of $\Psi_0(\boldsymbol{\omega})$, we have $\nabla\Psi_0(\boldsymbol{\omega}) = \lambda(0)(\boldsymbol{\omega} - \bar{\boldsymbol{v}}(0))$. Now, we assume that (5.99) is true for t. This combined with (5.97), leads us to

$$\begin{aligned}
\nabla\Psi_{t+1}(\boldsymbol{\omega}) &= (1-\alpha(t))\lambda(t)(\boldsymbol{\omega} - \bar{\boldsymbol{v}}(t)) + \alpha(t)\bar{\mathbf{z}}(t) + \frac{\alpha(t)}{\nu(t)}\left(\bar{\mathbf{e}}_g(t) - \alpha(t)\bar{\mathbf{e}}_v(t)\right) \\
&= \lambda(t+1)(\boldsymbol{\omega} - \bar{\boldsymbol{v}}(t+1) + \bar{\boldsymbol{v}}(t+1) - \bar{\boldsymbol{v}}(t)) + \alpha(t)\bar{\mathbf{z}}(t) + \frac{\alpha(t)}{\nu(t)}\left(\bar{\mathbf{e}}_g(t) - \alpha(t)\bar{\mathbf{e}}_v(t)\right) \\
&\overset{(a)}{=} \lambda(t+1)\left(\boldsymbol{\omega} - \bar{\boldsymbol{v}}(t+1) + \bar{\mathbf{e}}_v(t) - \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right) - \lambda(t+1)\frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t) + \alpha(t)\bar{\mathbf{z}}(t) \\
&\quad + \frac{\alpha(t)}{\nu(t)}\left(\bar{\mathbf{e}}_g(t) - \alpha(t)\bar{\mathbf{e}}_v(t)\right) \\
&\overset{(b)}{=} \lambda(t+1)(\boldsymbol{\omega} - \bar{\boldsymbol{v}}(t+1)),
\end{aligned} \tag{5.100}$$

where (a) follows from (5.46) and (b) holds due to $\alpha^2(t) = \nu(t)\lambda(t+1)$, which we will prove in the following using induction. $\alpha^2(t) = \nu(t)\lambda(t+1)$ is true for $t = 0$ by the definition of $\lambda(0)$. Additionally, from (5.7) note that

$$\alpha^2(t+1) = \frac{\nu(t+1)}{\nu(t)}(1-\alpha(t+1))\alpha^2(t) = \frac{\nu(t+1)}{\nu(t)}(1-\alpha(t+1))\nu(t)\lambda(t+1) = \nu(t+1)\lambda(t+2). \tag{5.101}$$

Now that (5.99) is proven, thus we can write

$$\Psi_t(\boldsymbol{\omega}) = \psi_t + \frac{\lambda(t)}{2}\left\|\boldsymbol{\omega} - \bar{\boldsymbol{v}}(t)\right\|^2. \tag{5.102}$$

Starting from (5.102) and using $\Psi_{t+1}(\bar{\mathbf{y}}(t)) = (1-\alpha(t))\Psi_t(\bar{\mathbf{y}}(t)) + \alpha(t)\hat{f}(t)$, we arrive at

$$\begin{aligned}
\psi_{t+1} &= \Psi_{t+1}(\bar{\mathbf{y}}(t)) - \frac{\lambda(t+1)}{2}\left\|\bar{\mathbf{y}}(t) - \bar{\boldsymbol{v}}(t+1)\right\|^2 \\
&= (1-\alpha(t))\psi_t + \alpha(t)\hat{f}(t) + (1-\alpha(t))\frac{\lambda(t)}{2}\left\|\bar{\mathbf{y}}(t) - \bar{\boldsymbol{v}}(t)\right\|^2 - \frac{\lambda(t+1)}{2}\left\|\bar{\mathbf{y}}(t) - \bar{\boldsymbol{v}}(t+1)\right\|^2 \\
&= (1-\alpha(t))\psi_t + \alpha(t)\hat{f}(t) + \frac{\lambda(t+1)}{2}\left(\left\|\bar{\mathbf{y}}(t) - \bar{\boldsymbol{v}}(t)\right\|^2 - \left\|\bar{\mathbf{y}}(t) - \bar{\boldsymbol{v}}(t+1)\right\|^2\right) \\
&\overset{(a)}{=} (1-\alpha(t))\psi_t + \alpha(t)\hat{f}(t) \\
&\quad + \frac{\lambda(t+1)}{2}\left(\left\|\bar{\mathbf{y}}(t) - \bar{\boldsymbol{v}}(t)\right\|^2 - \left\|\bar{\mathbf{y}}(t) - \bar{\boldsymbol{v}}(t) + \frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t) - \bar{\mathbf{e}}_v(t) + \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\|^2\right) \\
&= (1-\alpha(t))\psi_t + \alpha(t)\hat{f}(t) + \lambda(t+1)\left\langle \bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t), \frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t) - \bar{\mathbf{e}}_v(t) + \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\rangle \\
&\quad - \frac{\lambda(t+1)}{2}\left\|\frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t) - \bar{\mathbf{e}}_v(t) + \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\|^2, \qquad (5.103)
\end{aligned}$$

where (a) follows from (5.46). Using (5.68) and (5.69), we get

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t), \frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t) - \bar{\mathbf{e}}_v(t) + \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\rangle\right] \\
&= \frac{\nu(t)}{\alpha(t)}\mathbb{E}\left[\langle \bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t), \bar{\mathbf{z}}(t)\rangle\right] - \mathbb{E}\left[\langle \bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t), \bar{\mathbf{e}}_v(t)\rangle\right] + \frac{1}{\alpha(t)}\mathbb{E}\left[\langle \bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t), \bar{\mathbf{e}}_g(t)\rangle\right] \\
&= \frac{\nu(t)}{\alpha(t)}\mathbb{E}\left[\langle \bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t), \bar{\mathbf{z}}(t)\rangle\right]. \qquad (5.104)
\end{aligned}$$

Moreover, from (5.70), we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t) - \bar{\mathbf{e}}_v(t) + \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\|^2\right] &= \frac{\nu^2(t)}{\alpha^2(t)}\mathbb{E}\left[\left\|\bar{\mathbf{z}}(t)\right\|^2\right] + \mathbb{E}\left[\left\|\bar{\mathbf{e}}_v(t) - \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\|^2\right] \\
&\quad - 2\frac{\nu(t)}{\alpha(t)}\mathbb{E}\left[\langle \bar{\mathbf{z}}(t), \bar{\mathbf{e}}_v(t)\rangle\right] + 2\frac{\nu(t)}{\alpha^2(t)}\mathbb{E}\left[\langle \bar{\mathbf{z}}(t), \bar{\mathbf{e}}_g(t)\rangle\right] \\
&= \frac{\nu^2(t)}{\alpha^2(t)}\mathbb{E}\left[\left\|\bar{\mathbf{z}}(t)\right\|^2\right] + \mathbb{E}\left[\left\|\bar{\mathbf{e}}_v(t) - \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\|^2\right]. \qquad (5.105)
\end{aligned}$$

Using Lemma (5.2) with $\theta = 1$, for the second term of (5.105), we can write

$$\mathbb{E}\left[\left\|\bar{\mathbf{e}}_v(t) - \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\|^2\right] \le 2\mathbb{E}\left[\left\|\bar{\mathbf{e}}_v(t)\right\|^2\right] + \frac{2}{\alpha^2(t)}\mathbb{E}\left[\left\|\bar{\mathbf{e}}_g(t)\right\|^2\right] \le 4\zeta_1\frac{\nu^2(t)}{\alpha^2(t)}. \qquad (5.106)$$

where the last inequality follows from (5.52). Plugging (5.106) into (5.105), we arrive at

$$\mathbb{E}\left[\left\|\frac{\nu(t)}{\alpha(t)}\bar{\mathbf{z}}(t)-\bar{\mathbf{e}}_v(t)+\frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\|^2\right]\leq\frac{\nu^2(t)}{\alpha^2(t)}\left(\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]+4\zeta_1\right). \tag{5.107}$$

Taking expectations of both sides of (5.103) and using (5.104) and (5.107), we get

$$\begin{aligned}\mathbb{E}\left[\psi_{t+1}\right]\geq&(1-\alpha(t))\mathbb{E}\left[\psi_t\right]+\alpha(t)\mathbb{E}\left[\hat{f}(t)\right]+\lambda(t+1)\frac{\nu(t)}{\alpha(t)}\mathbb{E}\left[\langle\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t),\bar{\mathbf{z}}(t)\rangle\right]\\&-\frac{\lambda(t+1)}{2}\frac{\nu^2(t)}{\alpha^2(t)}\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]-2\zeta_1\lambda(t+1)\frac{\nu^2(t)}{\alpha^2(t)}\\=&(1-\alpha(t))\mathbb{E}\left[\psi_t\right]+\alpha(t)\mathbb{E}\left[\hat{f}(t)+\langle\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t),\bar{\mathbf{z}}(t)\rangle\right]-\frac{1}{2}\nu(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]-2\zeta_1\nu(t),\end{aligned} \tag{5.108}$$

where the last step holds due to $\alpha^2(t)=\nu(t)\lambda(t+1)$. We continue with bounding the second term in (5.108). Starting from (5.23), we can write

$$\begin{aligned}\alpha(t)\left\langle\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t),\bar{\mathbf{z}}(t)\right\rangle&=-(1-\alpha(t))\left\langle\bar{\mathbf{x}}(t)-\bar{\mathbf{y}}(t),\bar{\mathbf{z}}(t)\right\rangle\\&\geq(1-\alpha(t))(\hat{f}(t)-f(\bar{\mathbf{x}}(t))).\end{aligned} \tag{5.109}$$

where the last inequality follows from (5.13). Plugging (5.109) into (5.108), we arrive at

$$\mathbb{E}\left[\psi_{t+1}\right]\geq(1-\alpha(t))\mathbb{E}\left[\psi_t-f(\bar{\mathbf{x}}(t))\right]+\mathbb{E}\left[\hat{f}(t)\right]-\frac{1}{2}\nu(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]-2\zeta_1\nu(t). \tag{5.110}$$

From (5.47) and (5.70), we arrive at

$$\begin{aligned}\mathbb{E}\left[\langle\bar{\mathbf{x}}(t+1)-\bar{\mathbf{y}}(t),\bar{\mathbf{z}}(t)\rangle\right]&=-\nu(t)\mathbb{E}\left[\langle\bar{\mathbf{z}}(t),\bar{\mathbf{z}}(t)\rangle\right]+\mathbb{E}\left[\langle\bar{\mathbf{e}}_y(t),\bar{\mathbf{z}}(t)\rangle\right]-\mathbb{E}\left[\langle\bar{\mathbf{e}}_g(t),\bar{\mathbf{z}}(t)\rangle\right]\\&=-\nu(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right].\end{aligned} \tag{5.111}$$

Moreover, we can write

$$\begin{aligned}\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1)-\bar{\mathbf{y}}(t)\|^2\right]&=\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]+\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)-\bar{\mathbf{e}}_g(t)\|^2\right]\\&\overset{(a)}{\leq}\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]+2\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)\|^2\right]+2\mathbb{E}\left[\|\bar{\mathbf{e}}_g(t)\|^2\right]\\&\overset{(b)}{\leq}\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]+4\zeta_1\nu^2(t),\end{aligned} \tag{5.112}$$

where in (a) we used Lemma (5.2) with $\theta = 1$ and step (b) holds due to (5.51) and (5.52). From (5.14) with $\boldsymbol{\omega} = \bar{\mathbf{x}}(t+1)$, we get

$$f(\bar{\mathbf{x}}(t+1)) \le \hat{f}(t) + \langle \bar{\mathbf{x}}(t+1) - \bar{\mathbf{y}}(t), \bar{\mathbf{z}}(t)\rangle + K\|\bar{\mathbf{x}}(t+1) - \bar{\mathbf{y}}(t)\|^2 + \frac{K}{n}\|Y(t) - \mathbf{1}\bar{\mathbf{y}}(t)\|^2, \tag{5.113}$$

Taking expectation of both sides of (5.113), we have

$$\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}(t+1))] \\
&\le \mathbb{E}\left[\hat{f}(t)\right] + \mathbb{E}\left[\langle \bar{\mathbf{x}}(t+1) - \bar{\mathbf{y}}(t), \bar{\mathbf{z}}(t)\rangle\right] + K\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1) - \bar{\mathbf{y}}(t)\|^2\right] + \frac{K}{n}\mathbb{E}\left[\|\Delta Y(t)\|^2\right] \\
&\le \mathbb{E}\left[\hat{f}(t)\right] - \nu(t)(1 - \nu(t)K)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right] + \frac{K}{n}\mathbb{E}\left[\|\Delta Y(t)\|^2\right] + 4\zeta_1 K\nu^2(t),
\end{aligned} \tag{5.114}$$

where the last inequality follows from (5.111) and (5.112). Finally, subtracting (5.110) from (5.114), we arrive at

$$\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}(t+1)) - \psi_{t+1}] \\
&\le (1-\alpha(t))\mathbb{E}[f(\bar{\mathbf{x}}(t)) - \psi_t] - \nu(t)\left(\frac{1}{2} - \nu(t)K\right)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right] + \frac{K}{n}\mathbb{E}\left[\|\Delta Y(t)\|^2\right] + 2\zeta_1\nu(t) \\
&\quad + 4\zeta_1 K\nu^2(t).
\end{aligned}$$

This combined with (5.91) leads us to

$$\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}(t+1)) - \psi_{t+1}] \\
&\le (1-\alpha(t))\mathbb{E}[f(\bar{\mathbf{x}}(t)) - \psi_t] - \nu(t)\left(\frac{1}{2} - \nu(t)K - \frac{96}{1-\sqrt{\rho}}K^3\zeta_4\zeta_6\nu(t)\Theta_2(\nu(t))\right)\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right] \\
&\quad + \frac{24}{1-\sqrt{\rho}}\zeta_6 K^3\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t) - \bar{\mathbf{y}}(t)\|^2\right]\Theta_2(\nu(t)) + K\left(4\zeta_1 + \frac{24}{1-\sqrt{\rho}}\zeta_5\zeta_6\Theta_2(\nu(t))\right)\nu^2(t) \\
&\quad + \frac{24}{1-\sqrt{\rho}}K\zeta_1\zeta_6\Theta_2(\nu(t)) + 2\zeta_1\nu(t).
\end{aligned} \tag{5.115}$$

From Lemma 5.6 and conditions (b)-(3),(9),(10) we get

$$\begin{aligned}
&\frac{96}{1-\sqrt{\rho}}K^3\zeta_4\zeta_6\nu(t)\Theta_2(\nu(t)) \le \frac{96K^3\zeta_4\zeta_6}{(\zeta_3^2-1)\rho(1-\sqrt{\rho})^3}\left(\frac{1}{(\zeta_3^2-1)\rho} + \frac{1}{\sqrt{\rho}}\right)\nu^3(t) \le \frac{1}{8}, \\
&\frac{24}{1-\sqrt{\rho}}\zeta_5\zeta_6\Theta_2(\nu(t)) \le \frac{24\zeta_5\zeta_6}{(\zeta_3^2-1)\rho(1-\sqrt{\rho})^3}\left(\frac{1}{(\zeta_3^2-1)\rho} + \frac{1}{\sqrt{\rho}}\right)\nu^2(t) \le \zeta_1, \\
&K\nu(t) \le \frac{1}{8}.
\end{aligned}$$

for every $t \geq 0$. This together with (5.115) arrives us at

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{\mathbf{x}}(t+1))-\psi_{t+1}\right] \\
&\leq (1-\alpha(t))\mathbb{E}\left[f(\bar{\mathbf{x}}(t))-\psi_t\right]-\frac{\nu(t)}{4}\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]+\frac{24}{1-\sqrt{\rho}}\zeta_6 K^3\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]\Theta_2(\nu(t)) \\
&\quad + 5\zeta_1 K\nu^2(t)+\frac{24}{1-\sqrt{\rho}}K\zeta_1\zeta_6\Theta_2(\nu(t))+2\zeta_1\nu(t) \\
&\leq (1-\alpha(t))\mathbb{E}\left[f(\bar{\mathbf{x}}(t))-\psi_t\right]+\frac{24}{1-\sqrt{\rho}}\zeta_6 K^3\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]\Theta_2(\nu(t))+5\zeta_1 K\nu^2(t) \\
&\quad + \frac{24}{1-\sqrt{\rho}}K\zeta_1\zeta_6\Theta_2(\nu(t))+2\zeta_1\nu(t). \qquad (5.116)
\end{aligned}$$

Moreover, from Lemma 5.6-(c), we reduce (5.116) to

$$\begin{aligned}
\mathbb{E}\left[f(\bar{\mathbf{x}}(t+1))-\psi_{t+1}\right] &\leq (1-\alpha(t))\mathbb{E}\left[f(\bar{\mathbf{x}}(t))-\psi_t\right]+\zeta_{12}\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right]\nu^2(t) \\
&\quad + \zeta_1\zeta_{13}\nu^2(t)+2\zeta_1\nu(t), \qquad (5.117)
\end{aligned}$$

where $\zeta_{12} \coloneqq \frac{24K^3\zeta_6}{(\zeta_3^2-1)\rho(1-\sqrt{\rho})^3}\left(\frac{1}{(\zeta_3^2-1)\rho}+\frac{1}{\sqrt{\rho}}\right)$ and

$$\zeta_{13} \coloneqq 5K+\frac{1}{K^2}\zeta_{12}. \qquad (5.118)$$

Using (5.23), we obtain

$$\alpha^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\bar{\mathbf{y}}(t)\|^2\right] = (1-\alpha(t))^2\,\mathbb{E}\left[\|\bar{\mathbf{y}}(t)-\bar{\mathbf{x}}(t)\|^2\right] \leq \mathbb{E}\left[\|\bar{\mathbf{y}}(t)-\bar{\mathbf{x}}(t)\|^2\right]. \qquad (5.119)$$

Plugging (5.119) into (5.117), we get

$$\begin{aligned}
\mathbb{E}\left[f(\bar{\mathbf{x}}(t+1))-\psi_{t+1}\right] &\leq (1-\alpha(t))\mathbb{E}\left[f(\bar{\mathbf{x}}(t))-\psi_t\right]+\zeta_{12}\mathbb{E}\left[\|\bar{\mathbf{y}}(t)-\bar{\mathbf{x}}(t)\|^2\right]\nu^2(t) \\
&\quad + \zeta_1\zeta_{13}\nu^2(t)+2\zeta_1\nu(t), \qquad (5.120)
\end{aligned}$$

Next, expanding (5.120) recursively, we get

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{\mathbf{x}}(t+1))-\psi_{t+1}\right] \\
&\leq \prod_{k=0}^{t}(1-\alpha(k))\left(\mathbb{E}\left[f(\bar{\mathbf{x}}(0))-\psi_0\right]\right) \\
&\quad + \sum_{k=0}^{t}\left(\zeta_{12}\mathbb{E}\left[\|\bar{\mathbf{y}}(k)-\bar{\mathbf{x}}(k)\|^2\right]\nu^2(k)+\zeta_1\zeta_{13}\nu^2(k)+2\zeta_1\nu(k)\right)\prod_{\ell=k+1}^{t}(1-\alpha(\ell))\,. \qquad (5.121)
\end{aligned}$$

For the first term in (5.121), from (5.102) we get

$$\begin{aligned}\psi_0 &= \Psi_0(\boldsymbol{\omega}) - \frac{\lambda(0)}{2}\|\boldsymbol{\omega}-\bar{\boldsymbol{v}}(0)\|^2 \\ &\overset{(a)}{=} f(\bar{\mathbf{x}}(0)) + \frac{\lambda_0}{2}\|\boldsymbol{\omega}-\bar{\boldsymbol{v}}(0)\|^2 - \frac{\lambda(0)}{2}\|\boldsymbol{\omega}-\bar{\boldsymbol{v}}(0)\|^2 \overset{(b)}{=} f(\bar{\mathbf{x}}(0)).\end{aligned} \tag{5.122}$$

where (a) follows from the definition $\Psi_t(\boldsymbol{\omega})$ and (b) holds due to $\lambda(0) = \lambda_0$. Plugging (5.122) into (5.121), we arrive at

$$\begin{aligned}&\mathbb{E}\left[f(\bar{\mathbf{x}}(t+1)) - \psi_{t+1}\right] \\ &\leq \sum_{k=0}^{t}\left(\zeta_{12}\mathbb{E}\left[\|\bar{\mathbf{y}}(k)-\bar{\mathbf{x}}(k)\|^2\right]\nu^2(k) + \zeta_1\zeta_{13}\nu^2(k) + 2\zeta_1\nu(k)\right)\prod_{\ell=k+1}^{t}(1-\alpha(\ell)),\end{aligned} \tag{5.123}$$

Using Lemma 5.11-(c) with $\nu(t) = \frac{\nu_0}{(t+\tau)^\sigma}$, for the last term in (5.123), we can write

$$\begin{aligned}\sum_{k=0}^{t}\nu(k)\prod_{\ell=k+1}^{t}(1-\alpha(\ell)) &= \sum_{k=0}^{t}\nu(k)\frac{\prod_{\ell=0}^{t}(1-\alpha(\ell))}{\prod_{s=0}^{k}(1-\alpha(s))} \\ &\overset{(a)}{=} \sum_{k=0}^{t}\nu(k)\frac{\prod_{\ell=0}^{t}\frac{\lambda(\ell+1)}{\lambda(\ell)}}{\prod_{s=0}^{k}\frac{\lambda(s+1)}{\lambda(s)}} \\ &= \sum_{k=0}^{t}\nu(k)\frac{\frac{\lambda(t+1)}{\lambda(0)}}{\frac{\lambda(k+1)}{\lambda(0)}} \\ &= \sum_{k=0}^{t}\nu(k)\frac{\lambda(t+1)}{\lambda(k+1)} \\ &\overset{(b)}{=} \lambda(t+1)\sum_{k=0}^{t}\frac{\nu^2(k)}{\alpha^2(k)} \\ &\overset{(c)}{\leq} \frac{\nu_0^2(4+\alpha_0)^2}{16(2-\sigma)^2}\lambda(t+1)\left(\sum_{k=0}^{t}(k+\tau)^{-2(\sigma-1)}\right) \\ &\leq \zeta_{14}\lambda(t+1),\end{aligned} \tag{5.124}$$

where

$$\zeta_{14} := \frac{\nu_0^2(4+\alpha_0)^2}{16(2-\sigma)^2}\sum_{k=0}^{\infty}(k+\tau)^{-2(\sigma-1)}, \tag{5.125}$$

is some constant for any $2(\sigma-1) > 1$, i.e., $\sigma \in (3/2, 2)$. Note that in the chain of inequalities, (a) follows from (5.98), step (b) holds due to (5.101), and in (c) we used

Lemma 5.11-(c). Similarly, we arrive at

$$\sum_{k=0}^{t}\nu^2(k)\prod_{\ell=k+1}^{t}(1-\alpha(\ell))\le\frac{\nu_0^3(4+\alpha_0)^2}{16(2-\sigma)^2}\lambda(t+1)\left(\sum_{k=0}^{t}(k+\tau)^{-(3\sigma-2)}\right)\le\zeta_{15}\lambda(t+1),\quad(5.126)$$

where

$$\zeta_{15}:=\frac{\nu_0^3(4+\alpha_0)^2}{16(2-\sigma)^2}\sum_{k=0}^{\infty}(k+\tau)^{-(3\sigma-2)}\quad(5.127)$$

is some constant for $\sigma\in(1,2)$. Plugging (5.124) and (5.126) into (5.123), we arrive at

$$\begin{aligned}&\mathbb{E}\left[f(\bar{\mathbf{x}}(t+1))-\psi_{t+1}\right]\\&\le\zeta_{12}\sum_{k=0}^{t}\mathbb{E}\left[\|\bar{\mathbf{y}}(k)-\bar{\mathbf{x}}(k)\|^2\right]\nu^2(k)\prod_{\ell=k+1}^{t}(1-\alpha(\ell))+\zeta_1\lambda(t+1)\left(2\zeta_{14}+\zeta_{13}\zeta_{15}\right),\end{aligned}\quad(5.128)$$

Next, having (5.128), we aim to show that

$$\mathbb{E}\left[f(\bar{\mathbf{x}}(t))-\psi_t\right]\le\zeta_{16}\lambda(t)+\zeta_1\lambda(t)\left(2\zeta_{14}+\zeta_{13}\zeta_{15}\right),\quad(5.129)$$

for every $t\ge0$ where ζ_{16} is given in (5.149). We use induction to prove this claim. Using (5.122) and the fact that $\zeta_1,\zeta_{14},\zeta_{15},\zeta_{16}>0$, we get (5.129) for $t=0$. Suppose that (5.129) is true for t. From (5.95) and (5.102), we have

$$\mathbb{E}\left[\psi_t\right]+\frac{\lambda(t)}{2}\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\mathbf{x}^\star\|^2\right]\le f^\star+\sigma(t)(\Psi_0(\mathbf{x}^\star)-f^\star).$$

This combined with the induction assumption leads us to

$$\begin{aligned}&\mathbb{E}\left[f(\bar{\mathbf{x}}(t))\right]+\frac{\lambda(t)}{2}\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\mathbf{x}^\star\|^2\right]\\&\le\zeta_{16}\lambda(t)+\zeta_1\lambda(t)\left(2\zeta_{14}+\zeta_{13}\zeta_{15}\right)+f^\star+\sigma(t)(\Psi_0(\mathbf{x}^\star)-f^\star).\end{aligned}\quad(5.130)$$

From the fact that $\mathbb{E}\left[f(\bar{\mathbf{x}}(t))-f^\star\right]\ge0$ and (5.130), we arrive at

$$\begin{aligned}\nu^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t)-\mathbf{x}^\star\|^2\right]&\le\frac{2\sigma(t)}{\lambda(t)}\nu^2(t)(\Psi_0(\mathbf{x}^\star)-f^\star)+\frac{2\nu^2(t)}{\lambda(t)}\left(\zeta_{16}\lambda(t)+\zeta_1\lambda(t)\left(2\zeta_{14}+\zeta_{13}\zeta_{15}\right)\right)\\&=\frac{2\sigma(t)}{\lambda(t)}\nu^2(t)(\Psi_0(\mathbf{x}^\star)-f^\star)+2\nu^2(t)\left(\zeta_{16}+\zeta_1\left(2\zeta_{14}+\zeta_{13}\zeta_{15}\right)\right)\\&\le\frac{2\sigma(t)}{\lambda(t)}\nu^2(t)(\Psi_0(\mathbf{x}^\star)-f^\star)+2\nu_0^2\left(\zeta_{16}+\zeta_1\left(2\zeta_{14}+\zeta_{13}\zeta_{15}\right)\right):=\zeta_{17},\end{aligned}\quad(5.131)$$

where the last step follows from the fact that $\{\nu(t)\}$ is a non-increasing sequence. From (5.23), we get $\bar{\boldsymbol{v}}(t) = \frac{1}{\alpha(t)}(\bar{\mathbf{y}}(t) - \bar{\mathbf{x}}(t)) + \bar{\mathbf{x}}(t)$ implying

$$\begin{aligned}\nu^2(t)\,\|\bar{\boldsymbol{v}}(t) - \mathbf{x}^\star\|^2 &= \nu^2(t)\left\|\frac{1}{\alpha(t)}(\bar{\mathbf{y}}(t) - \bar{\mathbf{x}}(t)) + \bar{\mathbf{x}}(t) - \mathbf{x}^\star\right\|^2 \\ &\geq \frac{\nu^2(t)}{2\alpha^2(t)}\|\bar{\mathbf{y}}(t) - \bar{\mathbf{x}}(t)\|^2 - \nu^2(t)\,\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2. \end{aligned} \tag{5.132}$$

where the last inequality follows from Corollary 5.2 with $\theta = 1$. Using Jensen's inequality, for $\sigma \in (1/2, 2)$ we get

$$\begin{aligned} f(\mathbb{E}\,[\bar{\mathbf{x}}(t)]) &\leq \mathbb{E}\,[f(\bar{\mathbf{x}}(t))] \\ &\overset{(a)}{\leq} \zeta_{16}\lambda(t) + \zeta_1\lambda(t)\,(2\zeta_{14} + \zeta_{13}\zeta_{15}) + f^\star + \sigma(t)\left(f(\bar{\mathbf{x}}(0)) + \frac{\lambda_0}{2}\|\mathbf{x}^\star - \bar{\boldsymbol{v}}(0)\|^2 - f^\star\right) \\ &\overset{(b)}{\leq} \zeta_{16}\lambda(t) + \zeta_1\lambda(t)\,(2\zeta_{14} + \zeta_{13}\zeta_{15}) + f^\star + \sigma(0)\left(f(\bar{\mathbf{x}}(0)) + \frac{\lambda_0}{2}\|\mathbf{x}^\star - \bar{\boldsymbol{v}}(0)\|^2 - f^\star\right) \\ &\overset{(c)}{=} \zeta_{16}\lambda(t) + \zeta_1\lambda(t)\,(2\zeta_{14} + \zeta_{13}\zeta_{15}) + f(\bar{\mathbf{x}}(0)) + \frac{\lambda_0}{2}\|\mathbf{x}^\star - \bar{\boldsymbol{v}}(0)\|^2 := \zeta_{18}, \end{aligned}$$

where (a) follows from (5.130) and $\Psi_0(\mathbf{x}^\star) = f(\bar{\mathbf{x}}(0)) + \frac{\lambda_0}{2}\|\mathbf{x}^\star - \bar{\boldsymbol{v}}(0)\|^2$, step (b) holds due to the fact that $\{\sigma(t)\}$ is non-increasing and $f(\bar{\mathbf{x}}(0)) \geq f^\star$. Finally, in (c) we used $\sigma(0) = 1$. Thus, we have $\mathbb{E}\,[\bar{\mathbf{x}}(t)]$ lies within the ζ_{18}-level set of f. By Assumption 5.2-(c) and Proposition B.9 in [154], the level set is compact. Thus, we get $\|\mathbb{E}\,[\bar{\mathbf{x}}(t) - \mathbf{x}^\star]\| \leq R$ where R is the diameter of that level set. This combined with Lemma 5.2 with $\theta = 1$ leads us to

$$\begin{aligned}\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2\right] &= \nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbb{E}\,[\bar{\mathbf{x}}(t)] + \mathbb{E}\,[\bar{\mathbf{x}}(t)] - \mathbf{x}^\star\|^2\right] \\ &\leq 2\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbb{E}\,[\bar{\mathbf{x}}(t)]\,\|^2\right] + 2\nu^2(t)\,\|\mathbb{E}\,[\bar{\mathbf{x}}(t) - \mathbf{x}^\star]\|^2 \\ &\leq 2\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbb{E}\,[\bar{\mathbf{x}}(t)]\,\|^2\right] + 2\nu_0^2 R^2. \end{aligned} \tag{5.133}$$

Now, we show that $\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{x}}(t)-\mathbb{E}\left[\bar{\mathbf{x}}(t)\right]\|^2\right]<\infty$. To this end, starting from (5.47), we get

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1)-\mathbb{E}\left[\bar{\mathbf{x}}(t+1)\right]\|^2\Big|\mathcal{F}_t\right]\\
&=\mathbb{E}\left[\|\bar{\mathbf{y}}(t)-\mathbb{E}\left[\bar{\mathbf{y}}(t)\right]-\nu(t)\left(\bar{\mathbf{z}}(t)-\mathbb{E}\left[\bar{\mathbf{z}}(t)\right]\right)+\bar{\mathbf{e}}_y(t)-\bar{\mathbf{e}}_g(t)\|^2\Big|\mathcal{F}_t\right]\\
&\overset{(a)}{=}\|\bar{\mathbf{y}}(t)-\mathbb{E}\left[\bar{\mathbf{y}}(t)\right]-\nu(t)\left(\bar{\mathbf{z}}(t)-\mathbb{E}\left[\bar{\mathbf{z}}(t)\right]\right)\|^2+\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)-\bar{\mathbf{e}}_g(t)\|^2\Big|\mathcal{F}_t\right]\\
&\leq(1+\nu(t))\|\bar{\mathbf{y}}(t)-\mathbb{E}\left[\bar{\mathbf{y}}(t)\right]\|^2+\left(1+\frac{1}{\nu(t)}\right)\nu^2(t)\|\bar{\mathbf{z}}(t)-\mathbb{E}\left[\bar{\mathbf{z}}(t)\right]\|^2+\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)-\bar{\mathbf{e}}_g(t)\|^2\Big|\mathcal{F}_t\right],
\end{aligned}\tag{5.134}$$

where (a) follows from (5.69), (5.70), and the last step holds due Lemma 5.2 with $\theta=\nu(t)$. For the second term in (5.134), using Lemma 5.2 with $\theta=1$, we can write

$$\begin{aligned}
\|\bar{\mathbf{z}}(t)-\mathbb{E}\left[\bar{\mathbf{z}}(t)\right]\|^2&\leq 2\|\bar{\mathbf{z}}(t)\|^2+2\|\mathbb{E}\left[\bar{\mathbf{z}}(t)\right]\|^2\\
&\overset{(a)}{\leq} 2\|\bar{\mathbf{z}}(t)\|^2+2\mathbb{E}\left[\|\bar{\mathbf{z}}(t)\|^2\right]\\
&=2\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{y}_i(t))\right\|^2+2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{y}_i(t))\right\|^2\right]\\
&\overset{(b)}{\leq}\frac{2}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{y}_i(t))\|^2+\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla f_i(\mathbf{y}_i(t))\|^2\right]\leq 4\zeta_2^2.
\end{aligned}\tag{5.135}$$

where in (a) we used Jensen's inequality, (b) holds due to Corollary 5.1, and the last step is true from Assumption 5.2-(b). Similarly, to bound the last term in (5.134), from Lemma 5.2 with $\theta=1$, we get

$$\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)-\bar{\mathbf{e}}_g(t)\|^2\Big|\mathcal{F}_t\right]\leq 2\mathbb{E}\left[\|\bar{\mathbf{e}}_y(t)\|^2\Big|\mathcal{F}_t\right]+2\mathbb{E}\left[\|\bar{\mathbf{e}}_g(t)\|^2\Big|\mathcal{F}_t\right]\leq 4\zeta_1\nu^2(t),\tag{5.136}$$

where the last inequality follows from (5.51) and (5.52). Plugging (5.135) and (5.136) into (5.134), we have

$$\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1)-\mathbb{E}\left[\bar{\mathbf{x}}(t+1)\right]\|^2\Big|\mathcal{F}_t\right]\leq(1+\nu(t))\|\bar{\mathbf{y}}(t)-\mathbb{E}\left[\bar{\mathbf{y}}(t)\right]\|^2+4\zeta_2^2\left(\nu^2(t)+\nu(t)\right)+4\zeta_1\nu^2(t).\tag{5.137}$$

To bound further the first in (5.137), from (5.48), we have

$$\begin{aligned}
&\|\bar{\mathbf{y}}(t) - \mathbb{E}[\bar{\mathbf{y}}(t)]\|^2 \\
&= \|(1-\alpha(t))\,(\bar{\mathbf{x}}(t) - \mathbb{E}[\bar{\mathbf{x}}(t)]) + \alpha(t)\,(\bar{\boldsymbol{v}}(t) - \mathbb{E}[\bar{\boldsymbol{v}}(t)])\|^2 \\
&\overset{(a)}{\leq} (1+\nu(t))(1-\alpha(t))^2\,\|\bar{\mathbf{x}}(t) - \mathbb{E}[\bar{\mathbf{x}}(t)]\|^2 + \left(1+\frac{1}{\nu(t)}\right)\alpha^2(t)\,\|\bar{\boldsymbol{v}}(t) - \mathbb{E}[\bar{\boldsymbol{v}}(t)]\|^2 \\
&\leq (1+\nu(t))\,\|\bar{\mathbf{x}}(t) - \mathbb{E}[\bar{\mathbf{x}}(t)]\|^2 + \left(1+\frac{1}{\nu(t)}\right)\alpha^2(t)\,\|\bar{\boldsymbol{v}}(t) - \mathbb{E}[\bar{\boldsymbol{v}}(t)]\|^2, \qquad (5.138)
\end{aligned}$$

where (a) follows from Lemma 5.2 with $\theta = \nu(t)$ and the last step holds due to $0 \leq \alpha(t) \leq 1$. Plugging (5.138) into (5.137), we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1) - \mathbb{E}[\bar{\mathbf{x}}(t+1)]\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\leq (1+\nu(t))^2\,\|\bar{\mathbf{x}}(t) - \mathbb{E}[\bar{\mathbf{x}}(t)]\|^2 + (1+\nu(t))\left(1+\frac{1}{\nu(t)}\right)\alpha^2(t)\,\|\bar{\boldsymbol{v}}(t) - \mathbb{E}[\bar{\boldsymbol{v}}(t)]\|^2 \\
&\quad + 4\zeta_2^2\left(\nu^2(t)+\nu(t)\right) + 4\zeta_1\nu^2(t). \qquad (5.139)
\end{aligned}$$

Using (5.139) and the fact that $\nu(t+1) \leq \nu(t)$, we can write

$$\begin{aligned}
&\nu^2(t+1)\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1) - \mathbb{E}[\bar{\mathbf{x}}(t+1)]\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\leq \nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1) - \mathbb{E}[\bar{\mathbf{x}}(t+1)]\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\leq (1+\nu(t))^2\nu^2(t)\,\|\bar{\mathbf{x}}(t) - \mathbb{E}[\bar{\mathbf{x}}(t)]\|^2 + (1+\nu(t))^2\nu(t)\alpha^2(t)\,\|\bar{\boldsymbol{v}}(t) - \mathbb{E}[\bar{\boldsymbol{v}}(t)]\|^2 \\
&\quad + 4\zeta_2^2\left(\nu^4(t)+\nu^3(t)\right) + 4\zeta_1\nu^4(t). \qquad (5.140)
\end{aligned}$$

Now, we focus on the third term in (5.140). Starting from (5.46), using (5.68), and (5.70), we can write

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t+1) - \mathbb{E}[\bar{\boldsymbol{v}}(t+1)]\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|\bar{\boldsymbol{v}}(t) - \mathbb{E}[\bar{\boldsymbol{v}}(t)] + \bar{\mathbf{e}}_v(t) - \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t) - \frac{\nu(t)}{\alpha(t)}(\bar{\mathbf{z}}(t) - \mathbb{E}[\bar{\mathbf{z}}(t)])\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&= \left\|\bar{\boldsymbol{v}}(t) - \mathbb{E}[\bar{\boldsymbol{v}}(t)] - \frac{\nu(t)}{\alpha(t)}(\bar{\mathbf{z}}(t) - \mathbb{E}[\bar{\mathbf{z}}(t)])\right\|^2 + \mathbb{E}\left[\left\|\bar{\mathbf{e}}_v(t) - \frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\|^2 \,\middle|\, \mathcal{F}_t\right]. \qquad (5.141)
\end{aligned}$$

For the first term in (5.141), from Lemma 5.2 with $\theta = \nu(t)$, we have

$$\begin{aligned}
&\left\|\bar{\boldsymbol{v}}(t)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t)\right]-\frac{\nu(t)}{\alpha(t)}\left(\bar{\mathbf{z}}(t)-\mathbb{E}\left[\bar{\mathbf{z}}(t)\right]\right)\right\|^2 \\
&\leq (1+\nu(t))\left\|\bar{\boldsymbol{v}}(t)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t)\right]\right\|^2+\left(1+\frac{1}{\nu(t)}\right)\frac{\nu^2(t)}{\alpha^2(t)}\left\|\bar{\mathbf{z}}(t)-\mathbb{E}\left[\bar{\mathbf{z}}(t)\right]\right\|^2 \\
&\leq (1+\nu(t))\left\|\bar{\boldsymbol{v}}(t)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t)\right]\right\|^2+4\zeta_2^2\left(1+\frac{1}{\nu(t)}\right)\frac{\nu^2(t)}{\alpha^2(t)}, \qquad (5.142)
\end{aligned}$$

where in the last inequality we used (5.135). To bound the second term in (5.141), from Lemma 5.2 with $\theta = \nu(t)$, we get

$$\mathbb{E}\left[\left\|\bar{\mathbf{e}}_v(t)-\frac{1}{\alpha(t)}\bar{\mathbf{e}}_g(t)\right\|^2\middle|\mathcal{F}_t\right]\leq 2\mathbb{E}\left[\left\|\bar{\mathbf{e}}_v(t)\right\|^2\middle|\mathcal{F}_t\right]+\frac{2}{\alpha^2(t)}\mathbb{E}\left[\left\|\bar{\mathbf{e}}_g(t)\right\|^2\middle|\mathcal{F}_t\right]\leq 4\zeta_1\frac{\nu^2(t)}{\alpha^2(t)}, \qquad (5.143)$$

where the last step holds due to (5.52). Plugging (5.142) and (5.143) into (5.141), we can write

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{\boldsymbol{v}}(t+1)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t+1)\right]\right\|^2\middle|\mathcal{F}_t\right] \\
&\leq (1+\nu(t))\left\|\bar{\boldsymbol{v}}(t)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t)\right]\right\|^2+4\zeta_2^2\left(1+\frac{1}{\nu(t)}\right)\frac{\nu^2(t)}{\alpha^2(t)}+4\zeta_1\frac{\nu^2(t)}{\alpha^2(t)}.
\end{aligned}$$

This combined with the fact that $\alpha(t+1)\leq\alpha(t)$ arrives us to

$$\begin{aligned}
&\alpha^2(t+1)\mathbb{E}\left[\left\|\bar{\boldsymbol{v}}(t+1)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t+1)\right]\right\|^2\middle|\mathcal{F}_t\right] \\
&\leq \alpha^2(t)\mathbb{E}\left[\left\|\bar{\boldsymbol{v}}(t+1)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t+1)\right]\right\|^2\middle|\mathcal{F}_t\right] \\
&\leq (1+\nu(t))\alpha^2(t)\left\|\bar{\boldsymbol{v}}(t)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t)\right]\right\|^2+4\zeta_2^2\left(\nu^2(t)+\nu(t)\right)+4\zeta_1\nu^2(t). \qquad (5.144)
\end{aligned}$$

Now, we apply Lemma 4.10 with $w(t)=\alpha^2(t)\left\|\bar{\boldsymbol{v}}(t)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t)\right]\right\|^2$, $p(t)=\nu(t)$, $u(t)=0$, and $q(t)=4\zeta_2^2\left(\nu^2(t)+\nu(t)\right)+4\zeta_1\nu^2(t)$. Moreover, for $\sigma\in(1,2)$ and using $\nu(t)=\frac{\nu_0}{(t+\tau)^\sigma}$, we have $\sum_{t=0}^{\infty}\nu^2(t)\leq\sum_{t=0}^{\infty}\nu(t)<\infty$. Thus, we get $\alpha^2(t)\left\|\bar{\boldsymbol{v}}(t)-\mathbb{E}\left[\bar{\boldsymbol{v}}(t)\right]\right\|^2\leq\zeta_{18}<\infty$ for some constant ζ_{18}. This together with (5.140) leads us to

$$\begin{aligned}
&\nu^2(t+1)\mathbb{E}\left[\left\|\bar{\mathbf{x}}(t+1)-\mathbb{E}\left[\bar{\mathbf{x}}(t+1)\right]\right\|^2\middle|\mathcal{F}_t\right] \\
&\leq (1+\nu(t))^2\nu^2(t)\left\|\bar{\mathbf{x}}(t)-\mathbb{E}\left[\bar{\mathbf{x}}(t)\right]\right\|^2+(1+\nu(t))^2\nu(t)\zeta_{18}+4\zeta_2^2\left(\nu^4(t)+\nu^3(t)\right)+4\zeta_1\nu^4(t). \qquad (5.145)
\end{aligned}$$

Applying Lemma 4.10 with $w(t) = \nu^2(t)\,\|\bar{\mathbf{x}}(t) - \mathbb{E}\,[\bar{\mathbf{x}}(t)]\|^2$, $p(t) = 2\nu(t)+\nu^2(t)$, $u(t) = 0$, and $q(t) = (1+\nu(t))^2\nu(t)\zeta_{18} + 4\zeta_2^2\left(\nu^4(t)+\nu^3(t)\right)+4\zeta_1\nu^4(t)$ and using the above equation, we get $\nu^2(t)\,\|\bar{\mathbf{x}}(t) - \mathbb{E}\,[\bar{\mathbf{x}}(t)]\|^2 \le \zeta_{19} < \infty$ for some constant ζ_{19}. This together with (5.133) arrives us at

$$\begin{aligned}\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2\right] &\le 2\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbb{E}\,[\bar{\mathbf{x}}(t)]\,\|^2\right] + 2\nu_0^2R^2\\ &\le 2(\zeta_{19} + \nu_0^2R^2).\end{aligned} \tag{5.146}$$

Taking expectations of both sides of (5.132), we get

$$\mathbb{E}\left[\|\bar{\mathbf{y}}(t) - \bar{\mathbf{x}}(t)\|^2\right] \le \frac{\alpha^2(t)}{\nu^2(t)}\left(2\nu^2(t)\mathbb{E}\left[\|\bar{\boldsymbol{v}}(t) - \mathbf{x}^\star\|^2\right]+2\nu^2(t)\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2\right]\right)\le\zeta_{20}\frac{\alpha^2(t)}{\nu^2(t)}, \tag{5.147}$$

where $\zeta_{20} \coloneqq 2(\zeta_{17} + 2\zeta_{19} + 2\nu_0^2R^2)$ and the last step follows from (5.131) and (5.146). This combined with Lemma 5.11-(c) leads us to

$$\begin{aligned}\sum_{k=0}^{t}\mathbb{E}\left[\|\bar{\mathbf{y}}(k)-\bar{\mathbf{x}}(k)\|^2\right]\nu^2(k)\prod_{\ell=k+1}^{t}(1-\alpha(\ell)) &\le \zeta_{20}\sum_{k=0}^{t}\alpha^2(k)\prod_{\ell=k+1}^{t}(1-\alpha(\ell))\\ &= \zeta_{20}\sum_{k=0}^{t}\alpha^2(k)\frac{\prod_{\ell=0}^{t}(1-\alpha(\ell))}{\prod_{s=0}^{k}(1-\alpha(s))}\\ &= \zeta_{20}\sum_{k=0}^{t}\alpha^2(k)\frac{\lambda(t+1)}{\lambda(k+1)}\\ &= \zeta_{20}\lambda(t+1)\sum_{k=0}^{t}\nu(k) \le \zeta_{20}\lambda(t+1)\sum_{k=0}^{\infty}\nu(k).\end{aligned} \tag{5.148}$$

Plugging (5.148) into (5.128) we arrive at (5.129) for $t+1$ with

$$\zeta_{16} \coloneqq \zeta_{12}\zeta_{20}\sum_{k=0}^{\infty}\nu(k) = \zeta_{12}\zeta_{20}\nu_0\sum_{k=0}^{\infty}(k+\tau)^{-\sigma}. \tag{5.149}$$

This completes the proof of (5.129) for every $t \ge 0$. Starting from (5.129), we obtain

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{x}}(t))\right]-f^\star &\le \mathbb{E}\left[\psi_t\right] - f^\star + \zeta_{16}\lambda(t) + \zeta_1\lambda(t)\left(2\zeta_{14} + \zeta_{13}\zeta_{15}\right)\\ &\overset{(a)}{\le} \mathbb{E}\left[\Psi_t(\mathbf{x}^\star)\right]-\frac{\lambda(t)}{2}\mathbb{E}\left[\|\mathbf{x}^\star-\bar{\boldsymbol{v}}(t)\|^2\right]-f^\star+\zeta_{16}\lambda(t)+\zeta_1\lambda(t)\left(2\zeta_{14}+\zeta_{13}\zeta_{15}\right)\\ &\le \mathbb{E}\left[\Psi_t(\mathbf{x}^\star)\right] - f^\star + \zeta_{16}\lambda(t) + \zeta_1\lambda(t)\left(2\zeta_{14} + \zeta_{13}\zeta_{15}\right)\\ &\le \sigma(t)(\Psi_0(\mathbf{x}^\star) - f^\star) + \zeta_{16}\lambda(t) + \zeta_1\lambda(t)\left(2\zeta_{14} + \zeta_{13}\zeta_{15}\right),\end{aligned} \tag{5.150}$$

where (a) follows from (5.102) with $\boldsymbol{\omega} = \mathbf{x}^\star$ and the last inequality holds due to (5.95). Using (5.96), we get

$$\sigma(t) = \sigma(0)\prod_{\ell=0}^{t-1}(1-\alpha(\ell)) = \frac{\sigma(0)}{\lambda(0)}\lambda(0)\prod_{\ell=0}^{t-1}(1-\alpha(\ell)) \overset{\text{(a)}}{=} \frac{\sigma(0)}{\lambda(0)}\lambda(t) = \frac{1}{\lambda_0}\lambda(t), \tag{5.151}$$

where (a) is true from (5.98) and the last equality follows from $\sigma(0) = 1$. Let

$$\zeta_{21} \coloneqq \frac{1}{\lambda_0}(\Psi_0(\mathbf{x}^\star) - f^\star) + \zeta_{16}. \tag{5.152}$$

Plugging (5.151) into (5.150), we arrive at

$$\begin{aligned}
\mathbb{E}\left[f(\bar{\mathbf{x}}(t))\right] - f^\star &\leq \zeta_{21}\lambda(t) + \zeta_1\lambda(t)\left(2\zeta_{14} + \zeta_{13}\zeta_{15}\right), \\
&\overset{\text{(a)}}{=} \frac{\alpha^2(t)}{\nu(t)}\left(\zeta_{21} + \zeta_1\left(2\zeta_{14} + \zeta_{13}\zeta_{15}\right)\right) \\
&\overset{\text{(b)}}{\leq} \frac{4}{\nu_0}\left(\zeta_{21} + \zeta_1\left(2\zeta_{14} + \zeta_{13}\zeta_{15}\right)\right)\frac{(t+\tau)^\sigma}{(t+1)^2} \\
&\leq \frac{4\tau^2}{\nu_0}\left(\zeta_{21} + \zeta_1\left(2\zeta_{14} + \zeta_{13}\zeta_{15}\right)\right)(t+\tau)^{-(2-\sigma)},
\end{aligned} \tag{5.153}$$

where (a) holds due to (5.101), in (b) we used Lemma 5.11-(c), and the last step follows from $t + \tau \leq \tau(t+1)$ for every $t \geq 0$. This completes the proof of Theorem 5.2 with

$$\zeta \coloneqq \zeta_{21} + \zeta_1\left(2\zeta_{14} + \zeta_{13}\zeta_{15}\right). \tag{5.154}$$

5.10 Proof of Preliminaries

In this section, we provide the proofs of auxiliary lemmas.

Proof of Corollary 5.1: We use induction to prove the claim of the corollary. We note that for the induction base $n = 1$, we have $\|\boldsymbol{u}_1\|^2 \leq \|\boldsymbol{u}_1\|^2$. Now, assume that (5.11) holds for some $n > 1$. Using Lemma 5.2 with $\boldsymbol{u} = \sum_{i=1}^n \boldsymbol{u}_i$, $\mathbf{r} = \boldsymbol{u}_{n+1}$, and $\theta = \frac{1}{n}$, we arrive at

$$\begin{aligned}
\left\|\sum_{i=1}^{n+1}\boldsymbol{u}_i\right\|^2 &\leq \left(1+\frac{1}{n}\right)\left\|\sum_{i=1}^{n}\boldsymbol{u}_i\right\|^2 + (n+1)\left\|\boldsymbol{u}_{n+1}\right\|^2 \\
&\overset{\text{(a)}}{\leq} \left(1+\frac{1}{n}\right)n\sum_{i=1}^{n}\left\|\boldsymbol{u}_i\right\|^2 + (n+1)\left\|\boldsymbol{u}_{n+1}\right\|^2 \\
&= (n+1)\sum_{i=1}^{n+1}\left\|\boldsymbol{u}_i\right\|^2,
\end{aligned}$$

where (a) follows from the induction assumption.

Proof of Corollary 5.2: Using Lemma 5.2 with $\boldsymbol{u} = \boldsymbol{u}_1 + \mathbf{r}_1$ and $\mathbf{r} = -\mathbf{r}_1$, for any $\theta > 0$, we arrive at

$$\|\boldsymbol{u}_1\| \leq (1+\theta)\|\boldsymbol{u}_1 + \mathbf{r}_1\|^2 + (1+\theta^{-1})\|\mathbf{r}_1\|^2,$$

or equivalently,

$$\|\boldsymbol{u}_1 + \mathbf{r}_1\|^2 \geq \frac{1}{1+\theta}\|\boldsymbol{u}_1\|^2 - \theta^{-1}\|\mathbf{r}_1\|^2.$$

This completes the proof of the corollary.

Proof of Lemma 5.5: Using the fact $\|U^T\| = \|U\|$, we have

$$\begin{aligned}
&\left\|((1-\gamma)I + \gamma W)\left(A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right)\right\| \\
&= \left\|\left(A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right)^T ((1-\gamma)I + \gamma W)^T\right\| \\
&\overset{(a)}{=} \left\|\left(A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right)^T ((1-\gamma)I + \gamma W)\right\| \\
&\overset{(b)}{\leq} (1-\gamma)\left\|\left(A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right)^T I\right\| + \gamma\left\|\left(A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right)^T W\right\| \\
&= (1-\gamma)\left\|A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right\| + \gamma\left\|\left(A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right)^T \left(W - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right\| \\
&\overset{(c)}{\leq} (1-\gamma)\left\|A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right\| + \gamma(1-\beta)\left\|\left(A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right)^T\right\| \\
&= (1-\gamma)\left\|A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right\| + \gamma(1-\beta)\left\|A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right\| \\
&= (1-\gamma\beta)\left\|A - \frac{1}{n}\mathbf{1}\mathbf{1}^T A\right\|,
\end{aligned}$$

where (a) is true since W is symmetric, (b) follows from the triangle inequality, and step (c) holds due to Lemma 5.4.

Proof of Corollary 5.3: Starting from $\|U^T\| = \|U\|$, we get

$$\begin{aligned}
\left\|\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)A\right\| &= \left\|A^T\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T\right\| \\
&= \left\|A^T\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right\| \\
&\overset{(a)}{\leq} \left\|A^T\right\| = \|A\|,
\end{aligned}$$

where (a) follows from Lemma 5.5 with $\beta = 0$.

Proof of Lemma 5.6: We start with the characteristic polynomial of $A(\nu)$, given by

$$p(\tau) = (\tau-\rho)^2(\tau-c\rho-\delta_3) - \delta_3(2+\sqrt{\rho})(\tau-\rho) - 2\delta_3\sqrt{\rho}, \tag{5.155}$$

where $\delta_3 := \delta_0\nu^2$. Using the conditions $\delta_3 < (c-1)^2\rho^2/5$ and $\rho < c\rho < 1$, we have $\delta_3 < 1$.

Now, we evaluate $p(\cdot)$ on $c\rho+\delta_3$ and get

$$p(c\rho+\delta_3) = -\delta_3(2+\sqrt{\rho})((c-1)\rho+\delta_3) - 2\delta_3\sqrt{\rho} < 0. \tag{5.156}$$

Moreover, we have

$$\begin{aligned}
p(c\rho+4\delta_3^{1/3}) &= (c\rho+4\delta_3^{1/3}-\rho)^2\left(c\rho+4\delta_3^{1/3}-c\rho-\delta_3\right) - \delta_3(2+\sqrt{\rho})(c\rho+4\delta_3^{1/3}-\rho) - 2\delta_3\sqrt{\rho} \\
&\overset{(a)}{\geq} (c\rho+4\delta_3^{1/3}-\rho)^2\left(4\delta_3^{1/3}-\delta_3\right) - 3\delta_3(c\rho+4\delta_3^{1/3}-\rho) - 2\delta_3 \\
&\overset{(b)}{>} 16\delta_3^{2/3}\left(4\delta_3^{1/3}-\delta_3\right) - 3\delta_3(1+4\delta_3^{1/3}) - 2\delta_3 \\
&= 64\delta_3 - 16\delta_3^{5/3} - 3\delta_3 - 12\delta_3^{4/3} - 2\delta_3 \\
&\overset{(c)}{\geq} 64\delta_3 - 16\delta_3 - 3\delta_3 - 12\delta_3 - 2\delta_3 \\
&= 12\delta_3 \\
&> 0,
\end{aligned} \tag{5.157}$$

where (a) follows from $\rho < 1$, step (b) holds due to $0 < \rho < c\rho < 1$, and in (c) we used $\delta_3^{5/3} \leq \delta_3^{4/3} \leq \delta_3 < 1$. Therefore, (5.156) and (5.157) imply that $p(\tau)$ has at least root in $(c\rho+\delta_3, c\rho+4\delta_3^{1/3})$. Next, we show that it has no root greater than $c\rho+\delta_3^{1/3}$. Taking derivative of (5.155) w.r.t. τ, we can write

$$p'(\tau) = 2(\tau-\rho)(\tau-c\rho-\delta_3) + (\tau-\rho)^2 - \delta_3(\sqrt{\rho}+2).$$

Any $\tau \geq c\rho+\delta_3$ can be represented as $\tau = c\rho+\delta_3+\sigma$ for some $\sigma \geq 0$. Defining $\delta_4 :=$

$c\rho + \delta_3 - \rho > \delta_3$, we have

$$\begin{aligned} p'(c\rho+\delta_3+\sigma) &= 2(\sigma+\delta_4)\sigma + (\sigma+\delta_4)^2 - \delta_3\left(2+\sqrt{\rho}\right) \\ &= 3\sigma^2 + 4\delta_4\sigma + \delta_4^2 - \delta_3\left(2+\sqrt{\rho}\right) \\ &= 3\sigma^2 + 4\delta_4\sigma + \left(\delta_3^2 + (c-1)^2\rho^2 + 2(c-1)\rho\delta_3\right) - \delta_3\left(2+\sqrt{\rho}\right) \\ &\overset{(a)}{>} 3\sigma^2 + 4\delta_4\sigma + \delta_3^2 + 5\delta_3 + 2(c-1)\rho\delta_3 - \delta_3\left(2+\sqrt{\rho}\right) \\ &= 3\sigma^2 + 4\delta_4\sigma + \delta_3^2 + \left(5 + 2(c-1)\rho - 2 - \sqrt{\rho}\right)\delta_3 \\ &= 3\sigma^2 + 4\delta_4\sigma + \delta_3^2 + \left(2c\rho + \left(3-\rho-\sqrt{\rho}\right)\right)\delta_3 \overset{(b)}{\geq} 0, \end{aligned} \tag{5.158}$$

where (a) follows from $\delta_3 < (c-1)^2\rho^2/5$, holds since $\rho < 1$. Therefore, we have $p(\tau) \geq p(c\rho + 4\delta_3^{1/3}) > 0$ for every $\tau \geq c\rho + 4\delta_3^{1/3}$. Hence, the largest root of $p(t)$ should lie in the interval $(c\rho+\delta_3, c\rho+4\delta_3^{1/3})$. This proves the first claim of the lemma.

Before we prove part (b), we need to show that the eigenvector of $A(\nu)$ can be normalized and written as $[\Theta_1(\nu), \Theta_2(\nu), 1]$. To this end, we need to show that $\Theta_3(\nu) \neq 0$. We prove this by contradiction. Assume that $\Theta_3(\nu) = 0$. From the first row of $A(\nu)\boldsymbol{\Theta}(\nu) = \mu(\nu)\boldsymbol{\Theta}(\nu)$, we get $\rho\Theta_1(\nu) = \mu(\nu)\ \Theta_1(\nu)$. However, from part (a) we know that $\mu(\nu) > c\rho + \delta_3 > c > \rho$. Therefore, we have $\Theta_1(\nu) = 0$.

Then, the second row of $A(\nu)\boldsymbol{\Theta}(\nu) = \mu(\nu)\boldsymbol{\Theta}(\nu)$ implies $\rho\Theta_2(\nu) = \mu(\nu)\ \Theta_2(\nu)$, which leads to $\Theta_2(\nu) = 0$, or $\boldsymbol{\Theta}(\nu) = \mathbf{0}$. Hence $\boldsymbol{\Theta}(\nu)$ cannot be an eigenvector. This contradiction implies $\Theta_3(\nu) \neq 0$.

Now, consider a normalized eigenvector, so that $\boldsymbol{\Theta}(\nu) = [\Theta_1(\nu), \Theta_2(\nu), 1]^T$. From the first row of $A(\nu)\boldsymbol{\Theta}(\nu) = \mu(\nu)\boldsymbol{\Theta}(\nu)$, we get $\rho\Theta_1(\nu) + \frac{\nu^2}{1-\sqrt{\rho}} = \mu(\nu)\Theta_1(\nu)$, or equivalently,

$$\Theta_1(\nu) = \frac{\nu^2}{(1-\sqrt{\rho})(\mu(\nu)-\rho)}. \tag{5.159}$$

From part (a), and the facts that $\delta_3 < (c-1)^2\rho^2/5$ and $\rho < c\rho < 1$, we have

$$\mu(\nu) - \rho < c\rho + 4\delta_3^{1/3} - \rho \leq (c-1)\rho + \frac{4}{5^{1/3}}(c-1)^{2/3}\rho^{2/3} < 4.$$

This together with (5.159) implies $\Theta_1(\nu) > \frac{\nu^2}{4(1-\sqrt{\rho})}$.

Similarly, we have $\mu(\nu) - \rho > c\rho - \rho = (c-1)\rho$, thus with (5.159) we get $\Theta_1(\nu) < \frac{\nu^2}{(c-1)\rho(1-\sqrt{\rho})}$. This completes the proof of part (b).

To prove part (c), from the second row of $A(\nu)\mathbf{\Theta}(\nu) = \mu(\nu)\mathbf{\Theta}(\nu)$, we arrive at

$$\frac{1}{1-\sqrt{\rho}}\Theta_1(\nu) + \rho\Theta_2(\nu) + \frac{\nu^2}{\sqrt{\rho}(1-\sqrt{\rho})^2} = \mu(\nu)\Theta_2(\nu).$$

This combined with (5.159) leads us to

$$\Theta_2(\nu) = \frac{\nu^2}{(1-\sqrt{\rho})^2(\mu(\nu)-\rho)}\left(\frac{1}{\mu(\nu)-\rho} + \frac{1}{\sqrt{\rho}}\right). \tag{5.160}$$

Thus, from $c\rho < c\rho + \delta_3 < \mu(\nu)$, we have

$$\Theta_2(\nu) < \frac{\nu^2}{(1-\sqrt{\rho})^2(c-1)\rho}\left(\frac{1}{(c-1)\rho} + \frac{1}{\sqrt{\rho}}\right),$$

which completes the proof of the lemma.

Proof of Lemma 5.7: We first show that $\mu(\nu) > c\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_3$. To this end, using $(c-1)^{2/3}\rho^{2/3} < 1$ we get

$$\begin{aligned}
&p\left(c\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_3\right)\\
&= \left(c\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_3 - \rho\right)^2\left(c\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_3 - c\rho - \delta_3\right)\\
&\quad - \delta_3(2+\sqrt{\rho})\left(c\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_3 - \rho\right) - 2\delta_3\sqrt{\rho}\\
&< \left(\frac{2}{(c-1)^{3/2}\rho^{3/2}}\left(c\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_3 - \rho\right) - (2+\sqrt{\rho})\right)\delta_3\left(c\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_3 - \rho\right)\\
&< \left(\frac{6}{(c-1)^3\rho^3}\delta_3 - \sqrt{\rho}\right)\delta_3\left(c\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_3 - \rho\right)\\
&< 0,
\end{aligned}$$

where the last step follows from $\delta_3 < \sqrt{\rho}(c-1)^3\rho^3/64 < \sqrt{\rho}(c-1)^3\rho^3/6$. Starting (5.160), we can write

$$\begin{aligned}
\Theta_2(\nu) &> \frac{\nu^2}{(1-\sqrt{\rho})^2(\mu(\nu)-\rho)^2}\\
&\overset{(a)}{>} \frac{\nu^2}{(1-\sqrt{\rho})^2\left(c\rho - \rho + 4\delta_0^{1/3}\nu^{2/3}\right)^2}\\
&> \frac{\nu^2}{4(1-\sqrt{\rho})^2(c-1)^2\rho^2},
\end{aligned}$$

where (a) follows from $\mu(\nu) < c\rho + 4\delta_3^{1/3}$ and the last inequality holds due to $\delta_3 < \sqrt{\rho}(c-1)^3\rho^3/64 < (c-1)^3\rho^3/64$.

To prove part (c), from Lemma 5.6 and part (b), we can write

$$\mathbf{\Theta}(\nu) > \begin{bmatrix} \frac{\nu^2}{4(1-\sqrt{\rho})} \\ \frac{\nu^2}{4(1-\sqrt{\rho})^2(c-1)^2\rho^2} \\ 1 \end{bmatrix} > \frac{1}{4}\begin{bmatrix} \nu^2 \\ \nu^2 \\ 1 \end{bmatrix},$$

where in the last inequality we used $0 < \rho < c\rho < 1$. This completes the proof of the lemma.

Proof of Lemma 5.8:

Using the mean value theorem for the function $h_1(x) = \ln\Theta_1(e^x)$ and $h_2(x) = \ln\Theta_2(e^x)$ we have

$$\frac{h_1(\ln\nu_2) - h_1(\ln\nu_1)}{\ln\nu_2 - \ln\nu_1} = h_1'(\nu_3),$$
$$\frac{h_2(\ln\nu_2) - h_2(\ln\nu_1)}{\ln\nu_2 - \ln\nu_1} = h_2'(\nu_4),$$

for some $\nu_3, \nu_4 \in [\ln\nu_1, \ln\nu_2]$ which implies

$$\frac{h_1(\ln\nu_2) - h_1(\ln\nu_1)}{\ln\nu_2 - \ln\nu_1} = \frac{\ln\Theta_1(\nu_2) - \ln\Theta_1(\nu_1)}{\ln\nu_2 - \ln\nu_1} = \frac{\ln\left(\frac{\Theta_1(\nu_2)}{\Theta_1(\nu_1)}\right)}{\ln\left(\frac{\nu_2}{\nu_1}\right)} = h_1'(\nu_3), \tag{5.161}$$

$$\frac{h_2(\ln\nu_2) - h_2(\ln\nu_1)}{\ln\nu_2 - \ln\nu_1} = \frac{\ln\Theta_2(\nu_2) - \ln\Theta_2(\nu_1)}{\ln\nu_2 - \ln\nu_1} = \frac{\ln\left(\frac{\Theta_2(\nu_2)}{\Theta_2(\nu_1)}\right)}{\ln\left(\frac{\nu_2}{\nu_1}\right)} = h_2'(\nu_4). \tag{5.162}$$

From the definitions $h_1(x)$ and $h_2(x)$, we can write

$$h_1'(x) = \frac{e^x\Theta_1'(e^x)}{\Theta_1(e^x)}, \tag{5.163}$$

$$h_2'(x) = \frac{e^x\Theta_2'(e^x)}{\Theta_2(e^x)}, \tag{5.164}$$

Now, we evaluate $\Theta_1'(\nu)$. Taking derivative of (5.159) w.r.t ν, we have

$$\Theta_1'(\nu) = \frac{2\nu(\mu(\nu)-\rho) - \mu'(\nu)\nu^2}{(1-\sqrt{\rho})(\mu(\nu)-\rho)^2}. \tag{5.165}$$

Plugging (5.165) and (5.159) into (5.163), we arrive at

$$|h_1'(x)| = \left| e^x \frac{\frac{2e^x(\mu(e^x)-\rho)-\mu'(e^x)e^{2x}}{(1-\sqrt{\rho})(\mu(e^x)-\rho)^2}}{\frac{e^{2x}}{(1-\sqrt{\rho})(\mu(e^x)-\rho)}} \right| = \left| 2 - e^x \frac{\mu'(e^x)}{\mu(e^x)-\rho} \right| \leq 2 + e^x \frac{|\mu'(e^x)|}{|\mu(e^x)-\rho|}. \tag{5.166}$$

To derive an upper bound on $\mu'(\nu)$, we exploit the fact $p(\mu(\nu)) = 0$. Taking derivative w.r.t ν on both sides of $p(\mu(\nu)) = 0$, we arrive at

$$\begin{aligned} &2(\mu(\nu)-\rho)(\mu(\nu)-c\rho-\delta_0\nu^2)\mu'(\nu) + (\mu(\nu)-\rho)^2\left(\mu'(\nu)-2\delta_0\nu\right) \\ &-2\delta_0\nu(2+\sqrt{\rho})(\mu(\nu)-\rho) - \delta_0\nu^2(2+\sqrt{\rho})\mu'(\nu) - 4\delta_0\nu\sqrt{\rho} = 0. \end{aligned}$$

Hence, we can write

$$\mu'(\nu) = \frac{((\mu(\nu)-\rho)^2 + (2+\sqrt{\rho})(\mu(\nu)-\rho) + 2\sqrt{\rho})}{2(\mu(\nu)-\rho)(\mu(\nu)-c\rho-\delta_0\nu^2) + (\mu(\nu)-\rho)^2 - \delta_0\nu^2(2+\sqrt{\rho})}(2\delta_0\nu). \tag{5.167}$$

For the nominator part of (5.167), from Lemma 5.6-(a), we get

$$\begin{aligned} 0 <&(\mu(\nu)-\rho)^2 + (2+\sqrt{\rho})(\mu(\nu)-\rho) + 2\sqrt{\rho} \\ &\leq ((c-1)\rho + 4\delta_0^{1/3}\nu^{2/3})^2 + (2+\sqrt{\rho})((c-1)\rho + 4\delta_0^{1/3}\nu^{2/3}) + 2\sqrt{\rho} \\ &\overset{(a)}{<} \left((c-1)\rho + \rho^{1/6}(c-1)\rho\right)^2 + (2+\sqrt{\rho})\left((c-1)\rho + \rho^{1/6}(c-1)\rho\right) + 2\sqrt{\rho} \\ &< 4(c-1)^2\rho^2 + 6(c-1)\rho + 2\sqrt{\rho} \\ &< 10(c-1)\rho + 2\sqrt{\rho}, \end{aligned} \tag{5.168}$$

where (a) follows from $\delta_0\nu^2 < \sqrt{\rho}(c-1)^3\rho^3/64$ and the last step holds due to $(c-1)^2\rho^2 < (c-1)\rho$.

For the denominator part of (5.167), from $\mu(\nu) > c\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_0\nu^2$ we have

$$\begin{aligned} &2(\mu(\nu)-\rho)(\mu(\nu)-c\rho-\delta_0\nu^2) + (\mu(\nu)-\rho)^2 - \delta_0\nu^2(2+\sqrt{\rho}) \\ &> 2(\mu(\nu)-\rho)\frac{2}{(c-1)^{3/2}\rho^{3/2}}\delta_0\nu^2 + (\mu(\nu)-\rho)^2 - \delta_0\nu^2(2+\sqrt{\rho}) \\ &> \frac{4\delta_0\nu^2}{(c-1)^{1/2}\rho^{1/2}} + (\mu(\nu)-\rho)^2 - 3\delta_0\nu^2 \\ &\overset{(a)}{>} \frac{\delta_0\nu^2}{(c-1)^{1/2}\rho^{1/2}} + \left((c-1)\rho + \frac{3}{(c-1)^{3/2}\rho^{3/2}}\delta_0\nu^2\right)^2 \\ &> \frac{7\delta_0\nu^2}{(c-1)^{1/2}\rho^{1/2}}, \end{aligned} \tag{5.169}$$

where (a) follows from $0 < \rho < c\rho < 1$. Plugging (5.168) and (5.169) into (5.167), we arrive at

$$0 < \mu'(\nu) < \left(\frac{20}{7}(c-1)\rho + \frac{4}{7}\sqrt{\rho}\right)(c-1)^{1/2}\rho^{1/2}\nu^{-1} < 4(c-1)^{1/2}\rho^{1/2}\nu^{-1}. \tag{5.170}$$

This together with (5.166), $\mu(\nu) > c\rho$, and $0 < \rho < c\rho < 1$, we can write

$$|h_1'(x)| \le 2 + e^x \frac{\mu'(e^x)}{\mu(e^x) - \rho} \le 2 + 4(c-1)^{-1/2}\rho^{-1/2} < 6(c-1)^{-1/2}\rho^{-1/2}.$$

Using this and (5.161), we arrive at the first part of the lemma.

Taking derivative of (5.160) w.r.t ν, we get

$$\Theta_2'(\nu) = \frac{1}{(1-\sqrt{\rho})^2(\mu(\nu)-\rho)^2}\left((2\nu(\mu(\nu)-\rho) - \mu'(\nu)\nu^2)\left(\frac{1}{\mu(\nu)-\rho} + \frac{1}{\sqrt{\rho}}\right) - \frac{\mu'(\nu)\nu^2}{\mu(\nu)-\rho}\right).$$

Moreover, we have

$$|\Theta_2'(\nu)| = \frac{1}{(1-\sqrt{\rho})^2(\mu(\nu)-\rho)^2}\left|(2\nu(\mu(\nu)-\rho) - \mu'(\nu)\nu^2)\left(\frac{1}{\mu(\nu)-\rho} + \frac{1}{\sqrt{\rho}}\right) - \frac{\mu'(\nu)\nu^2}{\mu(\nu)-\rho}\right|. \tag{5.171}$$

Using the triangle inequality, we get

$$\begin{aligned}
&\left|(2\nu(\mu(\nu)-\rho) - \mu'(\nu)\nu^2)\left(\frac{1}{\mu(\nu)-\rho} + \frac{1}{\sqrt{\rho}}\right) - \frac{\mu'(\nu)\nu^2}{\mu(\nu)-\rho}\right| \\
&\le (2\nu(\mu(\nu)-\rho) + |\mu'(\nu)|\nu^2)\left(\frac{1}{\mu(\nu)-\rho} + \frac{1}{\sqrt{\rho}}\right) + \frac{|\mu'(\nu)|\nu^2}{\mu(\nu)-\rho} \\
&\overset{(a)}{<} \left(2\nu(c\rho + 4\delta_0^{1/3}\nu^{2/3} - \rho) + |\mu'(\nu)|\nu^2\right)\left(\frac{1}{(c-1)\rho} + \frac{1}{\sqrt{\rho}}\right) + \frac{|\mu'(\nu)|\nu^2}{(c-1)\rho} \\
&\overset{(b)}{<} \left(\left(2(c\rho + 4\delta_0^{1/3}\nu^{2/3} - \rho) + 4(c-1)^{1/2}\rho^{1/2}\right)\left(\frac{1}{(c-1)\rho} + \frac{1}{\sqrt{\rho}}\right) + \frac{4}{(c-1)^{1/2}\rho^{1/2}}\right)\nu \\
&\overset{(c)}{<} \left(\left(2(c\rho + \rho^{1/6}(c-1)\rho - \rho) + 4(c-1)^{1/2}\rho^{1/2}\right)\left(\frac{1}{(c-1)\rho} + \frac{1}{\sqrt{\rho}}\right) + \frac{4}{(c-1)^{1/2}\rho^{1/2}}\right)\nu \\
&\overset{(d)}{<} \left(\left(8(c-1)^{1/2}\rho^{1/2}\left(\frac{1}{(c-1)\rho} + \frac{1}{\sqrt{\rho}}\right) + \frac{4}{(c-1)^{1/2}\rho^{1/2}}\right)\right)\nu \\
&\overset{(e)}{<} \frac{8c+4}{(c-1)^{1/2}\rho^{1/2}}\nu,
\end{aligned} \tag{5.172}$$

where (a) follows from $c\rho < \mu(\nu) < c\rho + 4\delta_0^{1/3}\nu^{2/3}$, in (b) we used (5.170). In step (c), we exploited the fact $\delta_0\nu^2 < \sqrt{\rho}(c-1)^3\rho^3/64$, and the step in (d) holds due to

$(c-1)\rho<(c-1)^{1/2}\rho^{1/2}$. Finally, step (e) follows from

$$(c-1)^{1/2}\rho^{1/2}\left(\frac{1}{(c-1)\rho}+\frac{1}{\sqrt{\rho}}\right)<(c-1)^{1/2}\rho^{1/2}\left(\frac{1}{(c-1)\rho}+\frac{1}{\rho}\right)=\frac{c}{(c-1)^{1/2}\rho^{1/2}}.$$

Plugging (5.172) into (5.171) we get

$$|\Theta_2'(\nu)|<\frac{1}{(1-\sqrt{\rho})^2(\mu(\nu)-\rho)^2}\left(\frac{8c+4}{(c-1)^{1/2}\rho^{1/2}}\nu\right)<\frac{8c+4}{(1-\sqrt{\rho})^2(c-1)^{5/2}\rho^{5/2}}\nu. \quad (5.173)$$

where in the last inequality we used $\mu(\nu)>c\rho$. Using (5.164), (5.173) and Lemma 5.7-(a), we have

$$|h_2'(x)|=\frac{e^x|\Theta_2'(e^x)|}{|\Theta_2(e^x)|}<\frac{e^{2x}\frac{8c+4}{(1-\sqrt{\rho})^2(c-1)^{5/2}\rho^{5/2}}}{e^{2x}\frac{1}{4(1-\sqrt{\rho})^2(c-1)^2\rho^2}}<\frac{32c+16}{(c-1)^{1/2}\rho^{1/2}}.$$

This combined with (5.161) leads us to

$$\ln\left(\frac{\Theta_2(\nu_2)}{\Theta_2(\nu_1)}\right)<\frac{32c+16}{(c-1)^{1/2}\rho^{1/2}}\ln\left(\frac{\nu_2}{\nu_1}\right),$$

which completes the proof of the lemma.

Proof of Lemma 5.9: We first note that

$$\frac{d}{dx}\frac{16x+8}{(x-1)^{1/2}\rho^{1/2}}=\frac{8x-4}{(x-1)^{3/2}\rho^{1/2}}\geq 0,$$

for $x>1$. Therefore, both the base and exponent in $h(x)$ are increasing functions of x, implying that $h(x)$ is an increasing function for $x\geq 1$. We aim to find $\lim_{x\to 1^+}h(x)$. To this end, we first evaluate $\lim_{x\to 1^+}\log h(x)$, where $\log h(x)=2\log x+\frac{16x+8}{(x-1)^{1/2}\rho^{1/2}}\log x$. Using the L'Hôpital's rule, we have

$$\begin{aligned}\lim_{x\to 1^+}\frac{(16x+8)\log x}{(x-1)^{1/2}\rho^{1/2}}&=\lim_{x\to 1^+}\frac{\frac{d}{dx}(16x+8)\log x}{\frac{d}{dx}(x-1)^{1/2}\rho^{1/2}}\\&=\lim_{x\to 1^+}\frac{16\log x+16+\frac{8}{x}}{\frac{1}{2}(x-1)^{-1/2}\rho^{1/2}}\\&=\lim_{x\to 1^+}\frac{16(x-1)^{1/2}(2\log x+2+1/x)}{\rho^{1/2}}=0.\end{aligned}$$

Thus, $\lim_{x\to 1^+}\log h(x)=0$, or equivalently, $\lim_{x\to 1^+}h(x)=1$. It is also straight-forward to see that $\lim_{x\to\infty}h(x)=\infty$. That means $h(x)$ is a continuous and monotonically

increasing function over $x \in (1, \infty)$, that takes any value from 1 to ∞. Hence, for any $\delta_5 > 0$, there exists some $c_0 > 1$ such that $h(c_0) = 1 + \delta_5$. Therefore, $h(x) \le h(c_0) = 1 + \delta_5$ for $1 \le x \le c_0$. This completes the proof of the lemma.

Proof of Lemma 5.11: We start with proof of part (a). First note that $0 \le \alpha(0) = \sqrt{\nu_0 K} \le 1$. Next we show that $0 \le \alpha(t+1) \le \alpha(t)$ for every $t \ge 1$. Note that $\alpha^2(t+1) = \frac{\nu(t+1)}{\nu(t)}(1 - \alpha(t+1))\alpha^2(t)$ provides a quadratic equation in $\alpha(t+1)$, leading to

$$
\begin{aligned}
0 \le \alpha(t+1) &= \frac{1}{2}\left(-\frac{\nu(t+1)}{\nu(t)}\alpha^2(t) + \sqrt{\frac{\nu^2(t+1)}{\nu^2(t)}\alpha^4(t) + 4\frac{\nu(t+1)}{\nu(t)}\alpha^2(t)}\right) \qquad (5.174)\\
&\overset{(a)}{\le} \frac{1}{2}\left(-\frac{\nu(t+1)}{\nu(t)}\alpha^2(t) + \frac{\nu(t+1)}{\nu(t)}\alpha^2(t) + 2\sqrt{\frac{\nu(t+1)}{\nu(t)}}\alpha(t)\right)\\
&= \sqrt{\frac{\nu(t+1)}{\nu(t)}}\alpha(t)\\
&\overset{(b)}{\le} \alpha(t) \le 1,
\end{aligned}
$$

where (a) follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and (b) holds since $\{\nu(t)\}$ is a non-increasing step-size sequence. This proves the claim in part (a).

Using the mean value theorem for $\nu(t) = \frac{\nu_0}{(t+\tau)^\sigma}$, we get $\nu(t+1) - \nu(t) = \nu'(t_0)$ for some $t_0 \in [t, t+1]$. Hence, we arrive at

$$
-\Delta\nu(t) = \nu(t) - \nu(t+1) = -\nu'(t_0) = \sigma\frac{\nu_0}{(t_0+\tau)^{\sigma+1}} \le \sigma\frac{\nu_0}{(t+\tau)^{\sigma+1}} = \sigma\frac{\nu(t)}{t+\tau}.
$$

This shows the claim in part (b).

We refer readers to Lemma 16 of [121] for the proof of $\alpha(t) \le \frac{2}{t+1}$. Starting from (5.174), we have

$$
\alpha(t+1) = \frac{2}{1 + \sqrt{1 + 4\frac{\nu(t)}{\nu(t+1)\alpha^2(t)}}}.
$$

This implies

$$\begin{aligned}\frac{1}{\alpha(t+1)}-\sqrt{\frac{\nu(t)}{\nu(t+1)}}\frac{1}{\alpha(t)} &= \frac{1}{2}+\frac{1}{2}\left[\sqrt{1+4\frac{\nu(t)}{\nu(t+1)\alpha^2(t)}}-\sqrt{4\frac{\nu(t)}{\nu(t+1)\alpha^2(t)}}\right]\\ &= \frac{1}{2}+\frac{1}{2}\frac{1}{\sqrt{1+4\frac{\nu(t)}{\nu(t+1)\alpha^2(t)}}+\sqrt{4\frac{\nu(t)}{\nu(t+1)\alpha^2(t)}}}\\ &\leq \frac{1}{2}+\frac{1}{2}\frac{1}{2\sqrt{4\frac{\nu(t)}{\nu(t+1)\alpha^2(t)}}}\\ &= \frac{1}{2}+\frac{1}{8}\sqrt{\frac{\nu(t+1)}{\nu(t)}}\alpha(t)\leq\frac{1}{2}+\frac{\alpha_0}{8}. \end{aligned} \tag{5.175}$$

We can rewrite (5.175) as

$$\frac{\sqrt{\nu(t+1)}}{\alpha(t+1)}-\frac{\sqrt{\nu(t)}}{\alpha(t)}\leq\left(\frac{1}{2}+\frac{\alpha_0}{8}\right)\sqrt{\nu(t+1)}. \tag{5.176}$$

Applying a telescopic summation in (5.176), we arrive at

$$\begin{aligned}\frac{\sqrt{\nu(t)}}{\alpha(t)} &\leq \frac{\sqrt{\nu_0}}{\alpha_0}+\left(\frac{1}{2}+\frac{\alpha_0}{8}\right)\sum_{k=1}^{t}\sqrt{\nu(k)}\\ &< \frac{\sqrt{\nu_0}}{\alpha_0}+\left(\frac{1}{2}+\frac{\alpha_0}{8}\right)\sqrt{\nu_0}\sum_{k=1}^{t}\frac{1}{(k+\tau)^{\sigma/2}}\\ &\leq \frac{\sqrt{\nu_0}}{\alpha_0}+\left(\frac{1}{2}+\frac{\alpha_0}{8}\right)\sqrt{\nu_0}\int_0^t\frac{1}{(x+\tau)^{\sigma/2}}dx\\ &= \frac{\sqrt{\nu_0}}{\alpha_0}+\left(\frac{1}{2}+\frac{\alpha_0}{8}\right)\frac{2\sqrt{\nu_0}}{2-\sigma}\left[(t+\tau)^{1-\sigma/2}-\tau^{1-\sigma/2}\right]\\ &\leq \left(\frac{1}{2}+\frac{\alpha_0}{8}\right)\frac{2\sqrt{\nu_0}}{2-\sigma}(t+\tau)^{1-\sigma/2}, \end{aligned}$$

where the last inequality holds for $\tau>\left(\frac{4(2-\sigma)}{\alpha_0(4+\alpha_0)}\right)^{\frac{1}{1-\sigma/2}}$. This together with the fact that $\nu(t)=\frac{\nu_0}{(t+\tau)^\sigma}$ leads to the desired lower bound on $\alpha(t)$. This completes the proof of the lemma.

References

[1] Sandra Pattison. 35 streaming services statistics you need to know in 2024. https://www.cloudwards.net/streaming-services-statistics/, July 2024. Accessed: 2024-7-8.

[2] AI funding united states 2011-2019. https://www.statista.com/statistics/672712/ai-funding-united-states/. Accessed: 2024-7-8.

[3] Nicholas Mitsakos. Distributed machine learning can bring healthcare breakthroughs. https://arcadiacapitalgroup.com/2020/03/distributed-machine-learning-can-bring-healthcare-breakthroughs/, March 2020. Accessed: 2024-7-8.

[4] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *NeurIPS*, pages 1709–1720, 2017.

[5] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *NeurIPS*, pages 1509–1519, 2017.

[6] Amirhossein Reisizadeh, Hossein Taheri, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. Robust and communication-efficient collaborative learning. In *NeurIPS*, pages 8388–8399, 2019.

[7] Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*, 2017.

[8] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *arXiv preprint arXiv:1809.07599*, 2018.

[9] Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations. In *NeurIPS*, pages 14695–14706, 2019.

[10] Tianyi Chen, Georgios Giannakis, Tao Sun, and Wotao Yin. Lag: Lazily aggregated gradient for communication-efficient distributed learning. *Advances in neural information processing systems*, 31, 2018.

[11] Robert Hönig, Yiren Zhao, and Robert Mullins. Dadaquant: Doubly-adaptive quantization for communication-efficient federated learning. In *International Conference on Machine Learning*, pages 8852–8866. PMLR, 2022.

[12] Divyansh Jhunjhunwala, Advait Gadhikar, Gauri Joshi, and Yonina C Eldar. Adaptive quantization of model updates for communication-efficient federated learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3110–3114. IEEE, 2021.

[13] Yuzhu Mao, Zihao Zhao, Guangfeng Yan, Yang Liu, Tian Lan, Linqi Song, and Wenbo Ding. Communication-efficient federated learning with adaptive quantization. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–26, 2022.

[14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[15] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[16] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

[17] Andrew C Yao. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*, pages 160–164. IEEE, 1982.

[18] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.

[19] Payman Mohassel and Yupeng Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE symposium on security and privacy (SP)*, pages 19–38. IEEE, 2017.

[20] Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 169–178, 2009.

[21] Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, 51(4):1–35, 2018.

[22] Mohammad Ali Maddah-Ali and Urs Niesen. Fundamental limits of caching. *IEEE Transactions on Information Theory*, 60(5):2856–2867, 2014.

[23] Mohammad Ali Maddah-Ali and Urs Niesen. Decentralized coded caching attains order-optimal memory-rate tradeoff. *IEEE/ACM Transactions On Networking*, 23(4):1029–1040, 2014.

[24] Sung Hoon Lim, Chien-Yi Wang, and Michael Gastpar. Information-theoretic caching: The multi-user case. *IEEE Transactions on Information Theory*, 63(11):7018–7037, 2017.

[25] Qian Yu, Mohammad Ali Maddah-Ali, and A Salman Avestimehr. The exact rate-memory tradeoff for caching with uncoded prefetching. *IEEE Transactions on Information Theory*, 64(2):1281–1296, 2018.

[26] Ramtin Pedarsani, Mohammad Ali Maddah-Ali, and Urs Niesen. Online coded caching. *IEEE/ACM Transactions on Networking*, 24(2):836–845, 2015.

[27] Mingyue Ji, Giuseppe Caire, and Andreas F Molisch. Wireless device-to-device caching networks: Basic principles and system performance. *IEEE Journal on Selected Areas in Communications*, 34(1):176–189, 2015.

[28] Mingyue Ji, Giuseppe Caire, and Andreas F Molisch. Fundamental limits of caching in wireless d2d networks. *IEEE Transactions on Information Theory*, 62(2):849–869, 2015.

[29] Urs Niesen and Mohammad Ali Maddah-Ali. Coded caching with nonuniform demands. *IEEE Transactions on Information Theory*, 63(2):1146–1158, 2016.

[30] Jinbei Zhang, Xiaojun Lin, and Xinbing Wang. Coded caching under arbitrary popularity distributions. *IEEE Transactions on Information Theory*, 64(1):349–366, 2017.

[31] Hadi Reisizadeh, Mohammad Ali Maddah-Ali, and Soheil Mohajer. Erasure coding for decentralized coded caching. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1715–1719. IEEE, 2018.

[32] Hadi Reisizadeh, Mohammad Ali Maddah-Ali, and Soheil Mohajer. Subspace coding for coded caching: Decentralized and centralized placements meet for three users. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 677–681. IEEE, 2019.

[33] Zhi Chen, Pingyi Fan, and K Ben Letaief. Fundamental limits of caching: Improved bounds for small buffer users. *arXiv preprint arXiv:1407.1935*, 2014.

[34] Yi-Peng Wei and Sennur Ulukus. Novel decentralized coded caching through coded prefetching. In *2017 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2017.

[35] Mohammad Ali Maddah-Ali and Urs Niesen. Cache-aided interference channels. *IEEE Transactions on Information Theory*, 65(3):1714–1724, 2019.

[36] Navid Naderializadeh, Mohammad Ali Maddah-Ali, and Amir Salman Avestimehr. Fundamental limits of cache-aided interference management. *IEEE Transactions on Information Theory*, 63(5):3092–3107, 2017.

[37] Jad Hachem, Urs Niesen, and Suhas Diggavi. A layered caching architecture for the interference channel. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 415–419. IEEE, 2016.

[38] Roy Timo and Michele Wigger. Joint cache-channel coding over erasure broadcast channels. In *2015 International Symposium on Wireless Communication Systems (ISWCS)*, pages 201–205. IEEE, 2015.

[39] Shirin Saeedi Bidokhti, Michèle Wigger, and Roy Timo. Erasure broadcast networks with receiver caching. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1819–1823. IEEE, 2016.

[40] Shirin Saeedi Bidokhti, Michele Wigger, and Aylin Yener. Benefits of cache assignment on degraded broadcast channels. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1222–1226. IEEE, 2017.

[41] Mohammad Mohammadi Amiri and Deniz Gündüz. Cache-aided content delivery over erasure broadcast channels. *IEEE Transactions on Communications*, 66(1):370–381, 2018.

[42] Khac-Hoang Ngo, Sheng Yang, and Mari Kobayashi. Scalable content delivery with coded caching in multi-antenna fading channels. *IEEE Transactions on Wireless Communications*, 17(1):548–562, 2018.

[43] Jingjing Zhang, Felix Engelmann, and Petros Elia. Coded caching for reducing csit-feedback in wireless communications. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1099–1105. IEEE, 2015.

[44] Jingjing Zhang and Petros Elia. Fundamental limits of cache-aided wireless bc: Interplay of coded-caching and csit feedback. *IEEE Transactions on Information Theory*, 63(5):3142–3160, 2017.

[45] A Salman Avestimehr, Suhas N Diggavi, and NC David. Wireless network information flow: A deterministic approach. *IEEE Transactions on Information theory*, 57(4):1872–1905, 2011.

[46] NC David and Roy D Yates. Fading broadcast channels with state information at the receivers. *IEEE Transactions on Information Theory*, 58(6):3453–3471, 2012.

[47] Roy D Yates and David Tse. K user fading broadcast channels with csi at the receivers. In *2011 Information Theory and Applications Workshop*, pages 1–6. IEEE, 2011.

[48] Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

[49] Robert M Gray and Aaron D Wyner. Source coding for a simple network. *Bell System Technical Journal*, 53(9):1681–1721, 1974.

[50] Michael Gastpar, Bixio Rimoldi, and Martin Vetterli. To code, or not to code: Lossy source-channel communication revisited. *IEEE Transactions on Information Theory*, 49(5):1147–1158, 2003.

[51] Angelia Nedić, Alex Olshevsky, Asuman Ozdaglar, and John N Tsitsiklis. On distributed averaging algorithms and quantization effects. *IEEE Trans. Automat. Contr.*, 54(11):2506–2517, 2009.

[52] Angelia Nedić, Asuman Ozdaglar, and Pablo A Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Trans. Automat. Contr.*, 55(4):922–938, 2010.

[53] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NeurIPS*, pages 315–323, 2013.

[54] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[55] Jakub KoneVcnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[56] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, et al. Large scale distributed deep networks. In *NeurIPS*, pages 1223–1231, 2012.

[57] DuVsan Jakovetić, Joao Xavier, and José MF Moura. Fast distributed gradient methods. *IEEE Trans. Automat. Contr.*, 59(5):1131–1146, 2014.

[58] Adel Aghajan and Behrouz Touri. Distributed optimization over dependent random networks. *preprint arXiv:2010.01956*, 2020.

[59] Hadi Reisizadeh, Behrouz Touri, and Soheil Mohajer. Distributed optimization over time-varying graphs with imperfect sharing of information. *IEEE Transactions on Automatic Control*, 68(7):4420–4427, 2022.

[60] Hadi Reisizadeh, Behrouz Touri, and Soheil Mohajer. Dimix: Diminishing mixing for sloppy agents. *SIAM Journal on Optimization*, 33(2):978–1005, 2023.

[61] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *NeurIPS*, pages 1299–1309, 2018.

[62] Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In *NeurIPS*, pages 9850–9861, 2018.

[63] Min Ye and Emmanuel Abbe. Communication-computation efficient gradient coding. *arXiv preprint arXiv:1802.03475*, 2018.

[64] Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In *NeurIPS*, pages 7652–7662, 2018.

[65] Toby Berger, Zhen Zhang, and Harish Viswanathan. The ceo problem [multiterminal source coding]. *IEEE Transactions on Information Theory*, 42(3):887–902, 1996.

[66] Jun Chen, Xin Zhang, Toby Berger, and Stephen B Wicker. An upper bound on the sum-rate distortion function and its corresponding rate allocation schemes for

the ceo problem. *IEEE Journal on Selected Areas in Communications*, 22(6):977–987, 2004.

[67] Yasutada Oohama. The rate-distortion function for the quadratic gaussian ceo problem. *IEEE Transactions on Information Theory*, 44(3):1057–1070, 1998.

[68] Ziyue Huang, Yilei Wang, Ke Yi, et al. Optimal sparsity-sensitive bounds for distributed mean estimation. *Proc. of NeurIPS*, 32:6371–6381, 2019.

[69] Mikhail Borisovich Nevelson and Rafail Zalmanovich Hasminski. *Stochastic approximation and recursive estimation*, volume 47. American Mathematical Soc., 1976.

[70] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

[71] Tuncer Can Aysal, Mehmet Ercan Yildiz, Anand D Sarwate, and Anna Scaglione. Broadcast gossip algorithms for consensus. *IEEE Trans. Signal Process.*, 57(7):2748–2761, 2009.

[72] Shahin Nikookhoy, Jie Lu, and Choon Yik Tang. Distributed convex optimization with identical constraints. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 2926–2931. IEEE, 2011.

[73] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pages 2021–2031, 2020.

[74] Michael Rabbat and Robert Nowak. Distributed optimization in sensor networks. In *IPSN*, pages 20–27, 2004.

[75] Soummya Kar and José MF Moura. Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise. *IEEE Trans. Signal Process.*, 57(1):355–369, 2008.

[76] Giovanni Neglia, Giuseppe Reina, and Sara Alouf. Distributed gradient optimization for epidemic routing: A preliminary evaluation. In *2009 2nd IFIP wireless days (WD)*, pages 1–6. IEEE, 2009.

[77] Konstantinos Tsianos, Sean Lawlor, and Michael Rabbat. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In *Annu. Allerton Conf. Commun. Control Comput.*, pages 1543–1550. IEEE, 2012.

[78] Sundhar Srinivasan Ram, Venugopal V Veeravalli, and Angelia Nedic. Distributed non-autonomous power control through distributed convex optimization. In *IEEE INFOCOM 2009*, pages 3001–3005. IEEE, 2009.

[79] Lin Xiao and Stephen Boyd. Optimal scaling of a gradient method for distributed resource allocation. *Journal of optimization theory and applications*, 129(3):469–488, 2006.

[80] Alejandro Ribeiro. Ergodic stochastic optimization algorithms for wireless communication and networking. *IEEE Trans. Signal Process.*, 58(12):6369–6386, 2010.

[81] Ioannis D Schizas, Alejandro Ribeiro, and Georgios B Giannakis. Consensus in ad hoc WSNs with noisy links—part I: Distributed estimation of deterministic signals. *IEEE Trans. Signal Process.*, 56(1):350–364, 2007.

[82] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[83] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[84] Alexandros G Dimakis, Soummya Kar, José MF Moura, Michael G Rabbat, and Anna Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

[85] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Trans. Automat. Contr.*, 57(3):592–606, 2011.

[86] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

[87] Chenguang Xi and Usman A Khan. Dextra: A fast algorithm for optimization over directed graphs. *IEEE Trans. Automat. Contr.*, 62(10):4980–4993, 2017.

[88] Tatiana Tatarenko and Behrouz Touri. Non-convex distributed optimization. *IEEE Trans. Automat. Contr.*, 62(8):3744–3757, 2017.

[89] Jinshan Zeng and Wotao Yin. On nonconvex decentralized gradient descent. *IEEE Trans. Signal Process.*, 66(11):2834–2848, 2018.

[90] Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Trans. Automat. Contr.*, 54(1):48–61, 2009.

[91] Angelia Nedić and Alex Olshevsky. Distributed optimization of strongly convex functions on directed time-varying graphs. In *IEEE Glob. Conf. Signal Inf. Process. Proc.*, pages 329–332, 2013.

[92] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Trans. Automat. Contr.*, 60(3):601–615, 2014.

[93] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. An exact quantized decentralized gradient descent algorithm. *IEEE Trans. Signal Process.*, 67(19):4934–4947, 2019.

[94] Marcos M Vasconcelos, Thinh T Doan, and Urbashi Mitra. Improved convergence rate for a distributed two-time-scale gradient method under random quantization. In *Proc. IEEE Conf. Decis. Control*, pages 3117–3122. IEEE, 2021.

[95] Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *ICML*, pages 3478–3487, 2019.

[96] Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *ICLR*, 2020.

[97] Kunal Srivastava and Angelia Nedic. Distributed asynchronous constrained stochastic optimization. *IEEE journal of selected topics in signal processing*, 5(4):772–790, 2011.

[98] Paolo Di Lorenzo and Gesualdo Scutari. NEXT: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

[99] Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Convergence of asynchronous distributed gradient methods over stochastic networks. *IEEE Trans. Automat. Contr.*, 63(2):434–448, 2017.

[100] Rick Durrett. *Probability: theory and examples*. Cambridge University Press, 2010.

[101] Peter M DeMarzo, Dimitri Vayanos, and Jeffrey Zwiebel. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly journal of economics*, 118(3):909–968, 2003.

[102] Behrouz Touri and Angelia Nedić. Product of random stochastic matrices. *IEEE Trans. Automat. Contr.*, 59(2):437–448, 2013.

[103] Behrouz Touri and Cedric Langbort. On endogenous random consensus and averaging dynamics. *IEEE Transactions on Control of Network Systems*, 1(3):241–248, 2014.

[104] Adel Aghajan and Behrouz Touri. Distributed optimization over dependent random networks. *IEEE Trans. Automat. Contr.*, 2022.

[105] Vijay R Konda and John N Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability*, 14(2):796–819, 2004.

[106] Thinh T Doan. Nonlinear two-time-scale stochastic approximation convergence and finite-time performance. *IEEE Trans. Automat. Contr.*, 2022.

[107] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[108] S Sundhar Ram, Angelia Nedić, and Venu V Veeravalli. Asynchronous gossip algorithm for stochastic optimization: Constant stepsize analysis. In *Recent Advances in Optimization and its Applications in Engineering*, pages 51–60. Springer, 2010.

[109] Soomin Lee and Angelia Nedić. Asynchronous gossip-based random projection algorithms over networks. *IEEE Trans. Automat. Contr.*, 61(4):953–968, 2015.

[110] Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

[111] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[112] Ashwin Verma, Marcos M Vasconcelos, Urbashi Mitra, and Behrouz Touri. Maximal dissent: a fast way to agree in distributed convex optimization. *arXiv preprint arXiv:2205.00647*, 2022.

[113] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[114] Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

[115] Behrouz Touri and Angelia Nedić. On existence of a quadratic comparison function for random weighted averaging dynamics and its implications. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 3806–3811. IEEE, 2011.

[116] Yongqiang Wang and H Vincent Poor. Decentralized stochastic optimization with inherent privacy protection. *IEEE Transactions on Automatic Control*, 68(4):2293–2308, 2022.

[117] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

[118] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–964. IEEE, 2022.

[119] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

[120] Rudrajit Das, Satyen Kale, Zheng Xu, Tong Zhang, and Sujay Sanghavi. Beyond uniform lipschitz condition in differentially private optimization. In *International Conference on Machine Learning*, pages 7066–7101. PMLR, 2023.

[121] Guannan Qu and Na Li. Accelerated distributed nesterov gradient descent. *IEEE Trans. Automat. Contr.*, 65(6):2566–2581, 2019.

[122] Yongqiang Wang and Tamer Başar. Quantization enabled privacy protection in decentralized stochastic optimization. *IEEE Transactions on Automatic Control*, 2022.

[123] Ilan Lobel and Asuman Ozdaglar. Convergence analysis of distributed subgradient methods over random networks. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 353–360. IEEE, 2008.

[124] Ilan Lobel, Asuman Ozdaglar, and Diego Feijer. Distributed multi-agent optimization with state-dependent communication. *Mathematical programming*, 129(2):255–284, 2011.

[125] Ion Matei and John S Baras. Performance evaluation of the consensus-based distributed subgradient method under random communication topologies. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):754–771, 2011.

[126] Minghui Zhu and Sonia Martínez. On distributed convex optimization under inequality and equality constraints. *IEEE Transactions on Automatic Control*, 57(1):151–164, 2011.

[127] Alex Olshevsky. Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control. *arXiv preprint arXiv:1411.4186*, 2014.

[128] Ermin Wei and Asuman Ozdaglar. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450. IEEE, 2012.

[129] Tsung-Hui Chang, Mingyi Hong, and Xiangfeng Wang. Multi-agent distributed optimization via inexact consensus admm. *IEEE Transactions on Signal Processing*, 63(2):482–497, 2014.

[130] Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

[131] Aryan Mokhtari, Wei Shi, Qing Ling, and Alejandro Ribeiro. Dqm: Decentralized quadratically approximated alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 64(19):5158–5173, 2016.

[132] Nikolaos Chatzipanagiotis, Darinka Dentcheva, and Michael M Zavlanos. An augmented lagrangian method for distributed optimization. *Mathematical Programming*, 152:405–434, 2015.

[133] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[134] Aryan Mokhtari and Alejandro Ribeiro. Dsa: Decentralized double stochastic averaging gradient algorithm. *The Journal of Machine Learning Research*, 17(1):2165–2199, 2016.

[135] Zonghao Huang, Rui Hu, Yuanxiong Guo, Eric Chan-Tin, and Yanmin Gong. Dp-admm: Admm-based distributed learning with differential privacy. *IEEE Transactions on Information Forensics and Security*, 15:1002–1012, 2019.

[136] Lichao Sun, Jianwei Qian, and Xun Chen. Ldp-fl: Practical private aggregation in federated learning with local differential privacy. *arXiv preprint arXiv:2007.15789*, 2020.

[137] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[138] Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2521–2529. PMLR, 2021.

[139] Muah Kim, Onur Günlü, and Rafael F Schaefer. Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2650–2654. IEEE, 2021.

[140] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

[141] Jorge Cortés, Geir E Dullerud, Shuo Han, Jerome Le Ny, Sayan Mitra, and George J Pappas. Differential privacy in control and network systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 4252–4272. IEEE, 2016.

[142] Huan Gao, Yongqiang Wang, and Angelia Nedić. Dynamics based privacy preservation in decentralized optimization. *Automatica*, 151:110878, 2023.

[143] Yongqiang Wang and Angelia Nedić. Tailoring gradient methods for differentially-private distributed optimization. *IEEE Transactions on Automatic Control*, 2023.

[144] Junlong Zhu, Changqiao Xu, Jianfeng Guan, and Dapeng Oliver Wu. Differentially private distributed online algorithms over time-varying directed networks. *IEEE Transactions on Signal and Information Processing over Networks*, 4(1):4–17, 2018.

[145] Yongyang Xiong, Jinming Xu, Keyou You, Jianxing Liu, and Ligang Wu. Privacy-preserving distributed online optimization over unbalanced digraphs via subgradient rescaling. *IEEE Transactions on Control of Network Systems*, 7(3):1366–1378, 2020.

[146] Dongyu Han, Kun Liu, Yeming Lin, and Yuanqing Xia. Differentially private distributed online learning over time-varying digraphs via dual averaging. *International Journal of Robust and Nonlinear Control*, 32(5):2485–2499, 2022.

[147] Tie Ding, Shanying Zhu, Jianping He, Cailian Chen, and Xinping Guan. Differentially private distributed optimization via state and direction perturbation in multiagent systems. *IEEE Transactions on Automatic Control*, 67(2):722–737, 2021.

[148] Zhenqi Huang, Sayan Mitra, and Nitin Vaidya. Differentially private distributed optimization. In *Proceedings of the 16th International Conference on Distributed Computing and Networking*, pages 1–10, 2015.

[149] Yongqiang Wang and Angelia Nedić. Differentially-private distributed optimization with guaranteed optimality. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 4162–4169. IEEE, 2023.

[150] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[151] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

[152] Aurélien Bellet, Rachid Guerraoui, Mahsa Taziki, and Marc Tommasi. Personalized and private peer-to-peer machine learning. In *International conference on artificial intelligence and statistics*, pages 473–481. PMLR, 2018.

[153] Edwige Cyffers, Mathieu Even, Aurélien Bellet, and Laurent Massoulié. Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging. *Advances in Neural Information Processing Systems*, 35:15889–15902, 2022.

[154] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.