Title
Learning under Imperfections by Networked Agents

Author
Zhao, Xiaochuan

Publication Date
2014

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Learning under Imperfections by Networked Agents

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Electrical Engineering

by

Xiaochuan Zhao

2014

Abstract of the Dissertation

Learning under Imperfections by Networked Agents

by

Xiaochuan Zhao

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2014

Professor Ali H. Sayed, Chair

Distributed learning deals with the problem of optimizing aggregate cost functions by networked agents from streaming data. This scenario arises in many contexts including distributed estimation, machine learning, resource allocation, and in the modeling of flocking and swarming behavior by biological networks. Among several available solutions such as consensus and incremental strategies, the class of diffusion strategies has proven to be particularly attractive because these techniques are scalable, robust, fully-distributed, and endow networks with real-time adaptation and learning abilities.

One key challenge in real applications is that networked agents generally face many types of asynchronous imperfections, such as random link failures, random data arrival times, noisy links, random topology changes, agents turning on and off randomly, and even drifting objectives. This dissertation provides a detailed analysis of the stability and performance of asynchronous diffusion strategies for solving distributed optimization and adaptation problems over networks in the presence of such imperfections. Conditions are developed to ensure the stability of the mean-square and mean-fourth-order moments of the network error vectors;

closed-form expressions are derived to reveal how the network parameters influence the learning behavior; and the performance of the asynchronous networks is then compared against centralized solutions and synchronous networks. One notable conclusion is that the mean-square performance of asynchronous networks degrades only in the order of μ, which is a small step-size parameter, while the convergence rate remains largely unaltered. A second notable conclusion is that even under the influence of asynchronous events, all agents in the adaptive network can still reach an $O(\mu^{1+\gamma})$ near-agreement with some constant $\gamma > 0$, while approaching the desired solution within $O(\mu)$ accuracy. These theoretical results provide a solid justification for the remarkable resilience of cooperative networks in the face of random imperfections at multiple levels: agents, links, data arrivals, and topology.

The dissertation also examines a second challenging form of uncertainty arising from agents in a network pursuing different objectives or observing data arising from different unknown models. In these cases, indiscriminate cooperation will lead to undesired results. A useful adaptive clustering and learning strategy is developed in order to allow agents to learn which neighbors should be trusted and which other neighbors should be ignored. The resulting procedure enables agents to identify their grouping and to attain improved learning performance.

The dissertation of Xiaochuan Zhao is approved.

Mario Gerla

Lara Dolecek

Gregory Pottie

Ali H. Sayed, Committee Chair

University of California, Los Angeles

2014

To my parents and my dear wife.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost I want to thank my advisor, Professor Ali H. Sayed, who has supported and guided me through my PhD studies in the past five years. His curiosity, hard-working attitude, high standards in research, and rich experience are much more than what I had expected from an advisor. He always advised us to keep curiosity alive and never stop exploiting interesting questions and research topics. I remember the many days and nights we exchanged emails and phone calls to discuss technical questions in my work. I have been very impressed and inspired by his persistent pursuit of high-quality articles — I feel that he put much more effort himself on revising and perfecting our articles than we did. I feel truly thankful for being able to work with him during my PhD study. I believe what I learned from him will help me march forward with confidence.

I would also like to thank my colleagues and friends in the Adaptive Systems Laboratory (ASL): Zaid Towfic, Jianshu Chen, Sheng-Yuan Tu, Shang-Kee Ting, and Chung-Kai Yu. We spent so much time together discussing interesting problems on the white board — which, in my opinion, is the most productive way to conduct research. We helped each other and learned from each other. I am also grateful for having the opportunity to make friends from around the world and meeting many visitors to the lab: Paolo Di Lorenzo from Italy, Alexander Bertrand from Belgium, Victor Lora from France, Jae-Woo Lee from Korea, Oyvind L. Rortveit from Norway, Ricardo Merched and Cassio G. Lopes from Brazil, Reza Abdolee and Milad A. Toutounchian from Canada, Mohammad-Reza from Sweden, and Sergio Valcarcel Macua and Jesus F. Bes from Spain. I really enjoyed the discussions with all of them.

Lastly and most importantly, I would like to thank my family for all their love and support. I am thankful for having my dear wife, Yan Fang. She brought so much joy into my life, and gave me the most support when I struggled. Recently, she has given birth to our son — how amazing my life can be! I must thank my beloved parents. I owe them too much in these years for not having enough time to spend with them. Without their deep love and endless support, I would have never had the opportunity to pursue my dream. I love them with all my heart.

The work in this dissertation is based upon work supported by the National Science Foundation under grants CCF-0942936 and CCF-1011918. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Vita

2002–2006	B.S. in Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), China.
2006–2009	M.S. in Communication and Information Systems, BUPT, China.
2008–2009	Intern, Intel China Research Center, Beijing, China.
2009–present	Research and Teaching Assistant, Department of Electrical Engineering, University of California, Los Angeles.
2012	Intern, Qualcomm, Santa Clara, California.

CHAPTER 1

Introduction

Distributed learning arises when a global objective needs to be achieved through local cooperation among networked agents that are subjected to streaming data. This problem is prevalent in many contexts, including distributed estimation [1–8], distributed machine learning [9–12], resource allocation [13, 14], and in the modeling of flocking and swarming behavior by biological networks [15–19]. Several useful decentralized solutions, such as consensus strategies [20–28], incremental strategies [29–33], and diffusion strategies [5, 6, 8, 12, 34, 35], have been developed for this purpose. The diffusion strategies are particularly attractive because they are scalable, robust, fully-distributed, and endow networks with real-time adaptation and learning abilities.

One key challenge in real applications is that networked agents generally face many types of asynchronous imperfections, such as random link failures, random data arrival times, noisy links, random topology changes, agents turning on and off randomly, and even drifting objectives. This dissertation provides a detailed analysis of the stability and performance of asynchronous diffusion strategies for solving distributed optimization and adaptation problems over networks in the presence of such imperfections. The dissertation also examines another challenging form of uncertainty arising from agents in the network pursuing different objectives or observing data arising from different unknown models. In these cases, indiscriminate cooperation generally leads to undesired or even catastrophic re-

sults.

In this chapter, we will first briefly review the distributed optimization problem and describe strategies for its solution. We will then describe the various forms of uncertainties that may occur. This chapter is concluded with a summary of the main contributions in the dissertation.

1.1 Distributed Optimization Over Networks

We consider a connected network consisting of N agents as shown in Fig. 1.1. The objective is to minimize, in a distributed manner, an aggregate cost function of the form:

$$\underset{w}{\text{minimize}} \quad J^{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_k(w) \tag{1.1}$$

where the $\{J_k(w)\}$ denote individual cost functions. The costs $\{J_k(w) : \mathbb{C}^M \mapsto \mathbb{R}; k = 1, 2, \ldots, N\}$ will be assumed to satisfy a certain smoothness condition (which will be described in Assumption 2.2 later in Chapter 2) and to be strongly convex over $\mathbb{C}^M$. They are also assumed to share a *common* and *unique* minimizer at $w^o \in \mathbb{C}^M$. Using arguments similar to [8, 12], we can motivate the following diffusion strategy for solving the distributed optimization problem (1.1) with *constant* step-sizes — see Fig. 1.1:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}) \quad \text{(adaptation)} \tag{1.2a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k}\, \boldsymbol{\psi}_{\ell,i} \quad \text{(combination)} \tag{1.2b}$$

where (1.2a) is a stochastic gradient approximation step for self-learning and (1.2b) is a convex combination step for social-learning. The iterate $\boldsymbol{w}_{k,i}$ is the estimate for w^o that is computed by agent k at iteration i. The iterate $\boldsymbol{\psi}_{k,i}$ is an intermediate solution that results from the adaptation step and will be shared

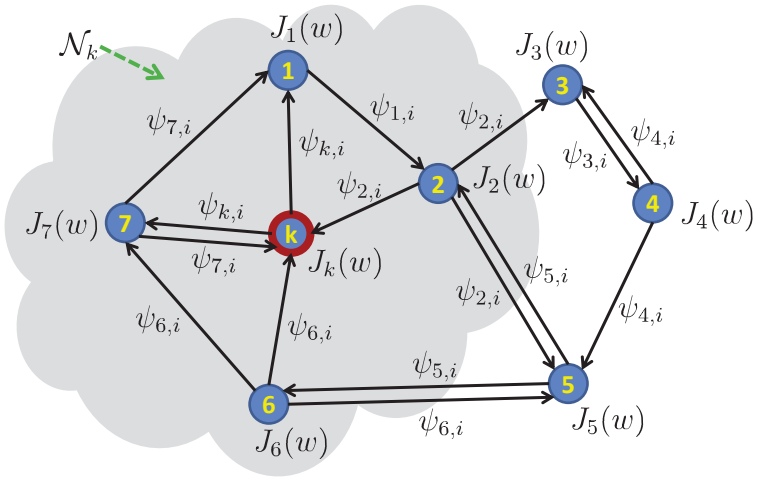


Figure 1.1: An illustration of a connected network with individual costs associated with the various agents.

with the neighbors in the combination step. The factor μ_k is a positive step-size parameter and the combination coefficients $\{a_{\ell k}\}$ are nonnegative parameters and are required to satisfy the following constraints:

$$\sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1, \quad \text{and} \quad \begin{cases} a_{\ell k} > 0, & \text{if } \ell \in \mathcal{N}_k \\ a_{\ell k} = 0, & \text{otherwise} \end{cases} \tag{1.3}$$

where $\mathcal{N}_k$ denotes the set of neighbors of agent k including k itself. If we collect these coefficients into an $N \times N$ matrix such that $[A]_{\ell k} = a_{\ell k}$, then condition (1.3) implies that A is a left-stochastic matrix, written as

$$A^{\mathsf{T}} \mathbb{1}_N = \mathbb{1}_N \tag{1.4}$$

where $\mathbb{1}_N$ is the $N \times 1$ vector with all entries equal to one. In (1.2a), a stochastic *approximation* for the true gradient vector is used because, in general, agents do not have sufficient information to acquire the true gradients. The difference between the true and approximate gradients is called gradient noise, which is

random in nature and will seep into the operation of the algorithm. Accordingly, the variables $\{\boldsymbol{w}_{k,i}\}$ in (1.2a)–(1.2b) are random quantities since they are subject to gradient noise, and we are denoting them by using the boldface notation. We will model the gradient noise, denoted by $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})$, as an additive random perturbation to the true gradient vector, i.e.,

$$\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}) = \nabla_{w^*} J_k(\boldsymbol{w}_{k,i-1}) + \boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1}) \tag{1.5}$$

It has been established in [6, 8, 12] that the diffusion network can achieve the desired objective, w^o, within $O(\mu_{\max})$ *regardless* of which cooperation policy, A, is used, where $\mu_{\max}$ denotes the largest step-size. It has also been shown in [36] for the common minimizer case that diffusion strategies can significantly reduce the average network error performance in comparison to non-cooperative agents. This is a major benefit for in-network cooperation. To measure the performance of the learning algorithms, we will choose the average network mean-square-deviation (MSD) as the metric, which is defined by

$$\text{MSD}^{\text{net}} \triangleq \lim_{i\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\,\|w^o - \boldsymbol{w}_{k,i}\|^2 \tag{1.6}$$

1.2 Asynchronous Distributed Learning over Networks

From the distributed solution (1.2a)–(1.2b), it can be observed that there are several implicit conditions assumed. It is assumed that all agents in the network operate in a perfect *synchronous* manner: the approximate gradients are acquired at the beginning of each iteration by all agents; the self-learning step is performed by all agents before the social learning step; the intermediate estimates are shared among neighboring agents and are assumed to be delivered successfully; and all agents are assumed to make use all the information received from their neighbors to update their local estimates. This ideal scenario is not always applicable

in real applications, where the measurement data may not arrive in time, the agents may turn on and off randomly to save power or suffer from malfunction, the communication links between agents may be subject to loss, etc. All these factors may prevent the distributed solutions (1.2a)–(1.2b) from operating properly. Therefore, it is important to investigate the performance of the diffusion strategy (1.2a)–(1.2b) under such uncertainties and imperfections.

There already exist several insightful studies in the literature on the performance of consensus and gossip type strategies in the presence of asynchronous events [21, 23, 26, 27] or changing topologies [2, 23, 26, 27, 37–42]. There are also some limited studies in the context of diffusion strategies [43, 44]. However, with the exception of the latter two works, the earlier references investigated pure averaging algorithms *without* the ability to respond to streaming data, assumed noise-free data, or relied on the use of diminishing step-size sequences. These conditions are problematic for adaptation and learning purposes when data is continually streaming in, since decaying step-sizes turn off adaptation eventually, and noise (including gradient noise) is always present.

In this dissertation, we remove these limitations. We allow for fairly general sources of uncertainties and random failures and permit them to occur simultaneously. To model the *asynchronous* behavior of the network, we modify the diffusion strategy (1.2a)–(1.2b) to the following form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \boldsymbol{\mu}_k(i)\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}) \tag{1.7a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\, \boldsymbol{\psi}_{\ell,i} \tag{1.7b}$$

where the $\{\boldsymbol{\mu}_k(i), \boldsymbol{a}_{\ell\, k}(i)\}$ are now *time-varying* and *random* step-sizes and combination coefficients, and $\boldsymbol{\mathcal{N}}_{k,i}$ denotes the *random* neighborhood of agent k at time i. The step-size parameters $\{\boldsymbol{\mu}_k(i)\}$ are nonnegative random variables, and the

combination coefficients $\{\boldsymbol{a}_{\ell k}(i)\}$ are also nonnegative random variables, which are required to satisfy the following constraints (compare with (1.3)):

$$\sum_{\ell\in\boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i) = 1, \;\; \text{and} \;\; \begin{cases} \boldsymbol{a}_{\ell k}(i) > 0, & \text{if } \ell \in \boldsymbol{\mathcal{N}}_{k,i} \\ \boldsymbol{a}_{\ell k}(i) = 0, & \text{otherwise} \end{cases} \tag{1.8}$$

The asynchronous strategy (1.7a)–(1.7b) described above is general enough to cover many situations of practical interest.

Note that the model does not impose any specific probabilistic distribution on the step-sizes, network topologies, or combination coefficients. For example, we can choose the sample space of each step-size $\boldsymbol{\mu}_k(i)$ to be the binary choice $\{0, \mu\}$ to model a random "on-off" behavior at each agent k for the purpose of saving power, waiting for data, or even due to random agent failures. Similarly, we can choose the sample space of each combination coefficient $\boldsymbol{a}_{\ell\,k}(i)$, $\ell \in \mathcal{N}_k\backslash\{k\}$, to be $\{0, a_{\ell k}\}$ to model a random "on-off" status for the link from agent ℓ to agent k at time i for the purpose of either saving communication cost or due to random link failures. If links are randomly chosen by agents such that at every time i there is only one other neighboring agent being communicated with, then we effectively mimic the random gossip strategies [2, 23, 38, 39, 42]. Note that the convex constraint (1.8) can be satisfied by adjusting the value of $\boldsymbol{a}_{kk}(i)$ according to the realizations of $\{\boldsymbol{a}_{\ell k}(i); \ell \in \boldsymbol{\mathcal{N}}_{k,i}\backslash\{k\}\}$. If the underlying topology is changing over time and the combination weights are also selected in a random manner, then we obtain the probabilistic diffusion strategy studied in [43, 44] or the random link or topology model studied in [26, 37, 40].

Since the parameter matrices $\boldsymbol{M}_i$ and $\boldsymbol{A}_i$ are assumed to be independent of each other and of any other random variable, the statistical dependency among the random variables $\{\boldsymbol{w}_i, \boldsymbol{\psi}_i, \boldsymbol{A}_i, \boldsymbol{M}_i\}$ is illustrated in Fig. 1.2.

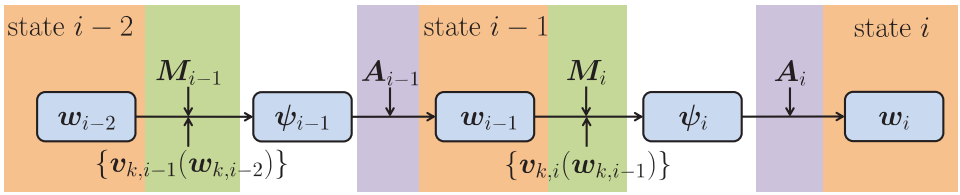


Figure 1.2: An illustration of the statistical dependency for the asynchronous diffusion strategy (1.7a)–(1.7b)

1.2.1 Asynchronous Centralized or Batch Learning

We will also examine the operation of batch or centralized implementations. An *asynchronous* centralized (batch) strategy seeks the optimal solution w^o of (1.1) by running a stochastic gradient batch algorithm of the following form:

$$\boldsymbol{w}_{\mathrm{c},i} = \boldsymbol{w}_{\mathrm{c},i-1} - \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\boldsymbol{\mu}_k(i)\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{\mathrm{c},i-1}) \tag{1.9}$$

where $\boldsymbol{w}_{\mathrm{c},i}$ denotes the iterate at time i, the $\{\boldsymbol{\pi}_k(i)\}$ are nonnegative convex fusion coefficients such that

$$\sum_{k=1}^{N} \boldsymbol{\pi}_k(i) = 1, \qquad \boldsymbol{\pi}_k(i) \geq 0 \tag{1.10}$$

for all $i \geq 0$, and the $\{\boldsymbol{\mu}_k(i)\}$ are the random step-sizes. In the batch algorithm (1.9), we use random step-sizes $\{\boldsymbol{\mu}_k(i)\}$ to account for random activity by the agents, which may be caused by random data arrival times or by some power saving strategies that turn agents on and off randomly. We also use random fusion coefficients $\{\boldsymbol{\pi}_k(i)\}$ to model the random status of the communication links connecting the agents to the fusion center. This source of randomness may be caused by random fading effects over the communication channels or by random data feeding/fetching strategies. Therefore, the batch algorithm described in (1.9) is able to accommodate various forms of asynchronous events as well. The

synchronous implementation of (1.9) is given by

$$\boldsymbol{w}_{\mathrm{c},i} = \boldsymbol{w}_{\mathrm{c},i-1} - \sum_{k=1}^{N} \pi_k \mu_k \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{\mathrm{c},i-1}) \tag{1.11}$$

where the $\{\mu_k\}$ are now deterministic nonnegative step-sizes and the $\{\pi_k\}$ are nonnegative fusion coefficients that satisfy $\sum_{k=1}^{N} \pi_k = 1$. It is worth noting that the centralized batch algorithm (1.9) admits a decentralized, though not fully-distributed, implementation of the following form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{\mathrm{c},i-1} - \boldsymbol{\mu}_k(i) \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{\mathrm{c},i-1}) \quad \text{(adaptation)} \tag{1.12a}$$

$$\boldsymbol{w}_{\mathrm{c},i} = \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\, \boldsymbol{\psi}_{\ell,i} \qquad \text{(fusion)} \tag{1.12b}$$

In this description, each agent k uses the local gradient data to calculate the intermediate iterate $\boldsymbol{\psi}_{k,i}$ and feeds its value to a fusion center; the fusion center fuses all intermediate updates $\{\boldsymbol{\psi}_{k,i}\}$ according to (1.12b) to obtain $\boldsymbol{w}_{\mathrm{c},i}$ and then forwards the results to all agents. This process repeats itself at every iteration. Implementation (1.12a)–(1.12b) is not fully distributed because, for example, all agents require knowledge of the same global iterate $\boldsymbol{w}_{\mathrm{c},i}$ to perform the adaptation step (1.12a). It is worth noting that the decentralized batch solution (1.12a)–(1.12b) can also be viewed as distributed solutions over *fully*-connected networks [45].

1.2.2 Questions to be Addressed in This Dissertation

Based on the asynchronous strategy (1.7a)–(1.7b), we shall address the following questions in this dissertation:

1. How does asynchronous behavior affect network stability? Can mean-square stability still be ensured under non-vanishing step-sizes?

2. How is the convergence rate of the algorithm affected? Is it altered relative to synchronous networks?

3. Are agents still able to reach some sort of agreement in steady-state despite the random nature of their interactions and despite data arriving at possibly different rates?

4. How close do the steady-state iterates of the various agents get to each other and to the optimal solution that the network is seeking?

5. Compared with synchronous networks, under what conditions and by how much does the asynchronous behavior generate a *net* negative effect in performance?

6. How close can the performance of an asynchronous network get to that of a stochastic-gradient centralized solution?

We shall answer these questions in detail in Chapters 2–4. Before that, we briefly summarize the main results here.

For the general asynchronous diffusion strategy (1.7a)–(1.7b), which allows for random topologies, random link failures, random data arrival times, and random agents turning on and off, we will establish in Chapter 2 the remarkable conclusion that despite these uncertainties, which could even occur simultaneously, the adaptation process remains mean-square stable for sufficiently small step-sizes. Specifically, we will show that if the first and second moments of the step-size parameters satisfy an upper bound of the form

$$\frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{\lambda_{k,\max}^2 + \alpha} \tag{1.13}$$

for all k, then mean-square stability is ensured in the sense that

$$\limsup_{i\to\infty} \mathbb{E}\,\|w^o - \boldsymbol{w}_{k,i}\|^2 = O(\nu) \tag{1.14}$$

for all k, where

$$\bar{\mu}_k^{(m)} \triangleq \mathbb{E}[\boldsymbol{\mu}_k(i)]^m \tag{1.15}$$

denotes the m-th moment of the random step-size parameter $\boldsymbol{\mu}_k(i)$, $\lambda_{k,\min}$ and $\lambda_{k,\max}$ are positive constant parameters that relate to the Hessian of the individual cost $J_k(w)$, α is a positive constant relating to the variance of the gradient noise, and

$$\nu \triangleq \max_k \frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} \tag{1.16}$$

We will further show that under a strengthened condition, namely,

$$\frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha} \tag{1.17}$$

for all k, the MSD performance metric is given by

$$\text{MSD}^{\text{net}} = \frac{1}{4}\underbrace{\text{Tr}(H^{-1}R)}_{=O(\nu)} + O(\nu^{1+\gamma_o}) \tag{1.18}$$

and the estimates at the individual agents coalesce to satisfy

$$\limsup_{i\to\infty} \mathbb{E}\,\|\boldsymbol{w}_{k,i} - \boldsymbol{w}_{\ell,i}\|^2 = O(\nu^{1+\gamma_o'}) \tag{1.19}$$

for all k and ℓ, where $\gamma_o > 0$ and $\gamma_o' > 0$ are some constants that relate to the covariances of the gradient noise, H is related to the Hessian of the individual costs, and R is related to the covariance of the gradient noise and the moments of the combination coefficients. Expressions (1.14) and (1.19) show that all agents are able to reach a level of $O(\nu^{1+\gamma_o'})$ agreement with each other and to get $O(\nu)$ close to w^o in steady-state (see Fig. 1.3). These results establish that asynchronous networks can operate in a stable manner under fairly general asynchronous events and, importantly, are able to adapt and learn well.

We indicated earlier that studies exist in the literature that examine the performance of distributed strategies in the presence of some forms of asynchronous

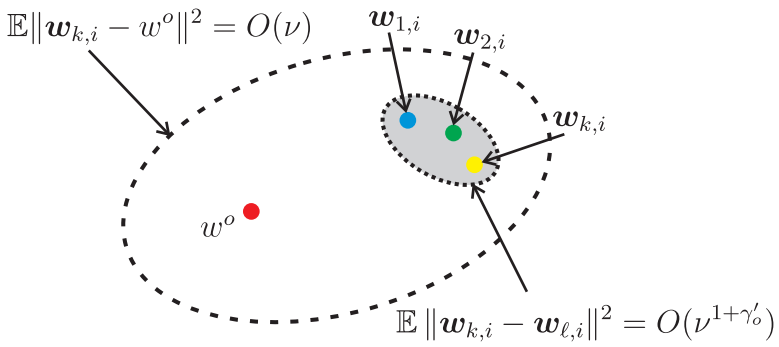


Figure 1.3: Illustration of results (1.14) and (1.19): the solutions by the agents do not only get $O(\nu)$ close to the target w^o but they also cluster next to each other within $O(\nu^{1+\gamma_o'})$ for some $\gamma_o' > 0$.

uncertainties [21, 23, 26, 27] or changing topologies [2, 23, 26, 27, 37–41], albeit for decaying step-sizes. We also explained how the general asynchronous model that we used in (1.7a)–(1.7b) covers broader situations of practical interest, including adaptation and learning under constant step-sizes, and how it allows for the simultaneous occurrence of multiple random events from various sources. Still, these earlier studies did not address the two questions posed earlier on how asynchronous networks compare in performance to synchronous networks and to centralized (batch) solutions. If it can be argued that asynchronous networks are still able to deliver performance similar to synchronous implementations where no uncertainty occurs, or similar to batch solutions where all information is aggregated and available for processing in a centralized fashion, then such a conclusion would be of significant practical relevance. The same conclusion would provide a clear theoretical justification for another critical benefit of cooperation by networked agents, namely, that cooperation does not only enhance performance in comparison to stand-alone single-agent processing, as already demonstrated in prior works in the literature (see, e.g., [8, 34, 35] and the references therein), but it also endows the network with remarkable resilience to various forms of un-

certainties and is still able to deliver performance that is as powerful as batch solutions.

We will therefore further compare the performance of asynchronous networks against synchronous networks, and also compare the performance of distributed solutions against centralized (batch) solutions. The results will show that the performance of adaptive networks are surprisingly immune to the effect of asynchronous uncertainties: the mean and mean-square convergence rates and the asymptotic bias values are not degraded relative to synchronous or centralized implementations. Only the steady-state MSD suffers a degradation in the order of ν. The results will also show that an adaptive network can always matches the performance of a centralized solution. These main conclusions are summarized in Table 1.1, which compares various performance metrics across different implementations. The notation in Table 1.1 will be explained in the sequel. For now, we simply remark that the results in Table 1.1 show that the distributed and centralized implementations have almost the same mean-square performance in either the synchronous or asynchronous modes of operation, i.e., the asynchronous distributed implementation approaches the asynchronous centralized implementation, and the synchronous distributed implementation approaches the synchronous centralized implementation.

1.3 Imperfect Information Exchange and Drifting Model

One useful application of the diffusion strategy (1.2a)–(1.2b) is in the context of the mean-square-error estimation, where the individual costs, $\{J_k(w)\}$, are mean-square-error costs:

$$J_k(w) \triangleq \mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2 \tag{1.20}$$

Table 1.1: Comparison of synchronous vs. asynchronous and distributed vs. centralized solutions

	Synchronous Distributed	**Asynchronous Distributed**	**Synchronous Centralized**	**Asynchronous Centralized**
Algs.	(1.2a) and (1.2b)	(1.7a) and (1.7b)	(1.11)	(1.9)
Vars. [a]	$\boldsymbol{w}_{i,\text{sync}}^{\text{diff}}$	$\boldsymbol{w}_{i,\text{async}}^{\text{diff}}$	$\boldsymbol{w}_{i,\text{sync}}^{\text{cent}}$	$\boldsymbol{w}_{i,\text{async}}^{\text{cent}}$
Paras. [b]	$\{\bar{a}_{\ell k}, \bar{\mu}_k\}$	$\{\boldsymbol{a}_{\ell k}(i), \boldsymbol{\mu}_k(i)\}$	$\{\bar{\pi}_k, \bar{\mu}_k\}$	$\{\boldsymbol{\pi}_k(i), \boldsymbol{\mu}_k(i)\}$
Mn. Rate [c]	$\rho(\bar{\mathcal{B}}) = \rho_o + O(\nu^{1+1/N})$	$\rho(\bar{\mathcal{B}}) = \rho_o + O(\nu^{1+1/N})$	$\rho(\bar{B}) = \rho_o$	$\rho(\bar{B}) = \rho_o$
M.S. Rate [d]	$\rho(\mathcal{F}_{\text{sync}}) = \rho_o^2 + O(\nu^{1+1/N})$	$\rho(\mathcal{F}_{\text{async}}) = \rho_o^2 + O(\nu^{1+1/N^2})$	$\rho(F_{\text{sync}}) = \rho_o^2$	$\rho(F_{\text{async}}) = \rho_o^2 + O(\nu^2)$
MSD [e]	$\frac{1}{4}\text{Tr}(H^{-1}R_{\text{sync}}) + O(\nu^{1+\gamma_o})$	$\frac{1}{4}\text{Tr}(H^{-1}R_{\text{async}}) + O(\nu^{1+\gamma_o})$	$\frac{1}{4}\text{Tr}(H^{-1}R_{\text{sync}}) + O(\nu^{1+\gamma_o})$	$\frac{1}{4}\text{Tr}(H^{-1}R_{\text{async}}) + O(\nu^{1+\gamma_o})$

[a] Variables. The variables for synchronous diffusion strategies are denoted in the table by $\boldsymbol{w}_{i,\text{sync}}^{\text{diff}} \triangleq \text{col}\{\boldsymbol{w}_{1,i,\text{sync}}^{\text{diff}}, \boldsymbol{w}_{2,i,\text{sync}}^{\text{diff}}, \ldots, \boldsymbol{w}_{N,i,\text{sync}}^{\text{diff}}\}$, where $\boldsymbol{w}_{k,i,\text{sync}}^{\text{diff}}$ denotes the iterate of agent k at time i. The variables for asynchronous diffusion strategies are defined in the same manner.

[b] Parameters. The parameters used by the four strategies satisfy:

1. $\bar{\mu}_k = \mathbb{E}\,[\boldsymbol{\mu}_k(i)]$ and $c_{\mu,k,\ell} = \mathbb{E}\,[(\boldsymbol{\mu}_k(i) - \bar{\mu}_k)(\boldsymbol{\mu}_\ell(i) - \bar{\mu}_\ell)]$.
2. $\bar{a}_{\ell k} = \mathbb{E}\,[\boldsymbol{a}_{\ell k}(i)]$, $\bar{\pi}_k = \mathbb{E}\,[\boldsymbol{\pi}_k(i)]$, and $\bar{\pi}_k = \bar{p}_k$, where $\bar{A} = [\bar{a}_{\ell k}]_{\ell,k=1}^N$, $\bar{p} = [\bar{p}_k]_{k=1}^N$, $\bar{A}\bar{p} = \bar{p}$, and $\bar{p}^{\mathsf{T}}\mathbb{1}_N = 1$.
3. $c_{a,\ell k,nm} = \mathbb{E}\,[(\boldsymbol{a}_{\ell k}(i) - \bar{a}_{\ell k})(\boldsymbol{a}_{nm}(i) - \bar{a}_{nm})]$, $c_{\pi,k,\ell} = \mathbb{E}\,[(\boldsymbol{\pi}_k(i) - \bar{\pi}_k)(\boldsymbol{\pi}_\ell(i) - \bar{\pi}_\ell)]$, and $C_\pi = P_p - \bar{p}\bar{p}^{\mathsf{T}}$, where $C_A = [c_{a,\ell k,nm}]_{\ell,k,n,m=1}^N$, $C_\pi = [c_{\pi,k,\ell}]_{\ell,k=1}^N$, $p = \text{vec}(P_p)$, $(\bar{A} \otimes \bar{A} + C_A)p = p$, and $p^{\mathsf{T}}\mathbb{1}_{N^2} = 1$.

[c] Mean convergence rates. The $\{\bar{\mathcal{B}}, \bar{B}\}$ are some constant matrices that are related to the network parameters. Moreover, $\rho_o \triangleq 1 - \lambda_{\min}(H) = 1 - O(\nu)$, where H relates the Hessians of individual costs.

[d] Mean-Square convergence rates. The matrices $\{\bar{\mathcal{F}}_{\text{sync}}, \bar{\mathcal{F}}_{\text{async}}, F_{\text{sync}}, F_{\text{async}}\}$ are some constant matrices that are related to the network parameters.

[e] Mean-Square-Deviations. The matrices $\{R_{\text{sync}}, R_{\text{async}}\}$ are some constant matrices that are related to the network parameters, and γ_o is related to the covariance matrices of gradient noise. Moreover, $R_{\text{async}} - R_{\text{sync}} = O(\nu^2) > 0$.

The scalar measurement data $\boldsymbol{d}_k(i) \in \mathbb{C}$ and the row-vector regressor $\boldsymbol{u}_{k,i} \in \mathbb{C}^{1\times M}$ are assumed to be related via the linear regression model [46]:

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i} w^o + \boldsymbol{\xi}_k(i) \tag{1.21}$$

where $w^o \in \mathbb{C}^{M\times 1}$ is some unknown model parameter and $\boldsymbol{\xi}_k(i) \in \mathbb{C}$ is the additive model noise. The diffusion strategy (1.2a)–(1.2b) reduces in this case to the following form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} + \mu_k \boldsymbol{u}_{k,i}^* [\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}] \tag{1.22a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_{k,i}} a_{\ell k}\, \boldsymbol{\psi}_{\ell,i} \tag{1.22b}$$

When agents in the network perform the social learning step (1.22b), they need to transmit and receive information from their neighbors. This information exchange process can be subject to quantization errors and additive noise over the communication links. Studying the degradation in mean-square performance that results from these particular perturbations can be pursued, for both incremental and diffusion strategies, by extending the mean-square analysis from [5, 6], in the same manner that the tracking analysis of conventional stand-alone adaptive filters was obtained from the counterpart results in the stationary case (as explained in [46, Ch. 21]). Useful results along these lines, which study the effect of link noise during the exchange of the weight estimates, already appear for the traditional diffusion algorithm in the works [47–49] and for consensus-based algorithms in [26,50]. In this dissertation, our objective is to go beyond these earlier studies by taking into account additional effects, and by considering a more general algorithmic structure. The reason for this level of generality is because the analytical results will help reveal which noise sources influence the network performance more seriously, in what manner, and at what stage of the adaptation process. The results will suggest important remedies and mechanisms to adapt

the combination weights in real-time. Some of these insights are hard to get if one focuses solely on noise during the exchange of the weight estimates. The analysis will show that noise during the exchange of the regression data plays a more critical role than noises that arise from the exchange of other pieces of information. The noise that is related to the exchange of the regression data will be shown to alters the learning dynamics and modes of the network, and to bias the weight estimates. Noises related to the exchange of other pieces of information do not alter the dynamics of the network but still contribute to the deterioration of the network performance.

To arrive at these results, we will consider a generalized analysis that applies to a broad class of diffusion adaptation strategies [6]:

$$\boldsymbol{\phi}_{k,i-1} = \sum_{\ell \in \mathcal{N}_k} a_{\ell l k} \boldsymbol{w}_{\ell,i-1} \tag{1.23a}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \mu_k \sum_{\ell \in \mathcal{N}_k} c_{\ell k} \boldsymbol{u}_{\ell,i}^{*} [\boldsymbol{d}_{\ell}(i) - \boldsymbol{u}_{\ell,i} \boldsymbol{\phi}_{k,i-1}] \tag{1.23b}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \boldsymbol{\psi}_{\ell,i} \tag{1.23c}$$

This class includes the original diffusion strategy (1.22a)–(1.22b) as a special case. The analysis based on the diffusion strategy (1.23a)–(1.23c) allows us to account for various sources of information noise over the communication links. Each of the steps in (1.23a)–(1.23c) involves the sharing of information between node k and its neighbors. For example, in the first step (1.23a), all neighbors of node k send their estimates $\boldsymbol{w}_{l,i-1}$ to node k. This transmission is generally subject to additive noise and possibly quantization errors. Likewise, steps (1.23b) and (1.23c) involve the sharing of other pieces of information with node k. These exchange steps can all be subject to perturbations (such as additive noise and quantization errors). We model the data received by node k from its neighbor ℓ

as — see Fig. 1.4:

$$\boldsymbol{w}_{\ell k,i-1} \triangleq \boldsymbol{w}_{\ell,i-1} + \boldsymbol{v}_{\ell k,i-1}^{(w)} \tag{1.24}$$

$$\boldsymbol{\psi}_{\ell k,i} \triangleq \boldsymbol{\psi}_{\ell,i} + \boldsymbol{v}_{\ell k,i}^{(\psi)} \tag{1.25}$$

$$\boldsymbol{d}_{\ell k}(i) \triangleq \boldsymbol{d}_{\ell}(i) + \boldsymbol{v}_{\ell k}^{(d)}(i) \tag{1.26}$$

$$\boldsymbol{u}_{\ell k,i} \triangleq \boldsymbol{u}_{\ell,i} + \boldsymbol{v}_{\ell k,i}^{(u)} \tag{1.27}$$

where $\boldsymbol{v}_{\ell k,i-1}^{(w)}$ and $\boldsymbol{v}_{\ell k,i}^{(\psi)}$ are $M \times 1$ noise signals, $\boldsymbol{v}_{\ell k,i}^{(u)}$ is an $M \times M$ noise signal, and $\boldsymbol{v}_{\ell k}^{(d)}(i)$ is a scalar noise signal (see Fig. 1.4). Using the perturbed data model (1.24)–(1.27), the diffusion algorithm (1.23a)–(1.23c) becomes

$$\boldsymbol{\phi}_{k,i-1} = \sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \boldsymbol{w}_{\ell,i-1} + \boldsymbol{v}_{k,i-1}^{(w)} \tag{1.28}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \mu_k \sum_{\ell \in \mathcal{N}_k} c_{\ell k} \boldsymbol{u}_{\ell k,i}^{*} [\boldsymbol{d}_{\ell k}(i) - \boldsymbol{u}_{\ell k,i} \boldsymbol{\phi}_{k,i-1}] \tag{1.29}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \boldsymbol{\psi}_{\ell,i} + \boldsymbol{v}_{k,i}^{(\psi)} \tag{1.30}$$

where $\boldsymbol{v}_{k,i-1}^{(w)}$ and $\boldsymbol{v}_{k,i}^{(\psi)}$ denote the aggregate $M \times 1$ zero-mean noise signals over the neighborhood $\mathcal{N}_k$:

$$\boldsymbol{v}_{k,i-1}^{(w)} \triangleq \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} a_{1,\ell k} \boldsymbol{v}_{\ell k,i-1}^{(w)} \tag{1.31}$$

$$\boldsymbol{v}_{k,i}^{(\psi)} \triangleq \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} a_{2,\ell k} \boldsymbol{v}_{\ell k,i}^{(\psi)} \tag{1.32}$$

Note that the $\{\boldsymbol{d}_{\ell k}(i), \boldsymbol{u}_{\ell k,i}\}$ in (1.29) are the data contaminated by the link noise.

The diffusion strategy (1.23a)–(1.23c) is adaptive in nature. One of the main benefits of adaptation (by using constant step-sizes) is that it endows networks with tracking abilities when the underlying weight vector w^o varies with time. We will adopt a random-walk model for w^o, which is commonly used in the literature, to describe the non-stationarity of the weight vector [46]:

$$\boldsymbol{w}_i^o = \boldsymbol{w}_{i-1}^o + \boldsymbol{\eta}_i \tag{1.33}$$

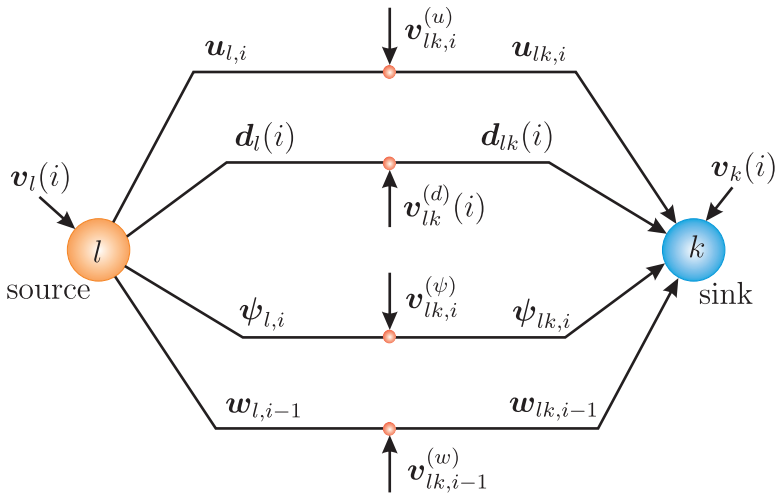


Figure 1.4: Several additive noise sources perturb the exchange of information from node ℓ to node k.

where the sequence $\{\boldsymbol{w}_i^o\}$ has a constant mean w^o for all i, and the $\{\boldsymbol{\eta}_i\}$ is an independent zero-mean random sequence.

One of the main conclusions in Chapter 3 regarding network performance under exchange noise (1.24)–(1.27) and the drifting model (1.33) is that a sufficient condition for the mean-square stability of the diffusion network requires the step-size parameters to be sufficiently small, namely,

$$\mu_k < \frac{2}{\max_{\ell\in\mathcal{N}_k}\left[\lambda_{\max}\left(R_{u,\ell}+R_{v,\ell k}^{(u)}\right)\right]} \tag{1.34}$$

where $R_{u,\ell}$ is the covariance of the regressor $\boldsymbol{u}_{\ell,i}$ and $R_{v,\ell k}^{(u)}$ is the covariance of the link noise $\boldsymbol{v}_{\ell k,i}^{(u)}$. The steady-state MSD will be shown to be given by

$$\begin{aligned}\text{MSD} \approx \frac{1}{N}&\left[\text{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2+\mathcal{R}_\eta+\mathcal{R}_v+\mathcal{Y}+\mathcal{Y}^*)\right]^*\\ &\times(I_{N^2M^2}-\mathcal{F})^{-1}\text{vec}(I_{NM})\end{aligned} \tag{1.35}$$

where the term $\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2$ is contributed by the model noise $\{\boldsymbol{v}_k(i)\}$, the secend term $\mathcal{R}_\eta$ is contributed by the driving noise $\boldsymbol{\eta}_i$ of the drifting model

(1.33), and the remaining terms $\{\mathcal{R}_v, \mathcal{Y}\}$ are contributed by the link noises $\{\boldsymbol{v}_{\ell k,i-1}^{(w)}, \boldsymbol{v}_{\ell k}^{(d)}(i), \boldsymbol{v}_{\ell k,i}^{(u)}, \boldsymbol{v}_{\ell k,i}^{(\psi)}\}$.

1.4 Clustered Networks with Multiple Objectives

A third class of uncertain environment that interferes with the operation of an adaptive network is the possibility of different agents being subjected to different models, with the models being unknown to the agents or their neighbors. There have been several useful works in this domain in the literature under various assumptions, including in the earlier version of this work in [51]. This early investigation dealt only with the case of two separate clusters in the network with each cluster interested in one parameter vector. One useful application of this formulation in the context of biological networks was considered in [52], where each agent was assumed to collect data arising from one of two models (e.g., the location of two separate food sources). The agents did not know which model generated their observations and, yet, they needed to reach agreement about which model to follow (i.e., which food source to move towards). Another important extension dealing with multiple (more than two) models appears in [53, 54] where multi-task problems are introduced. In this formulation, different clusters of the agents are again interested in estimating different parameter vectors (called "tasks") and the tasks of adjacent clusters are further assumed to be related to each other so that cooperation among clusters can still be beneficial. This formulation is useful in many scenarios, as already illustrated in [53], including in multiple target tracking [55, 56] and classification problems involving multiple models [57–62]. Other useful variations of multi-task problems appear in [63], which assumes fully-connected networks, and in [64] where the agents have two types of parameters to estimate (a local parameter and a global pa-

rameter). These various works focus on mean-square-error (MSE) design, where the parameters of interest are estimated by seeking the minimizer of an MSE cost. Moreover, with the exception of [51, 54], it is generally assumed in these works that the agents know beforehand which clusters they belong to or which parameters they are interested in estimating.

In this dissertation, we extend the approach of [51] and study multi-tasking adaptive networks under three conditions that are fundamentally different from previous studies. First, we go beyond mean-square-error estimation and allow for more general convex risk functions at the agents. This level of generality allows the framework to handle broader situations both in adaptation and learning, such as logistic regression for pattern classification purposes. Second, we do not assume any relation among the different objectives pursued by the clusters. In other words, we study the important problem where different components of the network are truly interested in different objectives and would like to avoid interference among clusters. And third, the agents do not know beforehand which clusters they belong to and which other agents are interested in the same objective.

For example, in an application involving a sensor network tracking multiple moving objects from various directions, it is reasonable to assume that the trajectories of these objects are independent of each other. In this case, only information shared within clusters is beneficial for learning; the information from agents in other clusters would amount to interference. This means that agents would need to cooperate with neighbors that belong to the same cluster and would need to cut their links to neighbors with different objectives. This task would be simple to achieve if agents were aware of their cluster information. However, we will not be making that assumption. The cluster information will need to be

learned as well. This point highlights one major feature of our formulation: we do not assume that agents have full knowledge about their clusters. This assumption is quite common in the context of unsupervised machine learning [57, 61], where the collected measurement data are not labeled and there are multiple candidate models. If two neighboring agents are interested in the same model and they are aware of this fact, then they should exchange data and cooperate. However, the agents may not know this fact, so they cannot be certain about whether or not they should cooperate. Accordingly, in this work, we will devise an adaptive clustering and learning strategy that allows agents to learn which neighbors they should cooperate with. In doing so, the resulting algorithm enables the agents in a network to be correctly clustered and to attain improved learning performance through enhanced intra-cluster cooperation.

For this problem, we shall assume that each individual cost function has a unique minimizer and they are categorized into G mutually exclusive groups, denoted by $\mathcal{G}_m$, $m = 1, 2, \ldots, G$, according to their minimizers.

Definition 1.1 (Cluster)**.** *Each cluster q, denoted by $\mathcal{C}_q$, consists of the collection of agents whose individual costs share the common minimizer $w_q^\star$, i.e., $w_k^o = w_q^\star$ for all $k \in \mathcal{C}_q$.* □

Since agents from different clusters do not share common minimizers, the network then aims to solve the *clustered* multi-task problem:

$$\underset{\{w_q\}_{q=1}^Q}{\text{minimize}} \quad J(w_1, \ldots, w_Q) \triangleq \sum_{q=1}^{Q} \sum_{k \in \mathcal{C}_q} J_k(w_q) \tag{1.36}$$

If the cluster information $\{\mathcal{C}_q\}$ is available to the agents, then problem (1.36) can be decomposed into Q separate optimization problems over the sub-networks

associated with the clusters:

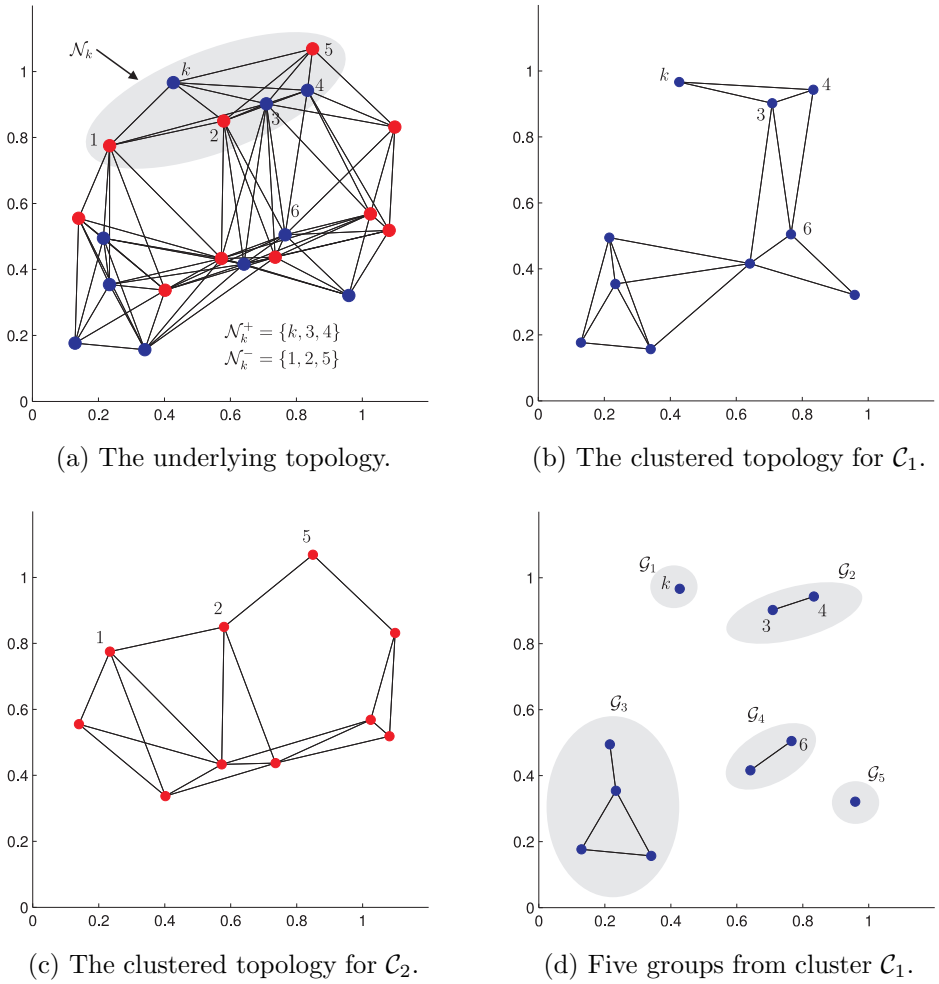
$$\underset{w}{\text{minimize}} \quad J_q^{\text{cluster}}(w) \triangleq \sum_{k \in \mathcal{C}_q} J_k(w) \tag{1.37}$$

for $q = 1, 2, \ldots, Q$. Assuming the cluster topologies are connected, the corresponding minimizers $\{w_q^\star\}$ can be sought by employing diffusion strategies over each cluster. In this case, collaborative learning will only occur *within* each cluster without any interaction across clusters. This means that for every agent k that belongs to a particular cluster $\mathcal{C}_q$, i.e., $k \in \mathcal{C}_q$, its neighbors, which belong to the set denoted by $\mathcal{N}_k$, will need to be segmented into two sets: one set is denoted by $\mathcal{N}_k^+$ and consists of neighbors that belong to the same cluster $\mathcal{C}_q$, and the other set is denoted by $\mathcal{N}_k^-$ and consists of neighbors that belong to other clusters. It is clear that

$$\mathcal{N}_k^+ \triangleq \mathcal{N}_k \cap \mathcal{C}_q, \qquad \mathcal{N}_k^- \triangleq \mathcal{N}_k \backslash \mathcal{N}_k^+ \tag{1.38}$$

We illustrate a two-cluster network with a total of $N = 20$ agents in Fig. 1.5a. The agents in the clusters are denoted by blue and red circles, and are interconnected by the underlying topology, so that agents may have in-cluster neighbors as well as neighbors from other clusters. For example, agent k from blue cluster $\mathcal{C}_1$ has the in-cluster sub-neighborhood $\mathcal{N}_k^+ = \{k, 3, 4\}$, which is a subset of its neighborhood $\mathcal{N}_k = \{k, 1, 2, 3, 4, 5\}$. If the cluster information is available to all agents, then the network can be split into two sub-networks, one for each cluster, as illustrated in Figs. 1.5b and 1.5c.

However, in this dissertation we consider the more challenging scenario in which the cluster information $\{\mathcal{C}_q\}$ is only *partially* available to the agents beforehand, or even completely unavailable. When the cluster information is completely absent, each agent k must first identify neighbors belonging to $\mathcal{N}_k^+$. When the cluster information is partially known, meaning that some agents from the



(a) The underlying topology.

(b) The clustered topology for $\mathcal{C}_1$.

(c) The clustered topology for $\mathcal{C}_2$.

(d) Five groups from cluster $\mathcal{C}_1$.

Figure 1.5: A network with $N = 20$ nodes and $Q = 2$ clusters. Cluster $\mathcal{C}_1$ consists of 10 agents in blue. Cluster $\mathcal{C}_2$ consists of another 10 agents in red. Agent k belongs to Cluster $\mathcal{C}_1$, and its neighborhood is denoted by $\mathcal{N}_k = \{k, 1, 2, 3, 4, 5\}$ with $\mathcal{N}_k^+ = \{k, 3, 4\}$. With perfect cluster information, the underlying topology splits into two sub-networks, one for each cluster. With partial cluster information, cluster $\mathcal{C}_1$ breaks down into five groups: two singleton groups $\mathcal{G}_1$ and $\mathcal{G}_5$, and three non-trivial groups $\mathcal{G}_2$, $\mathcal{G}_3$, and $\mathcal{G}_4$. Through adaptive learning and clustering, the five groups in (b) will end up merging into one largest group corresponding to the entire cluster in (c).

same cluster already know each other, then these agents can cooperate to identify the other members in their cluster. In order to study these two scenarios in a uniform manner, we introduce the concept of a group.

Definition 1.2 (Group)**.** *A group m, denoted by $\mathcal{G}_m$, consists of a collection of connected agents from the same cluster and knowing that they belong to this same cluster.* □

Figure 1.5d illustrates the concept of groups when cluster information is only partially available to the agents in the network from Fig. 1.5a. If an agent has no information about its neighbors, then it falls into a singleton group, such as groups $\mathcal{G}_1$ and $\mathcal{G}_5$ in Fig. 1.5d. If some neighboring agents know the cluster information of each other, then they form a non-trivial group, such as groups $\mathcal{G}_2$, $\mathcal{G}_3$, and $\mathcal{G}_4$. If every agent in a cluster knows the cluster information of all its neighbors, then all cluster members form one group and this group coincides with the cluster itself, as shown in Fig. 1.5b.

Since cooperation among neighbors belonging to different clusters can lead to biased results [53,65,66], agents should only cooperate within clusters. However, when agents have access to partial cluster information, then they only know their group neighbors but not *all* cluster neighbors. Therefore, at this stage, agents can only cooperate within groups, leaving behind some potential opportunity for cooperation with neighbors from the same cluster. We shall devise a procedure to enable agents to identify all of their cluster neighbors, such that small groups from the same cluster can merge automatically into larger groups. At the same time, the procedure needs to be able to turn off links between different clusters in order to avoid interference. By using such a procedure, agents in multi-task networks with *partial* cluster information will be able to cluster themselves in an *adaptive* manner, and then solve problem (1.36) by solving (1.37) collaboratively

within each cluster. The proposed strategy is summarized in the following listing.

Distributed clustering and learning over networks

Initialization: $\boldsymbol{w}_{k,-1} = \boldsymbol{w}'_{k,-1} = 0$ and $\boldsymbol{\mathcal{N}}^+_{k,-1} = \mathcal{N}_k \cap \mathcal{G}_m$ for all $k \in \mathcal{G}_m$ and $m = 1, 2, \ldots, G$.

for $i \geq 0$ **do**

(1) Each agent k updates $\boldsymbol{w}_{k,i}$ according to diffusion strategies over $\mathcal{N}_k \cap \mathcal{G}_m$:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k \widehat{\nabla J_k}(\boldsymbol{w}_{k,i-1}) \tag{1.39a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k \cap \mathcal{G}_m} a_{\ell k} \boldsymbol{\psi}_{\ell,i} \tag{1.39b}$$

(2) Each agent k updates $\boldsymbol{w}'_{k,i}$ according to diffusion strategies over $\boldsymbol{\mathcal{N}}^+_{k,i-1}$:

$$\boldsymbol{\psi}'_{k,i} = \boldsymbol{w}'_{k,i-1} - \mu_k \widehat{\nabla J_k}(\boldsymbol{w}'_{k,i-1}) \tag{1.40a}$$

$$\boldsymbol{w}'_{k,i} = \sum_{\ell \in \boldsymbol{\mathcal{N}}^+_{k,i-1}} \boldsymbol{a}'_{\ell k}(i-1) \boldsymbol{\psi}'_{\ell,i} \tag{1.40b}$$

(3) Each agent k updates $\boldsymbol{\mathcal{N}}^+_{k,i}$ by using:

$$\boldsymbol{\mathcal{N}}^+_{k,i} \triangleq \{\ell \in \mathcal{N}_k; \|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 < \theta_{k,\ell} \ \text{ or } \ \ell \in \mathcal{G}_m\} \tag{1.41}$$

with $\{\boldsymbol{w}_{\ell,i}; \ell \in \mathcal{N}_k\}$ from step (1).

end for

The algorithm enables agents to identify their grouping and to attain improved learning and estimation performance over networks. The last step (1.41) is actually a hypothesis test for agents ℓ and k to determine whether or not they are in the same group:

$$\|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 \underset{\mathbb{H}_1}{\overset{\mathbb{H}_0}{\lessgtr}} \theta_{k,\ell} \tag{1.42}$$

where $\mathbb{H}_0$ denotes the hypothesis of $w^o_\ell = w^o_k$, $\mathbb{H}_1$ denotes the hypothesis of $w^o_\ell \neq w^o_k$, and $\theta_{k,\ell} > 0$ is a predefined threshold. Both agents ℓ and k will test on

(1.42) in order to reach a symmetric pattern for cooperation. Since $\boldsymbol{w}_{k,i}$ and $\boldsymbol{w}_{\ell,i}$ are accessible through local interactions within neighborhoods, the hypothesis test (1.42) can be carried out in a distributed manner. We shall establish later in Chapter 6 that the probabilities for both types of errors incurred in (1.42), i.e., the false alarm (Type-I) and the missing detection (Type-II), decay at exponential rates:

$$\text{Type-I error:} \qquad \mathbb{P}[\|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 > \theta_{k,\ell} | w_\ell^o = w_k^o] \leq O(e^{-c_1/\mu_{\max}}) \tag{1.43}$$

$$\text{Type-II error:} \qquad \mathbb{P}[\|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 < \theta_{k,\ell} | w_\ell^o \neq w_k^o] \leq O(e^{-c_2/\mu_{\max}}) \tag{1.44}$$

for some constants $c_1 > 0$ and $c_2 > 0$. Therefore, for long enough iterations and small enough step-sizes, agents are able to successfully infer the group information with very high probability.

1.5 Organization

The organization of the dissertation is summarized as follows.

- **Chapter 2**: We examine asynchronous networks that are subject to fairly general sources of uncertainties, such as changing topologies, random link failures, random data arrival times, and agents turning on and off randomly. Under this model, agents in the network may stop updating their solutions or may stop sending or receiving information in a random manner and without coordination with other agents. We establish in this chapter conditions on the first and second-order moments of the relevant parameter distributions to ensure mean-square stable behavior. One notable conclusion is that the mean-square-error performance of asynchronous networks shows a degradation only of the order of $O(\nu)$, where ν is a small step-size parameter, while the convergence rate remains largely unaltered. The

results provide a solid justification for the remarkable resilience of cooperative networks in the face of random failures at multiple levels: agents, links, data arrivals, and topology.

- **Chapter 3**: We carry out a detailed analysis of the mean-square-error performance of asynchronous strategies for solving distributed optimization and adaptation problems over networks. We derive analytical expressions for the mean-square convergence rate and the steady-state mean-square-deviation. The expressions reveal how the various parameters of the asynchronous behavior influence network performance. In the process, we establish the interesting conclusion that even under the influence of asynchronous events, all agents in the adaptive network can still reach an $O(\nu^{1+\gamma_o'})$ near-agreement with some $\gamma_o' > 0$ while approaching the desired solution within $O(\nu)$ accuracy, where ν is proportional to the small step-size parameter for adaptation.

- **Chapter 4**: In this chapter, we compare the performance of synchronous and asynchronous networks. We also compare the performance of decentralized adaptation against centralized stochastic-gradient (batch) solutions. Two interesting conclusions stand out. First, the results establish that the performance of adaptive networks is largely immune to the effect of asynchronous events: the mean and mean-square convergence rates and the asymptotic bias values are not degraded relative to synchronous or centralized implementations. Only the steady-state mean-square-deviation suffers a degradation in the order of ν, which represents the small step-size parameters used for adaptation. Second, the results show that the adaptive distributed network matches the performance of the centralized solution. These conclusions highlight another critical benefit of cooperation by net-

worked agents: cooperation does not only enhance performance in comparison to stand-alone single-agent processing, but it also endows the network with remarkable resilience to various forms of random failure events and is able to deliver performance that is as powerful as batch solutions.

- **Chapter 5**: This chapter investigates the mean-square performance of general adaptive diffusion algorithms in the presence of various sources of imperfect information exchanges, quantization errors, and model non-stationarities. Among other results, the analysis reveals that link noise over the regression data modifies the dynamics of the network evolution in a distinct way, and leads to biased estimates in steady-state. The analysis also reveals how the network mean-square performance is dependent on the combination weights. We use these observations to show how the combination weights can be optimized and adapted. Simulation results illustrate the theoretical findings and match well with theory.

- **Chapter 6**: In this chapter, we consider the situation where agents belong to different groups that pursue different objectives. We propose an adaptive clustering and learning scheme that allows agents to learn which neighbors should be cooperate with and which other neighbors should be ignored. In doing so, the resulting algorithm enables the agents to identify their grouping and to attain improved learning and estimation performance over networks.

- **Chapter 7**: In this chapter, we starts from an aggregate cost function defined over a network of agents with different minimizers. We extend the aggregate cost into a higher-dimensional cost and regularize it by adding a weighted quadratic term whose nullspace coincides with the agreement subspace for all agents. Under a local balance condition on the combination

coefficients over the edges, we are able to relate consensus and diffusion strategies to diagonally-weighted gradient-descent iterations. In this case, stability and performance analysis for single-agent implementations become applicable to the distributed solutions. The results in this chapter can be further used to interpret the two-phase transient behavior of consensus and diffusion strategies. When the local balance condition is not satisfied (e.g., when more general left-stochastic combination policies are employed), the study of the behavior of distributed solutions becomes more demanding than stand-alone gradient-descent iterations [66, 67].

1.6 Notation

We use lowercase letters to denote vectors, uppercase letters for matrices, plain letters for deterministic variables, and boldface letters for random variables. We also use $(\cdot)^{\mathsf{T}}$ to denote transposition, $(\cdot)^*$ to denote conjugate transposition, $(\cdot)^{-1}$ for matrix inversion, $\mathrm{Tr}(\cdot)$ for the trace of a matrix, $\lambda(\cdot)$ for the eigenvalues of a matrix, and $\|\cdot\|$ for the 2-norm of a matrix or the Euclidean norm of a vector. Besides, we use $\otimes$ to denote the Kronecker product.

CHAPTER 2

Stability Analysis of Asynchronous Networks

In this chapter, we provide a rather detailed analysis for the stability and performance of the asynchronous strategy (1.7a)–(1.7b) for solving the distributed optimization and adaptation problem (1.1) presented in Chapter 1. We shall examine asynchronous networks that are subject to fairly general sources of uncertainties, such as changing topologies, random link failures, random data arrival times, and agents turning on and off randomly. Under this model, agents in the network may stop updating their solutions or may stop sending or receiving information in a random manner and without coordination with other agents. We shall establish conditions on the first and second-order moments of the relevant parameter distributions to ensure mean-square stable behavior. We shall also derive analytical expressions for the mean-square convergence rate and the steady-state mean-square-deviation. The expressions will reveal how the various parameters of the asynchronous behavior influence network performance. In this process, we shall establish the interesting conclusion that even under the influence of asynchronous events, all agents in the adaptive network can still reach an $O(\nu^{1+\gamma_o'})$ near-agreement with some $\gamma_o' > 0$ while approaching the desired solution within $O(\nu)$ accuracy, where ν is proportional to the small step-size parameter for adaptation. The results in this chapter are based on material from [68].

2.1 Preliminaries

In the problem (1.1) presented in Chapter 1, we allow the argument w to be complex-valued so that the results are applicable to a wide range of problems, especially in the fields of communications and signal processing where complex parameters are fairly common (e.g., in modeling wireless channels, power grid models, beamforming weights, etc.). To facilitate the analysis, and before describing the distributed strategies, we first introduce two alternative ways for representing real-valued functions of complex arguments.

2.1.1 Equivalent Representations

The first representation is based on the *1-to-1* mapping $\bar{\mathbb{T}} : \mathbb{C}^M \longmapsto \mathbb{R}^{2M}$:

$$\bar{w} \triangleq \bar{\mathbb{T}}(w) = \begin{bmatrix} \mathfrak{Re}(w) \\ \mathfrak{Im}(w) \end{bmatrix} \tag{2.1}$$

which replaces the $M \times 1$ complex vector w by the $2M \times 1$ extended vector $\bar{w}$ composed of the real and imaginary components of w. In this way, we can interpret each $J_k(w)$ as a function of the real-valued variable $\bar{w}$ and write $J_k(\bar{w}) \triangleq J_k(w)$ as well as

$$J^{\text{glob}}(\bar{w}) \triangleq J^{\text{glob}}(w) = \sum_{k=1}^{N} J_k(w) = \sum_{k=1}^{N} J_k(\bar{w}) \tag{2.2}$$

The second representation for functions of complex arguments is based on another *1-to-1* mapping $\underline{\mathbb{T}} : \mathbb{C}^M \longmapsto \mathbb{C}_M^{2M}$ (where $\mathbb{C}_M^{2M}$ is a sub-manifold of complex dimension M and is isomorphic to $\mathbb{R}^{2M}$ [69]) defined as

$$\underline{w} \triangleq \underline{\mathbb{T}}(w) = \begin{bmatrix} w \\ (w^*)^{\mathsf{T}} \end{bmatrix} \tag{2.3}$$

in terms of the entries of w and their complex conjugates. In this case, we can interpret each $J_k(w)$ as a function defined over the extended variable $\underline{w} \in \mathbb{C}_M^{2M}$ and write $J_k(\underline{w}) \triangleq J_k(w)$ as well as

$$J^{\text{glob}}(\underline{w}) \triangleq J^{\text{glob}}(w) = \sum_{k=1}^{N} J_k(w) = \sum_{k=1}^{N} J_k(\underline{w}) \tag{2.4}$$

Most of our analysis will be based on the second representation (2.3)–(2.4); the first representation (2.1)–(2.2) will be used when we need to exploit some analytic properties of real functions. Note from (2.1) and (2.3) that the variables $\{\bar{w}, \underline{w}\}$ are related linearly as follows:

$$\underbrace{\begin{bmatrix} w \\ (w^*)^{\mathsf{T}} \end{bmatrix}}_{=\underline{w}} = \underbrace{\begin{bmatrix} I_M & jI_M \\ I_M & -jI_M \end{bmatrix}}_{\triangleq D} \underbrace{\begin{bmatrix} \mathfrak{Re}(w) \\ \mathfrak{Im}(w) \end{bmatrix}}_{=\bar{w}} \iff \underline{w} = D \cdot \bar{w} \tag{2.5}$$

where the matrix D satisfies $DD^* = D^*D = 2 \cdot I_{2M}$ and I_{2M} denotes the $2M \times 2M$ identity matrix. It follows that

$$\bar{w} = D^{-1} \cdot \underline{w} = \frac{1}{2} D^* \cdot \underline{w} \tag{2.6}$$

Using the real representation $\{J_k(\bar{w})\}$, we introduce the following assumption on the analytic properties of $\{J_k(w)\}$.

Assumption 2.1 (Properties of cost functions)**.** *The individual cost functions* $\{J_k(\bar{w}) : \mathbb{R}^{2M} \mapsto \mathbb{R}\,; k = 1, 2, \ldots, N\}$ *are assumed to be at least twice-differentiable and strongly convex over* $\mathbb{R}^{2M}$*. They are also assumed to share a* common *and* unique *minimizer at* $\bar{w}^o \triangleq \bar{\mathbb{T}}(w^o)$*, where* $w^o \in \mathbb{C}^M$*.* □

The situation involving a common minimizer for the cost functions $\{J_k(\bar{w})\}$ is frequent in practice, especially when agents need to cooperate with each other in order to attain a *common* objective. For example, in biological networks, it is

usual for agents in a school of fish to interact while searching for a common food source or avoiding a common predator [19]. Likewise, in wireless sensor networks, it is common for sensors to survey the same physical environment, to interact with each other to estimate a common modeling parameter, or to track the same target [14]. Furthermore, in machine learning applications [70], it is common for all agents to minimize the same cost function (for example, the expected risk) which automatically satisfies the condition of a common minimizer. It follows from Assumption 2.1 that the real global cost function, $J^{\text{glob}}(\bar{w})$, has a unique minimizer at $\bar{w}^o$, or equivalently, that the original global cost function, $J^{\text{glob}}(w)$, has a unique minimizer at w^o. Accordingly, the unique minimizer for $J^{\text{glob}}(\underline{w})$ and for each $J_k(\underline{w})$ is given by $\underline{w}^o = \underline{\mathbb{T}}(w^o)$.

The strong convexity assumption on each cost $J_k(\bar{w})$ ensures that their Hessian matrices are sufficiently bounded away from zero, which avoids situations involving ill-conditioning in recursive implementations based on streaming data. Strong convexity is not a serious limitation because it is common practice in adaptation and learning to incorporate regularization into the cost functions, and it is well-known that regularization helps enforce strong convexity [46, 71]. We may add though that many of the results in this work would still hold if we only require the aggregate cost function $J^{\text{glob}}(\bar{w})$ to be strongly convex by following arguments similar to those used in [65]; in that case, it would be sufficient to require only one of the individual costs $J_k(\bar{w})$ to be strongly convex while the remaining costs can be simply convex. Nevertheless, some of the derivations will become more technical under these more relaxed conditions. For this reason, and since the arguments in Parts I–III are already demanding and lengthy, we opt to convey the main ideas and results by working under Assumption 2.1.

2.1.2 Hessian Matrices

We explain in Appendix 2.A how to compute the complex gradient vector and the complex Hessian matrix of the cost $J_k(w)$, and its equivalent representations, with respect to their arguments. The strong convexity condition from Assumption 2.1 translates into the existence of a lower bound on the Hessian matrices as shown below in (2.7). In addition, we shall assume that the Hessian matrices are also bounded from above. This requirement relaxes conditions from prior studies in the literature where it has been customary to bound the gradient vector as *opposed* to the Hessian matrix [25, 27]; bounding the gradient vector limits the class of cost functions to those with linear growth — see [12] for an explanation.

Assumption 2.2 (Bounded Hessian and Lipschitz condition)**.** *The eigenvalues of the complex Hessian* $\{\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})\}$ *(defined by* (2.109) *in Appendix 2.A) are bounded from below and from above by*

$$\lambda_{k,\min} \leq \lambda(\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})) \leq \lambda_{k,\max} \tag{2.7}$$

where $0 < \lambda_{k,\min} \leq \lambda_{k,\max}$*. Moreover, the complex Hessian functions* $\{\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}\}$ *are assumed to be* locally *Lipschitz continuous [72] at* $\underline{w}^o$*, i.e.,*

$$\|\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o) - \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})\| \leq \tau_k \cdot \|\underline{w}^o - \underline{w}\| \tag{2.8}$$

where $\tau_k \geq 0$*,* $\underline{w}^o = \mathbb{T}(w^o)$ *and* $\underline{w} = \mathbb{T}(w)$ *for any* $w \in \mathbb{B}(w^o, \delta_k)$ *with* $\mathbb{B}(w^o, \delta_k)$ *denoting the 2-norm ball* $\mathbb{B}(w^o, \delta_k) \triangleq \{w \in \mathbb{C}^M; \|w^o - w\| \leq \delta_k\}$*, which is centered at* w^o *with radius* δ_k*.* □

Lemma 2.1 (Global Lipschitz continuity)**.** *When conditions* (2.7) *and* (2.8) *hold, the Hessian matrix functions* $\{\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})\}$ *are globally Lipschitz continuous at* w^o*, i.e.,*

$$\|\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o) - \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})\| \leq \tau'_k \cdot \|\underline{w}^o - \underline{w}\| \tag{2.9}$$

for any w, and

$$\tau_k' \triangleq \max\left\{\tau_k, \frac{\lambda_{k,\max} - \lambda_{k,\min}}{\sqrt{2}\delta_k}\right\} \tag{2.10}$$

Proof. We first note that

$$\|\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o) - \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})\| \leq \lambda_{k,\max} - \lambda_{k,\min} \tag{2.11}$$

for any $\underline{w} = \mathbb{T}(w)$, because for any $2M \times 1$ vector x,

$$\begin{aligned} x^*[\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o) - \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})]x &= x^*[\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o)]x - x^*[\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})]x \\ &\leq (\lambda_{k,\max} - \lambda_{k,\min})\|x\|^2 \end{aligned} \tag{2.12}$$

Now, if $w \in \mathbb{B}(w^o, \delta_k)$, by condition (2.8), we have

$$\|\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o) - \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})\| \leq \tau_k' \cdot \|\underline{w}^o - \underline{w}\| \tag{2.13}$$

On the other hand, if $w \notin \mathbb{B}(w^o, \delta_k)$, i.e., $\|w^o - w\| > \delta_k$ or $\|\underline{w}^o - \underline{w}\| > \sqrt{2}\delta_k$, then we have

$$\begin{aligned} \|\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o) - \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})\| &\leq \frac{\lambda_{k,\max} - \lambda_{k,\min}}{\sqrt{2}\delta_k} \cdot \sqrt{2}\delta_k \\ &\leq \tau_k' \cdot \|\underline{w}^o - \underline{w}\| \end{aligned} \tag{2.14}$$

by condition (2.11). □

2.2 Asynchronous Diffusion Networks

We first describe the traditional synchronous diffusion network studied in [8,12], then we introduce the asynchronous network and derive some useful properties.

2.2.1 Synchronous Diffusion Networks

References [8, 12] deal with the optimization of aggregate real functions of the form $J^{\text{glob}}(\bar{w})$. Starting from equations (12)–(14) from [12] and using (2.5) we

can derive the following diffusion strategy for solving the distributed optimization problem (1.1) with *constant* step-sizes:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}) \quad \text{(adaptation)} \tag{2.15a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k}\, \boldsymbol{\psi}_{\ell,i} \quad \text{(combination)} \tag{2.15b}$$

where (2.15a) is a stochastic gradient approximation step for self-learning and (2.15b) is a convex combination step for social-learning. The iterate $\boldsymbol{w}_{k,i}$ is the estimate for w^o that is computed by agent k at iteration i. The iterate $\boldsymbol{\psi}_{k,i}$ is an intermediate solution that results from the adaptation step and will be shared with the neighbors in the combination step. The factor μ_k is a positive step-size parameter and the combination coefficients $\{a_{\ell k}\}$ are nonnegative parameters and are required to satisfy the following constraints:

$$\sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1, \;\; \text{and} \;\; \begin{cases} a_{\ell k} > 0, & \text{if } \; \ell \in \mathcal{N}_k \\ a_{\ell k} = 0, & \text{otherwise} \end{cases} \tag{2.16}$$

where $\mathcal{N}_k$ denotes the set of neighbors of agent k including k itself. If we collect these coefficients into an $N \times N$ matrix such that $[A]_{\ell k} = a_{\ell k}$, then condition (2.16) implies that A is a left-stochastic matrix, written as $A^{\mathsf{T}} \mathbb{1}_N = \mathbb{1}_N$ where $\mathbb{1}_N$ is the $N \times 1$ vector with all entries equal to one.

In (2.15a), the stochastic approximation for the true gradient vector is used because, in general, agents do not have sufficient information to acquire the true gradients. The difference between the true and approximate gradients is called gradient noise, which is random in nature and seeps into the algorithm. That is why the variables $\{\boldsymbol{w}_{k,i}\}$ in (2.15a)–(2.15b) are random and are represented in boldface. We model the gradient noise, denoted by $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})$, as an additive random perturbation to the true gradient vector, i.e.,

$$\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}) = \nabla_{w^*} J_k(\boldsymbol{w}_{k,i-1}) + \boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1}) \tag{2.17}$$

Let $\mathbb{F}_{i-1}$ denote the filtration to represent all information available up to iteration $i-1$. The conditional covariance of the individual gradient noise $\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})$ is given by

$$\begin{aligned} R_{k,i}(\boldsymbol{w}_{k,i-1}) &\triangleq \mathbb{E}[\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\underline{\boldsymbol{v}}_{k,i}^{*}(\boldsymbol{w}_{k,i-1})|\mathbb{F}_{i-1}] \\ &= \begin{bmatrix} R_{v,k,i}(\boldsymbol{w}_{k,i-1}) & R'_{v,k,i}(\boldsymbol{w}_{k,i-1}) \\ R'^{*}_{v,k,i}(\boldsymbol{w}_{k,i-1}) & R^{\mathsf{T}}_{v,k,i}(\boldsymbol{w}_{k,i-1}) \end{bmatrix} \end{aligned} \tag{2.18}$$

by using (2.3), where

$$R_{v,k,i}(\boldsymbol{w}_{k,i-1}) \triangleq \mathbb{E}[\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})\boldsymbol{v}_{k,i}^{*}(\boldsymbol{w}_{k,i-1})|\mathbb{F}_{i-1}] \tag{2.19}$$

$$R'_{v,k,i}(\boldsymbol{w}_{k,i-1}) \triangleq \mathbb{E}[\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})\boldsymbol{v}_{k,i}^{\mathsf{T}}(\boldsymbol{w}_{k,i-1})|\mathbb{F}_{i-1}] \tag{2.20}$$

so that $R_{v,k,i}(\boldsymbol{w}_{k,i-1})$ is Hermitian positive semi-definite and $R'_{v,k,i}(\boldsymbol{w}_{k,i-1})$ is symmetric. Let further

$$\underline{\boldsymbol{v}}_{i}(\boldsymbol{w}_{i-1}) \triangleq \mathrm{col}\{\underline{\boldsymbol{v}}_{1,i}(\boldsymbol{w}_{1,i-1}), \ldots, \underline{\boldsymbol{v}}_{N,i}(\boldsymbol{w}_{N,i-1})\} \tag{2.21}$$

The conditional covariance of $\underline{\boldsymbol{v}}_{i}(\boldsymbol{w}_{i-1})$ is denoted by

$$\mathcal{R}_{i}(\boldsymbol{w}_{i-1}) \triangleq \mathbb{E}[\underline{\boldsymbol{v}}_{i}(\boldsymbol{w}_{i-1})\underline{\boldsymbol{v}}_{i}^{*}(\boldsymbol{w}_{i-1})|\mathbb{F}_{i-1}] \tag{2.22}$$

Assumption 2.3 (Gradient noise model)**.** *The gradient noise $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})$, conditioned on $\mathbb{F}_{i-1}$, is assumed to be independent of any other random sources including topology, links, combination coefficients, and step-sizes. The conditional mean and variance of $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})$ satisfy:*

$$\mathbb{E}\left[\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})|\mathbb{F}_{i-1}\right] = 0 \tag{2.23}$$

$$\mathbb{E}\left[\|\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})\|^{2}|\mathbb{F}_{i-1}\right] \leq \alpha\left\|w^{o} - \boldsymbol{w}_{k,i-1}\right\|^{2} + \sigma_{v}^{2} \tag{2.24}$$

for some $\alpha \geq 0$ and $\sigma_{v}^{2} \geq 0$. □

Let $\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1}) \triangleq \mathbb{T}(\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1}))$. From Assumption 2.3, the extended gradient noise $\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})$, conditioned on $\mathbb{F}_{i-1}$, is independent of other random sources including topology, links, combination coefficients, and step-sizes. The conditional mean and variance of $\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})$ satisfy

$$\mathbb{E}[\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})|\mathbb{F}_{i-1}] = 0 \tag{2.25}$$

$$\mathbb{E}[\|\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\mathbb{F}_{i-1}] \leq \alpha\|\underline{w}^o - \underline{\boldsymbol{w}}_{k,i-1}\|^2 + 2\sigma_v^2 \tag{2.26}$$

Conditions similar to (2.25) and (2.26) appeared in the works [12, 72, 73] on distributed algorithms. However, they are more relaxed than those employed in [72,73], as already explained in [12]. Conditions (2.25) and (2.26) are satisfied in several useful scenarios of practical relevance such as those involving quadratic costs or logistic costs.

2.2.2 Asynchronous Diffusion Networks

To model the *asynchronous* behavior of the network, we modify the diffusion strategy (2.15a)–(2.15b) to the following form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \boldsymbol{\mu}_k(i)\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}) \tag{2.27a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\,\boldsymbol{\psi}_{\ell,i} \tag{2.27b}$$

where the $\{\boldsymbol{\mu}_k(i), \boldsymbol{a}_{\ell\,k}(i)\}$ are now *time-varying* and *random* step-sizes and combination coefficients, and $\boldsymbol{\mathcal{N}}_{k,i}$ denotes the *random* neighborhood of agent k at time i. The step-size parameters $\{\boldsymbol{\mu}_k(i)\}$ are nonnegative random variables, and the combination coefficients $\{\boldsymbol{a}_{\ell k}(i)\}$ are also nonnegative random variables, which are required to satisfy the following constraints (compare to (2.16)):

$$\sum_{\ell\in\boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i) = 1, \;\; \text{and} \;\; \begin{cases} \boldsymbol{a}_{\ell k}(i) > 0, & \text{if } \ell \in \boldsymbol{\mathcal{N}}_{k,i} \\ \boldsymbol{a}_{\ell k}(i) = 0, & \text{otherwise} \end{cases} \tag{2.28}$$

Let

$$\boldsymbol{w}_i \triangleq \mathrm{col}\{\boldsymbol{w}_{1,i}, \boldsymbol{w}_{2,i}, \ldots, \boldsymbol{w}_{N,i}\} \tag{2.29}$$

$$\boldsymbol{\psi}_i \triangleq \mathrm{col}\{\boldsymbol{\psi}_{1,i}, \boldsymbol{\psi}_{2,i}, \ldots, \boldsymbol{\psi}_{N,i}\} \tag{2.30}$$

denote the collections of the iterates from across the network at time i. Let also

$$\boldsymbol{M}_i \triangleq \mathrm{diag}\{\boldsymbol{\mu}_1(i), \boldsymbol{\mu}_2(i), \ldots, \boldsymbol{\mu}_N(i)\} \tag{2.31}$$

be the diagonal random step-size matrix at time i. We further collect the combination coefficients $\{\boldsymbol{a}_{\ell k}(i)\}$ at time i into the matrix $\boldsymbol{A}_i \in \mathbb{R}^{N\times N}$. The asynchronous network model consists of the following conditions on $\{\boldsymbol{M}_i, \boldsymbol{A}_i; i \geq 0\}$:

1. The stochastic process $\{\boldsymbol{M}_i, i \geq 0\}$ consists of a sequence of diagonal random matrices with *bounded* nonnegative entries, $\{\boldsymbol{\mu}_k(i) \in [0, \mu_k]\}$, where the upper bound $\mu_k > 0$ is a constant. The random matrix $\boldsymbol{M}_i$ is assumed to have constant mean $\bar{M}$ of size $N \times N$ and constant Kronecker-covariance matrix C_M of size $N^2 \times N^2$, i.e.,

$$\bar{M} \triangleq \mathbb{E}\,\boldsymbol{M}_i = \mathrm{diag}\{\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_N\} \tag{2.32}$$

$$\bar{\mu}_k \triangleq \mathbb{E}\,\boldsymbol{\mu}_k(i) \tag{2.33}$$

$$\begin{aligned} C_M &\triangleq \mathbb{E}\,[(\boldsymbol{M}_i - \bar{M}) \otimes (\boldsymbol{M}_i - \bar{M})] \\ &= \mathrm{diag}\{C_{\mu,1}, C_{\mu,2}, \ldots, C_{\mu,N}\} \end{aligned} \tag{2.34}$$

$$\begin{aligned} C_{\mu,k} &\triangleq \mathbb{E}[(\boldsymbol{\mu}_k(i) - \bar{\mu}_k)(\boldsymbol{M}_i - \bar{M})] \\ &= \mathrm{diag}\{c_{\mu,k,1}, c_{\mu,k,2}, \ldots, c_{\mu,k,N}\} \end{aligned} \tag{2.35}$$

where $\bar{\mu}_k$ denotes the k-th entry on the diagonal of $\bar{M}$, $C_{\mu,k}$ is a diagonal matrix and denotes the k-th block of size $N \times N$ on the diagonal of C_M, and $c_{\mu,k,\ell}$ denotes the ℓ-th entry on the diagonal of $C_{\mu,k}$. The scalar $c_{\mu,k,\ell}$

represents the covariance between the step-sizes $\boldsymbol{\mu}_k(i)$ and $\boldsymbol{\mu}_\ell(i)$:

$$c_{\mu,k,\ell} \triangleq \mathbb{E}[(\boldsymbol{\mu}_k(i) - \bar{\mu}_k)(\boldsymbol{\mu}_\ell(i) - \bar{\mu}_\ell)] \tag{2.36}$$

When $\ell = k$, the scalar $c_{\mu,k,k}$ becomes the variance of $\boldsymbol{\mu}_k(i)$. Since the $\{\bar{\mu}_k\}$ are all finite positive numbers, the condition number of the matrix $\bar{M}$ is bounded by some finite positive constant $\kappa > 0$, i.e.,

$$\frac{\max_k\{\bar{\mu}_k\}}{\min_k\{\bar{\mu}_k\}} \leq \kappa \tag{2.37}$$

2. The stochastic process $\{\boldsymbol{A}_i, i \geq 0\}$ consists of a sequence of left-stochastic random matrices, whose entries satisfy the conditions in (2.28) at every time i. The mean and Kronecker-covariance matrices of $\boldsymbol{A}_i$ are assumed to be constant and are denoted by the $N \times N$ matrix $\bar{A}$ and the $N^2 \times N^2$ matrix C_A, respectively,

$$\bar{A} \triangleq \mathbb{E}\,\boldsymbol{A}_i = [\bar{a}_{\ell k}]_{\ell,k=1}^N \tag{2.38}$$

$$\bar{a}_{\ell k} \triangleq \mathbb{E}\,\boldsymbol{a}_{\ell k}(i) \tag{2.39}$$

$$C_A \triangleq \mathbb{E}[(\boldsymbol{A}_i - \bar{A}) \otimes (\boldsymbol{A}_i - \bar{A})] = [C_{a,\ell k}]_{\ell,k=1}^N \tag{2.40}$$

$$C_{a,\ell k} \triangleq \mathbb{E}[(\boldsymbol{a}_{\ell k}(i) - \bar{a}_{\ell k})(\boldsymbol{A}_i - \bar{A})] = [c_{a,\ell k,nm}]_{n,m=1}^N \tag{2.41}$$

where $\bar{a}_{\ell k}$ denotes the (ℓ, k)-th element of $\bar{A}$, $C_{a,\ell k}$ denotes the (ℓ, k)-th block with size $N \times N$ of C_A, and $c_{a,\ell k,nm}$ denotes the (n, m)-th element of $C_{a,\ell k}$. The scalar $c_{a,\ell k,nm}$ represents the covariance between the combination coefficients $\boldsymbol{a}_{\ell k}(i)$ and $\boldsymbol{a}_{nm}(i)$:

$$c_{a,\ell k,nm} \triangleq \mathbb{E}[(\boldsymbol{a}_{\ell k}(i) - \bar{a}_{\ell k})(\boldsymbol{a}_{nm}(i) - \bar{a}_{nm})] \tag{2.42}$$

When $\ell = n$ and $k = m$, the scalar $c_{a,\ell k,\ell k}$ becomes the variance of $\boldsymbol{a}_{\ell k}(i)$.

3. The random matrices $\boldsymbol{M}_i$ and $\boldsymbol{A}_i$ are mutually-independent and are independent of any other random variable.

4. We refer to the topology that corresponds to the average combination matrix $\bar{A}$ as the *mean* graph, which is fixed over time. For each agent k, the neighborhood defined by the mean graph is denoted by $\mathcal{N}_k$. The mean combination coefficients $\bar{a}_{\ell k} > 0$ satisfy the following constraints (compare with (2.16) and (2.28)):

$$\sum_{\ell \in \mathcal{N}_k} \bar{a}_{\ell k} = 1, \quad \text{and} \quad \begin{cases} \bar{a}_{\ell k} > 0, & \text{if } \ell \in \mathcal{N}_k \\ \bar{a}_{\ell k} = 0, & \text{otherwise} \end{cases} \tag{2.43}$$

The asynchronous network model described above is general enough to cover many situations of practical interest — note that the model does not impose any specific probabilistic distribution on the step-sizes, network topologies, or combination coefficients. The upper bounds $\{\mu_k\}$ are arbitrary and are independent of the constant step-size parameters used in synchronous diffusion networks (2.15a)–(2.15b). For example, we can choose the sample space of each step-size $\boldsymbol{\mu}_k(i)$ to be the binary choice $\{0, \mu\}$ to model a random "on-off" behavior at each agent k for the purpose of saving power, waiting for data, or even due to random agent failures. Similarly, we can choose the sample space of each combination coefficient $\boldsymbol{a}_{\ell\, k}(i)$, $\ell \in \mathcal{N}_k \backslash \{k\}$, to be $\{0, a_{\ell k}\}$ to model a random "on-off" status for the link from agent ℓ to agent k at time i for the purpose of either saving communication cost or due to random link failures. If links are randomly chosen by agents such that at every time i there is only one other neighboring agent being communicated with, then we effectively mimic the random gossip strategies [2,23,38,39,42]. Note that the convex constraint (2.28) can be satisfied by adjusting the value of $\boldsymbol{a}_{kk}(i)$ according to the realizations of $\{\boldsymbol{a}_{\ell k}(i); \ell \in \boldsymbol{\mathcal{N}}_{k,i} \backslash \{k\}\}$. If the underlying topology is changing over time and the combination weights are also selected in a random manner, then we obtain the probabilistic diffusion strategy studied in [43, 44] or the random link or topology model studied in [26, 37, 40]. Since the

parameter matrices $\boldsymbol{M}_i$ and $\boldsymbol{A}_i$ are assumed to be independent of each other and of any other random variable, the statistical dependency among the random variables $\{\boldsymbol{w}_i, \boldsymbol{\psi}_i, \boldsymbol{A}_i, \boldsymbol{M}_i\}$ is illustrated in Fig. 1.2 in Chapter 1. The filtration $\mathbb{F}_{i-1}$ now also includes information about $\boldsymbol{A}_{i-1}$ and $\boldsymbol{M}_{i-1}$ to represent *all* information available up to iteration $i-1$.

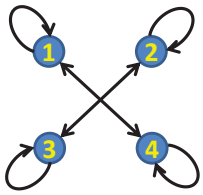
2.2.3 Properties of the Asynchronous Model

The randomness of the combination coefficient matrix $\boldsymbol{A}_i$ arises from three different sources. The first source is the randomness in the topology. The random topology is used to model the rich dynamics of evolving adaptive networks. The second source arises once a certain topology is realized, where the links among agents are further allowed to drop randomly. This phenomenon may be caused by either random interference or fading that blocks the communication links, or by neighbor selection policies used to save resources such as energy and bandwidth. The third sources relates to the agents which are allowed to assign random combination coefficients to their links, as long as the constraint (2.28) is satisfied. An example of a random network with two equally probable realizations and its mean graph is shown in Fig. 2.1. The letter ω is used to index the sample space of the random matrix $\boldsymbol{A}_i$. A useful result relating the random neighborhoods $\{\boldsymbol{\mathcal{N}}_{k,i}\}$ from (2.27b) to the neighborhoods $\{\mathcal{N}_k\}$ from the mean network model is given in the following statement.

Lemma 2.2 (Neighborhoods)**.** *The neighborhood $\mathcal{N}_k$ defined by the mean graph of the asynchronous network model is equal to the* union *of all possible realizations for the random neighborhood $\boldsymbol{\mathcal{N}}_{k,i}$ in* (2.27b)*, i.e.,*

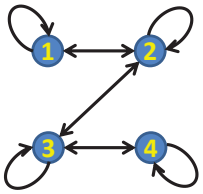
$$\mathcal{N}_k = \bigcup_{\omega \in \Omega} \boldsymbol{\mathcal{N}}_{k,i}(\omega) \tag{2.44}$$

$$\boldsymbol{A}_i(\omega_1) = \begin{bmatrix} 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \end{bmatrix}$$



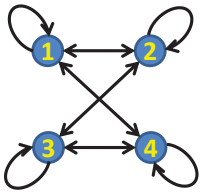
$$\boldsymbol{\mathcal{N}}_{1,i}(\omega_1) = \{1,4\}$$
$$\boldsymbol{\mathcal{N}}_{2,i}(\omega_1) = \{2,3\}$$
$$\boldsymbol{\mathcal{N}}_{3,i}(\omega_1) = \{2,3\}$$
$$\boldsymbol{\mathcal{N}}_{4,i}(\omega_1) = \{1,4\}$$

$$\boldsymbol{A}_i(\omega_2) = \begin{bmatrix} 0.5 & 0.25 & 0 & 0 \\ 0.5 & 0.5 & 0.25 & 0 \\ 0 & 0.25 & 0.5 & 0.5 \\ 0 & 0 & 0.25 & 0.5 \end{bmatrix}$$



$$\boldsymbol{\mathcal{N}}_{1,i}(\omega_2) = \{1,2\}$$
$$\boldsymbol{\mathcal{N}}_{2,i}(\omega_2) = \{1,2,3\}$$
$$\boldsymbol{\mathcal{N}}_{3,i}(\omega_2) = \{2,3,4\}$$
$$\boldsymbol{\mathcal{N}}_{4,i}(\omega_2) = \{3,4\}$$

$$\bar{A} = 0.5[\boldsymbol{A}_i(\omega_1) + \boldsymbol{A}_i(\omega_2)] = \begin{bmatrix} 0.5 & 0.125 & 0 & 0.25 \\ 0.25 & 0.5 & 0.375 & 0 \\ 0 & 0.375 & 0.5 & 0.25 \\ 0.25 & 0 & 0.125 & 0.5 \end{bmatrix}$$



$$\mathcal{N}_1 = \{1,2,4\}$$
$$\mathcal{N}_2 = \{1,2,3\}$$
$$\mathcal{N}_3 = \{2,3,4\}$$
$$\mathcal{N}_4 = \{1,3,4\}$$

Figure 2.1: The first two rows show two equally probable realizations with the respective neighborhoods. The last row shows the resulting mean graph.

for any k, where Ω denotes the sample space of $\boldsymbol{\mathcal{N}}_{k,i}$.

Proof. We first establish $\bigcup_{\omega\in\Omega} \boldsymbol{\mathcal{N}}_{k,i}(\omega) \subseteq \mathcal{N}_k$. By (2.28), we have $\boldsymbol{a}_{\ell k}(i) > 0$ for any $\ell \in \boldsymbol{\mathcal{N}}_{k,i}$. Since $\boldsymbol{a}_{\ell k}(i)$ is a nonnegative random variable, if the event $\boldsymbol{a}_{\ell k}(i) > 0$ occurs, then $\bar{a}_{\ell k} > 0$ by (2.39), which implies $\ell \in \mathcal{N}_k$. Thus, we get $\boldsymbol{\mathcal{N}}_{k,i} \subseteq \mathcal{N}_k$. This relation holds for any realization of $\boldsymbol{\mathcal{N}}_{k,i}$, so we have $\bigcup_{\omega\in\Omega} \boldsymbol{\mathcal{N}}_{k,i}(\omega) \subseteq \mathcal{N}_k$.

Now we establish $\mathcal{N}_k \subseteq \bigcup_{\omega\in\Omega} \boldsymbol{\mathcal{N}}_{k,i}(\omega)$. For any $\ell \in \mathcal{N}_k$, we have $\bar{a}_{\ell k} > 0$ by definition. This is only possible if there exists at least one realization of $\boldsymbol{a}_{\ell k}(i)$ assuming a positive value, which means that $\ell \in \boldsymbol{\mathcal{N}}_{k,i}(\omega)$ for a certain ω. Therefore, $\mathcal{N}_k \subseteq \bigcup_{\omega\in\Omega} \boldsymbol{\mathcal{N}}_{k,i}(\omega)$ holds as expected. □

Another useful property for the asynchronous model relates to the combina-

tion coefficient matrices $\{\bar{A}, \bar{A} \otimes \bar{A} + C_A\}$.

Lemma 2.3 (Left-stochastic matrices)**.** *The $N \times N$ matrix $\bar{A}$ and the $N^2 \times N^2$ matrix $\bar{A} \otimes \bar{A} + C_A$ are left-stochastic matrices, meaning that every element of $\bar{A}$ or $\bar{A} \otimes \bar{A} + C_A$ is nonnegative and*

$$\bar{A}^{\mathsf{T}} \mathbb{1}_N = \mathbb{1}_N, \qquad (\bar{A} \otimes \bar{A} + C_A)^{\mathsf{T}} \mathbb{1}_{N^2} = \mathbb{1}_{N^2} \tag{2.45}$$

Proof. Since $\boldsymbol{A}_i$ has nonnegative entries by the asynchronous network model, it is easy to verify that $\bar{A}$ and $\boldsymbol{A}_i \otimes \boldsymbol{A}_i$ also have nonnegative entries. Moreover, noting that

$$\begin{aligned} \mathbb{E}(\boldsymbol{A}_i \otimes \boldsymbol{A}_i) &= \bar{A} \otimes \bar{A} + \mathbb{E}[(\boldsymbol{A}_i - \bar{A}) \otimes (\boldsymbol{A}_i - \bar{A})] \\ &= \bar{A} \otimes \bar{A} + C_A \end{aligned} \tag{2.46}$$

it follows that $\bar{A} \otimes \bar{A} + C_A$ has nonnegative entries as well. Furthermore, observe that

$$\bar{A}^{\mathsf{T}} \mathbb{1}_N = \mathbb{E}(\boldsymbol{A}_i^{\mathsf{T}}) \mathbb{1}_N = \mathbb{E}(\boldsymbol{A}_i^{\mathsf{T}} \mathbb{1}_N) = \mathbb{1}_N \tag{2.47}$$

and

$$\begin{aligned} (\bar{A} \otimes \bar{A} + C_A)^{\mathsf{T}} \mathbb{1}_{N^2} &= \mathbb{E}(\boldsymbol{A}_i^{\mathsf{T}} \otimes \boldsymbol{A}_i^{\mathsf{T}})(\mathbb{1}_N \otimes \mathbb{1}_N) \\ &= \mathbb{E}[(\boldsymbol{A}_i^{\mathsf{T}} \mathbb{1}_N) \otimes (\boldsymbol{A}_i^{\mathsf{T}} \mathbb{1}_N)] \\ &= \mathbb{1}_{N^2} \end{aligned} \tag{2.48}$$

as desired. □

A useful special case of the asynchronous network model is the spatially-uncorrelated model, where the random step-sizes at the agents are uncorrelated with each other across the network, and the random combination coefficients assigned by each agent to its local neighbors *excluding* itself are also uncorrelated

with each other and with all other combinations weights assigned by other agents across the network. In the next subsection we provide two concrete examples for this model.

Lemma 2.4 (The spatially-uncorrelated model)**.** *Under the asynchronous network model, if at each iteration i, the random step-sizes $\{\boldsymbol{\mu}_k(i); k=1,2,\ldots,N\}$ are uncorrelated with each other across the network, and if the random combination coefficients $\{\boldsymbol{a}_{\ell k}(i); \ell \neq k, k=1,2,\ldots,N\}$ are also assumed to be uncorrelated with each other across the network, then the covariances $\{c_{\mu,k,\ell}\}$ in* (2.36) *and $\{c_{a,\ell k,nm}\}$ in* (2.42) *are now given by*

$$c_{\mu,k,\ell} = \begin{cases} c_{\mu,k,k}, & \text{if } \ell = k \\ 0, & \text{otherwise} \end{cases} \tag{2.49}$$

and

$$c_{a,\ell k,nm} = \begin{cases} c_{a,\ell k,\ell k}, & \text{if } k=m, \ell=n, \ell \in \mathcal{N}_k \backslash \{k\} \\ -c_{a,\ell k,\ell k}, & \text{if } k=m=n, \ell \in \mathcal{N}_k \backslash \{k\} \\ -c_{a,nk,nk}, & \text{if } k=m=\ell, n \in \mathcal{N}_k \backslash \{k\} \\ \displaystyle\sum_{j \in \mathcal{N}_k \backslash \{k\}} c_{a,jk,jk}, & \text{if } k=m=\ell=n \\ 0, & \text{otherwise} \end{cases} \tag{2.50}$$

Correspondingly, the block matrices $\{C_{\mu,k}, C_{a,\ell k}\}$ in (2.35) *and* (2.41) *are given by the following compact expressions:*

$$C_{\mu,k} = c_{\mu,k,k} \cdot E_{kk} \tag{2.51}$$

$$C_{a,\ell k} = c_{a,\ell k,\ell k} \cdot (E_{\ell k} - E_{kk}), \quad \ell \in \mathcal{N}_k \backslash \{k\} \tag{2.52}$$

$$C_{a,kk} = \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} c_{a,\ell k,\ell k} \cdot (E_{kk} - E_{\ell k}) \tag{2.53}$$

where $E_{\ell k}$ denotes the $N \times N$ matrix whose entries are all zero except for the (ℓ, k)-th entry, which is equal to one.

Proof. See Appendix 2.B. □

We remark that the matrices $\{C_M, C_A\}$ are Kronecker-covariance matrices defined by (2.34) and (2.40); they are *not* conventional covariance matrices and, therefore, are *not* necessarily Hermitian matrices.

2.2.4 Two Useful Network Models

In this subsection we describe two scenarios where the asynchronous behavior arises naturally. The first model below is referred to as the Bernoulli model, a special case of which was used before to model random link failures over consensus networks [26, 40]. The Bernoulli model given here is more general in that it also allows us to consider simultaneously on-off strategies for adaptation through equation (2.54).

2.2.4.1 The Bernoulli Model

We assume that at every time i, each agent k adopts a random "on-off" policy to reduce energy consumption. Specifically, agent k enters an active mode with probability $0 < q_k < 1$ and performs the self-learning step (2.27a), and it enters a sleep mode with probability $1 - q_k$ to save energy. This behavior can also be interpreted as the result of random data arrivals: at every time i, a new data becomes available to agent k with probability q_k. This situation can be modeled

by the following Bernoulli random step-size model:

$$\boldsymbol{\mu}_k(i) = \begin{cases} \mu_k, & \text{with probability } \ q_k \\ 0, & \text{with probability } \ 1-q_k \end{cases} \tag{2.54}$$

where μ_k is a fixed step-size. We further assume that the underlying topology is fixed. However, each agent k is allowed to randomly choose a *subset* of its neighbors to perform the social-learning step (2.27b). Specifically, agent k chooses neighbor ℓ with probability $0 < \eta_{\ell k} < 1$ to save on communication costs. This behavior can also be interpreted as resulting from random link failures: at every time i, the communication link from agent ℓ to agent k drops with probability $\eta_{\ell k}$. This situation can be modeled by the following Bernoulli random combination coefficients model:

$$\boldsymbol{a}_{\ell k}(i) = \begin{cases} a_{\ell k}, & \text{with probability } \ \eta_{\ell k} \\ 0, & \text{with probability } \ 1-\eta_{\ell k} \end{cases} \tag{2.55}$$

for any $\ell \in \boldsymbol{\mathcal{N}}_{k,i} \backslash \{k\}$, where $0 < a_{\ell k} < 1$ is a fixed combination coefficient. Based on Lemma 2.2, we require the values of $\boldsymbol{a}_{\ell k}(i)$ in (2.55) to ensure that $0 \leq \boldsymbol{a}_{kk}(i) \leq 1$ where

$$\boldsymbol{a}_{kk}(i) = 1 - \sum_{\ell \in \boldsymbol{\mathcal{N}}_{k,i} \backslash \{k\}} \boldsymbol{a}_{\ell k}(i) \tag{2.56}$$

Using Lemma 2.4, the relevant quantities introduced in the asynchronous network model are given by

$$\bar{\mu}_k = q_k \mu_k \tag{2.57}$$

$$c_{\mu,k,k} = q_k(1-q_k)\mu_k^2 \tag{2.58}$$

$$\bar{a}_{\ell k} = \eta_{\ell k} a_{\ell k} \tag{2.59}$$

$$\bar{a}_{kk} \triangleq 1 - \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} \eta_{\ell k} a_{\ell k} \tag{2.60}$$

$$c_{a,\ell k,\ell k} = \eta_{\ell k}(1-\eta_{\ell k})a_{\ell k}^2, \quad \ell \in \mathcal{N}_k \backslash \{k\} \tag{2.61}$$

2.2.4.2 The Beta Model

The other example involves continuous random variables modeled by Beta distributions, which can be viewed as extensions of binary Bernoulli distributions to the continuous domain when the probability mass is distributed over a bounded region. The family of Beta distributions takes values in the interval $[0, 1]$ and includes the uniform distribution over $[0, 1]$ as a special case [74]. The probability density function (PDF) of a Beta distribution is given by

$$B(x;\xi,\zeta) = \begin{cases} \dfrac{\Gamma(\xi+\zeta)}{\Gamma(\xi)\Gamma(\zeta)} x^{\xi-1}(1-x)^{\zeta-1}, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.62}$$

where $\xi, \zeta > 0$ are the shape parameters and $\Gamma(\cdot)$ denotes the Gamma function. Figure 2.2 plots $B(x;\xi,\zeta)$ for two values of ζ. The mean and variance of the Beta distribution (2.62) are given by

$$\bar{x} = \frac{\xi}{\xi+\zeta}, \quad \sigma_x^2 = \frac{\xi\zeta}{(\xi+\zeta)^2(\xi+\zeta+1)} \tag{2.63}$$

For the asynchronous network model, we assume that the step-size $\boldsymbol{\mu}_k(i)$ takes random values in the range $[0, \mu_k]$, where μ_k denotes the largest possible value for the k-th step-size. We further assume that the scaled parameter $\boldsymbol{\mu}_k(i)/\mu_k$ is governed by a Beta distribution:

$$\boldsymbol{x}_k(i) = \frac{\boldsymbol{\mu}_k(i)}{\mu_k} \sim B(x_k;\xi_k,\zeta_k) \tag{2.64}$$

where $\{\xi_k, \zeta_k > 0\}$ are the corresponding shape parameters. Likewise, we assume that the combination coefficient $\boldsymbol{a}_{\ell k}(i)$ for $\ell \in \boldsymbol{\mathcal{N}}_{k,i} \backslash \{k\}$ takes random values in the range $[0, a_{\ell k}]$ with $0 < a_{\ell k} < 1$. The scaled parameter $\boldsymbol{a}_{\ell k}(i)/a_{\ell k}$ is assumed to

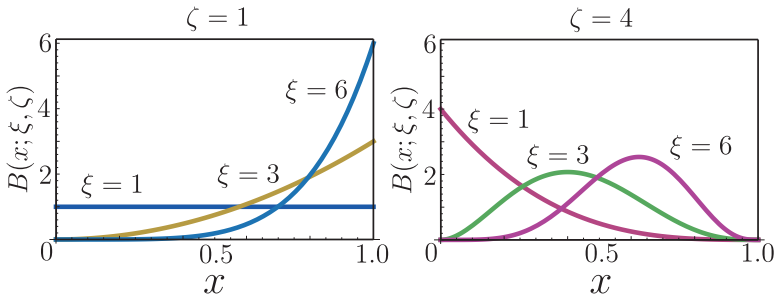


Figure 2.2: The PDFs of the Beta distribution $B(x;\xi,\zeta)$ for different values of ξ and ζ.

be governed by a Beta distribution:

$$\boldsymbol{y}_{\ell k}(i) = \frac{\boldsymbol{a}_{\ell k}(i)}{a_{\ell k}} \sim B(y_{\ell k};\xi_{\ell k},\zeta_{\ell k}) \tag{2.65}$$

where $\{\xi_{\ell k},\zeta_{\ell k} > 0\}$ are the shape parameters. We adopt the spatially uncorrelated model from Lemma 2.4. In order to guarantee that $\boldsymbol{a}_{kk}(i)$ always assumes values within the range $[0,1]$, we again require condition (2.56). Then, we can use (2.63) to calculate the relevant quantities introduced in the asynchronous network model:

$$\bar{\mu}_k = \frac{\xi_k}{\xi_k+\zeta_k}\mu_k \tag{2.66}$$

$$c_{\mu,k,k} = \frac{\xi_k\zeta_k}{(\xi_k+\zeta_k)^2(\xi_k+\zeta_k+1)}\mu_k^2 \tag{2.67}$$

$$\bar{a}_{\ell k} = \frac{\xi_{\ell k}}{\xi_{\ell k}+\zeta_{\ell k}}a_{\ell k} \tag{2.68}$$

$$\bar{a}_{kk} \triangleq 1 - \sum_{\ell\in\mathcal{N}\backslash\{k\}} \frac{\xi_{\ell k}}{\xi_{\ell k}+\zeta_{\ell k}}a_{\ell k} \tag{2.69}$$

$$c_{a,\ell k,\ell k} = \frac{\xi_{\ell k}\zeta_{\ell k}}{(\xi_{\ell k}+\zeta_{\ell k})^2(\xi_{\ell k}+\zeta_{\ell k}+1)}a_{\ell k}^2,\ \ell\in\mathcal{N}_k\backslash\{k\} \tag{2.70}$$

2.3 Mean-Square Stability

For each agent k, we introduce the error vectors:

$$\widetilde{\boldsymbol{\psi}}_{k,i} \triangleq w^o - \boldsymbol{\psi}_{k,i}, \qquad \widetilde{\boldsymbol{w}}_{k,i} \triangleq w^o - \boldsymbol{w}_{k,i} \tag{2.71}$$

where w^o is the desired optimal solution. Subtracting w^o from both sides of (2.15a)–(2.15b) and using (2.17) gives

$$\widetilde{\boldsymbol{\psi}}_{k,i} = \widetilde{\boldsymbol{w}}_{k,i-1} + \boldsymbol{\mu}_k(i)[\nabla_{w^*} J_k(\boldsymbol{w}_{k,i-1}) + \boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})] \tag{2.72a}$$

$$\widetilde{\boldsymbol{w}}_{k,i} = \sum_{\ell \in \boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\, \widetilde{\boldsymbol{\psi}}_{\ell,i} \tag{2.72b}$$

Applying the transformation $\mathbb{T}(\cdot)$ from (2.3) to both sides of these equations, we show in Appendix 2.C that the error recursion (2.72a)–(2.72b) becomes

$$\underline{\widetilde{\boldsymbol{\psi}}}_{k,i} = [I_{2M} - \boldsymbol{\mu}_k(i)\boldsymbol{H}_{k,i-1}]\underline{\widetilde{\boldsymbol{w}}}_{k,i-1} + \boldsymbol{\mu}_k(i)\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1}) \tag{2.73a}$$

$$\underline{\widetilde{\boldsymbol{w}}}_{k,i} = \sum_{\ell \in \boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\, \underline{\widetilde{\boldsymbol{\psi}}}_{\ell,i} \tag{2.73b}$$

where we introduced the $2M \times 2M$ matrix:

$$\boldsymbol{H}_{k,i-1} \triangleq \int_0^1 \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o - t\underline{\widetilde{\boldsymbol{w}}}_{k,i-1})\, dt \tag{2.74}$$

To proceed, we introduce the following network variables:

$$\underline{\widetilde{\boldsymbol{w}}}_i \triangleq \mathrm{col}\{\underline{\widetilde{\boldsymbol{w}}}_{1,i}, \underline{\widetilde{\boldsymbol{w}}}_{2,i}, \ldots, \underline{\widetilde{\boldsymbol{w}}}_{N,i}\} \tag{2.75}$$

$$\boldsymbol{\mathcal{M}}_i \triangleq \boldsymbol{M}_i \otimes I_{2M} \tag{2.76}$$

$$\boldsymbol{\mathcal{A}}_i \triangleq \boldsymbol{A}_i \otimes I_{2M} \tag{2.77}$$

$$\boldsymbol{\mathcal{H}}_i \triangleq \mathrm{diag}\{\boldsymbol{H}_{1,i}, \boldsymbol{H}_{2,i}, \ldots, \boldsymbol{H}_{N,i}\} \tag{2.78}$$

Using (2.73a)–(2.73b) we conclude that the network error vector (2.75) evolves according to the following dynamics:

$$\boxed{\underline{\widetilde{\boldsymbol{w}}}_i = \boldsymbol{\mathcal{A}}_i^{\mathsf{T}}(I_{2MN} - \boldsymbol{\mathcal{M}}_i\boldsymbol{\mathcal{H}}_{i-1})\underline{\widetilde{\boldsymbol{w}}}_{i-1} + \boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{\mathcal{M}}_i\underline{\boldsymbol{v}}_i(\boldsymbol{w}_{i-1})} \tag{2.79}$$

2.3.1 Condition for Mean-Square Stability

The main result in this Part I is to establish the mean-square stability of this recursion in Theorem 2.1 below. The difficulty lies in the fact that the error dynamics (2.79) is a time-variant stochastic recursion that also depends nonlinearly on the data. The parameters $\{\boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{M}}_i, \boldsymbol{\mathcal{H}}_{i-1}\}$ are random, time-varying, and multiplied together and by the error vector and noise variables. The statement and proof of Theorem 2.1 rely on the following quantities:

$$\epsilon^2(i) \triangleq \max_k \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 = \frac{1}{2} \cdot \max_k \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 \tag{2.80}$$

$$\gamma_k^2 \triangleq 1 - 2\bar{\mu}_k \lambda_{k,\min} + (\bar{\mu}_k^2 + c_{\mu,k,k})\lambda_{k,\max}^2 \tag{2.81}$$

$$\beta \triangleq \max_k \{\gamma_k^2 + \alpha(\bar{\mu}_k^2 + c_{\mu,k,k})\} \tag{2.82}$$

$$\theta \triangleq \max_k \{\bar{\mu}_k^2 + c_{\mu,k,k}\} \tag{2.83}$$

where the $\{\lambda_{k,\min}, \lambda_{k,\max}\}$ correspond to the lower and upper bounds on the Hessian matrices from (2.7).

Theorem 2.1 (Mean-square stability)**.** *The mean-square stability of the asynchronous diffusion strategy* (2.27a)–(2.27b) *reduces to studying the convergence of the recursive inequality:*

$$\epsilon^2(i) \leq \beta \cdot \epsilon^2(i-1) + \theta\sigma_v^2 \tag{2.84}$$

where σ_v^2 *is from* (2.24)*. The model* (2.84) *is stable if the mean* $\{\bar{\mu}_k\}$ *and the ratio* $\{(\bar{\mu}_k^2 + c_{\mu,k,k})/\bar{\mu}_k\}$ *satisfy the following relation:*

$$\boxed{\frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k} < \frac{\lambda_{k,\min}}{\alpha + \lambda_{k,\max}^2}} \tag{2.85}$$

for $k = 1, 2, \ldots, N$*, where the parameter* α *is from* (2.26)*. When condition* (2.85) *holds, an upper bound on the individual steady-state mean-square-deviation*

(MSD) for each agent k in the network is given by

$$\boxed{\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 \leq b \cdot \nu_o} \tag{2.86}$$

where

$$\nu_o \triangleq \max_k \frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k}, \quad b \triangleq \frac{\kappa\sigma_v^2}{\min_k\{\lambda_{k,\min}\}} \tag{2.87}$$

and the parameter κ is from (2.37)*.*

Proof. See Appendix 2.D. □

From the asynchronous network model, we know that $\boldsymbol{\mu}_k(i) \in [0, \mu_k]$. It follows that

$$\frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k} = \frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}} \leq \frac{\mathbb{E}\,[\boldsymbol{\mu}_k(i)\mu_k]}{\bar{\mu}_k} = \mu_k \tag{2.88}$$

From (2.88), a sufficient condition for (2.85) to hold is given by

$$\mu_k < \frac{\lambda_{k,\min}}{\alpha + \lambda_{k,\max}^2} \tag{2.89}$$

Condition (2.85) allows us to provide some insights about how the dispersion of $\boldsymbol{\mu}_k(i)$ affects mean-square stability. Note that condition (2.85) even allows the *random* step-sizes to assume some abnormally large values at a relatively low probability. This "hopping" behavior (resulting from infrequent large step-sizes) would not destroy the mean-square stability of the network; this fact reveals another useful form of robustness.

Since the constant coefficient b defined in (2.87) is a fixed bound, Theorem 2.1 implies that for sufficiently large i, the MSD of each individual agent's solution has a bounded value. The upper bound in (2.86) is proportional to the parameter ν_o across the network. Using the useful conclusion of (2.86), we will be able to derive in the sequel a condition for fourth-order stability of the error recursion (2.79).

2.3.2 Stability Conditions for Bernoulli and Beta Models

We specialize condition (2.86) for the asynchronous models described in Section 2.2.4.

2.3.2.1 The Bernoulli Model

Substituting (2.57) and (2.58) into (2.85) yields the condition

$$\mu_k < \frac{\lambda_{k,\min}}{\alpha + \lambda_{k,\max}^2} \tag{2.90}$$

which is identical to condition (2.89) on the upper limit of the range of random step-sizes.

2.3.2.2 The Beta Model

Without loss of generality, let $\zeta_k = \phi_k \cdot \xi_k$ with a constant factor $\phi_k > 0$. It follows from (2.66) that the mean value $\bar{\mu}_k$ can be expressed in terms of the factor ϕ_k and the upper limit μ_k:

$$\bar{\mu}_k = \frac{\mu_k}{1 + \phi_k} \tag{2.91}$$

Likewise, from (2.67), we have

$$c_{\mu,k,k} = \frac{\phi_k \mu_k^2}{(1 + \phi_k)^2(\xi_k + \xi_k \phi_k + 1)} \tag{2.92}$$

which is a monotonically decreasing function of the shape parameter $\xi_k \geq 1$. As the value of ξ_k becomes larger, the probability mass of $\boldsymbol{\mu}_k(i)$ will gradually concentrate around its mean (2.91), as shown in Fig. 2.3. Substituting (2.91) and (2.92) into (2.85) yields

$$\mu_k < \left(1 + \frac{\phi_k \xi_k}{1 + \xi_k}\right) \frac{\lambda_{k,\min}}{\alpha + \lambda_{k,\max}^2} \tag{2.93}$$

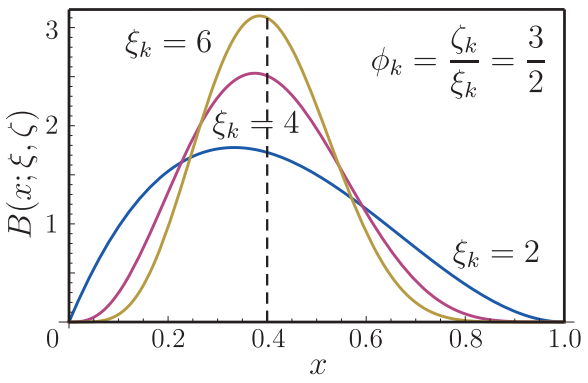


Figure 2.3: The PDFs of the Beta distribution $B(x;\xi_k,\zeta_k)$ for $\zeta_k = 1.5\xi_k$ and $\xi_k = 2, 4, 6$.

where μ_k is the largest possible value for $\boldsymbol{\mu}_k(i)$ defined by (2.64). In (2.93), the bound on μ_k is a monotonically increasing function of the shape parameter $\xi_k \geq 1$. As ξ_k becomes larger, the bound in (2.93) becomes larger. The net effect allows for a wider range for the realizations of the random step-sizes. Moreover, it is easy to verify that the upper bound in (2.93) is larger than that in (2.90).

2.3.3 Condition for Fourth-Order Stability

Result (2.86) establishes that the network is mean-square stable under the assumption of bounded second-order moments for the gradient noise process as in (2.26). If desired, under a similar condition on bounded fourth-order moments for the gradient noise, we can also establish by extending the arguments of Appendix 2.D and [67] that the error recursion (2.79) is stable in the fourth-order sense.

Theorem 2.2 (Stability of fourth-order error moments)**.** *Assume the fourth-order moments of the gradient noise components are bounded by*

$$\mathbb{E}[\|\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4|\mathbb{F}_{i-1}] \leq \alpha^2\|w^o - \boldsymbol{w}_{k,i-1}\|^4 + \sigma_v^4 \tag{2.94}$$

for some constants $\alpha \geq 0$ and $\sigma_v \geq 0$. If

$$\frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k} < \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha} \tag{2.95}$$

holds for all k, then the fourth-order moments of the individual errors are asymptotically bounded by

$$\boxed{\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^4 \leq b_4^2 \cdot \nu^2} \tag{2.96}$$

where the parameter ν is defined by

$$\nu \triangleq \max_k \frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k}, \quad b_4 \triangleq \frac{3\sigma_v^2(\kappa+1)}{\min_k \lambda_{k,\min}} \tag{2.97}$$

Proof. See Appendix 2.E. □

It is easy to verify that condition (2.94) implies a bound on the second-order moment of the gradient noise:

$$\mathbb{E}[\|\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\mathbb{F}_{i-1}] \leq \alpha\|w^o - \boldsymbol{w}_{k,i-1}\|^2 + \sigma_v^2 \tag{2.98}$$

although the converse is generally not true. In other words, it is redundant to assume both conditions (2.26) and (2.94). It can be verified that condition (2.95) implies (2.85) (see (2.192) and (2.193) in Appendix 2.E). Therefore, conditions (2.94) and (2.95) are sufficient to ensure both mean-square and fourth-order stability of error moments. Moreover, it is straightforward to verify that

$$\nu_o = \max_k \frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k} \leq \max_k \frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k} = \nu \tag{2.99}$$

Therefore, we can use ν to upper bound ν_o.

2.4 Conclusion

We introduced a fairly general model for *asynchronous* behavior over networks with random step-sizes, links, topologies, and combination coefficients. We then carried out a mean-square analysis and showed that, even under non-vanishing step-sizes, the asynchronous network remains mean-square stable for sufficiently small step-sizes. We derived a condition on the first and second-order moments of the random step-sizes to ensure stable behavior. We specialized the results to two models: a Bernoulli network and a Beta network. It was observed that the Beta network admits a wider range of step-sizes for stability. The results suggest that networks where step-sizes assume values randomly within a certain interval are robustly more stable than networks that have their step-sizes be turned on or off.

2.A Equivalent Complex-Domain Representations

First, we recall the definition of the *real* Jacobian of a real-valued function $J(w)$ with respect to a real column vector $w \in \mathbb{R}^M$ as

$$\frac{\partial J(w)}{\partial w} \triangleq \text{row}\left\{\frac{\partial J(w)}{\partial w_1}, \frac{\partial J(w)}{\partial w_2}, \ldots, \frac{\partial J(w)}{\partial w_M}\right\} \tag{2.100}$$

where $w_m \in \mathbb{R}$ denotes the m-th element of w. Using (2.1), the real gradient of the function $J_k(\bar{w})$ with respect to $\bar{w}$ is defined as

$$\nabla_{\bar{w}} J_k(\bar{w}) \triangleq \frac{\partial J_k(\bar{w})}{\partial \bar{w}} = \left[\frac{\partial J_k(w)}{\partial \mathfrak{Re}(w)} \quad \frac{\partial J_k(w)}{\partial \mathfrak{Im}(w)}\right] \tag{2.101}$$

and the real Hessian matrix of the same function $J_k(\bar{w})$ with respect to $\bar{w}$ is defined by

$$\nabla^2_{\bar{w}\bar{w}^{\mathsf{T}}} J_k(\bar{w}) \triangleq \frac{\partial}{\partial \bar{w}}\left[\frac{\partial J_k(\bar{w})}{\partial \bar{w}}\right]^{\mathsf{T}} \tag{2.102}$$

It is easy to verify that $\nabla^2_{\bar{w}\bar{w}^{\mathsf{T}}} J_k(\bar{w})$ is a symmetric matrix.

Then, we define the derivative of a real-valued function $J(z)$ with respect to the complex argument $z \in \mathbb{C}$ as [75, 76]:

$$\frac{\partial J(z)}{\partial z} \triangleq \frac{1}{2}\left(\frac{\partial J(z)}{\partial \mathfrak{Re}(z)} - j\frac{\partial J(z)}{\partial \mathfrak{Im}(z)}\right) \tag{2.103}$$

and the *complex* Jacobian of a real-valued function $J(w)$ with respect to the complex column vector $w \in \mathbb{C}^M$ is given by

$$\frac{\partial J(w)}{\partial w} \triangleq \text{row}\left\{\frac{\partial J(w)}{\partial w_1}, \frac{\partial J(w)}{\partial w_2}, \ldots, \frac{\partial J(w)}{\partial w_M}\right\} \tag{2.104}$$

where $w_m \in \mathbb{C}$ denotes the m-th element of w. The complex gradient of the real-valued function $J_k(w)$ with respect to the complex vector argument $w \in \mathbb{C}^M$ is defined as [76, eq. (20)] (compare with (2.101)):

$$\nabla_w J_k(w) \triangleq \frac{\partial J(w)}{\partial w} = \frac{1}{2}\left(\frac{\partial J(w)}{\partial \mathfrak{Re}(w)} - j\frac{\partial J(w)}{\partial \mathfrak{Im}(w)}\right) \tag{2.105}$$

and the complex conjugate gradient of $J_k(w)$ with respect to $w^* \in \mathbb{C}^M$ is defined by [76, eqs. (21, 22)]:

$$\nabla_{w^*} J_k(w) \triangleq \frac{\partial J(w)}{\partial w^*} = [\nabla_w J_k(w)]^* \tag{2.106}$$

Using (2.3), the complex gradient of $J_k(\underline{w})$ with respect to the column vector $\underline{w} \in \mathbb{C}^{2M}_M$ is then given by [76, eq. (18)]:

$$\nabla_{\underline{w}} J_k(\underline{w}) = \frac{\partial J_k(\underline{w})}{\partial \underline{w}} = \begin{bmatrix}\nabla_w J_k(w) & (\nabla_{w^*} J_k(w))^{\mathsf{T}}\end{bmatrix} \tag{2.107}$$

and the corresponding complex conjugate gradient is given by

$$\nabla_{\underline{w}^*} J_k(\underline{w}) = \left[\frac{\partial J_k(\underline{w})}{\partial \underline{w}}\right]^* = [\nabla_{\underline{w}} J_k(\underline{w})]^* \tag{2.108}$$

The complex Hessian of $J_k(\underline{w})$ with respect to $\underline{w} \in \mathbb{C}^{2M}_M$ is defined by [76, eq. (32)]:

$$\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}) \triangleq \begin{bmatrix}\nabla^2_{ww^*} J_k(w) & (\nabla^2_{ww^{\mathsf{T}}} J_k(w))^* \\ \nabla^2_{ww^{\mathsf{T}}} J_k(w) & (\nabla^2_{ww^*} J_k(w))^{\mathsf{T}}\end{bmatrix} \tag{2.109}$$

where

$$\nabla^2_{ww^*} J_k(w) \triangleq \frac{\partial}{\partial w}\left[\frac{\partial J_k(w)}{\partial w}\right]^* \tag{2.110}$$

$$\nabla^2_{ww^\mathsf{T}} J_k(w) \triangleq \frac{\partial}{\partial w}\left[\frac{\partial J_k(w)}{\partial w}\right]^\mathsf{T} \tag{2.111}$$

It is easy to verify that $\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})$ is a Hermitian matrix.

From (2.101) and (2.107), we have [76, eqs. (18, 19)]:

$$\nabla_{\bar{w}} J_k(\bar{w}) = \nabla_{\underline{w}} J_k(\underline{w}) \cdot D \tag{2.112}$$

$$\nabla_{\bar{w}} J_k(\bar{w}) \cdot \frac{1}{2} D^* = \nabla_{\underline{w}} J_k(\underline{w}) \tag{2.113}$$

Similarly, from (2.102) and (2.109), we have [76, eqs. (32, 33)]:

$$\nabla^2_{\bar{w}\bar{w}^\mathsf{T}} J_k(\bar{w}) = D^* \cdot [\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w})] \cdot D \tag{2.114}$$

$$\frac{1}{4} D \cdot [\nabla^2_{\bar{w}\bar{w}^\mathsf{T}} J_k(\bar{w})] \cdot D^* = \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}) \tag{2.115}$$

Identities (2.112)–(2.115) play an important role in our analysis.

2.B Proof of Lemma 2.4

Expression (2.49) is because $\boldsymbol{\mu}_k(i)$ and $\boldsymbol{\mu}_\ell(i)$ are uncorrelated when $k \neq \ell$. Expression (2.51) is obtained by using (2.49) and (2.35). Using (2.28) and Lemma 2.2, we have

$$\boldsymbol{a}_{kk}(i) = 1 - \sum_{\ell \in \boldsymbol{\mathcal{N}}_{k,i} \backslash \{k\}} \boldsymbol{a}_{\ell k}(i) = 1 - \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} \boldsymbol{a}_{\ell k}(i) \tag{2.116}$$

since $\boldsymbol{a}_{\ell k}(i) = 0$ for any $\ell \in \mathcal{N}_k \backslash \boldsymbol{\mathcal{N}}_{k,i}$. When $\ell \neq k$, all entries in $\boldsymbol{A}_i$ are uncorrelated with $\boldsymbol{a}_{\ell k}(i)$ except for the (ℓ, k)-th and (k, k)-th entries. It follows from Lemma 2.2 that

$$c_{a,\ell k,kk} = \mathbb{E}[(\boldsymbol{a}_{\ell k}(i) - \bar{a}_{\ell k})(\boldsymbol{a}_{kk}(i) - \bar{a}_{kk})]$$

$$
\begin{aligned}
&\overset{(a)}{=} -\sum_{n\in\mathcal{N}_k\backslash\{k\}} \mathbb{E}[(\boldsymbol{a}_{\ell k}(i)-\bar{a}_{\ell k})(\boldsymbol{a}_{nk}(i)-\bar{a}_{nk})] \\
&\overset{(b)}{=} -\mathbb{E}(\boldsymbol{a}_{\ell k}(i)-\bar{a}_{\ell k})^2 \\
&\overset{(c)}{=} -c_{a,\ell k,\ell k} \qquad (2.117)
\end{aligned}
$$

for any $\ell \in \mathcal{N}_k\backslash\{k\}$ since (2.117) holds for any realization of the random neighborhood $\boldsymbol{\mathcal{N}}_{k,i}$, and where step (a) is due to (2.116); step (b) is because $\{\boldsymbol{a}_{nk}(i); n \in \boldsymbol{\mathcal{N}}_{k,i}\backslash\{k\}\}$ are all uncorrelated with $\boldsymbol{a}_{\ell k}(i)$ except for $\boldsymbol{a}_{\ell k}(i)$ itself, and $\boldsymbol{a}_{\ell k}(i)=0$ for any $\ell \in \mathcal{N}_k\backslash\boldsymbol{\mathcal{N}}_{k,i}$; and step (c) is because of (2.42). From (2.117), we get (2.50). When $\ell = k$, all entries in $\boldsymbol{A}_i$ are uncorrelated with $\boldsymbol{a}_{kk}(i)$ except for the (ℓ,k)-th entries for all $\ell \in \boldsymbol{\mathcal{N}}_{k,i}$. It follows from Lemma 2.2 that

$$
\begin{aligned}
c_{a,kk,\ell k} &= \mathbb{E}[(\boldsymbol{a}_{kk}(i)-\bar{a}_{kk})(\boldsymbol{a}_{\ell k}(i)-\bar{a}_{\ell k})] \\
&= -c_{a,\ell k,\ell k}, \qquad \ell \in \mathcal{N}_k\backslash\{k\} \qquad (2.118) \\
c_{a,kk,kk} &\overset{(a)}{=} \sum_{\ell,n\in\mathcal{N}_k\backslash\{k\}} \mathbb{E}[(\boldsymbol{a}_{\ell k}(i)-\bar{a}_{\ell k})(\boldsymbol{a}_{nk}(i)-\bar{a}_{nk})] \\
&\overset{(b)}{=} \sum_{\ell\in\mathcal{N}_k\backslash\{k\}} \mathbb{E}(\boldsymbol{a}_{\ell k}(i)-\bar{a}_{\ell k})^2 \\
&\overset{(c)}{=} \sum_{\ell\in\mathcal{N}_k\backslash\{k\}} c_{a,\ell k,\ell k} \qquad (2.119)
\end{aligned}
$$

where (2.118) is because of (2.117); step (a) is because of (2.116); step (b) is because $\{\boldsymbol{a}_{\ell k}(i); \ell \in \boldsymbol{\mathcal{N}}_{k,i}\backslash\{k\}\}$ are mutually-uncorrelated, and $\boldsymbol{a}_{\ell k}(i)=0$ for any $\ell \in \mathcal{N}_k\backslash\boldsymbol{\mathcal{N}}_{k,i}$; and step (c) is because of (2.42). From (2.118) and (2.119), we get (2.50).

2.C Derivation of Error Recursion (2.73a)–(2.73b)

Applying the transformation $\underline{\mathbb{T}}$ from (2.3) to both sides of the error recursion (2.72a)–(2.72b), we get

$$\widetilde{\underline{\boldsymbol{\psi}}}_{k,i} = \widetilde{\underline{\boldsymbol{w}}}_{k,i-1} + \boldsymbol{\mu}_k(i)[\nabla_{\underline{w}^*} J_k(\underline{\boldsymbol{w}}_{k,i-1}) + \underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})] \tag{2.120a}$$

$$\widetilde{\underline{\boldsymbol{w}}}_{k,i} = \sum_{\ell \in \boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\, \widetilde{\underline{\boldsymbol{\psi}}}_{\ell,i} \tag{2.120b}$$

where, by definition,

$$\underline{\mathbb{T}}(\nabla_{w^*} J_k(w)) = \begin{bmatrix} \nabla_{w^*} J_k(w) \\ \nabla_{w^\mathsf{T}} J_k(w) \end{bmatrix} = \nabla_{\underline{w}^*} J_k(\underline{w}) \tag{2.121}$$

The *real* gradient defined by (2.101) can be expressed using the mean-value theorem as [72]:

$$\nabla_{\bar{w}^\mathsf{T}} J_k(\bar{w}) = \left[\int_0^1 \nabla^2_{\bar{w}\bar{w}^\mathsf{T}} J_k(\bar{w}^o - t(\bar{w}^o - \bar{w}))dt \right] (\bar{w} - \bar{w}^o) \tag{2.122}$$

since $\nabla_{\bar{w}^\mathsf{T}} J_k(\bar{w}^o) = 0$ by Assumption 2.1. From (2.5), (2.122), (2.113), and (2.115), we get

$$\begin{aligned} \nabla_{\underline{w}^*} J_k(\underline{w}) &= \frac{1}{2} D \cdot \nabla_{\bar{w}^\mathsf{T}} J_k(\bar{w}) \\ &= \int_0^1 \frac{1}{4} D \left[\nabla^2_{\bar{w}\bar{w}^\mathsf{T}} J_k(\bar{w}^o - t(\bar{w}^o - \bar{w})) \right] D^* dt \cdot D(\bar{w} - \bar{w}^o) \\ &= \left[\int_0^1 \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o - t(\underline{w}^o - \underline{w}))dt \right] \cdot (\underline{w} - \underline{w}^o) \end{aligned} \tag{2.123}$$

Letting $\underline{w} = \underline{\boldsymbol{w}}_{k,i-1}$, we get

$$\nabla_{\underline{w}^*} J_k(\underline{\boldsymbol{w}}_{k,i-1}) = -\left[\int_0^1 \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o - t\widetilde{\underline{\boldsymbol{w}}}_{k,i-1})\, dt \right] \widetilde{\underline{\boldsymbol{w}}}_{k,i-1} \tag{2.124}$$

Then, by (2.124), the error recursion (2.120a) and (2.120b) can be rewritten as (2.73a)–(2.73b).

2.D Proof of Theorem 2.1

We start from equation (2.73b). Since the squared Euclidean norm $\|\cdot\|^2$ is a convex function of its vector argument, using Jensen's inequality [71] we get

$$\|\widetilde{\underline{w}}_{k,i}\|^2 \leq \sum_{\ell\in\boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\|\widetilde{\underline{\psi}}_{\ell,i}\|^2 = \sum_{\ell\in\mathcal{N}_k} \boldsymbol{a}_{\ell k}(i)\|\widetilde{\underline{\psi}}_{\ell,i}\|^2 \tag{2.125}$$

since $\boldsymbol{a}_{\ell k}(i) = 0$ for any $\ell \in \mathcal{N}_k\backslash\boldsymbol{\mathcal{N}}_{k,i}$ by (2.28) and Lemma 2.2. Taking the expectation of both sides of (2.125) and using the asynchronous network model, we get

$$\mathbb{E}\|\widetilde{\underline{w}}_{k,i}\|^2 \leq \sum_{\ell\in\mathcal{N}_k} \bar{a}_{\ell k}\,\mathbb{E}\|\widetilde{\underline{\psi}}_{\ell,i}\|^2 \leq \max_{\ell}\{\mathbb{E}\|\widetilde{\underline{\psi}}_{\ell,i}\|^2\} \tag{2.126}$$

Conditioned on $\mathbb{F}_{i-1}$, the random matrix $\boldsymbol{H}_{k,i-1}$ defined by (2.74) becomes deterministic. Let

$$\boldsymbol{\Sigma}_{k,i} \triangleq [I_{2M} - \boldsymbol{\mu}_k(i)\boldsymbol{H}_{k,i-1}]^2 \tag{2.127}$$

From (2.73a), we get

$$\begin{aligned}
\mathbb{E}(\|\widetilde{\underline{\psi}}_{\ell,i}\|^2|\mathbb{F}_{i-1}) &\overset{(a)}{=} \mathbb{E}(\|\widetilde{\underline{w}}_{k,i-1}\|^2_{\boldsymbol{\Sigma}_{k,i}}|\mathbb{F}_{i-1}) + \mathbb{E}[\boldsymbol{\mu}_k^2(i)\|\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\mathbb{F}_{i-1}] \\
&\overset{(b)}{\leq} \mathbb{E}(\|\boldsymbol{\Sigma}_{k,i}\|\,\|\widetilde{\underline{w}}_{k,i-1}\|^2|\mathbb{F}_{i-1}) + (\bar{\mu}_k^2 + c_{\mu,k,k})\mathbb{E}[\|\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\mathbb{F}_{i-1}] \\
&\overset{(c)}{\leq} \mathbb{E}(\|\boldsymbol{\Sigma}_{k,i}\||\mathbb{F}_{i-1})\|\widetilde{\underline{w}}_{k,i-1}\|^2 + (\bar{\mu}_k^2 + c_{\mu,k,k})(\alpha\|\widetilde{\underline{w}}_{k,i-1}\|^2 + 2\sigma_v^2)
\end{aligned} \tag{2.128}$$

where step (a) is from (2.127) and cross terms are eliminated by using the conditional independence and zero-mean properties of $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})$ from Assumption 2.3; step (b) in (2.128) is due to the asynchronous network model and the sub-multiplicative property of the 2-norm; and step (c) is by conditioning and (2.26). Using Assumptions 2.2 and (2.74) we have

$$1-\boldsymbol{\mu}_k(i)\lambda_{k,\max} \leq \lambda(I_{2M}-\boldsymbol{\mu}_k(i)\boldsymbol{H}_{k,i-1}) \leq 1-\boldsymbol{\mu}_k(i)\lambda_{k,\min} \tag{2.129}$$

Then, from (2.127), we obtain

$$\begin{aligned}\lambda\left(\boldsymbol{\Sigma}_{k,i}\right) &\leq \max\{(1-\boldsymbol{\mu}_k(i)\lambda_{k,\min})^2, (1-\boldsymbol{\mu}_k(i)\lambda_{k,\max})^2\} \\ &= \max\{1-2\boldsymbol{\mu}_k(i)\lambda_{k,\min}+\boldsymbol{\mu}_k^2(i)\lambda_{k,\min}^2, 1-2\boldsymbol{\mu}_k(i)\lambda_{k,\max}+\boldsymbol{\mu}_k^2(i)\lambda_{k,\max}^2\} \\ &\leq 1-2\boldsymbol{\mu}_k(i)\lambda_{k,\min}+\boldsymbol{\mu}_k^2(i)\lambda_{k,\max}^2 \end{aligned} \tag{2.130}$$

because $\boldsymbol{\mu}_k(i)$ is nonnegative. Therefore, we have

$$\begin{aligned}\mathbb{E}(\|\boldsymbol{\Sigma}_{k,i}\||\mathbb{F}_{i-1}) &\overset{(a)}{=} \mathbb{E}[\lambda_{\max}\left(\boldsymbol{\Sigma}_{k,i}\right)|\mathbb{F}_{i-1}] \\ &\overset{(b)}{\leq} \mathbb{E}[1-2\boldsymbol{\mu}_k(i)\lambda_{k,\min}+\boldsymbol{\mu}_k^2(i)\lambda_{k,\max}^2] \\ &\overset{(c)}{=} \gamma_k^2 \end{aligned} \tag{2.131}$$

where step (a) is because $\boldsymbol{\Sigma}_{k,i}$ in (2.127) is Hermitian and positive semi-definite, and its largest singular value coincides with its largest eigenvalue; step (b) is by using (2.130) and the independence condition in the asynchronous model; and step (c) is by (2.81). Substituting (2.131) into (2.128), and taking the expectation of both sides with respect to $\widetilde{\underline{\boldsymbol{w}}}_{i-1}$ yields

$$\mathbb{E}\|\widetilde{\underline{\boldsymbol{\psi}}}_{k,i}\|^2 \leq [\gamma_k^2+\alpha(\bar{\mu}_k^2+c_{\mu,k,k})]\cdot\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^2+2(\bar{\mu}_k^2+c_{\mu,k,k})\sigma_v^2 \tag{2.132}$$

Combining (2.132) and (2.126) yields

$$\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i}\|^2 \leq \max_{\ell}\{[\gamma_\ell^2+\alpha(\bar{\mu}_\ell^2+c_{\mu,\ell,\ell})]\cdot\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{\ell,i-1}\|^2+2(\bar{\mu}_\ell^2+c_{\mu,\ell,\ell})\sigma_v^2\} \tag{2.133}$$

Dividing both sides of (2.133) by 2, and using the fact that $\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i-1}\|^2 = \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^2/2$, we get

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 \leq \left[\max_{\ell}\{\gamma_\ell^2+\alpha(\bar{\mu}_\ell^2+c_{\mu,\ell,\ell})\}\right]\left[\max_{\ell}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{\ell,i-1}\|^2\right]+\left[\max_{\ell}\{\bar{\mu}_\ell^2+c_{\mu,\ell,\ell}\}\right]\sigma_v^2 \tag{2.134}$$

Now since inequality (2.134) holds for every k, using (2.80), we conclude that (2.84) should hold. Propagating (2.84) backwards to the starting point yields

$$\epsilon^2(i) \leq \beta^{i+1}\cdot\epsilon^2(-1)+\theta\sigma_v^2\cdot\sum_{j=0}^{i}\beta^j \tag{2.135}$$

where $\epsilon^2(-1) \triangleq \max_k \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,-1}\|^2$ represents the initial error variance. In order to guarantee a convergent upper bound, we require $|\beta| < 1$, which, by (2.81) and (2.82), is equivalent to

$$|1 - 2\bar{\mu}_k\lambda_{k,\min} + (\bar{\mu}_k^2 + c_{\mu,k,k})(\lambda_{k,\max}^2 + \alpha)| < 1 \tag{2.136}$$

for any k. A sufficient condition for (2.136) is given by

$$\frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k} < \frac{2\lambda_{k,\min}}{\alpha + \lambda_{k,\max}^2} \tag{2.137}$$

It is easy to verify that condition (2.85) is a sufficient condition for (2.137). Therefore, if condition (2.85) holds, then $|\beta| < 1$.

Now under condition (2.137), we obtain from (2.84) that

$$\epsilon^2(i) \leq \beta^{i+1} \cdot \epsilon^2(-1) + \frac{\theta\sigma_v^2(1 - \beta^{i+1})}{1 - \beta} \tag{2.138}$$

When $i \to \infty$, we get an upper bound for the individual MSD:

$$\limsup_{i\to\infty} \epsilon^2(i) \leq \frac{\theta\sigma_v^2}{1 - \beta} \tag{2.139}$$

In the following we simplify the upper bound in (2.139). From (2.82) and (2.81), we get

$$\begin{aligned}
1 - \beta &= 1 - \max_k\{\gamma_k^2 + \alpha(\bar{\mu}_k^2 + c_{\mu,k,k})\} \\
&= 1 - \max_k\{1 - 2\bar{\mu}_k\lambda_{k,\min} + (\bar{\mu}_k^2 + c_{\mu,k,k})(\lambda_{k,\max}^2 + \alpha)\} \\
&= \min_k \left\{\bar{\mu}_k \cdot \left[2\lambda_{k,\min} - \frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k}(\alpha + \lambda_{k,\max}^2)\right]\right\} \\
&\geq \min_k\{\bar{\mu}_k\} \cdot \min_k \left[2\lambda_{k,\min} - \frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k}(\alpha + \lambda_{k,\max}^2)\right]
\end{aligned} \tag{2.140}$$

Using (2.85) again, we get

$$\frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k}(\alpha + \lambda_{k,\max}^2) < \lambda_{k,\min} \tag{2.141}$$

Hence, relation (2.140) can be further expressed as

$$1-\beta \geq \min_k\{\bar{\mu}_k\} \cdot \min_k\{\lambda_{k,\min}\} \tag{2.142}$$

From (2.83) we get

$$\theta \leq \max_k \frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k} \cdot \max_k \bar{\mu}_k \tag{2.143}$$

Therefore, when $i \to \infty$, using (2.142), (2.143), (2.37), and (2.87), we get from (2.139) that

$$\begin{aligned}
\limsup_{i\to\infty} \epsilon^2(i) &\leq \frac{\theta\sigma_v^2}{1-\beta} \\
&\leq \frac{\sigma_v^2}{\min_k\{\lambda_{k,\min}\}} \frac{\max_k\{\bar{\mu}_k\}}{\min_k\{\bar{\mu}_k\}} \max_k \frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k} \\
&\leq \frac{\kappa\sigma_v^2}{\min_k\{\lambda_{k,\min}\}} \cdot \max_k \frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k}
\end{aligned} \tag{2.144}$$

Substituting (2.87) into (2.144) completes the proof.

2.E Proof of Theorem 2.2

From (2.73b) and using Jensen's inequality, we obtain under expectation:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^4 \leq \sum_{\ell\in\mathcal{N}_k} \bar{a}_{\ell k}\, \mathbb{E}\|\widetilde{\underline{\psi}}_{\ell,i}\|^4 \tag{2.145}$$

for all k. Therefore, we have

$$\max_k \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^4 \leq \max_k \mathbb{E}\|\widetilde{\underline{\psi}}_{k,i}\|^4 \tag{2.146}$$

From (2.73a), we have

$$\|\widetilde{\underline{\psi}}_{k,i}\|^4 = \|[I_{2M} - \boldsymbol{\mu}_k(i)\boldsymbol{H}_{k,i-1}]\widetilde{\underline{w}}_{k,i-1} + \boldsymbol{\mu}_k(i)\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4 \tag{2.147}$$

Lemma 2.5 (Fourth-order inequality)**.** *For any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same size, it holds that*

$$\|\boldsymbol{x}+\boldsymbol{y}\|^4 \leq \|\boldsymbol{x}\|^4 + 8\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 + 3\|\boldsymbol{y}\|^4 + 4\|\boldsymbol{x}\|^2\mathfrak{Re}(\boldsymbol{x}^*\boldsymbol{y}) \tag{2.148}$$

Proof. It holds that

$$\begin{aligned}\|\boldsymbol{x}+\boldsymbol{y}\|^4 &= [\|\boldsymbol{x}\|^2 + 2\mathfrak{Re}(\boldsymbol{x}^*\boldsymbol{y}) + \|\boldsymbol{y}\|^2]^2 \\ &= \|\boldsymbol{x}\|^4 + 4[\mathfrak{Re}(\boldsymbol{x}^*\boldsymbol{y})]^2 + \|\boldsymbol{y}\|^4 + 2\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 \\ &\quad + 4\|\boldsymbol{x}\|^2\mathfrak{Re}(\boldsymbol{x}^*\boldsymbol{y}) + 4\mathfrak{Re}(\boldsymbol{x}^*\boldsymbol{y})\|\boldsymbol{y}\|^2 \end{aligned} \tag{2.149}$$

The result now follows by using the inequalities:

$$|\mathfrak{Re}(\boldsymbol{x}^*\boldsymbol{y})|^2 \le \|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2, \;\; 2\mathfrak{Re}(\boldsymbol{x}^*\boldsymbol{y}) \le \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 \tag{2.150}$$

□

Referring to (2.147), if we make the identifications

$$\boldsymbol{x} \equiv [I_{2M} - \boldsymbol{\mu}_k(i)\boldsymbol{H}_{k,i-1}]\widetilde{\boldsymbol{w}}_{k,i-1}, \;\; \boldsymbol{y} \equiv \boldsymbol{\mu}_k(i)\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1}) \tag{2.151}$$

then we obtain

$$\|\boldsymbol{x}\|^2 \le \boldsymbol{a}\cdot\boldsymbol{b}, \qquad \|\boldsymbol{y}\|^2 = \boldsymbol{c}\cdot\boldsymbol{d} \tag{2.152}$$

where

$$\boldsymbol{a} \triangleq 1 - 2\boldsymbol{\mu}_k(i)\lambda_{k,\min} + \boldsymbol{\mu}_k^2(i)\lambda_{k,\max}^2 \tag{2.153}$$

$$\boldsymbol{b} \triangleq \|\widetilde{\boldsymbol{w}}_{k,i-1}\|^2 \tag{2.154}$$

$$\boldsymbol{c} \triangleq \boldsymbol{\mu}_k^2(i) \tag{2.155}$$

$$\boldsymbol{d} \triangleq \|\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2 \tag{2.156}$$

Using Lemma 2.5, we obtain from (2.152) that

$$\|\boldsymbol{x}+\boldsymbol{y}\|^4 \le \boldsymbol{a}^2\cdot\boldsymbol{b}^2 + 8\boldsymbol{a}\cdot\boldsymbol{b}\cdot\boldsymbol{c}\cdot\boldsymbol{d} + 3\boldsymbol{c}^2\cdot\boldsymbol{d}^2 + 4\|\boldsymbol{x}\|^2\mathfrak{Re}(\boldsymbol{x}^*\boldsymbol{y}) \tag{2.157}$$

where

$$\boldsymbol{a}^2 = [1 - 2\boldsymbol{\mu}_k(i)\lambda_{k,\min} + \boldsymbol{\mu}_k^2(i)\lambda_{k,\max}^2]^2$$

$$
\begin{aligned}
&= 1 - 4\boldsymbol{\mu}_k(i)\lambda_{k,\min} + 2\boldsymbol{\mu}_k^2(i)(2\lambda_{k,\min}^2 + \lambda_{k,\max}^2) \\
&\quad - 4\boldsymbol{\mu}_k^3(i)\lambda_{k,\min}\lambda_{k,\max}^2 + \boldsymbol{\mu}_k^4(i)\lambda_{k,\max}^4 \\
&< 1 - 4\boldsymbol{\mu}_k(i)\lambda_{k,\min} + 2\boldsymbol{\mu}_k^2(i)(2\lambda_{k,\min}^2 + \lambda_{k,\max}^2) + \boldsymbol{\mu}_k^4(i)\lambda_{k,\max}^4 \qquad (2.158)
\end{aligned}
$$

$$
\boldsymbol{c}^2 = \boldsymbol{\mu}_k^4(i) \qquad (2.159)
$$

$$
\begin{aligned}
\boldsymbol{a}\cdot\boldsymbol{c} &= \boldsymbol{\mu}_k^2(i) - 2\boldsymbol{\mu}_k^3(i)\lambda_{k,\min} + \boldsymbol{\mu}_k^4(i)\lambda_{k,\max}^2 \\
&\le \boldsymbol{\mu}_k^2(i) + \boldsymbol{\mu}_k^4(i)\lambda_{k,\max}^2 \qquad (2.160)
\end{aligned}
$$

Taking the expectation of (2.157) conditioned on $\mathbb{F}_{i-1}$, we get

$$
\mathbb{E}[\|\boldsymbol{x}+\boldsymbol{y}\|^4|\mathbb{F}_{i-1}] \le \mathbb{E}[\boldsymbol{a}^2]\cdot\boldsymbol{b}^2 + 8\mathbb{E}[\boldsymbol{a}\cdot\boldsymbol{c}]\cdot\boldsymbol{b}\cdot\mathbb{E}[\boldsymbol{d}] + 3\mathbb{E}[\boldsymbol{c}^2]\cdot\mathbb{E}[\boldsymbol{d}^2] \qquad (2.161)
$$

where the last term disappears because $\boldsymbol{y}$ has the noise factor that is conditionally zero mean. From (2.158)–(2.160), we have

$$
\mathbb{E}[\boldsymbol{a}^2] \le 1 - 4\bar{\mu}_k^{(1)}\lambda_{k,\min} + 2\bar{\mu}_k^{(2)}(2\lambda_{k,\min}^2 + \lambda_{k,\max}^2) + \bar{\mu}_k^{(4)}\lambda_{k,\max}^4 \qquad (2.162)
$$

$$
\mathbb{E}[\boldsymbol{c}^2] = \bar{\mu}_k^{(4)} \qquad (2.163)
$$

$$
\mathbb{E}[\boldsymbol{a}\cdot\boldsymbol{c}] \le \bar{\mu}_k^{(2)} + \bar{\mu}_k^{(4)}\lambda_{k,\max}^2 \qquad (2.164)
$$

where $\bar{\mu}_k^{(m)} \triangleq \mathbb{E}[\boldsymbol{\mu}_k^m(i)]$ denotes the m-th moment of the random step-size parameter $\boldsymbol{\mu}_k(i)$. It follows from (2.94) that

$$
\mathbb{E}[\|\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4|\mathbb{F}_{i-1}] \le \alpha^2\cdot\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4 + 4\sigma_v^4 \qquad (2.165)
$$

where a factor of 4 appears because of the transform $\mathbb{T}(\cdot)$. Likewise, it follows from (2.98) that

$$
\mathbb{E}[\|\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\mathbb{F}_{i-1}] \le \alpha\cdot\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^2 + 2\sigma_v^2 \qquad (2.166)
$$

Using (2.165) and (2.166), we can bound the quantities $\mathbb{E}[\boldsymbol{d}^2]$ and $\mathbb{E}[\boldsymbol{d}]$ in (2.161) by

$$
\mathbb{E}[\boldsymbol{d}^2] \le \alpha^2\cdot\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4 + 4\sigma_v^4 = \alpha^2\cdot\boldsymbol{b}^2 + 4\sigma_v^4 \qquad (2.167)
$$

$$\mathbb{E}[\boldsymbol{d}] \leq \alpha \cdot \|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^2 + 2\sigma_v^2 = \alpha \cdot \boldsymbol{b} + 2\sigma_v^2 \tag{2.168}$$

Substituting (2.162)–(2.164) and (2.167)–(2.168) into (2.161), we end up with

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{x}+\boldsymbol{y}\|^4|\mathbb{F}_{i-1}] \\
&\leq [1 - 4\bar{\mu}_k^{(1)}\lambda_{k,\min} + 2\bar{\mu}_k^{(2)}(2\lambda_{k,\min}^2 + \lambda_{k,\max}^2) + \bar{\mu}_k^{(4)}\lambda_{k,\max}^4] \cdot \boldsymbol{b}^2 \\
&\quad + 8[\bar{\mu}_k^{(2)} + \bar{\mu}_k^{(4)}\lambda_{k,\max}^2] \cdot \boldsymbol{b} \cdot (\alpha \cdot \boldsymbol{b} + 2\sigma_v^2) + 3\bar{\mu}_k^{(4)} \cdot (\alpha^2 \cdot \boldsymbol{b}^2 + 4\sigma_v^4) \\
&= [1-4\bar{\mu}_k^{(1)}\lambda_{k,\min}+2\bar{\mu}_k^{(2)}(2\lambda_{k,\min}^2+\lambda_{k,\max}^2+4\alpha)+\bar{\mu}_k^{(4)}(\lambda_{k,\max}^4+8\alpha\lambda_{k,\max}^2+3\alpha^2)] \\
&\quad \times \boldsymbol{b}^2 + 16\sigma_v^2[\bar{\mu}_k^{(2)} + \bar{\mu}_k^{(4)}\lambda_{k,\max}^2] \cdot \boldsymbol{b} + 12\sigma_v^4\bar{\mu}_k^{(4)} \\
&\triangleq (1 - h_{k,1}) \cdot \boldsymbol{b}^2 + h_{k,2} \cdot \boldsymbol{b} + h_{k,3}
\end{aligned} \tag{2.169}$$

where

$$h_{k,1} \triangleq 4\bar{\mu}_k^{(1)}\lambda_{k,\min} - 2\bar{\mu}_k^{(2)}(2\lambda_{k,\min}^2 + \lambda_{k,\max}^2 + 4\alpha) - \bar{\mu}_k^{(4)}(\lambda_{k,\max}^4 + 8\alpha\lambda_{k,\max}^2 + 3\alpha^2) \tag{2.170}$$

$$h_{k,2} \triangleq 16\sigma_v^2(\bar{\mu}_k^{(2)} + \bar{\mu}_k^{(4)}\lambda_{k,\max}^2) \tag{2.171}$$

$$h_{k,3} \triangleq 12\sigma_v^4\bar{\mu}_k^{(4)} \tag{2.172}$$

Substituting (2.151), (2.147), and (2.154) into (2.169), we get

$$\mathbb{E}[\|\widetilde{\underline{\boldsymbol{\psi}}}_{k,i}\|^4|\mathbb{F}_{i-1}] \leq (1 - h_{k,1}) \cdot \|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4 + h_{k,2} \cdot \|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^2 + h_{k,3} \tag{2.173}$$

Taking the expectation with respect to $\mathbb{F}_{i-1}$ yields

$$\mathbb{E}\|\widetilde{\underline{\boldsymbol{\psi}}}_{k,i}\|^4 \leq (1 - h_{k,1}) \cdot \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4 + h_{k,2} \cdot \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^2 + h_{k,3} \tag{2.174}$$

From (2.86) in Theorem 2.1, we know for large enough i that

$$\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^2 \leq 2(b + \epsilon) \cdot \nu \tag{2.175}$$

where we used the fact that $\|\underline{w}\|^2 = 2\|w\|^2$, and $0 < \epsilon \ll 1$ is a small number. Therefore, we can bound $\mathbb{E}\|\widetilde{\underline{\boldsymbol{\psi}}}_{k,i}\|^4$ in (2.174) for large enough i by

$$\mathbb{E}\|\widetilde{\underline{\boldsymbol{\psi}}}_{k,i}\|^4 \leq (1 - h_{k,1}) \cdot \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4 + h_{k,2} \cdot 2(b + \epsilon) \cdot \nu + h_{k,3} \tag{2.176}$$

Substituting (2.176) into (2.146), we get

$$\max_k \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^4 \leq [\max_k(1-h_{k,1})] \cdot \max_k \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i-1}\|^4 + \max_k[h_{k,2} \cdot 2(b+\epsilon)\cdot\nu + h_{k,3}] \tag{2.177}$$

Let

$$\gamma_4 \triangleq \max_k(1-h_{k,1}) = 1 - \min_k h_{k,1} \tag{2.178}$$

$$\theta_4 \triangleq \max_k[h_{k,2} \cdot 2(b+\epsilon)\cdot\nu + h_{k,3}] \tag{2.179}$$

where b is from (2.87). We can then use (2.177) to write for large enough i that

$$\max_k \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^4 \leq \gamma_4 \cdot \max_k \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i-1}\|^4 + \theta_4 \tag{2.180}$$

Therefore, the fourth-order moment of the individual error is governed by (2.180). Whenever $|\gamma_4| < 1$, the quantity $\max_k \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^4$ will have a bounded value asymptotically. In order to guarantee $|\gamma_4| < 1$, it is sufficient to have

$$0 < 4\bar{\mu}_k^{(1)}\lambda_{k,\min} - 2\bar{\mu}_k^{(2)}(2\lambda_{k,\min}^2 + \lambda_{k,\max}^2 + 4\alpha) - \bar{\mu}_k^{(4)}(\lambda_{k,\max}^4 + 8\alpha\lambda_{k,\max}^2 + 3\alpha^2) < 2 \tag{2.181}$$

for all k. This condition can be guaranteed by the sufficient conditions:

$$4\bar{\mu}_k^{(1)}\lambda_{k,\min} < 2 \tag{2.182a}$$

$$\bar{\mu}_k^{(2)}(2\lambda_{k,\min}^2 + \lambda_{k,\max}^2 + 4\alpha) < \bar{\mu}_k^{(1)}\lambda_{k,\min} \tag{2.182b}$$

$$\bar{\mu}_k^{(4)}(\lambda_{k,\max}^4 + 8\alpha\lambda_{k,\max}^2 + 3\alpha^2) < \bar{\mu}_k^{(2)}(2\lambda_{k,\min}^2 + \lambda_{k,\max}^2 + 4\alpha) \tag{2.182c}$$

Condition (2.182a) is equivalent to

$$\bar{\mu}_k^{(1)} < \frac{1}{2\lambda_{k,\min}} \tag{2.183}$$

Condition (2.182b) holds if

$$\frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha} \tag{2.184}$$

Condition (2.182c) holds if

$$\frac{\bar{\mu}_k^{(4)}}{\bar{\mu}_k^{(2)}} < \frac{1}{\lambda_{k,\max}^2 + 4\alpha} \tag{2.185}$$

because

$$\frac{\lambda_{k,\max}^2 + 4\alpha}{(\lambda_{k,\max}^2 + 4\alpha)^2} < \frac{2\lambda_{k,\min}^2 + \lambda_{k,\max}^2 + 4\alpha}{\lambda_{k,\max}^4 + 8\alpha\lambda_{k,\max}^2 + 3\alpha^2} \tag{2.186}$$

Since, for any random variable $\boldsymbol{\mu}_k(i)$,

$$[\bar{\mu}_k^{(1)}]^2 \leq \bar{\mu}_k^{(2)}, \qquad [\bar{\mu}_k^{(2)}]^2 \leq \bar{\mu}_k^{(4)} \tag{2.187}$$

it is straightforward that

$$\max\left\{ [\bar{\mu}_k^{(1)}]^2, \left(\frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}}\right)^2, \frac{\bar{\mu}_k^{(4)}}{\bar{\mu}_k^{(2)}} \right\} \leq \frac{\bar{\mu}_k^{(4)}}{[\bar{\mu}_k^{(1)}]^2} \tag{2.188}$$

On the other hand, it can be verified that

$$\frac{\lambda_{k,\min}^2}{(3\lambda_{k,\max}^2 + 4\alpha)^2} < \min\left\{ \frac{1}{4\lambda_{k,\min}^2}, \frac{1}{\lambda_{k,\max}^2 + 4\alpha} \right\} \tag{2.189}$$

Therefore, if condition (2.95) holds for all k, then (2.183)–(2.185) hold, and $|\gamma_4| < 1$ holds. Using (2.188) and the new definition of ν in (2.97), we obtain

$$\bar{\mu}_k^{(1)} \leq \nu, \quad \bar{\mu}_k^{(2)} \leq \nu^2, \quad \bar{\mu}_k^{(4)} \leq \nu^4 \tag{2.190}$$

Using (2.190), we have

$$h_{k,2} \leq 16\sigma_v^2\nu^2(1 + \lambda_{k,\max}^2\nu^2), \quad h_{k,3} \leq 12\sigma_v^4\nu^4 \tag{2.191}$$

It is worth noting that the new definition of ν in (2.97) bounds the old definition in (2.87) from above since

$$\frac{\bar{\mu}_k^2 + c_{\mu,k,k}}{\bar{\mu}_k} = \frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}} \leq \frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} \tag{2.192}$$

due to (2.188). It is easy to verify that

$$\frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2+4\alpha} < \frac{\lambda_{k,\min}}{\alpha+\lambda_{k,\max}^2} \tag{2.193}$$

With (2.192) and (2.193), it is obvious that (2.95) implies (2.85).

When $|\gamma_4|<1$, the recursive inequality (2.180) leads to

$$\limsup_{i\to\infty}\left[\max_k \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^4\right] \le \frac{\theta_4}{1-\gamma_4} \tag{2.194}$$

Substituting (2.171) and (2.172) into (2.179) yields

$$\begin{aligned}\theta_4 &\le \max_k[16\sigma_v^2(\bar{\mu}_k^{(2)}+\bar{\mu}_k^{(4)}\lambda_{k,\max}^2)\cdot 2(b+\epsilon)\nu+12\sigma_v^4\bar{\mu}_k^{(4)}] \\ &= \max_k\left[32\sigma_v^2\bar{\mu}_k^{(2)}\left(1+\frac{\bar{\mu}_k^{(4)}}{\bar{\mu}_k^{(2)}}\lambda_{k,\max}^2\right)(b+\epsilon)\nu+12\sigma_v^4\bar{\mu}_k^{(4)}\right]\end{aligned} \tag{2.195}$$

where ν is given by (2.97). Using (2.95) and (2.188), we have

$$\frac{\bar{\mu}_k^{(4)}}{\bar{\mu}_k^{(2)}}\lambda_{k,\max}^2 < \frac{\lambda_{k,\max}^2\lambda_{k,\min}^2}{(3\lambda_{k,\max}^2+4\alpha)^2} \le \frac{\lambda_{k,\max}^4}{(3\lambda_{k,\max}^2)^2} = \frac{1}{9} \tag{2.196}$$

Substituting (2.196) into (2.195) yields

$$\begin{aligned}\theta_4 &\le \max_k\left[32\sigma_v^2\bar{\mu}_k^{(2)}\frac{10}{9}(b+\epsilon)\nu+12\sigma_v^4\bar{\mu}_k^{(4)}\right] \\ &\overset{(a)}{\le} \max_k\left[12\sigma_v^2\bar{\mu}_k^{(2)}\left(3b\nu+\sigma_v^2\frac{\bar{\mu}_k^{(4)}}{\bar{\mu}_k^{(2)}}\right)\right] \\ &\overset{(b)}{\le} \max_k\left[12\sigma_v^2\bar{\mu}_k^{(2)}(3b\nu+\sigma_v^2\nu^2)\right] \\ &= \max_k\left[12\sigma_v^2\bar{\mu}_k^{(1)}\frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}}(3b\nu+\sigma_v^2\nu^2)\right] \\ &\overset{(c)}{\le} \max_k\left[12\sigma_v^2\bar{\mu}_k^{(1)}\nu^2(3b+\sigma_v^2\nu)\right]\end{aligned} \tag{2.197}$$

where step (a) is by choosing $\epsilon \le b/80$; and steps (b) and (c) are by using (2.97) and (2.188). Substituting (2.182b) and (2.182c) into (2.170) yields

$$h_{k,1} \ge \bar{\mu}_k^{(1)}\lambda_{k,\min} \tag{2.198}$$

It follows from (2.178) and (2.198) that

$$1-\gamma_4=\min_k h_{k,1}\geq\min_k[\bar{\mu}_k^{(1)}\lambda_{k,\min}]\geq\min_k\bar{\mu}_k^{(1)}\cdot\min_k\lambda_{k,\min}\tag{2.199}$$

Substituting (2.197) and (2.199) into (2.194), we arrive at

$$\begin{aligned}\limsup_{i\to\infty}\left[\max_k\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^4\right]&\leq\frac{12\sigma_v^2\nu^2(3b+\sigma_v^2\nu)\cdot\max_k\bar{\mu}_k^{(1)}}{\min_k\bar{\mu}_k^{(1)}\cdot\min_k\lambda_{k,\min}}\\&\leq\frac{12\sigma_v^2(3b+\sigma_v^2\nu)}{\min_k\lambda_{k,\min}}\frac{\max_k\bar{\mu}_k^{(1)}}{\min_k\bar{\mu}_k^{(1)}}\nu^2\\&\leq\frac{12\kappa\sigma_v^2(3b+\sigma_v^2\nu)}{\min_k\lambda_{k,\min}}\nu^2\end{aligned}\tag{2.200}$$

where we used (2.37) in the last step. From (2.95) and (2.97), it is easy to verify that

$$\nu<\max_k\frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2+4\alpha}\leq\frac{1}{3\min_k\lambda_{k,\min}}\tag{2.201}$$

Then, from (2.87) and (2.201), we obtain

$$3b+\sigma_v^2\nu\leq\frac{3\kappa\sigma_v^2}{\min_k\lambda_{k,\min}}+\frac{\sigma_v^2}{3\min_k\lambda_{k,\min}}<\frac{3\sigma_v^2(\kappa+1)}{\min_k\lambda_{k,\min}}\tag{2.202}$$

Therefore, we obtain from (2.200) and (2.202) that

$$\limsup_{i\to\infty}\left[\max_k\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^4\right]\leq b_4^2\cdot\nu^2=O(\nu^2)\tag{2.203}$$

due to the identity $\|\underline{w}\|^4=4\cdot\|w\|^4$, where b_4 is given by (2.97).

CHAPTER 3

Mean-Square-Error Performance of Asynchronous Networks

In this chapter, we shall derive analytical expressions for the mean-square convergence rate and the steady-state mean-square-deviation. The expressions will reveal how the various parameters of the asynchronous behavior influence network performance. In this process, we shall establish the interesting conclusion that even under the influence of asynchronous events, all agents in the adaptive network can still reach an $O(\nu^{1+\gamma_o'})$ near-agreement with some $\gamma_o' > 0$ while approaching the desired solution within $O(\nu)$ accuracy, where ν is proportional to the small step-size parameter for adaptation. The results in this chapter are based on material from [77].

3.1 Network Error Dynamics

In order to study the mean-square-error performance in *steady-state*, it is necessary to strengthen the assumption on the stochastic gradient vectors, i.e., $\{\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1})\}$. We replace the gradient noise model described in Assumption 2.3 in Section 2.2.1 of Chapter 2 by the following one.

Assumption 3.1 (Gradient noise model)**.**

1. *The gradient noise $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})$, conditioned on $\mathbb{F}_{i-1}$, is assumed to be inde-*

pendent of any other random sources including topology, links, combination coefficients, and step-sizes. The conditional moments of $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})$ *satisfy:*

$$\mathbb{E}\,[\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})|\mathbb{F}_{i-1}]=0 \tag{3.1}$$

$$\mathbb{E}\,[\|\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4|\mathbb{F}_{i-1}]\leq\alpha^2\|w^o-\boldsymbol{w}_{k,i-1}\|^4+\sigma_v^4 \tag{3.2}$$

for some $\alpha\geq 0$ *and* $\sigma_v^2\geq 0$.

2. *The individual gradient noises* $\{\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})\}$ *are uncorrelated and circular across all agents such that*

$$\mathcal{R}_i(\boldsymbol{w}_{i-1}) = \mathrm{diag}\{R_{1,i}(\boldsymbol{w}_{1,i-1}),\ldots,R_{N,i}(\boldsymbol{w}_{N,i-1})\} \tag{3.3}$$

where $\mathcal{R}_i(\boldsymbol{w}_{i-1})$ *and* $\{R_{k,i}(\boldsymbol{w}_{k,i-1})\}$ *are from* (2.22) *and* (2.18) *both in Chapter 2.*

3. *The conditional covariance of* $\underline{\boldsymbol{v}}_i(\boldsymbol{w}_{i-1})$ *satisfies the Lipschitz condition*

$$\|\mathcal{R}_i(\mathbb{1}_N\otimes w^o)-\mathcal{R}_i(\boldsymbol{w}_{i-1})\|\leq\kappa_v\|\mathbb{1}_N\otimes w^o-\boldsymbol{w}_{i-1}\|^{\gamma_v} \tag{3.4}$$

for some constants $\kappa_v\geq 0$ *and* $0<\gamma_v\leq 4$.

4. *The covariance of* $\underline{\boldsymbol{v}}_i(\mathbb{1}_N\otimes w^o)$ *converges to a constant matrix:*

$$\mathcal{R}\triangleq\lim_{i\to\infty}\mathcal{R}_i(\mathbb{1}_N\otimes w^o)\triangleq\mathrm{diag}\{R_1,\ldots,R_N\} \tag{3.5}$$

where

$$R_k\triangleq\lim_{i\to\infty}R_{k,i}(w^o) \tag{3.6}$$

□

From Assumption 3.1, the conditional moments of $\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})$ satisfy

$$\mathbb{E}[\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})|\mathbb{F}_{i-1}]=0 \tag{3.7}$$

$$\mathbb{E}[\|\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4|\mathbb{F}_{i-1}]\leq\alpha^2\|\underline{w}^o-\underline{\boldsymbol{w}}_{k,i-1}\|^4+4\sigma_v^4 \tag{3.8}$$

where a factor of 4 appeared due to the transform $\mathbb{T}(\cdot)$ from (2.3) of Chapter 2.

3.1.1 Long Term Error Dynamics

The error recursion for the asynchronous network (1.7a)–(1.7b) from Chapter 1 is given by (2.79) from Chapter 2, where $\boldsymbol{\mathcal{H}}_{i-1} = \text{diag}\{\boldsymbol{H}_{1,i-1}, \boldsymbol{H}_{2,i-1}, \ldots, \boldsymbol{H}_{N,i-1}\}$ and

$$\boldsymbol{H}_{k,i-1} \triangleq \int_0^1 \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o - t\widetilde{\underline{\boldsymbol{w}}}_{k,i-1})\, dt \tag{3.9}$$

The dependency of $\boldsymbol{\mathcal{H}}_{i-1}$ on the previous iterate $\boldsymbol{w}_{i-1}$ complicates the mean-square analysis. Recall though from Lemma 2.1 in Chapter 2 that the Hessian matrices of the costs $\{J_k(\underline{w})\}$ are globally Lipschitz around $\underline{w}^o$. Let

$$H_k \triangleq \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o), \quad \mathcal{H} \triangleq \text{diag}\{H_1, \ldots, H_N\} \tag{3.10}$$

Recursion (2.79) from Chapter 2 can then be rewritten as

$$\widetilde{\underline{\boldsymbol{w}}}_i = \boldsymbol{\mathcal{A}}_i^{\mathsf{T}}(I_{2MN} - \boldsymbol{\mathcal{M}}_i\mathcal{H})\widetilde{\underline{\boldsymbol{w}}}_{i-1} + \boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{\mathcal{M}}_i\underline{\boldsymbol{v}}_i(\boldsymbol{w}_{i-1}) + \boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{d}_i \tag{3.11}$$

where the perturbation factor $\boldsymbol{d}_i$ is given by

$$\boldsymbol{d}_i \triangleq \boldsymbol{\mathcal{M}}_i(\mathcal{H} - \boldsymbol{\mathcal{H}}_{i-1})\widetilde{\underline{\boldsymbol{w}}}_{i-1} \triangleq \text{col}\{\boldsymbol{d}_{1,i}, \ldots, \boldsymbol{d}_{N,i}\} \tag{3.12}$$

$$\boldsymbol{d}_{k,i} \triangleq \boldsymbol{\mu}_k(i)(H_k - \boldsymbol{H}_{k,i-1})\widetilde{\boldsymbol{w}}_{k,i-1} \tag{3.13}$$

Let $\bar{\mu}_k^{(n)} \triangleq \mathbb{E}[\boldsymbol{\mu}_k(i)]^n$ denote the n-th moment of the random step-size parameter $\boldsymbol{\mu}_k(i)$; we also use $\bar{\mu}_k \equiv \bar{\mu}_k^{(1)}$ from (2.33) of Chapter 2 for the mean and $c_{\mu,k,\ell} = \mathbb{E}[(\boldsymbol{\mu}_k(i) - \bar{\mu}_k)(\boldsymbol{\mu}_\ell(i) - \bar{\mu}_\ell)]$ from (2.36) of Chapter 2 for the cross-covariance.

Lemma 3.1 (Size of perturbation)**.** *If condition* (2.95) *in Chapter 2, namely,*

$$\frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha} \tag{3.14}$$

holds for all k, then

$$\limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{d}_i\|^2 \leq O(\nu^4) \tag{3.15}$$

where

$$\nu \triangleq \max_k \frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} \tag{3.16}$$

Proof. See Appendix 3.A. □

Assumption 3.2 (Small step-sizes)**.** *The parameter ν from* (3.16) *is sufficiently small such that*

$$\nu < \min_k \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha} < 1 \tag{3.17}$$

□

Under Assumption 3.2, condition (3.14) holds. It was shown in (2.192) and (2.193) from Chapter 2 that condition (3.14) in this Part implies condition (2.85) from Chapter 2, i.e.,

$$\frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{\alpha + \lambda_{k,\max}^2} \tag{3.18}$$

for all k.

Since we are interested in examining the asymptotic performance of the asynchronous network, result (3.15) indicates that the network error recursion (2.79) from Chapter 2 can be expressed for large enough i by using the following *long-term* model:

$$\widetilde{\underline{\boldsymbol{w}}}_i' = \boldsymbol{\mathcal{A}}_i^{\mathsf{T}}(I_{2MN} - \boldsymbol{\mathcal{M}}_i\mathcal{H})\widetilde{\underline{\boldsymbol{w}}}_{i-1}' + \boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{\mathcal{M}}_i\underline{\boldsymbol{v}}_i(\boldsymbol{w}_{i-1}) \tag{3.19}$$

where we ignore the $O(\nu^2)$ term $\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{d}_i$ according to (3.15), and we use $\boldsymbol{w}_{i-1}'$ to denote the estimate obtained from this long-term model. It is worth noting that the gradient noise $\underline{\boldsymbol{v}}_i(\boldsymbol{w}_{i-1})$ in (3.19) is an extraneous noise that is imported from the original model (2.79) from Chapter 2; it only depends on the original estimate $\boldsymbol{w}_{i-1}$ but not on $\boldsymbol{w}_{i-1}'$. We will now use recursion (3.19) to determine expressions (rather than bounds) for the steady-state individual MSD and for the average

network MSD. One advantage of model (3.19) is that the random matrix $\boldsymbol{\mathcal{H}}_{i-1}$ from (2.79) from Chapter 2 has been replaced by the constant matrix $\mathcal{H}$. More formally, under Assumption 3.1 on the fourth-order moment of the gradient noise, and by extending the arguments of Appendices 2.D and 2.E from Chapter 2 and the arguments of [67], we will establish later in (3.71) that the MSD expression resulting from (3.19) is within $O(\nu^{3/2})$ of the MSD expression for the original recursion (2.79) from Chapter 2; this conclusion will rely on the following useful result.

Theorem 3.1 (Bounded mean-square gap)**.** *Under Assumptions 3.1 and 3.2, the mean-square gap from the original error recursion* (2.79) *from Chapter 2 to the long-term model* (3.19) *is then asymptotically bounded by*

$$\limsup_{i\to\infty}\left[\max_{k}\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i}-\widetilde{\underline{\boldsymbol{w}}}'_{k,i}\|^2\right]\leq O(\nu^2) \tag{3.20}$$

for any k.

Proof. See Appendix 3.B. □

To proceed with the mean-square-error performance analysis, we introduce the following auxiliary variables:

$$\boldsymbol{D}_{k,i}\triangleq I_{2M}-\boldsymbol{\mu}_k(i)H_k \tag{3.21}$$

$$\boldsymbol{\mathcal{D}}_i\triangleq I_{2MN}-\boldsymbol{\mathcal{M}}_i\mathcal{H}=\mathrm{diag}\{\boldsymbol{D}_{k,i}\} \tag{3.22}$$

$$\boldsymbol{\mathcal{B}}_i\triangleq\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{\mathcal{D}}_i \tag{3.23}$$

$$\underline{\boldsymbol{s}}_i\triangleq\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{\mathcal{M}}_i\underline{\boldsymbol{v}}_i(\boldsymbol{w}_{i-1}) \tag{3.24}$$

Based on the gradient noise model in Assumption 3.1 and the asynchronous network model described in Section 2.2.2 of Chapter 2, it is easy to verify that the (conditional) means of $\{\boldsymbol{\mathcal{A}}_i,\boldsymbol{\mathcal{M}}_i,\boldsymbol{D}_{k,i},\boldsymbol{\mathcal{D}}_i,\boldsymbol{\mathcal{B}}_i,\underline{\boldsymbol{s}}_i\}$ are given by:

$$\bar{\mathcal{A}}\triangleq\mathbb{E}(\boldsymbol{\mathcal{A}}_i)=\bar{A}\otimes I_{2M} \tag{3.25}$$

$$\bar{\mathcal{M}} \triangleq \mathbb{E}(\boldsymbol{\mathcal{M}}_i) = \bar{M} \otimes I_{2M} \tag{3.26}$$

$$\bar{D}_k \triangleq \mathbb{E}(\boldsymbol{D}_{k,i}) = I_{2M} - \bar{\mu}_k H_k \tag{3.27}$$

$$\bar{\mathcal{D}} \triangleq \mathbb{E}(\boldsymbol{\mathcal{D}}_i) = I_{2MN} - \bar{\mathcal{M}}\mathcal{H} = \mathrm{diag}\{\bar{D}_k\} \tag{3.28}$$

$$\bar{\mathcal{B}} \triangleq \mathbb{E}(\boldsymbol{\mathcal{B}}_i) = \bar{\mathcal{A}}^{\mathsf{T}}\bar{\mathcal{D}} \tag{3.29}$$

$$\bar{s} \triangleq \mathbb{E}(\underline{\boldsymbol{s}}_i|\mathbb{F}_{i-1}) = 0 \tag{3.30}$$

It can be verified that the block-Kronecker-covariance matrices of several random quantities are given by:

$$\mathcal{C}_A \triangleq \mathbb{E}[(\boldsymbol{\mathcal{A}}_i - \bar{\mathcal{A}}) \otimes_b (\boldsymbol{\mathcal{A}}_i - \bar{\mathcal{A}})] = C_A \otimes I_{4M^2} \tag{3.31}$$

$$\mathcal{C}_M \triangleq \mathbb{E}[(\boldsymbol{\mathcal{M}}_i - \bar{\mathcal{M}}) \otimes_b (\boldsymbol{\mathcal{M}}_i - \bar{\mathcal{M}})] = C_M \otimes I_{4M^2} \tag{3.32}$$

$$\mathcal{C}_D \triangleq \mathbb{E}[(\boldsymbol{\mathcal{D}}_i^* - \bar{\mathcal{D}}^*)^{\mathsf{T}} \otimes_b (\boldsymbol{\mathcal{D}}_i - \bar{\mathcal{D}})] = \mathcal{C}_M(\mathcal{H}^{\mathsf{T}} \otimes_b \mathcal{H}) \tag{3.33}$$

$$\begin{aligned}\mathcal{C}_B &\triangleq \mathbb{E}[(\boldsymbol{\mathcal{B}}_i^* - \bar{\mathcal{B}}^*)^{\mathsf{T}} \otimes_b (\boldsymbol{\mathcal{B}}_i - \bar{\mathcal{B}})] \\ &= (\bar{\mathcal{A}}^{\mathsf{T}} \otimes_b \bar{\mathcal{A}}^{\mathsf{T}})\mathcal{C}_D + \mathcal{C}_A^{\mathsf{T}}(\bar{\mathcal{D}}^{\mathsf{T}} \otimes_b \bar{\mathcal{D}} + \mathcal{C}_D)\end{aligned} \tag{3.34}$$

where the symbol $\otimes_b$ denotes the block-Kronecker operation of block size $2M \times 2M$ (see Appendix 3.C). Moreover, it can be verified by using property (3.128) from Appendix 3.C that

$$\mathbb{E}[(\boldsymbol{\mathcal{X}}^* - \bar{\mathcal{X}}^*)^{\mathsf{T}} \otimes_b (\boldsymbol{\mathcal{X}} - \bar{\mathcal{X}})] = \mathbb{E}[(\boldsymbol{\mathcal{X}}^*)^{\mathsf{T}} \otimes_b \boldsymbol{\mathcal{X}}] - (\bar{\mathcal{X}}^*)^{\mathsf{T}} \otimes_b \bar{\mathcal{X}} \tag{3.35}$$

for any random block matrix $\boldsymbol{\mathcal{X}}$ with appropriate block size and with mean $\bar{\mathcal{X}} \triangleq \mathbb{E}\boldsymbol{\mathcal{X}}$. The $\{C_A, C_M\}$ that appear in (3.31)–(3.34) relate to the second-order moments of $\{\boldsymbol{a}_{\ell k}(i)\}$ and $\{\boldsymbol{\mu}_k(i)\}$. Using (3.23) and (3.24), the long-term model (3.19) can be rewritten as

$$\widetilde{\underline{\boldsymbol{w}}}_i' = \boldsymbol{\mathcal{B}}_i \cdot \widetilde{\underline{\boldsymbol{w}}}_{i-1}' + \underline{\boldsymbol{s}}_i \tag{3.36}$$

3.1.2 Mean Error Recursion

Taking the expectation of both sides of (3.36), we end up with the mean error recursion for large i:

$$\mathbb{E}\,\widetilde{\boldsymbol{w}}_i' = \bar{\mathcal{B}} \cdot \mathbb{E}\,\widetilde{\boldsymbol{w}}_{i-1}' \tag{3.37}$$

The stability of recursion (3.37) requires the stability of $\bar{\mathcal{B}}$. A condition on the step-sizes to ensure the stability of $\bar{\mathcal{B}}$ can be derived as follows. Using the fact that $\bar{\mathcal{A}}$ is block left-stochastic and $\bar{\mathcal{D}}$ is block diagonal and Hermitian, and following the same argument in [78, App. A] [35], we obtain

$$\rho(\bar{\mathcal{B}}) \leq \rho(\bar{\mathcal{D}}) \tag{3.38}$$

where $\rho(\cdot)$ denotes the spectral radius of its matrix argument. It follows from (3.28) and (3.38) that asymptotic mean stability is guaranteed if the mean step-size $\bar{\mu}_k$ satisfies

$$\bar{\mu}_k \equiv \bar{\mu}_k^{(1)} < \frac{2}{\rho(H_k)} \tag{3.39}$$

for all k. Since H_k is a positive semi-definite matrix, its spectral radius coincides with its largest eigenvalue. Using (2.7) from Chapter 2, we have $\rho(H_k) \leq \lambda_{k,\max}$. If condition (3.14) holds, then from (2.188) of Chapter 2, we have

$$\bar{\mu}_k^{(1)} \leq \frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha} \leq \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2} \leq \frac{2}{\rho(H_k)} \tag{3.40}$$

since $\alpha > 0$. Therefore, condition (3.39) holds if condition (3.14) does so. With Assumption 3.2, we have

$$\lim_{i\to\infty} \mathbb{E}\,\widetilde{\boldsymbol{w}}_{k,i}' = 0 \tag{3.41}$$

for all k. From (3.41), we conclude that the long-term model (3.19) or, equivalently, (3.36), is the asymptotically centered version of the original error recursion (2.79) in Chapter 2.

3.1.3 Error Covariance Recursion

We proceed to examine the evolution of the covariance matrix of the network error vector $\widetilde{\underline{\boldsymbol{w}}}_i'$ in the long-term model (3.36). Let

$$r_i(\boldsymbol{w}_{i-1}) \triangleq \text{bvec}(\mathcal{R}_i(\boldsymbol{w}_{i-1})) = \mathbb{E}[(\underline{\boldsymbol{v}}_i^*(\boldsymbol{w}_{i-1}))^{\mathsf{T}} \otimes_b \underline{\boldsymbol{v}}_i(\boldsymbol{w}_{i-1})|\mathbb{F}_{i-1}] \tag{3.42}$$

$$y_i \triangleq (\bar{\mathcal{A}}^{\mathsf{T}} \otimes_b \bar{\mathcal{A}}^{\mathsf{T}} + \mathcal{C}_A^{\mathsf{T}})(\bar{\mathcal{M}} \otimes_b \bar{\mathcal{M}} + \mathcal{C}_M)\mathbb{E}[r_i(\boldsymbol{w}_{i-1})] \tag{3.43}$$

$$z_i \triangleq \text{bvec}(\mathbb{E}\,(\widetilde{\underline{\boldsymbol{w}}}_i'\widetilde{\underline{\boldsymbol{w}}}_i'^*)) = \mathbb{E}\,[(\widetilde{\underline{\boldsymbol{w}}}_i'^*)^{\mathsf{T}} \otimes_b \widetilde{\underline{\boldsymbol{w}}}_i'] \tag{3.44}$$

$$\mathcal{G} \triangleq \mathbb{E}[(\boldsymbol{\mathcal{D}}_i^*)^{\mathsf{T}} \otimes_b \boldsymbol{\mathcal{D}}_i] = \bar{\mathcal{D}}^{\mathsf{T}} \otimes_b \bar{\mathcal{D}} + \mathcal{C}_D \tag{3.45}$$

$$\mathcal{F} \triangleq \mathbb{E}[(\boldsymbol{\mathcal{B}}_i^*)^{\mathsf{T}} \otimes_b \boldsymbol{\mathcal{B}}_i]^* = \bar{\mathcal{B}}^{\mathsf{T}} \otimes_b \bar{\mathcal{B}}^* + \mathcal{C}_B^* = \mathcal{G}(\bar{\mathcal{A}} \otimes_b \bar{\mathcal{A}} + \mathcal{C}_A) \tag{3.46}$$

where the notations $\text{bvec}(\cdot)$ and $\otimes_b$ denote block vectorization and block Kronecker products, respectively, both of size $2M \times 2M$ (see Appendix 3.C). We note that the second equalities in (3.42) and (3.44) are due to property (3.125) and the second equalities in (3.45) and (3.46) are by using (3.28), (3.29), and (3.33)–(3.35). Using (3.42)–(3.46), we obtain the following recursion for the block-vectorized covariance matrix of the network error vector $\widetilde{\underline{\boldsymbol{w}}}_i'$.

Theorem 3.2 (Network error covariance recursion)**.** *The vector z_i evolves according to the following recursion:*

$$z_i = \mathcal{F}^* z_{i-1} + y_i \tag{3.47}$$

Recursion (3.47) *converges if condition* (3.14) *holds, and its convergence rate is determined by $\rho(\mathcal{F})$.*

Proof. See Appendix 3.D. □

The vector z_i can be used to compute useful error metrics. For example, we can examine any weighted MSE measure for $\widetilde{\underline{\boldsymbol{w}}}_i'$ by evaluating quantities of the

form:

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|_{\Sigma}^2 = \mathbb{E}\,[\mathrm{Tr}(\widetilde{\boldsymbol{w}}_i'\widetilde{\boldsymbol{w}}_i'^{*}\Sigma)] = z_i^{*}\cdot \mathrm{bvec}(\Sigma) \tag{3.48}$$

where Σ is an arbitrary positive semi-definite weight matrix. To guarantee the convergence of $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|_{\Sigma}^2$ for any weighting matrix Σ, it is sufficient and necessary to guarantee the convergence of z_i. It follows from Theorem 3.2 that under Assumption 3.2, the spectral radius of the matrix $\mathcal{F}$ in (3.47) determines the mean-square stability and convergence rate of the asynchronous diffusion strategy (1.7a)–(1.7b) from Chapter 1.

Before proceeding we comment on the reason why we choose to use the block vectorization operation $\mathrm{bvec}(\cdot)$ in (3.44) instead of the traditional vectorization operation $\mathrm{vec}(\cdot)$. This is because $\mathrm{bvec}(\cdot)$ allows us to track each block of its matrix argument *after* vectorization. By the definition in (3.120) and the illustration in Fig. 3.1, operation $\mathrm{bvec}(\cdot)$ preserves the locality of every block in the original matrix argument whereas operation $\mathrm{vec}(\cdot)$ blends different blocks together. Therefore, whenever we need to vectorize a network matrix whose blocks relate to individual agents, it is more natural to use the block vectorization operation $\mathrm{bvec}(\cdot)$; on the other hand, whenever we need to vectorize a matrix that only relates to a single agent, we can use the conventional vectorization operation $\mathrm{vec}(\cdot)$. A useful property of the conventional vectorization operation $\mathrm{vec}(\cdot)$ is

$$\mathrm{vec}(ABC) = (C^{\mathsf{T}}\otimes A)\cdot \mathrm{vec}(B) \tag{3.49}$$

for matrices $\{A, B, C\}$ of compatible sizes. A similar property holds for the $\mathrm{bvec}(\cdot)$ operation:

$$\mathrm{bvec}(ABC) = (C^{\mathsf{T}}\otimes_b A)\cdot \mathrm{bvec}(B) \tag{3.50}$$

for block matrices $\{A, B, C\}$ with appropriate block sizes. In Fig. 3.2, we compare the structures of $A\otimes B$ and $A\otimes_b B$, where $\{A, B\}$ are a pair of block matrices.

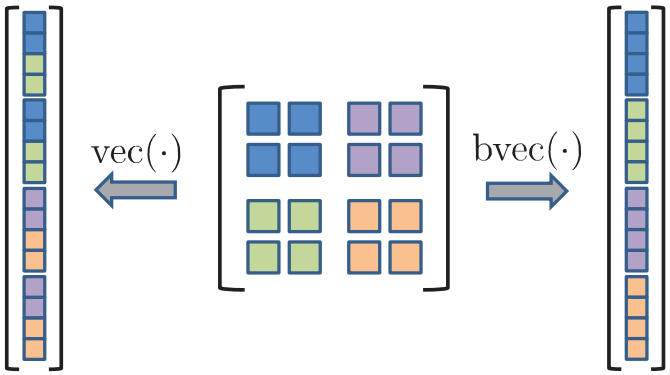


Figure 3.1: Comparing two vectorization operations: vec(·) versus bvec(·). The operation vec(·) destroys the locality of the blocks in the original matrix argument while the operation bvec(·) preserves it.

The observation is that the operation $\otimes$ destroys the locality of the blocks from matrix B, whereas the operation $\otimes_b$ preserves the locality of the blocks from both matrices A and B.

Using properties of the block operations bvec(·) and $\otimes_b$, we can derive from Theorem 3.2 a useful relation between the blocks of the network error covariance matrix, $\mathbb{E}\widetilde{\boldsymbol{w}}'_i\widetilde{\boldsymbol{w}}'^*_i$, and the blocks of the vector z_i. Let us partition the $4M^2N^2$-dimensional vector z_i as

$$z_i = \mathrm{col}\{z_i^{(1)}, \ldots, z_i^{(N)}\}, z_i^{(\ell)} \triangleq \mathrm{col}\{z_i^{(\ell,1)}, \ldots, z_i^{(\ell,N)}\} \tag{3.51}$$

where $z_i^{(\ell)}$ is the ℓ-th sub-vector of z_i with dimension $4M^2N$ and $z_i^{(\ell,k)}$ is the k-th block of $z_i^{(\ell)}$ with dimension $4M^2$. From (3.44) and (3.120), we find that these vectors have the following useful interpretations for $k, \ell = 1, 2, \ldots, N$:

$$z_i = \mathbb{E}[\mathrm{bvec}(\widetilde{\boldsymbol{w}}'_i\widetilde{\boldsymbol{w}}'^*_i)] = \mathrm{col}\{\mathbb{E}\,[(\widetilde{\boldsymbol{w}}'^*_{\ell,i})^{\mathsf{T}} \otimes \widetilde{\boldsymbol{w}}'_{k,i}]\}_{\ell,k=1}^{N} \tag{3.52}$$

$$z_i^{(\ell,k)} \triangleq \mathrm{vec}(\mathbb{E}[\widetilde{\boldsymbol{w}}'_{k,i}\widetilde{\boldsymbol{w}}'^*_{\ell,i}]) = \mathbb{E}[(\widetilde{\boldsymbol{w}}'^*_{\ell,i})^{\mathsf{T}} \otimes \widetilde{\boldsymbol{w}}'_{k,i}] \tag{3.53}$$

where $\mathbb{E}\widetilde{\boldsymbol{w}}'_{k,i}\widetilde{\boldsymbol{w}}'^*_{\ell,i}$ is the (k,ℓ)-th block of $\mathbb{E}\widetilde{\boldsymbol{w}}'_i\widetilde{\boldsymbol{w}}'^*_i$ with size $2M \times 2M$. The block entries of the vector z_i in (3.53) do not only allow us to recover the covariance

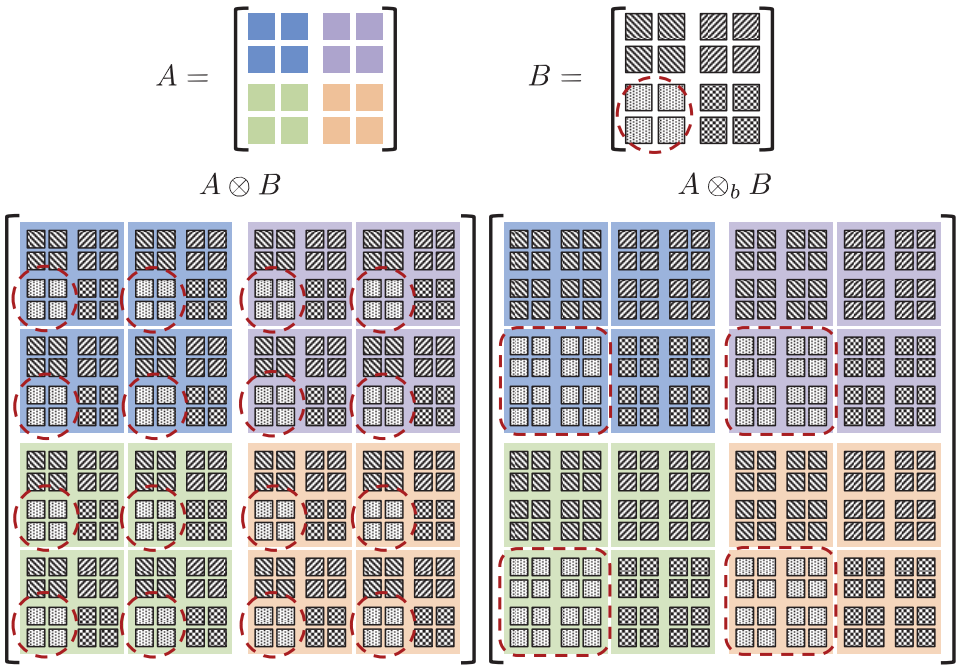


Figure 3.2: Comparing two Kronecker product operations: $\otimes$ versus $\otimes_b$. The operation $\otimes$ destroys the locality of the blocks from matrix B while the operation $\otimes_b$ preserves the locality of the blocks from both matrices A and B.

matrices of any individual error vectors, $\mathbb{E}\widetilde{\boldsymbol{w}}'_{k,i}\widetilde{\boldsymbol{w}}'^{*}_{k,i}$, but they also allow us to recover the *cross-covariance* matrices, $\mathbb{E}\widetilde{\boldsymbol{w}}'_{k,i}\widetilde{\boldsymbol{w}}'^{*}_{\ell,i}$, for *any* pair of agents $\{k,\ell\}$. Therefore, by studying the evolution of the entire covariance vector in (3.47), we are able to extract some detailed information about the dynamics of the asynchronous diffusion network, as we shall show in Theorem 3.4 and Corollary 3.3 in Section 3.3.

3.2 Steady-State Performance

When $i \to \infty$, and by the fact that $\mathcal{F}$ is stable, we obtain from (3.47) that

$$z_\infty \triangleq \lim_{i\to\infty} z_i$$

$$
\begin{aligned}
&= (I_{4M^2N^2} - \mathcal{F}^*)^{-1} \lim_{i\to\infty} y_i \\
&= (I_{4M^2N^2} - \mathcal{F}^*)^{-1}(\bar{\mathcal{A}}^{\mathsf{T}} \otimes_b \bar{\mathcal{A}}^{\mathsf{T}} + \mathcal{C}_A^{\mathsf{T}})(\bar{\mathcal{M}} \otimes_b \bar{\mathcal{M}} + \mathcal{C}_M) \lim_{i\to\infty} \mathrm{bvec}(\mathbb{E}\mathcal{R}_i(\boldsymbol{w}_{i-1}))
\end{aligned} \tag{3.54}
$$

where we also used (3.42) and (3.43). Now note that

$$
\begin{aligned}
\|\mathcal{R}_i(\mathbb{1}_N \otimes w^o) - \mathbb{E}\mathcal{R}_i(\boldsymbol{w}_{i-1})\| &\overset{(a)}{\leq} \mathbb{E}\|\mathcal{R}_i(\mathbb{1}_N \otimes w^o) - \mathcal{R}_i(\boldsymbol{w}_{i-1})\| \\
&\overset{(b)}{\leq} \kappa_v \cdot \mathbb{E}\,\|\mathbb{1}_N \otimes w^o - \boldsymbol{w}_{i-1}\|^{\gamma_v} \\
&= \kappa_v \cdot \mathbb{E}\,[\|\widetilde{\boldsymbol{w}}_{i-1}\|^4]^{\gamma_v/4} \\
&\overset{(c)}{\leq} \kappa_v \cdot [\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4]^{\gamma_v/4}
\end{aligned} \tag{3.55}
$$

where step (a) is by Jensen's inequality; step (b) is by (3.4) in Assumption 3.1; and step (c) is by Jensen's inequality and the fact that $|\cdot|^{\gamma_v/4}$ is concave due to $0 < \gamma_v/4 \leq 1$. Now, from Theorem 2.2 in Chapter 2, we know that $\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 \leq O(\nu^2)$ under Assumption 3.2. Therefore, we obtain from (3.55) that

$$
\limsup_{i\to\infty} \|\mathcal{R}_i(\mathbb{1}_N \otimes w^o) - \mathbb{E}\mathcal{R}_i(\boldsymbol{w}_{i-1})\| \leq O(\nu^{\gamma_v/2}) \tag{3.56}
$$

According to (3.56), we can replace $\mathbb{E}\mathcal{R}_i(\boldsymbol{w}_{i-1})$ in (3.54) by $\mathcal{R}_i(\mathbb{1}_N \otimes w^o)$ with an error in the order of $\nu^{\gamma_v/2}$. Let

$$
z \triangleq (I_{4M^2N^2} - \mathcal{F}^*)^{-1}(\bar{\mathcal{A}}^{\mathsf{T}} \otimes_b \bar{\mathcal{A}}^{\mathsf{T}} + \mathcal{C}_A^{\mathsf{T}})(\bar{\mathcal{M}} \otimes_b \bar{\mathcal{M}} + \mathcal{C}_M)\mathrm{bvec}(\mathcal{R}) \tag{3.57}
$$

From (2.190) in Chapter 2, we know that the second-order moments of $\{\boldsymbol{\mu}_k(i)\}$ are in the order of ν^2. Hence, by (2.76) from Chapter 2, (3.26), and (3.32), it is easy to verify that

$$
\|\bar{\mathcal{M}} \otimes_b \bar{\mathcal{M}} + \mathcal{C}_M\| = O(\nu^2) \tag{3.58}
$$

Using (3.57), (3.58), and the fact that $\|(I_{4M^2N^2} - \mathcal{F}^*)^{-1}\| = O(\nu^{-1})$ from Lemma 3.5 further ahead, we conclude that

$$
\|z\| = O(\nu) \tag{3.59}
$$

Then, by using (3.5) and (3.56)–(3.59), we obtain from (3.54) that

$$z_\infty = z + O(\nu^{1+\gamma_v/2}), \qquad \|z_\infty\| = O(\nu) \tag{3.60}$$

Define the steady-state average network MSD by

$$\mathrm{MSD}^{\mathrm{net}} \triangleq \lim_{i\to\infty} \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 \tag{3.61}$$

and the steady-state individual MSD for agent k by

$$\mathrm{MSD}_k \triangleq \lim_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 \tag{3.62}$$

Theorem 3.3 (Steady-state MSD)**.** *It holds that*

$$MSD^{net} = \frac{1}{2N} z^* \mathrm{bvec}(I_{2MN}) + O(\nu^{1+\gamma_o}) \tag{3.63}$$

$$MSD_k = \frac{1}{2} z^* \mathrm{bvec}(E_{kk}\otimes I_{2M}) + O(\nu^{1+\gamma_o}) \tag{3.64}$$

where z is given by (3.57)*,*

$$\gamma_o \triangleq \frac{1}{2}\min\{1,\gamma_v\} \tag{3.65}$$

and E_{kk} is the $N\times N$ basis matrix that only has one non-zero element, which is equal to 1*, at the (k,k)-th entry.*

Proof. From (3.48) by selecting $\Sigma = I_{2MN}$, and also using (3.54) and (3.60), we get

$$\begin{aligned}\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i'\|^2 &= z_\infty^* \mathrm{bvec}(I_{2MN}) \\ &= z^*\mathrm{bvec}(I_{2MN}) + O(\nu^{1+\gamma_v/2}) \\ &= O(\nu)\end{aligned} \tag{3.66}$$

Likewise, by selecting $\Sigma = E_{kk}\otimes I_{2M}$, we get

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i'\|^2_{E_{kk}\otimes I_{2M}} = z_\infty^* \mathrm{bvec}(E_{kk}\otimes I_{2M})$$

$$= z^{*}\text{bvec}(E_{kk} \otimes I_{2M}) + O(\nu^{1+\gamma_v/2})$$
$$= O(\nu) \tag{3.67}$$

Note further that

$$\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_i\|^2 = \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}'_i\|^2 + \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_i - \widetilde{\underline{\boldsymbol{w}}}'_i\|^2 + 2\mathfrak{Re}\mathbb{E}[\widetilde{\underline{\boldsymbol{w}}}'^{*}_i(\widetilde{\underline{\boldsymbol{w}}}_i - \widetilde{\underline{\boldsymbol{w}}}'_i)] \tag{3.68}$$

Using the Cauchy-Schwartz inequality, it can be verified that

$$|\mathfrak{Re}\mathbb{E}[\widetilde{\underline{\boldsymbol{w}}}'^{*}_i(\widetilde{\underline{\boldsymbol{w}}}_i - \widetilde{\underline{\boldsymbol{w}}}'_i)]| \leq \sqrt{\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}'_i\|^2 \cdot \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_i - \widetilde{\underline{\boldsymbol{w}}}'_i\|^2} \tag{3.69}$$

From Theorem 3.1, we have

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_i - \widetilde{\underline{\boldsymbol{w}}}'_i\|^2 \leq O(\nu^2) \tag{3.70}$$

Substituting (3.66) and (3.70) into (3.68), and using (3.69), we get

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_i\|^2 = \lim_{i\to\infty} \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}'_i\|^2 + O(\nu^2) + 2\sqrt{O(\nu)\cdot O(\nu^2)}$$
$$= \lim_{i\to\infty} \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}'_i\|^2 + O(\nu^{3/2}) \tag{3.71}$$

Results (3.63) and (3.64) follow from (3.66), (3.67), and (3.71). □

Result (3.63) generalizes its counterpart (276) from [35] for the synchronous diffusion strategy. Since expressions (3.66) and (3.67) are both related to the vector z in (3.57), let us examine z more closely to reveal the implications of asynchronous adaptation and learning on performance. Theorem 3.4 in the following section will lead to powerful alternative expressions for (3.66) and (3.67). The new expressions will highlight some important properties about the behavior of the asynchronous network in steady-state, such as the behavior that was illustrated earlier in Fig. 1.3 from Chapter 1. The subsequent analysis relies on a useful low-rank factorization result.

3.3 Low-Rank Factorization

From (3.54) we see that the structure of z depends on the structure of the matrix $(I_{4M^2N^2} - \mathcal{F}^*)^{-1}$. In the following, we show that by retaining the dominant eigenspace of $(I_{4M^2N^2} - \mathcal{F}^*)^{-1}$, we can obtain a more revealing MSD expression than (3.63) that is still accurate to the order of $O(\nu^{1+\gamma_o})$.

3.3.1 Perron Eigenvectors

To proceed, we introduce the following condition on the matrix $\bar{A} \otimes \bar{A} + C_A$.

Assumption 3.3 (Primitiveness of $\bar{A} \otimes \bar{A} + C_A$)**.** *The matrix $\bar{A} \otimes \bar{A} + C_A$ is assumed to be primitive [79, p. 45], namely, that there exists a finite positive integer j such that all entries of $(\bar{A} \otimes \bar{A} + C_A)^j$ are positive.* □

Lemma 3.2 (Primitiveness of $\bar{A}$)**.** *The matrix $\bar{A}$ is primitive if $\bar{A} \otimes \bar{A} + C_A$ is primitive.*

Proof. See Appendix 3.F. □

Assumption 3.3 is guaranteed if the directed graph (digraph) associated with the matrix $\bar{A} \otimes \bar{A} + C_A$ is strongly-connected with as least one self-loop [79, pp. 30,34]. The digraph associated with $\bar{A} \otimes \bar{A} + C_A$ is the union of all possible digraphs associated with the realizations of $\boldsymbol{A}_i \otimes \boldsymbol{A}_i$ [80, p. 29]. Each possible digraph associated with $\boldsymbol{A}_i \otimes \boldsymbol{A}_i$ is a Kronecker graph of order 2 generated by the initiator $\boldsymbol{A}_i$ [81]. Therefore, Assumption 3.3 amounts to an assumption that the *union* of all possible digraphs associated with the realizations of $\boldsymbol{A}_i \otimes \boldsymbol{A}_i$ is strongly-connected with at least one self-loop. As illustrated in Fig. 3.3, this condition still allows the digraphs associated with $\boldsymbol{A}_i$ to be weakly-connected with or without self-loops or even to be disconnected. Important cases such as

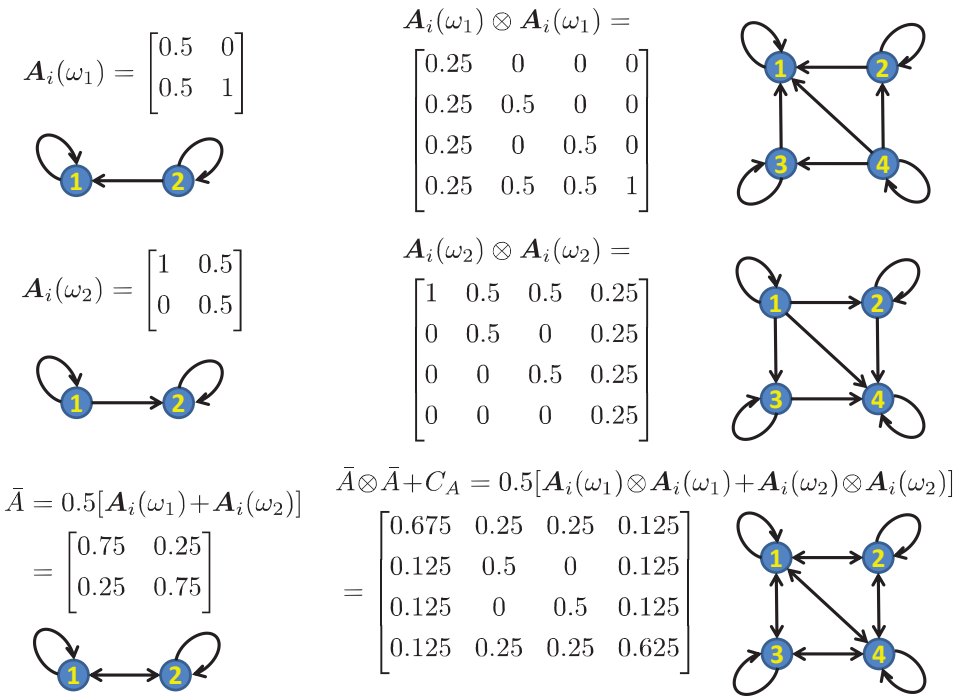


Figure 3.3: An illustration of the digraph associated with $\mathbb{E}\left(\boldsymbol{A}_i \otimes \boldsymbol{A}_i | \boldsymbol{w}_{i-1}\right) = \bar{A} \otimes \bar{A} + C_A$, where $\boldsymbol{A}_i$ has two equally probable realizations $\boldsymbol{A}_i(\omega_1)$ and $\boldsymbol{A}_i(\omega_2)$. It can be observed that *neither* of the digraphs associated with $\boldsymbol{A}_i(\omega_j) \otimes \boldsymbol{A}_i(\omega_j)$, $j = 1, 2$, is strongly-connected due to the existence of the source and sink nodes, where information can only flow in *one* direction through the network. However, the digraph associated with $\mathbb{E}\left(\boldsymbol{A}_i \otimes \boldsymbol{A}_i | \boldsymbol{w}_{i-1}\right)$, which is the union of the first two digraphs, is strongly-connected, where information can flow in *any* direction through the network.

random gossip [2, 23, 38, 39, 42] or probabilistic diffusion [43, 44] are therefore not ruled out by this condition. It can be verified that the converse of Lemma 3.2 is generally not true: when the digraph associated with $\bar{A}$ is primitive, the digraph associated with $\bar{A} \otimes \bar{A} + C_A$ does not even need to be connected.

By Lemma 2.3 from Chapter 2 and the above Assumption 3.3, the matrix $\bar{A} \otimes \bar{A} + C_A$ is left-stochastic and primitive. It follows from the Perron-Frobenius theorem [79] [82] that this matrix has a unique eigenvalue at one and a pair of eigenvectors $\{\mathbb{1}_{N^2}, p\}$ with *positive* entries satisfying:

$$(\bar{A} \otimes \bar{A} + C_A) \cdot p = p, \qquad p^{\mathsf{T}} \cdot \mathbb{1}_{N^2} = 1 \tag{3.72}$$

Likewise, the matrix $\bar{A}$ is also left-stochastic and primitive. It has a unique eigenvalue at one and a pair of eigenvectors $\{\mathbb{1}_N, \bar{p}\}$ with positive entries satisfying:

$$\bar{A} \cdot \bar{p} = \bar{p}, \qquad \bar{p}^{\mathsf{T}} \cdot \mathbb{1}_N = 1 \tag{3.73}$$

All other eigenvalues of $\bar{A} \otimes \bar{A} + C_A$ and $\bar{A}$ are inside the unit circle. To simplify the presentation, we shall use the name "Perron eigenvector" to refer to the unique eigenvectors p and $\bar{p}$ in the sequel. Since the vector p is of dimension $N^2 \times 1$, we partition it into N sub-vectors of dimension $N \times 1$ each:

$$p = \mathrm{col}\{p_1, p_2, \ldots, p_N\} \tag{3.74}$$

where p_k denotes the k-th sub-vector. We further define an $N \times N$ matrix P_p whose columns are the sub-vectors $\{p_k\}$:

$$P_p \triangleq \mathrm{unvec}(p) = \begin{bmatrix} p_1 & p_2 & \ldots & p_N \end{bmatrix} \tag{3.75}$$

We use $p_{\ell,k}$ to denote the (ℓ, k)-th element of matrix P_p, which is equal to the ℓ-th element of p_k.

Lemma 3.3 (Properties of P_p)**.** *The matrix P_p in* (3.75) *is symmetric positive semi-definite and it satisfies $P_p \mathbb{1}_N = \bar{p}$, where the $\bar{p}$ is the Perron eigenvector in* (3.73)*.*

Proof. See Appendix 3.G. □

From Lemma 3.3, we get the following useful relations:

$$p_{\ell,k} = p_{k,\ell}, \quad \sum_{k=1}^{N} p_{\ell,k} = \bar{p}_\ell, \quad \sum_{\ell=1}^{N} p_{\ell,k} = \bar{p}_k \tag{3.76}$$

3.3.2 Low-Rank Approximation

We return to our earlier objective of seeking a low-rank factorization for the matrix $(I_{4M^2N^2} - \mathcal{F}^*)^{-1}$. For this purpose, we first introduce the $4M^2 \times 4M^2$ Hermitian matrix:

$$F \triangleq \sum_{k=1}^{N} \sum_{\ell=1}^{N} p_{\ell,k} [\bar{D}_\ell^{\mathsf{T}} \otimes \bar{D}_k + c_{\mu,\ell,k} (H_\ell^{\mathsf{T}} \otimes H_k)] \tag{3.77}$$

where $\bar{D}_k$ is given by (3.27).

Lemma 3.4 (Spectral radius of F)**.** *The matrix F in* (3.77) *is stable if condition* (3.18) *is satisfied. Moreover,*

$$\rho(F) = [1 - \lambda_{\min}(H)]^2 + O(\nu^2) = 1 - O(\nu) \tag{3.78}$$

where

$$H \triangleq \sum_{k=1}^{N} \bar{p}_k \bar{\mu}_k H_k \tag{3.79}$$

It can be verified that $\|H\| = O(\nu)$ and the $O(\nu^2)$ term in (3.78) *is negligible by Assumption 3.2.*

Proof. See Appendix 3.H. □

Lemma 3.5 (Low-rank approximation)**.** *Under Assumptions 3.2 and 3.3, it holds that*

$$(I_{4N^2M^2} - \mathcal{F})^{-1} = (p\mathbb{1}_{N^2}^{\mathsf{T}}) \otimes (I_{4M^2} - F)^{-1} + O(1) \tag{3.80}$$

$$(p\mathbb{1}_{N^2}^{\mathsf{T}}) \otimes (I_{4M^2} - F)^{-1} = O(\nu^{-1}) \tag{3.81}$$

Under Assumption 3.2 where $\nu \ll 1$, the term in (3.81) *dominates the $O(1)$ term in* (3.80)*. Moreover,*

$$\rho(\mathcal{F}) = \rho(F) + O(\nu^{1+1/N^2}) \tag{3.82}$$

where the $\rho(F)$ from (3.78) *dominates the $O(\nu)$ term in* (3.82)*.*

Proof. See Appendix 3.I. □

In expression (3.77) we observe that the matrix F is dependent on the first and second-order moments of the random step-sizes, i.e., $\{\bar{\mu}_k\}$ and $\{c_{\mu,\ell,k}\}$, and is also dependent on the first and second order moments of the random combination coefficient matrix, i.e., $\bar{A}$ and C_A, through the dependence on the Perron eigenvector p. Let us introduce two $4M^2 \times 4M^2$ matrices:

$$R \triangleq \sum_{k=1}^{N} p_{k,k} \, (\bar{\mu}_k^2 + c_{\mu,k,k}) R_k \tag{3.83}$$

$$Z \triangleq \text{unvec}\left((I_{4M^2} - F)^{-1}\text{vec}(R)\right) \tag{3.84}$$

where R_k is given by (3.6). Using (3.181)–(3.183) and (3.209) in Appendix 3.I, we can verify that

$$\|R\| = O(\nu^2), \qquad \|Z\| = O(\nu) \tag{3.85}$$

Then, using Lemma 3.5, we can establish the following useful result about the structure of the steady-state network error covariance matrix.

Theorem 3.4 (Network error covariance matrix)**.** *In steady-state, the covariance matrix of the network error $\widetilde{\boldsymbol{w}}_i'$ from the long-term model* (3.19) *can be approximated by*

$$\lim_{i\to\infty} \mathbb{E}\widetilde{\underline{\boldsymbol{w}}}_i'\widetilde{\underline{\boldsymbol{w}}}_i'^{*} = (\mathbb{1}_N\mathbb{1}_N^{\mathsf{T}}) \otimes Z + O(\nu^{1+\gamma_o'}) \tag{3.86}$$

where Z is from (3.84)*,*

$$\gamma_o' \triangleq \frac{1}{2}\min\{2, \gamma_v\} \tag{3.87}$$

and the first term on the RHS is dominant.

Proof. See Appendix 3.J. □

According to Theorem 3.4, the (cross-) covariance matrices of $\{\widetilde{\boldsymbol{w}}_{k,i}'\}$, which are uniformly expressed by $\mathbb{E}\widetilde{\underline{\boldsymbol{w}}}_{k,i}'\widetilde{\underline{\boldsymbol{w}}}_{\ell,i}'^{*} = \mathrm{unvec}(z_i^{(\ell,k)})$ for all k and ℓ according to (3.53), can be approximated by Z in steady-state. However, this result is useful only if Z is a valid complex-Hessian-type matrix.

Definition 3.1 (Complex-Hessian-type matrices)**.** *Let X be an $M\times M$ positive semi-definite Hermitian matrix and let Y be an $M\times M$ symmetric matrix. Then, a positive semi-definite block matrix of the form*

$$H \triangleq \begin{bmatrix} X & Y \\ Y^{*} & X^{\mathsf{T}} \end{bmatrix} \geq 0 \tag{3.88}$$

will be referred to as a complex-Hessian-type matrix. □

The following result explains the reason for introducing this definition.

Lemma 3.6 (Complex-Hessian-type covariance matrices)**.** *Let $\boldsymbol{x}$ denote an $M\times 1$ zero-mean complex random vector and let $R_x \triangleq \mathbb{E}\boldsymbol{x}\boldsymbol{x}^{*}$ and $R_x' \triangleq \mathbb{E}\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}$. Then, the covariance matrix of $\underline{\boldsymbol{x}} = \mathbb{T}(\boldsymbol{x})$ is given by*

$$\mathbb{E}\underline{\boldsymbol{x}}\underline{\boldsymbol{x}}^{*} = \begin{bmatrix} \mathbb{E}\boldsymbol{x}\boldsymbol{x}^{*} & \mathbb{E}\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \\ \mathbb{E}(\boldsymbol{x}^{*})^{\mathsf{T}}\boldsymbol{x}^{*} & \mathbb{E}(\boldsymbol{x}^{*})^{\mathsf{T}}\boldsymbol{x}^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} R_x & R_x' \\ R_x'^{*} & R_x^{\mathsf{T}} \end{bmatrix} \tag{3.89}$$

and this matrix is a complex-Hessian-type matrix.

Proof. It follows from comparing (3.89) to (3.88). □

By Lemma 3.6, for *any* zero-mean complex random vector $\boldsymbol{x}$, the covariance matrix of $\underline{\boldsymbol{x}} = \mathbb{T}(\boldsymbol{x})$ must be a complex-Hessian-type matrix. Therefore, in order to approximate $\{\text{unvec}(z_i^{(\ell,k)})\}$ by Z according to (3.86), we establish the following useful result for the matrix Z.

Lemma 3.7 (Properties of Z)**.** *The matrix Z in* (3.84) *is a positive semi-definite complex-Hessian-type matrix.*

Proof. See Appendix 3.K. □

Using (3.86) and Lemmas 3.6 and 3.7, we arrive at the following result for the covariance and cross-covariance matrices of the steady-state error vectors $\{\widetilde{\boldsymbol{w}}'_{k,i}\}$ from the long-term model (3.19).

Corollary 3.1 (Covariance and cross-covariance matrices)**.** *The steady-state covariance and cross-covariance matrices of individual errors $\{\widetilde{\boldsymbol{w}}'_{k,i}\}$ from the long-term model* (3.19) *can be approximated by*

$$\lim_{i\to\infty} \mathbb{E}\widetilde{\boldsymbol{w}}'_{k,i}\widetilde{\boldsymbol{w}}'^{*}_{\ell,i} = Z + O(\nu^{1+\gamma'_o}) \tag{3.90}$$

for all k and ℓ, where Z is given by (3.84) *and is dominant due to* (3.85)*, and γ'_o is given by* (3.87)*.*

Proof. By Lemma 3.7, the Z is a complex-Hessian-type matrix. According to Lemma 3.6, it is a valid covariance matrix for complex random vectors obtained via the transform $\mathbb{T}(\cdot)$. The approximation (3.90) then follows from Theorem 3.4. □

3.3.3 Steady-State MSD

Using Corollary 3.1, we obtain two useful results about the steady-state MSD for asynchronous diffusion solutions.

Corollary 3.2 (Steady-state MSD)**.** *Based on the same assumptions as Theorem 3.4, the steady-state MSD (either the network MSD in* (3.61) *or the individual MSD in* (3.62)*) can be approximated by*

$$MSD^{net} = \frac{1}{4}\mathrm{Tr}(H^{-1}R) + O(\nu^{1+\gamma_o}) \tag{3.91}$$

$$MSD_k = \frac{1}{4}\mathrm{Tr}(H^{-1}R) + O(\nu^{1+\gamma_o}) \tag{3.92}$$

where $\{H, R\}$ *are given by* (3.79) *and* (3.83)*, respectively,* γ_o *is from* (3.65)*, and* $\mathrm{Tr}(H^{-1}R)$ *is of the order of* ν*.*

Proof. See Appendix 3.L. □

Corollary 3.3 (Clustered solutions)**.** *The steady-state* relative *MSD between any two agents* k *and* ℓ*, i.e.,* $\mathbb{E}\,\|\boldsymbol{w}_{k,i} - \boldsymbol{w}_{\ell,i}\|^2$*, is negligible compared to their steady-state* absolute *MSD with respect to* w^o*, i.e.,*

$$\lim_{i\to\infty} \mathbb{E}\|\boldsymbol{w}_{k,i} - \boldsymbol{w}_{\ell,i}\|^2 = O(\nu^{1+\gamma_o'}) \ll \max_k MSD_k = O(\nu) \tag{3.93}$$

where γ_o' *is given by* (3.87)*.*

Proof. First, from Corollary 3.1, we have

$$\begin{aligned}
\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}'_{k,i} - \widetilde{\boldsymbol{w}}'_{\ell,i}\|^2 &= \lim_{i\to\infty} \mathbb{E}[\|\widetilde{\boldsymbol{w}}'_{k,i}\|^2 + \|\widetilde{\boldsymbol{w}}'_{\ell,i}\|^2 - \widetilde{\boldsymbol{w}}'^{*}_{k,i}\widetilde{\boldsymbol{w}}'_{\ell,i} - \widetilde{\boldsymbol{w}}'^{*}_{\ell,i}\widetilde{\boldsymbol{w}}'_{k,i}] \\
&= \mathrm{Tr}(Z) + \mathrm{Tr}(Z) - \mathrm{Tr}(Z) - \mathrm{Tr}(Z) + O(\nu^{1+\gamma_o'}) \\
&= O(\nu^{1+\gamma_o'})
\end{aligned} \tag{3.94}$$

From Theorem 3.1 and using (3.94), we get

$$\begin{aligned}\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}-\widetilde{\boldsymbol{w}}_{\ell,i}\|^2 &= \lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}-\widetilde{\boldsymbol{w}}'_{k,i}+\widetilde{\boldsymbol{w}}'_{k,i}-\widetilde{\boldsymbol{w}}'_{\ell,i}+\widetilde{\boldsymbol{w}}'_{\ell,i}-\widetilde{\boldsymbol{w}}_{\ell,i}\|^2 \\ &\leq \lim_{i\to\infty}3\mathbb{E}\left(\|\widetilde{\boldsymbol{w}}_{k,i}-\widetilde{\boldsymbol{w}}'_{k,i}\|^2+\|\widetilde{\boldsymbol{w}}'_{k,i}-\widetilde{\boldsymbol{w}}'_{\ell,i}\|^2+\|\widetilde{\boldsymbol{w}}'_{\ell,i}-\widetilde{\boldsymbol{w}}_{\ell,i}\|^2\right) \\ &\leq O(\nu^2)+O(\nu^{1+\gamma'_o})+O(\nu^2) \\ &\leq O(\nu^{1+\gamma'_o}) \end{aligned} \tag{3.95}$$

Using (2.5) from Chapter 2, (3.95), and Corollary 3.2 completes the proof. □

We illustrated Corollaries 3.2 and 3.3 earlier in Fig. 1.3.

3.4 Conclusion

We studied in some detail the MSE performance of *asynchronous* networks with random step-sizes, links, topologies, and combination coefficients. Assuming sufficiently small step-sizes, we showed that at steady-state, the error vector for every individual agent tends to cluster within $O(\nu^{1+\gamma_o})$ from each other, which means that the MSD performance is essentially uniform across the entire network. The result in Corollary 3.2 shows explicitly how the MSD performance of the network is affected by the asynchronous behavior. Quantities that relate to the first and second-order moments of the distribution of the random step-sizes and combination coefficients appear in these expressions. These results can be used to guide strategies for adjusting the combination weights and the rate at which the agents update their solutions and to ensure that the performance (in terms of MSD and rate of convergence) does not degrade below desirable levels.

3.A Proof of Lemma 3.1

Using Lemma 2.1 in Chapter 2, we get from (3.9)–(3.10) that

$$\begin{aligned}\|H_k - \boldsymbol{H}_{k,i-1}\| &= \left\| \int_0^1 [\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o) - \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o - t\widetilde{\underline{w}}_{k,i-1})]dt \right\| \\ &\leq \int_0^1 \|\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o) - \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o - t\widetilde{\underline{w}}_{k,i-1})\| dt \\ &\leq \int_0^1 \tau'_k \cdot t \cdot \|\widetilde{\underline{w}}_{k,i-1}\| \, dt = \frac{\tau'_k}{2} \cdot \|\widetilde{\underline{w}}_{k,i-1}\| \end{aligned} \tag{3.96}$$

Using (3.96), we get from (3.13) that

$$\begin{aligned}\|\boldsymbol{d}_{k,i}\|^2 &\leq [\boldsymbol{\mu}_k(i)]^2 \cdot \|H_k - \boldsymbol{H}_{k,i-1}\|^2 \cdot \|\widetilde{\underline{w}}_{k,i-1}\|^2 \\ &\leq [\boldsymbol{\mu}_k(i)]^2 \cdot \frac{\tau'^2_k}{4} \cdot \|\widetilde{\underline{w}}_{k,i-1}\|^4 \end{aligned} \tag{3.97}$$

Taking the expectation of both sides of (3.97) yields

$$\mathbb{E}\|\boldsymbol{d}_{k,i}\|^2 \leq \bar{\mu}_k^{(2)} \frac{\tau'^2_k}{4} \cdot \mathbb{E}\|\widetilde{\underline{w}}_{k,i-1}\|^4 \tag{3.98}$$

From Theorem 2.2 of Chapter 2, it holds for large enough i that

$$\mathbb{E}\|\widetilde{\underline{w}}_{k,i}\|^4 \leq 2b_4^2 \cdot \nu^2 \tag{3.99}$$

Using the fact from (2.190) of Chapter 2 that $\bar{\mu}_k^{(2)} \leq \nu^2$ for any k, and letting

$$\tau' \triangleq \max_k \tau'_k \tag{3.100}$$

we obtain from (3.98) and (3.99) that

$$\mathbb{E}\|\boldsymbol{d}_{k,i}\|^2 \leq 2\tau'^2 b_4^2 \cdot \nu^4 = O(\nu^4) \tag{3.101}$$

where a factor of 4 appeared due to the conversion $\mathbb{T}(\cdot)$ from (2.3) of Chapter 2. Then,

$$\mathbb{E}\|\boldsymbol{\mathcal{A}}_i^{\mathsf{T}} \boldsymbol{d}_i\|^2 = \sum_{k=1}^N \mathbb{E} \left\| \sum_{\ell \in \boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i) \boldsymbol{d}_{\ell,i} \right\|^2$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \sum_{k=1}^{N} \mathbb{E}\left[\sum_{\ell\in\boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\|\boldsymbol{d}_{\ell,i}\|^2\right] \\
&= \sum_{k=1}^{N}\sum_{\ell\in\mathcal{N}_k} \bar{a}_{\ell k}\mathbb{E}\|\boldsymbol{d}_{\ell,i}\|^2 \\
&\leq N\cdot \max_{\ell}\mathbb{E}\|\boldsymbol{d}_{\ell,i}\|^2 \\
&\overset{(b)}{\leq} 2N\tau'^2 b_4^2\cdot\nu^4
\end{aligned} \tag{3.102}
$$

where step (a) is by using Jensen's inequality; and step (b) is by using (3.101).

3.B Proof of Theorem 3.1

We rewrite the original error recursion (3.11) and the long-term model (3.19) respectively as follows:

$$
\widetilde{\underline{\boldsymbol{\psi}}}_{k,i} = [I_{2M} - \boldsymbol{\mu}_k(i)H_k]\widetilde{\underline{\boldsymbol{w}}}_{k,i-1} + \boldsymbol{\mu}_k(i)\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1}) + \boldsymbol{d}_{k,i} \tag{3.103}
$$

$$
\widetilde{\underline{\boldsymbol{w}}}_{k,i} = \sum_{\ell\in\boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\,\widetilde{\underline{\boldsymbol{\psi}}}_{\ell,i} \tag{3.104}
$$

and

$$
\widetilde{\underline{\boldsymbol{\psi}}}'_{k,i} = [I_{2M} - \boldsymbol{\mu}_k(i)H_k]\widetilde{\underline{\boldsymbol{w}}}'_{k,i-1} + \boldsymbol{\mu}_k(i)\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1}) \tag{3.105}
$$

$$
\widetilde{\underline{\boldsymbol{w}}}'_{k,i} = \sum_{\ell\in\boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\,\widetilde{\underline{\boldsymbol{\psi}}}'_{\ell,i} \tag{3.106}
$$

with the prime notation for quantities associated with the long-term model (3.19). From (3.104) and (3.106), and using Jensen's inequality, the squared 2-norm of the difference between the two models is given by

$$
\|\widetilde{\underline{\boldsymbol{w}}}_{k,i} - \widetilde{\underline{\boldsymbol{w}}}'_{k,i}\|^2 \leq \sum_{\ell\in\boldsymbol{\mathcal{N}}_{k,i}} \boldsymbol{a}_{\ell k}(i)\,\|\widetilde{\underline{\boldsymbol{\psi}}}_{\ell,i} - \widetilde{\underline{\boldsymbol{\psi}}}'_{\ell,i}\|^2 \tag{3.107}
$$

Taking the expectation of both sides yields

$$
\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i} - \widetilde{\underline{\boldsymbol{w}}}'_{k,i}\|^2 \leq \sum_{\ell\in\mathcal{N}_k} \bar{a}_{\ell k}\,\mathbb{E}\|\widetilde{\underline{\boldsymbol{\psi}}}_{\ell,i} - \widetilde{\underline{\boldsymbol{\psi}}}'_{\ell,i}\|^2
$$

$$\leq \max_{\ell} \mathbb{E}\|\widetilde{\underline{\psi}}_{\ell,i} - \widetilde{\underline{\psi}}'_{\ell,i}\|^2 \tag{3.108}$$

for all k. Then,

$$\max_{k} \mathbb{E}\|\widetilde{\underline{w}}_{k,i} - \widetilde{\underline{w}}'_{k,i}\|^2 \leq \max_{k} \mathbb{E}\|\widetilde{\underline{\psi}}_{k,i} - \widetilde{\underline{\psi}}'_{k,i}\|^2 \tag{3.109}$$

From (3.103) and (3.105), we have

$$\widetilde{\underline{\psi}}_{k,i} - \widetilde{\underline{\psi}}'_{k,i} = [I_{2M} - \boldsymbol{\mu}_k(i)H_k](\widetilde{\underline{w}}_{k,i-1} - \widetilde{\underline{w}}'_{k,i-1}) + \boldsymbol{d}_{k,i} \tag{3.110}$$

Taking the expected squared 2-norm of both sides, we have

$$\begin{aligned}
\mathbb{E}\|\widetilde{\underline{\psi}}_{k,i} - \widetilde{\underline{\psi}}'_{k,i}\|^2 &\leq \mathbb{E}\|[I_{2M} - \boldsymbol{\mu}_k(i)H_k](\widetilde{\underline{w}}_{k,i-1} - \widetilde{\underline{w}}'_{k,i-1}) + \boldsymbol{d}_{k,i}\|^2 \\
&= \mathbb{E}\left\|(1-t)\frac{I_{2M} - \boldsymbol{\mu}_k(i)H_k}{1-t}(\widetilde{\underline{w}}_{k,i-1} - \widetilde{\underline{w}}'_{k,i-1}) + t\frac{\boldsymbol{d}_{k,i}}{t}\right\|^2 \\
&\leq \frac{1}{1-t}\mathbb{E}\|I_{2M} - \boldsymbol{\mu}_k(i)H_k\|^2\mathbb{E}\|\widetilde{\underline{w}}_{k,i-1} - \widetilde{\underline{w}}'_{k,i-1}\|^2 + \frac{1}{t}\mathbb{E}\|\boldsymbol{d}_{k,i}\|^2
\end{aligned} \tag{3.111}$$

for any $0 < t < 1$, where we used Jensen's inequality in the second inequality. By condition (3.14), it can be verified that

$$\begin{aligned}
\mathbb{E}\|I_{2M} - \boldsymbol{\mu}_k(i)H_k\|^2 &\leq 1 - 2\bar{\mu}_k^{(1)}\lambda_{k,\min} + \bar{\mu}_k^{(2)}\lambda_{k,\max}^2 \\
&\overset{(a)}{\leq} 1 - \bar{\mu}_k^{(1)}\lambda_{k,\min} \\
&\leq \left(1 - \frac{1}{2}\bar{\mu}_k^{(1)}\lambda_{k,\min}\right)^2 \\
&< 1
\end{aligned} \tag{3.112}$$

where step (a) is from (2.184) of Chapter 2. Substituting $t = \frac{1}{2}\bar{\mu}_k^{(1)}\lambda_{k,\min} < 1$ and (3.112) into (3.111) yields

$$\mathbb{E}\|\widetilde{\underline{\psi}}_{k,i} - \widetilde{\underline{\psi}}'_{k,i}\|^2 \leq \left(1 - \frac{1}{2}\bar{\mu}_k^{(1)}\lambda_{k,\min}\right)\mathbb{E}\|\widetilde{\underline{w}}_{k,i-1} - \widetilde{\underline{w}}'_{k,i-1}\|^2 + \frac{2}{\lambda_{k,\min}\bar{\mu}_k^{(1)}}\mathbb{E}\|\boldsymbol{d}_{k,i}\|^2 \tag{3.113}$$

Using (3.98), the second term on the RHS of (3.113) can be bounded for large enough i by

$$\begin{aligned}\frac{2}{\lambda_{k,\min}\bar{\mu}_k^{(1)}}\cdot\mathbb{E}\|\boldsymbol{d}_{k,i}\|^2 &\le \frac{2}{\lambda_{k,\min}\bar{\mu}_k^{(1)}}\cdot\bar{\mu}_k^{(2)}\frac{\tau_k'^2}{4}\cdot\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4\\ &= \frac{\tau_k'^2}{2\lambda_{k,\min}}\cdot\frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}}\cdot\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4\\ &\le \frac{\tau_k'^2\nu}{2\lambda_{k,\min}}\cdot\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4 \qquad (3.114)\end{aligned}$$

where we used the fact from (2.192) of Chapter 2 that $\nu \ge \bar{\mu}_k^{(2)}/\bar{\mu}_k^{(1)}$. Substituting (3.114) into (3.113) yields

$$\mathbb{E}\|\widetilde{\underline{\boldsymbol{\psi}}}_{k,i}-\widetilde{\underline{\boldsymbol{\psi}}}'_{k,i}\|^2 \le \left(1-\frac{1}{2}\bar{\mu}_k^{(1)}\lambda_{k,\min}\right)\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}-\widetilde{\underline{\boldsymbol{w}}}'_{k,i-1}\|^2+\frac{\tau_k'^2\nu}{2\lambda_{k,\min}}\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4 \qquad (3.115)$$

Therefore,

$$\begin{aligned}\max_k \mathbb{E}\|\widetilde{\underline{\boldsymbol{\psi}}}_{k,i}-\widetilde{\underline{\boldsymbol{\psi}}}'_{k,i}\|^2 \le{}& \max_k\left(1-\frac{1}{2}\bar{\mu}_k^{(1)}\lambda_{k,\min}\right)\max_k\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}-\widetilde{\underline{\boldsymbol{w}}}'_{k,i-1}\|^2\\ &+\max_k\left[\frac{\tau_k'^2\nu}{2\lambda_{k,\min}}\cdot\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4\right] \qquad (3.116)\end{aligned}$$

Substituting (3.116) into (3.109) yields

$$\max_k\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i}-\widetilde{\underline{\boldsymbol{w}}}'_{k,i}\|^2 \le \gamma\max_k\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}-\widetilde{\underline{\boldsymbol{w}}}'_{k,i-1}\|^2+\frac{\tau'^2\nu}{2\min_k\lambda_{k,\min}}\max_k\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4 \qquad (3.117)$$

where

$$\gamma \triangleq \max_k\left(1-\frac{1}{2}\bar{\mu}_k^{(1)}\lambda_{k,\min}\right)=1-\frac{1}{2}\min_k\{\bar{\mu}_k^{(1)}\cdot\lambda_{k,\min}\} \qquad (3.118)$$

When condition (3.14) holds, it can be verified by using (2.183) from Chapter 2 that $|\gamma|<1$. Then, we get from (3.117) that

$$\limsup_{i\to\infty}\left[\max_k\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i}-\widetilde{\underline{\boldsymbol{w}}}'_{k,i}\|^2\right] \le \frac{\tau'^2\nu}{(1-\gamma)2\min_k\lambda_{k,\min}}\limsup_{i\to\infty}\left[\max_k\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{k,i-1}\|^4\right]$$

$$\leq \frac{4\tau'^2 b_4^2 \nu^3}{\min_k \bar{\mu}_k^{(1)} \cdot \min_k \lambda_{k,\min}^2} \leq O(\nu^2) \tag{3.119}$$

where we used Theorem 2.2 from Chapter 2 and the fact from (2.190) of Chapter 2 that $\bar{\mu}_k^{(1)} = O(\nu)$.

3.C Block Operations

Consider a block matrix $\mathcal{X}$ of size $NM \times NM$ and partition it into $N \times N$ blocks where $X_{k\ell}$ denotes its (k,ℓ)-th sub-matrix of size $M \times M$. The block vectorization of $\mathcal{X}$ with block size $M \times M$ is defined as follows [83]:

$$\begin{aligned}\text{bvec}(\mathcal{X}) \triangleq \text{col}\{&\text{vec}(X_{11}), \text{vec}(X_{21}), \ldots, \text{vec}(X_{N1}), \ldots, \\ &\text{vec}(X_{1N}), \text{vec}(X_{2N}), \ldots, \text{vec}(X_{NN})\}\end{aligned} \tag{3.120}$$

Let $\mathcal{Y}$ denote another block matrix of size $NM \times NM$ and let $Y_{k\ell}$ denote its (k,ℓ)-th sub-matrix of size $M \times M$. Then, the block Kronecker product of $\mathcal{X}$ and $\mathcal{Y}$ with block size $M \times M$ is defined by [83]:

$$\mathcal{X} \otimes_b \mathcal{Y} \triangleq [Z_{k\ell}]_{k,\ell=1}^N \tag{3.121}$$

where

$$Z_{k\ell} \triangleq [X_{k\ell} \otimes Y_{mn}]_{m,n=1}^N \tag{3.122}$$

For any matrices $\{X, Y, A, B\}$ of compatible dimensions and with blocks of size $M \times M$, it holds that

$$(X \otimes A) \otimes_b (Y \otimes B) = (X \otimes Y) \otimes (A \otimes B) \tag{3.123}$$

where $\otimes$ denotes the traditional Kronecker product operation. Other useful properties for the $\otimes_b$ operation can be found in [83, pp. 176-179] and are listed here for ease of reference:

$$\text{bvec}(\mathcal{A}\mathcal{B}\mathcal{C}) = (\mathcal{C}^{\mathsf{T}} \otimes_b \mathcal{A}) \cdot \text{bvec}(\mathcal{B}) \tag{3.124}$$

$$\mathrm{bvec}(xy^{\mathsf{T}}) = y \otimes_b x \tag{3.125}$$

$$\mathrm{Tr}(\mathcal{A}\mathcal{B}) = [\mathrm{bvec}(\mathcal{A}^{\mathsf{T}})]^{\mathsf{T}} \cdot \mathrm{bvec}(\mathcal{B}) = [\mathrm{bvec}(\mathcal{A}^{*})]^{*} \cdot \mathrm{bvec}(\mathcal{B}) \tag{3.126}$$

$$(\mathcal{A}\mathcal{C}) \otimes_b (\mathcal{B}\mathcal{D}) = (\mathcal{A} \otimes_b \mathcal{B})(\mathcal{C} \otimes_b \mathcal{D}) \tag{3.127}$$

$$(\mathcal{A} + \mathcal{B}) \otimes_b (\mathcal{C} + \mathcal{D}) = \mathcal{A} \otimes_b \mathcal{C} + \mathcal{B} \otimes_b \mathcal{C} + \mathcal{A} \otimes_b \mathcal{D} + \mathcal{B} \otimes_b \mathcal{D} \tag{3.128}$$

$$(\mathcal{A} \otimes_b \mathcal{B})^{*} = \mathcal{A}^{*} \otimes_b \mathcal{B}^{*} \tag{3.129}$$

$$(\mathcal{A} \otimes_b \mathcal{B})^{\mathsf{T}} = \mathcal{A}^{\mathsf{T}} \otimes_b \mathcal{B}^{\mathsf{T}} \tag{3.130}$$

for any block matrices $\{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}\}$ and any block vectors $\{x, y\}$ with appropriate sizes.

3.D Proof of Theorem 3.2

From the long-term model (3.36), we obtain that

$$\mathbb{E}(\widetilde{\underline{\boldsymbol{w}}}_i'\widetilde{\underline{\boldsymbol{w}}}_i'^{*}|\mathbb{F}_{i-1}) = \mathbb{E}(\boldsymbol{\mathcal{B}}_i\widetilde{\underline{\boldsymbol{w}}}_{i-1}'\widetilde{\underline{\boldsymbol{w}}}_{i-1}'^{*}\boldsymbol{\mathcal{B}}_i^{*}|\mathbb{F}_{i-1}) + \mathbb{E}(\underline{\boldsymbol{s}}_i\underline{\boldsymbol{s}}_i^{*}|\mathbb{F}_{i-1}) \tag{3.131}$$

where the cross terms that involve $\underline{\boldsymbol{s}}_i$ disappear because $\mathbb{E}(\boldsymbol{\mathcal{B}}_i\widetilde{\underline{\boldsymbol{w}}}_{i-1}\underline{\boldsymbol{s}}_i^{*}|\mathbb{F}_{i-1}) = 0$ by (3.24) and (3.30). Performing the block vectorization of block size $2M$ for both sides of (3.131), and using (3.124) and (3.125) yield

$$\mathbb{E}[(\widetilde{\underline{\boldsymbol{w}}}_i'^{*})^{\mathsf{T}} \otimes_b \widetilde{\underline{\boldsymbol{w}}}_i'|\mathbb{F}_{i-1}] = \mathbb{E}[(\boldsymbol{\mathcal{B}}_i^{*})^{\mathsf{T}} \otimes_b \boldsymbol{\mathcal{B}}_i][(\widetilde{\underline{\boldsymbol{w}}}_{i-1}'^{*})^{\mathsf{T}} \otimes_b \widetilde{\underline{\boldsymbol{w}}}_{i-1}'] + \mathbb{E}[(\underline{\boldsymbol{s}}_i^{*})^{\mathsf{T}} \otimes_b \underline{\boldsymbol{s}}_i|\mathbb{F}_{i-1}] \tag{3.132}$$

Using (3.24), (3.127), and (3.130), the second term on the RHS of (3.132) can be expressed as

$$\mathbb{E}[(\underline{\boldsymbol{s}}_i^{*})^{\mathsf{T}} \otimes_b \underline{\boldsymbol{s}}_i|\mathbb{F}_{i-1}] = (\bar{\mathcal{A}} \otimes_b \bar{\mathcal{A}} + \mathcal{C}_A)^{\mathsf{T}}(\bar{\mathcal{M}} \otimes_b \bar{\mathcal{M}} + \mathcal{C}_M)r_i(\boldsymbol{w}_{i-1}) \tag{3.133}$$

Substituting (3.46) and (3.133) into (3.132), taking the expectation with respect to $\mathbb{F}_{i-1}$, and then using (3.43), we arrive at the desired recursion (3.47), namely,

$$\mathbb{E}\,[(\widetilde{\underline{\boldsymbol{w}}}_i'^{*})^{\mathsf{T}} \otimes_b \widetilde{\underline{\boldsymbol{w}}}_i'] = \mathcal{F}^{*}\mathbb{E}\,[(\widetilde{\underline{\boldsymbol{w}}}_{i-1}'^{*})^{\mathsf{T}} \otimes_b \widetilde{\underline{\boldsymbol{w}}}_{i-1}'] + y_i \tag{3.134}$$

From (3.46), we know that $\mathcal{G}$ in (3.45) is a factor of $\mathcal{F}$. Hence, we use the following result to examine the stability of $\mathcal{F}$.

Lemma 3.8 (Properties of $\mathcal{G}$)**.** *The matrix $\mathcal{G}$ in* (3.45) *satisfies the following properties:*

1. *Block diagonal and Hermitian matrix: it holds that*

$$\mathcal{G} = \mathrm{diag}\{G_1, G_2, \ldots, G_N\} \tag{3.135}$$

$$G_\ell = \mathrm{diag}\{G_{\ell,1}, G_{\ell,2}, \ldots, G_{\ell,N}\} \tag{3.136}$$

   where G_ℓ denotes the ℓ-th block on the diagonal of $\mathcal{G}$ with block size $4M^2N \times 4M^2N$ and $G_{\ell,k}$ denotes the k-th block on the diagonal of G_ℓ with block size $4M^2 \times 4M^2$. The block $G_{\ell,k}$ is Hermitian and is given by

$$G_{\ell,k} \triangleq \bar{D}_\ell^{\mathsf{T}} \otimes \bar{D}_k + c_{\mu,\ell,k}(H_\ell^{\mathsf{T}} \otimes H_k) \tag{3.137}$$

   where $\bar{D}_k$ is given by (3.27)*.*

2. *Norms and spectral radius: it can be verified that*

$$\rho(\mathcal{G}) = \max_{k,m}\{(1 - \bar{\mu}_k\lambda_{k,m})^2 + c_{\mu,k,k}\lambda_{k,m}^2\} \tag{3.138}$$

   where $\lambda_{k,m}$ denotes the m-th eigenvalue of H_k, $m = 1, 2, \ldots, 2M$.

3. *Stability: if condition* (3.18) *holds, then*

$$\rho(\mathcal{G}) < 1 \tag{3.139}$$

Proof. See Appendix 3.E. □

Using the fact that $\mathcal{G}$ is block diagonal and Hermitian, and that $\bar{A} \otimes \bar{A} + C_A$ is block left-stochastic, result (153) from [78, App. A] implies that

$$\rho(\mathcal{F}) \leq \rho(\mathcal{G}) \tag{3.140}$$

By (3.139) and (3.140), we conclude that $\mathcal{F}$ is stable if condition (3.18) holds.

3.E Proof of Lemma 3.8

The first property relating to the block diagonal and Hermitian structure of (3.135)–(3.137) is established by using the definition of $\otimes_b$ and (3.45). Because the matrix $\boldsymbol{\mathcal{D}}_i$ is block diagonal with block size $2M \times 2M$, the block Kronecker product:

$$\boldsymbol{\mathcal{G}}_i \triangleq \boldsymbol{\mathcal{D}}_i^{\mathsf{T}} \otimes_b \boldsymbol{\mathcal{D}}_i \tag{3.141}$$

is block diagonal with block size $4M^2N \times 4M^2N$ and each block is itself block diagonal with block size $4M^2 \times 4M^2$. Let us denote the ℓ-th block on the diagonal of $\boldsymbol{\mathcal{G}}_i$ with block size $4M^2N \times 4M^2N$ by

$$\boldsymbol{G}_{\ell,i} \triangleq \boldsymbol{D}_{\ell,i}^{\mathsf{T}} \otimes \boldsymbol{\mathcal{D}}_i \tag{3.142}$$

and the k-th block on the diagonal of $\boldsymbol{G}_{\ell,i}$ with block size $4M^2 \times 4M^2$ by

$$\boldsymbol{G}_{\ell,k,i} \triangleq \boldsymbol{D}_{\ell,i}^{\mathsf{T}} \otimes \boldsymbol{D}_{k,i} \tag{3.143}$$

where we used the fact that $\boldsymbol{D}_{\ell,i}$ is Hermitian. Then, we have

$$\boldsymbol{\mathcal{G}}_i = \mathrm{diag}\{\boldsymbol{G}_{1,i}, \boldsymbol{G}_{2,i}, \ldots, \boldsymbol{G}_{N,i}\} \tag{3.144}$$

$$\boldsymbol{G}_{\ell,i} = \mathrm{diag}\{\boldsymbol{G}_{\ell,1,i}, \boldsymbol{G}_{\ell,2,i}, \ldots, \boldsymbol{G}_{\ell,N,i}\} \tag{3.145}$$

Using (3.45) and taking the expectation of both sides of (3.144) and (3.145), we get (3.135) and (3.136) by identifying:

$$G_\ell = \mathbb{E}[\boldsymbol{G}_{\ell,i}], \qquad G_{\ell,k} = \mathbb{E}[\boldsymbol{G}_{\ell,k,i}] \tag{3.146}$$

Equation (3.137) follows from (3.146), (3.143), (3.21), and (3.27). Since the matrices $\{G_{\ell,k}\}$ are all Hermitian, by (3.135) and (3.136), the matrix $\mathcal{G}$ is also Hermitian.

The second property in (3.138) is established by using the block diagonal and Hermitian properties of $\mathcal{G}$ to readily conclude that $\rho(\mathcal{G}) = \max_{\ell,k} \rho(G_{\ell,k})$. Furthermore, by (3.137), the eigenvalues of $G_{\ell,k}$ are given by

$$\lambda_{m,n}(G_{\ell,k}) = (1 - \bar{\mu}_\ell \lambda_{\ell,n})(1 - \bar{\mu}_k \lambda_{k,m}) + c_{\mu,\ell,k} \lambda_{\ell,n} \lambda_{k,m} \tag{3.147}$$

for any $\ell, k = 1, 2, \ldots, N$, where $\lambda_{k,m}$ denotes the m-th eigenvalue of H_k and $m, n = 1, 2, \ldots, 2M$. It is straightforward to verify that

$$\lambda_{m,n}(G_{\ell,k}) = \mathbb{E}[(1 - \boldsymbol{\mu}_\ell(i) \lambda_{\ell,n})(1 - \boldsymbol{\mu}_k(i) \lambda_{k,m})] \tag{3.148}$$

Using Cauchy-Schwarz inequality, we get

$$\begin{aligned} |\mathbb{E}[(1 - \boldsymbol{\mu}_\ell(i) \lambda_{\ell,n})(1 - \boldsymbol{\mu}_k(i) \lambda_{k,m})]| &\leq \sqrt{\mathbb{E}[(1 - \boldsymbol{\mu}_\ell(i) \lambda_{\ell,n})^2] \mathbb{E}[(1 - \boldsymbol{\mu}_k(i) \lambda_{k,m})^2]} \\ &\leq \max_{k,m} \{\mathbb{E}[(1 - \boldsymbol{\mu}_k(i) \lambda_{k,m})^2]\} \\ &= \max_{k,m} \{(1 - \bar{\mu}_k \lambda_{k,m})^2 + c_{\mu,k,k} \lambda_{k,m}^2\} \end{aligned} \tag{3.149}$$

where the first inequality becomes equality when $\ell = k$ and $n = m$. From (3.147)–(3.149) we get

$$|\lambda_{m,n}(G_{\ell,k})| \leq \max_{k,m} \{(1 - \bar{\mu}_k \lambda_{k,m})^2 + c_{\mu,k,k} \lambda_{k,m}^2\} \tag{3.150}$$

for any ℓ, k, m, and n. Since the above inequality applies to *all* eigenvalues of $G_{\ell,k}$, and since $G_{\ell,k}$ is Hermitian, we get

$$\rho(G_{\ell,k}) \leq \max_{k,m} \{(1 - \bar{\mu}_k \lambda_{k,m})^2 + c_{\mu,k,k} \lambda_{k,m}^2\} \tag{3.151}$$

Furthermore, from (3.147) we know that

$$\lambda_{m,m}(G_{k,k}) = (1 - \bar{\mu}_k \lambda_{k,m})^2 + c_{\mu,k,k} \lambda_{k,m}^2 \tag{3.152}$$

so that equality in (3.151) is achievable for some k and m.

For the third property in (3.139), we introduce the quadratic function

$$f(x) \triangleq (1 - \bar{\mu}_k x)^2 + c_{\mu,k,k} x^2 \tag{3.153}$$

with $x \in [\lambda_{k,\min}, \lambda_{k,\max}]$. It is easy to verify that $f(x)$ achieves its maximum value at either one of its boundaries:

$$\begin{aligned} f(x) &\leq \max\{f(\lambda_{k,\min}), f(\lambda_{k,\max})\} \\ &\leq 1 - 2\bar{\mu}_k \lambda_{k,\min} + (\bar{\mu}_k^2 + c_{\mu,k,k})\lambda_{k,\max}^2 \end{aligned} \tag{3.154}$$

From Assumption 2.2 in Chapter 2 we have $\lambda_{k,\min} \leq \lambda_{k,m} \leq \lambda_{k,\max}$ for any k and m. We then deduce from (3.154) that

$$f(\lambda_{k,m}) \leq 1 - 2\bar{\mu}_k \lambda_{k,\min} + (\bar{\mu}_k^2 + c_{\mu,k,k})\lambda_{k,\max}^2 \tag{3.155}$$

for any k and m. Using (3.138), (3.153), and (3.155), we get

$$\begin{aligned} \rho(\mathcal{G}) &= \max_{k,m}\{f(\lambda_{k,m})\} \\ &\leq \max_{k}\{f(\lambda_{k,\min}), f(\lambda_{k,\max})\} \\ &< \max_{k}\{f(\lambda_{k,\min}), f(\lambda_{k,\max})\} + \alpha(\bar{\mu}_k^2 + c_{\mu,k,k}) \end{aligned} \tag{3.156}$$

where $\alpha > 0$ by Assumption 3.1. When condition (3.18) holds, using (2.136) from Chapter 2, we have

$$\max_{k}\{1 - 2\bar{\mu}_k \lambda_{k,\min} + (\bar{\mu}_k^2 + c_{\mu,k,k})(\lambda_{k,\max}^2 + \alpha)\} < 1 \tag{3.157}$$

Therefore, by (3.156) and (3.157), if condition (3.18) holds, then $\rho(\mathcal{G}) < 1$, which completes the proof.

3.F Proof of Lemma 3.2

From Lemma 2.3 in Chapter 2 we know that the matrices $\bar{A} \otimes \bar{A} + C_A$ and $\bar{A}$ are both left-stochastic. To establish the desired result, we only need to show that

the matrix $\bar{A} \otimes \bar{A}$ is primitive if $\bar{A} \otimes \bar{A} + C_A$ is primitive. This is because if $\bar{A} \otimes \bar{A}$ is primitive, then for some finite positive integer $j > 0$, the matrix $(\bar{A} \otimes \bar{A})^j$ has strictly positive entries. Since $(\bar{A} \otimes \bar{A})^j = \bar{A}^j \otimes \bar{A}^j$ and $\bar{A}$ has nonnegative entries, $\bar{A}^j$ must have strictly positive entries. Therefore, $\bar{A}$ is primitive.

In order to prove that the matrix $\bar{A} \otimes \bar{A}$ is primitive if $\bar{A} \otimes \bar{A} + C_A$ is primitive, we first introduce the following concept.

Definition 3.2 (Comparing sparsity)**.** *For any two matrices $\{A, B\}$ with nonnegative entries and of the same size, the matrix A is called* sparser *than B, or, equivalently, B is called* denser *than A, if, and only if, $[B]_{\ell k} > 0$ whenever $[A]_{\ell k} > 0$ for any k and ℓ.* □

It is straightforward to verify the following three useful properties related to Definition 3.2.

Lemma 3.9 (Denser product)**.** *For any $M \times N$ matrices $\{A, B\}$ and any $N \times P$ matrices $\{C, D\}$ all with nonnegative entries, if B is denser than A and D is denser than C, then BD is denser than AC.* □

Lemma 3.10 (Denser Kronecker product)**.** *For any $M \times N$ matrices $\{A, B\}$ and any $P \times Q$ matrices $\{C, D\}$ all with nonnegative entries, if B is denser than A and D is denser than C, then $B \otimes D$ is denser than $A \otimes C$.* □

Lemma 3.11 (Sum is not denser)**.** *For any set of $M \times N$ matrices $\{A_i\}$ with nonnegative entries, where $i \in \mathcal{I}$ and $\mathcal{I}$ is an index set (which can be uncountable), if there exists an $M \times N$ matrix B with nonnegative entries such that B is denser than every A_i, $i \in \mathcal{I}$, and assuming that the sum $S \triangleq \sum_{i \in \mathcal{I}} A_i$ exists, then B is also denser than S.* □

Now, from Lemma 2.2 in Chapter 2, we know that $\bar{A}$ is denser than any realization of $\boldsymbol{A}_i$, say, $\boldsymbol{A}_i(\omega)$ where $\omega \in \Omega$ and Ω is the sample space of $\boldsymbol{A}_i$. Using

Lemma 3.10, we get that $\bar{A} \otimes \bar{A}$ is denser than any $\boldsymbol{A}_i(\omega) \otimes \boldsymbol{A}_i(\omega)$. Using Lemma 3.11 and the fact that the probability measures only take nonnegative values, we get that $\bar{A} \otimes \bar{A}$ is denser than $\bar{A} \otimes \bar{A} + C_A = \mathbb{E}\left[\boldsymbol{A}_i \otimes \boldsymbol{A}_i\right]$. If $\bar{A} \otimes \bar{A} + C_A$ is primitive, then there exists a finite positive integer $j > 0$ such that $(\bar{A} \otimes \bar{A} + C_A)^j$ has strictly positive entries. Using Lemma 3.9, we know that $(\bar{A} \otimes \bar{A})^j$ must be denser than $(\bar{A} \otimes \bar{A} + C_A)^j$. Therefore, $(\bar{A} \otimes \bar{A})^j$ must also have strictly positive entries, which means that $\bar{A} \otimes \bar{A}$ must be primitive.

3.G Proof of Lemma 3.3

We first show that $P_p = P_p^{\mathsf{T}}$, or equivalently,

$$p = \mathrm{vec}(P_p^{\mathsf{T}}) \tag{3.158}$$

Lemma 3.12 (Vec-permutation matrix)**.** *The $N^2 \times N^2$ vec-permutation matrix Π is a matrix whose columns are formed from the basis vectors in $\mathbb{R}^{N^2}$ and it satisfies:*

$$\mathrm{vec}(A) = \Pi \cdot \mathrm{vec}(A^{\mathsf{T}}) \tag{3.159}$$

for any $N \times N$ matrix A. Then, for any $N \times N$ matrices $\{A, B\}$,

$$A \otimes B = \Pi(B \otimes A)\Pi \tag{3.160}$$

In addition, $\Pi = \Pi^{\mathsf{T}} = \Pi^* = \Pi^{-1}$.

Proof. See [84, Tabs. I and II] [85, Eqs. (5) and (6)]. □

Let Π be the permutation matrix that satisfies

$$\mathrm{vec}(P_p^{\mathsf{T}}) = \Pi \cdot \mathrm{vec}(P_p) \tag{3.161}$$

From (3.75) and (3.161), proving (3.158) is equivalent to proving

$$p = \Pi \cdot p \tag{3.162}$$

To establish (3.162), we only need to show that $\Pi \cdot p$ is the Perron eigenvector of $\bar{A} \otimes \bar{A} + C_A$. In that case, we can obtain (3.162) directly from the uniqueness of the Perron eigenvector, which is p. Thus, note that

$$\begin{aligned} \Pi(\bar{A} \otimes \bar{A} + C_A)\Pi &= \Pi[\mathbb{E}(\boldsymbol{A}_i \otimes \boldsymbol{A}_i]\Pi \\ &\overset{(a)}{=} \mathbb{E}\,(\boldsymbol{A}_i \otimes \boldsymbol{A}_i) \\ &= \bar{A} \otimes \bar{A} + C_A \end{aligned} \tag{3.163}$$

where step (a) is by (3.160). Then, we deduce from (3.72) that

$$\Pi \cdot p = \Pi(\bar{A} \otimes \bar{A} + C_A)p = (\bar{A} \otimes \bar{A} + C_A)(\Pi \cdot p) \tag{3.164}$$

$$\mathbb{1}_{N^2}^{\mathsf{T}} \cdot \Pi \cdot p = \mathbb{1}_{N^2}^{\mathsf{T}} \cdot p = 1 \tag{3.165}$$

where we used the fact that $\Pi^2 = I_{N^2}$ by Lemma 3.12 and $\Pi \cdot \mathbb{1}_{N^2} = \mathbb{1}_{N^2}$. Results (3.164) and (3.165) establish that $\Pi \cdot p$ is the Perron eigenvector of $\bar{A} \otimes \bar{A} + C_A$ and proves (3.162).

We next establish that P_p is positive semi-definite. Note that for any vector $x \in \mathbb{R}^N$:

$$x^{\mathsf{T}} P_p x = \mathrm{vec}(x^{\mathsf{T}} P_p x) = \frac{1}{N^2}(x^{\mathsf{T}} \otimes x^{\mathsf{T}})p \cdot \mathbb{1}_{N^2}^{\mathsf{T}} \mathbb{1}_{N^2} \tag{3.166}$$

by using (3.75) and the fact that $\mathbb{1}_{N^2}^{\mathsf{T}} \mathbb{1}_{N^2} = N^2$. Since $\bar{A} \otimes \bar{A} + C_A = \mathbb{E}(\boldsymbol{A}_j \otimes \boldsymbol{A}_j)$, we can introduce a series of fictitious random combination matrices $\{\boldsymbol{A}'_j; j \geq 1\}$ such that they are mutually-independent and satisfy $\mathbb{E}(\boldsymbol{A}'_j \otimes \boldsymbol{A}'_j) = \bar{A} \otimes \bar{A} + C_A$ for any $j \geq 1$. Let $\boldsymbol{\Phi}_i \triangleq \prod_{j=1}^{i} \boldsymbol{A}'_j$ for any $i \geq 1$. Then,

$$\lim_{i \to \infty} \mathbb{E}(\boldsymbol{\Phi}_i \otimes \boldsymbol{\Phi}_i) \overset{(a)}{=} \lim_{i \to \infty} \prod_{j=1}^{i} \mathbb{E}(\boldsymbol{A}'_j \otimes \boldsymbol{A}'_j)$$

$$= \lim_{i\to\infty} (\bar{A}\otimes\bar{A} + C_A)^i$$
$$\overset{(b)}{=} p\cdot \mathbb{1}_{N^2}^{\mathsf{T}} \tag{3.167}$$

where step (a) is by using the fact that the $\{\boldsymbol{A}_j'\}$ are mutually-independent, and step (b) is by using the Perron-Frobenius Theorem [79]. Substituting (3.167) into (3.166) and using the fact that $\mathbb{1}_{N^2} = \mathbb{1}_N \otimes \mathbb{1}_N$, we get

$$x^{\mathsf{T}} P_p x = \frac{1}{N^2} \lim_{i\to\infty} \mathbb{E}\,[(x^{\mathsf{T}}\boldsymbol{\Phi}_i \mathbb{1}_N)^2] \geq 0 \tag{3.168}$$

which shows that P_p is positive semi-definite.

Now we show that $P_p \mathbb{1}_N = \bar{p}$. Note from (3.75) and (3.72) that

$$P_p = \mathbb{E}\,[\boldsymbol{A}_i \cdot P_p \cdot \boldsymbol{A}_i^{\mathsf{T}}] \tag{3.169}$$

by switching the order of unvec$(\cdot)$ and $\mathbb{E}(\cdot)$ and applying unvec$(\cdot)$ to the identity $\text{vec}(ABC) = (C^{\mathsf{T}}\otimes A)\cdot \text{vec}(B)$. Furthermore, we get from (3.169) that

$$P_p \cdot \mathbb{1}_N = \mathbb{E}\,(\boldsymbol{A}_i P_p \boldsymbol{A}_i^{\mathsf{T}})\cdot \mathbb{1}_N = \bar{A}(P_p\cdot \mathbb{1}_N) \tag{3.170}$$

which implies that the vector $P_p\cdot\mathbb{1}_N$ is the Perron eigenvector of $\bar{A}$, which is $\bar{p}$. Because the Perron eigenvector is unique, by (3.73), equation $P_p\cdot\mathbb{1}_N = \bar{p}$ must hold.

3.H Proof of Lemma 3.4

We first establish that F is stable if condition (3.18) is satisfied. From (3.137) and (3.77), we get

$$F = \sum_{\ell=1}^{N}\sum_{k=1}^{N} p_{\ell,k} G_{\ell,k} \tag{3.171}$$

By (3.72) and (3.75), the elements $\{p_{\ell,k}\}$ of P_p satisfy:

$$\sum_{\ell=1}^{N}\sum_{k=1}^{N} p_{\ell,k} = 1, \quad \text{and} \quad p_{\ell,k} > 0 \tag{3.172}$$

Then, in terms of the 2-induced norm, we have

$$\|F\| \overset{(a)}{\leq} \sum_{k=1}^{N}\sum_{\ell=1}^{N} p_{\ell,k}\|G_{\ell,k}\| \overset{(b)}{\leq} \max_{k,\ell}\|G_{\ell,k}\| \overset{(c)}{=} \rho(\mathcal{G}) \tag{3.173}$$

where step (a) is from the triangle inequality of norms; step (b) is by using (3.172); and step (c) is by (3.138). Using (3.139), (3.173), and the fact that $\rho(F) = \|F\|$ for the Hermitian matrix F, we conclude that matrix F is stable if condition (3.18) holds.

We now establish expression (3.78). Introduce the Hermitian matrix

$$F' \triangleq \sum_{k=1}^{N}\sum_{\ell=1}^{N} \bar{p}_\ell \bar{p}_k (\bar{D}_\ell^{\mathsf{T}} \otimes \bar{D}_k) \tag{3.174}$$

From (3.27), we can rewrite F' as $F' = (I_{2M} - H)^{\mathsf{T}} \otimes (I_{2M} - H)$, where we used (3.73) and (3.79). Therefore, the eigenvalues of F' are equal to the products of any two of the eigenvalues of $I_{2M} - H$, which are given by $1 - \lambda(H)$. Since $\{\bar{p}_k\}$ and $\{\bar{\mu}_k\}$ are all positive and $\{H_k\}$ are all positive definite, it is easy to verify that H in (3.79) is also positive definite. Then, from (3.79), (3.73), and (3.10), and using (2.7) from Chapter 2 as well as Jensen's inequality [71], we get

$$0 < \lambda(H) \leq \|H\| \leq \max_k\{\bar{\mu}_k \lambda_{k,\max}\} \tag{3.175}$$

for all eigenvalues of H. When condition (3.18) holds, we get from (2.187) of Chapter 2 that

$$\bar{\mu}_k \leq \frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k} < \frac{\lambda_{k,\min}}{\lambda_{k,\max}^2 + \alpha} < \frac{1}{\lambda_{k,\max}} \tag{3.176}$$

for any k. This implies that $\max_k\{\bar{\mu}_k\lambda_{k,\max}\} < 1$ and therefore, $0 < \lambda(H) < 1$ for all eigenvalues of H. From (3.79) and (3.181), we obtain

$$0 < \lambda(H) = O(\nu) < 1 \tag{3.177}$$

for any eigenvalue of H. Therefore, we get

$$\lambda(I_{2M} - H) = 1 - O(\nu), \qquad \rho(I_{2M} - H) = 1 - \lambda_{\min}(H) \tag{3.178}$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of its Hermitian matrix argument. We further get from (3.178) that

$$\lambda(F') = 1 - O(\nu), \qquad \rho(F') = [1 - \lambda_{\min}(H)]^2 \tag{3.179}$$

Using Lemma 3.3, (3.72), (3.73), and (3.27), the difference between F in (3.77) and F' in (3.174) is given by

$$F - F' = \sum_{k=1}^{N} \sum_{\ell=1}^{N} \{[(p_{\ell,k} - \bar{p}_\ell \bar{p}_k)\bar{\mu}_\ell \bar{\mu}_k + p_{\ell,k} c_{\mu,\ell,k}](H_\ell^{\mathsf{T}} \otimes H_k)\} \tag{3.180}$$

which is also Hermitian. From (2.190) in Chapter 2, we get

$$\bar{\mu}_k \equiv \bar{\mu}_k^{(1)} \leq \nu \tag{3.181}$$

$$c_{\mu,k,k} \leq \bar{\mu}_k^{(2)} \leq \nu^2 \tag{3.182}$$

$$|c_{\mu,\ell,k}| \leq \sqrt{c_{\mu,\ell,\ell} \cdot c_{\mu,k,k}} \leq \nu^2 \tag{3.183}$$

where (3.183) is by using the Cauchy-Schwartz inequality. By (3.181)–(3.183), we get $\|F - F'\| = O(\nu^2)$. Using a corollary of the Wielandt-Hoffman theorem [86, Corollary 8.1.6, p. 396], we then conclude that

$$|\lambda_m(F) - \lambda_m(F')| \leq \|F - F'\| = O(\nu^2) \tag{3.184}$$

where $\lambda_m(\cdot)$ denotes the m-th eigenvalue of its Hermitian matrix argument; the eigenvalues are assumed to be ordered from largest to smallest in each case. Result (3.184) implies that for every eigenvalue of F' there is an eigenvalue of F that is $O(\nu^2)$ close to it. From (3.184) and (3.179) we immediately deduce that

$$\lambda_m(F) = 1 - O(\nu), \qquad \rho(F) = \rho(F') + O(\nu^2) \tag{3.185}$$

where $\rho(F')$ from (3.179) dominates the $O(\nu^2)$ term.

3.I Proof of Lemma 3.5

We first establish (3.80). Introduce the Jordan decomposition:

$$\bar{A} \otimes \bar{A} + C_A \triangleq PJQ^{\mathsf{T}} = \begin{bmatrix} p & P' \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & J' \end{bmatrix} \begin{bmatrix} \mathbb{1}_{N^2} & Q' \end{bmatrix}^{\mathsf{T}} \tag{3.186}$$

where J is the Jordan canonical form of $\bar{A} \otimes \bar{A} + C_A$ and J' is a sub-matrix of J containing its stable eigenvalues, P' and Q' are sub-matrices of P and Q, and $P^{-1} = Q^{\mathsf{T}}$. Then, the Jordan decomposition of $\bar{\mathcal{A}} \otimes_b \bar{\mathcal{A}} + \mathcal{C}_A$ is given by

$$\bar{\mathcal{A}} \otimes_b \bar{\mathcal{A}} + \mathcal{C}_A = \mathcal{P}\mathcal{J}\mathcal{Q}^{\mathsf{T}} = \begin{bmatrix} \mathcal{P}_1 & \mathcal{P}' \end{bmatrix} \begin{bmatrix} I_{4M^2} & 0 \\ 0 & \mathcal{J}' \end{bmatrix} \begin{bmatrix} \mathcal{Q}_1 & \mathcal{Q}' \end{bmatrix}^{\mathsf{T}} \tag{3.187}$$

where

$$\mathcal{P} \triangleq P \otimes I_{4M^2}, \qquad \mathcal{P}' \triangleq P' \otimes I_{4M^2} \tag{3.188}$$

$$\mathcal{J} \triangleq J \otimes I_{4M^2}, \qquad \mathcal{J}' \triangleq J' \otimes I_{4M^2} \tag{3.189}$$

$$\mathcal{Q} \triangleq Q \otimes I_{4M^2}, \qquad \mathcal{Q}' \triangleq Q' \otimes I_{4M^2} \tag{3.190}$$

$$\mathcal{P}_1 \triangleq p \otimes I_{4M^2}, \qquad \mathcal{Q}_1 \triangleq \mathbb{1}_{N^2} \otimes I_{4M^2} \tag{3.191}$$

Let

$$\mathcal{X} \triangleq I_{4M^2N^2} - \mathcal{G} \tag{3.192}$$

where $\mathcal{G}$ is given by (3.45). Then, by (3.46),

$$\begin{aligned} \mathcal{Q}^{\mathsf{T}}\mathcal{F}\mathcal{P} &= \mathcal{Q}^{\mathsf{T}}[\mathcal{G} \cdot (\bar{\mathcal{A}} \otimes_b \bar{\mathcal{A}} + \mathcal{C}_A)]\mathcal{P} \\ &= \mathcal{Q}^{\mathsf{T}}(I_{4M^2N^2} - \mathcal{X})(\mathcal{P}\mathcal{J}\mathcal{Q}^{\mathsf{T}})\mathcal{P} \\ &= \mathcal{J} - \mathcal{Q}^{\mathsf{T}}\mathcal{X}\mathcal{P}\mathcal{J} \\ &= \begin{bmatrix} I_{4M^2} - \mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1 & -\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}' \\ -\mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}_1 & \mathcal{J}' - \mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}' \end{bmatrix} \end{aligned} \tag{3.193}$$

From (3.193), we further get

$$(I_{4M^2N^2} - \mathcal{Q}^{\mathsf{T}}\mathcal{F}\mathcal{P})^{-1} = \begin{bmatrix} \mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1 & \mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}' \\ \mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}_1 & I - \mathcal{J}' + \mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}' \end{bmatrix}^{-1} \tag{3.194}$$

where the I denotes the $4M^2(N^2-1) \times 4M^2(N^2-1)$ identity matrix. The quantity $\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1$ in (3.193) can be expressed as

$$\begin{aligned} \mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1 &\overset{(a)}{=} \mathcal{Q}_1^{\mathsf{T}} \cdot \mathcal{P}_1 - \mathcal{Q}_1^{\mathsf{T}}\mathcal{G}\mathcal{P}_1 \\ &\overset{(b)}{=} (\mathbb{1}_{N^2}^{\mathsf{T}} p) \otimes I_{4M^2} - (\mathbb{1}_{N^2}^{\mathsf{T}} \otimes I_{4M^2})(\mathrm{diag}\{G_{\ell,k}\})(p \otimes I_{4M^2}) \\ &\overset{(c)}{=} I_{4M^2} - F \end{aligned} \tag{3.195}$$

where step (a) is by (3.192); step (b) is by (3.135)–(3.136); and step (c) is by (3.171). We already know that the matrices F and $\mathcal{F}$ are stable for sufficiently small step-sizes. Thus, the matrices $I_{4M^2N^2} - \mathcal{F}$ and $I_{4M^2} - F$ are invertible. It follows that the quantity $\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1$ is invertible. Moreover, the Schur complement with respect to $\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1$ in (3.194) is also invertible. Let us denote the inverse of this Schur complement by

$$\Delta \triangleq [I - \mathcal{J}' + \mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}' - \mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}_1(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}']^{-1} \tag{3.196}$$

Then, by using a formula for the inversion of block matrices [87, Eq. (7), p. 48], equality (3.194) can be expressed as

$$(I_{4M^2N^2} - \mathcal{Q}^{\mathsf{T}}\mathcal{F}\mathcal{P})^{-1} = \begin{bmatrix} (\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1} + \Delta' & -(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}'\Delta \\ -\Delta\mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}_1(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1} & \Delta \end{bmatrix} \tag{3.197}$$

where

$$\Delta' \triangleq (\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}'\Delta\mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}_1(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1} \tag{3.198}$$

Now, from (3.195), (3.77), (3.27), (3.76), and (3.79), we can also write

$$\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1 = H \otimes I_{2M} + I_{2M} \otimes H - \sum_{\ell,k=1}^{N} p_{\ell,k}(\bar{\mu}_\ell\bar{\mu}_k + c_{\mu,\ell,k})(H_\ell^{\mathsf{T}} \otimes H_k) \tag{3.199}$$

It follows from (3.199) and (3.181)–(3.183) that $\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1$ is Hermitian and

$$\|\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1\| = O(\nu), \quad \|(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\| = O(\nu^{-1}) \tag{3.200}$$

Likewise, from (3.26) and (3.181)–(3.183), we get that

$$\|\bar{\mathcal{M}}\| = O(\nu), \qquad \|\mathcal{C}_M\| = O(\nu^2) \tag{3.201}$$

and from (3.192), (3.45), (3.28), and (3.33), we further get

$$\|\mathcal{X}\| = \|(\bar{\mathcal{M}}\mathcal{H})^{\mathsf{T}} \otimes_b I_{2MN} + I_{2MN} \otimes_b (\bar{\mathcal{M}}\mathcal{H}) + O(\nu^2)\| = O(\nu) \tag{3.202}$$

since matrix $\mathcal{H}$ is constant and independent of ν. Furthermore, it follows from (3.202) that

$$\|\mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}'\| = O(\nu), \quad \|\mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}_1\| = O(\nu), \quad \|\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}'\| = O(\nu) \tag{3.203}$$

From (3.187), matrix $I - \mathcal{J}'$ is invertiable and is independent of ν. Therefore,

$$\|\mathcal{J}'\| = O(1), \;\; \|I - \mathcal{J}'\| = O(1), \;\; \|(I - \mathcal{J}')^{-1}\| = O(1) \tag{3.204}$$

Then, by (3.196), (3.203), and (3.204), we get

$$\|\Delta\| = \|(I - \mathcal{J}' + O(\nu))^{-1}\| = O(1) \tag{3.205}$$

By (3.198), (3.200), (3.203), and (3.205), we further get

$$\begin{gathered}\|\Delta'\| = O(1), \qquad \|(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}'\mathcal{J}'\Delta\| = O(1), \\ \|\Delta\mathcal{Q}'^{\mathsf{T}}\mathcal{X}\mathcal{P}_1(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\| = O(1)\end{gathered} \tag{3.206}$$

Using (3.206) and Assumption 3.2, we get from (3.197) that

$$(I_{4M^2N^2} - \mathcal{Q}^{\mathsf{T}}\mathcal{F}\mathcal{P})^{-1} = \begin{bmatrix} (\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1} & 0 \\ 0 & 0 \end{bmatrix} + O(1) \tag{3.207}$$

Then,

$$
\begin{aligned}
(I_{4M^2N^2}-\mathcal{F})^{-1} &\overset{(a)}{=} \mathcal{P}\begin{bmatrix}(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1} & 0\\ 0 & 0\end{bmatrix}\mathcal{Q}^{\mathsf{T}}+O(1)\\
&\overset{(b)}{=} \mathcal{P}_1(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\mathcal{Q}_1^{\mathsf{T}}+O(1)\\
&\overset{(c)}{=} (p\mathbb{1}_{N^2}^{\mathsf{T}})\otimes(I_{4M^2}-F)^{-1}+O(1)
\end{aligned}
\tag{3.208}
$$

where step (a) is by using the fact that $\mathcal{P}^{-1}=\mathcal{Q}^{\mathsf{T}}$; step (b) is by using the block division in (3.187); step (c) is by using (3.191); and by (3.195) and (3.200),

$$
\|(I_{4M^2}-F)^{-1}\| = O(\nu^{-1}) \tag{3.209}
$$

Under Assumption 3.2, the parameter $\nu \ll 1$. Therefore, $\nu^{-1} \gg 1$ and $(p\mathbb{1}_{N^2}^{\mathsf{T}})\otimes(I_{4M^2}-F)^{-1}$ dominates the $O(1)$ term in (3.208).

Finally, we establish result (3.82). Let

$$
\mathcal{F}_s \triangleq \mathcal{Q}^{\mathsf{T}}\mathcal{F}\mathcal{P} = \begin{bmatrix}F & O(\nu)\\ O(\nu) & \mathcal{J}'+O(\nu)\end{bmatrix} \triangleq \begin{bmatrix}F & \mathcal{F}_{12}\\ \mathcal{F}_{21} & \mathcal{F}_{22}\end{bmatrix} \tag{3.210}
$$

by using (3.193), (3.195), (3.203), and (3.204). Since $\mathcal{F}_s$ is similar to $\mathcal{F}$, they have the same eigenvalues [87]. Since F is Hermitian, let us introduce its eigenvalue decomposition as

$$
F = U\Lambda U^* \tag{3.211}
$$

where U is a $4M^2\times 4M^2$ unitary matrix and Λ is a $4M^2\times 4M^2$ diagonal matrix. The $(N^2-1)\times(N^2-1)$ matrix J', which contains the stable eigenvalues of $\bar{A}\otimes\bar{A}+C_A$ in (3.186), can be generally expressed as

$$
J' = \begin{bmatrix}\lambda_{a,2} & & T'\\ & \ddots & \\ 0 & & \lambda_{a,N^2}\end{bmatrix} \tag{3.212}
$$

where $\{\lambda_{a,n}\}$ are the eigenvalues of $\bar{A}\otimes\bar{A}+C_A$ with $\lambda_{a,1}=1$ and $|\lambda_{a,n}|<1$ for all $n=2,3,\ldots,N^2$. In (3.212), the elements in the strictly upper triangular region T' are either 1 or 0, which depend on the Jordan blocks in J'. Using (3.212) and (3.189), we can express the $(2,2)$ block in (3.210) as

$$\mathcal{J}'+O(\nu)=\begin{bmatrix}\lambda_{a,2}I_{4M^2}+O(\nu) & & \mathcal{T}'+O(\nu)\\ & \ddots & \\ O(\nu) & & \lambda_{a,N^2}I_{4M^2}+O(\nu)\end{bmatrix} \tag{3.213}$$

where the elements in the strictly upper triangular region $\mathcal{T}'$ are either 1 or 0, which depend on the elements of T' in (3.212). We now apply a similarity transformation to $\mathcal{F}_s$ in (3.210) by multiplying

$$\mathcal{D}\triangleq\begin{bmatrix}\nu^{\epsilon}U & \\ & \mathcal{D}_2\end{bmatrix} \tag{3.214}$$

and its inverse $\mathcal{D}^{-1}$ on either side of (3.210), where $\epsilon=1/N^2$ and

$$\mathcal{D}_2\triangleq\mathrm{diag}\{\nu^{2\epsilon},\nu^{3\epsilon},\ldots,\nu^{N^2\epsilon}\}\otimes I_{4M^2}\triangleq D\otimes I_{4M^2} \tag{3.215}$$

Using (3.210) and (3.213), we end up with

$$\begin{aligned}\mathcal{D}^{-1}\mathcal{F}_s\mathcal{D}&=\begin{bmatrix}\nu^{-\epsilon}U^* & \\ & \mathcal{D}_2^{-1}\end{bmatrix}\begin{bmatrix}F & \mathcal{F}_{12}\\ \mathcal{F}_{21} & \mathcal{F}_{22}\end{bmatrix}\begin{bmatrix}\nu^{\epsilon}U & \\ & \mathcal{D}_2\end{bmatrix}\\ &=\begin{bmatrix}U^*FU & \nu^{-\epsilon}U^*\mathcal{F}_{12}\mathcal{D}_2\\ \nu^{\epsilon}\mathcal{D}_2^{-1}\mathcal{F}_{21}U & \mathcal{D}_2^{-1}\mathcal{F}_{22}\mathcal{D}_2\end{bmatrix}\end{aligned} \tag{3.216}$$

Using (3.211), the $(1,1)$-block in (3.216) is given by

$$U^*FU=U^*U\Lambda U^*U=\Lambda \tag{3.217}$$

From (3.215), we have

$$\|\mathcal{D}_2\|=\|D\|=\nu^{2\epsilon}\qquad\text{and}\qquad\|\mathcal{D}_2^{-1}\|=\|D^{-1}\|=\nu^{-N^2\epsilon}=\nu^{-1} \tag{3.218}$$

where we used the fact that $0 < \nu < 1$ and $\epsilon = 1/N^2$. Using (3.218) and (3.210), the $(1,2)$- and $(2,1)$-blocks in (3.216) satisfy

$$\|\nu^{-\epsilon} U^* \mathcal{F}_{12} \mathcal{D}_2\| = O(\nu^{1+\epsilon}), \qquad \|\nu^{\epsilon} \mathcal{D}_2^{-1} \mathcal{F}_{21} U\| = O(\nu^{\epsilon}) \tag{3.219}$$

For the $(2,2)$-block in (3.216), we first split the $\mathcal{J}'$ from (3.213) into two parts:

$$\mathcal{J}' \triangleq \Lambda' + \Upsilon \tag{3.220}$$

where Λ' consists of the diagonal entries of $\mathcal{J}'$, namely, $\{\lambda_{a,k}; k = 2, 3, \ldots, N^2\}$, and Υ consists of the first upper off-diagonal entries of $\mathcal{J}'$, namely, the 1 and 0 entries in the Jordan blocks. Then, the $\mathcal{F}_{22}$ in (3.210) can be expressed as

$$\mathcal{F}_{22} = \Lambda' + \Upsilon + O(\nu) \mathbb{1}_{4(N^2-1)M^2} \mathbb{1}_{4(N^2-1)M^2}^{\mathsf{T}} \tag{3.221}$$

where we used the third term on the RHS of (3.221) to explicitly indicate that every entry in $\mathcal{F}_{22}$ is perturbed by a term at least in the order of ν. Now, the $(2,2)$-block in (3.216) can be expressed as

$$\begin{aligned} \mathcal{D}_2^{-1} \mathcal{F}_{22} \mathcal{D}_2 &= \mathcal{D}_2^{-1} [\Lambda' + \Upsilon + O(\nu) \mathbb{1}_{4(N^2-1)M^2} \mathbb{1}_{4(N^2-1)M^2}^{\mathsf{T}}] \mathcal{D}_2 \\ &= \Lambda' + \mathcal{D}_2^{-1} \Upsilon \mathcal{D}_2 + O(\nu) \mathcal{D}_2^{-1} \mathbb{1}_{4(N^2-1)M^2} \mathbb{1}_{4(N^2-1)M^2}^{\mathsf{T}} \mathcal{D}_2 \end{aligned} \tag{3.222}$$

Since Υ only has non-zero entries, which are equal to one, on the first upper off-diagonal, it is straightforward to verify that

$$\|\mathcal{D}_2^{-1} \Upsilon \mathcal{D}_2\| = O(\nu^{\epsilon}) \tag{3.223}$$

Let

$$d \triangleq \operatorname{col}\{\nu^{2\epsilon}, \nu^{3\epsilon}, \ldots, \nu^{N^2\epsilon}\} = D \mathbb{1}_{N^2-1} \tag{3.224}$$

$$e \triangleq \operatorname{col}\{\nu^{-2\epsilon}, \nu^{-3\epsilon}, \ldots, \nu^{-N^2\epsilon}\} = D^{-1} \mathbb{1}_{N^2-1} \tag{3.225}$$

where D is from (3.215). Then, using (3.215), (3.224), and (3.225), the third term on the RHS of (3.222) can be expressed as

$$\begin{aligned} O(\nu)\mathcal{D}_2^{-1}\mathbb{1}_{4(N^2-1)M^2}\mathbb{1}_{4(N^2-1)M^2}^{\mathsf{T}}\mathcal{D}_2 \\ &= O(\nu)(D\otimes I_{4M^2})^{-1}(\mathbb{1}_{N^2-1}\otimes\mathbb{1}_{4M^2})(\mathbb{1}_{N^2-1}\otimes\mathbb{1}_{4M^2})^{\mathsf{T}}(D\otimes I_{4M^2}) \\ &= O(\nu)(D^{-1}\mathbb{1}_{N^2-1}\mathbb{1}_{N^2-1}^{\mathsf{T}}D)\otimes(\mathbb{1}_{4M^2}\mathbb{1}_{4M^2}^{\mathsf{T}}) \\ &= O(\nu)(ed^{\mathsf{T}})\otimes(\mathbb{1}_{4M^2}\mathbb{1}_{4M^2}^{\mathsf{T}}) \end{aligned} \tag{3.226}$$

In the rank-one matrix ed^{T}, due to the fact that $\nu < 1$, the entries on the diagonal are equal to one; the entries above the diagonal are at least in the order of ν^{ϵ}; the entries below the diagonal are at least in the order of $\nu^{-(N^2-2)\epsilon}$. Therefore, it follows from (3.226) that

$$O(\nu)\mathcal{D}_2^{-1}\mathbb{1}_{4(N^2-1)M^2}\mathbb{1}_{4(N^2-1)M^2}^{\mathsf{T}}\mathcal{D}_2 = \begin{bmatrix} O(\nu) & & O(\nu^{1+\epsilon}) \\ & \ddots & \\ O(\nu^{2\epsilon}) & & O(\nu) \end{bmatrix} \tag{3.227}$$

where we used the fact that $\epsilon = 1/N^2$. Substituting (3.222), (3.223), and (3.227) into (3.221) yields

$$\mathcal{F}_{22} = \begin{bmatrix} \lambda_{a,2}I_{4M^2} + O(\nu) & & O(\nu^{\epsilon}) \\ & \ddots & \\ O(\nu^{2\epsilon}) & & \lambda_{a,N^2}I_{4M^2} + O(\nu) \end{bmatrix} \tag{3.228}$$

where we meant that in the matrix $\mathcal{F}_{22}$, the entries above the diagonal are at least in the order of ν^{ϵ}, and the entries below the diagonal are at least in the

order of $\nu^{2\epsilon}$. Substituting (3.217), (3.219), and (3.228) into (3.216) yields

$$\mathcal{D}^{-1}\mathcal{F}_s\mathcal{D} = \left[\begin{array}{c|ccc} \Lambda & \multicolumn{3}{c}{O(\nu^{1+\epsilon})} \\ \hline & \lambda_{a,2}I_{4M^2}+O(\nu) & & O(\nu^{\epsilon}) \\ O(\nu^{\epsilon}) & & \ddots & \\ & O(\nu^{2\epsilon}) & & \lambda_{a,N^2}I_{4M^2}+O(\nu) \end{array}\right] \tag{3.229}$$

From (3.229), we know that all off-diagonal entries of $\mathcal{D}^{-1}\mathcal{F}_s\mathcal{D}$ are *at least* of the order of ν^{ϵ}. Therefore, using Gershgorin Theorem [86, p. 320] under Assumption 3.2, and since $\mathcal{F}$ and $\mathcal{F}_s$ have the same eigenvalues due to similarity, we get

$$|\lambda(\mathcal{F})-\lambda(F)| \leq O(\nu^{1+\epsilon}) \text{ or } |\lambda(\mathcal{F})-\lambda_{a,k}| \leq O(\nu^{\epsilon}) \tag{3.230}$$

where $\lambda(\mathcal{F})$ denotes the eigenvalue of $\mathcal{F}$ and $k = 2, 3, \ldots, N^2$. Result (3.230) implies that the eigenvalues of $\mathcal{F}$ are either located in the Gershgorin circles that are centered at the eigenvalues of F with radii $O(\nu^{1+\epsilon})$ or in the Gershgorin circles that are centered at $\{\lambda_{a,k}; k = 2, 3, \ldots, N^2\}$ with radii $O(\nu^{\epsilon})$. From (3.185), we have

$$\rho(F) = 1 - O(\nu) < 1 \tag{3.231}$$

By Assumption 3.3 and Perron-Frobenius Theorem [79], we have

$$\rho(J') \triangleq \max_{k=2,3,\ldots,N^2} |\lambda_{a,k}| < 1 \tag{3.232}$$

By Assumption 3.2, if the parameter ν is small enough to satisfy

$$O(\nu^{\epsilon}) + O(\nu) < 1 - \rho(J') \tag{3.233}$$

such that the inequality

$$\rho(J') + O(\nu^{\epsilon}) < 1 - O(\nu) = \rho(F) \tag{3.234}$$

holds, then the Gershgorin circles centered at the eigenvalues of F are isolated from those centered at $\{\lambda_{a,k}; k = 2, 3, \ldots, N^2\}$. According to Gershgorin Theorem [88, p. 181], there are precisely $4M^2$ eigenvalues of $\mathcal{F}$ satisfying

$$|\lambda(\mathcal{F}) - \lambda(F)| \leq O(\nu^{1+\epsilon}) \tag{3.235}$$

while all the other eigenvalues satisfy

$$|\lambda(\mathcal{F}) - \lambda_{a,k}| \leq O(\nu^{\epsilon}), \quad k = 2, 3, \ldots, N^2 \tag{3.236}$$

By (3.234), the eigenvalues $\lambda(\mathcal{F})$ satisfying (3.235) are greater than those satisfying (3.236) in magnitude. Furthermore, when ν is sufficiently small, the Gershgorin circles centered at $\lambda_{\max}(F)$ with radius $O(\nu^{1+\epsilon})$ will become disjoint from the other circles (see Fig. 3.4). Then, by using (3.231) and Gershgorin Theorem again, we conclude from (3.235) that

$$\rho(\mathcal{F}) = \rho(F) + O(\nu^{1+\epsilon}) \tag{3.237}$$

It is worth noting that from (3.78) we get

$$\rho(\mathcal{F}) = 1 - O(\nu) + O(\nu^{1+\epsilon}) < 1 \tag{3.238}$$

for $\nu \ll 1$ because $\epsilon = 1/N^2 > 1$.

3.J Proof of Theorem 3.4

From (3.187) and (3.208), we first have

$$\begin{aligned}(\bar{\mathcal{A}} \otimes_b \bar{\mathcal{A}} + \mathcal{C}_A)(I_{4M^2N^2} - \mathcal{F})^{-1} \\ = \begin{bmatrix} \mathcal{P}_1 & \mathcal{P}' \end{bmatrix} \begin{bmatrix} I_{4M^2} & 0 \\ 0 & \mathcal{J}' \end{bmatrix} \begin{bmatrix} (\mathcal{Q}_1^{\mathsf{T}} \mathcal{X} \mathcal{P}_1)^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathcal{Q}_1 & \mathcal{Q}' \end{bmatrix}^{\mathsf{T}} + O(1)\end{aligned}$$

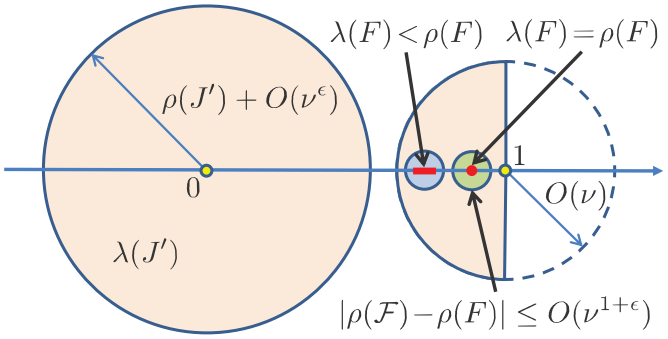


Figure 3.4: An illustration of the locations of the eigenvalues of $\mathcal{F}$. The eigenvalues of J' are all in the left big circle, so the eigenvalues of $\mathcal{F}$ satisfying (3.236) are also in the left big circle. The eigenvalues of F are all in the right big circle, so the eigenvalues of $\mathcal{F}$ satisfying (3.235) are also in the right big circle. Specifically, the eigenvalues of F with $\lambda(F) < \rho(F)$ are all on the red segment on the horizontal line, so the eigenvalues of $\mathcal{F}$ that satisfy (3.235) are all in the small blue circle on the left; the eigenvalues of F with $\lambda(F) = \rho(F)$ are on the red dot on the horizontal line, so the eigenvalues of $\mathcal{F}$ that satisfy (3.235) are all in the small green circle on the right.

$$= \mathcal{P}_1(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\mathcal{Q}_1^{\mathsf{T}} + O(1) \tag{3.239}$$

Since $\mathcal{P}_1(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\mathcal{Q}_1^{\mathsf{T}} = O(\nu^{-1})$ by (3.200), the first term on the RHS of (3.239) dominates the second term under Assumption 3.2. By (3.26), (3.5), (3.32), and (3.120), we get

$$\begin{aligned}\mathcal{P}_1^{\mathsf{T}}(\bar{\mathcal{M}} \otimes_b \bar{\mathcal{M}} + \mathcal{C}_M)\mathrm{bvec}(\mathcal{R}) &= (p^{\mathsf{T}} \otimes I_{4M^2})[(\bar{M} \otimes \bar{M} + C_M) \otimes I_{4M^2}]\mathrm{bvec}(\mathcal{R}) \\ &= \mathrm{vec}(R)\end{aligned} \tag{3.240}$$

where R is defined by (3.83) and is of the order of ν^2. We then get a low-rank expression for z in (3.57):

$$z \overset{(a)}{=} [\mathcal{P}_1(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}\mathcal{Q}_1^{\mathsf{T}} + O(1)]^*(\bar{\mathcal{M}} \otimes_b \bar{\mathcal{M}} + \mathcal{C}_M)\mathrm{bvec}(\mathcal{R})$$

$$\begin{aligned}
&\overset{(b)}{=} \mathcal{Q}_1[(\mathcal{Q}_1^{\mathsf{T}}\mathcal{X}\mathcal{P}_1)^{-1}]^*\mathcal{P}_1^{\mathsf{T}}(\bar{\mathcal{M}}\otimes_b\bar{\mathcal{M}}+\mathcal{C}_M)\mathrm{bvec}(\mathcal{R})+O(\nu^2)\\
&\overset{(c)}{=} \mathcal{Q}_1[(I_{4M^2}-F)^{-1}]^*\mathrm{vec}(R)+O(\nu^2)\\
&\overset{(d)}{=} \mathbb{1}_{N^2}\otimes[(I_{4M^2}-F)^{-1}\mathrm{vec}(R)]+O(\nu^2) \qquad (3.241)
\end{aligned}$$

where step (a) is by (3.239); step (b) is by (3.58); step (c) is by (3.195) and (3.240); and step (d) is by (3.191) and the fact that F is Hermitian. The first term on the RHS of (3.241) is dominant due to (3.209) and (3.85). Applying unbvec$(\cdot)$ to both sides of (3.241) and using (3.84) yields

$$\mathrm{unbvec}(z)=(\mathbb{1}_N\mathbb{1}_N^{\mathsf{T}})\otimes Z+O(\nu^2) \qquad (3.242)$$

where Z is given by (3.84) and is of the order of ν by (3.85). From (3.44) and (3.54), we know that unbvec(z_∞) is the steady-state covariance matrix of $\widetilde{\boldsymbol{w}}_i'$. Using (3.60) and (3.242), the steady-state covariance matrix of $\widetilde{\boldsymbol{w}}_i'$ can be approximated by

$$\begin{aligned}
\lim_{i\to\infty}\mathbb{E}\widetilde{\boldsymbol{w}}_i'\widetilde{\boldsymbol{w}}_i'^{*} &= \mathrm{unbvec}(z_\infty)\\
&= (\mathbb{1}_N\mathbb{1}_N^{\mathsf{T}})\otimes Z+O(\nu^{1+\min\{2,\gamma_v\}/2})\\
&= (\mathbb{1}_N\mathbb{1}_N^{\mathsf{T}})\otimes Z+O(\nu^{1+\gamma_o'}) \qquad (3.243)
\end{aligned}$$

where γ_o' is given by (3.87), and the first term on the RHS is dominant due to (3.85).

3.K Proof of Lemma 3.7

From Lemma 3.4 and Assumption 3.2, we know that matrix F is stable. From (3.84), we get

$$Z=\mathrm{unvec}\left(\sum_{j=0}^{\infty}F^j\mathrm{vec}(R)\right)=\sum_{j=0}^{\infty}\mathrm{unvec}\left(F^j\mathrm{vec}(R)\right) \qquad (3.244)$$

Let

$$R^{(j)} \triangleq \text{unvec}\left(F^j \text{vec}(R)\right), \qquad j \geq 0 \tag{3.245}$$

where $R^{(0)} = R$. Then, since F is stable, the $2M \times 2M$ matrix sequence $\{R^{(j)}; j \geq 0\}$ converges to zero. Substituting (3.245) into (3.244) yields

$$Z = \sum_{j=0}^{\infty} R^{(j)} \tag{3.246}$$

Lemma 3.13 (Condition for complex-Hessian-type matrices)**.** *A sufficient and necessary condition for any $2M \times 2M$ positive semi-definite matrix H to be a complex-Hessian-type matrix in Definition 3.1 is to require $LH^{\mathsf{T}}L = H$, where*

$$L \triangleq \begin{bmatrix} 0 & I_M \\ I_M & 0 \end{bmatrix} \tag{3.247}$$

satisfies $L = L^{\mathsf{T}} = L^{-1}$.

Proof. Let the $2M \times 2M$ positive semi-definite matrix be

$$H = \begin{bmatrix} A & B \\ B^* & D \end{bmatrix} \tag{3.248}$$

where $\{A, B, D\}$ are $M \times M$ submatrices satisfying $A = A^*$ and $D = D^*$. Then,

$$LH^{\mathsf{T}}L \triangleq \begin{bmatrix} D^{\mathsf{T}} & B^{\mathsf{T}} \\ (B^*)^{\mathsf{T}} & A^{\mathsf{T}} \end{bmatrix} \tag{3.249}$$

By Definition 3.1, the matrix H is a complex-Hessian-type matrix if, and only if, $A = D^{\mathsf{T}}$ and $B = B^{\mathsf{T}}$. It is straightforward to verify that these conditions are equivalent to the equality $LH^{\mathsf{T}}L = H$. □

Using Lemma 3.13, it is easy to verify that if each $R^{(j)}$, $j \geq 0$, in (3.246) is Hermitian positive semi-definite and of complex-Hessian-type, then so is Z. Now,

from (3.245), we have

$$\text{vec}\left(R^{(j)}\right) = F^j \text{vec}(R) = FF^{j-1}\text{vec}(R) = F\text{vec}\left(R^{(j-1)}\right) \tag{3.250}$$

From (3.77) and using the property $\text{vec}(ABC) = (C^{\mathsf{T}} \otimes A)\text{vec}(B)$, we get the following recursion:

$$R^{(j)} = \sum_{k=1}^{N}\sum_{\ell=1}^{N} p_{\ell,k}[\bar{D}_k R^{(j-1)} \bar{D}_\ell + c_{\mu,\ell,k} H_k R^{(j-1)} H_\ell] \tag{3.251}$$

We can now verify by mathematical induction that each $R^{(j)}$ is Hermitian positive semi-definite and of complex-Hessian-type. Obviously, from (3.83), we know that $R^{(0)} = R$ is Hermitian positive semi-definite and of complex-Hessian-type. Now, assuming that $R^{(j-1)}$ is Hermitian positive semi-definite and of complex-Hessian-type, let us verify that the same applies to $R^{(j)}$.

Since $\{\bar{D}_k, H_k\}$ are all Hermitian matrices, it is easy to verify that

$$\sum_{k=1}^{N}\sum_{\ell=1}^{N} p_{\ell,k}\bar{D}_k R^{(j-1)}\bar{D}_\ell = (\mathbb{1}_N \otimes I_{2M})^* \bar{\mathcal{D}}^* [P_p \otimes R^{(j-1)}]\bar{\mathcal{D}}(\mathbb{1}_N \otimes I_{2M}) \tag{3.252}$$

and

$$\sum_{k=1}^{N}\sum_{\ell=1}^{N} p_{\ell,k} c_{\mu,\ell,k} H_k R^{(j-1)} H_\ell = (\mathbb{1}_N \otimes I_{2M})^* \mathcal{H}^* [(P_p \odot C_\mu) \otimes R^{(j-1)}]\mathcal{H}(\mathbb{1}_N \otimes I_{2M}) \tag{3.253}$$

where $\bar{\mathcal{D}}$ is from (3.28), $\mathcal{H}$ is from (3.19), $C_\mu \triangleq [c_{\mu,\ell,k}]_{\ell,k=1}^N$, and $\odot$ denotes the Hadamard product (the element-wise product) of matrices [89]. Since P_p is Hermitian positive semi-definite by Lemma 3.3, and $R^{(j-1)}$ is also Hermitian positive semi-definite by the induction hypothesis, the Kronecker product $P_p \otimes R^{(j-1)}$ must be Hermitian positive semi-definite [89, p. 245]. Therefore, the term on the LHS of (3.252) must be Hermitian positive semi-definite. From (2.42) of Chapter 2, it is obvious that the C_μ is the covariance matrix of $\{\boldsymbol{\mu}_k(i)\}$ and it must be Hermitian positive semi-definite. Since P_p and C_μ are both Hermitian

positive semi-definite, the Hadamard product $P_p \odot C_\mu$ is also Hermitian positive semi-definite by the Schur product Theorem [89, p. 309]. Then, the Kronecker product $(P_p \odot C_\mu) \otimes R^{(j-1)}$ must be Hermitian positive semi-definite [89, p. 245], and in turn, the term on the LHS of (3.253) must also be Hermitian positive semi-definite. From (3.252) and (3.253), we conclude that the $R^{(j)}$ in (3.251) must be Hermitian positive semi-definite.

Finally, we show that if $R^{(j-1)}$ is of complex-Hessian-type, then so is $R^{(j)}$. It is easy to verify that $\{\bar{D}_k\}$ in (3.27) and $\{H_k\}$ in (3.10) are all of complex-Hessian-type such that

$$L\bar{D}_k L = \bar{D}_k, \quad LH_k L = H_k, \quad k = 1, 2, \ldots, N \tag{3.254}$$

From (3.245), we have

$$\begin{aligned} LR^{(j)}L &\overset{(a)}{=} \sum_{k=1}^{N}\sum_{\ell=1}^{N} p_{\ell,k}[(L\bar{D}_k L)(LR^{(j-1)}L)(L\bar{D}_\ell L) \\ &\qquad + c_{\mu,\ell,k}(LH_k L)(LR^{(j-1)}L)(LH_\ell L)] \\ &\overset{(b)}{=} \sum_{k=1}^{N}\sum_{\ell=1}^{N} p_{\ell,k}[\bar{D}_k R^{(j-1)}\bar{D}_\ell + c_{\mu,\ell,k}H_k R^{(j-1)}H_\ell] \\ &= R^{(j)} \end{aligned} \tag{3.255}$$

where step (a) is by the fact that $LL = I_{2M}$ and step (b) is by (3.254) and the induction hypothesis. Therefore, the matrix $R^{(j)}$ is also of complex-Hessian-type.

3.L Proof of Corollary 3.2

Following an argument similar to the proof of Theorem 3.3, we can obtain

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 = \lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}'_{k,i}\|^2 + O(\nu^{3/2}) \tag{3.256}$$

From (3.64), (3.256), and Corollary 3.1, and using the fact that $\gamma_o = \min\{1/2, \gamma_o'\}$, we can express the individual MSD by

$$\mathrm{MSD}_k = \frac{1}{2}\mathrm{Tr}(Z) + O(\nu^{1+\gamma_o}) \tag{3.257}$$

where the first term on the RHS is of the order of ν and dominates the other term. Then, by (3.61), we immediately get

$$\mathrm{MSD}^{\mathrm{net}} = \frac{1}{N}\sum_{k=1}^{N}\mathrm{MSD}_k = \frac{1}{2}\mathrm{Tr}(Z) + O(\nu^{1+\gamma_o}) \tag{3.258}$$

Let

$$S \triangleq H^{\mathsf{T}} \otimes I_{2M} + I_{2M} \otimes H = O(\nu) \tag{3.259}$$

$$Y \triangleq \sum_{\ell,k=1}^{N} p_{\ell,k}(\bar{\mu}_\ell \bar{\mu}_k + c_{\mu,\ell,k})(H_\ell^{\mathsf{T}} \otimes H_k) = O(\nu^2) \tag{3.260}$$

by (3.181)–(3.183) from Chapter 2, where H is given by (3.79). It is worth noting that S is invertible when condition (3.14) holds and Y is always invertible by Assumption 2.2 in Chapter 2. By using (3.195), (3.199), (3.259), and (3.260), we get

$$I_{4M^2} - F = S + Y \tag{3.261}$$

Using the matrix inversion lemma [87], we get from (3.261) that

$$(I_{4M^2} - F)^{-1} = S^{-1} - S^{-1}(Y^{-1} + S^{-1})^{-1}S^{-1} \tag{3.262}$$

By (3.259) and (3.260), we know that $\|S^{-1}\| = O(\nu^{-1})$ and $\|Y^{-1}\| = O(\nu^{-2})$. Then,

$$(I_{4M^2} - F)^{-1} = S^{-1} + O(1) \tag{3.263}$$

By (3.83), we have $\|R\| = O(\nu^2)$. Using (3.84) and (3.263), we get

$$\mathrm{Tr}(Z) = [\mathrm{vec}(Z)]^* \mathrm{vec}(I_{2M})$$

$$= [\mathrm{vec}(R)]^*(I_{4M^2} - F)^{-1}\mathrm{vec}(I_{2M})$$
$$= [\mathrm{vec}(R)]^* S^{-1}\mathrm{vec}(I_{2M}) + O(\nu^2) \tag{3.264}$$

where the first term on the RHS is of the order of ν and, therefore, is the dominant term. To further simplify (3.264), we consider the Lyapunov equation with respect to the unknown matrix $X \in \mathbb{C}^{2M\times 2M}$: $XH + HX = I_{2M}$, where H is given by (3.79). By applying the $\mathrm{vec}(\cdot)$ operation to both sides, the Lyapunov equation is equivalent to the linear equation: $S \cdot \mathrm{vec}(X) = \mathrm{vec}(I_{2M})$, where $\mathrm{vec}(X) \in \mathbb{C}^{4M^2\times 1}$. Since S is invertible, the Lyapunov equation has a unique solution, which is given by $X = \frac{1}{2}H^{-1}$, or $\mathrm{vec}(X) = S^{-1}\mathrm{vec}(I_{2M}) = \frac{1}{2}\mathrm{vec}(H^{-1})$. It then follows from (3.264) that

$$\mathrm{Tr}(Z) = \frac{1}{2}\mathrm{Tr}(H^{-1}R) + O(\nu^2) \tag{3.265}$$

where the first term on the RHS is of the order of ν and dominates the other term. Substituting (3.265) into (3.257) and (3.258) completes the proof.

CHAPTER 4

Comparison with Synchronous Networks and Batch Implementations

In this chapter, we compare the performance of synchronous and asynchronous networks. We also compare the performance of decentralized adaptation against centralized stochastic-gradient (batch) solutions. Two interesting conclusions stand out. First, the results establish that the performance of adaptive networks is largely immune to the effect of asynchronous events: the mean and mean-square convergence rates and the asymptotic bias values are not degraded relative to synchronous or centralized implementations. Only the steady-state mean-square-deviation suffers a degradation in the order of ν, which represents the small step-size parameters used for adaptation. Second, the results show that the adaptive distributed network matches the performance of the centralized solution. These conclusions highlight another critical benefit of cooperation by networked agents: cooperation does not only enhance performance in comparison to stand-alone single-agent processing, but it also endows the network with remarkable resilience to various forms of random failure events and is able to deliver performance that is as powerful as batch solutions.

4.1 Centralized Batch Solution

We first describe and examine the centralized (batch) solution. In order to allow for a fair comparison among the various implementations, we assume that the centralized solution is also running a stochastic-gradient approximation algorithm albeit one that has access to the *entire* set of data at each iteration. Obviously, centralized solutions can be more powerful and run more complex optimization procedures. Our purpose is to examine the various implementations under similar algorithmic structures and complexity. The results in this chapter are based on material from [90].

4.1.1 Centralized Solution in Two Forms

We thus consider a scenario where there is a fusion center that regularly collects the data from across the network and is interested in solving the same minimization problem (1.1) as in Chapter 1. The fusion center seeks the optimal solution w^o of (1.1) by running an *asynchronous* stochastic gradient batch algorithm of the following form (later in (4.60) we consider a synchronous version of this batch solution):

$$\boldsymbol{w}_{\mathrm{c},i} = \boldsymbol{w}_{\mathrm{c},i-1} - \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\boldsymbol{\mu}_k(i)\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{\mathrm{c},i-1}) \tag{4.1}$$

where $\boldsymbol{w}_{\mathrm{c},i}$ denotes the iterate at time i, the $\{\boldsymbol{\pi}_k(i)\}$ are nonnegative convex fusion coefficients such that

$$\sum_{k=1}^{N} \boldsymbol{\pi}_k(i) = 1, \qquad \boldsymbol{\pi}_k(i) \geq 0 \tag{4.2}$$

for all $i \geq 0$, and the $\{\boldsymbol{\mu}_k(i)\}$ are the random step-sizes. In the batch algorithm (4.1), we use *random* step-sizes $\{\boldsymbol{\mu}_k(i)\}$ to account for random activity by the agents, which may be caused by random data arrival times or by some power saving strategies that turn agents on and off randomly. We also use *random*

fusion coefficients $\{\boldsymbol{\pi}_k(i)\}$ to model the random status of the communication links connecting the agents to the fusion center. This source of randomness may be caused by random fading effects over the communication channels or by random data feeding/fetching strategies. Therefore, the batch algorithm described in (4.1) is able to accommodate various forms of asynchronous events as well.

It is worth noting that the centralized (batch) algorithm (4.1) admits a decentralized, though not fully-distributed, implementation of the following form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{\mathrm{c},i-1} - \boldsymbol{\mu}_k(i)\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{\mathrm{c},i-1}) \quad \text{(adaptation)} \tag{4.3a}$$

$$\boldsymbol{w}_{\mathrm{c},i} = \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\, \boldsymbol{\psi}_{k,i} \qquad \text{(fusion)} \tag{4.3b}$$

In this description, each agent k uses the local gradient data to calculate the intermediate iterate $\boldsymbol{\psi}_{k,i}$ and feeds its value to a fusion center; the fusion center fuses all intermediate updates $\{\boldsymbol{\psi}_{k,i}\}$ according to (4.3b) to obtain $\boldsymbol{w}_{\mathrm{c},i}$ and then forwards the results to all agents. This process repeats itself at every iteration. Implementation (4.3a)–(4.3b) is not fully distributed because, for example, all agents require knowledge of the same global iterate $\boldsymbol{w}_{\mathrm{c},i}$ to perform the adaptation step (4.3a). Since the one-step centralized implementation (4.1) and the two-step equivalent (4.3a)–(4.3b) represent the same algorithm, we shall use them interchangeably to facilitate the analysis whenever necessary. One advantage of the decentralized representation (4.3a)–(4.3b) is that it can be viewed as a distributed solution over *fully*-connected networks [45].

4.1.2 Gradient Noise and Asynchronous Models

We assume that the approximate gradient vector $\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{\mathrm{c},i-1})$ in (4.1) follows the same model described by (2.17) in Chapter 2, namely,

$$\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{\mathrm{c},i-1}) = \nabla_{w^*} J_k(\boldsymbol{w}_{\mathrm{c},i-1}) + \boldsymbol{v}_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1}) \tag{4.4}$$

where the first term on the RHS is the true gradient and the second term models the uncertainty about the true gradient. We continue to assume that the gradient noise $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1})$ satisfies Assumption 3.1 from Chapter 3.

From Assumption 3.1 of Chapter 3, the conditional moments of $\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})$ satisfy

$$\mathbb{E}[\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})|\mathbb{F}_{i-1}] = 0 \tag{4.5}$$

$$\mathbb{E}[\|\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4|\mathbb{F}_{i-1}] \leq \alpha^2\|\underline{w}^o - \underline{\boldsymbol{w}}_{k,i-1}\|^4 + 4\sigma_v^4 \tag{4.6}$$

where a factor of 4 appeared due to the transform $\mathbb{T}(\cdot)$ from (2.3) of Chapter 2.

To facilitate the comparison in the sequel, we further assume the following asynchronous model for the centralized batch solution (4.1):

1. The random step-sizes $\{\boldsymbol{\mu}_k(i)\}$ satisfy the same properties as the asynchronous model for distributed diffusion networks described in Section 2.2.2 of Chapter 2. In particular, the first and second-order moments of $\{\boldsymbol{\mu}_k(i)\}$ are constant and denoted by

$$\bar{\mu}_k \triangleq \mathbb{E}\,[\boldsymbol{\mu}_k(i)] \tag{4.7}$$

$$c_{\mu,k,\ell} \triangleq \mathbb{E}[(\boldsymbol{\mu}_k(i) - \bar{\mu}_k)(\boldsymbol{\mu}_\ell(i) - \bar{\mu}_\ell)] \tag{4.8}$$

for all k, ℓ, and $i \geq 0$, where the values of these moments are the same as those in (2.33) and (2.36) from Chapter 2.

2. The random fusion coefficients $\{\boldsymbol{\pi}_k(i)\}$ satisfy condition (4.2) at every iteration i. Moreover, the first and second-order moments of $\{\boldsymbol{\pi}_k(i)\}$ are denoted by

$$\bar{\pi}_k \triangleq \mathbb{E}\,[\boldsymbol{\pi}_k(i)] \tag{4.9}$$

$$c_{\pi,k,\ell} \triangleq \mathbb{E}[(\boldsymbol{\pi}_k(i)-\bar{\pi}_k)(\boldsymbol{\pi}_\ell(i)-\bar{\pi}_\ell)] \tag{4.10}$$

for all k, ℓ, and $i \geq 0$.

3. The random parameters $\{\boldsymbol{\mu}_k(i)\}$ and $\{\boldsymbol{\pi}_k(i)\}$ are mutually-independent and independent of any other random variable.

We collect the fusion coefficients into the vector:

$$\boldsymbol{\pi}_i \triangleq \mathrm{col}\{\boldsymbol{\pi}_1(i), \boldsymbol{\pi}_2(i), \ldots, \boldsymbol{\pi}_N(i)\} \tag{4.11}$$

Then, condition (4.2) implies that $\boldsymbol{\pi}_i^{\mathsf{T}}\mathbb{1}_N = 1$. By (4.9) and (4.10), the mean and covariance matrix of $\boldsymbol{\pi}_i$ are given by

$$\bar{\pi} \triangleq \mathbb{E}\,(\boldsymbol{\pi}_i) = \mathrm{col}\{\bar{\pi}_1, \bar{\pi}_2, \ldots, \bar{\pi}_N\} \tag{4.12}$$

$$C_\pi \triangleq \mathbb{E}\,[(\boldsymbol{\pi}_i-\bar{\pi})(\boldsymbol{\pi}_i-\bar{\pi})^{\mathsf{T}}] = \begin{bmatrix} c_{\pi,1,1} & \cdots & c_{\pi,1,N} \\ \vdots & \ddots & \vdots \\ c_{\pi,N,1} & \cdots & c_{\pi,N,N} \end{bmatrix} \tag{4.13}$$

Lemma 4.1 (Properties of moments of $\{\boldsymbol{\pi}_k(i)\}$)**.** *The first and second-order moments of* $\{\boldsymbol{\pi}_k(i)\}$ *defined in* (4.9) *and* (4.10) *satisfy*

$$\sum_{k=1}^{N}\bar{\pi}_k = 1, \quad \bar{\pi}_k \geq 0, \quad \sum_{k=1}^{N} c_{\pi,k,\ell} = 0, \quad \sum_{\ell=1}^{N} c_{\pi,k,\ell} = 0 \tag{4.14}$$

for any k *and* ℓ*.*

Proof. Using (4.12) and (4.13) and the fact that C_π is symmetric, conditions (4.14) require

$$\bar{\pi}^{\mathsf{T}}\mathbb{1}_N = 1, \qquad C_\pi \mathbb{1}_N = 0 \tag{4.15}$$

The first equation in (4.15) is straightforward from (4.2). The second condition in (4.15) is true because

$$C_\pi \mathbb{1}_N = [\mathbb{E}(\boldsymbol{\pi}_i \boldsymbol{\pi}_i^{\mathsf{T}} | \boldsymbol{w}_{\mathrm{c},i-1}) - \bar{\pi}\bar{\pi}^{\mathsf{T}}]\mathbb{1}_N = 0 \tag{4.16}$$

where we used the fact that $\boldsymbol{\pi}_i^{\mathsf{T}}\mathbb{1}_N = 1$ and $\bar{\pi}^{\mathsf{T}}\mathbb{1}_N = 1$. □

We next examine the stability and steady-state performance of the asynchronous batch algorithm (4.1), and then compare its performance with that of the asynchronous distributed diffusion strategy.

4.2 Performance of the Centralized Solution

Following an argument similar to that given in Section 2.3 of Chapter 2, we can derive from (4.3a)–(4.3b) the following error recursion for the asynchronous centralized implementation:

$$\widetilde{\underline{\boldsymbol{\psi}}}_{k,i} = [I_{2M} - \boldsymbol{\mu}_k(i)\boldsymbol{H}_{k,i-1}]\widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i-1} + \underline{\boldsymbol{s}}_{k,i} \tag{4.17a}$$

$$\widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i} = \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\, \widetilde{\underline{\boldsymbol{\psi}}}_{k,i} \tag{4.17b}$$

where

$$\widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i} \triangleq \mathbb{T}(\widetilde{\boldsymbol{w}}_{\mathrm{c},i}) \tag{4.18}$$

$$\widetilde{\underline{\boldsymbol{\psi}}}_{k,i} \triangleq \mathbb{T}(\widetilde{\boldsymbol{\psi}}_{k,i}) \tag{4.19}$$

$$\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1}) \triangleq \mathbb{T}(\boldsymbol{v}_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1})) \tag{4.20}$$

and the mapping $\mathbb{T}(\cdot)$ is from (2.3) in Chapter 2. Moreover,

$$\boldsymbol{H}_{k,i-1} \triangleq \int_0^1 \nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}^o - t\widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i-1})\, dt \tag{4.21}$$

$$\underline{\boldsymbol{s}}_{k,i} \triangleq \boldsymbol{\mu}_k(i)\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1}) \tag{4.22}$$

We can merge (4.17a) and (4.17b) to find that the error dynamics of (4.1) evolves according to the following recursion:

$$\widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i} = \left[I_{2M} - \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\boldsymbol{\mu}_k(i)\boldsymbol{H}_{k,i-1}\right] \widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i-1} + \underline{\boldsymbol{s}}_i \tag{4.23}$$

where

$$\underline{\boldsymbol{s}}_i \triangleq \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\underline{\boldsymbol{s}}_{k,i} = \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\boldsymbol{\mu}_k(i)\underline{\boldsymbol{v}}_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1}) \tag{4.24}$$

4.2.1 Mean-Square and Mean-Fourth-Order Stability

To maintain consistency with the notation used in Chapters 2 and 3, we shall employ the same auxiliary quantities in these parts for the centralized batch solution (4.1) with minor adjustments whenever necessary. For example, the error quantity $\widetilde{\boldsymbol{w}}_{k,i}$ used before in Chapter 2 for the error vector at agent k at time i in the distributed implementation is now replaced by $\widetilde{\boldsymbol{w}}_{\mathrm{c},i}$, with a subscript c, for the error vector of the centralized solution at time i. Thus, we let

$$\epsilon^2(i) \triangleq \mathbb{E}\|\widetilde{\boldsymbol{w}}_{\mathrm{c},i}\|^2 = \frac{1}{2}\mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i}\|^2 \tag{4.25}$$

denote the MSD for the centralized solution $\widetilde{\boldsymbol{w}}_{\mathrm{c},i}$.

Theorem 4.1 (Mean-square stability)**.** *The mean-square stability of the asynchronous centralized implementation* (4.1) *reduces to studying the convergence of the recursive inequality:*

$$\epsilon^2(i) \leq \beta \cdot \epsilon^2(i-1) + \theta\sigma_v^2 \tag{4.26}$$

where the parameters $\{\beta, \theta, \sigma_v^2\}$ *are from* (2.82), (2.83), *and* (2.24) *in Chapter 2, respectively. The model* (4.26) *is stable if condition*

$$\boxed{\frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{\lambda_{k,\max}^2 + \alpha}} \tag{4.27}$$

holds for all k, *where the parameters* $\{\lambda_{k,\min}, \lambda_{k,\max}, \alpha\}$ *are from Assumptions 2.2 and 2.3 of Chapter 2. When condition* (4.27) *holds, an upper bound on the steady-state MSD is given by*

$$\boxed{\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \leq b \cdot \nu} \tag{4.28}$$

where ν *is given by* (3.16) *from Chapter 3 and* b *is a constant defined by* (2.87) *from Chapter 2.*

Proof. Since the centralized solution (4.1), or, equivalently, (4.3a)–(4.3b), can be viewed as a distributed solution over *fully*-connected networks [45], Theorem 2.1 from Chapter 2 can be applied directly. The result then follows from the fact that $\nu_o \leq \nu$ by (2.99) in Chapter 2. □

Comparing the above result to Theorem 2.1 in Chapter 2, we observe that the mean-square stability of the centralized solution (4.1) and the distributed asynchronous solution (2.27a)–(2.27b) from Chapter 2 is governed by the same model (4.26). Therefore, the same condition (4.27) guarantees the stability for both strategies and leads to the same MSD bound (4.28).

Theorem 4.2 (Stability of fourth-order error moment)**.** *If*

$$\boxed{\frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha}} \tag{4.29}$$

holds for all k, then the fourth-order moment of the error $\widetilde{\boldsymbol{w}}_{c,i}$ is asymptotically bounded by

$$\boxed{\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^4 \leq b_4^2 \cdot \nu^2} \tag{4.30}$$

where the parameter ν is given by (3.16) *from Chapter 3, and b_4 is a constant defined by* (2.97) *of Chapter 2.*

Proof. This result follows from Theorem 2.2 of Chapter 2 because the centralized solution (4.1), or, equivalently, (4.3a)–(4.3b), can be viewed as a distributed solution over *fully*-connected networks [45]. □

An alternative method to investigate the stability conditions for the centralized solution (4.1) is to view it as a stochastic gradient descent iteration for a *standalone* agent (i.e., a singleton network with $N = 1$) [8, 34, 35].

4.2.2 Long Term Error Dynamics

Using an argument similar to the one in Section 3.1.1 from Chapter 3, the original error recursion (4.23) can be rewritten as

$$\widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i} = \left[I_{2M} - \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\boldsymbol{\mu}_k(i)H_k\right] \widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i-1} + \underline{\boldsymbol{s}}_i + \boldsymbol{d}_i \tag{4.31}$$

where

$$\boldsymbol{d}_i \triangleq \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\boldsymbol{\mu}_k(i)(H_k - \boldsymbol{H}_{k,i-1})\widetilde{\underline{\boldsymbol{w}}}_{\mathrm{c},i-1} \tag{4.32}$$

Then, under condition (4.29),

$$\limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{d}_i\|^2 \leq O(\nu^4) \tag{4.33}$$

where ν is given by (3.16) from Chapter 3.

Assumption 4.1 (Small step-sizes)**.** *The parameter ν from* (3.16) *in Chapter 3 is sufficiently small such that*

$$\nu < \min_k \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha} < 1 \tag{4.34}$$

□

Under Assumption 4.1, condition (4.29) holds. Let

$$\boldsymbol{B}_i \triangleq \sum_{k=1}^{N} \boldsymbol{\pi}_k(i) \boldsymbol{D}_{k,i} \tag{4.35}$$

$$\boldsymbol{D}_{k,i} \triangleq I_{2M} - \boldsymbol{\mu}_k(i) H_k \tag{4.36}$$

where $\boldsymbol{B}_i$ is Hermitian positive semi-definite. Since we are interested in examining the asymptotic performance of the asynchronous batch solution, we can again call upon the same argument from Section 3.1.1 of Chapter 3 and use result (4.33) to conclude that we can assess the performance of (4.31) by working with the following *long-term* model, which holds for large enough i:

$$\widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i} = \boldsymbol{B}_i \cdot \widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i-1} + \underline{\boldsymbol{s}}_i, \quad i \gg 1 \tag{4.37}$$

In model (4.37), we ignored the $O(\nu^2)$ term $\boldsymbol{d}_i$ according to (4.33), and we are using $\boldsymbol{w}'_{c,i}$ to denote the estimate obtained from this long-term model. Note that the driving noise term $\underline{\boldsymbol{s}}_i$ in (4.37) is extraneous and imported from the original error recursion (4.23).

Theorem 4.3 (Bounded mean-square gap)**.** *Under Assumption 4.1, the mean-square gap from the original error recursion* (4.23) *to the long-term model* (4.37) *is asymptotically bounded by*

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\underline{\boldsymbol{w}}}_{c,i} - \widetilde{\underline{\boldsymbol{w}}}'_{c,i}\|^2 \leq O(\nu^2) \tag{4.38}$$

where ν is given by (3.16) *from Chapter 3.*

Proof. This result follows from Theorem 3.1 of Chapter 3 since the centralized solution (4.1) can be viewed as a distributed solution over *fully*-connected networks [45]. □

4.2.3 Mean Error Recursion

By taking the expectation of both sides of (4.37), and using the fact that $\mathbb{E}(\underline{\boldsymbol{s}}_i) = 0$, we conclude that the mean error satisfies the recursion:

$$\mathbb{E}\,\widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i} = \bar{B} \cdot \mathbb{E}\,\widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i-1} \tag{4.39}$$

for large enough i, where

$$\bar{B} \triangleq \mathbb{E}(\boldsymbol{B}_i) = \sum_{k=1}^{N} \bar{\pi}_k \bar{D}_k \tag{4.40}$$

$$\bar{D}_k \triangleq \mathbb{E}(\boldsymbol{D}_{k,i}) = I_{2M} - \bar{\mu}_k H_k \tag{4.41}$$

The convergence of recursion (4.39) requires the stability of $\bar{B}$. It is easy to verify that $\{\bar{B}, \bar{D}_k\}$ are Hermitian. Using (4.14) and Jensen's inequality, we get from (4.40) that $\rho(\bar{B}) \leq \max_k \rho(\bar{D}_k)$. As we showed in (3.40) from Chapter 3, if condition (4.27) holds, then $\rho(\bar{D}_k) < 1$ for all k. Therefore, it follows from Assumption 4.1 that

$$\lim_{i\to\infty} \mathbb{E}\,\widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i} = 0 \tag{4.42}$$

which implies that the long-term model (4.37) is the asymptotically centered version of the original error recursion (4.23).

4.2.4 Error Covariance Recursion

Let $\mathbb{F}_{i-1}$ denote the filtration that represents all information available up to iteration $i-1$. Then we deduce from (4.37) that for large enough i:

$$\mathbb{E}(\widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i}\widetilde{\underline{\boldsymbol{w}}}'^{*}_{\mathrm{c},i}|\mathbb{F}_{i-1}) = \mathbb{E}(\boldsymbol{B}_i\widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i-1}\widetilde{\underline{\boldsymbol{w}}}'^{*}_{\mathrm{c},i-1}\boldsymbol{B}_i|\mathbb{F}_{i-1}) + \mathbb{E}(\underline{\boldsymbol{s}}_i\underline{\boldsymbol{s}}_i^{*}|\mathbb{F}_{i-1}) \tag{4.43}$$

where the cross terms that involve $\underline{\boldsymbol{s}}_i$ disappear because $\mathbb{E}(\underline{\boldsymbol{s}}_i^* \boldsymbol{B}_i \widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i-1} | \mathbb{F}_{i-1}) = 0$ by the gradient noise model from Assumption 3.1 of Chapter 3. Vectorizing both sides of (4.43) and taking expectation, we obtain

$$\mathbb{E}\,[(\widetilde{\underline{\boldsymbol{w}}}'^*_{\mathrm{c},i})^{\mathsf{T}} \otimes \widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i}] = F_{\mathrm{c}} \cdot \mathbb{E}\,[(\widetilde{\underline{\boldsymbol{w}}}'^*_{\mathrm{c},i-1})^{\mathsf{T}} \otimes \widetilde{\underline{\boldsymbol{w}}}'_{\mathrm{c},i-1}] + y_{\mathrm{c},i} \tag{4.44}$$

where

$$F_{\mathrm{c}} \triangleq \mathbb{E}\,[\boldsymbol{B}_i^{\mathsf{T}} \otimes \boldsymbol{B}_i] \tag{4.45}$$

$$y_{\mathrm{c},i} \triangleq \mathbb{E}[(\underline{\boldsymbol{s}}_i^*)^{\mathsf{T}} \otimes \underline{\boldsymbol{s}}_i] \tag{4.46}$$

Let further

$$H_{\mathrm{c}} \triangleq \sum_{k=1}^{N} \bar{\pi}_k \bar{\mu}_k H_k = O(\nu) \tag{4.47}$$

where $\{H_k\}$ are from (3.9) of Chapter 3.

Lemma 4.2 (Properties of F_{c})**.** *The matrix F_c defined by* (4.45) *is Hermitian and can be expressed as*

$$F_c = \sum_{\ell=1}^{N} \sum_{k=1}^{N} (\bar{\pi}_\ell \bar{\pi}_k + c_{\pi,\ell,k})(\bar{D}_\ell^{\mathsf{T}} \otimes \bar{D}_k + c_{\mu,\ell,k} H_\ell^{\mathsf{T}} \otimes H_k) \tag{4.48}$$

If condition (4.27) *holds, then F_c is stable and*

$$\rho(F_c) = [1 - \lambda_{\min}(H_c)]^2 + O(\nu^2) \tag{4.49}$$

where H_c is given by (4.47)*, and $[1 - \lambda_{\min}(H_c)]^2 = 1 - O(\nu)$ under Assumption 4.1. Moreover,*

$$\|(I_{4M^2} - F_c)^{-1}\| = O(\nu^{-1}) \tag{4.50}$$

Proof. See Appendix 4.A. □

Theorem 4.4 (Error covariance recursion)**.** *For sufficiently large i, the vectorized error covariance for the long-term model* (4.37) *satisfies the following relation:*

$$z_{c,i} = F_c \cdot z_{c,i-1} + y_{c,i}, \;\; i \gg 1 \tag{4.51}$$

where F_c and $y_{c,i}$ are from (4.45) *and* (4.46)*, respectively, and*

$$z_{c,i} \triangleq \mathbb{E}\,[(\widetilde{\boldsymbol{w}}_{c,i}'^{*})^{\mathsf{T}} \otimes \widetilde{\boldsymbol{w}}_{c,i}'] \tag{4.52}$$

Recursion (4.51) *is convergent if condition* (4.27) *holds, and its convergence rate is dominated by* $[1-\lambda_{\min}(H_c)]^2 = 1 - O(\nu)$ *under Assumption 4.1.*

Proof. Equation (4.51) follows from (4.44). Recursion (4.51) converges if, and only if, the matrix $F_{\rm c}$ is stable. By Lemma 4.2, we know that $\rho(F_{\rm c}) < 1$ if condition (4.27) holds and, moreover, the convergence rate of recursion (4.51) is determined by $\rho(F_{\rm c}) = [1-\lambda_{\min}(H_{\rm c})]^2 + O(\nu^2)$. □

4.2.5 Steady-State MSD

At steady-state as $i \to \infty$, we get from (4.50) and (4.51) that

$$z_{\rm c,\infty} \triangleq \mathrm{vec}\left(\lim_{i\to\infty} \mathbb{E}\widetilde{\boldsymbol{w}}_{{\rm c},i}'\widetilde{\boldsymbol{w}}_{{\rm c},i}'^{*}\right) = (I_{4M^2} - F_{\rm c})^{-1} \cdot \lim_{i\to\infty} y_{{\rm c},i} \tag{4.53}$$

Using $z_{\rm c,\infty}$, we can determine the value of any steady-state weighted mean-square-error metric for the long-term model (4.37) as follows:

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{{\rm c},i}'\|_{\Sigma}^2 = \frac{1}{2}\lim_{i\to\infty} \mathrm{Tr}[\mathbb{E}(\widetilde{\boldsymbol{w}}_{{\rm c},i}'\widetilde{\boldsymbol{w}}_{{\rm c},i}'^{*})\Sigma] = \frac{1}{2} z_{\rm c,\infty}^{*}\mathrm{vec}(\Sigma) \tag{4.54}$$

where we used the fact that $\mathrm{Tr}(AB) = [\mathrm{vec}(A^*)]^*\mathrm{vec}(B)$, and Σ is an arbitrary Hermitian positive semi-definite weighting matrix. The steady-state MSD for the original error recursion (4.31) is defined by

$$\mathrm{MSD}^{\rm cent} \triangleq \lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{{\rm c},i}\|^2 = \lim_{i\to\infty} \frac{1}{2}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{{\rm c},i}\|^2 \tag{4.55}$$

Therefore, by setting $\Sigma = I_{2M}$ in (4.54) and using Theorem 4.3, it is easy to verify by following an argument similar to the proof of Theorem 3.3 from Chapter 3 that

$$\mathrm{MSD}^{\rm cent} = \lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{{\rm c},i}'\|^2 + O(\nu^{3/2}) \tag{4.56}$$

Introduce

$$R_{\rm c} \triangleq \sum_{k=1}^{N} (\bar{\pi}_k^2 + c_{\pi,k,k})(\bar{\mu}_k^2 + c_{\mu,k,k}) R_k = O(\nu^2) \tag{4.57}$$

where $\{R_k\}$ are from (3.6) of Chapter 3. Then, using (4.54) and (4.56), we arrive the following result.

Theorem 4.5 (Steady-state MSD)**.** *The steady-state MSD for the asynchronous centralized (batch) solution* (4.1) *is given by*

$$MSD^{cent} = \frac{1}{2}[\mathrm{vec}(R_c)]^* (I_{4M^2} - F_c)^{-1} \mathrm{vec}(I_{2M}) + O(\nu^{1+\gamma_o}) \tag{4.58}$$

where $0 < \gamma_o \leq 1/2$ *is from* (3.65) *of Chapter 3. Expression* (4.58) *can be further reworked to yield*

$$MSD^{cent} = \frac{1}{4}\mathrm{Tr}(H_c^{-1} R_c) + O(\nu^{1+\gamma_o}) \tag{4.59}$$

where the first term on the RHS is in the order of ν *and therefore dominates the* $O(\nu^{1+\gamma_o})$ *term under Assumption 4.1.*

Proof. See Appendix 4.B. □

4.2.6 Results for the Synchronous Centralized Solution

We may also consider a synchronous centralized (batch) implementation for solving the same problem (1.1). It would take the following form:

$$\boldsymbol{w}_{{\rm c},i} = \boldsymbol{w}_{{\rm c},i-1} - \sum_{k=1}^{N} \pi_k \mu_k \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{{\rm c},i-1}) \tag{4.60}$$

where the $\{\mu_k\}$ are now deterministic nonnegative step-sizes and the $\{\pi_k\}$ are nonnegative fusion coefficients that satisfy $\sum_{k=1}^{N} \pi_k = 1$. The synchronous batch solution can be viewed as a special case of the asynchronous batch solution (4.1) when the random step-sizes and fusion coefficients assume constant values. If

the covariances $\{c_{\mu,k,k}\}$ and $\{c_{\pi,k,k}\}$ are set to zero, then the asynchronous solution (4.1) will reduce into a synchronous solution that employs the constant parameters $\{\bar{\mu}_k\}$ and $\{\bar{\pi}_k\}$. The previous stability and performance results can be specialized to the synchronous batch implementation under these conditions.

It is easy to verify that the mean error recursion for the synchronous solution with parameters $\{\bar{\mu}_k\}$ and $\{\bar{\pi}_k\}$ is identical to (4.39). The mean convergence rate for the long-term model is still determined by $\rho(\bar{B})$, where $\bar{B}$ is given by (4.40). The mean square convergence rate for the long-term model is determined by $\rho(F_{\rm c}')$ where

$$F_{\rm c}' \triangleq \sum_{k=1}^{N}\sum_{\ell=1}^{N}\bar{\pi}_\ell\bar{\pi}_k(\bar{D}_\ell^{\mathsf{T}}\otimes\bar{D}_k) \tag{4.61}$$

It follows that

$$\rho(F_{\rm c}') = [1-\lambda_{\min}(H_c)]^2 = 1-O(\nu) \tag{4.62}$$

The steady-state MSD is given by

$$\mathrm{MSD}_{\rm sync}^{\rm cent} = \frac{1}{4}\mathrm{Tr}(H_{\rm c}^{-1}R_{\rm c}') + O(\nu^{1+\gamma_o}) \tag{4.63}$$

where

$$R_{\rm c}' \triangleq \sum_{k=1}^{N}\bar{\pi}_k^2\bar{\mu}_k^2R_k, \quad \|R_{\rm c}'\| = O(\nu^2) \tag{4.64}$$

and $\mathrm{Tr}(H_{\rm c}^{-1}R_{\rm c}') = O(\nu)$.

4.3 Comparison I: Distributed vs. Centralized Strategies

In this section, we compare the mean-square performance of the distributed diffusion strategy (2.27a)–(2.27b) from Chapter 2, namely,

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \boldsymbol{\mu}_k(i)\widehat{\nabla_{w^*}J_k}(\boldsymbol{w}_{k,i-1}) \tag{4.65a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\boldsymbol{\mathcal{N}}_{k,i}}\boldsymbol{a}_{\ell k}(i)\,\boldsymbol{\psi}_{\ell,i} \tag{4.65b}$$

with the centralized (batch) solution described by (4.1). We establish the important conclusion that if the combination matrix is primitive (Assumption 3.3 in Chapter 3), then the asynchronous network is able to achieve almost the same mean-square performance as the centralized (batch) solution for sufficiently small step-sizes. In other words, diffusion strategies are *efficient* mechanisms to perform continuous adaptation and learning tasks over networks even in the presence of various sources of random failures.

4.3.1 Adjusting Relevant Parameters

First, however, we need to describe the conditions that are necessary for a *fair* and meaningful comparison between the distributed and centralized implementations. This is because the two implementations use different parameters. Recall that the agents in the distributed network (4.65a)–(4.65b) employ random combination coefficients $\{\boldsymbol{a}_{\ell k}(i)\}$ to aggregate information from neighborhoods using random step-sizes $\{\boldsymbol{\mu}_k(i)\}$. The random parameters $\{\boldsymbol{a}_{\ell k}(i), \boldsymbol{\mu}_k(i)\}$ are assumed to satisfy the model described in Section 2.2.2 from Chapter 2. On the other hand, the centralized batch solution (4.1) uses random combination coefficients $\{\boldsymbol{\pi}_k(i)\}$ to fuse the information from all agents in the network, and then performs updates using random step-sizes $\{\boldsymbol{\mu}_k(i)\}$. The random parameters $\{\boldsymbol{\pi}_k(i), \boldsymbol{\mu}_k(i)\}$ are assumed to satisfy the conditions specified in Section 4.1.2 of this part. In general, the two sets of random parameters, i.e., $\{\boldsymbol{a}_{\ell k}(i), \boldsymbol{\mu}_k(i)\}$ for distributed strategies and $\{\boldsymbol{\pi}_k(i), \boldsymbol{\mu}_k(i)\}$ for centralized strategies, are not necessarily related. Therefore, in order to make a meaningful comparison between the distributed and centralized strategies, we need to introduce connections between these two sets of parameters. This is possible because the parameters play similar roles.

From the previous analysis in Section 3.3 of Chapter 3, we know that the first

and second-order moments of $\{\boldsymbol{a}_{\ell k}(i), \boldsymbol{\mu}_k(i)\}$ determine the mean-square performance of diffusion networks. Likewise, from the analysis in Section 4.2 of this part, we know that the first and second-order moments of $\{\boldsymbol{\pi}_k(i), \boldsymbol{\mu}_k(i)\}$ determine the mean-square performance of centralized solutions. Therefore, it is sufficient to introduce connections between the first and second-order moments of these random parameters. For the random step-size parameters, we assumed in (2.33) and (2.36) from Chapter 2 and in (4.7) and (4.8) from this part that their first and second-order moments are *constant* and that their values coincide with each other, i.e., $\bar{\mu}_k$ from (2.33) in Chapter 2 coincides with $\bar{\mu}_k$ from (4.7) in this part, and similarly for $c_{\mu,k,\ell}$. This requirement is obviously reasonable.

The connection that we need to enforce between the moments of the combination coefficients $\{\boldsymbol{a}_{\ell k}(i)\}$ and $\{\boldsymbol{\pi}_k(i)\}$, while reasonable again, is less straightforward to explain. This is because the $\{\boldsymbol{a}_{\ell k}(i)\}$ form a random matrix $\boldsymbol{A}_i = [\boldsymbol{a}_{\ell k}(i)]_{k,\ell=1}^N$ of size $N \times N$, while the $\{\boldsymbol{\pi}_k(i)\}$ only form a random vector $\boldsymbol{\pi}_i = [\boldsymbol{\pi}_k(i)]_{k=1}^N$ of size $N \times 1$. From the result of Corollary 3.2 in Chapter 3 though, we know that the mean-square performance of the *primitive* diffusion network does not *directly* depend on the moments of $\boldsymbol{A}_i$, namely, its mean $\bar{A}$ and its Kronecker covariance C_A; instead, the performance depends on the Perron eigenvector (the unique right eigenvector corresponding to the eigenvalue at one for primitive left-stochastic matrices [79, 82]). If, for example, we compare expression (3.91) from Chapter 3 for asynchronous networks with expression (4.48) from this part, we conclude that it is sufficient to relate the vectors $\{\bar{p}, p\}$ defined in (3.72) and (3.73) from Chapter 3 to the moments $\{\bar{\pi}_k, c_{\pi,k,\ell}\}$. Since $\bar{p}$ is the Perron eigenvector of the mean matrix $\bar{A}$, and the $\{\bar{\pi}_k\}$ are the means of $\{\boldsymbol{\pi}_k(i)\}$, we connect them by requiring

$$\bar{\pi}_k \equiv \bar{p}_k \tag{4.66}$$

for all k, where the $\{\bar{p}_k\}$ are the elements of $\bar{p}$. Likewise, since p is the Perron eigenvector of the matrix $\bar{A} \otimes \bar{A} + C_A = \mathbb{E}\,(\boldsymbol{A}_i \otimes \boldsymbol{A}_i)$, which consists of the second-order moments, and $\{\bar{\pi}_k \bar{\pi}_\ell + c_{\pi,k,\ell} = \mathbb{E}\,[\boldsymbol{\pi}_k(i)\boldsymbol{\pi}_\ell(i)]\}$ are also the second-order moments, we connect them by requiring

$$\bar{\pi}_k \bar{\pi}_\ell + c_{\pi,k,\ell} \equiv p_{k,\ell} \tag{4.67}$$

for all k and ℓ, where the $\{p_{k,\ell}\}$ are the elements of p defined after (3.75) in Chapter 3. When conditions (4.66) and (4.67) are satisfied, then the mean-square convergence rates and steady-state MSD for the distributed and centralized solutions become identical. We establish this result in the sequel. Using (4.12) and (4.13), conditions (4.66) and (4.67) can be rewritten as

$$\bar{\pi} \equiv \bar{p}, \qquad C_\pi + \bar{\pi}\bar{\pi}^{\mathsf{T}} \equiv P_p \tag{4.68}$$

where

$$P_p = \begin{bmatrix} p_{1,1} & \cdots & p_{1,N} \\ \vdots & \ddots & \vdots \\ p_{N,1} & \cdots & p_{N,N} \end{bmatrix} \tag{4.69}$$

is the symmetric matrix defined by (3.75) of Chapter 3. It is worth noting that, since the Perron eigenvectors $p = \mathrm{vec}(P_p)$ and $\bar{p}$ consist of positive entries, the corresponding quantities $\bar{\pi}$ and $C_\pi + \bar{\pi}\bar{\pi}^{\mathsf{T}}$ must also consist of positive entries — we shall refer to the centralized solutions that satisfy this condition as *primitive* centralized solutions. Clearly, the second requirement in (4.68) is meaningful only if the difference $P_p - \bar{p}\bar{p}^{\mathsf{T}}$ results in a symmetric positive semi-definite matrix (and, hence, a covariance matrix) that also satisfies $C_\pi \mathbb{1}_N = 0$.

4.3.2 Constructing Primitive Batch Solutions

Before comparing the performance of the centralized and distributed solutions under (4.68), we first answer the following important inquiry. Given a distributed primitive network with parameters $\{\bar{p}, P_p\}$, is it possible to determine a batch solution with parameters $\{\bar{\pi}, C_\pi\}$ satisfying (4.68) such that the resulting C_π is a symmetric and positive semi-definite matrix (and, therefore, has the interpretation of a valid covariance matrix)? The answer is in the affirmative as we proceed to explain. The following are auxiliary results in this direction.

Lemma 4.3 (Positive semi-definite property)**.** *The matrix difference $P_p - \bar{p}\bar{p}^{\mathsf{T}}$ is symmetric positive semi-definite and satisfies $(P_p - \bar{p}\bar{p}^{\mathsf{T}})\mathbb{1}_N = 0$ for any $\bar{p}$ and P_p defined by* (3.73) *and* (3.75) *from Chapter 3.*

Proof. See Appendix 4.C. □

Therefore, starting from an asynchronous diffusion network with parameters $\{\bar{p}, P_p\}$, there exists an asynchronous batch solution with valid parameters $\{\bar{\pi}, C_\pi\}$ that satisfy (4.68). We now explain one way by which a random variable $\boldsymbol{\pi}_i$ can be constructed with the pre-specified moments $\{\bar{\pi}, C_\pi\}$. We first observe that in view of condition (4.11), the random variable $\boldsymbol{\pi}_i$ is actually defined on the probability simplex in $\mathbb{R}^{N\times 1}$ [71, p. 33]:

$$\Delta_N \triangleq \{x \in \mathbb{R}^{N\times 1}; x^{\mathsf{T}}\mathbb{1}_N = 1, x_k \geq 0, k = 1, \ldots, N\} \tag{4.70}$$

If the moments $\{\bar{\pi}, C_\pi\}$ obtained from (4.68) satisfy certain conditions, then there are several models in the literature that can be used to generate random vectors $\{\boldsymbol{\pi}_i\}$ according to these moments such as using the Dirichlet distribution [91], the Generalized Dirichlet distribution [92–99], the Logistic-Normal distribution [93,100,101], or the Generalized inverse Gaussian distribution [94,102].

Unfortunately, if the conditions for these models are not satisfied, no *closed-form* probabilistic model is available for us to generate random variables on the probability simplex with pre-specified means and covariance matrices.

Nevertheless, inspired by the Markov Chain Monte Carlo (MCMC) method [103], we describe one procedure to construct random variables *indirectly* so that they are able to meet the desired moment requirements. In a manner similar to the argument used in Appendix 4.C, we introduce a series of fictitious random combination matrices $\{\boldsymbol{A}'_j; j \geq 1\}$ that satisfy the asynchronous model introduced in Chapter 2. We assume that the $\{\boldsymbol{A}'_j; j \geq 1\}$ are independently, identically distributed (i.i.d.) random matrices, and they are independent of any other random variable. Then, the mean and Kronecker-covariance matrices of $\boldsymbol{A}'_j$ for any j are given by $\bar{A}$ and C_A, respectively. We further introduce the random matrix

$$\boldsymbol{\Phi}_{i,t} \triangleq \prod_{j=1}^{t} \boldsymbol{A}'_j \tag{4.71}$$

Similar to (4.155) and (4.157), we can verify that

$$\lim_{t\to\infty} \mathbb{E}(\boldsymbol{\Phi}_{i,t}) = \bar{p}\mathbb{1}_N^{\mathsf{T}}, \quad \lim_{t\to\infty} \mathbb{E}(\boldsymbol{\Phi}_{i,t} \otimes \boldsymbol{\Phi}_{i,t}) = p\mathbb{1}_{N^2}^{\mathsf{T}} \tag{4.72}$$

Let

$$\boldsymbol{\phi}_i \triangleq \frac{1}{N}\left(\lim_{t\to\infty} \boldsymbol{\Phi}_{i,t}\right)\mathbb{1}_N \tag{4.73}$$

Then, the entries of $\boldsymbol{\phi}_i$ are nonnegative since the entries of $\boldsymbol{\Phi}_{i,t}$ are nonnegative. Using (4.71) and (4.73), we have

$$\mathbb{1}_N^{\mathsf{T}}\boldsymbol{\phi}_i = \frac{1}{N}\lim_{t\to\infty} \mathbb{1}_N^{\mathsf{T}}\left(\prod_{j=1}^{t} \boldsymbol{A}'_j\right)\mathbb{1}_N = 1 \tag{4.74}$$

since each $\boldsymbol{A}'_j$ is left-stochastic. Therefore, $\boldsymbol{\phi}_i$ is a random variable defined on the probability simplex Δ_N. By using (4.72) and the fact that $\mathbb{1}_N \otimes \mathbb{1}_N = \mathbb{1}_{N^2}$,

we have

$$\mathbb{E}(\boldsymbol{\phi}_i) = \frac{1}{N}(\bar{p} \cdot \mathbb{1}_N^{\mathsf{T}})\mathbb{1}_N = \bar{p} \tag{4.75}$$

$$\mathbb{E}(\boldsymbol{\phi}_i \otimes \boldsymbol{\phi}_i) = \frac{1}{N^2}(p \cdot \mathbb{1}_{N^2}^{\mathsf{T}})\mathbb{1}_{N^2} = p \tag{4.76}$$

Therefore,

$$\mathbb{E}(\boldsymbol{\phi}_i) = \bar{p}, \qquad \mathrm{Cov}(\boldsymbol{\phi}_i) = P_p - \bar{p}\bar{p}^{\mathsf{T}} \tag{4.77}$$

where $P_p = \mathrm{unvec}(p)$. In this way, we have been able to construct a random variable $\boldsymbol{\phi}_i$ whose support is the probability simplex Δ_N and whose mean vector and covariance matrix match the specification. The random variable $\boldsymbol{\phi}_i$ can then be used by the asynchronous centralized solution at time i, which would then enable a meaningful comparison with the asynchronous distributed solution.

Although unnecessary for our development, it is instructive to pose the converse question: Given a *primitive* batch solution with parameters $\{\bar{\pi}, C_\pi\}$, is it always possible to determine a distributed solution with parameters $\{\bar{p}, P_p\}$ satisfying (4.68) such that these parameters have the properties of Perron eigenvectors? In other words, given a primitive centralized solution, is it possible to determine a distributed solution on a *partially*-connected network (otherwise the problem is trivial since fully-connected networks are equivalent to centralized solutions [45]) with equivalent performance levels? The answer to this question remains open. The challenge stems from the fact mentioned earlier that, in general, there is no systematic solution to generate distributions on the probability simplex with pre-specified first and second-order moments. The method of moments [104], which is an iterative solution, does not generally guarantee convergence and therefore, cannot ensure that a satisfactory distribution can be generated eventually.

4.3.3 Comparing Performance

From the mean error recursion in (3.37) of Chapter 3, the mean convergence rate for the long-term model of the distributed diffusion strategy is determined by $\rho(\bar{\mathcal{B}})$, where $\bar{\mathcal{B}}$ is defined by (3.29) of Chapter 3. From the mean error recursion (4.39) in this part, the mean convergence rate for the long-term model of the centralized batch solution is determined by $\rho(\bar{B})$, where $\bar{B}$ is given by (4.40).

Lemma 4.4 (Matching mean convergence rates)**.** *The mean convergence rates for the asynchronous distributed strategy and the centralized batch solution are almost the same. Specifically, it holds that*

$$|\rho(\bar{\mathcal{B}}) - \rho(\bar{B})| \leq O(\nu^{1+1/N}) \tag{4.78}$$

where $\rho(\bar{\mathcal{B}})$ and $\rho(\bar{B})$ are of the order of $1 - O(\nu)$.

Proof. See Appendix 4.D. □

Likewise, from Theorem 3.2 of Chapter 3, the mean-square convergence rate of the distributed diffusion strategy for large enough i is determined by $\rho(\mathcal{F})$, where $\mathcal{F}$ is from (3.46) of Chapter 3. From Theorem 4.4 of this part, the mean-square convergence rate of the centralized (batch) solution is determined by $\rho(F_{\rm c})$, where $F_{\rm c}$ is from (4.45).

Lemma 4.5 (Matching mean-square convergence rates)**.** *The mean-square convergence rates for the asynchronous distributed strategy and the centralized batch solution are almost the same. Specifically, it holds that*

$$|\rho(\mathcal{F}) - \rho(F_c)| \leq O(\nu^{1+1/N^2}) \tag{4.79}$$

where $\rho(\mathcal{F})$ and $\rho(F_c)$ are of the order of $1 - O(\nu)$.

Proof. From (4.45) and (4.67), it is easy to verify that $F_{\rm c} = F$, where F is from (3.77) of Chapter 3. Using Lemmas 3.4 and 3.5 from Chapter 3 then completes the proof. □

The steady-state network MSD for the distributed diffusion strategy is given by (3.91) of Chapter 3:

$$\mathrm{MSD}^{\rm dist} = \frac{1}{4}\mathrm{Tr}(H^{-1}R) + O(\nu^{1+\gamma_o}) \tag{4.80}$$

for some $0 < \gamma_o \leq 1/2$ given by (3.65) of Chapter 3. The steady-state MSD for the centralized batch solution is given by (4.59).

Lemma 4.6 (Matching MSD performance)**.** *The network MSD for the asynchronous distributed strategy and the MSD for the centralized batch solution are close to each other in steady-state. Specifically, we have*

$$|MSD^{dist} - MSD^{cent}| \leq O(\nu^{1+\gamma_o}) \tag{4.81}$$

where both MSD^{dist} and MSD^{cent} are in the order of ν.

Proof. From (4.47), (4.57), and (4.67), it is easy to verify that $H_{\rm c} = H$ and $R_{\rm c} = R$, where $\{H, R\}$ are given by (3.79) and (3.83) of Chapter 3. Using (4.59) and (4.80) then completes the proof. □

4.4 Comparison II: Asynchronous vs. Synchronous Networks

Synchronous diffusion networks run (2.15a)–(2.15b) from Chapter 2, namely,

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}) \quad \text{(adaptation)} \tag{4.82a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k}\, \boldsymbol{\psi}_{\ell,i} \qquad \text{(combination)} \tag{4.82b}$$

These networks can be viewed as a special case of asynchronous networks running (4.65a)–(4.65b) when the random step-sizes and combination coefficients assume constant values. If we set the covariances $\{c_{\mu,k,k}\}$ and $\{c_{a,\ell k,\ell k}\}$ to zero, then the asynchronous network (4.65a)–(4.65b) will reduce to the synchronous network (4.82a)–(4.82b) with the parameters $\{\mu_k, a_{\ell k}\}$ replaced by $\{\bar{\mu}_k, \bar{a}_{\ell k}\}$. We can therefore specialize the results obtained for asynchronous networks to the synchronous case by using $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$ and assuming $c_{\mu,k,k} = 0$ and $c_{a,\ell k,\ell k} = 0$ for all k and ℓ. For example, it is easy to verify that the mean error recursion of the long term model for the synchronous solution with $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$ is identical to (3.37) from Chapter 3 for the asynchronous solution.

Under Assumption 4.1, the asynchronous network with the random parameters $\{\boldsymbol{\mu}_k(i)\}$ and $\{\boldsymbol{a}_{\ell k}(i)\}$ and the synchronous network with the constant parameters $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$ have similar mean-square convergence rates for large i, but the steady-state MSD performance of the former is larger than that of the latter by a small amount. This result is established as follows. From Theorem 3.2 in Chapter 3, the mean-square convergence rate for the asynchronous network with large i is determined by $\rho(\mathcal{F}_{\text{async}})$ where

$$\mathcal{F}_{\text{async}} = \mathbb{E}(\boldsymbol{\mathcal{B}}_i^{\mathsf{T}} \otimes_b \boldsymbol{\mathcal{B}}_i^{*}) \tag{4.83}$$

and $\boldsymbol{\mathcal{B}}_i$ is given by (3.23) of Chapter 3. We are adding the subscript "async" to quantities that are related to asynchronous networks. Correspondingly, the mean-square convergence rate for the synchronous network with the constant parameters $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$ will be determined by $\rho(\mathcal{F}_{\text{sync}})$ where

$$\mathcal{F}_{\text{sync}} \triangleq \bar{\mathcal{B}}^{\mathsf{T}} \otimes_b \bar{\mathcal{B}}^{*} \tag{4.84}$$

and $\bar{\mathcal{B}}$ is given by (3.29) of Chapter 3.

Lemma 4.7 (Matching mean-square convergence rates)**.** *For large i, the mean-square convergence rate of the asynchronous diffusion strategy is close to that of the synchronous diffusion strategy:*

$$|\rho(\mathcal{F}_{async}) - \rho(\mathcal{F}_{sync})| = O(\nu^{1+1/N^2}) \tag{4.85}$$

where $\rho(\mathcal{F}_{async})$ and $\rho(\mathcal{F}_{sync})$ are both dominated by $[1-\lambda_{\min}(H)]^2 = 1-O(\nu)$ for small ν by Assumption 4.1.

Proof. By Lemma 3.5 of Chapter 3, we have

$$\rho(\mathcal{F}_{\text{async}}) = \rho(F_{\text{async}}) + O(\nu^{1+1/N^2}) \tag{4.86}$$

where F_{async} is given by (3.77) of Chapter 3. Correspondingly, we will also have

$$\rho(\mathcal{F}_{\text{sync}}) = \rho(F_{\text{sync}}) + O(\nu^{1+1/N^2}) \tag{4.87}$$

where F_{sync} is given by

$$F_{\text{sync}} = \sum_{k=1}^{N}\sum_{\ell=1}^{N} \bar{p}_\ell \bar{p}_k (\bar{D}_\ell^{\mathsf{T}} \otimes \bar{D}_k) \tag{4.88}$$

Noting that F_{sync} is identical to F' in (3.174) of Chapter 3, then from Lemma 3.4 of Chapter 3 we obtain

$$\rho(F_{\text{async}}) = \rho(F_{\text{sync}}) + O(\nu^2) \tag{4.89}$$

Using (4.86), (4.87), and (4.89), we get

$$\begin{aligned} |\rho(\mathcal{F}_{\text{async}}) - \rho(\mathcal{F}_{\text{sync}})| &= |\rho(F_{\text{async}}) - \rho(F_{\text{sync}}) + O(\nu^{1+1/N^2})| \\ &= |O(\nu^2) + O(\nu^{1+1/N^2})| \\ &= O(\nu^{1+1/N^2}) \end{aligned} \tag{4.90}$$

Using (3.78) from Chapter 3 and (4.90) completes the proof. □

Likewise, assuming $c_{\mu,k,k} = 0$ and $c_{a,\ell k,\ell k} = 0$ for all k and ℓ for the synchronous strategy, it is easy to verify from (3.72)–(3.75) of Chapter 3 that $p = \bar{p} \otimes \bar{p}$. Then, we obtain the following expression for the steady-state MSD of the synchronous network with the constant parameters $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$:

$$\mathrm{MSD}_{\mathrm{sync}}^{\mathrm{dist}} = \frac{1}{4}\mathrm{Tr}(H^{-1}R_{\mathrm{sync}}) + O(\nu^{1+\gamma_o}) \tag{4.91}$$

where $H = O(\nu)$ and $0 < \gamma_o \leq 1/2$ are given by (3.79) and (3.65) from Chapter 3, respectively, and

$$R_{\mathrm{sync}} \triangleq \sum_{k=1}^{N} \bar{p}_k^2 \, \bar{\mu}_k^2 R_k = O(\nu^2) \tag{4.92}$$

Since $\mathrm{Tr}(H^{-1}R_{\mathrm{sync}}) = O(\nu)$, the first term on the RHS of (4.91) dominates the other term, $O(\nu^{1+\gamma_o})$. From (4.80) and (4.91), we observe that the network MSDs of asynchronous and synchronous networks are both in the order of ν.

Lemma 4.8 (Degradation in MSD is $O(\nu)$)**.** *The network MSD* (4.80) *for the asynchronous diffusion strategy is greater than the network MSD* (4.91) *for the synchronous diffusion strategy by a difference in the order of* ν*.*

Proof. The difference between R_{async} and R_{sync} is

$$R_{\mathrm{async}} - R_{\mathrm{sync}} = \sum_{k=1}^{N} [(p_{k,k} - \bar{p}_k^2)\bar{\mu}_k^2 + p_{k,k}c_{\mu,k,k}]R_k \tag{4.93}$$

where R_{async} is given by (3.83) of Chapter 3. Since $p_{k,k} - \bar{p}_k^2$ is the k-th entry on the diagonal of $P_p - \bar{p}\bar{p}^{\mathsf{T}}$, from Lemma 4.3, we know that all entries on the diagonal of $P_p - \bar{p}\bar{p}^{\mathsf{T}}$ are nonnegative, which implies that $p_{k,k} - \bar{p}_k^2 \geq 0$. Moreover, by Perron-Frobenius Theorem [79], all entries of the Perron eigenvector p must be positive, which implies that $p_{k,k} > 0$. We also know that $c_{\mu,k,k}$ must be positive in the asynchronous model. Therefore, we get

$$(p_{k,k} - \bar{p}_k^2)\bar{\mu}_k^2 + p_{k,k}c_{\mu,k,k} > 0 \tag{4.94}$$

Moreover, by using (3.181)–(3.183) from Chapter 3, we have

$$(p_{k,k} - \bar{p}_k^2)\bar{\mu}_k^2 + p_{k,k}c_{\mu,k,k} = O(\nu^2) \tag{4.95}$$

Then, using the fact that the $\{R_k\}$ are positive semi-definite, we conclude from (4.93)–(4.95) that

$$\|R_{\text{async}} - R_{\text{sync}}\| = O(\nu^2) > 0 \tag{4.96}$$

From (3.79) of Chapter 3, we know that $H^{-1} = O(\nu^{-1})$. Therefore, we get

$$\begin{aligned} \text{MSD}_{\text{async}}^{\text{dist}} - \text{MSD}_{\text{sync}}^{\text{dist}} &= \frac{1}{4}\text{Tr}[H^{-1}(R_{\text{async}} - R_{\text{sync}})] + O(\nu^{1+\gamma_o}) \\ &= O(\nu) + O(\nu^{1+\gamma_o}) = O(\nu) \end{aligned} \tag{4.97}$$

and

$$\text{MSD}_{\text{async}}^{\text{dist}} - \text{MSD}_{\text{sync}}^{\text{dist}} \geq 0 \tag{4.98}$$

which complete the proof. □

We observe from the above results that when the step-sizes are sufficiently small, the mean-square convergence rate of the asynchronous network tends to be immune from the uncertainties caused by random topologies, links, agents, and data arrival time. However, there is an $O(\nu)$ degradation in the steady-state MSD level for the asynchronous network – refer to Table 1.1 for a summary of the main conclusions.

4.5 A Case Study: MSE Estimation

The previous results apply to arbitrary strongly-convex costs $\{J_k(w)\}$ whose Hessian functions are locally Lipschitz continuous at w^o. In this section we specialize the results to the case of MSE estimation over networks, where the costs $\{J_k(w)\}$ become quadratic in $w \in \mathbb{C}^{M\times 1}$.

4.5.1 Problem Formulation and Modeling

We now assume that each agent k has access to streaming data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ related via the linear regression model:

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i} w^o + \boldsymbol{\xi}_k(i) \tag{4.99}$$

where $\boldsymbol{d}_k(i) \in \mathbb{C}$ is the observation, $\boldsymbol{u}_{k,i} \in \mathbb{C}^{1\times M}$ is the regressor, $w^o \in \mathbb{C}^{M\times 1}$ is the desired parameter vector, and $\boldsymbol{\xi}_k(i)$ is additive noise.

Assumption 4.2 (Data model)**.**

1. *The regressors $\{\boldsymbol{u}_{k,i}\}$ are temporally white and spatially independent circular symmetric complex random variables with zero mean and covariance matrix $R_{u,k} > 0$.*

2. *The noise signals $\{\boldsymbol{\xi}_k(i)\}$ are temporally white and spatially independent circular symmetric complex random variables with zero mean and variance $\sigma^2_{\xi,k} > 0$.*

3. *The random variables $\{\boldsymbol{u}_{k,i}, \boldsymbol{\xi}_\ell(j)\}$ are mutually independent for any k and ℓ, i and j, and they are independent of any other random variable.* □

The objective for the network is to estimate w^o by minimizing the aggregate mean-square-error cost defined by

$$\underset{w}{\text{minimize}} \;\; \sum_{k=1}^{N} J_k(w) \triangleq \sum_{k=1}^{N} \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} w|^2 \tag{4.100}$$

It can be verified that this problem satisfies Assumptions 2.1 and 2.2 introduced in Chapter 2.

4.5.2 Distributed Diffusion Solutions

The asynchronous diffusion solution (4.65a)–(4.65b) will then reduce to the following form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} + \boldsymbol{\mu}_k(i)\boldsymbol{u}_{k,i}^*[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}] \tag{4.101a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_{k,i}} \boldsymbol{a}_{\ell k}(i)\,\boldsymbol{\psi}_{\ell,i} \tag{4.101b}$$

and the synchronous network (4.82a)–(4.82b) will become

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} + \bar{\mu}_k\boldsymbol{u}_{k,i}^*[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}] \tag{4.102a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_{k,i}} \bar{a}_{\ell k}\,\boldsymbol{\psi}_{\ell,i} \tag{4.102b}$$

We assume that the network is under the Bernoulli model described in Chapter 2. For illustration purposes only, we assume that the parameters $\{\mu_k\}$ in (2.54) of Chapter 2 are uniform, $\mu_k \equiv \mu$, and that the parameters $\{a_{\ell k}; \ell \in \mathcal{N}_k\backslash\{k\}\}$ in (2.55) of Chapter 2 are given by $a_{\ell k} = |\mathcal{N}_k|^{-1}$.

Substituting (4.99) into (4.101a) and comparing with (2.17) of Chapter 2, we find that the approximate gradient, $\widehat{\nabla_{w^*}J_k}(\boldsymbol{w}_{k,i-1})$, and the corresponding gradient noise, $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})$, in this case are given by

$$\begin{aligned}\widehat{\nabla_{w^*}J_k}(\boldsymbol{w}_{k,i-1}) &= -\boldsymbol{u}_{k,i}^*[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}] \\ &= -\boldsymbol{u}_{k,i}^*\boldsymbol{u}_{k,i}\widetilde{\boldsymbol{w}}_{k,i-1} - \boldsymbol{u}_{k,i}^*\boldsymbol{\xi}_k(i) \\ &= -R_{u,k}\widetilde{\boldsymbol{w}}_{k,i-1} - \boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})\end{aligned} \tag{4.103}$$

where

$$\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1}) = (\boldsymbol{u}_{k,i}^*\boldsymbol{u}_{k,i} - R_{u,k})\widetilde{\boldsymbol{w}}_{k,i-1} + \boldsymbol{u}_{k,i}^*\boldsymbol{\xi}_k(i) \tag{4.104}$$

It can be verified that the gradient noise $\boldsymbol{v}_{k,i}(\boldsymbol{w}_{k,i-1})$ in (4.104) satisfies Assumption 3.1 of Chapter 3 and that the covariance matrix of $\underline{\boldsymbol{v}}_{k,i}(w^o) = \mathbb{T}(\boldsymbol{u}_{k,i}^*\boldsymbol{\xi}_k(i))$,

where $\mathbb{T}(\cdot)$ is from (2.3) of Chapter 2, is given by

$$R_k = \text{diag}\{\sigma_{\xi,k}^2 R_{u,k}, \sigma_{\xi,k}^2 R_{u,k}^{\mathsf{T}}\} \triangleq \sigma_{\xi,k}^2 H_k \tag{4.105}$$

Moreover, the complex Hessian of the cost $J_k(w)$ is given by

$$\nabla^2_{\underline{w}\underline{w}^*} J_k(\underline{w}) \triangleq H_k = \text{diag}\{R_{u,k}, R_{u,k}^{\mathsf{T}}\} \tag{4.106}$$

We further note that for the Bernoulli network under study,

$$\bar{\mu}_k^{(1)} = q_k \mu, \quad \bar{\mu}_k^{(2)} = q_k \mu^2 \tag{4.107}$$

Therefore, the parameter $\nu = \mu$ in this case. If μ is small enough and satisfies Assumption 4.1, then from (4.80), the network MSD of the asynchronous network is given by

$$\text{MSD}_{\text{async}}^{\text{diff}} = \frac{\mu}{2}\text{Tr}\left[\left(\sum_{k=1}^{N} \bar{p}_k q_k R_{u,k}\right)^{-1} \left(\sum_{k=1}^{N} p_{k,k} q_k \sigma_{\xi,k}^2 R_{u,k}\right)\right] + O(\mu^{1+\gamma_o}) \tag{4.108}$$

Likewise, the network MSD of the synchronous network from (4.91) is given by

$$\text{MSD}_{\text{sync}}^{\text{diff}} = \frac{\mu}{2}\text{Tr}\left[\left(\sum_{k=1}^{N} \bar{p}_k q_k R_{u,k}\right)^{-1} \left(\sum_{k=1}^{N} \bar{p}_k^2 q_k^2 \sigma_{\xi,k}^2 R_{u,k}\right)\right] + O(\mu^{1+\gamma_o}) \tag{4.109}$$

Clearly, since $q_k \leq 1$ and $p_{k,k} \geq \bar{p}_k^2$ for all k, the MSD in (4.108) is always greater than the MSD in (4.109) and the difference is in the order of μ.

4.5.3 Centralized Solution

The asynchronous batch solution (4.1) will now reduce to

$$\boldsymbol{w}_{\text{c},i} = \boldsymbol{w}_{\text{c},i-1} + \sum_{k=1}^{N} \boldsymbol{\pi}_k(i)\boldsymbol{\mu}_k(i)\boldsymbol{u}_{k,i}^*[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{\text{c},i-1}] \tag{4.110}$$

and the synchronous batch solution (4.60) will become

$$\boldsymbol{w}_{\text{c},i} = \boldsymbol{w}_{\text{c},i-1} + \sum_{k=1}^{N} \bar{\pi}_k \bar{\mu}_k \boldsymbol{u}_{k,i}^*[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{\text{c},i-1}] \tag{4.111}$$

We continue to assume that the random step-size parameters $\{\boldsymbol{\mu}_k(i)\}$ satisfy the same Bernoulli model described in Chapter 2 with a uniform profile $\mu_k \equiv \mu$. We use the procedure described in Section 4.3.2 to generate the random fusion coefficients $\{\boldsymbol{\pi}_k(i)\}$. Specifically, we have $\boldsymbol{\pi}_k(i) = \boldsymbol{\phi}_k(i)$, where $\boldsymbol{\phi}_k(i)$ denotes the k-th entry of $\boldsymbol{\phi}_i$ from (4.73).

4.5.4 Simulation Results

We consider a network consisting of $N = 100$ agents with the connected topology shown in Fig. 4.1 where each link is assumed to be bidirectional. The length of the unknown parameter w^o is set to $M = 2$. The regressors are assumed to be white, i.e., $R_{u,k} = \sigma_{u,k}^2 I_M$. The values of $\{\sigma_{u,k}^2, \sigma_{v,k}^2\}$ are randomly generated and shown in Fig. 4.2. The step-size parameter is set to $\mu = 0.002$. We randomly select the values for the probabilities $\{\eta_{\ell k}\}$ in (2.55) of Chapter 2 within the range $(0.4, 0.8)$, and randomly select the values for the probabilities $\{q_k\}$ in (2.54) of Chapter 2 within the set $\{0.3, 0.5, 0.7, 0.9\}$. The asynchronous distributed strategy (4.101a)–(4.101b), the synchronous distributed strategy (4.102a)–(4.102b), the asynchronous centralized solution (4.110), and the synchronous centralized solution (4.111) are all simulated over 100 trials and 6000 iterations for each trial. The random fusion coefficients $\{\boldsymbol{\pi}_k(i)\}$ are obtained by sampling $\boldsymbol{\phi}_i$ from (4.73). The $\boldsymbol{\phi}_i$ is constructed by consecutively multiplying 100 independent realizations of $\boldsymbol{A}_i$. The averaged learning curves (MSD) as well as the theoretical MSD results (4.108) for asynchronous solutions and (4.109) for synchronous solutions are plotted in Fig. 4.3. We observe a good match between theory and simulation. We also observe that both synchronous and asynchronous solutions converge at a similar rate but that the former attains a lower MSD level at steady-state as predicted by (4.108) and (4.109).

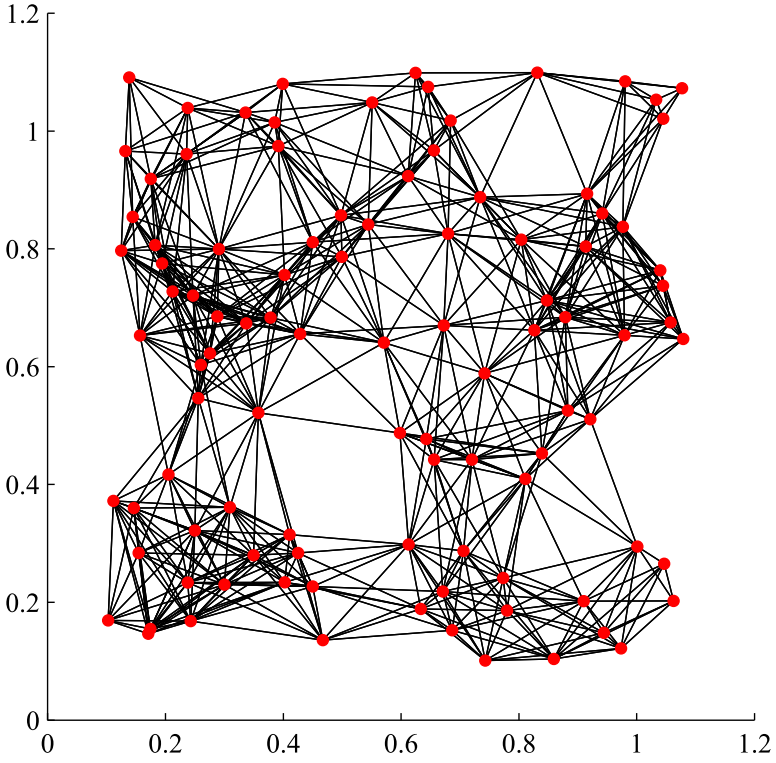


Figure 4.1: A topology with 100 nodes.

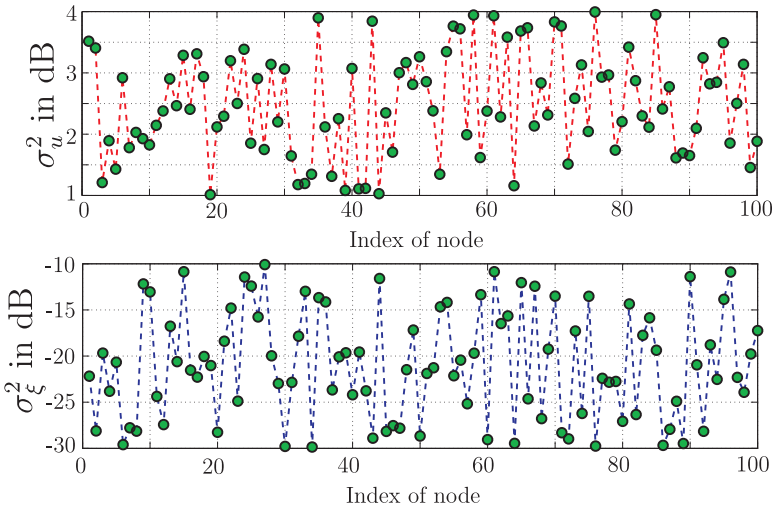


Figure 4.2: Values of $\{\sigma_{u,k}^2\}$ and $\{\sigma_{\xi,k}^2\}$.

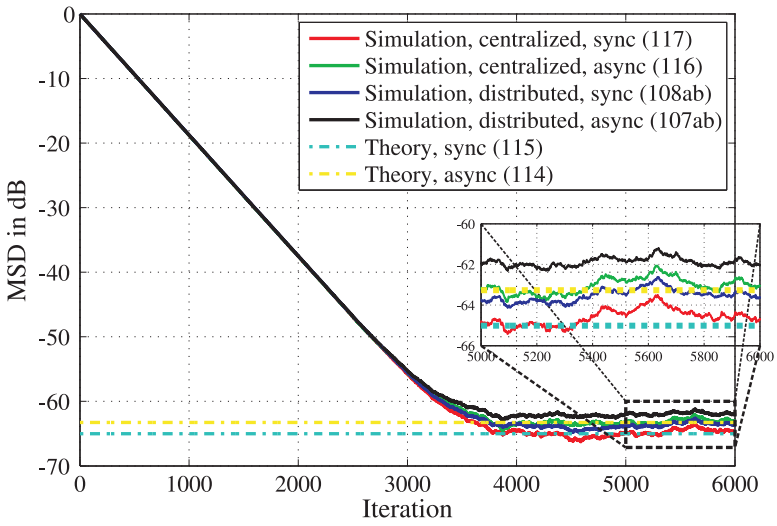


Figure 4.3: MSD learning curves for the asynchronous and synchronous modes of operation.

4.6 Conclusion

In this chapter, we compared the performance of distributed and centralized solutions under two modes of operation: synchronous and asynchronous implementations. We derived explicit comparisons for the mean and mean-square rates of convergence, as well as for the steady-state mean-square error performance. The main results are captured by Table 1. It is seen that diffusion networks are remarkably resilient to asynchronous or random failures: the convergence continues to occur at the same rate as synchronous or centralized solutions while the MSD level suffers a degradation in the order of $O(\nu)$ relative to synchronous diffusion networks. The results in the chapter highlight yet another benefit of cooperation: remarkable resilience to random failures and asynchronous events.

4.A Proof of Lemma 4.2

From (4.35), we get

$$\begin{aligned} F_{\rm c} &= \mathbb{E}\left[\left(\sum_{\ell=1}^{N}\boldsymbol{\pi}_\ell(i)\boldsymbol{D}_{\ell,i}\right)^{\mathsf{T}}\otimes\left(\sum_{k=1}^{N}\boldsymbol{\pi}_k(i)\boldsymbol{D}_{k,i}\right)\right] \\ &\overset{(a)}{=} \sum_{k=1}^{N}\sum_{\ell=1}^{N}\mathbb{E}[\boldsymbol{\pi}_\ell(i)\boldsymbol{\pi}_k(i)]\cdot\mathbb{E}(\boldsymbol{D}_{\ell,i}^{\mathsf{T}}\otimes\boldsymbol{D}_{k,i}) \\ &\overset{(b)}{=} \sum_{k=1}^{N}\sum_{\ell=1}^{N}(\bar{\pi}_\ell\bar{\pi}_k + c_{\pi,\ell,k})(\bar{D}_\ell^{\mathsf{T}}\otimes\bar{D}_k + c_{\mu,\ell,k}H_\ell^{\mathsf{T}}\otimes H_k) \end{aligned} \tag{4.112}$$

where step (a) is by using the independence condition from the asynchronous model; and step (b) is by using (4.7)–(4.10). Since $\{\bar{D}_k, H_k\}$ are all Hermitian, it is straightforward to verify that $F_{\rm c}$ is also Hermitian.

Using Jensen's inequality and the convexity of the 2-induced norm, $\|\cdot\|$, we obtain from (4.45) that

$$\rho(F_{\rm c}) \le \mathbb{E}\,\|\boldsymbol{B}_i^{\mathsf{T}}\otimes\boldsymbol{B}_i\| = \mathbb{E}\,\|\boldsymbol{B}_i\|^2 \tag{4.113}$$

where we used the identities $\|A\otimes B\| = \|A\|\cdot\|B\|$ [89, p. 245] and $\|A^{\mathsf{T}}\| = \|A\|$. Using Jensen's inequality again with respect to the convex coefficients $\{\boldsymbol{\pi}_k(i)\}$ and the fact that $\|\cdot\|^2$ is also a convex function, we get from (4.35) that

$$\|\boldsymbol{B}_i\|^2 = \left\|\sum_{k=1}^{N}\boldsymbol{\pi}_k(i)\boldsymbol{D}_{k,i}\right\|^2 \le \sum_{k=1}^{N}\boldsymbol{\pi}_k(i)\|\boldsymbol{D}_{k,i}\|^2 \tag{4.114}$$

Substituting (4.114) into (4.113), we obtain

$$\rho(F_{\rm c}) \le \sum_{k=1}^{N}\bar{\pi}_k\mathbb{E}\|\boldsymbol{D}_{k,i}\|^2 \le \max_k\,\mathbb{E}\|\boldsymbol{D}_{k,i}\|^2 \tag{4.115}$$

From (4.35) and from condition (2.7) in Chapter 2, we have

$$1-\boldsymbol{\mu}_k(i)\lambda_{k,\max} \le \lambda(\boldsymbol{D}_{k,i}) \le 1-\boldsymbol{\mu}_k(i)\lambda_{k,\min} \tag{4.116}$$

for every eigenvalue of $\boldsymbol{D}_{k,i}$ and for every k and $i \geq 0$. Since $\boldsymbol{D}_{k,i}$ is Hermitian, we conclude from (4.116) that for every k and $i \geq 0$,

$$\begin{aligned}\|\boldsymbol{D}_{k,i}\|^2 &\leq \max\{[1-\boldsymbol{\mu}_k(i)\lambda_{k,\min}]^2, [1-\boldsymbol{\mu}_k(i)\lambda_{k,\max}]^2\} \\ &\leq 1 - 2\boldsymbol{\mu}_k(i)\lambda_{k,\min} + \boldsymbol{\mu}_k^2(i)\lambda_{k,\max}^2 \end{aligned} \tag{4.117}$$

Substituting (4.117) into (4.115) yields

$$\begin{aligned}\rho(F_{\rm c}) &\leq \max_k \mathbb{E}\,[1 - 2\boldsymbol{\mu}_k(i)\lambda_{k,\min} + \boldsymbol{\mu}_k^2(i)\lambda_{k,\max}^2 | \widetilde{\boldsymbol{w}}_{{\rm c},i-1}] \\ &\leq \max_k \{1 - 2\bar{\mu}_k\lambda_{k,\min} + (\bar{\mu}_k^2 + c_{\mu,k,k})\lambda_{k,\max}^2\} \\ &< \max_k \{\gamma_k^2 + \alpha(\bar{\mu}_k^2 + c_{\mu,k,k})\} \\ &= \beta \end{aligned} \tag{4.118}$$

where $\alpha > 0$ and $\{\gamma_k^2, \beta\}$ are from (2.81) and (2.82) of Chapter 2, respectively. In (2.136) from Chapter 2, we established that $|\beta| < 1$ if condition (4.27) holds. Therefore, by (4.118) we conclude that $\rho(F_{\rm c}) < 1$ when condition (4.27) holds.

Since $c_{\mu,\ell,k} = O(\nu^2)$ by using (3.182) and (3.183) from Chapter 3, we get from (4.112) and (4.40) that

$$F_{\rm c} = \sum_{k=1}^{N}\sum_{\ell=1}^{N}(\bar{\pi}_\ell\bar{\pi}_k + c_{\pi,\ell,k})(\bar{D}_\ell^{\mathsf{T}} \otimes \bar{D}_k) + O(\nu^2) \tag{4.119}$$

Furthermore, we have

$$\begin{aligned}&\sum_{k=1}^{N}\sum_{\ell=1}^{N} c_{\pi,\ell,k}(\bar{D}_\ell^{\mathsf{T}} \otimes \bar{D}_k) \\ &\quad \overset{(a)}{=} \sum_{k=1}^{N}\sum_{\ell=1}^{N} c_{\pi,\ell,k}(I_{2M} - \bar{\mu}_\ell H_\ell)^{\mathsf{T}} \otimes (I_{2M} - \bar{\mu}_k H_k) \\ &\quad = \sum_{k=1}^{N}\sum_{\ell=1}^{N} c_{\pi,\ell,k}(I_{4M^2} - \bar{\mu}_\ell H_\ell^{\mathsf{T}} \otimes I_{2M} - I_{2M} \otimes \bar{\mu}_k H_k + \bar{\mu}_\ell\bar{\mu}_k H_\ell^{\mathsf{T}} \otimes H_k) \\ &\quad \overset{(b)}{=} \sum_{k=1}^{N}\sum_{\ell=1}^{N} c_{\pi,\ell,k}\bar{\mu}_\ell\bar{\mu}_k(H_\ell^{\mathsf{T}} \otimes H_k)\end{aligned}$$

$$\stackrel{(c)}{=} O(\nu^2) \tag{4.120}$$

where step (a) is by using (4.41); step (b) is by using (4.14); and step (c) is by using (3.182) and (3.183) from Chapter 3. From (4.119) and (4.120), we have

$$F_{\rm c} = \sum_{\ell=1}^{N}\sum_{k=1}^{N} \bar{\pi}_\ell \bar{\pi}_k (\bar{D}_\ell^{\mathsf{T}} \otimes \bar{D}_k) + O(\nu^2) \tag{4.121}$$

Now, consider the matrix $F_{\rm c}'$ defined in (4.61); it is easy to verify by using (4.40) that

$$F_{\rm c}' = \sum_{\ell=1}^{N}\sum_{k=1}^{N} \bar{\pi}_\ell \bar{\pi}_k (\bar{D}_\ell^{\mathsf{T}} \otimes \bar{D}_k) = \bar{B}^{\mathsf{T}} \otimes \bar{B} \tag{4.122}$$

Since $\bar{B}$ is Hermitian, so is $F_{\rm c}'$. From (4.121) and (4.122), we get $\|F_{\rm c} - F_{\rm c}'\| = O(\nu^2)$. Since both $F_{\rm c}$ and $F_{\rm c}'$ are Hermitian, their difference $F_{\rm c} - F_{\rm c}'$ is also Hermitian. Then, using a corollary of the Wielandt-Hoffman Theorem [86], we conclude that

$$|\lambda_m(F_{\rm c}) - \lambda_m(F_{\rm c}')| \leq \|F_{\rm c} - F_{\rm c}'\| = O(\nu^2) \tag{4.123}$$

where $\lambda_m(\cdot)$ denotes the m-th eigenvalue of its Hermitian matrix argument; the eigenvalues are assumed to be ordered from largest to smallest in each case. From (4.123), we immediately deduce that

$$|\rho(F_{\rm c}) - \rho(F_{\rm c}')| \leq O(\nu^2) \tag{4.124}$$

From (4.40)–(4.41) and (4.47), we have

$$\bar{B} = I_{2M} - H_{\rm c}, \qquad \lambda(\bar{B}) = 1 - \lambda(H_{\rm c}) \tag{4.125}$$

Since $H_{\rm c}$ is symmetric positive definite, and since the $\{\bar{\pi}_k\}$ are convex coefficients by (4.14), we get from Jensen's inequality that

$$0 < \lambda(H_{\rm c}) \leq \|H_{\rm c}\| \leq \sum_{k=1}^{N} \bar{\pi}_k \|\bar{\mu}_k H_k\| \leq \max_k \{\bar{\mu}_k \lambda_{k,\max}\} \tag{4.126}$$

for all eigenvalues of $H_{\rm c}$. When condition (4.27) holds, we have

$$\bar{\mu}_k \leq \bar{\mu}_k(1+\rho_k^2) < \frac{\lambda_{k,\min}}{\alpha+\lambda_{k,\max}^2} < \frac{1}{\lambda_{k,\max}} \tag{4.127}$$

for any k. This implies that $\max_k\{\bar{\pi}_k\lambda_{k,\max}\} < 1$ and therefore, $0 < \lambda(H_{\rm c}) < 1$ for all eigenvalues of $H_{\rm c}$. From (3.181) of Chapter 3 and (4.126), we get

$$0 < \lambda(H_{\rm c}) = O(\nu) < 1 \tag{4.128}$$

for any eigenvalue of $H_{\rm c}$. Therefore, we get from (4.125) that

$$\lambda(\bar{B}) = 1 - O(\nu), \quad \rho(\bar{B}) = 1 - \lambda_{\min}(H_{\rm c}) \tag{4.129}$$

Then, from (4.122) and (4.129), we have

$$\rho(F_{\rm c}') = [1-\lambda_{\min}(H_{\rm c})]^2 \tag{4.130}$$

It then follows from (4.124) and (4.130) that

$$\rho(F_{\rm c}) = [1-\lambda_{\min}(H_{\rm c})]^2 + O(\nu^2) \tag{4.131}$$

where $\lambda_{\min}(H_{\rm c}) = O(\nu)$. Under Assumption 4.1, we have

$$[1-\lambda_{\min}(H_{\rm c})]^2 = 1 - 2\lambda_{\min}(H_{\rm c}) + O(\nu^2) = 1 - O(\nu) \tag{4.132}$$

which therefore dominates the $O(\nu^2)$ in (4.131).

From (4.125) and (4.122), we get

$$F_{\rm c}' = I_{4M^2} - H_{\rm c}^{\mathsf{T}} \otimes I_{2M} - I_{2M} \otimes H_{\rm c} + H_{\rm c}^{\mathsf{T}} \otimes H_{\rm c} \tag{4.133}$$

Then, using (4.121), (4.122), and (4.133), we have

$$I_{4M^2} - F_{\rm c} = \underbrace{H_{\rm c}^{\mathsf{T}} \otimes I_{2M} + I_{2M} \otimes H_{\rm c}}_{=O(\nu)} + O(\nu^2) \tag{4.134}$$

where we used the fact that $H_{\rm c}^{\mathsf{T}} \otimes H_{\rm c} = O(\nu^2)$ since $H_{\rm c} = O(\nu)$ by (4.47). Using the fact that $H_{\rm c}$ is positive definite and is of the order of ν, we eventually get

$$\|(I_{4M^2} - F_{\rm c})^{-1}\| = O(\nu^{-1}) \tag{4.135}$$

4.B Proof of Theorem 4.5

We start with the $\lim_{i\to\infty} y_{c,i}$ in (4.53). From (4.46), we have

$$\lim_{i\to\infty} y_{c,i} = \lim_{i\to\infty} \operatorname{vec}\left(\mathbb{E}\underline{\boldsymbol{s}}_i\underline{\boldsymbol{s}}_i^*\right) \tag{4.136}$$

Using the gradient noise model from Section 3.1 of Chapter 3, it can be verified that $\underline{\boldsymbol{s}}_i$ is zero mean and that its conditional covariance matrix is given by

$$\begin{aligned}\mathbb{E}\left[\underline{\boldsymbol{s}}_i\underline{\boldsymbol{s}}_i^*|\mathbb{F}_{i-1}\right] &\overset{(a)}{=} \sum_{\ell=1}^{N}\sum_{k=1}^{N}\mathbb{E}\left[\boldsymbol{\pi}_\ell(i)\boldsymbol{\pi}_k(i)\right]\cdot\mathbb{E}\left[\boldsymbol{\mu}_\ell(i)\boldsymbol{\mu}_k(i)\right]\\ &\qquad\times\mathbb{E}\left[\underline{\boldsymbol{v}}_{\ell,i}(\boldsymbol{w}_{\mathrm{c},i-1})\underline{\boldsymbol{v}}_{k,i}^*(\boldsymbol{w}_{\mathrm{c},i-1})|\mathbb{F}_{i-1}\right]\\ &\overset{(b)}{=}\sum_{k=1}^{N}(\bar{\pi}_k^2+c_{\pi,k,k})(\bar{\mu}_k^2+c_{\mu,k,k})R_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1})\end{aligned} \tag{4.137}$$

where step (a) is by using the independence condition from the asynchronous model in Chapter 2; and step (b) is from (2.18) in Chapter 2, (3.3) in Chapter 3, and (4.7)–(4.10). Therefore,

$$\mathbb{E}\underline{\boldsymbol{s}}_i\underline{\boldsymbol{s}}_i^* = \sum_{k=1}^{N}(\bar{\pi}_k^2+c_{\pi,k,k})(\bar{\mu}_k^2+c_{\mu,k,k})\mathbb{E}R_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1}) \tag{4.138}$$

Note that

$$\begin{aligned}\|R_{k,i}(w^o)-\mathbb{E}R_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1})\| &\overset{(a)}{\leq} \|\mathcal{R}_i(\mathbb{1}_N\otimes w^o)-\mathbb{E}\mathcal{R}_i(\mathbb{1}_N\otimes\boldsymbol{w}_{\mathrm{c},i-1})\|\\ &\overset{(b)}{\leq}\kappa_v\cdot\left[\mathbb{E}\,\|\mathbb{1}_N\otimes\widetilde{\boldsymbol{w}}_{\mathrm{c},i-1}\|^4\right]^{\gamma_v/4}\\ &\overset{(c)}{=}\kappa_v N^{\gamma_v/2}\cdot\left[\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{\mathrm{c},i-1}\|^4\right]^{\gamma_v/4}\end{aligned} \tag{4.139}$$

where step (a) is due to (3.3) from Chapter 3; step (b) is by using (3.55) also from Chapter 3; and step (c) is by the fact that $\|\mathbb{1}_N\otimes x\|^4=[N\cdot\|x\|^2]^2=N^2\cdot\|x\|^4$ for any x. Under Assumption 4.1, we can get from Theorem 4.2 that

$$\limsup_{i\to\infty}\|R_{k,i}(w^o)-\mathbb{E}R_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1})\|\leq\kappa_v N^{\gamma_v/2}\cdot[b_4^2\nu^2]^{\gamma_v/4}=O(\nu^{\gamma_v/2}) \tag{4.140}$$

which means that, asymptotically, we can replace $\mathbb{E}R_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1})$ by R_k from (3.6) of Chapter 3 within an error in the order of $\nu^{\gamma_v/2}$. Therefore, it follows from (4.136) that

$$\begin{aligned}\lim_{i\to\infty} y_{c,i} &= \mathrm{vec}\left(\lim_{i\to\infty} \mathbb{E}\boldsymbol{s}_i\boldsymbol{s}_i^*\right)\\ &\overset{(a)}{=} \mathrm{vec}\left(\sum_{k=1}^{N}(\bar{\pi}_k^2+c_{\pi,k,k})(\bar{\mu}_k^2+c_{\mu,k,k})\lim_{i\to\infty}\mathbb{E}R_{k,i}(\boldsymbol{w}_{\mathrm{c},i-1})\right)\\ &\overset{(b)}{=} \mathrm{vec}\left(\sum_{k=1}^{N}(\bar{\pi}_k^2+c_{\pi,k,k})(\bar{\mu}_k^2+c_{\mu,k,k})[R_k+O(\nu^{\gamma_v/2})]\right)\\ &\overset{(c)}{=} \mathrm{vec}(R_{\mathrm{c}})+O(\nu^{2+\gamma_v/2}) \end{aligned} \tag{4.141}$$

where step (a) is by using (4.138); step (b) is by using (4.140); and step (c) is by using (4.57) and the fact from (2.190) of Chapter 2 that $\bar{\mu}_k^2+c_{\mu,k,k}=\bar{\mu}_k^{(2)}=O(\nu^2)$. Substituting (4.141) into (4.53) yields

$$z_{\mathrm{c},\infty} = (I_{4M^2}-F_{\mathrm{c}})^{-1}\cdot\mathrm{vec}(R_{\mathrm{c}})+O(\nu^{1+\gamma_v/2}) \tag{4.142}$$

where we used Lemma 4.2. Substituting (4.53) and $\Sigma=I_{2M}$ into (4.54), and using (4.50) as well as the fact that F_{c} and R_{c} are Hermitian, we obtain

$$\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}'_{\mathrm{c},i}\|^2 = \frac{1}{2}[\mathrm{vec}(R_{\mathrm{c}})]^*(I_{4M^2}-F_{\mathrm{c}})^{-1}\mathrm{vec}(I_{2M})+O(\nu^{1+\gamma_v/2}) \tag{4.143}$$

Substituting (4.143) into (4.56) yields (4.58).

We establish (4.59) next. From (4.134), we know that

$$I_{4M^2}-F_{\mathrm{c}} = S_{\mathrm{c}}+O(\nu^2) \tag{4.144}$$

where

$$S_{\mathrm{c}} \triangleq H_{\mathrm{c}}^{\mathsf{T}}\otimes I_{2M}+I_{2M}\otimes H_{\mathrm{c}} = O(\nu) \tag{4.145}$$

Since H_{c} is symmetric and positive definite by (4.128), it is easy to verify that S_{c} is also symmetric and positive definite. Therefore, S_{c} is invertible. Using the

matrix inversion lemma [87], we get from (4.144) that

$$(I_{4M^2} - F_{\rm c})^{-1} = S_{\rm c}^{-1} + O(1) \tag{4.146}$$

where we used the fact that $\|S_{\rm c}^{-1}\| = O(\nu^{-1})$. Substituting (4.146) into (4.58) yields:

$$\begin{aligned}
\mathrm{MSD}^{\rm cent} &= \frac{1}{2}[\mathrm{vec}(R_{\rm c})]^*[S_{\rm c}^{-1} + O(1)]\mathrm{vec}(I_{2M}) + O(\nu^{1+\gamma_o}) \\
&= \frac{1}{2}[\mathrm{vec}(R_{\rm c})]^* S_{\rm c}^{-1}\mathrm{vec}(I_{2M}) + O(\nu^2) + O(\nu^{1+\gamma_o}) \\
&= \frac{1}{2}[\mathrm{vec}(R_{\rm c})]^* S_{\rm c}^{-1}\mathrm{vec}(I_{2M}) + O(\nu^{1+\gamma_o})
\end{aligned} \tag{4.147}$$

where we used the fact from (4.57) that $\|R_{\rm c}\| = O(\nu^2)$ and $\gamma_o < 1/2$ from (3.65) of Chapter 3. Since the first term on the RHS of (4.147) is of the order of ν, it is the dominant term under Assumption 4.1. To further simplify (4.147), we introduce the Lyapunov equation with respect to the unknown square matrix X:

$$XH_{\rm c} + H_{\rm c}X = I_{2M} \tag{4.148}$$

where $H_{\rm c}$ is given by (4.47). Vectorizing both sides and using (4.145), the Lyapunov equation is equivalent to the linear system of equations:

$$S_{\rm c}\mathrm{vec}(X) = \mathrm{vec}(I_{2M}) \tag{4.149}$$

Since $S_{\rm c}$ is invertible, the linear equation (4.149) has a unique solution, which is given by $X = \frac{1}{2}H_{\rm c}^{-1}$. From the Lyapunov equation (4.148) we get

$$\begin{aligned}
[\mathrm{vec}(R_{\rm c})]^* S_{\rm c}^{-1}\mathrm{vec}(I_{2M}) &= \frac{1}{2}[\mathrm{vec}(R_{\rm c})]^*\mathrm{vec}(H_{\rm c}^{-1}) \\
&= \frac{1}{2}\mathrm{Tr}(H_{\rm c}^{-1}R_{\rm c})
\end{aligned} \tag{4.150}$$

where we used the fact that $R_{\rm c}$ is Hermitian. Result (4.59) then follows from (4.147) and (4.150). The term $\mathrm{Tr}(H_{\rm c}^{-1}R_{\rm c}) = O(\nu)$ in (4.59) is the dominant term under Assumption 4.1.

4.C Proof of Lemma 4.3

From Lemma 3.3 of Chapter 3, we know that P_p is symmetric and, therefore, the matrix difference $C_p \triangleq P_p - \bar{p}\bar{p}^{\mathsf{T}}$ is also symmetric. We also know from Lemma 3.3 of Chapter 3 that $C_p \mathbb{1}_N = 0$. To establish that C_p is positive semi-definite, we consider the following quadratic expression:

$$x^{\mathsf{T}} C_p x = x^{\mathsf{T}}(P_p - \bar{p}\bar{p}^{\mathsf{T}})x = x^{\mathsf{T}} P_p x - (x^{\mathsf{T}}\bar{p})^2 \tag{4.151}$$

for any vector $x \in \mathbb{R}^N$. Note that

$$x^{\mathsf{T}} P_p x = \mathrm{vec}(x^{\mathsf{T}} P_p x) = \frac{1}{N^2}(x^{\mathsf{T}} \otimes x^{\mathsf{T}})p \cdot \mathbb{1}_{N^2}^{\mathsf{T}} \mathbb{1}_{N^2} \tag{4.152}$$

by using the relation $p = \mathrm{vec}(P_p)$ from (3.75) of Chapter 3 and the fact that $\mathbb{1}_{N^2}^{\mathsf{T}} \mathbb{1}_{N^2} = N^2$. Since

$$\bar{A} \otimes \bar{A} + C_A = \mathbb{E}(\boldsymbol{A}_j \otimes \boldsymbol{A}_j) \tag{4.153}$$

we can introduce a series of fictitious random combination matrices $\{\boldsymbol{A}'_j; j \geq 1\}$ such that they are mutually-independent and satisfy

$$\mathbb{E}(\boldsymbol{A}'_j \otimes \boldsymbol{A}'_j) = \bar{A} \otimes \bar{A} + C_A \tag{4.154}$$

for any $j \geq 1$. Let $\boldsymbol{\Phi}_i \triangleq \prod_{j=1}^{i} \boldsymbol{A}'_j$ for any $i \geq 1$. Then,

$$\lim_{i\to\infty} \mathbb{E}(\boldsymbol{\Phi}_i \otimes \boldsymbol{\Phi}_i) \overset{(a)}{=} \lim_{i\to\infty} \prod_{j=1}^{i} \mathbb{E}(\boldsymbol{A}'_j \otimes \boldsymbol{A}'_j) \overset{(b)}{=} p \cdot \mathbb{1}_{N^2}^{\mathsf{T}} \tag{4.155}$$

where step (a) is by using the fact that the $\{\boldsymbol{A}'_j\}$ are mutually-independent, and step (b) is by using (4.153) and the Perron-Frobenius Theorem [79]. Substituting (4.155) into (4.152) and using $\mathbb{1}_{N^2} = \mathbb{1}_N \otimes \mathbb{1}_N$, we get

$$x^{\mathsf{T}} P_p x = \frac{1}{N^2} \lim_{i\to\infty} \mathbb{E}\left[(x^{\mathsf{T}} \boldsymbol{\Phi}_i \mathbb{1}_N)^2\right] \tag{4.156}$$

Moreover, since $\bar{A} = \mathbb{E}(\boldsymbol{A}_j|\boldsymbol{w}_{j-1})$, we have

$$\lim_{i\to\infty} \mathbb{E}(\boldsymbol{\Phi}_i) = \lim_{i\to\infty} \prod_{j=1}^{i} \mathbb{E}(\boldsymbol{A}'_j) = \lim_{i\to\infty} (\bar{A})^i = \bar{p} \cdot \mathbb{1}_N^{\mathsf{T}} \tag{4.157}$$

Then, using (4.157) and the fact that $\mathbb{1}_N^{\mathsf{T}}\mathbb{1}_N = N$, we have

$$x^{\mathsf{T}}\bar{p} = \frac{1}{N} x^{\mathsf{T}}\bar{p} \cdot \mathbb{1}_N^{\mathsf{T}}\mathbb{1}_N = \frac{1}{N} \lim_{i\to\infty} \mathbb{E}(x^{\mathsf{T}}\boldsymbol{\Phi}_i \mathbb{1}_N) \tag{4.158}$$

Substituting (4.156) and (4.158) into (4.151) yields

$$x^{\mathsf{T}} C_p x = \frac{1}{N^2} \lim_{i\to\infty} \left\{ \mathbb{E}\left[(x^{\mathsf{T}}\boldsymbol{\Phi}_i \mathbb{1}_N)^2\right] - \left[\mathbb{E}(x^{\mathsf{T}}\boldsymbol{\Phi}_i \mathbb{1}_N)\right]^2 \right\} \geq 0 \tag{4.159}$$

which confirms that C_p is positive semi-definite.

4.D Proof of Lemma 4.4

We prove Lemma 4.3 by using a procedure similar to the one given in Appendix 3.I of Chapter 3. Introduce the Jordan decomposition [87]:

$$\bar{A} = \bar{P}\bar{J}\bar{Q}^{\mathsf{T}} = \begin{bmatrix} \bar{p} & \bar{P}' \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \bar{J}' \end{bmatrix} \begin{bmatrix} \mathbb{1}_N & \bar{Q}' \end{bmatrix}^{\mathsf{T}} \tag{4.160}$$

where $\bar{J}'$ is a sub-matrix of $\bar{J}$ containing its stable eigenvalues, $\bar{P}'$ and $\bar{Q}'$ are sub-matrices of $\bar{P}$ and $\bar{Q}$, and $\bar{P}^{-1} = \bar{Q}^{\mathsf{T}}$. Then, the Jordan decomposition of $\bar{\mathcal{A}} = \bar{A} \otimes I_{2M}$ from (3.25) of Chapter 3 is given by

$$\bar{\mathcal{A}} = \bar{\mathcal{P}}\bar{\mathcal{J}}\bar{\mathcal{Q}}^{\mathsf{T}} = \begin{bmatrix} \bar{p}' & \bar{\mathcal{P}}' \end{bmatrix} \begin{bmatrix} I_{2M} & 0 \\ 0 & \bar{\mathcal{J}}' \end{bmatrix} \begin{bmatrix} \bar{q}' & \bar{\mathcal{Q}}' \end{bmatrix}^{\mathsf{T}} \tag{4.161}$$

where

$$\bar{\mathcal{P}} = \bar{P} \otimes I_{2M}, \qquad \bar{\mathcal{P}}' \triangleq \bar{P}' \otimes I_{2M} \tag{4.162}$$

$$\bar{\mathcal{J}} = \bar{J} \otimes I_{2M}, \qquad \bar{\mathcal{J}}' \triangleq \bar{J}' \otimes I_{2M} \tag{4.163}$$

$$\bar{\mathcal{Q}} = \bar{Q} \otimes I_{2M}, \qquad \bar{\mathcal{Q}}' \triangleq \bar{Q}' \otimes I_{2M} \tag{4.164}$$

$$\bar{p}' = \bar{p} \otimes I_{2M}, \qquad \bar{q}' \triangleq \mathbb{1}_N \otimes I_{2M} \tag{4.165}$$

Let

$$\bar{\mathcal{X}} \triangleq I_{2MN} - \bar{\mathcal{D}} = \bar{\mathcal{M}}\mathcal{H} = O(\nu) \tag{4.166}$$

where $\{\bar{\mathcal{D}}, \bar{\mathcal{M}}, \mathcal{H}\}$ are from (3.28), (3.26), and (3.10) of Chapter 3, respectively. Then, by (3.29) from Chapter 3 and using the fact that $\bar{\mathcal{A}}$ is real and $\bar{\mathcal{D}}$ is Hermitian, we get

$$\bar{\mathcal{Q}}^{\mathsf{T}}\bar{\mathcal{B}}^*\bar{\mathcal{P}} = \bar{\mathcal{Q}}^{\mathsf{T}}\bar{\mathcal{D}}\bar{\mathcal{A}}\bar{\mathcal{P}} = \begin{bmatrix} I_{2M} - \bar{q}'^{\mathsf{T}}\bar{\mathcal{X}}\bar{p}' & -\bar{q}'^{\mathsf{T}}\bar{\mathcal{X}}\bar{\mathcal{P}}'\bar{\mathcal{J}}' \\ -\bar{\mathcal{Q}}'^{\mathsf{T}}\bar{\mathcal{X}}\bar{p}' & \bar{\mathcal{J}}' - \bar{\mathcal{Q}}'^{\mathsf{T}}\bar{\mathcal{X}}\bar{\mathcal{P}}'\bar{\mathcal{J}}' \end{bmatrix} \tag{4.167}$$

Using (4.165) and (4.166) above and (3.10) and (3.26) from Chapter 3, we obtain

$$\bar{q}'^{\mathsf{T}}\bar{\mathcal{X}}\bar{p}' = \bar{q}'^{\mathsf{T}}\bar{\mathcal{M}}\mathcal{H}\bar{p}' = \sum_{k=1}^{N} \bar{p}_k\bar{\mu}_k H_k = H = O(\nu) \tag{4.168}$$

where H is given by (3.79) of Chapter 3. By (4.166), we get

$$\|\bar{q}'^{\mathsf{T}}\bar{\mathcal{X}}\bar{\mathcal{P}}'\bar{\mathcal{J}}'\| = O(\nu), \quad \|\bar{\mathcal{Q}}'^{\mathsf{T}}\bar{\mathcal{X}}\bar{p}'\| = O(\nu), \quad \|\bar{\mathcal{Q}}'^{\mathsf{T}}\bar{\mathcal{X}}\bar{\mathcal{P}}'\bar{\mathcal{J}}'\| = O(\nu) \tag{4.169}$$

Therefore, we get from (4.167)–(4.169) that

$$\bar{\mathcal{Q}}^{\mathsf{T}}\bar{\mathcal{B}}^*\bar{\mathcal{P}} = \begin{bmatrix} \bar{B}_{\mathrm{d}} & O(\nu) \\ O(\nu) & \bar{\mathcal{J}}' + O(\nu) \end{bmatrix} \tag{4.170}$$

where

$$\bar{B}_{\mathrm{d}} \triangleq I_{2M} - H \tag{4.171}$$

is Hermitian. From (3.178) of Chapter 3, we immediately get

$$\lambda(\bar{B}_{\mathrm{d}}) = \lambda(I_{2M} - H) = 1 - O(\nu) > 0 \tag{4.172}$$

$$\rho(\bar{B}_{\rm d}) = 1 - \lambda_{\min}(H) = 1 - O(\nu) \tag{4.173}$$

for sufficiently small ν under Assumption 4.1. Conjugating both sides of (4.170) and using the fact that $\bar{B}_{\rm d}$ is Hermitian, we get

$$\bar{\mathcal{B}}_s \triangleq (\bar{\mathcal{Q}}^{\mathsf{T}}\bar{\mathcal{B}}^*\bar{\mathcal{P}})^* = \bar{\mathcal{P}}^*\bar{\mathcal{B}}(\bar{\mathcal{Q}}^*)^{\mathsf{T}} = \begin{bmatrix} \bar{B}_{\rm d} & O(\nu) \\ O(\nu) & \bar{\mathcal{J}}'^* + O(\nu) \end{bmatrix} \tag{4.174}$$

Since $\bar{\mathcal{B}}_s$ is similar to $\bar{\mathcal{B}}$, they have the same eigenvalues [87]. Since $\bar{B}_d$ is Hermitian, let us introduce its eigenvalue decomposition as

$$\bar{B}_d = \bar{U}\bar{\Lambda}\bar{U}^* \tag{4.175}$$

where $\bar{U}$ is a $2M \times 2M$ unitary matrix and $\bar{\Lambda}$ is a $2M \times 2M$ diagonal matrix. The $(N-1) \times (N-1)$ matrix $\bar{J}'$, which contains the stable eigenvalues of $\bar{A}$ in (4.160), can be generally expressed as

$$\bar{J}' = \begin{bmatrix} \bar{\lambda}_{a,2} & & \bar{T}' \\ & \ddots & \\ 0 & & \bar{\lambda}_{a,N} \end{bmatrix} \tag{4.176}$$

where $\{\bar{\lambda}_{a,n}\}$ are the eigenvalues of $\bar{A}$ with $\bar{\lambda}_{a,1} = 1$ and $|\bar{\lambda}_{a,n}| < 1$ for all $n = 2, 3, \ldots, N$. In (4.176), the elements in the strictly upper triangular region $\bar{T}'$ are either 1 or 0, which depend on the Jordan blocks in $\bar{J}'$. Using (4.176) and (4.163), we can express the $(2,2)$ block in (4.174) as

$$\bar{\mathcal{J}}'^* + O(\nu) = \begin{bmatrix} \bar{\lambda}^*_{a,2} I_{2M} + O(\nu) & & O(\nu) \\ & \ddots & \\ \bar{\mathcal{T}}'^* + O(\nu) & & \bar{\lambda}^*_{a,N} I_{2M} + O(\nu) \end{bmatrix} \tag{4.177}$$

where the elements in the strictly lower triangular region $\bar{\mathcal{T}}'^*$ are either 1 or 0, which depend on the elements of $\bar{T}'$ in (4.176). We now apply a similarity

transformation to $\bar{\mathcal{B}}_s$ by multiplying

$$\bar{\mathcal{D}} \triangleq \text{diag}\{\nu^{\epsilon}\bar{U}, \nu^{2\epsilon}I_{2M}, \nu^{3\epsilon}I_{2M}, \ldots, \nu^{N\epsilon}I_{2M}\} \tag{4.178}$$

and its inverse $\bar{\mathcal{D}}^{-1}$ on either side of (4.174), where $\epsilon = 1/N$. Using (4.174) and (4.177), we end up with

$$\bar{\mathcal{D}}\bar{\mathcal{B}}_s\bar{\mathcal{D}}^{-1} = \left[\begin{array}{c|ccc} \bar{\Lambda} & \multicolumn{3}{c}{O(\nu^{\epsilon})} \\ \hline & \bar{\lambda}^*_{a,2}I_{2M} + O(\nu) & & O(\nu^{\epsilon}) \\ O(\nu^{1+\epsilon}) & & \ddots & \\ & O(\nu^{\epsilon}) & & \bar{\lambda}^*_{a,N}I_{2M} + O(\nu) \end{array}\right] \tag{4.179}$$

From (4.179), we know that all off-diagonal entries of $\bar{\mathcal{D}}\bar{\mathcal{B}}_s\bar{\mathcal{D}}^{-1}$ are *at least* of the order of ν^{ϵ}. Therefore, using Gershgorin Theorem [86, p. 320] under Assumption 4.1, and since $\bar{\mathcal{B}}$ and $\bar{\mathcal{B}}_s$ have the same eigenvalues due to similarity, we get

$$|\lambda(\bar{\mathcal{B}}) - \lambda(\bar{B}_d)| \leq O(\nu^{1+\epsilon}) \quad \text{or} \quad |\lambda(\bar{\mathcal{B}}) - \bar{\lambda}^*_{a,k}| \leq O(\nu^{\epsilon}) \tag{4.180}$$

where $\lambda(\bar{\mathcal{B}})$ denotes the eigenvalue of $\bar{\mathcal{B}}$ and $k = 2, 3, \ldots, N$. Result (4.180) implies that the eigenvalues of $\bar{\mathcal{B}}$ are either located in the Gershgorin circles that are centered at the eigenvalues of $\bar{B}_d$ with radii $O(\nu^{1+\epsilon})$ or in the Gershgorin circles that are centered at $\{\bar{\lambda}^*_{a,k}; k = 2, 3, \ldots, N\}$ with radii $O(\nu^{\epsilon})$. From (4.173), we have

$$\rho(\bar{B}_d) = 1 - O(\nu) < 1 \tag{4.181}$$

By Assumption 3.3 from Chapter 3 and Perron-Frobenius Theorem [79], we have

$$\rho(\bar{J}'^*) \triangleq \max_{k=2,3,\ldots,N} |\bar{\lambda}^*_{a,k}| = \rho(\bar{J}') < 1 \tag{4.182}$$

By Assumption 4.1, if the parameter ν is small enough such that

$$\rho(\bar{J}') + O(\nu^{\epsilon}) < 1 - O(\nu) = \rho(\bar{B}_d) \tag{4.183}$$

holds, then the Gershgorin circles centered at the eigenvalues of $\bar{B}_d$ are isolated from those centered at $\{\bar{\lambda}^*_{a,k}; k = 2, 3, \ldots, N\}$. According to Gershgorin Theorem [88, p. 181], there are precisely $2M$ eigenvalues of $\bar{\boldsymbol{\mathcal{B}}}$ satisfying

$$|\lambda(\bar{\boldsymbol{\mathcal{B}}}) - \lambda(\bar{B}_d)| \leq O(\nu^{1+\epsilon}) \tag{4.184}$$

while all the other eigenvalues satisfy

$$|\lambda(\bar{\boldsymbol{\mathcal{B}}}) - \bar{\lambda}^*_{a,k}| \leq O(\nu^{\epsilon}), \quad k = 2, 3, \ldots, N \tag{4.185}$$

By (4.183), the eigenvalues $\lambda(\bar{\boldsymbol{\mathcal{B}}})$ satisfying (4.184) are greater than those satisfying (4.185) in magnitude. Furthermore, when ν is sufficiently small, the Gershgorin circles centered at $\lambda_{\max}(\bar{B}_d)$ with radius $O(\nu^{1+\epsilon})$ will become disjoint from the other circles. Then, by using Gershgorin Theorem again, we conclude from (4.184) that

$$|\rho(\bar{\boldsymbol{\mathcal{B}}}) - \rho(\bar{B}_d)| \leq O(\nu^{1+\epsilon}) \tag{4.186}$$

It is worth noting that from (4.181) and (4.186) we get

$$\rho(\bar{\boldsymbol{\mathcal{B}}}) \leq 1 - O(\nu) + O(\nu^{1+\epsilon}) < 1 \tag{4.187}$$

for $\nu \ll 1$ because $\epsilon = 1/N > 1$. Eventually, using (4.40), (4.67), and (4.171), it is straightforward to verify that

$$\bar{B} = \bar{B}_{\mathrm{d}} \tag{4.188}$$

Using (4.173), (4.186), and (4.188) completes the proof.

CHAPTER 5

Imperfect Information Exchange and Drifting Objectives

In this chapter we investigate another type of imperfections that affect the performance of the diffusino strategies over networks. Adaptive networks rely on in-network and collaborative processing among distributed agents to deliver enhanced performance in estimation and inference tasks. Information is exchanged among the nodes, usually over noisy links. The combination weights that are used by the nodes to fuse information from their neighbors play a critical role in influencing the adaptation and tracking abilities of the network. We first investigates the mean-square performance of general adaptive diffusion algorithms in the presence of various sources of imperfect information exchanges, quantization errors, and model non-stationarities. Among other results, the analysis reveals that link noise over the regression data modifies the dynamics of the network evolution in a distinct way, and leads to biased estimates in steady-state. The analysis also reveals how the network mean-square performance is dependent on the combination weights. We then use these observations to show how the combination weights can be optimized and adapted. Simulation results illustrate the theoretical findings and match well with theory.

In the original diffusion least-mean-squares (LMS) strategy [5, 6], the weight estimates that are exchanged among the nodes can be subject to quantization

errors and additive noise over the communication links. Studying the degradation in mean-square performance that results from these particular perturbations can be pursued, for both incremental and diffusion strategies, by extending the mean-square analysis already presented in [5, 6], in the same manner that the tracking analysis of conventional stand-alone adaptive filters was obtained from the counterpart results in the stationary case (as explained in [46, Ch. 21]). Useful results along these lines, which study the effect of link noise during the exchange of the weight estimates, already appear for the traditional diffusion algorithm in the works [47–49, 105] and for consensus-based algorithms in [26, 50]. In this chapter, our objective is to go beyond these earlier studies by taking into account additional effects, and by considering a more general algorithmic structure. The reason for this level of generality is because the analytical results will help reveal which noise sources influence the network performance more seriously, in what manner, and at what stage of the adaptation process. The results will suggest important remedies and mechanisms to adapt the combination weights in real-time. Some of these insights are hard to get if one focuses solely on noise during the exchange of the weight estimates. The analysis will further show that noise during the exchange of the regression data plays a more critical role than other sources of imperfection: this particular noise alters the learning dynamics and modes of the network, and biases the weight estimates. Noises related to the exchange of other pieces of information do not alter the dynamics of the network but contribute to the deterioration of the network performance.

To arrive at these results, we first consider a generalized analysis that applies to a broad class of diffusion adaptation strategies (see (5.5)–(5.7) further ahead; this class includes the original diffusion strategies (5.3) and (5.4) as two special cases). The analysis allows us to account for various sources of information noise over the communication links. We allow for noisy exchanges during *each* of the

three processing steps of the adaptive diffusion algorithm (the two combination steps (5.5) and (5.7) and the adaptation step (5.6)). In this way, we are able to examine how the three sets of combination coefficients $\{a_{1,\ell k}, c_{\ell k}, a_{2,\ell k}\}$ in (5.5)–(5.7) influence the propagation of the noise signals through the network dynamics. Our results further reveal how the network mean-square-error performance is dependent on these combination weights. Following this line of reasoning, the analysis leads to algorithms (5.124) and (5.128) further ahead for choosing the combination coefficients to improve the steady-state network performance.

It should be noted that several combination rules, such as the Metropolis rule [106] and the maximum degree rule [107], were proposed previously in the literature — especially in the context of consensus-based iterations [22, 41, 107]. These schemes, however, usually suffer performance degradation in the presence of noisy information exchange since they ignore the network noise profile [108]. When the noise variance differs across the nodes, it becomes necessary to design combination rules that are aware of this variation as outlined further ahead in Section VI-B. Moreover, in a mobile network [19] where nodes are on the move and where neighborhoods evolve over time, it is even more critical to employ adaptive combination strategies that are able to track the variations in the noise profile in order to cope with such dynamic environments. This issue is taken up in Section VI-C. The results in this chapter are based on material from [78].

5.1 Diffusion Algorithms with Imperfect Information Exchange

We consider a connected network consisting of N nodes. Each node k collects scalar measurements $\boldsymbol{d}_k(i)$ and $1 \times M$ regression data vectors $\boldsymbol{u}_{k,i}$ over successive

time instants $i \geq 0$. Note that we use parenthesis to refer to the time-dependence of scalar variables, as in $\boldsymbol{d}_k(i)$, and subscripts to refer to the time-dependence of vector variables, as in $\boldsymbol{u}_{k,i}$. The measurements across all nodes are assumed to be related to an unknown $M \times 1$ vector w^o via a linear regression model of the form [46]:

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i} w^o + \boldsymbol{v}_k(i) \tag{5.1}$$

where $\boldsymbol{v}_k(i)$ denotes the measurement or model noise with zero mean and variance $\sigma_{v,k}^2$. The vector w^o in (5.1) denotes the parameter of interest, such as the parameters of some underlying physical phenomenon, the taps of a communication channel, or the location of food sources or predators. Such data models are also useful in studies on hybrid combinations of adaptive filters [109–113].

The nodes in the network would like to estimate w^o by solving the following minimization problem:

$$\underset{w}{\text{minimize}} \quad \sum_{k=1}^{N} \mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} w|^2 \tag{5.2}$$

In previous works [5,6,114], we introduced and studied several distributed strategies of the diffusion type that allow nodes to cooperate with each other in order to solve problems of the form (5.2) in an adaptive manner. These diffusion strategies endow networks with adaptation and learning abilities, and enable information to diffuse through the network in real-time. We review the adaptive diffusion strategies below.

5.1.1 Diffusion Adaptation with Perfect Information Exchange

In [5, 6], two classes of diffusion algorithms were proposed. One class is the so-called Combine-then-Adapt (CTA) strategy:

$$\begin{cases} \boldsymbol{\phi}_{k,i-1} = \displaystyle\sum_{\ell\in\mathcal{N}_k} a_{1,\ell k}\boldsymbol{w}_{\ell,i-1} \\ \boldsymbol{w}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \mu_k \displaystyle\sum_{\ell\in\mathcal{N}_k} c_{\ell k}\boldsymbol{u}_{\ell,i}^{*}[\boldsymbol{d}_\ell(i) - \boldsymbol{u}_{\ell,i}\boldsymbol{\phi}_{k,i-1}] \end{cases} \tag{5.3}$$

and the second class is the so-called Adapt-then-Combine (ATC) strategy:

$$\begin{cases} \boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} + \mu_k \displaystyle\sum_{\ell\in\mathcal{N}_k} c_{\ell k}\boldsymbol{u}_{\ell,i}^{*}[\boldsymbol{d}_\ell(i) - \boldsymbol{u}_{\ell,i}\boldsymbol{w}_{k,i-1}] \\ \boldsymbol{w}_{k,i} = \displaystyle\sum_{\ell\in\mathcal{N}_k} a_{2,\ell k}\boldsymbol{\psi}_{\ell,i} \end{cases} \tag{5.4}$$

where the $\{a_{1,\ell k}, c_{\ell k}, a_{2,\ell k}\}$ are nonnegative entries of the $N \times N$ matrices A_1, C, and A_2, respectively. The coefficients $\{a_{1,\ell k}, c_{\ell k}, a_{2,\ell k}\}$ are zero whenever node ℓ is not connected to node k, i.e., $\ell \notin \mathcal{N}_k$, where $\mathcal{N}_k$ denotes the neighborhood of node k and $k \in \mathcal{N}_k$. The two strategies (5.3) and (5.4) can be integrated into one broad class of diffusion adaptation [6]:

$$\boldsymbol{\phi}_{k,i-1} = \sum_{\ell\in\mathcal{N}_k} a_{1,\ell k}\boldsymbol{w}_{\ell,i-1} \tag{5.5}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \mu_k \sum_{\ell\in\mathcal{N}_k} c_{\ell k}\boldsymbol{u}_{\ell,i}^{*}[\boldsymbol{d}_\ell(i) - \boldsymbol{u}_{\ell,i}\boldsymbol{\phi}_{k,i-1}] \tag{5.6}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_k} a_{2,\ell k}\boldsymbol{\psi}_{\ell,i} \tag{5.7}$$

Several diffusion strategies can be obtained as special cases of (5.5)–(5.7) through proper selection of the coefficients $\{a_{1,\ell k}, c_{\ell k}, a_{2,\ell k}\}$. For example, to recover the CTA strategy (5.3), we set $A_2 = I_N$, and to recover the ATC strategy (5.4), we set $A_1 = I_N$, where I_N denotes the $N \times N$ identity matrix. In the general diffusion strategy (5.5)–(5.7), each node k evaluates its estimate $\boldsymbol{w}_{k,i}$ at time i

by relying solely on the data collected from its neighbors through steps (5.5) and (5.7) and on its local measurements through step (5.6). The matrices A_1, A_2, and C are required to be left or right-stochastic, i.e.,

$$A_1^{\mathsf{T}}\mathbb{1}_N = \mathbb{1}_N, \qquad A_2^{\mathsf{T}}\mathbb{1}_N = \mathbb{1}_N, \qquad C\mathbb{1}_N = \mathbb{1}_N \tag{5.8}$$

This means that each node performs a convex combination of the estimates received from its neighbors at every iteration i.

The mean-square performance and convergence properties of the diffusion algorithm (5.5)–(5.7) have already been studied in detail in [5,6]. For the benefit of the analysis in the subsequent sections, we present below in (5.21) the recursion describing the evolution of the weight error vectors across the network. To do so, we introduce the error vectors:

$$\widetilde{\boldsymbol{\phi}}_{k,i-1} \triangleq w^o - \boldsymbol{\phi}_{k,i-1} \tag{5.9}$$

$$\widetilde{\boldsymbol{\psi}}_{k,i} \triangleq w^o - \boldsymbol{\psi}_{k,i} \tag{5.10}$$

$$\widetilde{\boldsymbol{w}}_{k,i} \triangleq w^o - \boldsymbol{w}_{k,i} \tag{5.11}$$

and substitute the linear model (5.1) into the adaptation step (5.6) to find that

$$\widetilde{\boldsymbol{\psi}}_{k,i} = (I_M - \mu_k \boldsymbol{R}_{k,i})\widetilde{\boldsymbol{\phi}}_{k,i-1} - \mu_k \sum_{\ell\in\mathcal{N}_k} c_{\ell k}\boldsymbol{s}_{l,i} \tag{5.12}$$

where the $M \times M$ matrix $\boldsymbol{R}_{k,i}$ and the $M \times 1$ vector $\boldsymbol{s}_{k,i}$ are defined as:

$$\boldsymbol{R}_{k,i} \triangleq \sum_{\ell\in\mathcal{N}_k} c_{\ell k}\boldsymbol{u}_{\ell,i}^*\boldsymbol{u}_{\ell,i} \tag{5.13}$$

$$\boldsymbol{s}_{k,i} \triangleq \boldsymbol{u}_{k,i}^*\boldsymbol{v}_k(i) \tag{5.14}$$

We further collect the various quantities across all nodes in the network into the following block vectors and matrices:

$$\boldsymbol{\mathcal{R}}_i \triangleq \mathrm{diag}\,\{\boldsymbol{R}_{1,i},\ldots,\boldsymbol{R}_{N,i}\} \tag{5.15}$$

$$\boldsymbol{s}_i \triangleq \mathrm{col}\left\{\boldsymbol{s}_{1,i},\ldots,\boldsymbol{s}_{N,i}\right\} \tag{5.16}$$

$$\mathcal{M} \triangleq \mathrm{diag}\left\{\mu_1 I_M,\ldots,\mu_N I_M\right\} \tag{5.17}$$

$$\widetilde{\boldsymbol{\phi}}_i \triangleq \mathrm{col}\left\{\widetilde{\boldsymbol{\phi}}_{1,i},\ldots,\widetilde{\boldsymbol{\phi}}_{N,i}\right\} \tag{5.18}$$

$$\widetilde{\boldsymbol{\psi}}_i \triangleq \mathrm{col}\left\{\widetilde{\boldsymbol{\psi}}_{1,i},\ldots,\widetilde{\boldsymbol{\psi}}_{N,i}\right\} \tag{5.19}$$

$$\widetilde{\boldsymbol{w}}_i \triangleq \mathrm{col}\left\{\widetilde{\boldsymbol{w}}_{1,i},\ldots,\widetilde{\boldsymbol{w}}_{N,i}\right\} \tag{5.20}$$

Then, from (5.5), (5.7), and (5.12), the recursion for the network error vector $\widetilde{\boldsymbol{w}}_i$ is given by

$$\boxed{\widetilde{\boldsymbol{w}}_i = \mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}_i)\mathcal{A}_1^{\mathsf{T}}\widetilde{\boldsymbol{w}}_{i-1} - \mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\boldsymbol{s}_i} \tag{5.21}$$

where

$$\mathcal{A}_1 \triangleq A_1 \otimes I_M, \quad \mathcal{C} \triangleq C \otimes I_M, \quad \mathcal{A}_2 \triangleq A_2 \otimes I_M \tag{5.22}$$

5.1.2 Noisy Information Exchange

Each of the steps in (5.5)–(5.7) involves the sharing of information between node k and its neighbors. For example, in the first step (5.5), all neighbors of node k send their estimates $\boldsymbol{w}_{\ell,i-1}$ to node k. This transmission is generally subject to additive noise and possibly quantization errors. Likewise, steps (5.6) and (5.7) involve the sharing of other pieces of information with node k. These exchange steps can all be subject to perturbations (such as additive noise and quantization errors). One of the objectives of this work is to analyze the *aggregate* effect of these perturbations on general diffusion strategies of the type (5.5)–(5.7) and to propose choices for the combination weights in order to enhance the mean-square performance of the network in the presence of these disturbances. So let us examine what happens when information is exchanged over links with additive

noise. We model the data received by node k from its neighbor ℓ as

$$\boldsymbol{w}_{\ell k,i-1} \triangleq \boldsymbol{w}_{\ell,i-1} + \boldsymbol{v}_{\ell k,i-1}^{(w)} \tag{5.23}$$

$$\boldsymbol{\psi}_{\ell k,i} \triangleq \boldsymbol{\psi}_{\ell,i} + \boldsymbol{v}_{\ell k,i}^{(\psi)} \tag{5.24}$$

$$\boldsymbol{d}_{\ell k}(i) \triangleq \boldsymbol{d}_{\ell}(i) + \boldsymbol{v}_{\ell k}^{(d)}(i) \tag{5.25}$$

$$\boldsymbol{u}_{\ell k,i} \triangleq \boldsymbol{u}_{\ell,i} + \boldsymbol{v}_{\ell k,i}^{(u)} \tag{5.26}$$

where $\boldsymbol{v}_{\ell k,i-1}^{(w)}$ and $\boldsymbol{v}_{\ell k,i}^{(\psi)}$ are $M \times 1$ noise signals, $\boldsymbol{v}_{\ell k,i}^{(u)}$ is a $1 \times M$ noise signal, and $\boldsymbol{v}_{\ell k}^{(d)}(i)$ is a scalar noise signal. Observe further that in (5.23)–(5.26), we are including several sources of information exchange noise. In comparison, references [47, 48, 105] only considered the noise source $\boldsymbol{v}_{\ell k,i-1}^{(w)}$ in (5.23) and one set of combination coefficients $\{a_{1,\ell k}\}$; the other coefficients were set to $c_{\ell k} = a_{2,\ell k} = 0$ for $\ell \neq k$ and $c_{kk} = a_{2,kk} = 1$. In other words, these references only considered (5.23) and the following traditional CTA strategy without exchange of the data $\{\boldsymbol{d}_{\ell}(i), \boldsymbol{u}_{\ell,i}\}$ — compare with (5.3); note that the second step in (5.27) only uses $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$:

$$\begin{cases} \boldsymbol{\phi}_{k,i-1} = \displaystyle\sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \boldsymbol{w}_{\ell,i-1} \\ \boldsymbol{w}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \mu_k \boldsymbol{u}_{k,i}^{*} [\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} \boldsymbol{\phi}_{k,i-1}] \end{cases} \tag{5.27}$$

The analysis that follows examines the aggregate effect of all four noise sources appearing in (5.23)–(5.26), in addition to the three sets of combination coefficients appearing in (5.5)–(5.7). We introduce the following assumption on the statistical properties of the measurement data and noise signals.

Assumption 5.1 (Statistical properties of the variables)**.**

1. *The regression data $\boldsymbol{u}_{k,i}$ are temporally white and spatially independent random variables with zero mean and covariance matrix $R_{u,k} \triangleq \mathbb{E}\, \boldsymbol{u}_{k,i}^{*} \boldsymbol{u}_{k,i} \geq 0$.*

2. *The noise signals* $\boldsymbol{v}_k(i)$, $\boldsymbol{v}_{\ell k,i-1}^{(w)}$, $\boldsymbol{v}_{\ell k}^{(d)}(i)$, $\boldsymbol{v}_{\ell k,i}^{(u)}$, *and* $\boldsymbol{v}_{\ell k,i}^{(\psi)}$ *are temporally white and spatially independent random variables with zero mean and covariances* $\sigma_{v,k}^2$, $R_{v,\ell k}^{(w)}$, $\sigma_{v,\ell k}^2$, $R_{v,\ell k}^{(u)}$, *and* $R_{v,\ell k}^{(\psi)}$, *respectively. In addition,* $R_{v,\ell k}^{(w)}$, $\sigma_{v,\ell k}^2$, $R_{v,\ell k}^{(u)}$, *and* $R_{v,\ell k}^{(\psi)}$ *are all zero if* $\ell \notin \mathcal{N}_k$ *or* $\ell = k$.

3. *The regression data* $\{\boldsymbol{u}_{m,i_1}\}$, *the model noise signals* $\{\boldsymbol{v}_n(i_2)\}$, *and the link noise signals* $\{\boldsymbol{v}_{\ell_1 k_1,j_1}^{(w)}\}$, $\{\boldsymbol{v}_{\ell_2 k_2}^{(d)}(j_2)\}$, $\{\boldsymbol{v}_{\ell_3 k_3,j_3}^{(u)}\}$, *and* $\{\boldsymbol{v}_{\ell_4 k_4,j_4}^{(\psi)}\}$ *are mutually-independent random variables for all indexes:* i_1, i_2, j_1, j_2, j_3, j_4, m, n, ℓ_1, ℓ_2, ℓ_3, ℓ_4, k_1, k_2, k_3, k_4. □

Using the perturbed data (5.23)–(5.26), the diffusion algorithm (5.5)–(5.7) becomes

$$\boldsymbol{\phi}_{k,i-1} = \sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \boldsymbol{w}_{\ell k,i-1} \tag{5.28}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \mu_k \sum_{\ell \in \mathcal{N}_k} c_{\ell k} \boldsymbol{u}_{\ell k,i}^* [\boldsymbol{d}_{\ell k}(i) - \boldsymbol{u}_{\ell k,i} \boldsymbol{\phi}_{k,i-1}] \tag{5.29}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \boldsymbol{\psi}_{\ell k,i} \tag{5.30}$$

where we continue to use the symbols $\{\boldsymbol{\phi}_{k,i-1}, \boldsymbol{\psi}_{k,i}, \boldsymbol{w}_{k,i}\}$ to avoid an explosion of notation. From (5.23)–(5.24), expressions (5.28)–(5.30) can be rewritten as

$$\boldsymbol{\phi}_{k,i-1} = \sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \boldsymbol{w}_{\ell,i-1} + \boldsymbol{v}_{k,i-1}^{(w)} \tag{5.31}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \mu_k \sum_{\ell \in \mathcal{N}_k} c_{\ell k} \boldsymbol{u}_{\ell k,i}^* [\boldsymbol{d}_{\ell k}(i) - \boldsymbol{u}_{\ell k,i} \boldsymbol{\phi}_{k,i-1}] \tag{5.32}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \boldsymbol{\psi}_{\ell,i} + \boldsymbol{v}_{k,i}^{(\psi)} \tag{5.33}$$

where we are introducing the symbols $\boldsymbol{v}_{k,i-1}^{(w)}$ and $\boldsymbol{v}_{k,i}^{(\psi)}$ to denote the aggregate $M \times 1$ zero-mean noise signals defined over the neighborhood of node k:

$$\boldsymbol{v}_{k,i-1}^{(w)} \triangleq \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} a_{1,\ell k} \boldsymbol{v}_{\ell k,i-1}^{(w)} \tag{5.34}$$

$$\boldsymbol{v}_{k,i}^{(\psi)} \triangleq \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} a_{2,\ell k} \boldsymbol{v}_{\ell k,i}^{(\psi)} \tag{5.35}$$

with covariance matrices

$$R_{v,k}^{(w)} \triangleq \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} a_{1,\ell k}^2 R_{v,\ell k}^{(w)} \tag{5.36}$$

$$R_{v,k}^{(\psi)} \triangleq \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} a_{2,\ell k}^2 R_{v,\ell k}^{(\psi)} \tag{5.37}$$

It is worth noting that $R_{v,k}^{(w)}$ and $R_{v,k}^{(\psi)}$ depend on the combination coefficients $\{a_{1,\ell k}\}$ and $\{a_{2,\ell k}\}$, respectively. This property will be taken into account when optimizing over $\{a_{1,\ell k}\}$ and $\{a_{2,\ell k}\}$ in a later section. We further introduce the following scalar zero-mean noise signal:

$$\boldsymbol{v}_{\ell k}(i) \triangleq \boldsymbol{v}_{\ell}(i) + \boldsymbol{v}_{\ell k}^{(d)}(i) - \boldsymbol{v}_{\ell k,i}^{(u)} w^o \tag{5.38}$$

for $\ell \in \mathcal{N}_k \backslash \{k\}$, whose variance is

$$\sigma_{\ell k}^2 \triangleq \sigma_{v,\ell}^2 + \sigma_{v,\ell k}^2 + (w^o)^* R_{v,\ell k}^{(u)} w^o \tag{5.39}$$

To unify the notation, we define $\boldsymbol{v}_{kk}(i) \triangleq \boldsymbol{v}_k(i)$. Then, from (5.1) and (5.25)–(5.26), it is easy to verify that the noisy data $\{\boldsymbol{d}_{\ell k}(i), \boldsymbol{u}_{\ell k,i}\}$ are related via

$$\boldsymbol{d}_{\ell k}(i) = \boldsymbol{u}_{\ell k,i} w^o + \boldsymbol{v}_{\ell k}(i) \tag{5.40}$$

for $\ell \in \mathcal{N}_k$. Continuing with the adaptation step (5.32) and substituting (5.40), we get

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \mu_k \sum_{\ell \in \mathcal{N}_k} c_{\ell k} \boldsymbol{u}_{\ell k,i}^* \Big[\boldsymbol{u}_{\ell k,i} \widetilde{\boldsymbol{\phi}}_{k,i-1} + \boldsymbol{v}_{\ell k}(i) \Big] \tag{5.41}$$

Then, we can derive the following error recursion for node k (compare with (5.12)):

$$\widetilde{\boldsymbol{\psi}}_{k,i} = (I_M - \mu_k \boldsymbol{R}'_{k,i}) \widetilde{\boldsymbol{\phi}}_{k,i-1} - \mu_k \boldsymbol{z}_{k,i} \tag{5.42}$$

where the $M \times M$ matrix $\boldsymbol{R}'_{k,i}$ and the $M \times 1$ vector $\boldsymbol{z}_{k,i}$ are defined as (compare with (5.13) and (5.14)):

$$\boldsymbol{R}'_{k,i} \triangleq \sum_{\ell \in \mathcal{N}_k} c_{\ell k} \boldsymbol{u}^*_{\ell k,i} \boldsymbol{u}_{\ell k,i} \tag{5.43}$$

$$\boldsymbol{z}_{k,i} \triangleq \sum_{\ell \in \mathcal{N}_k} c_{\ell k} \boldsymbol{u}^*_{\ell k,i} \boldsymbol{v}_{\ell k}(i) \tag{5.44}$$

We further introduce the block vectors and matrices:

$$\boldsymbol{\mathcal{R}}'_i \triangleq \operatorname{diag}\left\{\boldsymbol{R}'_{1,i}, \ldots, \boldsymbol{R}'_{N,i}\right\} \tag{5.45}$$

$$\boldsymbol{z}_i \triangleq \operatorname{col}\left\{\boldsymbol{z}_{1,i}, \ldots, \boldsymbol{z}_{N,i}\right\} \tag{5.46}$$

$$\boldsymbol{v}_i^{(w)} \triangleq \operatorname{col}\left\{\boldsymbol{v}_{1,i}^{(w)}, \ldots, \boldsymbol{v}_{N,i}^{(w)}\right\} \tag{5.47}$$

$$\boldsymbol{v}_i^{(\psi)} \triangleq \operatorname{col}\left\{\boldsymbol{v}_{1,i}^{(\psi)}, \ldots, \boldsymbol{v}_{N,i}^{(\psi)}\right\} \tag{5.48}$$

and the corresponding covariance matrices for $\boldsymbol{v}_i^{(w)}$ and $\boldsymbol{v}_i^{(\psi)}$:

$$\mathcal{R}_v^{(w)} \triangleq \operatorname{diag}\left\{R_{v,1}^{(w)}, \ldots, R_{v,N}^{(w)}\right\} \tag{5.49}$$

$$\mathcal{R}_v^{(\psi)} \triangleq \operatorname{diag}\left\{R_{v,1}^{(\psi)}, \ldots, R_{v,N}^{(\psi)}\right\} \tag{5.50}$$

then, from (5.31), (5.33), and (5.42), we arrive at the following recursion for the network weight error vector in the presence of noisy information exchange:

$$\begin{aligned}
\widetilde{\boldsymbol{w}}_i &= \mathcal{A}_2^{\mathsf{T}} \widetilde{\boldsymbol{\psi}}_i - \boldsymbol{v}_i^{(\psi)} \\
&= \mathcal{A}_2^{\mathsf{T}} \left[(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}'_i)\, \widetilde{\boldsymbol{\phi}}_{i-1} - \mathcal{M}\boldsymbol{z}_i \right] - \boldsymbol{v}_i^{(\psi)} \\
&= \mathcal{A}_2^{\mathsf{T}} \left[(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}'_i) \left(\mathcal{A}_1^{\mathsf{T}} \widetilde{\boldsymbol{w}}_{i-1} - \boldsymbol{v}_{i-1}^{(w)} \right) - \mathcal{M}\boldsymbol{z}_i \right] - \boldsymbol{v}_i^{(\psi)}
\end{aligned} \tag{5.51}$$

That is,

$$\boxed{\widetilde{\boldsymbol{w}}_i = \mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}'_i)\mathcal{A}_1^{\mathsf{T}} \widetilde{\boldsymbol{w}}_{i-1} - \mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}'_i)\boldsymbol{v}_{i-1}^{(w)} - \mathcal{A}_2^{\mathsf{T}}\mathcal{M}\boldsymbol{z}_i - \boldsymbol{v}_i^{(\psi)}} \tag{5.52}$$

Compared to the previous error recursion (5.21), the noise terms in (5.52) consist of three parts:

- $\mathcal{A}_2^{\mathsf{T}}\left(I_{NM}-\mathcal{M}\boldsymbol{\mathcal{R}}_i'\right)\boldsymbol{v}_{i-1}^{(w)}$ is contributed by the noise introduced at the information exchange step (5.28) *before* adaptation.

- $\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\boldsymbol{z}_i$ is contributed by the noise introduced at the adaptation step (5.29).

- $\boldsymbol{v}_i^{(\psi)}$ is contributed by the noise introduced at the information-exchange step (5.30) *after* adaptation.

5.2 Convergence in the Mean with a Bias

Given the weight error recursion (5.52), we are now ready to study the mean convergence condition for the diffusion strategy (5.28) – (5.30) in the presence of disturbances during information exchange under Assumption 5.1. Taking expectations of both sides of (5.52), we get

$$\mathbb{E}\widetilde{\boldsymbol{w}}_i=\mathcal{B}\,\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1}-\mathcal{A}_2^{\mathsf{T}}\left(I_{NM}-\mathcal{M}\mathcal{R}'\right)\mathbb{E}\boldsymbol{v}_{i-1}^{(w)}-\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathbb{E}\boldsymbol{z}_i-\mathbb{E}\boldsymbol{v}_i^{(\psi)} \tag{5.53}$$

where

$$\mathcal{B}\triangleq\mathcal{A}_2^{\mathsf{T}}\left(I_{NM}-\mathcal{M}\mathcal{R}'\right)\mathcal{A}_1^{\mathsf{T}} \tag{5.54}$$

$$\mathcal{R}'\triangleq\mathbb{E}\boldsymbol{\mathcal{R}}_i'=\mathrm{diag}\left\{R_1',\ldots,R_N'\right\} \tag{5.55}$$

$$R_k'\triangleq\mathbb{E}\boldsymbol{R}_{k,i}'=\sum_{\ell\in\mathcal{N}_k}c_{\ell k}\left(R_{u,\ell}+R_{v,\ell k}^{(u)}\right) \tag{5.56}$$

From (5.34) – (5.35) and (5.47) – (5.48), it can be verified that

$$\mathbb{E}\boldsymbol{v}_{i-1}^{(w)}=\mathbb{E}\boldsymbol{v}_i^{(\psi)}=0 \tag{5.57}$$

whereas, from (5.44) and Assumption 5.1, we get

$$\mathbb{E}\boldsymbol{z}_{k,i}=\mathbb{E}\left[\sum_{\ell\in\mathcal{N}_k}c_{\ell k}\left(\boldsymbol{u}_{\ell,i}+\boldsymbol{v}_{\ell k,i}^{(u)}\right)^*\left(\boldsymbol{v}_\ell(i)+\boldsymbol{v}_{\ell k}^{(d)}(i)-\boldsymbol{v}_{\ell k,i}^{(u)}w^o\right)\right]$$

$$= -\left(\sum_{\ell\in\mathcal{N}_k} c_{\ell k} R_{v,\ell k}^{(u)}\right) w^o \tag{5.58}$$

Let us define an $NM \times NM$ matrix $\mathcal{R}_{v,c}^{(u)}$ that collects all covariance matrices $\{R_{v,\ell k}^{(u)}\}$, $k, \ell = 1, \ldots, N$, weighted by the corresponding combination coefficients $\{c_{\ell k}\}$, such that its (k,ℓ)th $M \times M$ submatrix is $c_{\ell k} R_{v,\ell k}^{(u)}$. Note that $\mathcal{R}_{v,c}^{(u)}$ itself is *not* a covariance matrix because $c_{kk} R_{v,kk}^{(u)} = 0$ for all k and, in general, $c_{\ell k} R_{v,\ell k}^{(u)} \neq c_{k\ell} R_{v,k\ell}^{(u)}$ when $k \neq \ell$. Then, from (5.46) and (5.58), we arrive at

$$z \triangleq \mathbb{E}\boldsymbol{z}_i = -\mathcal{R}_{v,c}^{(u)}(\mathbb{1}_N \otimes w^o) \tag{5.59}$$

Therefore, using (5.57) and (5.59), expression (5.53) becomes

$$\boxed{\mathbb{E}\widetilde{\boldsymbol{w}}_i = \mathcal{B}\,\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1} - \mathcal{A}_2^{\mathsf{T}}\mathcal{M}z} \tag{5.60}$$

with a driving term due to the presence of z. This driving term would disappear from (5.60) if there were no noise during the exchange of the regression data. To guarantee convergence of (5.60), the coefficient matrix $\mathcal{B}$ must be stable, i.e., $\rho(\mathcal{B}) < 1$. Since $\mathcal{A}_1^{\mathsf{T}}$ and $\mathcal{A}_2^{\mathsf{T}}$ are right-stochastic matrices, it can be shown that the matrix $\mathcal{B}$ is stable whenever $I_{NM} - \mathcal{M}\mathcal{R}'$ itself is stable (see Appendix 5.A). This fact leads to an upper bound on the step-sizes $\{\mu_k\}$ to guarantee the convergence of $\mathbb{E}\widetilde{\boldsymbol{w}}_i$ to a steady-state value, namely, we must have

$$\boxed{\mu_k < \frac{2}{\lambda_{\max}(R_k')}} \tag{5.61}$$

for $k = 1, 2, \ldots, N$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of its matrix argument. Note that the neighborhood covariance matrix R_k' in (5.56) is related to the combination weights $\{c_{\ell k}\}$. If we further assume that C is doubly - stochastic, i.e.,

$$C\mathbb{1}_N = \mathbb{1}_N, \qquad C^{\mathsf{T}}\mathbb{1}_N = \mathbb{1}_N \tag{5.62}$$

then, by Jensen's inequality [71],

$$\begin{aligned}\lambda_{\max}(R_k') &= \lambda_{\max}\left(\sum_{\ell\in\mathcal{N}_k} c_{\ell k}\left(R_{u,\ell} + R_{v,\ell k}^{(u)}\right)\right)\\ &\leq \sum_{\ell\in\mathcal{N}_k} c_{\ell k}\lambda_{\max}\left(R_{u,\ell} + R_{v,\ell k}^{(u)}\right)\\ &\leq \max_{\ell\in\mathcal{N}_k}\lambda_{\max}\left(R_{u,\ell} + R_{v,\ell k}^{(u)}\right) \end{aligned} \tag{5.63}$$

since (i) $\lambda_{\max}(\cdot)$ coincides with the induced 2-norm for any positive semi-definite Hermitian matrix; (ii) matrix norms are convex functions of their arguments [115]; and (iii) by (5.62), $\{c_{\ell k}\}$ are convex combination coefficients. Thus, we obtain a sufficient condition for the convergence of (5.60) in lieu of (5.61):

$$\boxed{\mu_k < \frac{2}{\max_{\ell\in\mathcal{N}_k}[\lambda_{\max}(R_{u,\ell} + R_{v,\ell k}^{(u)})]}} \tag{5.64}$$

for $k = 1, 2, \ldots, N$, where the upper bound for the step-size μ_k becomes independent of the combination weights $\{c_{\ell k}\}$. This bound can be determined solely from knowledge of the covariances of the regression data and the associated noise signals that are accessible to node k. It is worth noting that for traditional diffusion algorithms where information is perfectly exchanged, condition (5.64) reduces to

$$\mu_k < \frac{2}{\max_{\ell\in\mathcal{N}_k}[\lambda_{\max}(R_{u,\ell})]} \tag{5.65}$$

for $k = 1, 2, \ldots, N$. Comparing (5.64) with (5.65), we see that the link noise $\boldsymbol{v}_{\ell k,i}^{(u)}$ over regression data reduces the dynamic range of the step-sizes for mean stability. Now, under (5.61), and taking the limit of (5.60) as $i \to \infty$, we find that the mean error vector will converge to a steady-state value g:

$$g \triangleq \lim_{i\to\infty} \mathbb{E}\widetilde{\boldsymbol{w}}_i = -(I_{NM} - \mathcal{B})^{-1}\mathcal{A}_2^{\mathsf{T}}\mathcal{M}z \tag{5.66}$$

5.3 Mean-Square Convergence Analysis

It is well-known that studying the mean-square convergence of a single adaptive filter is a challenging task, since adaptive filters are nonlinear, time-variant, and stochastic systems. When a network of adaptive nodes is considered, the complexity of the analysis is compounded because the nodes now influence each other's behavior. In order to make the performance analysis more tractable, we rely on the energy conservation approach [46, 116], which was used successfully in [5,6] to study the mean-square performance of diffusion strategies under perfect information exchange conditions. That argument allows us to derive expressions for the mean-square-deviation (MSD) and the excess-mean-square-error (EMSE) of the network by analyzing how energy flows through the nodes.

From recursion (5.52) and under Assumption 5.1, we can obtain the following weighted variance relation for the global error vector $\widetilde{\boldsymbol{w}}_i$:

$$\boxed{\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{\Sigma}^2 &= \mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\Sigma'}^2 + \mathbb{E}\|\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\boldsymbol{z}_i\|_{\Sigma}^2 \\ &\quad - 2\,\mathfrak{Re}\{\mathbb{E}[\boldsymbol{z}_i^*\mathcal{M}\mathcal{A}_2\Sigma\mathcal{A}_2^{\mathsf{T}}(I_{NM}-\mathcal{M}\boldsymbol{\mathcal{R}}_i')\mathcal{A}_1^{\mathsf{T}}\widetilde{\boldsymbol{w}}_{i-1}]\} \\ &\quad + \mathbb{E}\|\mathcal{A}_2^{\mathsf{T}}(I_{NM}-\mathcal{M}\boldsymbol{\mathcal{R}}_i')\boldsymbol{v}_{i-1}^{(w)}\|_{\Sigma}^2 + \mathbb{E}\|\boldsymbol{v}_i^{(\psi)}\|_{\Sigma}^2\end{aligned}} \tag{5.67}$$

where Σ is an arbitrary $NM \times NM$ positive semi-definite Hermitian matrix that we are free to choose. Moreover, the notation $\|x\|_{\Sigma}^2$ stands for the quadratic term $x^*\Sigma x$. The weighting matrix Σ' in (5.67) can be expressed as

$$\Sigma' = \mathcal{B}^*\Sigma\mathcal{B} + O(\mathcal{M}^2) \tag{5.68}$$

where $\mathcal{B}$ is given by (5.54) and $O(\mathcal{M}^2)$ denotes a term on the order of $\mathcal{M}^2$. Evaluating the term $O(\mathcal{M}^2)$ requires knowledge of higher-order statistics of the regression data and link noises, which are not available under current assumptions. However, this term becomes negligible if we introduce a small step-size assumption.

Assumption 5.2 (Small step-sizes)**.** *The step-sizes are sufficiently small, i.e., $\mu_k \ll 1$, such that terms depending on higher-order powers of the step-sizes can be ignored.* □

Hence, in the sequel we use the approximation:

$$\Sigma' \approx \mathcal{B}^* \Sigma \mathcal{B} \tag{5.69}$$

Observe that on the right-hand side (RHS) of relation (5.67), only the first and third terms relate to the error vector $\widetilde{\boldsymbol{w}}_{i-1}$. By Assumption 5.1, the error vector $\widetilde{\boldsymbol{w}}_{i-1}$ is independent of $\boldsymbol{z}_i$ and $\boldsymbol{\mathcal{R}}_i'$. Thus, from (5.59), the third term on RHS of (5.67) can be expressed as

$$\begin{aligned}
&\text{Third term on RHS of (5.67)} \\
&\quad = -2\mathfrak{Re}\{\mathbb{E}[\boldsymbol{z}_i^* \mathcal{M}\mathcal{A}_2\Sigma\mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}_i')\mathcal{A}_1^{\mathsf{T}}]\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1}\} \\
&\quad = -2\mathfrak{Re}[z^* \mathcal{M}\mathcal{A}_2\Sigma\mathcal{A}_2^{\mathsf{T}}\mathcal{A}_1^{\mathsf{T}}\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1} + O(\mathcal{M}^2)]
\end{aligned} \tag{5.70}$$

Since we already showed in the previous section that $\mathbb{E}\widetilde{\boldsymbol{w}}_i$ converges to a fixed bias g, quantity (5.70) will converge to a fixed value as well when $i \to \infty$. Moreover, under Assumption 5.1, the second, fourth, and fifth terms on RHS of relation (5.67) are all fixed values. Therefore, the convergence of relation (5.67) depends on the behavior of the first term $\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\Sigma'}^2$. Although the weighting matrix Σ' of $\widetilde{\boldsymbol{w}}_{i-1}$ is different from the weighting matrix Σ of $\widetilde{\boldsymbol{w}}_i$, it turns out that the entries of these two matrices are approximately related by a linear equation shown ahead in (5.72). Introduce the vector notation [46]:

$$\sigma = \text{vec}(\Sigma), \qquad \sigma' = \text{vec}(\Sigma') \tag{5.71}$$

Then, by using the identity $\text{vec}(ABC) = (C^{\mathsf{T}} \otimes A)\text{vec}(B)$, it can be verified from (5.69) that

$$\sigma' \approx \mathcal{F}\sigma \tag{5.72}$$

where the $N^2M^2 \times N^2M^2$ matrix $\mathcal{F}$ is given by

$$\mathcal{F} \triangleq \mathcal{B}^{\mathsf{T}} \otimes \mathcal{B}^{*} \tag{5.73}$$

and the term $O(\mathcal{M}^2)$ is ignored due to Assumption 5.2. To guarantee mean-square convergence of the algorithm, the step-sizes should be selected to ensure that the matrix $\mathcal{F}$ is stable [46], i.e., $\rho(\mathcal{F}) < 1$, which is equivalent to the earlier condition $\rho(\mathcal{B}) < 1$. Although more specific conditions for mean-square stability can be determined without Assumption 5.2 [46], it is sufficient for our purposes here to conclude that the diffusion strategy (5.28)–(5.30) is stable in the mean and mean-square sense if the step-sizes $\{\mu_k\}$ satisfy (5.61) or (5.64) and are sufficiently small.

5.4 Steady-State Performance Analysis

The conclusion so far is that sufficiently small step-sizes ensure convergence of the diffusion strategy (5.28)–(5.30) in the mean and mean-square sense, even in the presence of exchange noises over the communication links. Let us now determine expressions for the error variances in steady-state. We start from the weighted variance relation (5.67). It shows that the error variance $\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2_{\Sigma}$ depends on the mean error $\mathbb{E}\widetilde{\boldsymbol{w}}_i$. We already determined the value of $\lim_{i\to\infty} \mathbb{E}\widetilde{\boldsymbol{w}}_i$ in (5.66).

5.4.1 Steady-State Variance Relation

We continue to use the vector notation (5.71) and proceed to evaluate all the terms, except the first one, on RHS of (5.67) in the following. For the *second* term, it can be expressed as

$$\text{Second term on RHS of (5.67)} = \text{Tr}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{R}_z\mathcal{M}\mathcal{A}_2\Sigma)$$

$$= \left[\mathrm{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{R}_z\mathcal{M}\mathcal{A}_2)\right]^*\sigma \tag{5.74}$$

where we used the identity $\mathrm{Tr}(W\Sigma) = [\mathrm{vec}(W)]^*\sigma$ for any Hermitian matrix W, and $\mathcal{R}_z$ denotes the autocorrelation matrix of $\boldsymbol{z}_i$. It is shown in Appendix 5.B that $\mathcal{R}_z$ is given by

$$\mathcal{R}_z \triangleq \mathbb{E}\,\boldsymbol{z}_i\boldsymbol{z}_i^* \approx \mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C} + \mathcal{T} + zz^* \tag{5.75}$$

where $\mathcal{C}$ is defined in (5.22), z is in (5.59), and $\{\mathcal{S},\mathcal{T}\}$ are two $NM\times NM$ positive semi-definite block diagonal matrices:

$$\mathcal{S} \triangleq \mathrm{diag}\left\{\sigma_{v,1}^2 R_{u,1},\ldots,\sigma_{v,N}^2 R_{u,N}\right\} \tag{5.76}$$

$$\mathcal{T} \triangleq \mathrm{diag}\left\{T_1,\ldots,T_N\right\} \tag{5.77}$$

$$T_k \triangleq \sum_{\ell\in\mathcal{N}_k} c_{\ell k}^2\left[\left(\sigma_{v,\ell}^2+\sigma_{v,\ell k}^2\right)R_{v,\ell k}^{(u)} + \left(\sigma_{v,\ell k}^2 + (w^o)^* R_{v,\ell k}^{(u)} w^o\right)R_{u,\ell}\right] \tag{5.78}$$

From expression (5.70) and Assumption 5.2, the *third* term on RHS of (5.67) is given by

Third term on RHS of (5.67)

$$\begin{aligned}
&\approx -z^*\mathcal{M}\mathcal{A}_2\Sigma\mathcal{A}_2^{\mathsf{T}}\mathcal{A}_1^{\mathsf{T}}(\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1}) - (\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1})^*\mathcal{A}_1\mathcal{A}_2\Sigma\mathcal{A}_2^{\mathsf{T}}\mathcal{M}z\\
&= -\mathrm{Tr}\left\{\left[\mathcal{A}_2^{\mathsf{T}}\mathcal{A}_1^{\mathsf{T}}(\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1})z^*\mathcal{M}\mathcal{A}_2 + \mathcal{A}_2^{\mathsf{T}}\mathcal{M}z(\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1})^*\mathcal{A}_1\mathcal{A}_2\right]\Sigma\right\}\\
&= -\left[\mathrm{vec}\left(\mathcal{A}_2^{\mathsf{T}}\mathcal{A}_1^{\mathsf{T}}(\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1})z^*\mathcal{M}\mathcal{A}_2 + \mathcal{A}_2^{\mathsf{T}}\mathcal{M}z(\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1})^*\mathcal{A}_1\mathcal{A}_2\right)\right]^*\sigma
\end{aligned} \tag{5.79}$$

Likewise, the *fourth* term on RHS of (5.67) is approximated by

Fourth term on RHS of (5.67)

$$\begin{aligned}
&= \left[\mathrm{vec}\left(\mathbb{E}[\mathcal{A}_2^{\mathsf{T}}(I_{NM}-\mathcal{M}\boldsymbol{\mathcal{R}}_i')\mathcal{R}_v^{(w)}(I_{NM}-\mathcal{M}\boldsymbol{\mathcal{R}}_i')\mathcal{A}_2]\right)\right]^*\sigma\\
&\approx \left[\mathrm{vec}\left(\mathcal{A}_2^{\mathsf{T}}\mathcal{R}_v^{(w)}\mathcal{A}_2\right)\right]^*\sigma
\end{aligned} \tag{5.80}$$

where the terms on the order of $\mathcal{M}$ and $\mathcal{M}^2$ are ignored due to Assumption 5.2, and the *fifth* term on RHS of (5.67) is given by

$$\text{Fifth term on RHS of (5.67)} = \left[\mathrm{vec}(\mathcal{R}_v^{(\psi)})\right]^*\sigma \tag{5.81}$$

Let us introduce

$$\mathcal{R}_v \triangleq \mathcal{A}_2^{\mathsf{T}}\mathcal{R}_v^{(w)}\mathcal{A}_2 + \mathcal{R}_v^{(\psi)} + \mathcal{A}_2^{\mathsf{T}}\mathcal{M}(\mathcal{T} + zz^*)\mathcal{M}\mathcal{A}_2 \tag{5.82}$$

$$\mathcal{Y} \triangleq -\mathcal{A}_2^{\mathsf{T}}\mathcal{A}_1^{\mathsf{T}}gz^*\mathcal{M}\mathcal{A}_2 = \mathcal{A}_2^{\mathsf{T}}\mathcal{A}_1^{\mathsf{T}}\left(I_{NM} - \mathcal{B}\right)^{-1}\mathcal{A}_2^{\mathsf{T}}\mathcal{M}zz^*\mathcal{M}\mathcal{A}_2 \tag{5.83}$$

At steady-state, as $i \to \infty$, by (5.66) and (5.74) – (5.83), the weighted variance relation (5.67) becomes

$$\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_\sigma^2 \approx \lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\mathcal{F}\sigma}^2 + \left[\mathrm{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v + \mathcal{Y} + \mathcal{Y}^*)\right]^*\sigma \tag{5.84}$$

where we are using the compact notation $\|x\|_\sigma^2$ to refer to $\|x\|_\Sigma^2$ — doing so allows us to represent Σ' by the more compact relation $\mathcal{F}\sigma$ on RHS of (5.84); we shall be using the weighting matrix Σ and its vector representation σ interchangeably for ease of notation (likewise, for Σ' and σ'). The steady-state weighted variance relation (5.84) can be rewritten as

$$\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{(I_{N^2M^2}-\mathcal{F})\sigma}^2 \approx \left[\mathrm{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v + \mathcal{Y} + \mathcal{Y}^*)\right]^*\sigma \tag{5.85}$$

where the term $\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2$ is contributed by the model noise $\{\boldsymbol{v}_k(i)\}$ while the remaining terms $\{\mathcal{R}_v, \mathcal{Y}\}$ are contributed by the link noises $\boldsymbol{v}_{\ell k,i-1}^{(w)}$, $\boldsymbol{v}_{\ell k}^{(d)}(i)$, $\boldsymbol{v}_{\ell k,i}^{(u)}$, and $\boldsymbol{v}_{\ell k,i}^{(\psi)}$. Recall that we are free to choose Σ and, hence, σ. Let $(I_{N^2M^2} - \mathcal{F})\sigma = \mathrm{vec}(\Omega)$, where Ω is another arbitrary positive semi-definite Hermitian matrix. Then, we arrive at the following theorem.

Theorem 5.1 (Steady-state variance relation)**.** *Under Assumptions 5.1 and 5.2, for any positive semi-definite Hermitian matrix Ω, the steady-state weighted error variance relation of the diffusion strategy* (5.28)–(5.30) *is approximately given by*

$$\boxed{\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_\Omega^2 \approx \left[\mathrm{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v + \mathcal{Y} + \mathcal{Y}^*)\right]^*\left(I_{N^2M^2} - \mathcal{F}\right)^{-1}\mathrm{vec}(\Omega)} \tag{5.86}$$

where $\mathcal{S}$ is given in (5.76)*, $\mathcal{R}_v$ in* (5.82)*, $\mathcal{Y}$ in* (5.83)*, and $\mathcal{F}$ in* (5.73)*.* □

5.4.2 Network MSD and EMSE

Each subvector of $\widetilde{\boldsymbol{w}}_i$ corresponds to the estimation error at a particular node, say, $\widetilde{\boldsymbol{w}}_{k,i}$ for node k. The network MSD is defined as [46]:

$$\text{MSD} \triangleq \lim_{i\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 \tag{5.87}$$

Since we are free to choose Ω, we select it as $\Omega = I_{NM}/N$. Then, expression (5.86) gives

$$\boxed{\text{MSD} \approx \frac{1}{N}\left[\text{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v + \mathcal{Y} + \mathcal{Y}^*)\right]^* (I_{N^2M^2} - \mathcal{F})^{-1}\text{vec}(I_{NM})} \tag{5.88}$$

Similarly, if we instead select $\Omega = \mathcal{R}_u/N$, where

$$\mathcal{R}_u \triangleq \text{diag}\,\{R_{u,1}, \ldots, R_{u,N}\} \tag{5.89}$$

then expression (5.86) would allow us to evaluate the network EMSE as:

$$\boxed{\text{EMSE} \approx \frac{1}{N}\left[\text{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v + \mathcal{Y} + \mathcal{Y}^*)\right]^* (I_{N^2M^2} - \mathcal{F})^{-1}\text{vec}(\mathcal{R}_u)} \tag{5.90}$$

where the network EMSE is defined as follows:

$$\begin{aligned}\text{EMSE} &\triangleq \lim_{i\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}|\boldsymbol{u}_{k,i}\widetilde{\boldsymbol{w}}_{k,i-1}|^2 \\ &= \lim_{i\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2_{\mathcal{R}_u}\end{aligned} \tag{5.91}$$

5.4.3 Simplifications when Regression Data are not Shared

We showed in the earlier sections that the link noise over regression data biases the weight estimators. In this section we examine how the results simplify when there is no sharing of regression data among the nodes.

Assumption 5.3 (No sharing of regression data)**.** *Nodes do not share regression data within neighborhoods, i.e., assume* $C = I_N$. □

By Assumptions 5.2 and 5.3, matrices $\{\mathcal{B}, \mathcal{R}_v, \mathcal{Y}\}$ in (5.54), (5.82), and (5.83) become

$$\mathcal{B} = \mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\mathcal{R}_u)\mathcal{A}_1^{\mathsf{T}} \tag{5.92}$$

$$\mathcal{R}_v = \mathcal{A}_2^{\mathsf{T}}\mathcal{R}_v^{(w)}\mathcal{A}_2 + \mathcal{R}_v^{(\psi)} \tag{5.93}$$

$$\mathcal{Y} = 0 \tag{5.94}$$

where $\mathcal{R}_u$ is given in (5.89). Then, the network MSD and EMSE expressions (5.88) and (5.90) simplify to:

$$\boxed{\text{MSD} \approx \frac{1}{N}\left[\text{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v)\right]^*(I_{N^2M^2} - \mathcal{F})^{-1}\text{vec}(I_{NM})} \tag{5.95}$$

and

$$\boxed{\text{EMSE} \approx \frac{1}{N}\left[\text{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v)\right]^*(I_{N^2M^2} - \mathcal{F})^{-1}\text{vec}(\mathcal{R}_u)} \tag{5.96}$$

5.4.4 Dependence of Performance on Combination Weights and Link Noise

Recalling that $\mathcal{R}_v$ and $\mathcal{F}$ are related to the combination matrices $\{\mathcal{A}_1, \mathcal{A}_2\}$, or, equivalently, $\{A_1, A_2\}$, results (5.95) and (5.96) express the network MSD and EMSE in terms of $\{A_1, A_2\}$. However, it is generally difficult to use these expressions to optimize over $\{A_1, A_2\}$ to reduce the impact of link noise. Instead, by substituting (5.73) into (5.95) and using the fact that $\mathcal{F}$ is stable, we can arrive at another useful expression for the network MSD:

$$\text{MSD} \approx \frac{1}{N}\left[\text{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v)\right]^*\sum_{j=0}^{\infty}\mathcal{F}^j\text{vec}(I_{NM})$$

$$= \frac{1}{N}\left[\mathrm{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2+\mathcal{R}_v)\right]^*\sum_{j=0}^{\infty}(\mathcal{B}^{\mathsf{T}}\otimes\mathcal{B}^*)^j\mathrm{vec}(I_{NM})$$

$$= \frac{1}{N}\left[\mathrm{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2+\mathcal{R}_v)\right]^*\sum_{j=0}^{\infty}\mathrm{vec}(\mathcal{B}^{*j}\mathcal{B}^j) \tag{5.97}$$

That is,

$$\boxed{\mathrm{MSD}\approx\frac{1}{N}\sum_{j=0}^{\infty}\mathrm{Tr}\left[\mathcal{B}^j(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2+\mathcal{R}_v)\mathcal{B}^{*j}\right]} \tag{5.98}$$

where $\mathcal{B}$ is given in (5.92). Similarly, the network EMSE can be expressed as

$$\boxed{\mathrm{EMSE}\approx\frac{1}{N}\sum_{j=0}^{\infty}\mathrm{Tr}\left[\mathcal{B}^j(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2+\mathcal{R}_v)\mathcal{B}^{*j}\mathcal{R}_u\right]} \tag{5.99}$$

Expressions (5.98) and (5.99) reveal in an interesting way how the noise sources originating from any particular node end up influencing the overall network performance. Let us denote

$$\boldsymbol{\mathcal{B}}_i \triangleq \mathcal{A}_2^{\mathsf{T}}(I_{NM}-\mathcal{M}\boldsymbol{\mathcal{R}}_i')\mathcal{A}_1^{\mathsf{T}} \tag{5.100}$$

$$\boldsymbol{\theta}_i \triangleq \mathcal{A}_2^{\mathsf{T}}(I_{NM}-\mathcal{M}\boldsymbol{\mathcal{R}}_i')\boldsymbol{v}_{i-1}^{(w)}+\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\boldsymbol{z}_i+\boldsymbol{v}_i^{(\psi)} \tag{5.101}$$

The error recursion (5.52) can be rewritten as

$$\widetilde{\boldsymbol{w}}_i=\boldsymbol{\mathcal{B}}_i\widetilde{\boldsymbol{w}}_{i-1}-\boldsymbol{\theta}_i=\boldsymbol{\Phi}_{0,i}\widetilde{\boldsymbol{w}}_{-1}-\sum_{m=0}^{i}\boldsymbol{\Phi}_{m+1,i}\boldsymbol{\theta}_m \tag{5.102}$$

where

$$\boldsymbol{\Phi}_{m,i}\triangleq\begin{cases}\boldsymbol{\mathcal{B}}_i\boldsymbol{\mathcal{B}}_{i-1}\ldots\boldsymbol{\mathcal{B}}_m, & i\geq m\\ I_{NM}, & i<m\end{cases} \tag{5.103}$$

Then,

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2=\mathbb{E}\,\|\boldsymbol{\Phi}_{0,i}\widetilde{\boldsymbol{w}}_{-1}\|^2+\mathbb{E}\left\|\sum_{m=0}^{i}\boldsymbol{\Phi}_{m+1,i}\boldsymbol{\theta}_m\right\|^2 \tag{5.104}$$

Under Assumption 5.3, $\{\boldsymbol{\mathcal{B}}_i,\boldsymbol{\theta}_i\}$ in (5.100)–(5.101) can be simplified as

$$\boldsymbol{\mathcal{B}}_i=\mathcal{A}_2^{\mathsf{T}}(I_{NM}-\mathcal{M}\boldsymbol{\mathcal{R}}_i)\mathcal{A}_1^{\mathsf{T}} \tag{5.105}$$

$$\boldsymbol{\theta}_i = \mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}_i)\boldsymbol{v}_{i-1}^{(w)} + \mathcal{A}_2^{\mathsf{T}}\mathcal{M}\boldsymbol{s}_i + \boldsymbol{v}_i^{(\psi)} \tag{5.106}$$

where $\{\boldsymbol{\mathcal{R}}_i, \boldsymbol{s}_i\}$ are given in (5.15) and (5.16). By Assumption 5.1, $\{\boldsymbol{\mathcal{B}}_i, \boldsymbol{\theta}_i\}$ are temporally independent for different i and

$$\mathbb{E}\,\boldsymbol{\mathcal{B}}_i = \mathcal{B}, \qquad \mathbb{E}\,\boldsymbol{\theta}_i = 0 \tag{5.107}$$

where $\mathcal{B}$ is given by (5.92). As $i \to \infty$, the first term on RHS of (5.104) becomes

$$\begin{aligned}
\text{First term on RHS of (5.104)} &= \lim_{i\to\infty} \operatorname{Tr}\left\{\mathbb{E}\left[\boldsymbol{\Phi}_{0,i}(\mathbb{E}\widetilde{\boldsymbol{w}}_{-1}\widetilde{\boldsymbol{w}}_{-1}^*)\boldsymbol{\Phi}_{0,i}^*\right]\right\} \\
&\overset{(a)}{\approx} \lim_{i\to\infty} \operatorname{Tr}\left[\left(\mathbb{E}\boldsymbol{\Phi}_{0,i}\right)\left(\mathbb{E}\widetilde{\boldsymbol{w}}_{-1}\widetilde{\boldsymbol{w}}_{-1}^*\right)\left(\mathbb{E}\boldsymbol{\Phi}_{0,i}\right)^*\right] \\
&= \lim_{i\to\infty} \operatorname{Tr}\left[\mathcal{B}^{i+1}(\mathbb{E}\,\widetilde{\boldsymbol{w}}_{-1}\widetilde{\boldsymbol{w}}_{-1}^*)\mathcal{B}^{(i+1)*}\right] \\
&\overset{(b)}{=} 0
\end{aligned} \tag{5.108}$$

where (a) is obtained by approximating the expectation of the product by the product of expectations and (b) is due to the stability of $\mathcal{B}$. Therefore, the steady-state value of (5.104) gives

$$\begin{aligned}
\lim_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 &= \lim_{i\to\infty} \sum_{m=0}^{i} \mathbb{E}\,\left\|\boldsymbol{\Phi}_{m+1,i}\boldsymbol{\theta}_m\right\|^2 \\
&\approx \lim_{i\to\infty} \sum_{m=0}^{i} \operatorname{Tr}\left[\left(\mathbb{E}\boldsymbol{\Phi}_{m+1,i}\right)\left(\mathbb{E}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^*\right)\left(\mathbb{E}\boldsymbol{\Phi}_{m+1,i}\right)^*\right] \\
&\overset{(a)}{\approx} \lim_{i\to\infty} \sum_{m=0}^{i} \operatorname{Tr}\left[\mathcal{B}^{i-m}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v)\mathcal{B}^{(i-m)*}\right] \\
&\overset{(b)}{=} \lim_{i\to\infty} \sum_{j=0}^{i} \operatorname{Tr}\left[\mathcal{B}^{j}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v)\mathcal{B}^{j*}\right] \\
&= \sum_{j=0}^{\infty} \operatorname{Tr}\left[\mathcal{B}^{j}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v)\mathcal{B}^{*j}\right]
\end{aligned} \tag{5.109}$$

where, by (5.93) and (5.106), (a) is due to

$$\mathbb{E}\boldsymbol{\theta}_m\boldsymbol{\theta}_m^* = \mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\mathcal{R}_u)\mathcal{R}_v^{(w)}(I_{NM} - \mathcal{M}\mathcal{R}_u)\mathcal{A}_2 + \mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v^{(\psi)}$$

$$\approx \mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2 + \mathcal{A}_2^{\mathsf{T}}\mathcal{R}_v^{(w)}\mathcal{A}_2 + \mathcal{R}_v^{(\psi)}$$
$$= \mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_v \tag{5.110}$$

and (b) is simply a change of variable: $j = i - m$. Since the jth term of the summation in (5.98) or (5.109) is contributed by the term $\mathbb{E}\boldsymbol{\theta}_{i-j}\boldsymbol{\theta}_{i-j}^*$, which consists of all the noise sources at time $i-j$, expression (5.98) shows how various sources of noises are involved and how they contribute to the MSD.

5.5 Optimizing the Combination Matrices

Before we optimize the combination matrices $\{A_1, A_2\}$, we first specialize the MSD expression (5.98) and the EMSE expression (5.99) for the ATC and CTA algorithms. For the ATC algorithm, we set $A_1 = I_N$ and $A_2 = A$, and for the CTA algorithm, we set $A_1 = A$ and $A_2 = I_N$. Let us denote

$$\mathcal{A} \triangleq A \otimes I_M \tag{5.111}$$
$$\mathcal{B}_{\text{atc}} \triangleq \mathcal{A}^{\mathsf{T}}(I_{NM} - \mathcal{M}\mathcal{R}_u) \tag{5.112}$$
$$\mathcal{B}_{\text{cta}} \triangleq (I_{NM} - \mathcal{M}\mathcal{R}_u)\mathcal{A}^{\mathsf{T}} \tag{5.113}$$

Then, we get

$$\text{MSD}_{\text{atc}} \approx \frac{1}{N}\sum_{j=0}^{\infty} \text{Tr}\left[\mathcal{B}_{\text{atc}}^j(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A} + \mathcal{R}_v^{(\psi)})\mathcal{B}_{\text{atc}}^{*j}\right] \tag{5.114}$$

$$\text{EMSE}_{\text{atc}} \approx \frac{1}{N}\sum_{j=0}^{\infty} \text{Tr}\left[\mathcal{B}_{\text{atc}}^j(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A} + \mathcal{R}_v^{(\psi)})\mathcal{B}_{\text{atc}}^{*j}\mathcal{R}_u\right] \tag{5.115}$$

and

$$\text{MSD}_{\text{cta}} \approx \frac{1}{N}\sum_{j=0}^{\infty} \text{Tr}\left[\mathcal{B}_{\text{cta}}^j(\mathcal{M}\mathcal{S}\mathcal{M} + \mathcal{R}_v^{(w)})\mathcal{B}_{\text{cta}}^{*j}\right] \tag{5.116}$$

$$\text{EMSE}_{\text{cta}} \approx \frac{1}{N}\sum_{j=0}^{\infty} \text{Tr}\left[\mathcal{B}_{\text{cta}}^j(\mathcal{M}\mathcal{S}\mathcal{M} + \mathcal{R}_v^{(w)})\mathcal{B}_{\text{cta}}^{*j}\mathcal{R}_u\right] \tag{5.117}$$

5.5.1 An Upper Bound for MSD

Minimizing the MSD expression (5.114) or the EMSE expression (5.115) for the ATC algorithm over left-stochastic matrices A is generally nontrivial. We pursue an approximate solution that relies on optimizing an upper bound and performs well in practice. Let us use $\|X\|_*$ to denote the nuclear norm (also known as the trace norm, or the Ky Fan n-norm) of matrix X [87], which is defined as the sum of the singular values of X. Therefore, $\|X\|_* = \|X^*\|_*$ for any X and $\|X\|_* = \mathrm{Tr}(X)$ when X is Hermitian and positive semi-definite. Let us also denote $\|X\|_{b,\infty}$ as the block maximum norm of matrix X (see Appendix 5.A). Then,

$$\begin{aligned}
&\mathrm{Tr}\left[\mathcal{B}_{\mathrm{atc}}^{j}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_v^{(\psi)})\mathcal{B}_{\mathrm{atc}}^{*j}\right] \\
&\quad=\left\|\mathcal{B}_{\mathrm{atc}}^{j}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_v^{(\psi)})\mathcal{B}_{\mathrm{atc}}^{*j}\right\|_* \\
&\quad\le \|\mathcal{B}_{\mathrm{atc}}^{j}\|_*\cdot\|\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_v^{(\psi)}\|_*\cdot\|\mathcal{B}_{\mathrm{atc}}^{*j}\|_* \\
&\quad\le c^2\cdot\|\mathcal{B}_{\mathrm{atc}}^{j}\|_{b,\infty}^2\cdot\mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_v^{(\psi)}) \\
&\quad\le c^2\cdot\|\mathcal{B}_{\mathrm{atc}}\|_{b,\infty}^{2j}\cdot\mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_v^{(\psi)}) \\
&\quad\le c^2\cdot(\|\mathcal{A}\|_{b,\infty}\cdot\|I_{NM}-\mathcal{M}\mathcal{R}_u\|_{b,\infty})^{2j}\mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_v^{(\psi)}) \\
&\quad= c^2\cdot\rho(I_{NM}-\mathcal{M}\mathcal{R}_u)^{2j}\cdot\mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_v^{(\psi)})
\end{aligned}\tag{5.118}$$

where c is some positive scalar such that $\|X\|_* \le c\,\|X\|_{b,\infty}$ because $\|X\|_*$ and $\|X\|_{b,\infty}$ are submultiplicative norms and all such norms are equivalent [115]. In the last step of (5.118) we used Lemmas 5.2 and 5.3 from Appendix 5.A. Thus, we can upper bound the network MSD (5.114) by

$$\begin{aligned}
\mathrm{MSD}_{\mathrm{atc}} &\le \frac{1}{N}\sum_{j=0}^{\infty} c^2\cdot\rho(I_{NM}-\mathcal{M}\mathcal{R}_u)^{2j}\mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_v^{(\psi)}) \\
&= \frac{c^2}{N}\cdot\frac{\mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_v^{(\psi)})}{1-[\rho(I_{NM}-\mathcal{M}\mathcal{R}_u)]^2}
\end{aligned}\tag{5.119}$$

where the combination matrix $\mathcal{A}$ appears only in the numerator.

5.5.2 Minimizing the Upper Bound

The result (5.119) motivates us to consider instead the problem of minimizing the upper bound, namely,

$$\begin{aligned} \underset{A}{\text{minimize}} \quad & \mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A} + \mathcal{R}_v^{(\psi)}) \\ \text{subject to} \quad & A^{\mathsf{T}}\mathbb{1} = \mathbb{1}, \;\; a_{\ell k} \geq 0, \;\; a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k \end{aligned} \tag{5.120}$$

Using (5.50), the cost function in (5.120) can be expressed as

$$\mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A} + \mathcal{R}_v^{(\psi)}) = \sum_{k=1}^{N} \sum_{\ell \in \mathcal{N}_k} a_{\ell k}^2 \left[\mu_\ell^2 \sigma_{v,\ell}^2 \mathrm{Tr}(R_{u,\ell}) + \mathrm{Tr}(R_{v,\ell k}^{(\psi)}) \right] \tag{5.121}$$

Problem (5.120) can therefore be decoupled into N separate optimization problems of the form:

$$\boxed{\begin{aligned} \underset{\{a_{\ell k},\, \ell \in \mathcal{N}_k\}}{\text{minimize}} \quad & \sum_{\ell \in \mathcal{N}_k} a_{\ell k}^2 \left[\mu_\ell^2 \sigma_{v,\ell}^2 \mathrm{Tr}(R_{u,\ell}) + \mathrm{Tr}(R_{v,\ell k}^{(\psi)}) \right] \\ \text{subject to} \quad & \sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1, \quad a_{\ell k} \geq 0, \quad a_{\ell k} = 0 \;\text{ if }\; \ell \notin \mathcal{N}_k \end{aligned}} \tag{5.122}$$

for $k = 1, \ldots, N$. With each node $\ell \in \mathcal{N}_k$, we associate the following nonnegative *variance product* meansure:

$$\gamma_{\ell k}^2 \triangleq \begin{cases} \mu_k^2 \sigma_{v,k}^2 \mathrm{Tr}(R_{u,k}), & \ell = k \\ \mu_\ell^2 \sigma_{v,\ell}^2 \mathrm{Tr}(R_{u,\ell}) + \mathrm{Tr}(R_{v,\ell k}^{(\psi)}), & \ell \in \mathcal{N}_k \backslash \{k\} \end{cases} \tag{5.123}$$

This measure incorporates information about the link noise covariances $\{R_{v,\ell k}^{(\psi)}\}$. The solution of (5.122) is then given by

$$\boxed{a_{\ell k} = \begin{cases} \dfrac{\gamma_{\ell k}^{-2}}{\sum_{m \in \mathcal{N}_k} \gamma_{mk}^{-2}}, & \text{if } \ell \in \mathcal{N}_k \\ 0, & \text{otherwise} \end{cases}} \tag{5.124}$$

We refer to this combination rule as the relative variance combination rule; it is an extension of the rule devised in [117] to the case of noisy information exchanges. In particular, the definition of the scalars $\{\gamma_{\ell k}^2\}$ in (5.123) is different and now depends on both subscripts ℓ and k.

Minimizing the EMSE expression (5.115) for the ATC algorithm over left-stochastic matrices A can be pursued in a similar manner by noting that

$$\begin{aligned}&\mathrm{Tr}\left[\mathcal{B}_{\mathrm{atc}}^{j}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_{v}^{(\psi)})\mathcal{B}_{\mathrm{atc}}^{*j}\mathcal{R}_{u}\right]\\&\quad\leq c^{2}[\rho(I_{NM}-\mathcal{M}\mathcal{R}_{u})]^{2j}\,\mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}+\mathcal{R}_{v}^{(\psi)})\,\mathrm{Tr}(\mathcal{R}_{u})\end{aligned}\tag{5.125}$$

Thus, minimizing the upper bound of the network EMSE leads to the same solution (5.124). Using the same argument, we can also show that the same result minimizes the upper bound of the network MSD or EMSE for the CTA algorithm.

5.5.3 Adaptive Combination Rule

To apply the relative variance combination rule (5.124), each node k needs to know the variance products, $\{\gamma_{\ell k}^2\}$, of their neighbors, which in general are not available since they require knowledge of the quantities $\{\sigma_{v,\ell}^2, \mathrm{Tr}(R_{u,\ell}), \mathrm{Tr}(R_{v,\ell k}^{(\psi)})\}$. Therefore, we now propose an adaptive combination rule, by using data that are available to the individual nodes. For the ATC algorithm, we first note from (5.24) and (5.29) that

$$\mathbb{E}\|\boldsymbol{\psi}_{\ell k,i}-\boldsymbol{w}_{\ell,i-1}\|^{2}\approx\mu_{\ell}^{2}\sigma_{v,\ell}^{2}\mathrm{Tr}(R_{u,\ell})+\mathrm{Tr}(R_{v,\ell k}^{(\psi)})=\gamma_{\ell k}^{2}\tag{5.126}$$

for $\ell\in\mathcal{N}_k\backslash\{k\}$. Since the algorithm converges in the mean and mean-square senses under Assumption 5.2, all the estimates $\{\boldsymbol{w}_{k,i}\}$ tend to w^o as $i\to\infty$. This allows us to estimate $\gamma_{\ell k}^2$ for node k by using instantaneous realizations of

$\|\boldsymbol{\psi}_{\ell k,i} - \boldsymbol{w}_{k,i-1}\|^2$, where we replace $\boldsymbol{w}_{\ell,i-1}$ by $\boldsymbol{w}_{k,i-1}$. Similarly, for node k itself, we can use realizations of $\|\boldsymbol{\psi}_{k,i} - \boldsymbol{w}_{k,i-1}\|^2$ to estimate γ_{kk}^2. To unify the notation, we define $\boldsymbol{\psi}_{kk,i} \triangleq \boldsymbol{\psi}_{k,i}$. Let $\widehat{\boldsymbol{\gamma}}_{\ell k}^2(i)$ denote an estimator for $\gamma_{\ell k}^2$ that is computed by node k at time i. Then, one way to evaluate $\widehat{\boldsymbol{\gamma}}_{\ell k}^2(i)$ is through the recursion:

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\ell k}^2(i) = (1-\nu_k)\widehat{\boldsymbol{\gamma}}_{\ell k}^2(i-1) + \nu_k \|\boldsymbol{\psi}_{\ell k,i} - \boldsymbol{w}_{k,i-1}\|^2} \tag{5.127}$$

for $\ell \in \mathcal{N}_k$, where $\nu_k \in (0,1)$ is a forgetting factor that is usually close to one. In this way, we arrive at the adaptive combination rule:

$$\boxed{\boldsymbol{a}_{\ell k}(i) = \begin{cases} \dfrac{[\widehat{\boldsymbol{\gamma}}_{\ell k}^2(i)]^{-1}}{\sum_{m\in\mathcal{N}_k}[\widehat{\boldsymbol{\gamma}}_{mk}^2(i)]^{-1}}, & \text{if } \ell \in \mathcal{N}_k \\ 0, & \text{otherwise} \end{cases}} \tag{5.128}$$

5.6 Mean-Square Tracking Behavior

The diffusion strategy (5.5)–(5.7) is adaptive in nature. One of the main benefits of adaptation (by using constant step-sizes) is that it endows networks with tracking abilities when the underlying weight vector w^o varies with time. In this section we analyze how well an adaptive network is able to track variations in w^o. To do so, we adopt a random-walk model for w^o that is commonly used in the literature to describe the non-stationarity of the weight vector [46].

Assumption 5.4 (Random - walk model)**.** *The weight vector w^o changes according to the model:*

$$\boldsymbol{w}_i^o = \boldsymbol{w}_{i-1}^o + \boldsymbol{\eta}_i \tag{5.129}$$

where $\{\boldsymbol{w}_i^o\}$ has a constant mean w^o for all i, $\{\boldsymbol{\eta}_i\}$ is an i.i.d. random sequence with zero mean and covariance matrix R_η; the sequence $\{\boldsymbol{\eta}_i\}$ is independent of the initial conditions $\{\boldsymbol{w}_{-1}^o, \boldsymbol{w}_{k,-1}\}$ and of all regression data and noise signals across the network for all time instants. □

We now define the error vector at node k as

$$\widetilde{\boldsymbol{w}}_{k,i} \triangleq \boldsymbol{w}_i^o - \boldsymbol{w}_{k,i} \tag{5.130}$$

so that the global error recursion (5.52) for the network is replaced by

$$\begin{aligned}\widetilde{\boldsymbol{w}}_i = {} & \mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}_i')\mathcal{A}_1^{\mathsf{T}}\widetilde{\boldsymbol{w}}_{i-1} + \mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}_i')\mathcal{A}_1^{\mathsf{T}}\boldsymbol{\zeta}_i \\ & - \mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}_i')\boldsymbol{v}_{i-1}^{(w)} - \mathcal{A}_2^{\mathsf{T}}\mathcal{M}\boldsymbol{z}_i - \boldsymbol{v}_i^{(\psi)}\end{aligned} \tag{5.131}$$

where the $NM \times 1$ vector $\boldsymbol{\zeta}_i$ is defined as

$$\boldsymbol{\zeta}_i \triangleq \operatorname{col}\{\boldsymbol{\eta}_i, \ldots, \boldsymbol{\eta}_i\} = \mathbb{1}_N \otimes \boldsymbol{\eta}_i \tag{5.132}$$

5.6.1 Convergence Conditions

By Assumptions 5.1 and 5.4, it can be verified that the condition for mean convergence continues to be $\rho(\mathcal{B}) < 1$, where $\mathcal{B}$ is defined in (5.54). In addition, it can also be verified that the error recursion (5.131) converges in the mean sense to the same non - zero bias vector g as in (5.66). From (5.131) and under Assumption 5.2, we can derive the weighted variance relation:

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_\sigma^2 \approx {} & \mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\mathcal{F}\sigma}^2 + \mathbb{E}\|\mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}_i')\mathcal{A}_1^{\mathsf{T}}\boldsymbol{\zeta}_i\|_\sigma^2 \\ & - 2\,\mathfrak{Re}\{\mathbb{E}[\boldsymbol{z}_i^*\mathcal{M}\mathcal{A}_2\Sigma\mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}_i')\mathcal{A}_1^{\mathsf{T}}\widetilde{\boldsymbol{w}}_{i-1}]\} \\ & + \mathbb{E}\|\mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{R}}_i')\boldsymbol{v}_{i-1}^{(w)}\|_\sigma^2 + \mathbb{E}\|\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\boldsymbol{z}_i\|_\sigma^2 + \mathbb{E}\|\boldsymbol{v}_i^{(\psi)}\|_\sigma^2\end{aligned} \tag{5.133}$$

where $\mathcal{F}$ is given in (5.73). If the step-sizes are sufficiently small, then we can assume that the network continues to be mean-square stable.

5.6.2 Steady-State Performance

The steady-state performance is affected by the non-stationarity of w^o. From Assumption 5.2, at steady-state, expression (5.133) becomes

$$\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{\Omega}^2 \approx \left[\mathrm{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2 + \mathcal{A}_2^{\mathsf{T}}\mathcal{A}_1^{\mathsf{T}}\mathcal{R}_\zeta\mathcal{A}_1\mathcal{A}_2 + \mathcal{R}_v + \mathcal{Y} + \mathcal{Y}^*)\right]^* \times (I_{N^2M^2} - \mathcal{F})^{-1}\mathrm{vec}(\Omega) \tag{5.134}$$

where $\mathcal{S}$ is given in (5.76), R_v in (5.82), $\mathcal{Y}$ in (5.83), $\mathcal{F}$ in (5.73), and $\mathcal{R}_\zeta$ is the covariance matrix of $\boldsymbol{\zeta}_i$:

$$\mathcal{R}_\zeta \triangleq \mathbb{E}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^* = (\mathbb{1}_N\mathbb{1}_N^{\mathsf{T}}) \otimes R_\eta \tag{5.135}$$

By (5.8), (5.22), and (5.135), we get

$$\begin{aligned}\mathcal{A}_2^{\mathsf{T}}\mathcal{A}_1^{\mathsf{T}}\mathcal{R}_\zeta\mathcal{A}_1\mathcal{A}_2 &= (A_2^{\mathsf{T}}A_1^{\mathsf{T}}\mathbb{1}_N\mathbb{1}_N^{\mathsf{T}}A_1A_2) \otimes R_\eta \\ &= (\mathbb{1}_N\mathbb{1}_N^{\mathsf{T}}) \otimes R_\eta \\ &= \mathcal{R}_\zeta\end{aligned} \tag{5.136}$$

Then, following the same argument that led to (5.88), we find that the network MSD is now given by:

$$\boxed{\begin{aligned}\mathrm{MSD}_{\mathrm{track}} \approx \frac{1}{N}[\mathrm{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_\zeta + \mathcal{R}_v + \mathcal{Y} + \mathcal{Y}^*)]^* \\ \times (I_{N^2M^2} - \mathcal{F})^{-1}\mathrm{vec}(I_{NM})\end{aligned}} \tag{5.137}$$

Similarly, the network EMSE is given by:

$$\boxed{\begin{aligned}\mathrm{EMSE}_{\mathrm{track}} \approx \frac{1}{N}[\mathrm{vec}(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{C}^{\mathsf{T}}\mathcal{S}\mathcal{C}\mathcal{M}\mathcal{A}_2 + \mathcal{R}_\zeta + \mathcal{R}_v + \mathcal{Y} + \mathcal{Y}^*)]^* \\ \times (I_{N^2M^2} - \mathcal{F})^{-1}\mathrm{vec}(\mathcal{R}_u)\end{aligned}} \tag{5.138}$$

where $\mathcal{R}_u$ is defined in (5.89). Observe that the main difference relative to (5.88) and (5.90) is the addition of the term $\mathcal{R}_\zeta$. Therefore, all the results that were

derived in the earlier section, such as (5.95) and (5.96), continue to hold by adding $\mathcal{R}_\zeta$. In particular, if Assumptions 5.2 and 5.3 are adopted, expressions (5.137) and (5.138) can be approximated as

$$\boxed{\mathrm{MSD}_{\mathrm{track}} \approx \frac{1}{N}\left[\mathrm{vec}\left(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2+\mathcal{R}_\zeta+\mathcal{R}_v\right)\right]^*\left(I_{N^2M^2}-\mathcal{F}\right)^{-1}\mathrm{vec}(I_{NM})} \tag{5.139}$$

and

$$\boxed{\mathrm{EMSE}_{\mathrm{track}} \approx \frac{1}{N}\left[\mathrm{vec}\left(\mathcal{A}_2^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2+\mathcal{R}_\zeta+\mathcal{R}_v\right)\right]^*\left(I_{N^2M^2}-\mathcal{F}\right)^{-1}\mathrm{vec}(\mathcal{R}_u)} \tag{5.140}$$

where $\mathcal{R}_v$ is now given in (5.93).

5.7 Simulation Results

We simulate two scenarios: noisy information exchanges and non-stationary environments. We consider a connected network with $N = 20$ nodes. The network topology is shown in Fig. 5.1.

5.7.1 Imperfect Information Exchange

The unknown complex parameter w^o of length $M = 2$ is randomly generated; its value is $[0.3750 + j2.0834, 0.7174 + j1.4123]$. We adopt uniform step-sizes, $\{\mu_k = 0.01\}$, and uniformly white Gaussian regression data with covariance matrices $\{R_{u,k} = \sigma_{u,k}^2 I_M\}$, where $\{\sigma_{u,k}^2\}$ are shown in Fig. 5.2a. The variances of the model noises, $\{\sigma_{v,k}^2\}$, are randomly generated and shown in Fig. 5.2b. We also use white Gaussian link noise signals such that $R_{v,\ell k}^{(w)} = \sigma_{w,\ell k}^2 I_M$, $R_{v,\ell k}^{(u)} = \sigma_{u,\ell k}^2 I_M$, and $R_{v,\ell k}^{(\psi)} = \sigma_{\psi,\ell k}^2 I_M$. All link noise variances, $\{\sigma_{w,\ell k}^2, \sigma_{v,\ell k}^2, \sigma_{u,\ell k}^2, \sigma_{\psi,\ell k}^2\}$, are randomly generated and illustrated in Fig. 5.3 from top to bottom. We assign the link number by the following procedure. We denote the link from node ℓ to node k

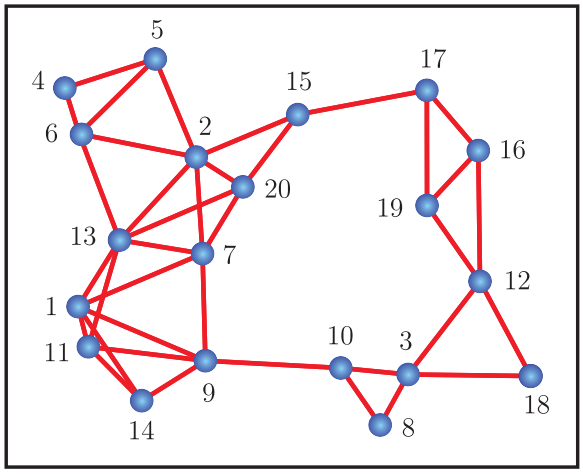


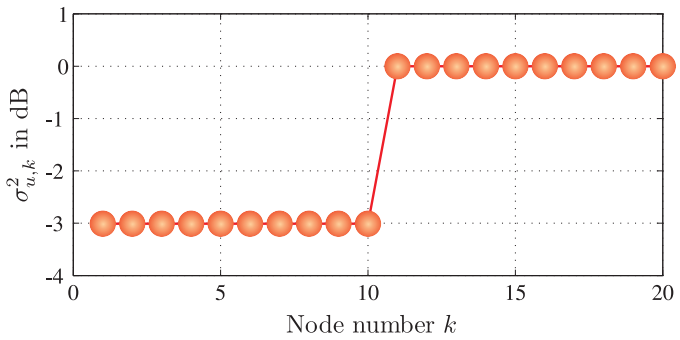
Figure 5.1: A network topology with $N = 20$ nodes.

as $\ell_{l,k}$, where $l \neq k$. Then, we collect the links $\{\ell_{l,k}, \ell \in \mathcal{N}_k \backslash \{k\}\}$ in an ascending order of ℓ in the list $\mathcal{L}_k$ (which is a set with *ordered* elements) for each node k. For example, for node $k = 2$ in Fig. 5.1, it has 6 links; the ordered links are then collected in $\mathcal{L}_2 \triangleq \{\ell_{5,2}, \ell_{6,2}, \ell_{7,2}, \ell_{13,2}, \ell_{15,2}, \ell_{20,2}\}$. We concatenate $\{\mathcal{L}_k\}$ in an ascending order of k to get the overall list $\mathcal{L} \triangleq \{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_N\}$. Eventually, the mth link in the network is given by the mth element in the list $\mathcal{L}$.

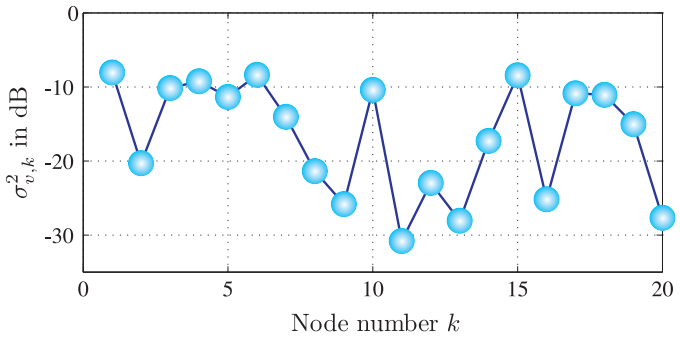
We examine the simplified CTA and ATC algorithms in (5.3) and (5.4), namely, no sharing of data among nodes (i.e., $C = I_N$), under various combination rules: (i) the relative variance rule in (5.124), (ii) the Metropolis rule in [106]:

$$\begin{cases} a_{\ell k} = \dfrac{1}{\max\{|\mathcal{N}_k|, |\mathcal{N}_l|\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ a_{\ell k} = 1 - \displaystyle\sum_{\ell \in \mathcal{N}_k \backslash \{k\}} a_{\ell k}, & \ell = k \\ a_{\ell k} = 0, & \ell \notin \mathcal{N}_k \end{cases} \tag{5.141}$$

where $|\mathcal{N}_k|$ denotes the degree of node k (including the node itself), (iii) the



(a) The variance profile of regression data.



(b) The variance profile of measurement noises.

Figure 5.2: The variance profiles for regression data and measurement noises.

uniform weighting rule:

$$\begin{cases} a_{\ell k} = \dfrac{1}{|\mathcal{N}_k|}, & \ell \in \mathcal{N}_k \\ a_{\ell k} = 0, & \ell \notin \mathcal{N}_k \end{cases} \tag{5.142}$$

and (iv) the adaptive rule in (5.128) with $\{\nu_k = 0.05\}$. We plot the network MSD and EMSE learning curves for ATC algorithms in Figs. 5.4a and 5.4c by averaging over 50 experiments. For CTA algorithms, we plot their network MSD and EMSE learning curves in Figs. 5.4b and 5.4d also by averaging over 50 experiments. Moreover, we also plot their theoretical results (5.95) and (5.96) in the same figures. From Fig. 5.4 we see that the relative variance rule makes diffusion algorithms achieve the lowest MSD and EMSE levels at steady-state, compared to the metropolis and uniform rules as well as the algorithm from [47]

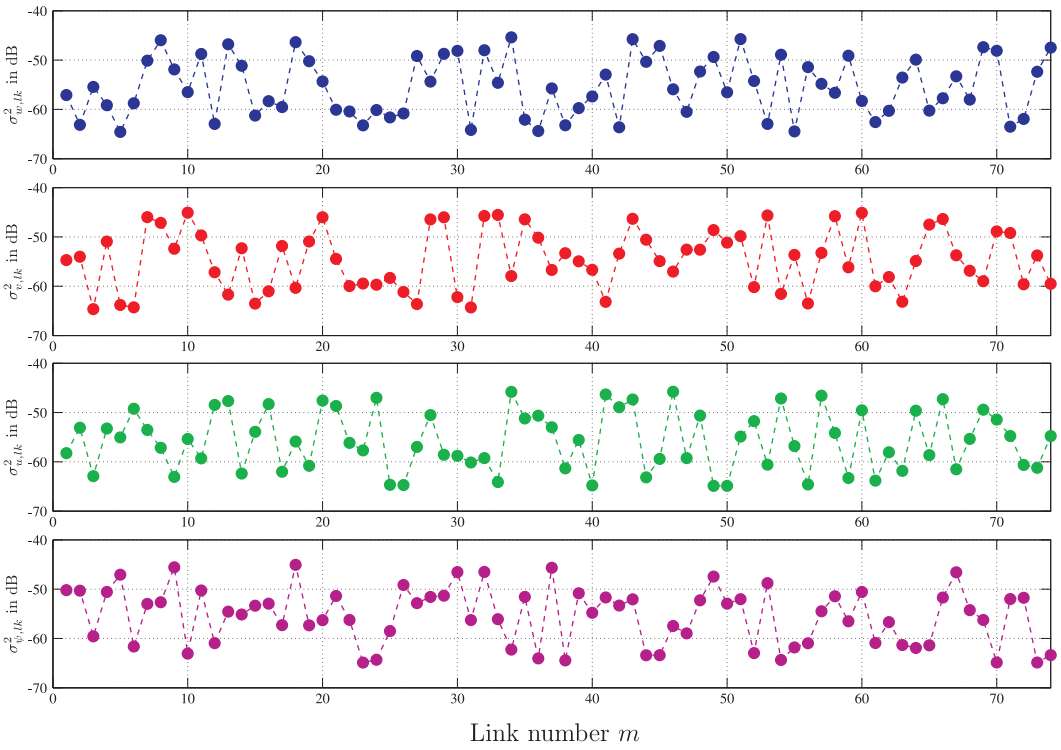
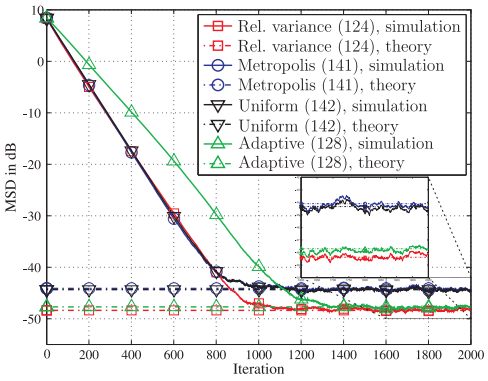


Figure 5.3: The variance profiles for various sources of link noises, including $\{\sigma^2_{w,\ell k}, \sigma^2_{v,\ell k}, \sigma^2_{u,\ell k}, \sigma^2_{\psi,\ell k}\}$.

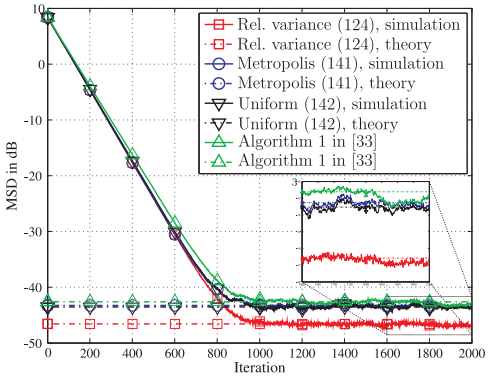
(which also requires knowledge of the noise variances). In addition, the adaptive rule attains MSD and EMSE levels that are only slightly larger than those of the relative variance rule, although, as expected, it converges slower due to the additional learning step (5.127).

5.7.2 Non-stationary Scenario

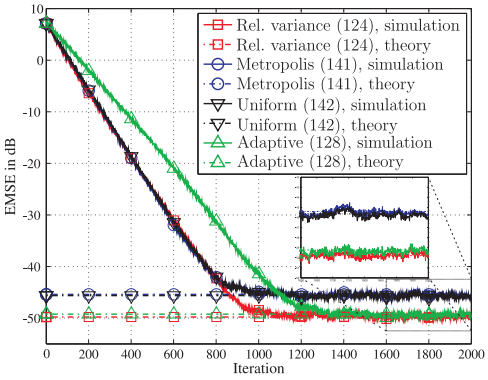
The value for each entry of the complex parameter $w_i^o = \text{col}\{w_{i,1}^o, w_{i,2}^o\}$ is assumed to be changing over time along a circular trajectory in the complex plane, as shown in Fig. 5.5. The dynamic model for w_i^o is expressed as $w_{i,m}^o = e^{j\omega} w_{i-1,m}^o$, where $m = 1, 2$, $\omega = 2\pi/6000$, and $w_{-1}^o = \text{col}\{1 + j, -1 - j\}$. The covariance matrices $\{R_{u,k}\}$ are randomly generated such that $R_{u,k} \neq R_{u,\ell}$ when $k \neq \ell$, but



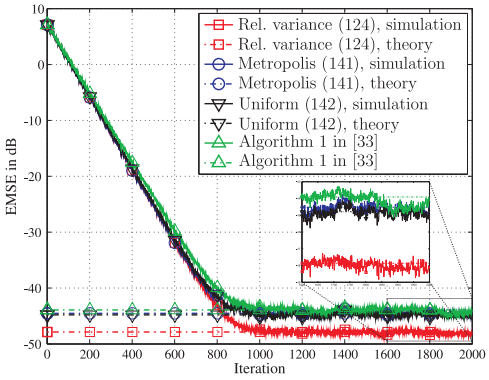
(a) Network MSD curves for ATC algorithms



(b) Network MSD curves for CTA algorithms



(c) Network EMSE curves for ATC algorithms



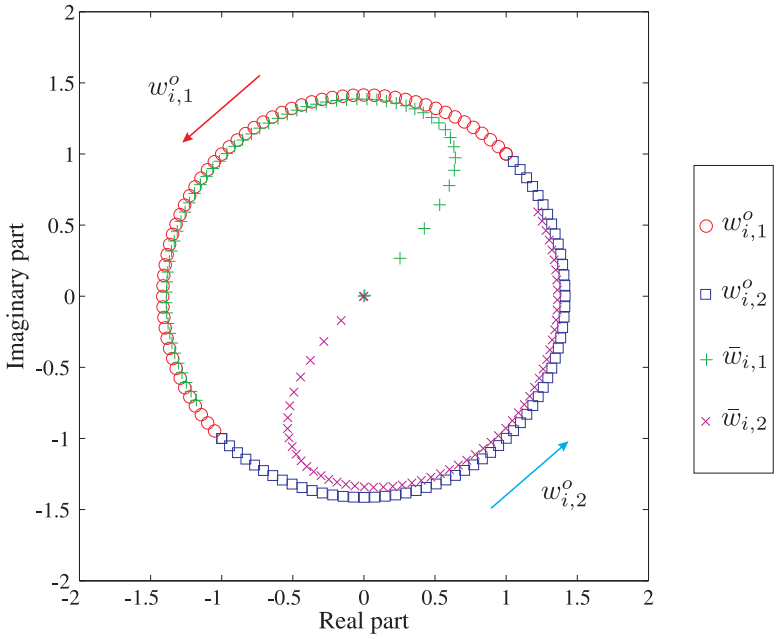
(d) Network EMSE curves for CTA algorithms

Figure 5.4: Simulated network MSD and EMSE curves and theoretical results (5.95) and (5.96) for diffusion algorithms with various combination rules under noisy information exchange.

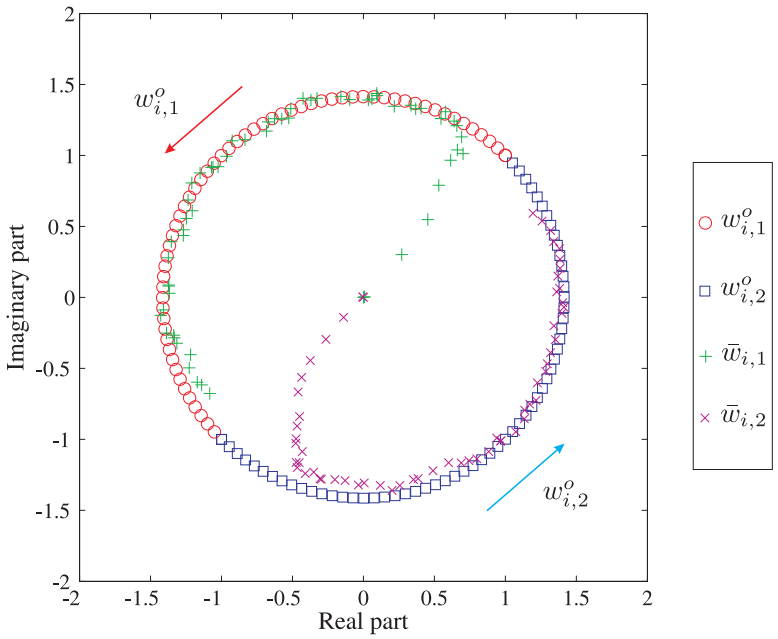
their traces are normalized to be one, i.e., $\mathrm{Tr}(R_{u,k}) = 1$, for all nodes. The variances for the model noises, $\{\sigma_{v,k}^2\}$, are also randomly generated. We examine two different scenarios: the low noise-level case where the average noise variance across the network is -5 dB and the noise variances are shown in Fig. 5.6a; and the high noise-level case where the average variance is 25 dB and the variances are shown in Fig. 5.6b. We simulate 3000 iterations and average over 20 experiments in Figs. 5.5a and 5.5b for each case. The step-size is 0.01 and uniform across the network. For simplicity, we adopt the simplified ATC algorithm where $C = I_N$, and only use the uniform weighting rule (5.142). The tracking behavior of the network, denoted as $\bar{w}_i = \mathrm{col}\{\bar{w}_{i,1}, \bar{w}_{i,2}\}$, is obtained by averaging over all the estimates, $\{w_{k,i}\}$, across the network. Figs. 5.5a and 5.5b depict the complex plane; the horizontal axis is the real axis and the vertical axis is the imaginary axis. Therefore, for every time i, each entry of w_i^o or $\bar{w}_i$ represents a point in the plane. When i is increasing, $w_{i,1}^o$ moves along the red trajectory (in $\circ$), $w_{i,2}^o$ along the blue trajectory (in $\square$), $\bar{w}_{i,1}$ along the green trajectory (in $+$), and $\bar{w}_{i,2}$ along the magenta trajectory (in $\times$). From Fig. 5.5, it can be seen that diffusion algorithms exhibit the tracking ability in both high and low noise-level environments.

5.8 Conclusions

In this chapter we investigated the performance of diffusion algorithms under several sources of noise during information exchange and under non-stationary environments. We first showed that, on one hand, the link noise over the regression data biases the estimators and deteriorates the conditions for mean and mean-square convergence. On the other hand, diffusion strategies can still stabilize the mean and mean-square convergence of the network with noisy information

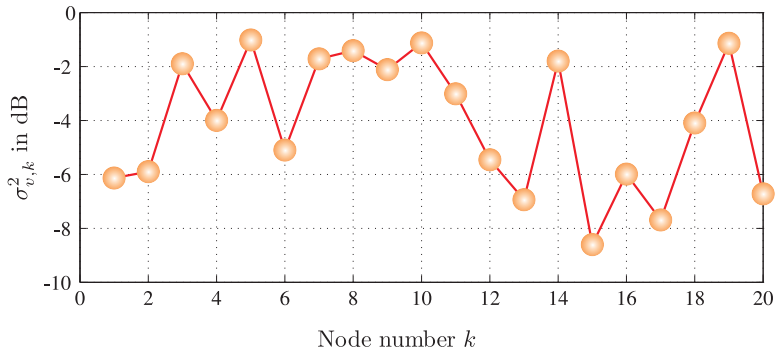


(a) The low noise-level case.

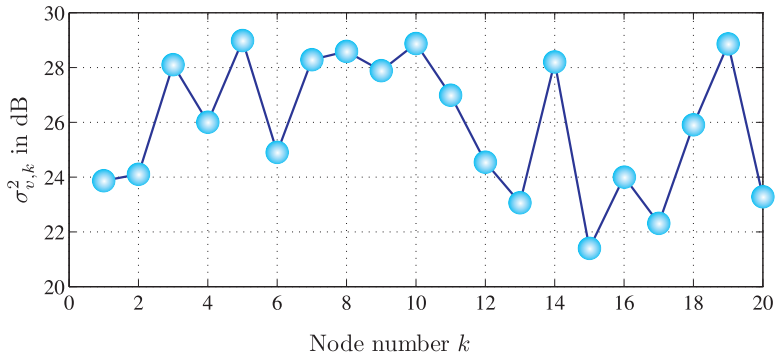


(b) The high noise-level case.

Figure 5.5: An adaptive network tracking a parameter vector $w^o \in \mathbb{C}^2$.



(a) The variance profile for low noise-level.



(b) The variance profile for high noise-level.

Figure 5.6: The noise variance profiles for two cases.

exchange. We derived analytic expressions for the network MSD and EMSE and used these expressions to motivate the choice of combination weights that help ameliorate the effect of information-exchange noise and improve network performance. We also extended the results to the non-stationary scenario where the unknown parameter w^o is changing over time. Simulation results illustrate the theoretical findings and how well they match with theory.

5.A Stability of $\mathcal{A}_2^{\mathsf{T}}\left(I_{NM}-\mathcal{M}\mathcal{R}'\right)\mathcal{A}_1^{\mathsf{T}}$

Following [108], we first define the block maximum norm of a vector.

Definition 5.1 (Block Maximum Norm)**.** *Given a vector* $x = \mathrm{col}\{x_1, \ldots, x_N\} \in \mathbb{C}^{MN}$ *consisting of* N *blocks* $\{x_k \in \mathbb{C}^M, k = 1, \ldots, N\}$*, the block maximum norm is the real function* $\|\cdot\|_{b,\infty} : \mathbb{C}^{MN} \to \mathbb{R}$*, defined as*

$$\|x\|_{b,\infty} \triangleq \max_{1 \le k \le N} \|x_k\|_2 \tag{5.143}$$

where $\|\cdot\|_2$ *denotes the standard 2-norm on* $\mathbb{C}^M$*.* □

Similarly, we define the matrix norm that is induced by the block maximum norm as follows:

Definition 5.2 (Block Maximum Matrix Norm)**.** *Given a block matrix* $\mathcal{A} \in \mathbb{C}^{MN \times MN}$ *with block size* $M \times M$*, then*

$$\|\mathcal{A}\|_{b,\infty} \triangleq \max_{x \in \mathbb{C}^{MN} \setminus \{0\}} \frac{\|\mathcal{A}x\|_{b,\infty}}{\|x\|_{b,\infty}} \tag{5.144}$$

denotes the induced block maximum (matrix) norm on $\mathbb{C}^{MN \times MN}$*.* □

Lemma 5.1. *The block maximum matrix norm is block unitary invariant, i.e., given a block diagonal unitary matrix* $\mathcal{U} \triangleq \mathrm{diag}\{U_1, \ldots, U_N\} \in \mathbb{C}^{MN \times MN}$ *consisting of* N *unitary blocks* $\{U_k \in \mathbb{C}^{M \times M}, k = 1, \ldots, N\}$*, where* $U_k U_k^* = U_k^* U_k = I_M$*, for any matrix* $\mathcal{A} \in \mathbb{C}^{MN \times MN}$*, then*

$$\|\mathcal{A}\|_{b,\infty} = \|\,\mathcal{U}\mathcal{A}\mathcal{U}^*\|_{b,\infty} \tag{5.145}$$

where $\|\cdot\|_{b,\infty}$ *denotes the block maximum matrix norm on* $\mathbb{C}^{MN \times MN}$ *with block size* $M \times M$*.* □

Lemma 5.2. *Let* $A \in \mathbb{C}^{N \times N}$ *be a right-stochastic matrix. Then, for block size* $M \times M$*,*

$$\|A \otimes I_M\|_{b,\infty} = 1 \tag{5.146}$$

Proof. From Definition 5.2, we get

$$\begin{aligned}
\|A \otimes I_M\|_{b,\infty} &= \max_{x \in \mathbb{C}^{MN} \setminus \{0\}} \frac{\max_l \| \sum_{k=1}^N [A]_{\ell k} x_k \|_2}{\max_k \|x_k\|_2} \\
&\leq \max_{x \in \mathbb{C}^{MN} \setminus \{0\}} \frac{\max_l \sum_{k=1}^N [A]_{\ell k} \|x_k\|_2}{\max_k \|x_k\|_2} \\
&\leq \max_{x \in \mathbb{C}^{MN} \setminus \{0\}} \frac{\max_l (\sum_{k=1}^N [A]_{\ell k}) \cdot \max_k \|x_k\|_2}{\max_k \|x_k\|_2} \\
&\leq \max_{x \in \mathbb{C}^{MN} \setminus \{0\}} \frac{\max_l 1 \cdot \max_k \|x_k\|_2}{\max_k \|x_k\|_2} \\
&= 1
\end{aligned} \tag{5.147}$$

where $x \triangleq \mathrm{col}\{x_1, \ldots, x_N\} \in \mathbb{C}^{MN}$ consists of N blocks $\{x_k \in \mathbb{C}^M, k = 1, \ldots, N\}$, and $[A]_{\ell k}$ denotes the (ℓ, k)-th entry of A. On the other hand, for any induced matrix norm, say, the block maximum norm, it is always lower bounded by the spectral radius of the matrix [115]:

$$\|A \otimes I_M\|_{b,\infty} \geq \rho(A \otimes I_M) = \rho(A) = 1 \tag{5.148}$$

Combining (5.147) and (5.148) completes the proof. □

Lemma 5.3. *Let $\mathcal{A} \in \mathbb{C}^{NM \times NM}$ be a block diagonal Hermitian matrix with block size $M \times M$. Then the block maximum norm of the matrix $\mathcal{A}$ is equal to its spectral radius, i.e.,*

$$\|\mathcal{A}\|_{b,\infty} = \rho(\mathcal{A}) \tag{5.149}$$

Proof. Denote the kth $M \times M$ submatrix on the diagonal of $\mathcal{A}$ by A_k. Let $A_k = U_k \Lambda_k U_k^*$ be the eigen - decomposition of A_k, where $U_k \in \mathbb{C}^{M \times M}$ is unitary and $\Lambda_k \in \mathbb{R}^{M \times M}$ is diagonal. Define the block unitary matrix $\mathcal{U} \triangleq \mathrm{diag}\{U_1, \ldots, U_N\}$ and the diagonal matrix $\Lambda \triangleq \mathrm{diag}\{\Lambda_1, \ldots, \Lambda_N\}$. Then, $\mathcal{A} = \mathcal{U} \Lambda \mathcal{U}^*$. By Lemma 5.1, the block maximum norm of $\mathcal{A}$ with block size $M \times M$ is

$$\|\mathcal{A}\|_{b,\infty} = \|\mathcal{U} \Lambda \mathcal{U}^*\|_{b,\infty}$$

$$
\begin{aligned}
&= \|\Lambda\|_{b,\infty} \\
&= \max_{x \in \mathbb{C}^{MN} \setminus \{0\}} \frac{\max_k \|\Lambda_k x_k\|_2}{\max_k \|x_k\|_2} \\
&\leq \max_{x \in \mathbb{C}^{MN} \setminus \{0\}} \frac{\max_k \|\Lambda_k\|_2 \cdot \|x_k\|_2}{\max_k \|x_k\|_2} \\
&\leq \max_{x \in \mathbb{C}^{MN} \setminus \{0\}} \frac{\max_k \|\Lambda_k\|_2 \cdot \max_k \|x_k\|_2}{\max_k \|x_k\|_2} \\
&= \max_k \|\Lambda_k\|_2 \\
&= \rho(\mathcal{A}) \qquad (5.150)
\end{aligned}
$$

where we used the fact that the induced 2-norm is identical to the spectral radius for Hermitian matrices [115]. On the other hand, any matrix norm is lower bounded by the spectral radius [115], i.e.,

$$\rho(\mathcal{A}) \leq \|\mathcal{A}\|_{b,\infty} \qquad (5.151)$$

Combining (5.150) and (5.151) completes the proof. □

Now we show that the matrix $\mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\mathcal{R}')\mathcal{A}_1^{\mathsf{T}}$ is stable if $I_{NM} - \mathcal{M}\mathcal{R}'$ is stable. For any induced matrix norm, say, the block maximum norm with block size $M \times M$, we have [115]

$$
\begin{aligned}
\rho\left(\mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\mathcal{R}')\mathcal{A}_1^{\mathsf{T}}\right) &\leq \|\mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\mathcal{R}')\mathcal{A}_1^{\mathsf{T}}\|_{b,\infty} \\
&\leq \|\mathcal{A}_2^{\mathsf{T}}\|_{b,\infty} \cdot \|I_{NM} - \mathcal{M}\mathcal{R}'\|_{b,\infty} \cdot \|\mathcal{A}_1^{\mathsf{T}}\|_{b,\infty} \\
&= \|I_{NM} - \mathcal{M}\mathcal{R}'\|_{b,\infty} \qquad (5.152)
\end{aligned}
$$

where, from (5.8) and (5.22), $\mathcal{A}_1^{\mathsf{T}}$ and $\mathcal{A}_2^{\mathsf{T}}$ satisfy Lemma 5.2. By (5.17) and (5.55), it is straightforward to see that $I_{NM} - \mathcal{M}\mathcal{R}'$ is block diagonal with block size $M \times M$. Then, by Lemma 5.3, expression (5.152) can be further expressed as

$$\rho\left(\mathcal{A}_2^{\mathsf{T}}(I_{NM} - \mathcal{M}\mathcal{R}')\mathcal{A}_1^{\mathsf{T}}\right) \leq \rho(I_{NM} - \mathcal{M}\mathcal{R}') \qquad (5.153)$$

which completes the proof.[1]

5.B Proof of expression (5.75)

Let us denote the (ℓ,k)th submatrix of R_z by $R_{z,\ell k} \in \mathbb{C}^{M\times M}$. By Assumptions 5.1 and expression (5.44), $R_{z,\ell k}$ can be evaluated as

$$R_{z,\ell k} = \mathbb{E}\,\boldsymbol{z}_{l,i}\boldsymbol{z}_{k,i}^* = \sum_{m\in\mathcal{N}_\ell}\sum_{n\in\mathcal{N}_k} c_{m\ell}c_{nk}\underbrace{\mathbb{E}\left(\boldsymbol{u}_{m\ell,i}^*\boldsymbol{v}_{m\ell}(i)\boldsymbol{v}_{nk}^*(i)\boldsymbol{u}_{nk,i}\right)}_{\triangleq R_{m\ell,nk}} \tag{5.154}$$

where, by expressions (5.26) and (5.38),

$$\begin{aligned} R_{m\ell,nk} = \mathbb{E}\left(\boldsymbol{u}_{m,i}+\boldsymbol{v}_{m\ell,i}^{(u)}\right)^*\left(\boldsymbol{v}_m(i)+\boldsymbol{v}_{m\ell}^{(d)}(i)-\boldsymbol{v}_{m\ell,i}^{(u)}w^o\right) \\ \times\left(\boldsymbol{v}_n(i)+\boldsymbol{v}_{nk}^{(d)}(i)-\boldsymbol{v}_{nk,i}^{(u)}w^o\right)^*\left(\boldsymbol{u}_{n,i}+\boldsymbol{v}_{nk,i}^{(u)}\right) \end{aligned} \tag{5.155}$$

When $m\neq n$, expression (5.155) reduces to

$$R_{m\ell,nk} = R_{v,m\ell}^{(u)}w^o(w^o)^*R_{v,nk}^{(u)} \tag{5.156}$$

When $m=n$, expression (5.155) becomes

$$\begin{aligned} R_{m\ell,nk} &= \mathbb{E}\left(\boldsymbol{v}_m(i)+\boldsymbol{v}_{m\ell}^{(d)}(i)-\boldsymbol{v}_{m\ell,i}^{(u)}w^o\right)\left(\boldsymbol{v}_m(i)+\boldsymbol{v}_{mk}^{(d)}(i)-\boldsymbol{v}_{mk,i}^{(u)}w^o\right)^* \\ &\quad\times\left(\boldsymbol{u}_{m,i}+\boldsymbol{v}_{m\ell,i}^{(u)}\right)^*\left(\boldsymbol{u}_{m,i}+\boldsymbol{v}_{mk,i}^{(u)}\right) \\ &= \mathbb{E}\left(\boldsymbol{v}_m(i)+\boldsymbol{v}_{m\ell}^{(d)}(i)\right)\left(\boldsymbol{v}_m(i)+\boldsymbol{v}_{mk}^{(d)}(i)\right)^*\mathbb{E}\left(\boldsymbol{u}_{m,i}+\boldsymbol{v}_{m\ell,i}^{(u)}\right)^*\left(\boldsymbol{u}_{m,i}+\boldsymbol{v}_{mk,i}^{(u)}\right) \\ &\quad+\mathbb{E}\,\boldsymbol{v}_{m\ell,i}^{(u)}w^o(w^o)^*\boldsymbol{v}_{mk,i}^{(u)*}\left(\boldsymbol{u}_{m,i}+\boldsymbol{v}_{m\ell,i}^{(u)}\right)^*\left(\boldsymbol{u}_{m,i}+\boldsymbol{v}_{mk,i}^{(u)}\right) \\ &= \left(\sigma_{v,m}^2+\delta_{\ell k}\sigma_{v,m\ell}^2\right)\left(R_{u,m}+\delta_{\ell k}R_{v,m\ell}^{(u)}\right) \\ &\quad+\delta_{\ell k}(w^o)^*R_{v,m\ell}^{(u)}w^oR_{u,m}+R_{v,m\ell}^{(u)}w^o(w^o)^*R_{v,mk}^{(u)} \end{aligned}$$

[1] This statement fixes the argument about block diagonal matrices $\mathcal{X}$ and $\mathcal{M}$ in Appendix I of [6] and Lemma 2 of [118], where the $\|\cdot\|_\rho$ norm used in these references should be replaced by the $\|\cdot\|_{b,\infty}$ norm used here.

$$+\delta_{\ell k}\left(\mathbb{E}\,\boldsymbol{v}_{m\ell,i}^{(u)*}\boldsymbol{v}_{m\ell,i}^{(u)}w^o(w^o)^*\boldsymbol{v}_{m\ell,i}^{(u)*}\boldsymbol{v}_{m\ell,i}^{(u)}-R_{v,m\ell}^{(u)}w^o(w^o)^*R_{v,m\ell}^{(u)}\right) \tag{5.157}$$

where $\delta_{\ell k}$ denotes the Kronecker delta function. Evaluating the last term on RHS of (5.157) requires knowledge of the excess kurtosis of $\boldsymbol{v}_{m\ell,i}^{(u)}$, which is generally not available. In order to proceed, we invoke a separation principle to approximate it as

$$\mathbb{E}\,\boldsymbol{v}_{m\ell,i}^{(u)*}\boldsymbol{v}_{m\ell,i}^{(u)}w^o(w^o)^*\boldsymbol{v}_{m\ell,i}^{(u)*}\boldsymbol{v}_{m\ell,i}^{(u)}\approx R_{v,m\ell}^{(u)}w^o(w^o)^*R_{v,m\ell}^{(u)} \tag{5.158}$$

Substituting (5.158) into (5.157) leads to

$$\begin{aligned}R_{m\ell,nk}&\approx\left(\sigma_{v,m}^2+\delta_{\ell k}\sigma_{v,m\ell}^2\right)\left(R_{u,m}+\delta_{\ell k}R_{v,m\ell}^{(u)}\right)\\&\quad+\delta_{\ell k}\left((w^o)^*R_{v,m\ell}^{(u)}w^o\right)R_{u,m}+R_{v,m\ell}^{(u)}w^o(w^o)^*R_{v,mk}^{(u)}\\&=\sigma_{v,m}^2R_{u,m}+R_{v,m\ell}^{(u)}w^o(w^o)^*R_{v,mk}^{(u)}\\&\quad+\delta_{\ell k}\left[(\sigma_{v,m\ell}^2+(w^o)^*R_{v,m\ell}^{(u)}w^o)R_{u,m}+(\sigma_{v,m}^2+\sigma_{v,m\ell}^2)R_{v,m\ell}^{(u)}\right]\end{aligned} \tag{5.159}$$

From (5.156) and (5.159), we get

$$\begin{aligned}R_{m\ell,nk}&\approx R_{v,m\ell}^{(u)}w^o(w^o)^*R_{v,mk}^{(u)}+\delta_{mn}\sigma_{v,m}^2R_{u,m}\\&\quad+\delta_{mn}\delta_{\ell k}\left[(\sigma_{v,m\ell}^2+(w^o)^*R_{v,m\ell}^{(u)}w^o)R_{u,m}+(\sigma_{v,m}^2+\sigma_{v,m\ell}^2)R_{v,m\ell}^{(u)}\right]\end{aligned} \tag{5.160}$$

Substituting (5.160) into (5.154), we obtain

$$\begin{aligned}R_{z,\ell k}&\approx\left(\sum_{m\in\mathcal{N}_\ell}c_{m\ell}R_{v,m\ell}^{(u)}w^o\right)\left(\sum_{n\in\mathcal{N}_k}c_{nk}R_{v,mk}^{(u)}w^o\right)^*+\sum_{m\in\mathcal{N}_\ell}\sum_{n\in\mathcal{N}_k}c_{m\ell}c_{nk}\delta_{mn}\sigma_{v,m}^2R_{u,m}\\&\quad+\delta_{\ell k}\sum_{m\in\mathcal{N}_\ell}c_{m\ell}^2\left[(\sigma_{v,m\ell}^2+(w^o)^*R_{v,m\ell}^{(u)}w^o)R_{u,m}+(\sigma_{v,m}^2+\sigma_{v,m\ell}^2)R_{v,m\ell}^{(u)}\right]\end{aligned} \tag{5.161}$$

From (5.58)–(5.59) and (5.76)–(5.78), we arrive at expression (5.75).

CHAPTER 6

Distributed Clustering and Learning Over Networks

In this chapter, we consider the situation where agents belong to different groups that pursue different objectives. Distributed processing over networks relies on in-network processing and cooperation among neighboring agents. Cooperation is beneficial when agents share a common objective. However, in many applications agents may belong to different clusters that pursue different objectives. Then, indiscriminate cooperation will lead to undesired results. In this chapter, we propose an adaptive clustering and learning scheme that allows agents to learn which neighbors they should cooperate with and which other neighbors they should ignore. In doing so, the resulting algorithm enables the agents to identify their clusters and to attain improved learning and estimation accuracy over networks. We carry out a detailed mean-square analysis and assess the error probabilities of Types I and II, i.e., false alarm and mis-detection, for the clustering mechanism. Among other results, we establish that these probabilities decay exponentially with the step-sizes so that the probability of correct clustering can be made arbitrarily close to one. The results in this chapter are based on material from [119].

6.1 Models and Assumptions

The clustered distributed learning problem was formulated in (1.36) in Chapter 1. The concepts of clusters and groups were also introduced in Definitions 1.1 and 1.2 in the same Chapter. To facilitate the performance analysis, we summarize the main conditions on the network topology in the following statement.

Assumption 6.1 (Topology, clusters, and groups)**.**

1. *The network consists of Q clusters, $\{\mathcal{C}_q; q = 1, 2, \ldots, Q\}$. The size of cluster $\mathcal{C}_q$ is denoted by N_q^c such that $|\mathcal{C}_q| = N_q^c$ and $\sum_{q=1}^{Q} N_q^c = N$.*

2. *The underlying topology for each cluster $\mathcal{C}_q$ is connected. Clusters are also inter-connected by some links so that agents from different clusters may still be neighbors of each other.*

3. *There is a total of G groups, denoted by $\{\mathcal{G}_m; m = 1, 2, \ldots, G\}$, in the network. The size of group $\mathcal{G}_m$ is denoted by N_m^g such that $|\mathcal{G}_m| = N_m^g$ and $\sum_{m=1}^{G} N_m^g = N$.* □

It is obvious that $Q \leq G \leq N$ because each cluster has at least one group and each group has at least one agent.

Definition 6.1 (Indexing rule)**.** *Without loss of generality, we index groups according to their cluster indexes such that groups from the same cluster will have consecutive indexes. Likewise, we index agents according to their group indexes such that agents from the same group will have consecutive indexes.* □

According to this indexing rule, if group $\mathcal{G}_m$ belongs to cluster $\mathcal{C}_q$, then the next group $\mathcal{G}_{m+1}$ will belong either to cluster $\mathcal{C}_q$ or the next cluster, $\mathcal{C}_{q+1}$; if agent

k belongs to group $\mathcal{G}_m$, then the next agent $k+1$ will belong either to group $\mathcal{G}_m$ or the next group, $\mathcal{G}_{m+1}$.

Based on the problem formulation in Section 1.4 from Chapter 1, although agents in the same cluster are connected, they are generally not aware of each other's cluster information, and therefore some agents in the same cluster may not cooperate in the initial stage of adaptation. On the other hand, agents in the same group are aware of each other's cluster information, so these agents can cooperate. As the learning process proceeds, agents from different groups in the same cluster will recognize each other through information sharing. Once cluster information is inferred, small groups will merge into larger groups, and agents will start cooperating with more neighbors. Through this adaptive clustering procedure, cooperative learning will grow until all agents within the same cluster become cooperative and the network performance is enhanced.

To proceed with the modeling assumptions, we introduce the following network Hessian matrix function:

$$\nabla^2 J(\mathcal{W}) \triangleq \operatorname{diag}\{\nabla^2 J_1(w_1), \ldots, \nabla^2 J_N(w_N)\} \tag{6.1}$$

where the vector $\mathcal{W}$ collects the parameters from across the network:

$$\mathcal{W} \triangleq \operatorname{col}\{w_1, \ldots, w_N\} \in \mathbb{R}^{NM\times 1} \tag{6.2}$$

We also collect the individual minimizers into a vector:

$$\mathcal{W}^o \triangleq \operatorname{col}\{w_1^o, \ldots, w_N^o\} = \operatorname{col}\{\mathbb{1}_{N_q^c} \otimes w_q^\star; q = 1, \ldots, Q\} \tag{6.3}$$

where the second equality is due to the indexing rule in Definition 6.1. We next list two standard assumptions for stochastic distributed learning over adaptive networks to guide the subsequent analysis in this work. One assumption relates to the analytical properties of the cost functions, and is meant to ensure well-defined

minima and well-posed problems. The second assumption relates to stochastic properties of the gradient noise processes that result from approximating the true gradient vectors. This assumption is meant to ensure that the gradient approximations are unbiased and with moments satisfying some regularity conditions. Explanations and motivation for these assumptions in the context of inference problems can be found in [8, 66, 72].

Assumption 6.2 (Cost functions)**.**

1. *Each individual cost $J_k(w)$ is assumed to be strictly convex, twice differentiable, and with bounded Hessian matrix function satisfying:*

$$\lambda_{k,L} I_M \leq \nabla^2 J_k(w) \leq \lambda_{k,U} I_M \tag{6.4}$$

   where $0 \leq \lambda_{k,L} \leq \lambda_{k,U} < \infty$.

2. *In each group $\mathcal{G}_m$, at least one individual cost, say, $J_{k^o}(w)$, is strongly-convex, meaning that the lower bound, $\lambda_{k^o,L}$, on the Hessian of this cost is positive.*

3. *The network Hessian function $\nabla^2 J(\mathcal{W})$ in (6.1) satisfies the Lipschitz condition:*

$$\|\nabla^2 J(\mathcal{W}_1) - \nabla^2 J(\mathcal{W}_2)\| \leq \kappa_H \|\mathcal{W}_1 - \mathcal{W}_2\| \tag{6.5}$$

   for any $\mathcal{W}_1, \mathcal{W}_2 \in \mathbb{R}^{NM \times 1}$ and some $\kappa_H \geq 0$. □

The second set of assumptions relate to conditions on the gradient noise processes. For this purpose, we introduce the filtration $\{\mathbb{F}_i; i \geq 0\}$ to represent the information flow that is available up to the i-th iteration of the learning process. The true network gradient function and its stochastic approximation are respectively denoted by

$$\nabla J(\mathcal{W}) \triangleq \text{col}\{\nabla J_1(w_1), \ldots, \nabla J_N(w_N)\} \tag{6.6}$$

$$\widehat{\nabla J}(w) \triangleq \mathrm{col}\{\widehat{\nabla J_1}(w_1), \ldots, \widehat{\nabla J_N}(w_N)\} \tag{6.7}$$

The gradient noise at iteration i and agent k is denoted by:

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) \triangleq \widehat{\nabla J_k}(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \tag{6.8}$$

where $\boldsymbol{w}_{k,i-1}$ denotes the estimate for w_k^o that is available to agent k at iteration $i-1$. The network gradient noise is denoted by $\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})$ and is the random process that is obtained by aggregating all noise processes from across the network into a vector:

$$\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}) \triangleq \mathrm{col}\{\boldsymbol{s}_{1,i}(\boldsymbol{w}_{1,i-1}), \ldots, \boldsymbol{s}_{N,i}(\boldsymbol{w}_{N,i-1})\} \tag{6.9}$$

Using (6.8), we can write

$$\widehat{\nabla J}(\boldsymbol{\mathcal{W}}_{i-1}) = \nabla J(\boldsymbol{\mathcal{W}}_{i-1}) + \boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}) \tag{6.10}$$

We denote the conditional covariance of $\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})$ by

$$\mathcal{R}_{s,i}(\boldsymbol{\mathcal{W}}_{i-1}) \triangleq \mathbb{E}[\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})\boldsymbol{s}_i^{\mathsf{T}}(\boldsymbol{\mathcal{W}}_{i-1})|\mathbb{F}_{i-1}] \tag{6.11}$$

where $\boldsymbol{\mathcal{W}}_{i-1}$ is in $\mathbb{F}_{i-1}$.

Assumption 6.3 (Gradient noise)**.** *It is assumed that the gradient noise process satisfies the following properties for any $\boldsymbol{\mathcal{W}}_{i-1}$ in $\mathbb{F}_{i-1}$:*

1. *Martingale difference [66, 120]:*

$$\mathbb{E}[\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})|\mathbb{F}_{i-1}] = 0 \tag{6.12}$$

2. *Bounded fourth-order moment [66–68]:*

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})\|^4|\mathbb{F}_{i-1}] \leq \alpha^2\|w^o - \boldsymbol{\mathcal{W}}_{i-1}\|^4 + \sigma_s^4 \tag{6.13}$$

for some $\alpha, \sigma_s \geq 0$, and where w^o is from (6.3)*.*

3. Lipschitz conditional covariance function [66–68]:

$$\|\mathcal{R}_{s,i}(w^o) - \mathcal{R}_{s,i}(\boldsymbol{w}_{i-1})\| \leq \kappa_s \|w^o - \boldsymbol{w}_{i-1}\|^{\gamma_s} \tag{6.14}$$

for some $\kappa_s \geq 0$ *and* $0 < \gamma_s \leq 4$.

4. Convergent conditional covariance matrix [66–68, 120]:

$$\mathcal{R}_s \triangleq \lim_{i\to\infty} \mathcal{R}_{s,i}(w^o) > 0 \tag{6.15}$$

where $\mathcal{R}_s$ *is symmetric and positive definite.* □

It is easy to verify from (6.13) that the second-order moment of the gradient noise process also satisfies:

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 | \mathbb{F}_{i-1}] \leq \alpha \|w^o - \boldsymbol{w}_{i-1}\|^2 + \sigma_s^2 \tag{6.16}$$

6.2 Proposed Algorithm and Main Results

In order to minimize all cluster cost functions $\{J_q^{\text{cluster}}(w); q = 1, 2, \ldots, Q\}$ defined by (1.37) in Chapter 1, agents need to cooperate only within their clusters. Although cluster information is in general not available beforehand, groups within each cluster are available according to Assumption 6.1. Therefore, based on this prior information, agents can instead focus on solving the following problem based on partitioning by groups rather than by clusters:

$$\underset{\{w_m\}_{m=1}^G}{\text{minimize}} \quad J'(w_1, \ldots, w_G) \triangleq \sum_{m=1}^{G} \sum_{k\in\mathcal{G}_m} J_k(w_m) \tag{6.17}$$

with one parameter vector w_m for each group $\mathcal{G}_m$. In the extreme case when prior clustering information is totally absent, groups will collapse into singletons and problem (6.17) will reduce to the individual non-cooperative case with each agent

running its own stochastic-gradient algorithm to minimize its cost function. In another extreme case when cluster information is completely available, groups will be equivalent to clusters and problem (6.17) will reduce to the formation in (1.36) from Chapter 1. Therefore, problem (6.17) is general and includes many scenarios of interest as special cases. We shall argue in the sequel that during the process of solving (6.17), agents will be able to gradually learn their neighbors' clustering information. This information will be exploited by a *separate* learning procedure by each group to dynamically involve more neighbors (from outside the group) in local cooperation. In this way, we will be able to establish analytically that, with high probability, agents will be able to successfully solve problem (1.36) from Chapter 1 (and not just (6.17)) even *without* having the complete clustering information in advance.

We motivate the algorithm by examining problem (6.17). Since the groups $\{\mathcal{G}_m\}$ are already formed and they are disjoint, problem (6.17) can be decomposed into G separate optimization problems, one for each group:

$$\operatorname*{minimize}_{w} \quad J_m^g(w) \triangleq \sum_{k\in\mathcal{G}_m} J_k(w) \tag{6.18}$$

with $m = 1, 2, \ldots, G$. For any agent k belonging to group $\mathcal{G}_m$ in cluster $\mathcal{C}_q$, i.e., $k \in \mathcal{G}_m \subseteq \mathcal{C}_q$, it is easy to verify that

$$\{k\} \subseteq \mathcal{N}_k \cap \mathcal{G}_m \subseteq \mathcal{N}_k \cap \mathcal{C}_q = \mathcal{N}_k^+ \tag{6.19}$$

Then, agents in group $\mathcal{G}_m$ can seek the solution of $J_m^g(w)$ in (6.18) by using the adapt-then-combine (ATC) diffusion learning strategy over $\mathcal{G}_m$, namely,

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k \widehat{\nabla J_k}(\boldsymbol{w}_{k,i-1}) \tag{6.20a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_k\cap\mathcal{G}_m} a_{\ell k}\boldsymbol{\psi}_{\ell,i} \tag{6.20b}$$

for all $k \in \mathcal{G}_m$, where $\mu_k > 0$ denotes the step-size parameter, and $\{a_{\ell k}\}$ are convex combination coefficients that satisfy

$$\begin{cases} a_{\ell k} > 0 \ \text{ if } \ \ell \in \mathcal{N}_k \cap \mathcal{G}_m \\ a_{\ell k} = 0 \ \text{ otherwise} \end{cases}, \ \text{ and } \ \sum_{\ell=1}^{N} a_{\ell k} = 1 \tag{6.21}$$

Moreover, $\boldsymbol{w}_{k,i}$ denotes the random estimate computed by agent k at iteration i, and $\boldsymbol{\psi}_{k,i}$ is the intermediate iterate. We collect the coefficients $\{a_{\ell k}\}$ into a matrix $A \triangleq [a_{\ell k}]_{\ell,k=1}^{N}$. Obviously, A is a left-stochastic matrix, namely,

$$A^{\mathsf{T}} \mathbb{1}_N = \mathbb{1}_N \tag{6.22}$$

We collect the iterates generated from (6.20a)–(6.20b) by group $\mathcal{G}_m$ into a vector:

$$\boldsymbol{\mathcal{W}}_{m,i} \triangleq \operatorname{col}\{\boldsymbol{w}_{k,i}; k \in \mathcal{G}_m\} \in \mathbb{R}^{N_m^g M \times 1} \tag{6.23}$$

where N_m^g is the size of $\mathcal{G}_m$. According to the indexing rule from Definition 6.1 for agents and groups, the estimate for the entire network from (6.20a)–(6.20b) can be obtained by stacking the group estimates $\{\boldsymbol{\mathcal{W}}_{m,i}\}$:

$$\boldsymbol{\mathcal{W}}_i \triangleq \operatorname{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\} = \operatorname{col}\{\boldsymbol{\mathcal{W}}_{1,i}, \ldots, \boldsymbol{\mathcal{W}}_{G,i}\} \tag{6.24}$$

The procedure used by the agents to enlarge their groups will be based on the following results to be established in later sections. We will show in Theorem 6.3 that after sufficient iterations, i.e., as $i \to \infty$, and for small enough step-sizes, i.e., $\mu_k \ll 1$ for all k, the network estimate $\boldsymbol{\mathcal{W}}_i$ defined by (6.24) exhibits a distribution that is *nearly* Gaussian:

$$\boldsymbol{\mathcal{W}}_i \sim \mathbb{N}(\mathcal{W}^o, \ \mu_{\max}\Pi) \tag{6.25}$$

where $\mathbb{N}(\phi, \Psi)$ denotes a Gaussian distribution with mean ϕ and covariance Ψ, $\mathcal{W}^o$ is from (6.3),

$$\mu_{\max} \triangleq \max_{k=1,\ldots,N} \mu_k \tag{6.26}$$

and $\Pi \in \mathbb{R}^{NM \times NM}$ is a symmetric, positive semi-definite matrix, independent of $\mu_{\max}$, and defined later by (6.115). In addition, we will show that for any pair of agents from two different groups, for example, $k \in \mathcal{G}_m$ and $\ell \in \mathcal{G}_n$, where the two groups $\mathcal{G}_m$ and $\mathcal{G}_n$ may or may not originate from the same cluster, the difference between their estimates will also be distributed approximately according to a Gaussian distribution:

$$\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i} \sim \mathbb{N}(w_\ell^o - w_k^o,\ \mu_{\max}\Delta_{\ell,k}) \tag{6.27}$$

where

$$\Delta_{\ell,k} \triangleq \Pi_{\ell,\ell} + \Pi_{k,k} - \Pi_{k,\ell} - \Pi_{\ell,k} \tag{6.28}$$

is a symmetric, positive semi-definite matrix, and $\Pi_{k,\ell}$ denotes the (k,ℓ)-th block of Π with block size $M \times M$. These results are useful for inferring the cluster information for agents k and ℓ. Indeed, since the covariance matrix in (6.27) is on the order of $\mu_{\max}$, the probability density function (pdf) of $\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}$ will concentrate around its mean, namely, $w_\ell^o - w_k^o$, when $\mu_{\max}$ is sufficiently small. Therefore, if these agents belong to the same cluster such that $w_\ell^o = w_k^o$, then we will be able to conclude from (6.27) that with high probability, $\|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 = O(\mu_{\max})$. On the other hand, if the agents belong to different clusters such that $w_\ell^o \neq w_k^o$, then it will hold with high probability that $\|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 = O(\mu_{\max}^0)$. This observation suggests that a hypothesis test can be formulated for agents ℓ and k to determine whether or not they are members of the same cluster:

$$\|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 \underset{\mathbb{H}_1}{\overset{\mathbb{H}_0}{\lessgtr}} \theta_{k,\ell} \tag{6.29}$$

where $\mathbb{H}_0$ denotes the hypothesis $w_\ell^o = w_k^o$, $\mathbb{H}_1$ denotes the hypothesis $w_\ell^o \neq w_k^o$, and $\theta_{k,\ell} > 0$ is a predefined threshold. Both agents ℓ and k will test (6.29) to reach a symmetric pattern of cooperation. Since $\boldsymbol{w}_{k,i}$ and $\boldsymbol{w}_{\ell,i}$ are accessible through local interactions within neighborhoods, the hypothesis test (6.29) can be carried

out in a distributed manner. We will further show that the probabilities for both types of errors incurred by (6.29), i.e., the false alarm (Type-I) and the missing detection (Type-II) errors, decay at exponential rates, namely,

$$\text{Type-I: } \mathbb{P}[\|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 > \theta_{k,\ell} | w_\ell^o = w_k^o] \leq O(e^{-c_1/\mu_{\max}})$$

$$\text{Type-II: } \mathbb{P}[\|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 < \theta_{k,\ell} | w_\ell^o \neq w_k^o] \leq O(e^{-c_2/\mu_{\max}})$$

for some constants $c_1 > 0$ and $c_2 > 0$. Therefore, for long enough iterations and small enough step-sizes, agents are able to successfully infer the cluster information with very high probability.

The clustering information acquired at each iteration i is used by the agents to dynamically adjust their *inferred* cluster neighborhoods. The $\boldsymbol{\mathcal{N}}_{k,i}^+$ for agent $k \in \mathcal{G}_m$ at iteration i consists of the neighbors that are accepted under hypothesis $\mathbb{H}_0$ and the other neighbors that are already in the same group:

$$\boldsymbol{\mathcal{N}}_{k,i}^+ \triangleq \{\ell \in \mathcal{N}_k; \|\boldsymbol{w}_{\ell,i} - \boldsymbol{w}_{k,i}\|^2 < \theta_{k,\ell} \text{ or } \ell \in \mathcal{G}_m\} \tag{6.30}$$

Using these dynamically-evolving cluster neighborhoods, we introduce a *separate* ATC diffusion learning strategy:

$$\boldsymbol{\psi}'_{k,i} = \boldsymbol{w}'_{k,i-1} - \mu_k \widehat{\nabla J_k}(\boldsymbol{w}'_{k,i-1}) \tag{6.31a}$$

$$\boldsymbol{w}'_{k,i} = \sum_{\ell \in \boldsymbol{\mathcal{N}}_{k,i-1}^+} \boldsymbol{a}'_{\ell k}(i-1) \boldsymbol{\psi}'_{\ell,i} \tag{6.31b}$$

where the combination coefficients $\{\boldsymbol{a}'_{\ell k}(i-1)\}$ become random because $\boldsymbol{\mathcal{N}}_{k,i-1}^+$ is random and may vary over iterations. The iteration index $i-1$ is used for these coefficients to enforce causality. Since $\mathcal{N}_k \cap \mathcal{G}_m$ denotes the neighbors of agent k that are already in the same group $\mathcal{G}_m$ as k, it is obvious that $\mathcal{N}_k \cap \mathcal{G}_m \subseteq \boldsymbol{\mathcal{N}}_{k,i-1}^+$ for any $i \geq 0$. This means that recursion (6.31a)–(6.31b) generally involves a larger range of interactions among agents than the first recursion (6.20a)–(6.20b). We summarize the algorithm in the following listing.

Distributed clustering and learning over networks

Initialization: $\boldsymbol{w}_{k,-1} = \boldsymbol{w}'_{k,-1} = 0$ and $\boldsymbol{\mathcal{N}}^{+}_{k,-1} = \mathcal{N}_k \cap \mathcal{G}_m$ for all $k \in \mathcal{G}_m$ and $m = 1, 2, \ldots, G$.

for $i \geq 0$ **do**

(1) Each agent k updates $\boldsymbol{w}_{k,i}$ according to the first recursion (6.20a)–(6.20b) over $\mathcal{N}_k \cap \mathcal{G}_m$.

(2) Each agent k updates $\boldsymbol{w}'_{k,i}$ according to the second recursion (6.31a)–(6.31b) over $\boldsymbol{\mathcal{N}}^{+}_{k,i-1}$.

(3) Each agent k updates $\boldsymbol{\mathcal{N}}^{+}_{k,i}$ by using (6.30) with $\{\boldsymbol{w}_{\ell,i}; \ell \in \mathcal{N}_k\}$ from step (1).

end for

6.3 Mean-Square-Error Analysis

In the previous section, we mentioned that Theorem 6.3 in Section 6.4.1 is the key result for the design of the clustering criterion. To arrive this theorem, we shall derive two useful intermediate results, Lemmas 6.1 and 6.2, in this section. These two results are related to the MSE analysis of the first recursion (6.20a)–(6.20b), which is used in step (1) of the proposed algorithm. We shall therefore examine the stability and the MSE performance of recursion (6.20a)–(6.20b) in the sequel. It is clear that the evolution of this recursion is not influenced by the other two steps. Thus, we can study recursion (6.20a)–(6.20b) independently.

6.3.1 Network Error Recursion

Using model (6.10), recursion (6.20a)–(6.20b) leads to

$$\boldsymbol{\mathcal{w}}_i = \mathcal{A}^{\mathsf{T}}\boldsymbol{\mathcal{w}}_{i-1} - \mathcal{A}^{\mathsf{T}}\mathcal{M}\,\nabla J(\boldsymbol{\mathcal{w}}_{i-1}) - \mathcal{A}^{\mathsf{T}}\mathcal{M}\boldsymbol{\mathcal{s}}_i(\boldsymbol{\mathcal{w}}_{i-1}) \tag{6.32}$$

where $\boldsymbol{\mathcal{w}}_i$ is from (6.24), $\nabla J(\cdot)$ is from (6.6), $\boldsymbol{\mathcal{s}}_i(\cdot)$ is from (6.9), and

$$\mathcal{M} \triangleq \text{diag}\{\mu_1, \ldots, \mu_N\} \otimes I_M \tag{6.33}$$

$$\mathcal{A} \triangleq A \otimes I_M \tag{6.34}$$

We introduce the network error vector:

$$\widetilde{\boldsymbol{\mathcal{w}}}_i \triangleq \mathcal{w}^o - \boldsymbol{\mathcal{w}}_i = \text{col}\{\widetilde{\boldsymbol{w}}_{1,i}, \ldots, \widetilde{\boldsymbol{w}}_{N,i}\} \tag{6.35}$$

where $\mathcal{w}^o$ is from (6.3), and the individual error vectors:

$$\widetilde{\boldsymbol{w}}_{k,i} \triangleq w_k^o - \boldsymbol{w}_{k,i} \tag{6.36}$$

Using the mean-value theorem [66, 72], we can write

$$\nabla J(\boldsymbol{\mathcal{w}}_{i-1}) = \nabla J(\mathcal{w}^o) - \left[\int_0^1 \nabla^2 J(\mathcal{w}^o - t\widetilde{\boldsymbol{\mathcal{w}}}_{i-1})dt\right]\widetilde{\boldsymbol{\mathcal{w}}}_{i-1} \tag{6.37}$$

where $\nabla^2 J(\cdot)$ is from (6.1). Since $\mathcal{w}^o$ consists of individual minimizers throughout the network, it follows that $\nabla J(\mathcal{w}^o) = 0$. Let

$$\boldsymbol{\mathcal{H}}_{i-1} \triangleq \int_0^1 \nabla^2 J(\mathcal{w}^o - t\widetilde{\boldsymbol{\mathcal{w}}}_{i-1})dt = \text{diag}\{\boldsymbol{H}_{k,i-1}\}_{k=1}^N \tag{6.38}$$

where

$$\boldsymbol{H}_{k,i-1} \triangleq \int_0^1 \nabla^2 J_k(w_k^o - t\widetilde{\boldsymbol{w}}_{k,i-1})dt \tag{6.39}$$

Then, expression (6.37) can be rewritten as

$$\nabla J(\boldsymbol{\mathcal{w}}_{i-1}) = -\boldsymbol{\mathcal{H}}_{i-1}\widetilde{\boldsymbol{\mathcal{w}}}_{i-1} \tag{6.40}$$

where it is worth noting that the random matrix $\boldsymbol{\mathcal{H}}_{i-1}$ is dependent on $\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}$. Substituting (6.40) into (6.32) yields:

$$\boldsymbol{\mathcal{W}}_i = \mathcal{A}^{\mathsf{T}}\boldsymbol{\mathcal{W}}_{i-1} + \mathcal{A}^{\mathsf{T}}\mathcal{M}\boldsymbol{\mathcal{H}}_{i-1}\widetilde{\boldsymbol{\mathcal{W}}}_{i-1} - \mathcal{A}^{\mathsf{T}}\mathcal{M}\boldsymbol{\mathcal{S}}_i(\boldsymbol{\mathcal{W}}_{i-1}) \tag{6.41}$$

By the indexing rule from Definition 6.1 and condition (6.21), the combination matrix A possesses a block diagonal structure:

$$A = \mathrm{diag}\{A_m; m = 1, \ldots, G\} \tag{6.42}$$

where each A_m collects the combination coefficients within group $\mathcal{G}_m$:

$$A_m \triangleq [a_{\ell k}; \ell, k \in \mathcal{G}_m] \tag{6.43}$$

From the same condition (6.21), we have that each A_m is itself an $N_m^g \times N_m^g$ left-stochastic matrix:

$$A_m^{\mathsf{T}}\mathbb{1}_{N_m^g} = \mathbb{1}_{N_m^g} \tag{6.44}$$

If group $\mathcal{G}_m$ is a subset of cluster $\mathcal{C}_q$, then the agents in $\mathcal{G}_m$ share the same minimizer at $w_q^\star$. Thus, for any $\mathcal{G}_m \subseteq \mathcal{C}_q$, let

$$\mathcal{W}_m^o \triangleq \mathrm{col}\{w_k^o; k \in \mathcal{G}_m\} = \mathbb{1}_{N_m^g} \otimes w_q^\star \tag{6.45}$$

It follows from (6.44) and (6.45) that

$$(A_m^{\mathsf{T}} \otimes I_M)\mathcal{W}_m^o = (A_m^{\mathsf{T}} \otimes I_M)(\mathbb{1}_{N_m^g} \otimes w_q^\star) = \mathcal{W}_m^o \tag{6.46}$$

Again, from the indexing rule in Definition 6.1, we have from (6.3) and (6.45) that

$$\mathcal{W}^o = \mathrm{col}\{\mathcal{W}_m^o; m = 1, \ldots, G\} \tag{6.47}$$

Then, it follows from (6.42) and (6.47) that

$$\mathcal{A}^{\mathsf{T}}\mathcal{W}^o = \begin{bmatrix} A_1^{\mathsf{T}} \otimes I_M & & \\ & \ddots & \\ & & A_G^{\mathsf{T}} \otimes I_M \end{bmatrix} \begin{bmatrix} \mathcal{W}_1^o \\ \vdots \\ \mathcal{W}_G^o \end{bmatrix} = \mathcal{W}^o \tag{6.48}$$

Accordingly, subtracting w^o from both sides of (6.41) and using (6.48) yields the network error recursion:

$$\widetilde{\boldsymbol{w}}_i = \mathcal{A}^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{H}}_{i-1})\widetilde{\boldsymbol{w}}_{i-1} + \mathcal{A}^{\mathsf{T}}\mathcal{M}\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{6.49}$$

We denote the coefficient matrix appearing in (6.49) by

$$\boldsymbol{\mathcal{B}}_{i-1} \triangleq \mathcal{A}^{\mathsf{T}}(I_{NM} - \mathcal{M}\boldsymbol{\mathcal{H}}_{i-1}) \tag{6.50}$$

Then, the network error recursion (6.49) can be rewritten as

$$\widetilde{\boldsymbol{w}}_i = \boldsymbol{\mathcal{B}}_{i-1}\widetilde{\boldsymbol{w}}_{i-1} + \mathcal{A}^{\mathsf{T}}\mathcal{M}\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{6.51}$$

We further introduce the group quantities:

$$\mathcal{A}_m \triangleq A_m \otimes I_M \tag{6.52}$$

$$\boldsymbol{w}_{m,i} \triangleq \mathrm{col}\{\boldsymbol{w}_{k,i}; k \in \mathcal{G}_m\} \in \mathbb{R}^{N_m^g M \times 1} \tag{6.53}$$

$$\mathcal{M}_m \triangleq \mathrm{diag}\{\mu_k; k \in \mathcal{G}_m\} \otimes I_M \tag{6.54}$$

$$\boldsymbol{\mathcal{H}}_{m,i-1} \triangleq \mathrm{diag}\{\boldsymbol{H}_{k,i-1}; k \in \mathcal{G}_m\} \tag{6.55}$$

$$\boldsymbol{s}_{m,i}(\boldsymbol{w}_{m,i-1}) \triangleq \mathrm{col}\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}); k \in \mathcal{G}_m\} \tag{6.56}$$

It follows from the indexing rule in Definition 6.1 that

$$\mathcal{A} = \mathrm{diag}\{\mathcal{A}_1, \ldots, \mathcal{A}_G\} \tag{6.57}$$

$$\boldsymbol{w}_i = \mathrm{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{G,i}\} \tag{6.58}$$

$$\mathcal{M} = \mathrm{diag}\{\mathcal{M}_1, \ldots, \mathcal{M}_G\} \tag{6.59}$$

$$\boldsymbol{\mathcal{H}}_{i-1} = \mathrm{diag}\{\boldsymbol{\mathcal{H}}_{1,i-1}, \ldots, \boldsymbol{\mathcal{H}}_{G,i-1}\} \tag{6.60}$$

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) = \mathrm{col}\{\boldsymbol{s}_{1,i}(\boldsymbol{w}_{1,i-1}), \ldots, \boldsymbol{s}_{G,i}(\boldsymbol{w}_{G,i-1})\} \tag{6.61}$$

Using (6.57)–(6.60), the matrix $\boldsymbol{\mathcal{B}}_{i-1}$ in (6.50) can be expressed by

$$\boldsymbol{\mathcal{B}}_{i-1} = \mathrm{diag}\{\boldsymbol{\mathcal{B}}_{1,i-1}, \ldots, \boldsymbol{\mathcal{B}}_{G,i-1}\} \tag{6.62}$$

where

$$\boldsymbol{\mathcal{B}}_{m,i-1} \triangleq \mathcal{A}_m^{\mathsf{T}}(I_{N_m^g M} - \mathcal{M}_m \boldsymbol{\mathcal{H}}_{m,i-1}) \tag{6.63}$$

Due to the block structures in (6.57)–(6.62), groups are isolated from each other. Therefore, using these group quantities, the network error recursion (6.51) is automatically decoupled into a total of G group error recursions, where the m-th recursion is given by

$$\widetilde{\boldsymbol{\mathcal{W}}}_{m,i} = \boldsymbol{\mathcal{B}}_{m,i-1}\widetilde{\boldsymbol{\mathcal{W}}}_{m,i-1} + \mathcal{A}_m^{\mathsf{T}}\mathcal{M}_m \boldsymbol{\mathcal{S}}_{m,i}(\boldsymbol{\mathcal{W}}_{m,i-1}) \tag{6.64}$$

6.3.2 Mean-Square and Mean-Fourth-Order Error Stability

The stability of the network error recursion (6.51) is now reduced to studying the stability of the group recursions (6.64). Recall that, by Definition 1.2 in Chapter 1, the agents in each group are connected. Moreover, condition (6.21) implies that agents in each group have non-trivial self-loops, meaning that $a_{kk} > 0$ for all $k \in \mathcal{G}_m$. It follows that each A_m is a primitive matrix [8, 79] (which is satisfied as long as there exists at least one $a_{kk} > 0$ in each group). Under these conditions, we are now able to ascertain the stability of the second and fourth-order error moments of the network error recursion (6.51) by appealing to results from [66].

Theorem 6.1 (Stability of error moments)**.** *For sufficiently small step-sizes, the network error recursion* (6.51) *is mean-square and mean-fourth-order stable in the sense that*

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2 = O(\mu_{\max}) \tag{6.65}$$

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^4 = O(\mu_{\max}^2) \tag{6.66}$$

Proof. It is obvious that the network error recursion (6.51) is mean-square and mean-fourth-order stable if, and only if, each group error recursion (6.64) is stable

in a similar sense. From Assumption 6.2, we know that there exists at least one strongly-convex cost in each group. Since the combination matrix A_m for each group is primitive and left-stochastic, we can now call upon Theorems 9.1 and 9.2 from [66, p. 508, p. 522] to conclude that every group error recursion is mean-square and mean-fourth-order stable, namely,

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}\|^2 = O(\mu_{\max}) \tag{6.67}$$

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}\|^4 = O(\mu_{\max}^2) \tag{6.68}$$

from which (6.65) and (6.66) follow. □

6.3.3 Long-Term Model

Once network stability is established, we can proceed to assess the performance of the adaptive clustering and learning procedure. To do so, it becomes more convenient to first introduce a long-term model for the error dynamics (6.51). Note that recursion (6.51) represents a non-linear, time-variant, and stochastic system that is driven by a state-dependent random noise process. Analysis of recursion (6.51) is facilitated by noting (see Lemma 6.1 below) that when the step-size parameter $\mu_{\max}$ is small enough, the mean-square behavior of (6.51) in steady-state, when $i \gg 1$, can be well approximated by the behavior of the following long-term model:

$$\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{lt}} = \mathcal{B}\,\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}^{\mathrm{lt}} + \mathcal{A}^{\mathsf{T}}\mathcal{M}\boldsymbol{\mathcal{S}}_i(\boldsymbol{\mathcal{W}}_{i-1}) \tag{6.69}$$

where we replaced the random matrix $\boldsymbol{\mathcal{B}}_{i-1}$ in (6.51) by the constant matrix

$$\mathcal{B} \triangleq \mathcal{A}^{\mathsf{T}}(I_{NM} - \mathcal{M}\mathcal{H}) \tag{6.70}$$

In (6.70), the matrix $\mathcal{H}$ is defined by

$$\mathcal{H} \triangleq \mathrm{diag}\{H_1, \ldots, H_N\} \tag{6.71}$$

where

$$H_k \triangleq \nabla^2 J_k(w_k^o) \tag{6.72}$$

Note that the long-term model (6.69) is now a *linear time-invariant* system, albeit one that continues to be driven by the *same* random noise process as in (6.51). Similarly to the original error recursion (6.51), the long-term recursion (6.69) can also be decoupled into G recursions, one for each group:

$$\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}} = \mathcal{B}_m \, \widetilde{\boldsymbol{\mathcal{W}}}_{m,i-1}^{\mathrm{lt}} + \mathcal{A}_m^{\mathsf{T}} \mathcal{M}_m \boldsymbol{\mathcal{S}}_{m,i}(\boldsymbol{\mathcal{W}}_{m,i-1}) \tag{6.73}$$

where

$$\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}} \triangleq \mathrm{col}\{\widetilde{\boldsymbol{w}}_{k,i}^{\mathrm{lt}}; k \in \mathcal{G}_m\} \in \mathbb{R}^{N_m^g M \times 1} \tag{6.74}$$

$$\mathcal{B}_m \triangleq \mathcal{A}_m^{\mathsf{T}}(I_{N_m^g M} - \mathcal{M}_m \mathcal{H}_m) \tag{6.75}$$

$$\mathcal{H}_m \triangleq \mathrm{diag}\{H_k; k \in \mathcal{G}_m\} \tag{6.76}$$

$$\mathcal{W}_m^o \triangleq \mathrm{col}\{w_k^o; k \in \mathcal{G}_m\} \tag{6.77}$$

Lemma 6.1 (Accuracy of long-term model)**.** *For sufficiently small step-sizes, the evolution of the long-term model* (6.69) *is close to the original error recursion* (6.51) *in MSE sense:*

$$\limsup_{i \to \infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i - \widetilde{\boldsymbol{\mathcal{W}}}_i^{lt}\|^2 = O(\mu_{\max}^2) \tag{6.78}$$

Proof. We call upon Theorem 10.2 from [66, p. 557] to conclude that the difference between each group error recursion (6.64) and its long-term model (6.73) satisfies:

$$\limsup_{i \to \infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_{m,i} - \widetilde{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}}\|^2 = O(\mu_{\max}^2) \tag{6.79}$$

for all m. It is then immediate to conclude that (6.78) holds. □

6.3.4 Low-Dimensional Model

Lemma 6.1 indicates that we can assess the MSE dynamics of the original network recursion (6.51) to first-order in $\mu_{\max}$ by working with the long-term model (6.69). It turns out that the state variable of the long-term model can be split into two parts, one consisting of the *centroids* of each group and the other consisting of in-group discrepancies. The details of this splitting are not important for our current discussion but interested readers can refer to Sec. V of [67] and Eq. (10.37) of [66, p. 558] for a detailed explanation. Here we only use this fact to motivate the introduction of the low-dimensional model. Moreover, it also turns out that the first part, i.e, the part corresponding to the centroids, is the dominant component in the evolution of the error dynamics and that the evolution of the two parts (centroids and in-group discrepancies) is weakly-coupled. By retaining the first part, we can therefore arrive at a low-dimensional model that will allow us to assess performance in closed-form to first-order in $\mu_{\max}$. To arrive at the low-dimensional model, we need to exploit the eigen-structure of the combination matrix A, or, equivalently, that of each A_m.

Recall that we indicated earlier prior to the statement of Theorem 6.1 that each A_m is a primitive and left-stochastic matrix. By the Perron-Frobenius theorem [66, 79, 82], it follows that each A_m has a simple eigenvalue at one with all other eigenvalues lying strictly inside the unit circle. Moreover, if we let $p_m^g \in \mathbb{R}^{N_m^g \times 1}$ denote the right-eigenvector of A_m that is associated with the eigenvalue at one, and normalize its entries to add up to one, then the same theorem ensures that all entries of p_m^g will be positive:

$$p_m^g \triangleq \mathrm{col}\{p_{m,k}^g\}_{k=1}^{N_m^g} \succ 0,\ A_m p_m^g = p_m^g,\ \mathbb{1}_{N_m^g}^{\mathsf{T}} p_m^g = 1 \tag{6.80}$$

where $p_{m,k}^g$ denotes the k-th entry of p_m^g. This means that we can express each

A_m in the form (see (6.165) further ahead):

$$A_m = p_m^g \mathbb{1}_{N_m^g}^{\mathsf{T}} + V_{m,R} J_{m,\epsilon} V_{m,L}^{\mathsf{T}} \tag{6.81}$$

for some eigenvector matrices $V_{m,R}$ and $V_{m,L}$, and where $J_{m,\epsilon}$ denotes the collection of the Jordan blocks with eigenvalues inside the unit circle and with their unit entries on the first lower sub-diagonal replaced by some arbitrarily small constant $0 < \epsilon \ll 1$. The first rank-one component on the RHS of (6.81) represents the contribution by the largest eigenvalue of A_m, and this component will be used further ahead to describe the centroid of group $\mathcal{G}_m$. The network Perron eigenvector is obtained by stacking the group Perron eigenvectors $\{p_m^g\}$:

$$p \triangleq \mathrm{col}\{p_1^g, \ldots, p_G^g\} \triangleq \mathrm{col}\{p_1, \ldots, p_N\} \tag{6.82}$$

where p_k denotes the k-th entry of $p \in \mathbb{R}^{N \times 1}$. According to the indexing rule from Definition 6.1, it is obvious that $p_m^g = \mathrm{col}\{p_k; k \in \mathcal{G}_m\}$.

Now, for each group $\mathcal{G}_m$, we introduce the low-dimensional (centroid) error recursion defined by (compare with (6.73)):

$$\widetilde{\boldsymbol{w}}_{m,i}^{\mathrm{ld}} = D_m \widetilde{\boldsymbol{w}}_{m,i-1}^{\mathrm{ld}} + (p_m^g \otimes I_M)^{\mathsf{T}} \mathcal{M}_m \boldsymbol{s}_{m,i}(\boldsymbol{\mathcal{w}}_{m,i-1}) \tag{6.83}$$

where $\widetilde{\boldsymbol{w}}_{m,i}^{\mathrm{ld}}$ is $M \times 1$, and D_m is $M \times M$ and defined by

$$D_m \triangleq I_M - \mu_{\max} \bar{H}_m \tag{6.84}$$

where

$$\begin{aligned} \bar{H}_m &\triangleq \mu_{\max}^{-1} (p_m^g \otimes I_M)^{\mathsf{T}} \mathcal{M}_m \mathcal{H}_m (\mathbb{1}_{N_m^g} \otimes I_M) \\ &= \sum_{k \in \mathcal{G}_m} \frac{p_k \mu_k}{\mu_{\max}} H_k = O(\mu_{\max}^0) \end{aligned} \tag{6.85}$$

The matrix $\bar{H}_m$ is positive definite since there is at least one Hessian matrix in $\{H_k; k \in \mathcal{G}_m\}$ that is positive definite according to Assumption 6.2. We collect

the low-rank recursions (6.83) for groups into one recursion for the entire network by stacking them on top of each other:

$$\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}} = \mathcal{D}\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}^{\mathrm{ld}} + \mathcal{P}^{\mathsf{T}}\mathcal{M}\boldsymbol{\mathcal{S}}_i(\boldsymbol{\mathcal{W}}_{i-1}) \tag{6.86}$$

where

$$\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}} \triangleq \mathrm{col}\{\widetilde{\boldsymbol{w}}_{1,i}^{\mathrm{ld}}, \ldots, \widetilde{\boldsymbol{w}}_{G,i}^{\mathrm{ld}}\} \in \mathbb{R}^{GM\times 1} \tag{6.87}$$

$$\mathcal{D} \triangleq \mathrm{diag}\{D_1, \ldots, D_G\} \in \mathbb{R}^{GM\times GM} \tag{6.88}$$

$$\mathcal{P} \triangleq \mathrm{diag}\{p_1^g, \ldots, p_G^g\} \otimes I_M \in \mathbb{R}^{NM\times GM} \tag{6.89}$$

Recursion (6.86) describes the joint dynamics of all the centroids (one for each group). Note that the dimension of $\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}$ in (6.86) is GM, which is lower than the dimension, NM, of $\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{lt}}$ in (6.69) or $\widetilde{\boldsymbol{\mathcal{W}}}_i$ in (6.51), because $G \leq N$ by Assumption 6.1. In order to measure the difference between the dynamics of the long-term model (6.69) and the low-dimensional model (6.86), we expand $\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}$ in the following manner (compare with (6.87)):

$$\bar{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}} \triangleq \mathrm{col}\{\bar{\boldsymbol{\mathcal{W}}}_{1,i}^{\mathrm{ld}}, \ldots, \bar{\boldsymbol{\mathcal{W}}}_{G,i}^{\mathrm{ld}}\} \in \mathbb{R}^{NM\times 1} \tag{6.90}$$

$$\bar{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{ld}} \triangleq \mathbb{1}_{N_m^g} \otimes \widetilde{\boldsymbol{w}}_{m,i}^{\mathrm{ld}} \in \mathbb{R}^{N_m^g M\times 1} \tag{6.91}$$

because $\sum_{m=1}^{G} N_m^g = N$ according to Assumption 6.1.

Lemma 6.2 (Accuracy of low-dimensional model)**.** *For sufficiently small step-sizes, the low-dimensional model* (6.86) *is close to the network long-term model* (6.69) *in the following sense:*

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i^{lt} - \bar{\boldsymbol{\mathcal{W}}}_i^{ld}\|^2 = O(\mu_{\max}^2) \tag{6.92}$$

where $\bar{\boldsymbol{\mathcal{W}}}_i^{ld}$ *is given by* (6.90) *and is related to* $\widetilde{\boldsymbol{\mathcal{W}}}_i^{ld}$ *via* (6.91)*.*

Proof. See Appendix 6.A. □

Lemma 6.3 (Low-dimensional error covariance)**.** *For sufficiently small step-sizes, the covariance matrix for* $\widetilde{\boldsymbol{\mathcal{W}}}_i^{ld}$ *satisfies*

$$\limsup_{i\to\infty} \|\mathbb{E}[\widetilde{\boldsymbol{\mathcal{W}}}_i^{ld}(\widetilde{\boldsymbol{\mathcal{W}}}_i^{ld})^{\mathsf{T}}] - \Theta\| = O(\mu_{\max}^{1+\gamma_s/2}) \tag{6.93}$$

where $\Theta \in \mathbb{R}^{GM\times GM}$ *is symmetric, positive-definite, and uniquely solves the discrete Lyapunov equation:*

$$\Theta = \mathcal{D}\Theta\mathcal{D} + \mathcal{P}^{\mathsf{T}}\mathcal{M}\mathcal{R}_s\mathcal{M}\mathcal{P} \tag{6.94}$$

Proof. See Appendix 6.B. □

6.3.5 Steady-State MSE Performance

From Theorem 6.1, we know that the limit superior of the MSE is bounded within $O(\mu_{\max})$. In order to define meaningful steady-state performance metrics, we consider the case in which the step-sizes approach zero asymptotically. Results obtained in this case are representative of operation in the slow adaptation regime (see Sec. 11.2 of [66, pp. 581–583]).

Lemma 6.4 (Steady-state normalized MSD)**.** *The normalized total MSD of* $\widetilde{\boldsymbol{\mathcal{W}}}_i$ *in* (6.51) *is given by*

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty} \mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2 = \sum_{m=1}^{G}\frac{N_m^g}{2\mu_{\max}}\mathrm{Tr}\left[\left(\sum_{k\in\mathcal{G}_m} p_k\mu_k H_k\right)^{-1}\left(\sum_{k\in\mathcal{G}_m} p_k^2\mu_k^2 R_k\right)\right] \tag{6.95}$$

where H_k *is from* (6.72) *and* R_k *is the* m*-th block on the diagonal of* $\mathcal{R}_s$ *from* (6.15) *with block size* $M\times M$*.*

Proof. The normalized total MSD is the sum of the normalized MSD for each group. From Lemma 11.3 of [66, p. 594], the normalized MSD for each group $\mathcal{G}_m$

is given by

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}\|^2=\frac{N_m^g}{2\mu_{\max}}\mathrm{Tr}\left[\left(\sum_{k\in\mathcal{G}_m}p_k\mu_k H_k\right)^{-1}\left(\sum_{k\in\mathcal{G}_m}p_k^2\mu_k^2 R_k\right)\right] \tag{6.96}$$

Note that we calculate the *normalized total* MSD rather than the *average* MSD in (6.95) and (6.96). □

In order to examine the statistical properties of the error vector $\widetilde{\boldsymbol{\mathcal{W}}}_i$, we need to strengthen the result in Lemma 6.4 by evaluating the full normalized error covariance matrix of $\widetilde{\boldsymbol{\mathcal{W}}}_i$ in steady-state. From Lemmas 6.1 and 6.2, it is clear that the mean-square dynamics of the original error recursion (6.51) can be well approximated by the low-dimensional model (6.86). And it was shown in Eq. (10.78) of [66, p. 563] that the variances of the centroids $\{\widetilde{\boldsymbol{w}}_{k,i}^{\mathrm{ld}}\}$ are in the order of $\mu_{\max}$ in steady-state, which implies that

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}\|^2=O(\mu_{\max}^0) \tag{6.97}$$

Since the induced-2 norm of the covariance matrix of any random vector is always bounded by its variance, i.e., $\|\mathbb{E}\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|\le\mathbb{E}\|\boldsymbol{x}\|^2$ by using Jensen's inequality, it follows from (6.97) that the normalized covariance matrix of $\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}$ is finite in steady-state. Moreover, since Lemma 6.3 applies to any positive value of $\mu_{\max}$ as long as it is small enough to ensure stability, we can take the limit of $\mu_{\max}$ in (6.93) by letting it approach zero asymptotically. That is,

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\|\mu_{\max}^{-1}\mathbb{E}[\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}(\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}})^{\mathsf{T}}]-\Phi\|=0 \tag{6.98}$$

where

$$\Phi\triangleq\lim_{\mu_{\max}\to 0}(\mu_{\max}^{-1}\Phi_i) \tag{6.99}$$

Due to (6.97) and (6.98), Φ is in the order of $\mu_{\max}^0$, i.e., $\|\Phi\| = O(\mu_{\max}^0)$. In fact, by introducing $\Phi_i \triangleq \mu_{\max}^{-1}\mathbb{E}[\widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm ld}(\widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm ld})^{\mathsf{T}}]$ and using the triangle inequality, we have

$$\|\Phi\| = \|\Phi - \Phi_i + \Phi_i\| \leq \|\Phi - \Phi_i\| + \|\Phi_i\| \tag{6.100}$$

$$\|\Phi_i\| = \|\Phi_i - \Phi + \Phi\| \leq \|\Phi_i - \Phi\| + \|\Phi\| \tag{6.101}$$

Taking $i \to \infty$ and $\mu_{\max} \to 0$ for both (6.100) and (6.101) yields:

$$\|\Phi\| \leq \lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \|\Phi_i\| \tag{6.102}$$

$$\|\Phi\| \geq \lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \|\Phi_i\| \tag{6.103}$$

by using (6.98). From (6.102) and (6.103), we get

$$\|\Phi\| = \lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \|\Phi_i\| \tag{6.104}$$

Since $\Phi_i \in \mathbb{R}^{GM\times GM}$ is positive semi-definite, it holds that

$$(GM)^{-1}\mathrm{Tr}(\Phi_i) \leq \|\Phi_i\| \leq \mathrm{Tr}(\Phi_i) \tag{6.105}$$

where we used the fact for any positive semi-definite matrix $X \geq 0$ that (i) all the eigenvalues of X are nonnegative, (ii) $\|X\|$ is equal to the largest eigenvalue of X, and (iii) $\mathrm{Tr}(X)$ is equal to the sum of all the eigenvalues of X. Moreover,

$$\mathrm{Tr}(\Phi_i) = \mathrm{Tr}(\mu_{\max}^{-1}\mathbb{E}[\widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm ld}(\widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm ld})^{\mathsf{T}}]) = \mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm ld}\|^2 \tag{6.106}$$

Using (6.97), it follows from (6.105) and (6.106) that

$$\lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \|\Phi_i\| = O(\mu_{\max}^0) \tag{6.107}$$

Substituting (6.107) into (6.104) yields the desired result, i.e., $\|\Phi\| = O(\mu_{\max}^0)$. Then, according to (6.99), Φ is the unique solution to equation (6.94) when $\mu_{\max} \to 0$ asymptotically. Introduce two $GM \times GM$ matrices:

$$\bar{\mathcal{H}} \triangleq \mathrm{diag}\{\bar{H}_1, \ldots, \bar{H}_G\} = O(\mu_{\max}^0) \tag{6.108}$$

$$\bar{\mathcal{R}} \triangleq \mu_{\max}^{-2}\mathcal{P}^{\mathsf{T}}\mathcal{M}\mathcal{R}_s\mathcal{M}\mathcal{P} = O(\mu_{\max}^{0}) \tag{6.109}$$

where $\bar{H}_m$ is from (6.85) and $\mathcal{R}_s$ is from (6.15). It is easy to verify that $\bar{\mathcal{H}}$ and $\bar{\mathcal{R}}$ are symmetric and positive-definite according to Assumptions 6.2 and 6.3. From (6.88), (6.108), and (6.84), we get

$$\mathcal{D} = I_{GM} - \mu_{\max}\bar{\mathcal{H}} \tag{6.110}$$

Using (6.99)–(6.110), equation (6.94) reduces to

$$\bar{\mathcal{H}}\Phi + \Phi\bar{\mathcal{H}} = \bar{\mathcal{R}} + \mu_{\max}\bar{\mathcal{H}}\Phi\bar{\mathcal{H}} \tag{6.111}$$

Since $\bar{\mathcal{H}}$ and $\bar{\mathcal{R}}$ are constant matrices, and Φ is finite, the last term on the RHS of (6.111) disappears as $\mu_{\max} \to 0$ asymptotically. Therefore, we conclude that Φ is the unique solution to the continuous Lyapunov equation:

$$\bar{\mathcal{H}}\Phi + \Phi\bar{\mathcal{H}} = \bar{\mathcal{R}} \tag{6.112}$$

Let us define the *normalized* network error covariance matrix for $\widetilde{\boldsymbol{\mathcal{W}}}_i$ from (6.51) by

$$\Pi_i \triangleq \mu_{\max}^{-1}\mathbb{E}(\widetilde{\boldsymbol{\mathcal{W}}}_i\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathsf{T}}) \tag{6.113}$$

Theorem 6.2 (Block structure)**.** *In steady-state, and as the step-sizes approach zero asymptotically, the normalized network error covariance matrix* Π_i *in* (6.113) *satisfies*

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\|\Pi_i - \Pi\| = 0 \tag{6.114}$$

where

$$\Pi \triangleq \begin{bmatrix} (\mathbb{1}_{N_1^g}\mathbb{1}_{N_1^g}^{\mathsf{T}}) \otimes \Phi_{1,1} & \cdots & (\mathbb{1}_{N_1^g}\mathbb{1}_{N_G^g}^{\mathsf{T}}) \otimes \Phi_{1,G} \\ \vdots & \ddots & \vdots \\ (\mathbb{1}_{N_G^g}\mathbb{1}_{N_1^g}^{\mathsf{T}}) \otimes \Phi_{G,1} & \cdots & (\mathbb{1}_{N_G^g}\mathbb{1}_{N_G^g}^{\mathsf{T}}) \otimes \Phi_{G,G} \end{bmatrix} \tag{6.115}$$

and $\Phi_{m,r}$ *denotes the* (m,r)*-th block of* Φ *from* (6.112) *with block size* $M \times M$*.*

Proof. See Appendix 6.C. □

6.4 Error Probability Analysis for Clustering

Using the results from the previous section, we now move on to assess the error probabilities for the hypothesis testing problem (6.29). To do so, we need to determine the probability distribution of the decision statistic that is generated by recursion (6.20a)–(6.20b).

6.4.1 Asymptotic Joint Distribution of Estimation Errors

Using (6.110), we rewrite the low-dimensional model (6.86) as

$$\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}} = \widetilde{\boldsymbol{\mathcal{W}}}_{i-1}^{\mathrm{ld}} - \mu_{\max}\bar{\mathcal{H}}\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}^{\mathrm{ld}} + \mu_{\max}\bar{\boldsymbol{s}}_i \tag{6.116}$$

where $\bar{\mathcal{H}}$ is from (6.108) and

$$\bar{\boldsymbol{s}}_i \triangleq \mu_{\max}^{-1}\mathcal{P}^{\mathsf{T}}\boldsymbol{\mathcal{M}}\boldsymbol{\mathcal{s}}_i(\boldsymbol{\mathcal{W}}_{i-1}) \in \mathbb{R}^{GM\times 1} \tag{6.117}$$

Lemma 6.5 (Rate of weak convergence)**.** *The normalized error vector sequence,* $\{\widetilde{\boldsymbol{\mathcal{W}}}_i^{ld}/\sqrt{\mu_{\max}}; i \geq 0\}$*, from* (6.116) *converges in distribution as* $i \to \infty$ *and* $\mu_{\max} \to 0$ *to the Gaussian random variable:*

$$\boldsymbol{\xi} \triangleq \mathrm{col}\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_G\} \sim \mathbb{N}(0, \Phi) \tag{6.118}$$

where $\boldsymbol{\xi}_m \in \mathbb{R}^{M\times 1}$ *for all* m*, and* $\Phi \in \mathbb{R}^{GM\times GM}$ *is the unique solution to the Lyapunov equation* (6.112)*.*

Proof. See Appendix 6.D. □

In the sequel we establish the main result that the distribution of the normalized error sequence from (6.51), $\{\widetilde{\boldsymbol{\mathcal{W}}}_i/\sqrt{\mu_{\max}}; i \geq 0\}$, asymptotically approaches a Gaussian distribution. According to Definition 4 from [121, p. 253], a random

sequence $\{\boldsymbol{\zeta}_i; i \geq 0\}$ converges in distribution to some random variable $\boldsymbol{\zeta}$ if, and only if,

$$\lim_{i\to\infty} \mathbb{E}\,|f(\boldsymbol{\zeta}_i) - f(\boldsymbol{\zeta})| = 0 \tag{6.119}$$

for *any* bounded continuous function $f(\cdot)$. We use this fact together with the following lemma to establish Theorem 6.3 further ahead.

Lemma 6.6 (Weak convergence)**.** *Let $\{\boldsymbol{\zeta}_i; i \geq 0\}$ and $\{\boldsymbol{\eta}_i; i \geq 0\}$ be two random sequences that are dependent on the parameter $\mu_{\max}$. If $\{\boldsymbol{\zeta}_i; i \geq 0\}$ approaches $\{\boldsymbol{\eta}_i; i \geq 0\}$ in mean-square sense:*

$$\lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\|^2 = 0 \tag{6.120}$$

and the variances of $\{\boldsymbol{\zeta}_i\}$ converge in the following sense:

$$\lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{\zeta}_i\|^2 = \sigma^2 \tag{6.121}$$

then it holds for any bounded continuous function $f(\cdot)$ that

$$\lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \mathbb{E}|f(\boldsymbol{\zeta}_i) - f(\boldsymbol{\eta}_i)| = 0 \tag{6.122}$$

Proof. See Appendix 6.E. □

Theorem 6.3 (Asymptotic normality)**.** *As $i \to \infty$ and $\mu_{\max} \to 0$, the normalized error sequence from* (6.51)*, $\{\widetilde{\boldsymbol{\mathcal{W}}}_i/\sqrt{\mu_{\max}}; i \geq 0\}$, converges in distribution* close *to the Gaussian random variable:*

$$\boldsymbol{\zeta} \triangleq \mathrm{col}\{\mathbb{1}_{N_1^g} \otimes \boldsymbol{\xi}_1, \ldots, \mathbb{1}_{N_G^g} \otimes \boldsymbol{\xi}_G\} \sim \mathbb{N}(0, \Pi) \tag{6.123}$$

in the following sense:

$$\lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \mathbb{E}\left| f\left(\frac{\widetilde{\boldsymbol{\mathcal{W}}}_i}{\sqrt{\mu_{\max}}}\right) - f(\boldsymbol{\zeta})\right| = 0 \tag{6.124}$$

for any bounded continuous function $f(\cdot) : \mathbb{R}^{NM\times 1} \mapsto \mathbb{R}$, where $\{\boldsymbol{\xi}_m\}$ are from (6.118)*, and Π is from* (6.115)*.*

Proof. Using the triangle inequality, we have

$$\begin{aligned}\mathbb{E}\left|f\left(\frac{\widetilde{\boldsymbol{\mathcal{W}}}_i}{\sqrt{\mu_{\max}}}\right)-f(\boldsymbol{\zeta})\right| \leq\ & \mathbb{E}\left|f\left(\frac{\widetilde{\boldsymbol{\mathcal{W}}}_i}{\sqrt{\mu_{\max}}}\right)-f\left(\frac{\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{lt}}}{\sqrt{\mu_{\max}}}\right)\right| \\ & +\mathbb{E}\left|f\left(\frac{\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{lt}}}{\sqrt{\mu_{\max}}}\right)-f\left(\frac{\bar{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}}{\sqrt{\mu_{\max}}}\right)\right| \\ & +\mathbb{E}\left|f\left(\frac{\bar{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}}{\sqrt{\mu_{\max}}}\right)-f(\boldsymbol{\zeta})\right| \end{aligned} \tag{6.125}$$

where $\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{lt}}$ is from the long-term model (6.69), and $\bar{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}$ is from (6.90) and is related to the low-dimensional model (6.86). By Lemma 6.4, the variances of the sequence $\{\widetilde{\boldsymbol{\mathcal{W}}}_i/\sqrt{\mu_{\max}}; i \geq 0\}$ converge to its normalized MSD in (6.95) in a sense similar to (6.121). Using Lemma 6.1, it is clear that $\{\widetilde{\boldsymbol{\mathcal{W}}}_i/\sqrt{\mu_{\max}}; i \geq 0\}$ approaches $\{\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{lt}}/\sqrt{\mu_{\max}}; i \geq 0\}$ in a sense similar to (6.120). Therefore, by calling upon Lemma 6.6, we conclude that the limit superior of the first term on the RHS of (6.125) vanishes. Likewise, using Lemmas 6.1 and 6.4, it can be verified that the variances of the sequence $\{\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{lt}}/\sqrt{\mu_{\max}}; i \geq 0\}$ also converge to the same normalized MSD in (6.95). Therefore, from Lemmas 6.2 and 6.6, the limit superior of the second term on the RHS of (6.125) vanishes. The limit superior of the third term vanishes since $\{\bar{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}/\sqrt{\mu_{\max}}; i \geq 0\}$ converges in distribution to $\boldsymbol{\zeta}$, which follows from Lemma 6.5. Therefore, the limit superior of the RHS of (6.125) vanishes when $i \to \infty$ and $\mu_{\max} \to 0$. □

Theorem 6.3 allows us to approximate the distribution of $\widetilde{\boldsymbol{\mathcal{W}}}_i/\sqrt{\mu_{\max}}$ by the Gaussian distribution $\mathbb{N}(0, \Pi)$ for large enough i and small enough $\mu_{\max}$.

6.4.2 Statistical Decision on Clustering

In Theorem 6.3, we established that for large enough i and for sufficiently small $\mu_{\max}$, the joint distribution of the individual estimators $\{\boldsymbol{w}_{k,i}; k = 1, 2, \ldots, N\}$

can be well approximated by a Gaussian distribution (6.123). Therefore, the marginal distribution for any pair of estimators, say, $\boldsymbol{w}_{k,i}$ and $\boldsymbol{w}_{\ell,i}$, can be well approximated by the Gaussian distribution:

$$\begin{bmatrix} \boldsymbol{w}_{k,i} \\ \boldsymbol{w}_{\ell,i} \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} w_k^o \\ w_\ell^o \end{bmatrix}, \mu_{\max} \begin{bmatrix} \Pi_{k,k} & \Pi_{k,\ell} \\ \Pi_{\ell,k} & \Pi_{\ell,\ell} \end{bmatrix} \right) \tag{6.126}$$

where w_k^o and w_ℓ^o are their individual minimizers, and $\Pi_{k,\ell}$ denotes the (k,ℓ)-th block of Π with block size $M \times M$. Without loss of generality, let us consider the scenario where agent k is from group $\mathcal{G}_m$ in cluster $\mathcal{C}_q$ and agent ℓ is from group $\mathcal{G}_n$ in cluster $\mathcal{C}_r$, i.e., $k \in \mathcal{G}_m \subseteq \mathcal{C}_q$ and $\ell \in \mathcal{G}_n \subseteq \mathcal{C}_r$. Then, we have from Definition 1.1 in Chapter 1 that

$$w_k^o = w_q^\star, \qquad w_\ell^o = w_r^\star \tag{6.127}$$

From Theorem 6.2, the covarince matrix Π possesses the block structure shown in (6.115). Using (6.115), and noticing that $k \in \mathcal{G}_m$ and $\ell \in \mathcal{G}_n$, it is obvious that

$$\Pi_{k,k} = \Phi_{m,m},\ \Pi_{k,\ell} = \Phi_{m,n},\ \Pi_{\ell,k} = \Phi_{n,m},\ \Pi_{\ell,\ell} = \Phi_{n,n} \tag{6.128}$$

Then, it follows from (6.126)–(6.128) that

$$\begin{bmatrix} \boldsymbol{w}_{k,i} \\ \boldsymbol{w}_{\ell,i} \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} w_q^\star \\ w_r^\star \end{bmatrix}, \mu_{\max} \begin{bmatrix} \Phi_{m,m} & \Phi_{m,n} \\ \Phi_{n,m} & \Phi_{n,n} \end{bmatrix} \right) \tag{6.129}$$

which means that the mean and covariance of the joint distribution for any pair of agents k and ℓ only depends on their groups. In other words, for any two agents k_1 and k_2 from the same group $\mathcal{G}_m$, the joint distribution of $\{k_1, \ell\}$ and the joint distribution of $\{k_2, \ell\}$ will be well approximated by the same Gaussian distribution in (6.129). Therefore, if both agents k_1 and k_2 need to decide whether agent ℓ is in the same cluster as they are, then they will have the same error probabilities in the hypothesis test (6.29).

Based on (6.129), the hypothesis test problem for clustering now becomes that of determining whether or not the two (near) Gaussian random vectors $\boldsymbol{w}_{k,i}$ and $\boldsymbol{w}_{\ell,i}$ have the same mean. Suppose the samples from the two variables are paired. The difference

$$\boldsymbol{d}_{k,\ell} \triangleq \boldsymbol{w}_{k,i} - \boldsymbol{w}_{\ell,i} \tag{6.130}$$

serves as a sufficient statistics [122]. Since $\boldsymbol{w}_{k,i}$ and $\boldsymbol{w}_{\ell,i}$ are jointly Gaussian in (6.129), their difference $\boldsymbol{d}_{k,\ell}$ is also Gaussian:

$$\boldsymbol{d}_{k,\ell} \sim \mathbb{N}(d^{\star}_{q,r}, \mu_{\max}\Delta_{m,n}) \tag{6.131}$$

where

$$d^{\star}_{q,r} \triangleq w^{\star}_{q} - w^{\star}_{r} \tag{6.132}$$

$$\Delta_{m,n} \triangleq \Phi_{m,m} + \Phi_{n,n} - \Phi_{m,n} - \Phi_{n,m} \geq 0 \tag{6.133}$$

If the agents k and ℓ are from the same cluster such that $q = r$, then hypothesis $\mathbb{H}_0$ in (6.29) is true and $d^{\star}_{q,r} = 0$; otherwise, hypothesis $\mathbb{H}_1$ in (6.29) is true and $d^{\star}_{q,r} \neq 0$. The hypothesis test for clustering becomes to test whether or not the difference $\boldsymbol{d}_{k,\ell}$ in (6.130) is zero mean *without* knowing its covariance matrix $\mu_{\max}\Delta_{m,n}$. If N_{sam} independent samples of $\boldsymbol{d}_{k,\ell}$ are available for testing, where $N_{\text{sam}} > M$, and $\Delta_{m,n}$ is non-singular, then according to the Neyman-Pearson criterion [123], the likelihood ratio test is given by [122, p. 164]

$$\boldsymbol{T}^2_{k,\ell} \triangleq N_{\text{sam}}\bar{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{S}^{-1}\bar{\boldsymbol{x}} \underset{\mathbb{H}_1}{\overset{\mathbb{H}_0}{\lessgtr}} \theta_{k,\ell} \tag{6.134}$$

where $\boldsymbol{T}^2_{k,\ell}$ is called Hotelling's T-square statistic, $\bar{\boldsymbol{x}}$ is the sample mean of $\boldsymbol{d}_{k,\ell}$, $\boldsymbol{S}$ is the unbiased sample covariance matrix, and $\theta_{k,\ell}$ is the predefined threshold from (6.29). The scaled T-square statistics $\frac{N_{\text{sam}}-M}{(N_{\text{sam}}-1)M} \cdot \boldsymbol{T}^2_{k,\ell}$ has a non-central F-distribution with M and $N_{\text{sam}} - M$ degrees of freedom and non-centrality parameter $N_{\text{sam}}\mu^{-1}_{\max}(d^{\star}_{q,r})^{\mathsf{T}}\Delta^{-1}_{m,n}d^{\star}_{q,r}$ [124, p. 480]. When $d^{\star}_{q,r} = 0$, it reduces to a central F-distribution [124, p. 322].

However, because stochastic iterative algorithms employ very small step-sizes, sampling their steady-state estimators over time does not produce independent samples. In many scenarios we only have one sample available for testing, where the sample mean reduces to the sample itself, and the sample covariance matrix is not even available. In order to carry out the hypothesis test, we replace the sample covariance matrix by the identity matrix. Then, the Hotelling's T-square test (6.134) becomes

$$\boldsymbol{\delta}_{k,\ell}^2 \triangleq \|\boldsymbol{d}_{k,\ell}\|^2 \underset{\mathbb{H}_1}{\overset{\mathbb{H}_0}{\lessgtr}} \theta_{k,\ell} \tag{6.135}$$

where we re-used $\boldsymbol{d}_{k,\ell}$ to denote the only available sample for testing. The decision statistic $\boldsymbol{\delta}_{k,\ell}^2$ is a quadratic form of the (near) Gaussian random vector $\boldsymbol{d}_{k,\ell}$. Using (6.131), the mean of $\boldsymbol{\delta}_{k,\ell}^2$ is given by

$$\mathbb{E}\boldsymbol{\delta}_{k,\ell}^2 = \mathbb{E}\|\boldsymbol{d}_{k,\ell}\|^2 = \mathbb{E}\mathrm{Tr}(\boldsymbol{d}_{k,\ell}\boldsymbol{d}_{k,\ell}^{\mathsf{T}}) = \mathrm{Tr}(\mathbb{E}\boldsymbol{d}_{k,\ell}\boldsymbol{d}_{k,\ell}^{\mathsf{T}}) = \|d_{q,r}^{\star}\|^2 + \mu_{\max}\mathrm{Tr}(\Delta_{m,n}) \tag{6.136}$$

and the variance of $\boldsymbol{\delta}_{k,\ell}^2$ is given by (see Appendix 6.F)

$$\mathrm{Var}(\boldsymbol{\delta}_{k,\ell}^2) = \mathbb{E}\|\boldsymbol{d}_{k,\ell}\|^4 - (\mathbb{E}\|\boldsymbol{d}_{k,\ell}\|^2)^2 = 4\mu_{\max}\|d_{q,r}^{\star}\|_{\Delta_{m,n}}^2 + 2\mu_{\max}^2\mathrm{Tr}(\Delta_{m,n}^2) \tag{6.137}$$

It is seen that the mean of $\boldsymbol{\delta}_{k,\ell}^2$ is dominated by $\|d_{q,r}^{\star}\|^2$ for sufficiently small step sizes. Since the variance of $\boldsymbol{\delta}_{k,\ell}^2$ is in the order of $\mu_{\max}$, according to Chebyshev's inequality [121, p. 47], we have

$$\mathbb{P}[|\boldsymbol{\delta}_{k,\ell}^2 - \mathbb{E}\boldsymbol{\delta}_{k,\ell}^2| \geq c] \leq \frac{\mathrm{Var}(\boldsymbol{\delta}_{k,\ell}^2)}{c} = O(\mu_{\max}) \tag{6.138}$$

for any constant $c > 0$. Therefore, for sufficiently small step sizes, the probability mass of $\boldsymbol{\delta}_{k,\ell}^2$ will highly concentrate around $\mathbb{E}\boldsymbol{\delta}_{k,\ell}^2$. When hypothesis $\mathbb{H}_0$ is true, we have $d_{q,r}^{\star} = 0$ and $\mathbb{E}\boldsymbol{\delta}_{k,\ell}^2 = \mu_{\max}\mathrm{Tr}(\Delta_{m,n}) = O(\mu_{\max}) \approx 0$; when hypothesis $\mathbb{H}_1$ is true, we have $d_{q,r}^{\star} \neq 0$ and $\mathbb{E}\boldsymbol{\delta}_{k,\ell}^2 = \|d_{q,r}^{\star}\|^2 + O(\mu_{\max}) \approx \|d_{q,r}^{\star}\|^2$. That is, the probability mass of $\boldsymbol{\delta}_{k,\ell}^2$ under $\mathbb{H}_0$ concentrates near 0 while the probability

mass of $\boldsymbol{\delta}_{k,\ell}^2$ under $\mathbb{H}_1$ concentrates near $\|d_{q,r}^\star\|^2 = \|w_q^\star - w_r^\star\|^2 > 0$ (which is a constant that is independent of $\mu_{\max}$). Obviously, the threshold $\theta_{k,\ell}$ should be chosen between 0 and $\|d_{q,r}^\star\|^2$. By doing so, the Type-I error will correspond to the right tail probability of $\boldsymbol{\delta}_{k,\ell}^2$ when $d_{q,r}^\star = 0$ (see (6.142) further ahead) and the Type-II error will correspond to the left tail probability of $\boldsymbol{\delta}_{k,\ell}^2$ when $d_{q,r}^\star \neq 0$ (see (6.143) further ahead).

In order to examine the statistical properties of $\boldsymbol{\delta}_{k,\ell}^2$ and to perform the analysis for error probabilities, let us introduce the eigen-decomposition of $\Delta_{m,n}$ in (6.133) and denote it by

$$\Delta_{m,n} = U_\Delta \Lambda_\Delta U_\Delta^{\mathsf{T}} \tag{6.139}$$

where U_Δ is orthonormal and Λ_Δ is diagonal and nonnegative. Let further

$$\boldsymbol{x} \triangleq \Lambda_\Delta^{-1/2} U_\Delta^{\mathsf{T}} \boldsymbol{d}_{k,\ell}, \quad \bar{x} \triangleq \Lambda_\Delta^{-1/2} U_\Delta^{\mathsf{T}} d_{q,r}^\star \tag{6.140}$$

Since $\boldsymbol{d}_{k,\ell} \sim \mathbb{N}(d_{q,r}^\star, \mu_{\max}\Delta_{m,n})$, it follows from (6.139) and (6.140) that $\boldsymbol{x} \sim \mathbb{N}(\bar{x}, \mu_{\max} I_M)$. Substituting (6.139) and (6.140) into (6.135) yields

$$\boldsymbol{\delta}_{k,\ell}^2 = \boldsymbol{x}^{\mathsf{T}} \Lambda_\Delta \boldsymbol{x} = \sum_{h=1}^{M} \lambda_h \boldsymbol{x}_h^2 \tag{6.141}$$

where $\boldsymbol{x}_h$ denotes the h-th elements of $\boldsymbol{x}$, and λ_h denotes the h-th element on the diagonal of Λ_Δ. From (6.141), it is obvious that $\boldsymbol{\delta}_{k,\ell}^2$ is a weighted sum of independent squared Gaussian random variables. When hypothesis $\mathbb{H}_0$ is true, we have $d_{q,r}^\star = 0$ and $\bar{x} = 0$ by (6.140). In this case, $\boldsymbol{\delta}_{k,\ell}^2$ reduces to a weighted sum of independent Gamma random variables (because squared zero-mean Gaussian random variables follow Gamma distributions [125, p. 337]), whose pdf is available in closed-form (but is very complicated) [126, 127]. When hypothesis $\mathbb{H}_1$ is true and $\|d_{q,r}^\star\|^2 > 0$, the pdf of $\boldsymbol{\delta}_{k,\ell}^2$ is generally not available in closed-form. Several procedures have been proposed in [128–132] for numerical evaluation of its

tail probability. Instead of relying on the precise pdf of $\boldsymbol{\delta}_{k,\ell}^2$, we shall provide some useful constructions in the sequel for the error probabilities in the hypothesis test problem (6.135).

6.4.3 Error Probabilities

For any $k \in \mathcal{G}_m \subseteq \mathcal{C}_q$ and $\ell \in \mathcal{G}_n \subseteq \mathcal{C}_r$, the Type-I error, namely, the false alarm for incorrect rejection of a true $\mathbb{H}_0$, is given by

$$\text{Type-I error}: \qquad \mathbb{P}[\boldsymbol{\delta}_{k,\ell}^2 > \theta_{k,\ell} | d_{q,r}^\star = 0] \tag{6.142}$$

and the Type-II error, namely, the missing detection for incorrect rejection of a true $\mathbb{H}_1$, is given by

$$\text{Type-II error}: \qquad \mathbb{P}[\boldsymbol{\delta}_{k,\ell}^2 < \theta_{k,\ell} | d_{q,r}^\star \neq 0] \tag{6.143}$$

It is seen that the Type-I error corresponds to the right tail probability of $\boldsymbol{\delta}_{k,\ell}^2$ with $d_{q,r}^\star = 0$ and the Type-II error corresponds to the left tail probability of $\boldsymbol{\delta}_{k,\ell}^2$ with $d_{q,r}^\star \neq 0$. This is a fundamental difference between the two types of errors and, therefore, different techniques are needed to approximate them. Specifically, for the Type-II error, the pdf of $\boldsymbol{\delta}_{k,\ell}^2$ is close to a bell shape and can be well approximated by a Gaussian pdf. Then, the Type-II error probability can be bounded by using Chernoff bound [133]. However, this technique does not apply to the Type-I error because when $d_{q,r}^\star = 0$, the pdf of $\boldsymbol{\delta}_{k,\ell}^2$ concentrates on the positive side of the origin point and is skewed with a long right tail. Consequently, we need to take a different approach to bound the Type-I error probability.

6.4.3.1 Type-I Error

We first note that

$$\boldsymbol{\delta}_{k,\ell}^2 = \boldsymbol{x}^\mathsf{T} \Lambda_\Delta \boldsymbol{x} \leq \|\Delta_{m,n}\| \cdot \|\boldsymbol{x}\|^2 \tag{6.144}$$

where Λ_Δ is from (6.139). This means that if $\boldsymbol{\delta}_{k,\ell}^2 > \theta_{k,\ell}$, then $\|\Delta_{m,n}\| \cdot \|\boldsymbol{x}\|^2 > \theta_{k,\ell}$ must be true, which further implies that the event $\{\boldsymbol{\delta}_{k,\ell}^2 > \theta_{k,\ell}\}$ is a subset of the event $\{\|\Delta_{m,n}\| \cdot \|\boldsymbol{x}\|^2 > \theta_{k,\ell}\}$. Therefore,

$$\mathbb{P}[\boldsymbol{\delta}_{k,\ell}^2 > \theta_{k,\ell} | d_{q,r}^\star = 0] \leq \mathbb{P}[\|\boldsymbol{x}\|^2 > \theta'_{k,\ell} | \bar{x} = 0] \tag{6.145}$$

where $\bar{x}$ is from (6.140), and

$$\theta'_{k,\ell} \triangleq \frac{\theta_{k,\ell}}{\|\Delta_{m,n}\|} \tag{6.146}$$

Since $\bar{x} = 0$, $\mu_{\max}^{-1}\|\boldsymbol{x}\|^2$ follows a central chi-square distribution with M degrees of freedom [125, p. 415]. Therefore, using the Chernoff bound for the central chi-square distribution [134, Lemma 1, p. 2500], we get from (6.145) that

$$\mathbb{P}[\boldsymbol{\delta}_{k,\ell}^2 > \theta_{k,\ell} | d_{q,r}^\star = 0] \leq 1 - \mathbb{P}[\|\boldsymbol{x}\|^2 \leq \theta'_{k,\ell} | \bar{x} = 0] \leq \left(\frac{\theta'_{k,\ell} e}{\mu_{\max} M}\right)^{M/2} \exp\left(-\frac{\theta'_{k,\ell}}{2\mu_{\max}}\right) \tag{6.147}$$

for $\mu_{\max} < \theta'_{k,\ell}/M$, where e is Euler's number. Therefore, when $\mu_{\max}$ is small enough, the Type-I error probability decays exponentially at a rate of $O(e^{-c_1/\mu_{\max}})$ for some constant $c_1 > 0$.

6.4.3.2 Type-II Error

We consider the characteristic function of $\boldsymbol{\delta}_{k,\ell}^2$. Since $\{\boldsymbol{x}_h\}$ are mutually independent, the characteristic function of $\boldsymbol{\delta}_{k,\ell}^2$ is given by

$$c_{\boldsymbol{\delta}_{k,\ell}^2}(t) \triangleq \mathbb{E}\left[e^{jt\boldsymbol{\delta}_{k,\ell}^2}\right] = \mathbb{E}\left[e^{jt\sum_{h=1}^M \lambda_h \boldsymbol{x}_h^2}\right] = \prod_{h=1}^M \mathbb{E}\left[e^{jt\lambda_h \boldsymbol{x}_h^2}\right] \tag{6.148}$$

where we used (6.144). Since $d_{q,r}^\star \neq 0$ in this case, $\boldsymbol{x}$ from (6.140) has nonzero mean $\bar{x} \neq 0$. Therefore, each $\mu_{\max}^{-1}\boldsymbol{x}_h^2$ is a non-central chi-square random variable with one degree of freedom and non-centrality $\mu_{\max}^{-1}\bar{x}_h^2$ [124, p. 433]. The characteristic function of $\boldsymbol{x}_h^2$ is then given by [124, p. 437]:

$$\mathbb{E}\left[e^{jt\boldsymbol{x}_h^2}\right] = \frac{1}{\sqrt{1 - 2jt\mu_{\max}}} \cdot e^{j\bar{x}_h^2 t/(1-2jt\mu_{\max})} \tag{6.149}$$

Substituting (6.149) into (6.148) yields:

$$c_{\boldsymbol{\delta}_{k,\ell}^2}(t) = \prod_{h=1}^{M} \frac{1}{\sqrt{1-2jt\mu_{\max}\lambda_h}} \cdot e^{j\bar{x}_h^2 t\lambda_h/(1-2jt\mu_{\max}\lambda_h)} \tag{6.150}$$

When $\mu_{\max}$ is sufficiently small, we have

$$\frac{1}{\sqrt{1-2jt\mu_{\max}\lambda_h}} \approx 1, \qquad \frac{1}{1-2jt\mu_{\max}\lambda_h} \approx 1+2jt\mu_{\max}\lambda_h \tag{6.151}$$

Using (6.153), we can approximate $c_{\boldsymbol{\delta}_{k,\ell}^2}(t)$ in (6.150) by

$$\begin{aligned} c_{\boldsymbol{\delta}_{k,\ell}^2}(t) &\approx \prod_{h=1}^{M} e^{j\bar{x}_h^2 t\lambda_h(1+2jt\mu_{\max}\lambda_h)} \\ &= e^{jt(\sum_{h=1}^{M}\lambda_h\bar{x}_h^2)-2t^2\mu_{\max}(\sum_{h=1}^{M}\lambda_h^2\bar{x}_h^2)} \\ &= e^{jt\|d_{q,r}^\star\|^2-2t^2\mu_{\max}\|d_{q,r}^\star\|_{\Lambda_\Delta}^2} \end{aligned} \tag{6.152}$$

where we used the fact that

$$\sum_{h=1}^{M}\lambda_h\bar{x}_h^2 = \|d_{q,r}^\star\|^2, \quad \sum_{h=1}^{M}\lambda_h^2\bar{x}_h^2 = \|d_{q,r}^\star\|_{\Lambda_\Delta}^2 \tag{6.153}$$

Note that the RHS of (6.152) coincides with the characteristic function of a Gaussian distribution with mean $\|d_{q,r}^\star\|^2$ and variance $4\mu_{\max}\|d_{q,r}^\star\|_{\Lambda_\Delta}^2$ [125, p. 89]. Since the distribution of a random variable is uniquely determined by its characteristic function, result (6.152) implies that $\boldsymbol{\delta}_{k,\ell}^2 \sim \mathbb{N}(\|d_{q,r}^\star\|^2, 4\mu_{\max}\|d_{q,r}^\star\|_{\Lambda_\Delta}^2)$ approximately for sufficiently small $\mu_{\max}$. Thus,

$$\mathbb{P}[\boldsymbol{\delta}_{k,\ell}^2 < \theta_{k,\ell} | d_{q,r}^\star \neq 0] \approx Q\left(\frac{\|d_{q,r}^\star\|^2-\theta_{k,\ell}}{2\mu_{\max}^{1/2}\|d_{q,r}^\star\|_{\Lambda_\Delta}}\right) \leq \frac{1}{2}e^{-(\|d_{q,r}^\star\|^2-\theta_{k,\ell})^2/8\mu_{\max}\|d_{q,r}^\star\|_{\Lambda_\Delta}^2} \tag{6.154}$$

where $Q(\cdot)$ denotes the Q-function, which is the tail probability of the standard Gaussian distribution, and the last step is by using the Chernoff bound [133, p. 380]. Therefore, when $\mu_{\max}$ is small enough, the Type-II error decays exponentially at a rate of $O(e^{-c_2/\mu_{\max}})$ for some constant $c_2 > 0$.

6.4.3.3 A Special Case

For the purpose of illustration only, we consider a special case where $\Delta_{m,n} = \sigma_{m,n}^2 I_M$. In this case, the pdf of $\boldsymbol{\delta}_{k,\ell}^2$ has a closed-form pdf. When $\mathbb{H}_1$ is true and $\|d_{q,r}^\star\|^2 > 0$, the quadratic form $\boldsymbol{\delta}_{k,\ell}^2/(\mu_{\max}\sigma_{m,n}^2)$ reduces to a non-central chi-square random variable with M degrees of freedom and non-centrality parameter $\|d_{q,r}^\star\|^2/\mu_{\max}\sigma_{m,n}^2$ [124, p. 433]. Let us denote the non-central chi-square distribution with d degrees of freedom and non-centrality parameter λ by $\chi_d^2(\lambda)$. The pdf of $\chi_d^2(\lambda)$ is then given by [124, p. 433]:

$$f_{\chi^2}(x; d, \lambda) = \frac{1}{2}\left(\frac{x}{\lambda}\right)^{(d-2)/4} e^{-(x+\lambda)/2} I_{(d-2)/2}(\sqrt{\lambda x}) \tag{6.155}$$

for $x \geq 0$, where $I_h(x)$ denotes the h-th order modified Bessel function of the first kind. Then,

$$\frac{\boldsymbol{\delta}_{k,\ell}^2}{\mu_{\max}\sigma_{m,n}^2} \sim \chi_M^2\left(\frac{\|d_{q,r}^\star\|^2}{\mu_{\max}\sigma_{m,n}^2}\right) \tag{6.156}$$

and the pdf of $\boldsymbol{\delta}_{k,\ell}^2$ is given by

$$f(z) = \frac{1}{\mu_{\max}\sigma_{m,n}^2} \cdot f_{\chi^2}\left(\frac{z}{\mu_{\max}\sigma_{m,n}^2}; M, \frac{\|d_{q,r}^\star\|^2}{\mu_{\max}\sigma_{m,n}^2}\right) \tag{6.157}$$

where $f_{\chi^2}(\cdot)$ is from (6.155). When $\mathbb{H}_0$ is true and $\|d_{q,r}^\star\|^2 = 0$, the pdf $f(z)$ in (6.157) reduces to a scaled central chi-square distribution [125, p. 415]:

$$f(z) = \frac{1}{\mu_{\max}\sigma_{m,n}^2} \cdot f_{\chi^2}\left(\frac{z}{\mu_{\max}\sigma_{m,n}^2}; M, 0\right) \tag{6.158}$$

We plot the pdf $f(z)$ from (6.157) and (6.158) in Fig. 6.1. It can be observed that when M, $\|d_{q,r}^\star\|^2$, and $\sigma_{m,n}^2$ are fixed, in both $\mathbb{H}_0$ (blue curves) and $\mathbb{H}_1$ (red curves) cases, the probability mass of $\boldsymbol{\delta}_{k,\ell}^2$ concentrates more around its mean as $\mu_{\max}$ decreases. When $q \neq r$ (i.e., $\mathbb{H}_1$ is true), the mean of $\boldsymbol{\delta}_{k,\ell}$ is close to $\|d_{q,r}^\star\|^2 = 1$ for sufficiently small $\mu_{\max}$; when $q = r$ (i.e., $\mathbb{H}_0$ is true), the mean is close to zero. The right tail probabilities of the blue curves (under $\mathbb{H}_0$) and the left tail probabilities

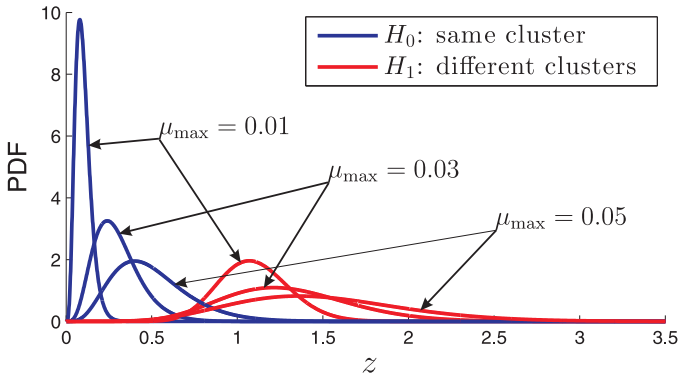


Figure 6.1: The pdf of $\boldsymbol{\delta}_{k,\ell}^2$ defined in (6.157) and (6.158) with $M = 10$, $\|d_{q,r}^{\star}\|^2 = 1$, $\sigma_{m,n}^2 = 1$, $\mu_{\max} = 0.01, 0.03, 0.05$.

of the red curves (under $\mathbb{H}_1$) all decay exponentially. In addition, it is seen that the pdf of $\boldsymbol{\delta}_{k,\ell}^2$ under $\mathbb{H}_1$ (the red curves with $\|d_{q,r}^{\star}\|^2 > 0$) is near symmetric and is in bell-shape, which agrees with the Gaussian approximation we made when evaluating the Type-II error (mis-detection) for the general case. On the other hand, the pdf of $\boldsymbol{\delta}_{k,\ell}^2$ under $\mathbb{H}_0$ (the blue curves with $\|d_{q,r}^{\star}\|^2 = 0$) concentrates close to zero and has large skewness with a long tail on the RHS, which distinguishes itself from Gaussian distributions; this demonstrates our previous statement that it is not appropriate to assess the Type-I error (false alarm) by approximating the pdf of $\boldsymbol{\delta}_{k,\ell}^2$ under $\mathbb{H}_0$ with Gaussian distributions.

6.4.4 Dynamics of Diffusion with Adaptive Clustering

Since both Type-I and Type-II errors decay exponentially with exponent proportional to $1/\mu_{\max}$, it is expected that incorrect clustering decisions will become rare as the iteration proceeds. We can therefore assume that enough iterations have elapsed and the first recursion (6.20a)–(6.20b) is operating in steady-state. Under these conditions, we can examine the dynamics of the second recursion (6.31a)–(6.31b) with adaptive clustering.

From Assumption 6.1, correct clustering decisions split the underlying topology into Q sub-networks one for each cluster. Within each cluster, correct clustering decisions merge all disjoint groups into a bigger group. Therefore, the resulting topology for the entire network will now consist of Q separate sub-networks and each sub-network will be strongly-connected. In addition, since the step-sizes are sufficiently small, the decision statistics $\|\boldsymbol{w}_{\ell,i}-\boldsymbol{w}_{k,i}\|^2$ generated by the first recursion (6.20a)–(6.20b) in steady-state will be nearly time-invariant. The clustering decisions will therefore also be nearly time-invariant. Then, with high probability, the cooperative sub-neighborhoods $\{\boldsymbol{\mathcal{N}}_{k,i}^+\}$ produced by (6.30) will become nearly time-invariant after the first recursion (6.20a)–(6.20b) reaches steady-state:

$$\boldsymbol{\mathcal{N}}_{k,i}^+ \to \mathcal{N}_k^+, \quad \text{as} \quad i \to \infty \tag{6.159}$$

for all k, where $\mathcal{N}_k^+$ is from (1.38) from Chapter 1.

In order to gain from enhanced cooperation via adaptive clustering, it is critical to choose proper combination policies for recursion (6.31a)–(6.31b). From the discussion in Chapter 12 of [66, p. 624-635], we know that doubly-stochastic combination policies are able to exploit the benefit of cooperation when more agents are included in cooperation. For example, one can choose the Metropolis rule [66, p. 664], i.e.,

$$\boldsymbol{a}'_{\ell k}(i) = \begin{cases} \dfrac{1}{\max\{|\boldsymbol{\mathcal{N}}_{\ell,i}^+|, |\boldsymbol{\mathcal{N}}_{k,i}^+|\}}, & \ell \in \boldsymbol{\mathcal{N}}_{k,i}^+ \backslash \{k\} \\ 1 - \displaystyle\sum_{n \in \boldsymbol{\mathcal{N}}_{k,i}^+ \backslash \{k\}} \boldsymbol{a}'_{nk}(i), & \ell = k \\ 0, & \ell \in \mathcal{N}_k \backslash \boldsymbol{\mathcal{N}}_{k,i}^+ \end{cases} \tag{6.160}$$

When the combination coefficients $\{\boldsymbol{a}'_{\ell k}(i)\}$ are chosen according to (6.160), their values are determined by the size of their cooperative sub-neighborhood $\boldsymbol{\mathcal{N}}_{k,i}^+$. It

is then obvious that coefficients $\{\boldsymbol{a}'_{\ell k}(i)\}$ will tend to be constant values:

$$\boldsymbol{a}'_{\ell k}(i) \to a'_{\ell k}, \quad \text{as} \quad i \to \infty \tag{6.161}$$

which will be determined by the size of $\mathcal{N}_k^+$. Therefore, we can rewrite the second recursion (6.31a)–(6.31b) for small enough $\mu_{\max}$ and large enough i as

$$\boldsymbol{\psi}'_{k,i} = \boldsymbol{w}'_{k,i-1} - \mu_k \widehat{\nabla J_k}(\boldsymbol{w}'_{k,i-1}) \tag{6.162a}$$

$$\boldsymbol{w}'_{k,i} = \sum_{\ell \in \mathcal{N}_k^+} a'_{\ell k} \boldsymbol{\psi}'_{\ell,i} \tag{6.162b}$$

by using (6.159) and (6.161). We collect the $\{a'_{\ell k}\}$ into a matrix and denote it by A'. The matrix A' is block diagonal and each block on its diagonal corresponds to a cluster. Recursion (6.162a)–(6.162b) only involves in-cluster cooperative learning for common minimizers, where all agents from a cluster form a single big group. Therefore, the performance analysis in Section 6.3 applies to this case as well.

6.5 Simulation Results

We first simulate a network consisting of $N = 200$ agents. Each agent observes a data stream $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}; i \geq 0\}$ that satisfies the linear regression model [46]:

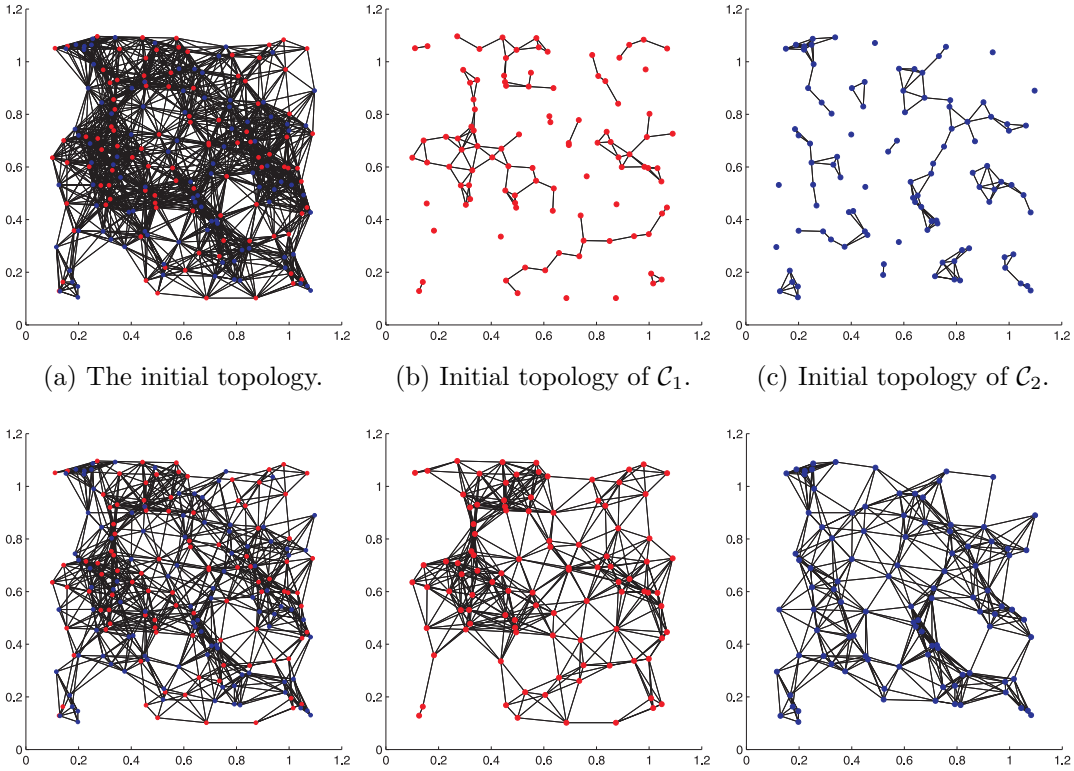
$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i} w_k^o + \boldsymbol{v}_k(i) \tag{6.163}$$

where $\boldsymbol{d}_k(i) \in \mathbb{R}$ is a scalar response variable and $\boldsymbol{u}_{k,i} \in \mathbb{R}^{1\times M}$ is a row vector feature variable with $M = 2$. The feature variable $\boldsymbol{u}_{k,i}$ is randomly generated at every iteration by using a Gaussian distribution with zero mean and scaled identity covariance matrix $\sigma_{u,k}^2 I_M$. The model noise $\boldsymbol{v}_k(i) \in \mathbb{R}$ is also randomly generated at every iteration by using another independent Gaussian distribution

with zero mean and variance $\sigma_{v,k}^2$. The values of $\{\sigma_{u,k}^2\}$ and $\{\sigma_{v,k}^2\}$ are positive and randomly generated.

There are $Q = 2$ clusters in the network. The first $N_1 = 100$ agents belong to cluster $\mathcal{C}_1$, i.e., $\mathcal{C}_1 = \{1, 2, \ldots, 100\}$. The second $N_2 = 100$ agents belong to cluster $\mathcal{C}_2$, i.e., $\mathcal{C}_2 = \{101, 102, \ldots, 200\}$. The loading factors for the two clusters, namely, $w_1^\star$ and $w_2^\star$, are randomly generated. The step-size is uniform and is set to $\mu = 0.05$. The underlying topology that connects all agents is shown in Fig. 6.2a. Agents from cluster $\mathcal{C}_1$ are in red and agents from $\mathcal{C}_2$ are in blue. We simulated the scenario where agents have some partial knowledge about the grouping at the beginning of the learning process. The partial knowledge is non-trivial, meaning that the groups $\{\mathcal{G}_m\}$ used in the first recursion (6.20a)–(6.20b) are not just singletons. The topologies that reflect the $\{\mathcal{G}_m\}$ are plotted in Figs. 6.2b and 6.2c for the two clusters. The Metropolis rule (6.160) is used in both recursions, (6.20a)–(6.20b) and (6.31a)–(6.31b).

As we explained before, in steady-state the clustering decisions become time-invariant and small groups in the same cluster merge into bigger groups. The links between neighbors within the same cluster are active while links to neighbors from different clusters are dropped. We plot the resulting topology in steady-state with active links in Fig. 6.2d. Compared to Fig. 6.2a, the underlying topology in Fig. 6.2d is trimmed and split into two disjoint sub-networks. This result implies that the interference between two clusters is suppressed. The two sub-networks are themselves connected at steady-state and are shown in Figs 6.2e and 6.2f. Comparing the resulting cluster topologies in Figs 6.2e and 6.2f with the initial cluster topologies in Figs. 6.2b and 6.2c, it can be observed that all separate small groups from the same cluster merge into a bigger group and collaborative learning involving more agents emerges.



(a) The initial topology. (b) Initial topology of $\mathcal{C}_1$. (c) Initial topology of $\mathcal{C}_2$.

(d) The final topology. (e) Resulting topology of $\mathcal{C}_1$. (f) Resulting topology of $\mathcal{C}_2$.

Figure 6.2: The underlying topology of the entire network where agents from different clusters are connected. As the learning process progresses, the disjoint groups in each cluster merge into a bigger group to enable collaborative learning among more agents. In steady-state, only in-cluster links remain active.

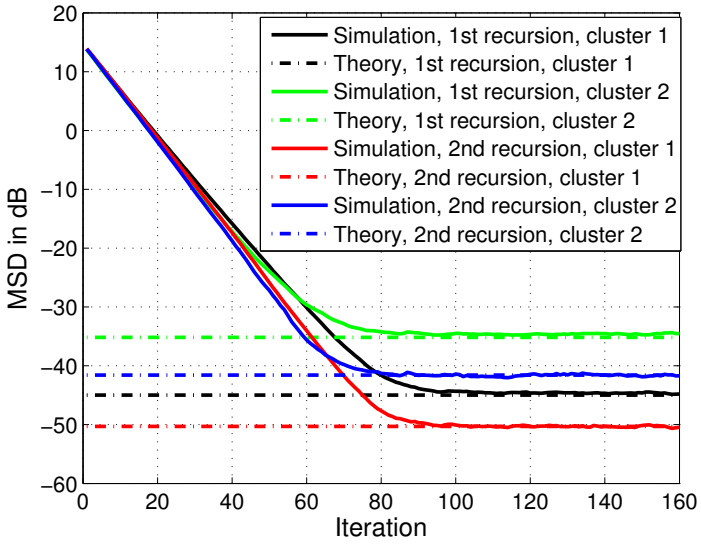
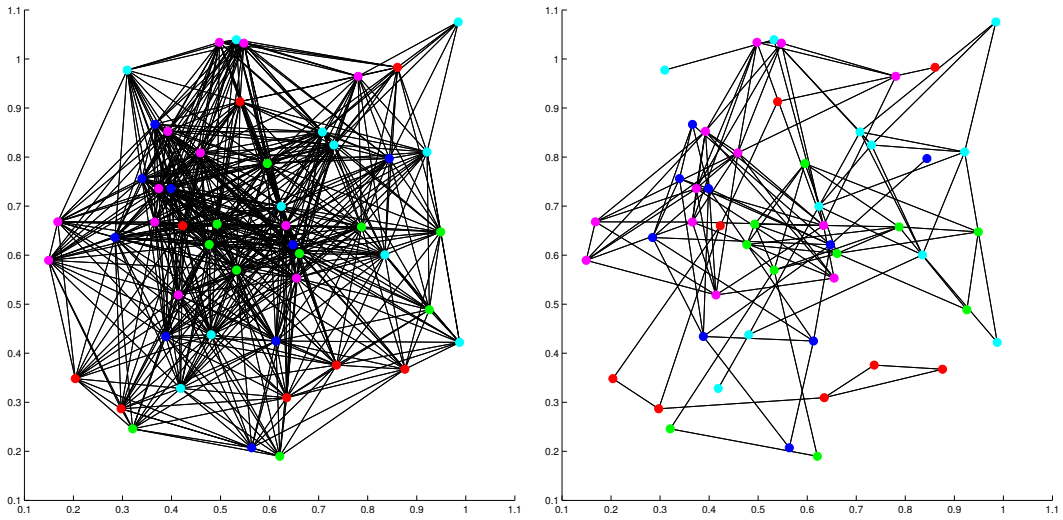


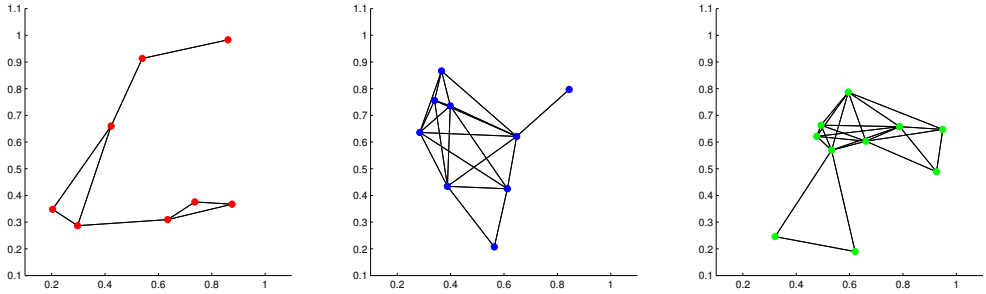
Figure 6.3: The steady-state cluster average MSDs for the first recursion (6.20a)–(6.20b) and the second recursion (6.31a)–(6.31b).

The MSD learning curves are plotted in Fig. 6.3 where the cluster MSDs are obtained by averaging over 100 trials. The cluster MSDs for the first recursion (6.20a)–(6.20b) are in black and green for clusters 1 and 2, respectively. The cluster MSDs for the second recursion (6.31a)–(6.31b) are in red and blue for clusters 1 and 2, respectively. Obviously both clusters improve their steady-state MSD performance on average by forming larger clusters for cooperation.

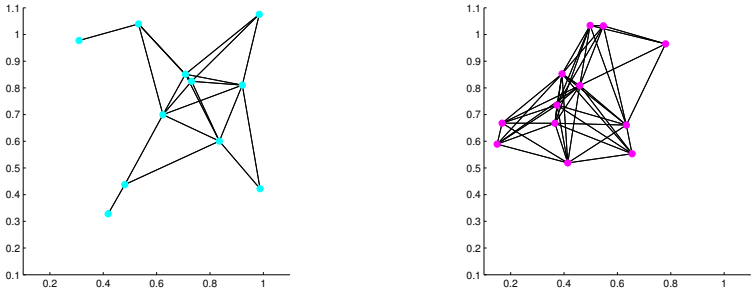
In the second simulation, we simulate a network with $N = 50$ nodes in $Q = 5$ clusters. The sizes of the five clusters are 8, 9, 10, 11, and 12, respectively. The initial topology is shown in Fig. 6.4a. We choose the uniform step-size $\mu = 0.01$. After 1000 iterations, the resulting topology is separated into five clusters and is shown in Fig. 6.4b, and the topologies for the five clusters are given in Figs. 6.4c–6.4g, respectively. The MSD learning curves that are obtained by averaging over 500 trials match the theory well, as shown in Figs. 6.5a and 6.5b.



(a) The initial topology with five clusters. (b) The remaining topology with five clusters.

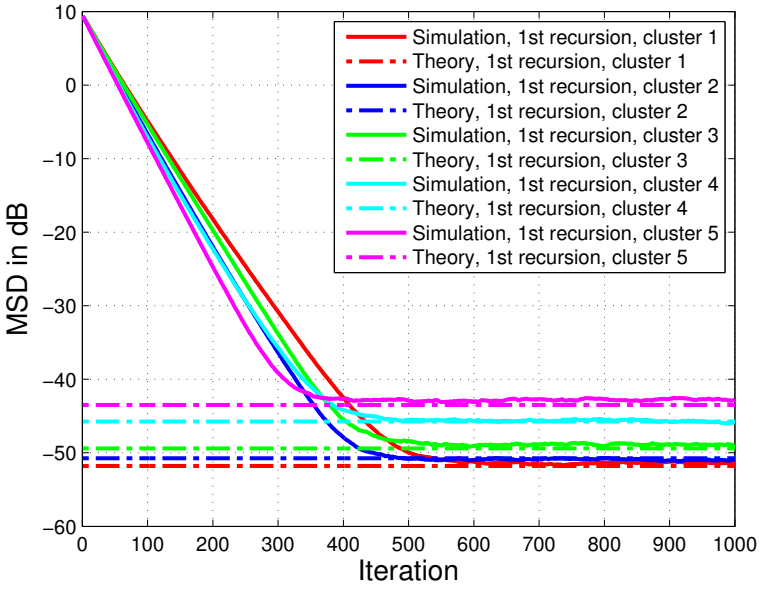


(c) Final topology of $\mathcal{C}_1$. (d) Final topology of $\mathcal{C}_2$. (e) Final topology of $\mathcal{C}_3$.

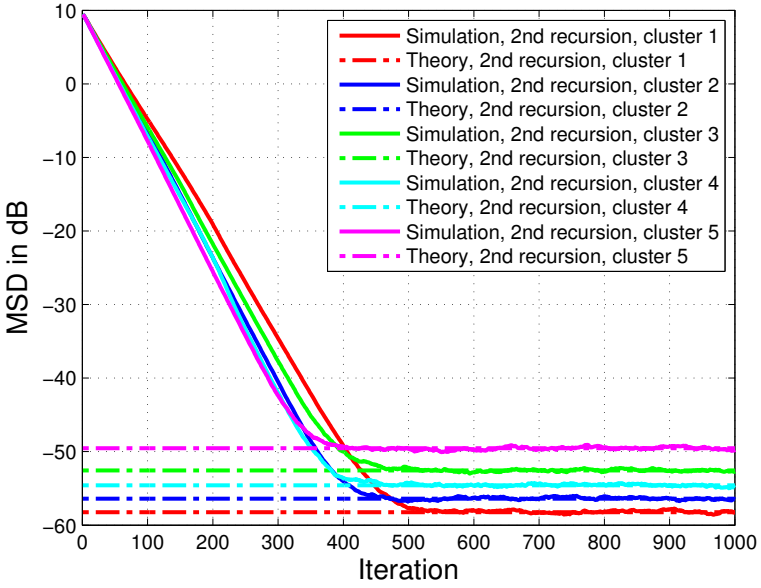


(f) Final topology of $\mathcal{C}_4$. (g) Final topology of $\mathcal{C}_5$.

Figure 6.4: The initial topology with $N = 50$ nodes and $Q = 5$ clusters. In steady-state, the five clusters are successfully separated from each other while each cluster remains connected.



(a) The MSD learning curves for the first recursion (6.20a)–(6.20b).



(b) The MSD learning curves for the second recursion (6.31a)–(6.31b).

Figure 6.5: The MSD learning curves for the proposed distributed clustering and learning algorithm.

6.6 Conclusions

In this chapter we proposed a distributed strategy for adaptive learning and clustering over multi-cluster networks. Detailed performance analysis is conducted and the results are supported by simulations. The proposed algorithm can be used in applications to segment heterogeneous networks into sub-networks to enhance in-cluster cooperation and suppress cross-cluster interference. It can also be applied to homogeneous networks to prevent intrusion or jamming by isolating malicious nodes from normal nodes. Furthermore, it can be used to trim and grow adaptive networks according to the objectives of the agents in the network.

6.A Proof of Lemma 6.2

Since both models, (6.86) and (6.69), can be decoupled into G separate recursions one for each group, it is sufficient to show that for sufficiently small step-sizes, and for any group $\mathcal{G}_m$, it holds that

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}} - \bar{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{ld}}\|^2 = O(\mu_{\max}^2) \tag{6.164}$$

where $\bar{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{ld}}$ is given by (6.91). We adopt a technique similar to the one used in the proof of Theorem 10.2 [66, p. 557] to establish (6.164) in the sequel. We introduce the Jordan decomposition of each A_m [66, 115]:

$$A_m = V_m J_m V_m^{-1} \triangleq \begin{bmatrix} p_m^g & V_{m,R} \end{bmatrix} \begin{bmatrix} 1 & \\ & J_{m,\epsilon} \end{bmatrix} \begin{bmatrix} \mathbb{1}_{N_m^g} & V_{m,L} \end{bmatrix}^{\mathsf{T}} \tag{6.165}$$

where $J_{m,\epsilon} \in \mathbb{C}^{(N_m^g-1)\times(N_m^g-1)}$ consists of all stable Jordan blocks with ϵ's on the first lower off-diagonal, and V_m is a non-singular complex matrix. Let

$$\mathcal{V}_m \triangleq V_m \otimes I_M \tag{6.166}$$

$$\mathcal{J}_m \triangleq J_m \otimes I_M \tag{6.167}$$

Multiplying $\mathcal{V}_m^{\mathsf{T}}$ to both sides of (6.73) yields:

$$\mathcal{V}_m^{\mathsf{T}}\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}} = \bar{\mathcal{B}}_m\mathcal{V}_m^{\mathsf{T}}\widetilde{\boldsymbol{\mathcal{W}}}_{m,i-1}^{\mathrm{lt}} + \mathcal{J}_m^{\mathsf{T}}\mathcal{V}_m^{\mathsf{T}}\mathcal{M}_m\boldsymbol{\mathcal{S}}_{m,i}(\boldsymbol{\mathcal{W}}_{m,i-1}) \tag{6.168}$$

where

$$\bar{\mathcal{B}}_m \triangleq \mathcal{V}_m^{\mathsf{T}}\mathcal{B}_m(\mathcal{V}_m^{\mathsf{T}})^{-1} = \mathcal{J}_m^{\mathsf{T}} - \mathcal{J}_m^{\mathsf{T}}\mathcal{V}_m^{\mathsf{T}}\mathcal{M}_m\mathcal{H}_m(\mathcal{V}_m^{\mathsf{T}})^{-1} \tag{6.169}$$

By (6.165) and (6.166), we have

$$\mathcal{V}_m^{\mathsf{T}}\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}} = \begin{bmatrix} (p_m^g \otimes I_M)^{\mathsf{T}}\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}} \\ (V_{m,R} \otimes I_M)^{\mathsf{T}}\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}} \end{bmatrix} \triangleq \begin{bmatrix} \bar{\boldsymbol{w}}_{m,i}^{\mathrm{lt}} \\ \check{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}} \end{bmatrix} \tag{6.170}$$

where $\bar{\boldsymbol{w}}_{m,i}^{\mathrm{lt}}$ is an $M \times 1$ vector, $\check{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}}$ is an $(N_m^g - 1)M \times 1$ vector. It follows from (6.166) and (6.91) that

$$\mathcal{V}_m^{\mathsf{T}}\bar{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{ld}} = (V_m^{\mathsf{T}}\mathbb{1}_{N_m^g}) \otimes \widetilde{\boldsymbol{w}}_{m,i}^{\mathrm{ld}} = \begin{bmatrix} \widetilde{\boldsymbol{w}}_{m,i}^{\mathrm{ld}} \\ 0 \end{bmatrix} \tag{6.171}$$

since $\mathbb{1}_{N_m^g}$ is the first column of $(V_m^{\mathsf{T}})^{-1}$ in (6.165). Using (6.170) and (6.171), we find that

$$\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}} - \bar{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{ld}}\|_{\Sigma_m}^2 = \mathbb{E}\|\bar{\boldsymbol{w}}_{m,i}^{\mathrm{lt}} - \widetilde{\boldsymbol{w}}_{m,i}^{\mathrm{ld}}\|^2 + \mathbb{E}\|\check{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}}\|^2 \tag{6.172}$$

where $\Sigma_m \triangleq \mathcal{V}_m\mathcal{V}_m^{\mathsf{T}}$ is a positive-definite weighting matrix. Since $\|\Sigma_m\|$ is independent of $\mu_{\max}$, result (6.164) holds if the following condition holds:

$$\limsup_{i\to\infty} \mathbb{E}\|\bar{\boldsymbol{w}}_{m,i}^{\mathrm{lt}} - \widetilde{\boldsymbol{w}}_{m,i}^{\mathrm{ld}}\|^2 + \mathbb{E}\|\check{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}}\|^2 = O(\mu_{\max}^2) \tag{6.173}$$

Using Eq. (10.78) in [66, p. 563], we know that

$$\limsup_{i\to\infty} \mathbb{E}\|\check{\boldsymbol{\mathcal{W}}}_{m,i}^{\mathrm{lt}}\|^2 = O(\mu_{\max}^2) \tag{6.174}$$

From (6.168) and (6.170), the evolution of $\bar{\boldsymbol{w}}_{m,i}^{\rm lt}$ is given by (see Eq. (9.61) from [66, p. 514] for a similar derivation):

$$\bar{\boldsymbol{w}}_{m,i}^{\rm lt} = D_m \bar{\boldsymbol{w}}_{m,i-1}^{\rm lt} - D_{21}^{\mathsf{T}} \check{\boldsymbol{w}}_{m,i-1}^{\rm lt} + (p_m^g \otimes I_M)^{\mathsf{T}} \mathcal{M}_m \boldsymbol{s}_{m,i}(\boldsymbol{\mathcal{W}}_{m,i-1}) \tag{6.175}$$

where $D_{21}^{\mathsf{T}} \triangleq (p_m^g \otimes I_M)^{\mathsf{T}} \mathcal{M}_m \mathcal{H}_m (V_{m,L} \otimes I_M)$. Using (6.175) and (6.83), we obtain

$$\bar{\boldsymbol{w}}_{m,i}^{\rm lt} - \widetilde{\boldsymbol{w}}_{m,i}^{\rm ld} = D_m(\bar{\boldsymbol{w}}_{m,i-1}^{\rm lt} - \widetilde{\boldsymbol{w}}_{m,i-1}^{\rm ld}) - D_{21}^{\mathsf{T}} \check{\boldsymbol{w}}_{m,i-1}^{\rm lt} \tag{6.176}$$

We recognize that recursion (6.176) has a form that is similar to the recursion for $\bar{\boldsymbol{b}}_i$ in Eq. (10.64) of [66, p. 561] except that here in (6.176) the driving noise term is absent. Therefore, we immediately get from Eq. (10.66) of [66, p. 562] that

$$\mathbb{E}\|\bar{\boldsymbol{w}}_{m,i}^{\rm lt} - \widetilde{\boldsymbol{w}}_{m,i}^{\rm ld}\|^2 \leq (1 - \sigma_{11}\mu_{\max})\mathbb{E}\|\bar{\boldsymbol{w}}_{m,i-1}^{\rm lt} - \widetilde{\boldsymbol{w}}_{m,i-1}^{\rm ld}\|^2 + \frac{\sigma_{21}^2 \mu_{\max}}{\sigma_{11}} \mathbb{E}\|\check{\boldsymbol{w}}_{m,i-1}^{\rm lt}\|^2 \tag{6.177}$$

for some constants $\sigma_{11} > 0$ and $\sigma_{21} > 0$. Substituting (6.174) into (6.177) yields

$$\mathbb{E}\|\bar{\boldsymbol{w}}_{m,i}^{\rm lt} - \widetilde{\boldsymbol{w}}_{m,i}^{\rm ld}\|^2 \leq (1 - \sigma_{11}\mu_{\max})\mathbb{E}\|\bar{\boldsymbol{w}}_{m,i-1}^{\rm lt} - \widetilde{\boldsymbol{w}}_{m,i-1}^{\rm ld}\|^2 + O(\mu_{\max}^3) \tag{6.178}$$

for large enough i. Therefore, it follows from (6.178) that

$$\limsup_{i \to \infty} \mathbb{E}\|\bar{\boldsymbol{w}}_{m,i}^{\rm lt} - \widetilde{\boldsymbol{w}}_{m,i}^{\rm ld}\|^2 = O(\mu_{\max}^2) \tag{6.179}$$

Combining (6.174) and (6.179) proves (6.173).

6.B Proof of Lemma 6.3

Let us examine the evolution of the covariance matrix of $\widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm ld}$, which is defined by

$$\Theta_i \triangleq \mathbb{E}[\widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm ld}(\widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm ld})^{\mathsf{T}}] \tag{6.180}$$

Using (6.11) and (6.12), we get from (6.86) that

$$\Theta_i = \mathcal{D}\Theta_{i-1}\mathcal{D} + \mathcal{P}^{\mathsf{T}}\mathcal{M}[\mathbb{E}\mathcal{R}_{s,i}(\boldsymbol{\mathcal{W}}_{i-1})]\mathcal{M}\mathcal{P} \tag{6.181}$$

We next introduce the fixed-point covariance recursion

$$\Theta_i^{\rm ss} = \mathcal{D}\Theta_{i-1}^{\rm ss}\mathcal{D} + \mathcal{P}^{\mathsf{T}}\mathcal{M}\mathcal{R}_{s,i}(\mathcal{W}^o)\mathcal{M}\mathcal{P} \tag{6.182}$$

Let

$$\Delta\Theta_i \triangleq \Theta_i - \Theta_i^{\rm ss}, \;\; \Delta\mathcal{R}_{s,i} \triangleq \mathbb{E}\mathcal{R}_{s,i}(\boldsymbol{\mathcal{W}}_{i-1}) - \mathcal{R}_{s,i}(\mathcal{W}^o) \tag{6.183}$$

The difference matrix $\Delta\Theta_i$ evolves by the following recursion:

$$\Delta\Theta_i = \mathcal{D}\Delta\Theta_{i-1}\mathcal{D} + \mathcal{P}^{\mathsf{T}}\mathcal{M}\Delta\mathcal{R}_{s,i}\mathcal{M}\mathcal{P} \tag{6.184}$$

We bound the difference matrix $\Delta\mathcal{R}_{s,i}$ by

$$\begin{aligned}
\|\Delta\mathcal{R}_{s,i}\| &\overset{(a)}{\leq} \mathbb{E}\|\mathcal{R}_{s,i}(\boldsymbol{\mathcal{W}}_{i-1}) - \mathcal{R}_{s,i}(\mathcal{W}^o)\| \\
&\overset{(b)}{\leq} \kappa_s\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}\|^{\gamma_s} \\
&\overset{(c)}{\leq} \kappa_s\left(\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}\|^4\right)^{\gamma_s/4}
\end{aligned} \tag{6.185}$$

where step (a) is by using Jensen's inequality; step (b) is by using (6.14) from Assumption 6.3; and step (c) is by applying Jensen's inequality again to the concave function $x^{\gamma_s/4}$ for $\gamma_s \leq 4$ and $x \geq 0$. As $i \to \infty$, we get from (6.185) that

$$\limsup_{i\to\infty} \|\Delta\mathcal{R}_{s,i}\| = O(\mu_{\max}^{\gamma_s/2}) \tag{6.186}$$

by using (6.66). From Eq. (9.286) in [66, p. 548], we have

$$\|\mathcal{D}\| = \max_m \|D_m\| \leq 1 - \sigma\mu_{\max} \tag{6.187}$$

for some $\sigma > 0$. Using the triangle inequality and the sub-multiplicativity property of norms, we have from (6.184) that

$$\begin{aligned}
\|\Delta\Theta_i\| &\leq \|\mathcal{D}\Delta\Theta_{i-1}\mathcal{D}\| + \|\mathcal{P}^{\mathsf{T}}\mathcal{M}\Delta\mathcal{R}_{s,i}\mathcal{M}\mathcal{P}\| \\
&\leq \|\mathcal{D}\|^2\|\Delta\Theta_{i-1}\| + \mu_{\max}^2\|\mathcal{P}\|^2\|\Delta\mathcal{R}_{s,i}\|
\end{aligned}$$

$$\leq (1-\sigma\mu_{\max})\|\Delta\Theta_{i-1}\| + \mu_{\max}^2\|\mathcal{P}\|^2\|\Delta\mathcal{R}_{s,i}\| \tag{6.188}$$

where in the last step we used (6.187) and the fact that $0 < 1-\sigma\mu_{\max} < 1$. Then, as $i \to \infty$, we get from (6.186) and (6.188) that

$$\limsup_{i\to\infty} \|\Delta\Theta_i\| \leq \sigma^{-1}\mu_{\max}\|\mathcal{P}\|^2(\limsup_{i\to\infty} \|\Delta\mathcal{R}_{s,i}\|) = O(\mu_{\max}^{1+\gamma_s/2}) \tag{6.189}$$

Now, since $\mathcal{D}$ is stable and in view of (6.15), the fixed-point recursion (6.182) converges as $i \to \infty$. At steady-state, the limit $\Theta_\infty^{\rm ss} \triangleq \lim_{i\to\infty} \Theta_i^{\rm ss}$ of (6.182) satisfies the discrete Lyapunov equation (6.94) by identifying $\Theta \equiv \Theta_\infty^{\rm ss}$.

6.C Proof of Theorem 6.2

From Lemmas 6.1 and 6.2,

$$\begin{aligned}
&\lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}\|^2 \\
&\leq \lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i - \widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm lt} + \widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm lt} - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}\|^2 \\
&\leq \lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} 2\mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i - \widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm lt}\|^2 + \lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} 2\mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i^{\rm lt} - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}\|^2 \\
&= 0
\end{aligned} \tag{6.190}$$

Let

$$\Pi_i^{\rm ld} \triangleq \mu_{\max}^{-1}\mathbb{E}\bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}(\bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld})^{\mathsf{T}} \tag{6.191}$$

Then, by Jensen's inequality,

$$\begin{aligned}
\mu_{\max}\|\Pi_i - \Pi_i^{\rm ld}\| &\leq \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathsf{T}} - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}(\bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld})^{\mathsf{T}}\| \\
&= \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathsf{T}} - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathsf{T}} + \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathsf{T}} - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}(\bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld})^{\mathsf{T}}\| \\
&\leq \mathbb{E}\|(\widetilde{\boldsymbol{\mathcal{W}}}_i - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld})\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathsf{T}}\| + \mathbb{E}\|\bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}(\widetilde{\boldsymbol{\mathcal{W}}}_i - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld})^{\mathsf{T}}\|
\end{aligned} \tag{6.192}$$

The second term on the RHS of (6.192) can be bounded by

$$\mathbb{E}\|\bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld}(\widetilde{\boldsymbol{\mathcal{W}}}_i - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld})^{\mathsf{T}}\| = \mathbb{E}\|(\bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld} - \widetilde{\boldsymbol{\mathcal{W}}}_i + \widetilde{\boldsymbol{\mathcal{W}}}_i)(\widetilde{\boldsymbol{\mathcal{W}}}_i - \bar{\boldsymbol{\mathcal{W}}}_i^{\rm ld})^{\mathsf{T}}\|$$

$$\leq \mathbb{E}\|(\bar{\boldsymbol{w}}_i^{\mathrm{ld}} - \widetilde{\boldsymbol{w}}_i)(\widetilde{\boldsymbol{w}}_i - \bar{\boldsymbol{w}}_i^{\mathrm{ld}})^{\mathsf{T}}\| + \mathbb{E}\|\widetilde{\boldsymbol{w}}_i(\widetilde{\boldsymbol{w}}_i - \bar{\boldsymbol{w}}_i^{\mathrm{ld}})^{\mathsf{T}}\|$$
$$= \mathbb{E}\|\bar{\boldsymbol{w}}_i^{\mathrm{ld}} - \widetilde{\boldsymbol{w}}_i\|^2 + \mathbb{E}\|\widetilde{\boldsymbol{w}}_i(\widetilde{\boldsymbol{w}}_i - \bar{\boldsymbol{w}}_i^{\mathrm{ld}})^{\mathsf{T}}\| \tag{6.193}$$

Substituting (6.193) into (6.192) yields:

$$\mu_{\max}\|\Pi_i - \Pi_i^{\mathrm{ld}}\| \leq 2\mathbb{E}\|(\widetilde{\boldsymbol{w}}_i - \bar{\boldsymbol{w}}_i^{\mathrm{ld}})\widetilde{\boldsymbol{w}}_i^{\mathsf{T}}\| + \mathbb{E}\|\bar{\boldsymbol{w}}_i^{\mathrm{ld}} - \widetilde{\boldsymbol{w}}_i\|^2 \tag{6.194}$$

The first term on the RHS of (6.194) can be bounded by

$$\mathbb{E}\|(\widetilde{\boldsymbol{w}}_i - \bar{\boldsymbol{w}}_i^{\mathrm{ld}})\widetilde{\boldsymbol{w}}_i^{\mathsf{T}}\| \leq \mathbb{E}(\|\widetilde{\boldsymbol{w}}_i - \bar{\boldsymbol{w}}_i^{\mathrm{ld}}\|\|\widetilde{\boldsymbol{w}}_i\|)$$
$$\leq \sqrt{\mathbb{E}\|\widetilde{\boldsymbol{w}}_i - \bar{\boldsymbol{w}}_i^{\mathrm{ld}}\|^2\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2} \tag{6.195}$$

by using the Cauchy-Schwarz inequality. Substituting (6.195) into (6.194) yields:

$$\|\Pi_i - \Pi_i^{\mathrm{ld}}\| \leq 2\sqrt{\mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i - \bar{\boldsymbol{w}}_i^{\mathrm{ld}}\|^2}\cdot\sqrt{\mu_{\max}^{-1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2} + \mu_{\max}^{-1}\mathbb{E}\|\bar{\boldsymbol{w}}_i^{\mathrm{ld}} - \widetilde{\boldsymbol{w}}_i\|^2 \tag{6.196}$$

Using (6.190) and Theorem 6.1, it follows from (6.196) that

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\|\Pi_i - \Pi_i^{\mathrm{ld}}\| = 0 \tag{6.197}$$

Noting that $\bar{\boldsymbol{w}}_i^{\mathrm{ld}}$ is obtained by extending $\widetilde{\boldsymbol{w}}_i^{\mathrm{ld}}$ via (6.90) and (6.91), we have

$$\mathbb{E}\bar{\boldsymbol{w}}_{m,i}^{\mathrm{ld}}(\bar{\boldsymbol{w}}_{n,i}^{\mathrm{ld}})^{\mathsf{T}} = (\mathbb{1}_{N_m^g}\mathbb{1}_{N_n^g}^{\mathsf{T}}) \otimes \mathbb{E}\widetilde{\boldsymbol{w}}_{m,i}^{\mathrm{ld}}(\widetilde{\boldsymbol{w}}_{n,i}^{\mathrm{ld}})^{\mathsf{T}} \tag{6.198}$$

for any m and n. From (6.98), we know that

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\|\mu_{\max}^{-1}\mathbb{E}\widetilde{\boldsymbol{w}}_{m,i}^{\mathrm{ld}}(\widetilde{\boldsymbol{w}}_{n,i}^{\mathrm{ld}})^{\mathsf{T}} - \Phi_{m,n}\| = 0 \tag{6.199}$$

where $\Phi_{m,n}$ denotes the (m,n)-th block of Φ with block size $M \times M$. It follows from (6.198) and (6.199) that

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\|\mu_{\max}^{-1}\mathbb{E}\bar{\boldsymbol{w}}_{m,i}^{\mathrm{ld}}(\bar{\boldsymbol{w}}_{n,i}^{\mathrm{ld}})^{\mathsf{T}} - (\mathbb{1}_{N_m^g}\mathbb{1}_{N_n^g}^{\mathsf{T}}) \otimes \Phi_{m,n}\| = 0 \tag{6.200}$$

Using (6.90), (6.115), and (6.191), we get from (6.200) that

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\|\Pi_i^{\mathrm{ld}} - \Pi\| = 0 \tag{6.201}$$

Combining (6.197) and (6.201), we arrive at (6.114).

6.D Proof of Lemma 6.5

We establish this result by calling upon Theorem 1.1 from [120, p. 319], which considers a stochastic recursion of the following form:

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} + \mu g(\boldsymbol{x}_{i-1}) + \mu \boldsymbol{v}_i \tag{6.202}$$

with step-size $\mu > 0$, update vector $g(\boldsymbol{x}_{i-1})$, and noise $\boldsymbol{v}_i$, satisfying the conditions:

1. The function $g(\cdot)$ is continuously differentiable and can be expanded as

$$g(x) = g(x^o) + [\nabla g(x^o)]^{\mathsf{T}}(x - x^o) + o(\|x - x^o\|) \tag{6.203}$$

   around a point x^o, where $\nabla g(\cdot)$ denotes the Jacobian of $g(\cdot)$, and $o(\cdot)$ is the "small-o" notation that represents higher order terms.

2. It holds that x^o is the unique point that satisfies:

$$g(x^o) = 0 \tag{6.204}$$

3. The Jacobian $A \triangleq \nabla g(x^o)$ is a Hurwitz matrix (i.e., the real parts of the eigenvalues of A are negative).

4. The noise process $\{\boldsymbol{v}_i; i \geq 0\}$ is a martingale difference, i.e.,

$$\mathbb{E}(\boldsymbol{v}_i | \mathbb{F}_{i-1}) = 0 \tag{6.205}$$

   where $\mathbb{F}_{i-1}$ is the filtration defined by $\{\boldsymbol{x}_i; i \geq 0\}$.

5. The noise $\boldsymbol{v}_i$ has an asymptotically bounded moment of order higher than 2, namely,

$$\lim_{\mu \to 0} \limsup_{i \to \infty} \mathbb{E}\|\boldsymbol{v}_i\|^{2+p} < \infty \tag{6.206}$$

   for some $p > 0$.

6. The covariance matrices of the noise process $\{\boldsymbol{v}_i; i \geq 0\}$ converge to a positive semi-definite matrix $\Sigma \geq 0$:

$$\lim_{\mu \to 0} \limsup_{i \to \infty} \|\mathbb{E}\boldsymbol{v}_i \boldsymbol{v}_i^{\mathsf{T}} - \Sigma\| = 0 \tag{6.207}$$

Under these conditions, it holds that as $i \to \infty$ and $\mu \to 0$ asymptotically, the sequence $\{\boldsymbol{x}_i/\sqrt{\mu}\}$ converges weakly to a Gaussian random distribution with mean x^o and covariance matrix C, which is the unique solution to the continuous Lyapunov equation $AC + CA^{\mathsf{T}} = \Sigma$.

These conditions are satisfied by our recursion (6.116) by identifying $\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}} \equiv \boldsymbol{x}_i$, $\mu_{\max} \equiv \mu$, $-\bar{\mathcal{H}}\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}^{\mathrm{ld}} \equiv g(\boldsymbol{x}_{i-1})$, $\boldsymbol{v}_i \equiv \bar{\boldsymbol{s}}_i$. First, since $\bar{\mathcal{H}}$ is positive-definite by (6.108) and (6.85), it is obvious that $x^o = 0$ is the unique point satisfying (6.204). Second, since $g(x) = -\bar{\mathcal{H}}x$ and $x^o = 0$, condition 1) holds automatically with $[\nabla g(x^o)]^{\mathsf{T}} = -\bar{\mathcal{H}}$. Third, it is easy to recognize that $A \equiv -\bar{\mathcal{H}}$ is Hurwitz since $\bar{\mathcal{H}}$ is positive-definite. Fourth, by (6.12) from Assumption 6.3, condition (6.205) holds. Fifth, by (6.13) from Assumption 6.3, we have

$$\begin{aligned}\mathbb{E}\|\bar{\boldsymbol{s}}_i\|^4 &\leq \|\mathcal{P}\|^4 \mathbb{E}\|\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})\|^4 \\ &\leq \|\mathcal{P}\|^4(\alpha^2 \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}\|^4 + \sigma_s^4)\end{aligned} \tag{6.208}$$

Using Theorem 6.1, we get from (6.208) that

$$\lim_{\mu_{\max} \to 0} \limsup_{i \to \infty} \mathbb{E}\|\bar{\boldsymbol{s}}_i\|^4 \leq \|\mathcal{P}\|^4(O(\mu_{\max}^2) + \sigma_s^4) < \infty \tag{6.209}$$

which satisfies condition (6.206). Sixth, we have from (6.117) and (6.11) that

$$\mathbb{E}\bar{\boldsymbol{s}}_i \bar{\boldsymbol{s}}_i^{\mathsf{T}} = \mu_{\max}^{-2} \mathcal{P}^{\mathsf{T}} \mathcal{M} \mathbb{E}\mathcal{R}_{s,i}(\boldsymbol{\mathcal{W}}_{i-1}) \mathcal{M}\mathcal{P} \tag{6.210}$$

Let

$$\Sigma_i \triangleq \mu_{\max}^{-2} \mathcal{P}^{\mathsf{T}} \mathcal{M} \mathcal{R}_{s,i}(\mathcal{W}^o) \mathcal{M}\mathcal{P} \tag{6.211}$$

Then, using Jensen's inequality and (6.14) from Assumption 6.3, we have from (6.210) that

$$\|\mathbb{E}\bar{\boldsymbol{s}}_i\bar{\boldsymbol{s}}_i^{\mathsf{T}} - \Sigma_i\| \leq \|\mathcal{P}\|^2\|\Delta\mathcal{R}_{s,i}\| \tag{6.212}$$

where $\Delta\mathcal{R}_{s,i}$ is from (6.183). Using (6.186), we further get

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\|\mathbb{E}\bar{\boldsymbol{s}}_i\bar{\boldsymbol{s}}_i^{\mathsf{T}} - \Sigma_i\| = 0 \tag{6.213}$$

Using (6.15), we have

$$\lim_{i\to\infty}\Sigma_i = \mu_{\max}^{-2}\mathcal{P}^{\mathsf{T}}\mathcal{M}\mathcal{R}_s\mathcal{M}\mathcal{P} = \bar{\mathcal{R}} \geq 0 \tag{6.214}$$

where $\bar{\mathcal{R}}$ is from (6.109). It follows from (6.213) and (6.214) that

$$\lim_{\mu_{\max}\to 0}\limsup_{i\to\infty}\|\mathbb{E}\bar{\boldsymbol{s}}_i\bar{\boldsymbol{s}}_i^{\mathsf{T}} - \bar{\mathcal{R}}\| = 0 \tag{6.215}$$

Therefore, we conclude that the sequence $\{\widetilde{\boldsymbol{\mathcal{W}}}_i^{\mathrm{ld}}/\sqrt{\mu_{\max}}; i \geq 0\}$ converges weakly to the Gaussian random variable with zero mean and covariance matrix Φ that satisfies (6.112).

6.E Proof of Lemma 6.6

We follow an argument similar to the proof of Theorem 2 from [121, p. 256] (which proves the result that convergence in moments implies convergence in distribution). Let $|f(x)| \leq c$, i.e., bounded. Because a continuous function $f(x)$ is also *uniformly* continuous in any *bounded* region [121, p. 54], for *any* constant $\epsilon > 0$ *and* for *any* constant $b > 0$, there exists some $\delta_{\epsilon,b} > 0$, which depends on the choices of ϵ *and* b, such that $|f(x) - f(y)| < \epsilon$ for $\|x\| < b$ *and* $\|x - y\| < \delta_{\epsilon,b}$. Now, setting $b \triangleq \sqrt{2c\sigma^2/\epsilon} > 0$, where σ^2 is from (6.121), and using conditional expectations, we have

$$\mathbb{E}|f(\boldsymbol{\zeta}_i) - f(\boldsymbol{\eta}_i)| = \mathbb{E}[|f(\boldsymbol{\zeta}_i) - f(\boldsymbol{\eta}_i)| \mid \|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\| < \delta_{\epsilon,b}, \|\boldsymbol{\zeta}_i\| < b]$$

$$
\begin{aligned}
&\times \mathbb{P}[\|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\| < \delta_{\epsilon,b}, \|\boldsymbol{\zeta}_i\| < b] \\
&+ \mathbb{E}[|f(\boldsymbol{\zeta}_i) - f(\boldsymbol{\eta}_i)| \mid \|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\| < \delta_{\epsilon,b}, \|\boldsymbol{\zeta}_i\| \geq b] \\
&\times \mathbb{P}[\|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\| < \delta_{\epsilon,b}, \|\boldsymbol{\zeta}_i\| \geq b] \\
&+ \mathbb{E}[|f(\boldsymbol{\zeta}_i) - f(\boldsymbol{\eta}_i)| \mid \|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\| \geq \delta_{\epsilon,b}] \\
&\times \mathbb{P}[\|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\| \geq \delta_{\epsilon,b}] \qquad (6.216)
\end{aligned}
$$

The first term on the RHS of (6.216) is bounded by

$$
\text{1st term} \leq \mathbb{E}[\epsilon \mid \|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\| < \delta, \|\boldsymbol{\zeta}_i\| < b] \times 1 = \epsilon \qquad (6.217)
$$

Using the fact that $|f(x) - f(y)| \leq |f(x)| + |f(y)| \leq 2c$, and also the fact that the joint probability is bounded by any one of the marginal probabilities, i.e., $\mathbb{P}[A \cap B] \leq \mathbb{P}[A]$ for any two events A and B, the second term on the RHS of (6.216) is bounded by

$$
\text{2nd term} \leq 2c\,\mathbb{P}[\|\boldsymbol{\zeta}_i\| \geq b] \leq \frac{2c\,\mathbb{E}\|\boldsymbol{\zeta}_i\|^2}{b^2} = \frac{\epsilon\,\mathbb{E}\|\boldsymbol{\zeta}_i\|^2}{\sigma^2} \qquad (6.218)
$$

where we used Chebyshev's inequality [121, p. 47]. Likewise, the third term on the RHS of (6.216) is bounded by

$$
\text{3rd term} \leq 2c\,\mathbb{P}[\|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\| \geq \delta] \leq \frac{2c\,\mathbb{E}\|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\|^2}{\delta^2} \qquad (6.219)
$$

Now, substituting (6.217)–(6.219) into (6.216), we have

$$
\mathbb{E}|f(\boldsymbol{\zeta}_i) - f(\boldsymbol{\eta}_i)| \leq \epsilon + \frac{\epsilon\,\mathbb{E}\|\boldsymbol{\zeta}_i\|^2}{\sigma^2} + \frac{2c\,\mathbb{E}\|\boldsymbol{\zeta}_i - \boldsymbol{\eta}_i\|^2}{\delta^2} \qquad (6.220)
$$

Using (6.120) and (6.121), we end up with

$$
\lim_{\mu_{\max} \to 0} \limsup_{i \to \infty} \mathbb{E}|f(\boldsymbol{\zeta}_i) - f(\boldsymbol{\eta}_i)| \leq 2\epsilon \qquad (6.221)
$$

Since ϵ is arbitrary, result (6.122) follows from (6.221).

6.F Proof of (6.137)

To simplify the notation, we drop the subscript of $\boldsymbol{d}_{k,\ell}$ and denote its mean by $\bar{d} \triangleq \mathbb{E}\boldsymbol{d}$ and its covariance by $C \triangleq \mathbb{E}(\boldsymbol{d}-\bar{d})(\boldsymbol{d}-\bar{d})^{\mathsf{T}}$. Since $\boldsymbol{d}$ is Gaussian, it holds that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{d}\|^4 &= \mathbb{E}\|\boldsymbol{d}-\bar{d}+\bar{d}\|^4 \\
&= \mathbb{E}[\|\boldsymbol{d}-\bar{d}\|^2 + 2(\boldsymbol{d}-\bar{d})^{\mathsf{T}}\bar{d} + \|\bar{d}\|^2]^2 \\
&= \mathbb{E}\|\boldsymbol{d}-\bar{d}\|^4 + 2\mathbb{E}\|\boldsymbol{d}-\bar{d}\|^2\|\bar{d}\|^2 + \|\bar{d}\|^4 + 4\bar{d}^{\mathsf{T}}\mathbb{E}[(\boldsymbol{d}-\bar{d})(\boldsymbol{d}-\bar{d})^{\mathsf{T}}]\bar{d} \\
&= \mathbb{E}\|\boldsymbol{d}-\bar{d}\|^4 + 2\mathrm{Tr}(C)\|\bar{d}\|^2 + \|\bar{d}\|^4 + 4\|\bar{d}\|_C^2 \qquad (6.222)
\end{aligned}$$

where we used the fact that the odd order moments of $\boldsymbol{d}-\bar{d}$ is zero. Likewise,

$$\begin{aligned}
(\mathbb{E}\|\boldsymbol{d}\|^2)^2 &= (\mathbb{E}\|\boldsymbol{d}-\bar{d}+\bar{d}\|^2)^2 \\
&= (\mathbb{E}\|\boldsymbol{d}-\bar{d}\|^2 + \|\bar{d}\|^2)^2 \\
&= [\mathrm{Tr}(C)]^2 + 2\mathrm{Tr}(C)\|\bar{d}\|^2 + \|\bar{d}\|^4 \qquad (6.223)
\end{aligned}$$

From (6.222) and (6.223), we have

$$\mathbb{E}\|\boldsymbol{d}\|^4 - (\mathbb{E}\|\boldsymbol{d}\|^2)^2 = \mathbb{E}\|\boldsymbol{d}-\bar{d}\|^4 - [\mathrm{Tr}(C)]^2 + 4\|\bar{d}\|_C^2 \qquad (6.224)$$

From Lemma A.2 of [46, p. 11], it can be verified that

$$\mathbb{E}\|\boldsymbol{d}-\bar{d}\|^4 = [\mathrm{Tr}(C)]^2 + 2\mathrm{Tr}(C^2) \qquad (6.225)$$

Substituting (6.225) into (6.224) yields:

$$\mathbb{E}\|\boldsymbol{d}\|^4 - (\mathbb{E}\|\boldsymbol{d}\|^2)^2 = 2\mathrm{Tr}(C^2) + 4\|\bar{d}\|_C^2 \qquad (6.226)$$

CHAPTER 7

Relating Consensus and Diffusion Strategies to Penalty Methods

In this chapter, we establish a connection between distributed consensus and diffusion strategies and classical diagonally-weighted gradient-descent iterations under a special local balance condition. We start from an aggregate cost function defined over a network of agents and regularize it by adding a weighted quadratic term whose nullspace coincides with the agreement subspace for all agents. Under a local balance condition on the combination coefficients over the edges, we show that consensus and diffusion strategies can be interpreted as diagonally-weighted gradient-descent iterations. In this case, stability and performance analysis for single-agent implementations become applicable to the distributed solutions. When the local balance condition is not satisfied, the dynamics of the distributed solutions become richer and their analysis becomes more demanding (see, e.g., [66, 67]). The results in this chapter can be used to provide an interpretation for the two-phase transient behavior of consensus and diffusion strategies.

7.1 Introduction and Problem Formulation

We consider a network consisting of N agents that are connected via some topology. Each agent k has an individual cost function $J_k(w) : \mathbb{R}^{M\times 1} \mapsto \mathbb{R}$. The cost $J_k(w)$ is assumed to be convex and twice-differentiable. A minimizer of $J_k(w)$ is denoted by w_k^o. In general, the minimizers $\{w_k^o\}$ across the agents do not necessarily coincide with each other, i.e., $w_k^o \neq w_\ell^o$ for $k \neq \ell$. In addition, we assume that there exists at least one individual cost, say, $J_m(w)$, that is *strongly*-convex, which means that its Hessian matrix is uniformly bounded away from zero, namely,

$$\nabla^2 J_m(w) \geq \lambda_{m,L} I_M \tag{7.1}$$

for any $w \in \mathbb{R}^{M\times 1}$ and some positive constant $\lambda_{m,L}$. The entire network of agents aims to seek the *unique* solution to the following minimization problem:

$$\underset{w}{\text{minimize}} \quad J^{\text{gen}}(w) \triangleq \sum_{k=1}^{N} q_k J_k(w) \tag{7.2}$$

where the $\{q_k\}$ is a set of convex coefficients that satisfy

$$q_k > 0, \qquad \sum_{k=1}^{N} q_k = 1 \tag{7.3}$$

Since the individual costs $\{J_k(k)\}$ in (7.2) are convex, and the weights $\{q_k\}$ are positive, the weighted aggregate cost $J^{\text{gen}}(w)$ is also convex. Furthermore, since there exists at least one strongly-convex individual cost $J_m(w)$, the aggregate cost $J^{\text{gen}}(w)$ is also strongly-convex, which ensures that $J^{\text{gen}}(w)$ has a *unique* minimizer. We denote this unique global minimizer by w^o, which satisfies the first-order condition for optimality [135]:

$$\nabla J^{\text{gen}}(w^o) = \sum_{k=1}^{N} q_k \nabla J_k(w^o) = 0 \tag{7.4}$$

Note that the general cost $J^{\text{gen}}(\cdot)$ in (7.2) is a linear combination of the individual costs $\{J_k(w)\}$ with convex weights $\{q_k\}$, so it can be regarded as the *scalarization* of the vector cost $J^{\text{vec}}(w) : \mathbb{R}^{M\times 1} \mapsto \mathbb{R}^{N\times 1}$ (i.e., a multi-objective cost) [71, p. 178]:

$$J^{\text{vec}}(w) \triangleq \text{col}\{J_1(w), \ldots, J_N(w)\} \tag{7.5}$$

with the weights $\{q_k\}$. Therefore, the minimizer w^o of the scalarized (aggregate) cost $J^{\text{gen}}(\cdot)$ is a Pareto optimal point for the vector cost $J^{\text{vec}}(w)$ [71, p. 177]. If the individual costs $\{J_k(w)\}$ are not minimized at the same point, then in general the Pareto optimal point w^o would be different from any one of the individual minimizers. If all individual costs happen to share a common minimizer such that $w_1^o = w_2^o = \cdots = w_N^o$, then the Pareto optimal point w^o will be identical to all the individual minimizers, i.e., $w_k^o = w^o$ for all k.

In this chapter, we use a regularized penalty method to motivate the consensus and diffusion strategies for solving problem (7.2) in a *distributed* manner over *any connected* topology.

7.2 Regularization for Distributed Processing

We assume that each agent in the network only has access to its own individual cost, $J_k(w)$, and is allowed to interact with its local neighbors (as defined by the topology) during the learning process. Since each agent will generate its own estimate for the global minimizer w^o, there will be N estimates in the network, which are denoted by $\{w_{k,i}\}_{k=1}^N$. The subscripts k and i indicate that $w_{k,i}$ is an estimate generated by agent k at time i. In order to emphasize the fact that N separate estimates will be evaluated across the network, we rewrite problem (7.2)

explicitly in the following *equivalent* form:

$$\begin{aligned} \underset{\{w_k\}_{k=1}^N}{\text{minimize}} \quad & J^{\text{dist}}(w_1,\ldots,w_N) \triangleq \sum_{k=1}^N q_k J_k(w_k) \\ \text{subject to} \quad & w_1 = \cdots = w_N \end{aligned} \tag{7.6}$$

The interpretation for this formulation is that each agent k is allowed to have its own estimate, w_k, of the parameter vector w^o, but all versions from across the network will need to be aligned. Compared with the original aggregate cost $J^{\text{gen}}(w)$ that is defined over $\mathbb{R}^{M\times 1}$, the distributed aggregate cost function $J^{\text{dist}}(w_1,\ldots,w_N)$ is defined over the higher-dimensional space $\mathbb{R}^{NM\times 1}$ since its argument is now given by the aggregate vector $\mathcal{w} \triangleq \text{col}\{w_1,\ldots,w_N\} \in \mathbb{R}^{NM\times 1}$. The dimension of the parameter space is expanded from M for w to NM for $\mathcal{w}$. Due to the agreement constraint in (7.6), the feasible region for $\mathcal{w}$ is the agreement subspace $\mathbb{S} \triangleq \{\mathbb{1}_N \otimes x; x \in \mathbb{R}^{M\times 1}\} \subset \mathbb{R}^{NM\times 1}$, where $\mathbb{1}_N$ denotes the $N\times 1$ vector with all entries equal to one. Using (7.6), the Hessian matrix of $J^{\text{dist}}(\mathcal{w})$ is given by

$$\nabla^2 J^{\text{dist}}(\mathcal{w}) = \text{diag}\{q_k \nabla^2 J_k(w_k)\}_{k=1}^N \tag{7.7}$$

where w_k is the k-th sub-vector of $\mathcal{w}$.

Lemma 7.1 (Strong-convexity of $J^{\text{dist}}(\mathcal{w})$)**.** *The distributed cost $J^{dist}(\mathcal{w})$ in (7.6) is strongly-convex in the agreement subspace $\mathbb{S}$, namely, $\nabla^2 J^{dist}(\mathcal{w}) \geq bI_{NM}$ for any $\mathcal{w} \in \mathbb{S}$ and some positive constant b.*

Proof. The statement is proven by showing that

$$x^{\mathsf{T}} \nabla^2 J^{\text{dist}}(\mathcal{w}) x \geq b > 0 \tag{7.8}$$

for *any* $\mathcal{w} \triangleq \mathbb{1}_N \otimes w \in \mathbb{S}$ and *any* $x \triangleq \mathbb{1}_N \otimes x_o \in \mathbb{S}$. This is true because

$$x^{\mathsf{T}} \nabla^2 J^{\text{dist}}(\mathcal{w}) x = \sum_{k=1}^N x_o^{\mathsf{T}} q_k \nabla^2 J_k(w) x_o$$

$$
\begin{aligned}
&= x_o^{\mathsf{T}} \left(\sum_{k=1}^{N} q_k \nabla^2 J_k(w) \right) x \\
&= x_o^{\mathsf{T}} \nabla^2 J^{\text{gen}}(w) x_o \\
&\geq b > 0
\end{aligned} \tag{7.9}
$$

for some positive constant b, where we used (7.7) and the fact that $J^{\text{gen}}(\cdot)$ is strongly-convex. $\square$

Due to the equivalency of problems (7.2) and (7.6), the cost $J^{\text{dist}}(\mathcal{w})$ has a unique minimizer within $\mathbb{S}$, which is given by

$$
\mathcal{w}^o \triangleq \mathbb{1}_N \otimes w^o \tag{7.10}
$$

In order to solve the *constrained* problem (7.6) in a distributed manner, we call upon the penalty method [71,72,135,136]. The choice of the penalty function plays an important role here because it will help reflect the network topology, as well as penalize solutions that violate the agreement constraint in (7.6). The penalty function will be related to two sets of parameters, namely, the step-sizes $\{\mu_k; k = 1, 2, \ldots, N\}$ and the combination coefficients $\{a_{\ell k}; k, \ell = 1, 2, \ldots, N\}$, which will appear in the distributed algorithms (7.33), (7.35), and (7.36).

The step-size parameters $\{\mu_k\}$ are a set of *positive constants*, one for each agent k. These parameters control the stability and the convergence rate of the distributed algorithms, as already explained in [66]. Using $\{\mu_k\}$, we introduce a set of auxiliary parameters:

$$
\beta_k \triangleq \frac{q_k/\mu_k}{\sum_{m=1}^{N} q_m/\mu_m} > 0 \tag{7.11}
$$

where $\{q_k\}$ are from (7.2). It is obvious that the $\{\beta_k\}$ satisfy $\sum_{k=1}^{N} \beta_k = 1$.

The selection of the second set of parameters, i.e., the combination coefficients $\{a_{\ell k}\}$, depends on the auxiliary parameters $\{\beta_k\}$. Specifically, we introduce a

combination coefficient matrix $A \triangleq [a_{\ell k}]_{\ell,k=1}^{N} \mathbb{R}^{N\times N}$ to represent a *weighted* graph based on the underlying topology. The entries on the k-th column of A, i.e., $\{a_{\ell k}\}$ for each k, will need to satisfy

$$\begin{cases} a_{\ell k} > 0 \ \text{ if } \ \ell \in \mathcal{N}_k \backslash \{k\}, & a_{kk} \geq 0 \\ a_{\ell k} = 0 \ \text{ if } \ \ell \notin \mathcal{N}_k, & \displaystyle\sum_{\ell=1}^{N} a_{\ell k} = 1 \end{cases} \tag{7.12}$$

where $\mathcal{N}_k$ denotes the neighborhood of agent k with the convention $k \in \mathcal{N}_k$. Among all diagonal entries of A, i.e., $\{a_{kk}\}$, there must be at least one positive entry. Since the network topology is assumed to be connected, and since at least one diagonal entry of A is positive, then the weighted graph corresponding to A is strongly-connected [66, p. 435], i.e., the graph is connected and contains at least one self-loop. In addition, the coefficients $\{a_{\ell k}\}$ will also need to satisfy the *local balance* condition:

$$a_{\ell k}\beta_k = a_{k\ell}\beta_\ell \tag{7.13}$$

for any pair of agents k and ℓ, where β_k and β_ℓ are from (7.11). Using (7.11), condition (7.13) can be rewritten as

$$\frac{a_{\ell k} q_k}{\mu_k} = \frac{a_{k\ell} q_\ell}{\mu_\ell} \tag{7.14}$$

For example, the Hastings rule proposed in [36, 66], and which was motivated by earlier results on Monte Carlo Markov Chains in [137], satisfies the local balance condition (7.13) and it applies to *any connected* topology. Specifically, the Hastings rule assigns the following weights to the links between agents:

$$a_{\ell k} = \begin{cases} \dfrac{\theta_k^2}{\max\{n_k\theta_k^2, n_\ell\theta_\ell^2\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ 1 - \displaystyle\sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}, & \ell = k \end{cases} \tag{7.15}$$

where $n_k = |\mathcal{N}_k|$ denotes the size of the neighborhood $\mathcal{N}_k$ and $\theta_k^2 = \mu_k/q_k$. Other constructions are also possible. We therefore assume that the combination coefficients $\{a_{\ell k}\}$ have been chosen to satisfy condition (7.13) or (7.14).

Since the weighted graph corresponding to A is strongly-connected, and the entries of A satisfy condition (7.12), the matrix A is left-stochastic and primitive. It then follows from the Perron-Frobenius theorem [66, 79] that A has a simple eigenvalue at one, while all other eigenvalues are inside the unit circle. We denote the left and right eigenvectors that are associated with the eigenvalue at one by

$$A^{\mathsf{T}}\mathbb{1}_N = \mathbb{1}_N, \quad Ap = p, \quad p^{\mathsf{T}}\mathbb{1}_N = 1 \tag{7.16}$$

and normalize the entries of p to add up to one. We refer to p as the Perron eigenvector of A. It further follows from the Perron-Frobenius theorem that all entries of p are strictly positive, written as $p \succ 0$. The entries of the Perron vector p can be identified as follows. Using the local balance condition (7.13) and the fact that $\sum_{k=1}^N a_{k\ell} = 1$ from (7.12), it holds that

$$\begin{aligned}
\sum_{k=1}^{N} a_{\ell k}\beta_k &= \sum_{k\neq\ell} a_{\ell k}\beta_k + a_{\ell\ell}\beta_\ell \\
&= \sum_{k\neq\ell} a_{k\ell}\beta_\ell + a_{\ell\ell}\beta_\ell \\
&= (1 - a_{\ell\ell})\beta_\ell + a_{\ell\ell}\beta_\ell \\
&= \beta_\ell
\end{aligned} \tag{7.17}$$

From (7.17) and the fact that $\beta_k > 0$ and $\sum_{k=1}^N \beta_k = 1$, we readily identify the entries of p as

$$p_k \equiv \beta_k \quad \text{or} \quad p \equiv \text{col}\{\beta_1, \ldots, \beta_N\} \tag{7.18}$$

where p_k denotes the k-th entry of p.

Now, using these Perron entries $\{p_k\}$ and the combination coefficients $\{a_{\ell k}\}$, we introduce a penalized cost function with a quadratic regularization term as

follows:

$$\underset{\{w_k\}_{k=1}^N}{\text{minimize}} \quad J^{\text{pen}}(w_1,\ldots,w_N) \triangleq \sum_{k=1}^N q_k J_k(w_k) + \eta \sum_{k=1}^N \sum_{\ell=1}^N p_k a_{\ell k} \|w_k - w_\ell\|^2 \quad (7.19)$$

where η is a positive factor chosen as

$$\eta \triangleq \frac{1}{4} \sum_{k=1}^N \frac{q_k}{\mu_k} > 0 \quad (7.20)$$

Compared with the *constrained* aggregate cost $J^{\text{dist}}(\mathcal{W})$ in (7.6), which enforces agreement among all agents, the penalized cost $J^{\text{pen}}(\mathcal{W})$ is *unconstrained* and, therefore, allows for some small disagreements among the agents. The level of tolerance for discrepancy is determined by the weighting factor η, which is inversely proportional to the step-size parameters. Therefore, when step-sizes $\{\mu_k\}$ are sufficiently small, which will correspond to operation in the slow adaptation regime, the regularization term weighted by η will be significant and will penalize discrepancies among agents. Let $P \triangleq \text{diag}(p)$. Note that the (ℓ,k)-th entry of AP is given by $a_{\ell k} p_k$, while the (ℓ,k)-th entry of PA^{T} is given by $a_{k\ell} p_\ell$. Then, it follows from the local balance condition (7.13) that

$$AP = PA^{\mathsf{T}} \quad (7.21)$$

Let $\mathcal{A} \triangleq A \otimes I_M$ and $\mathcal{P} \triangleq P \otimes I_M$. Then, the regularization term in (7.19) can be expressed as

$$\begin{aligned}
&\sum_{k=1}^N \sum_{\ell=1}^N p_k a_{\ell k} \|w_k - w_\ell\|^2 \\
&\qquad = \sum_{k=1}^N \sum_{\ell=1}^N p_k a_{\ell k} (\|w_k\|^2 + \|w_\ell\|^2 - 2 w_k^{\mathsf{T}} w_\ell) \\
&\qquad \overset{(a)}{=} \sum_{k=1}^N p_k \|w_k\|^2 \sum_{\ell=1}^N a_{\ell k} + \sum_{\ell=1}^N p_\ell \|w_\ell\|^2 \sum_{k=1}^N a_{k\ell} - 2 \sum_{k=1}^N \sum_{\ell=1}^N p_k a_{\ell k} w_k^{\mathsf{T}} w_\ell \\
&\qquad \overset{(b)}{=} 2 \sum_{k=1}^N p_k \|w_k\|^2 - 2 \sum_{k=1}^N \sum_{\ell=1}^N p_k a_{\ell k} w_k^{\mathsf{T}} w_\ell
\end{aligned}$$

$$= 2\mathcal{w}^{\mathsf{T}}\mathcal{P}\mathcal{w} - 2\mathcal{w}^{\mathsf{T}}(\mathcal{A}\mathcal{P})\mathcal{w}$$
$$= 2\mathcal{w}^{\mathsf{T}}(\mathcal{P} - \mathcal{A}\mathcal{P})\mathcal{w} \tag{7.22}$$

where step (a) is by using (7.13) to replace $p_k a_{\ell k}$ with $p_\ell a_{k\ell}$ in the second term on the RHS, and step (b) is by using (7.12). According to (7.21), the matrix $\mathcal{P}-\mathcal{A}\mathcal{P}$ is symmetric. Furthermore, we know from (7.22) that $\mathcal{w}^{\mathsf{T}}(\mathcal{P}-\mathcal{A}\mathcal{P})\mathcal{w} \geq 0$ for any $\mathcal{w} \in \mathbb{R}^{NM\times 1}$, which implies that $\mathcal{P}-\mathcal{A}\mathcal{P}$ is positive semi-definite. Using (7.22), the penalized cost $J^{\text{pen}}(\mathcal{w})$ in (7.19) can be rewritten in terms of the network variable $\mathcal{w}$ as

$$J^{\text{pen}}(\mathcal{w}) = J^{\text{dist}}(\mathcal{w}) + 2\eta\|\mathcal{w}\|^2_{\mathcal{P}-\mathcal{A}\mathcal{P}} \tag{7.23}$$

As a convex combination of convex individual costs $\{J_k(w_k)\}$, $J^{\text{dist}}(\mathcal{w})$ is also convex. Obviously, since $J^{\text{dist}}(\mathcal{w})$ and $2\eta\|\mathcal{w}\|^2_{\mathcal{P}-\mathcal{A}\mathcal{P}}$ are both twice-differentiable and convex, the penalized cost $J^{\text{pen}}(\mathcal{w})$ in (7.23) is also twice-differentiable and convex. Using (7.16), it is straightforward to verify that

$$(P - AP)\mathbb{1}_N = p - Ap = 0 \tag{7.24}$$

Lemma 7.2 (Nullspace of $\mathcal{P}-\mathcal{A}\mathcal{P}$)**.** *The weighting matrix $\mathcal{P}-\mathcal{A}\mathcal{P}$ in* (7.23) *is rank-deficient, and its nullspace coincides with the agreement subspace* $\mathbb{S}$*.*

Proof. Note that the diagonal matrix P is non-singular so that the rank of $P - AP = (I_N - A)P$ is identical to the rank of $I_N - A$. Since A is left-stochastic and primitive, its eigenvalue at one is simple by the Perron-Frobenius theorem [66,79]. Then, the eigenvalue of $I_N - A$ at zero is also simple, which implies that the rank of $P - AP$ is $N - 1$. Thus, the dimension of the null space of $P - AP$ is one. From (7.24), it is clear that $\mathbb{1}_N$ is in the nullspace of $P-AP$. Therefore, the nullspace of $P-AP$ is given by $\text{Span}(\mathbb{1}_N)$, and the nullspace of $\mathcal{P}-\mathcal{A}\mathcal{P}$ coincides with the agreement subspace $\mathbb{S}$. □

Using Lemma 7.2, it is clear that the quadratic regularization term, i.e., $2\eta\|\mathcal{W}\|^2_{\mathcal{P}-\mathcal{A}\mathcal{P}}$, in the cost $J^{\rm pen}(\mathcal{W})$ will generate a large penalty for solutions outside of the agreement subspace $\mathbb{S}$ and will have no impact on solutions inside $\mathbb{S}$. In this way, the penalized cost $J^{\rm pen}(\mathcal{W})$ *encourage* rather than *enforce* a consensus solution to the distributed cost $J^{\rm dist}(\mathcal{W})$.

Lemma 7.3 (Strong-convexity of $J^{\rm pen}(\mathcal{W})$)**.** *The penalized cost $J^{pen}(\mathcal{W})$ from (7.23) is strongly-convex in $\mathbb{R}^{NM\times 1}$, i.e., $\nabla^2 J^{pen}(\mathcal{W}) \geq cI_{NM}$ for any $\mathcal{W}\in\mathbb{R}^{NM\times 1}$ and some positive constant c.*

Proof. See Appendix 7.A. □

7.3 Distributed Gradient Descent Iteration

Now, we apply a diagonally-weighted gradient-descent algorithm [135] to solve problem (7.19), namely,

$$\mathcal{W}_i = \mathcal{W}_{i-1} - \mu\mathcal{P}^{-1}\nabla J^{\rm pen}(\mathcal{W}_{i-1}) \tag{7.25}$$

where μ is a (derived) positive step-size parameter chosen as

$$\mu \triangleq \frac{1}{4\eta} = \left(\sum_{k=1}^{N}\frac{q_k}{\mu_k}\right)^{-1} \tag{7.26}$$

It is worth noting that the μ in (7.26) is a weighted harmonic average of the step-sizes $\{\mu_k\}$. Let $\mu_{\max} \triangleq \max_k\{\mu_k\}$. Then, it follows from (7.3) and (7.26) that

$$\mu \leq \left(\sum_{k=1}^{N}\frac{q_k}{\mu_{\max}}\right)^{-1} = \mu_{\max} \tag{7.27}$$

Therefore, μ is consistent with all the step-sizes $\{\mu_k\}$; μ will become small when the $\{\mu_k\}$ do so. Since $\mathcal{P}$ is diagonal, all the sub-vectors of $\mathcal{W}_i$, i.e., $\{w_{k,i}\}$, in

(7.25) can be updated in parallel:

$$w_{k,i} = w_{k,i-1} - \mu p_k^{-1} \nabla_k J^{\text{pen}}(\mathcal{W}_{i-1}) \tag{7.28}$$

where $\nabla_k J^{\text{pen}}(\mathcal{W}_{i-1})$ denotes the partial gradient of $J^{\text{pen}}(\cdot)$ with respect to the k's sub-vector of $\mathcal{W}_{i-1}$, i.e., $w_{k,i-1}$, and is given by

$$\begin{aligned}\nabla_k J^{\text{pen}}(\mathcal{W}) &= q_k \nabla J_k(w_k) + 2\eta \sum_{\ell=1}^{N} (p_k a_{\ell k} + p_\ell a_{k\ell})(w_k - w_\ell) \\ &= q_k \nabla J_k(w_k) + 4\eta p_k \sum_{\ell=1}^{N} a_{\ell k}(w_k - w_\ell)\end{aligned} \tag{7.29}$$

where we used the local balance condition (7.13). Since the underlying topology is undirected, each incoming-neighbor of agent k is also an outgoing-neighbor of k. Substituting (7.29) into (7.28), and using (7.26), we obtain

$$w_{k,i} = w_{k,i-1} - \frac{\mu q_k}{p_k} \nabla J_k(w_{k,i-1}) - \sum_{\ell=1}^{N} a_{\ell k}(w_{k,i-1} - w_{\ell,i-1}) \tag{7.30}$$

Using (7.18) and (7.11), it can be verified that

$$p_k \mu_k = \beta_k \mu_k = \frac{q_k}{\sum_{m=1}^{N} q_m/\mu_m} = q_k \mu \tag{7.31}$$

Therefore, we get from (7.30) that

$$w_{k,i} = w_{k,i-1} - \mu_k \nabla J_k(w_{k,i-1}) - \sum_{\ell=1}^{N} a_{\ell k}(w_{k,i-1} - w_{\ell,i-1}) \tag{7.32}$$

Using condition (7.12), we readily arrive at the consensus strategy from (7.32), namely,

$$w_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} w_{\ell,i-1} - \mu_k \nabla J_k(w_{k,i-1}) \tag{7.33}$$

Alternatively, if we split the single gradient-descent step in (7.32) into two consecutive steps, then we would obtain

$$\begin{cases} \psi_{k,i} = w_{k,i-1} - \mu_k \nabla J_k(w_{k,i-1}) \\ w_{k,i} = \psi_{k,i} - \displaystyle\sum_{\ell=1}^{N} a_{\ell k}(w_{k,i-1} - w_{\ell,i-1}) \end{cases} \tag{7.34}$$

If we follow the incremental scheme [135] by replacing the $w_{k,i-1}$ and $w_{\ell,i-1}$ in the second step of (7.34) with the intermediate variables $\psi_{k,i}$ and $\psi_{\ell,i}$, respectively, then we would obtain the adapt-then-combine (ATC) diffusion strategy [6,12,66]:

$$\begin{cases} \psi_{k,i} = w_{k,i-1} - \mu_k \nabla J_k(w_{k,i-1}) \\ w_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \psi_{\ell,i} \end{cases} \tag{7.35}$$

In ATC diffusion adaptation (7.35), we use the post-combination variables to represent the state of each iteration. An alternative implementation is to use the post-adaptation variables, which leads to the combine-then-adapt (CTA) diffusion strategy [6, 12, 66]:

$$\begin{cases} \phi_{k,i-1} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} w_{\ell,i-1} \\ w_{k,i} = \phi_{k,i-1} - \mu_k \nabla J_k(\phi_{k,i-1}) \end{cases} \tag{7.36}$$

The stochastic implementations of these three strategies (7.33), (7.35), and (7.36) can be obtained by replacing the true gradient $\nabla J_k(\cdot)$ by its stochastic approximation $\widehat{\nabla J_k}(\cdot)$.

It is worth noting that the local balance condition (7.13) is a *necessary* condition to express consensus strategies in the form of diagonally-weighted gradient-descent iterations. This is because a diagonally-weighted gradient-descent iteration that minimizes some underlying cost $J(x)$ is given by the following generic form:

$$x_i = x_{i-1} - \mu \mathcal{D} \nabla J(x_{i-1}) \triangleq x_{i-1} - \mu g(x_{i-1}) \tag{7.37}$$

where μ is a step-size parameter, $\mathcal{D} \triangleq D \otimes I_M$ is a positive definite diagonal matrix, and $g(x) \triangleq \mathcal{D} \nabla J(x)$ is the update vector. Then, the Jacobian of the update vector in (7.37) is given by

$$\nabla g(x) = \mathcal{D} \nabla^2 J(x) \tag{7.38}$$

which implies that, up to a matrix factor $\mathcal{D}^{-1}$, the Jacobian of the update vector must be a Hessian matrix and, therefore, must be symmetric and positive semi-definite. Now, let us define $\mathcal{M} \triangleq \text{diag}\{\mu_1, \ldots, \mu_N\} \otimes I_M$. Then, the consensus strategy in (7.33) can be written into the following network aggregate form:

$$\mathcal{W}_i = \mathcal{A}^{\mathsf{T}} \mathcal{W}_{i-1} - \mathcal{M} \nabla J^{\text{dist}}(\mathcal{W}_{i-1}) \tag{7.39}$$

where $\nabla J^{\text{dist}}(\mathcal{W}) \triangleq \text{diag}\{q_k \nabla J_k(w_k)\}_{k=1}^{N}$. This iteration can be rewritten into a form similar to (7.37), i.e.,

$$\mathcal{W}_i = \mathcal{W}_{i-1} - \mu g(\mathcal{W}_{i-1}) \tag{7.40}$$

where the effective update vector is given by

$$g(\mathcal{W}) \triangleq \mu^{-1}[\mathcal{M} \nabla J^{\text{dist}}(\mathcal{W}) + (I_{NM} - \mathcal{A}^{\mathsf{T}})\mathcal{W}] \tag{7.41}$$

The Jacobian of the update vector $g(\mathcal{W})$ in (7.41) is given by

$$\nabla g(\mathcal{W}) = \mu^{-1}[\mathcal{M} \nabla^2 J^{\text{dist}}(\mathcal{W}) + (I_{NM} - \mathcal{A}^{\mathsf{T}})] \tag{7.42}$$

Therefore, if the update vector $g(\mathcal{W})$ from (7.41) is the gradient of some underlying cost, then up to a matrix factor $\mathcal{D}^{-1}$, the Jacobian $\nabla g(\mathcal{W})$ needs to be symmetric and positive semi-definite. Assume there exists such a matrix factor $\mathcal{D}^{-1}$. Then, the following condition for symmetry must hold:

$$\mathcal{D}^{-1} \nabla g(\mathcal{W}) = [\nabla g(\mathcal{W})]^{\mathsf{T}} \mathcal{D}^{-1} \tag{7.43}$$

Note that the first component in $g(\mathcal{W})$, i.e., $\mu^{-1}\mathcal{M}\nabla J^{\text{dist}}(\mathcal{W})$, is *block* diagonal and positive semi-definite, so it holds for any block diagonal matrix $\mathcal{D} = D \otimes I_M$, where D is diagonal and positive definite, that

$$\mathcal{D}^{-1}\mu^{-1}\mathcal{M} \nabla J^{\text{dist}}(\mathcal{W}) = [\mu^{-1}\mathcal{M}\nabla J^{\text{dist}}(\mathcal{W})]^{\mathsf{T}} \mathcal{D}^{-1} \tag{7.44}$$

Therefore, we only need to examine the following condition:

$$\mathcal{D}^{-1}\mu^{-1}(I_{NM}-\mathcal{A}^{\mathsf{T}})=[\mu^{-1}(I_{NM}-\mathcal{A}^{\mathsf{T}})]^{\mathsf{T}}\mathcal{D}^{-1} \tag{7.45}$$

which reduces to

$$D^{-1}A^{\mathsf{T}}=AD^{-1} \tag{7.46}$$

The (ℓ,k)-th entry of $D^{-1}A^{\mathsf{T}}$ is given by $a_{k\ell}/d_\ell$, while the (ℓ,k)-th entry of AD^{-1} is given by $a_{\ell k}/d_k$. Therefore, condition (7.46) holds if, and only if,

$$\frac{a_{k\ell}}{d_\ell}=\frac{a_{\ell k}}{d_k} \tag{7.47}$$

By identifying $\beta_k \equiv d_k^{-1}$, condition (7.47) is equivalent to the local balance condition (7.13). Therefore, the local balance condition (7.13), or (7.14), is a *necessary* condition to express the consensus strategy (7.33) in a form that is equivalent to a diagonally-weighted gradient-descent iteration for minimizing some underlying cost. It is also worth noting that, in addition to be *symmetric*, the matrix $\mathcal{D}^{-1}\nabla g(\mathcal{W})$ in (7.43) also needs to be *positive semi-definite* to be *sufficient* to express the consensus strategy (7.33) as a diagonally-weighted gradient-descent iteration.

7.4 Concluding Remarks

Based on the derivations from the previous section, if the combination coefficients $\{a_{\ell k}\}$ *satisfy* the local balance condition (7.14), then the consensus strategy (7.33) can be recognized as a standard gradient-descent iteration for minimizing the penalized cost $J^{\text{pen}}(\cdot)$ in (7.19); likewise, the diffusion strategies (7.35) and (7.36) can be recognized as an incremental iteration for minimizing the same penalized cost $J^{\text{pen}}(\cdot)$ starting from the same gradient-descent iteration. Moreover, we have shown in Lemma 7.3 that the penalized cost $J^{\text{pen}}(\cdot)$ is strongly-convex. Therefore, results for single-agent implementations [135] are applicable for distributed

consensus and diffusion strategies when the local balance condition (7.14) holds. For example, the stability of gradient-descent methods is established in [72] by assuming constant step-sizes with deterministic or stochastic gradient errors. The convergence to the limit point for gradient-descent methods is established in [73] by assuming diminishing step-sizes with either deterministic or stochastic gradient errors. The convergence for incremental methods is established in [138, 139] by assuming diminishing or constant step-sizes with deterministic or stochastic gradient errors. These results can be applied to distributed implementations under the local balance condition (7.14). If this condition is however not satisfied, then the stability and performance analysis of consensus and diffusion strategies become more demanding in comparison to standard gradient-descent implementations — see, e.g., the treatment in [8, 66].

The arguments in the previous section can also be used to interpret the two-phase transient behavior of consensus and diffusion strategies discovered in [67]. The gradient vector $\nabla J^{\text{pen}}(\mathcal{W})$ in (7.25) consists of two components: the first one is related the distributed cost $J^{\text{dist}}(\mathcal{W})$ and the second one is related to the regularization term. For small step-sizes with $\mu_k \ll 1$, and for *non-uniform* initialization where $\mathcal{W}_{-1} \notin \mathbb{S}$, i.e., $\mathcal{W}_{-1} \neq \mathbb{1}_N \otimes w_{-1}$ for any $w_{-1} \in \mathbb{R}^{M\times 1}$, the second component that is weighted by $\eta \gg 1$ dominates the first component. The first stage of the transient phase is therefore controlled by the second component in order to reduce the large penalty caused by the disagreement among the agents. After a number of iterations when $\mathcal{W}_i$ gets close enough to the agreement subspace such that the penalty is not significant any more, the gradient vector starts to reflect both components. The overall dynamics then enters the second stage of the transient phase.

7.A Proof of Lemma 7.3

The Hessian of $J^{\rm pen}(\mathcal{w})$ is given by

$$\nabla^2 J^{\rm pen}(\mathcal{w}) = \nabla^2 J^{\rm dist}(\mathcal{w}) + 4\eta(\mathcal{P} - \mathcal{A}\mathcal{P}) \tag{7.48}$$

where $\nabla^2 J^{\rm dist}(\mathcal{w})$ is given by (7.7). We argue in the sequel that the smallest eigenvalue of $\nabla^2 J^{\rm pen}(\mathcal{w})$ is lower bounded by some positive constant for *any* $\mathcal{w}$. Based on (7.48), this result can be established by showing that for any $x \in \mathbb{R}^{NM\times 1}$ with unit norm $\|x\| = 1$, and for any $\mathcal{w} \in \mathbb{R}^{NM\times 1}$,

$$\|x\|^2_{\nabla^2 J^{\rm pen}(\mathcal{W})} = \|x\|^2_{\nabla^2 J^{\rm dist}(\mathcal{W})} + 4\eta\|x\|^2_{\mathcal{P}-\mathcal{A}\mathcal{P}} \geq c \tag{7.49}$$

for some positive constant $c > 0$ that is *independent* of $\mathcal{w}$. Now note that the entire space $\mathbb{R}^{NM\times 1}$ can be orthogonally decomposed into two subspaces: $\mathbb{S}$ and its orthogonal compliment $\mathbb{S}^\perp$ [87]. Then, any vector $x \triangleq \mathrm{col}\{x_k\}_{k=1}^N$, $x_k \in \mathbb{R}^{M\times 1}$, with unit norm $\|x\| = 1$, can be uniquely and orthogonally decomposed as $x = \bar{x} + \tilde{x}$ with $\bar{x} \triangleq \mathrm{col}\{\bar{x}_k\}_{k=1}^N \in \mathbb{S}$ and $\tilde{x} \triangleq \mathrm{col}\{\tilde{x}_k\}_{k=1}^N \in \mathbb{S}^\perp$. Due to othogonality, it holds that $\|x\|^2 = \|\bar{x}\|^2 + \|\tilde{x}\|^2 = 1$. Since $\bar{x} \in \mathbb{S}$, it can be expressed as $\bar{x} = \mathbb{1}_N \otimes \bar{x}_o$, or, equivalently, $\bar{x}_k = \bar{x}_o$ for any k, and it holds that $\|\bar{x}\|^2 = N\|\bar{x}_o\|^2$. From Lemma 7.2, the nullspace of $\mathcal{P} - \mathcal{A}\mathcal{P}$ is $\mathbb{S}$. Since $\mathcal{P} - \mathcal{A}\mathcal{P}$ is symmetric, the range space of $\mathcal{P} - \mathcal{A}\mathcal{P}$ is then given by $\mathbb{S}^\perp$ [87]. Let us introduce the spectral decomposition for $\mathcal{P} - \mathcal{A}\mathcal{P}$ as follows:

$$\mathcal{P} - \mathcal{A}\mathcal{P} \triangleq \mathcal{U}\Lambda\mathcal{U}^{\mathsf{T}} \triangleq \begin{bmatrix}\mathcal{U}_1 & \mathcal{U}_2\end{bmatrix} \begin{bmatrix}\Lambda_1 & \\ & \Lambda_2\end{bmatrix} \begin{bmatrix}\mathcal{U}_1 & \mathcal{U}_2\end{bmatrix}^{\mathsf{T}} \tag{7.50}$$

where $\mathcal{U}$ is orthonormal and Λ is diagonal with nonnegative diagonal entries. In addition, $\Lambda_1 > 0$ corresponds to the range space of $\mathcal{P} - \mathcal{A}\mathcal{P}$, $\Lambda_2 = 0$ corresponds to the nullspace of $\mathcal{P} - \mathcal{A}\mathcal{P}$. Then, we have $\bar{x} \in \mathrm{Span}(\mathcal{U}_2)$ and $\tilde{x} \in \mathrm{Span}(\mathcal{U}_1)$.

Therefore,

$$\begin{aligned}\|x\|^2_{\mathcal{P}-\mathcal{A}\mathcal{P}} &= (\bar{x}+\tilde{x})^{\mathsf{T}}(\mathcal{P}-\mathcal{A}\mathcal{P})(\bar{x}+\tilde{x}) \\ &= \tilde{x}^{\mathsf{T}}(\mathcal{P}-\mathcal{A}\mathcal{P})\tilde{x} \\ &= \tilde{x}^{\mathsf{T}}\mathcal{U}_1\Lambda_1\mathcal{U}_1^{\mathsf{T}}\tilde{x} \\ &\geq \lambda_L\|\tilde{x}\|^2 \end{aligned} \tag{7.51}$$

where $\lambda_L > 0$ denotes the smallest entry on the diagonal of Λ_1 (which is the smallest nonzero eigenvalue of $\mathcal{P}-\mathcal{A}\mathcal{P}$). Meanwhile, we have from (7.7) that

$$\|x\|^2_{\nabla^2 J^{\rm dist}(\mathcal{W})} = \sum_{k=1}^{N} q_k\|x_k\|^2_{\nabla^2 J_k(w_k)} \geq q_m\lambda_{m,L}\|x_m\|^2 \tag{7.52}$$

where $J_m(\cdot)$ is strongly-convex by our previous assumptions and $\lambda_{m,L} > 0$ denotes the lower bound on $\nabla^2_m J_m(w_m)$. Then, we get from (7.49)–(7.52) that

$$\|x\|^2_{\nabla^2 J^{\rm pen}(\mathcal{W})} \geq 4\gamma\|x_m\|^2 + 4\eta\lambda_L\|\tilde{x}\|^2 \tag{7.53}$$

where $\gamma \triangleq q_m\lambda_{m,L}/4$. On one hand, it is straightforward from (7.52) to see that

$$\|x\|^2_{\nabla^2 J^{\rm pen}(\mathcal{W})} \geq 4\gamma\|x_m\|^2 \tag{7.54}$$

On the other hand, we have from (7.53) that

$$\begin{aligned}\|x\|^2_{\nabla^2 J^{\rm pen}(\mathcal{W})} &= 4\gamma\|x_m\|^2 + 4\eta\lambda_L\|\tilde{x}_m\|^2 + 4\eta\lambda_L(\|\tilde{x}\|^2 - \|\tilde{x}_m\|^2) \\ &\geq \min\{2\gamma, 2\eta\lambda_L\}(2\|x_m\|^2+2\|\tilde{x}_m\|^2)+4\eta\lambda_L(\|\tilde{x}\|^2-\|\tilde{x}_m\|^2) \\ &\overset{(a)}{\geq} \min\{2\gamma, 2\eta\lambda_L\}\|x_m-\tilde{x}_m\|^2 + 4\eta\lambda_L(\|\tilde{x}\|^2 - \|\tilde{x}_m\|^2) \\ &\overset{(b)}{=} \min\{2\gamma, 2\eta\lambda_L\}\|\bar{x}_m\|^2 + 4\eta\lambda_L\sum_{k\neq m}\|\tilde{x}_k\|^2 \\ &\overset{(c)}{=} \frac{\min\{\gamma,\eta\lambda_L\}}{N-1}\sum_{k\neq m} 2\|\bar{x}_k\|^2 + 2\eta\lambda_L\sum_{k\neq m}2\|\tilde{x}_k\|^2 \\ &\geq \frac{\min\{\gamma,\eta\lambda_L\}}{N-1}\sum_{k\neq m}(2\|\bar{x}_k\|^2 + 2\|\tilde{x}_k\|^2) \end{aligned}$$

$$
\begin{aligned}
&\overset{(d)}{\geq} \frac{\min\{\gamma, \eta\lambda_L\}}{N-1} \sum_{k\neq m} \|\bar{x}_k + \tilde{x}_k\|^2 \\
&\overset{(e)}{=} \frac{\min\{\gamma, \eta\lambda_L\}}{N-1} \sum_{k\neq m} \|x_k\|^2 \\
&\overset{(f)}{=} \frac{\min\{\gamma, \eta\lambda_L\}}{N-1} (1 - \|x_m\|^2)
\end{aligned} \tag{7.55}
$$

where steps (a) and (d) are due to the parallelogram law: $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$; steps (b) and (e) are because $x_k = \bar{x}_k + \tilde{x}_k$ for any k and $\|\tilde{x}\|^2 = \sum_{k=1}^N \|\tilde{x}_k\|^2$; step (c) is by using the fact that $\bar{x}_k = \bar{x}_o$ for all k; and step (f) is due to the unit norm: $\sum_{k=1}^N \|x_k\|^2 = 1$. Combining (7.54) and (7.55), and letting $c_1 = 4\gamma > 0$ and $c_2 = \min\{\gamma, \eta\lambda_L\}/(N-1) > 0$, we obtain

$$
\begin{aligned}
\|x\|^2_{\nabla^2 J^{\text{pen}}(\mathcal{W})} &\geq \max\left\{c_1\|x_m\|^2, c_2(1-\|x_m\|^2)\right\} \\
&\geq \min\{c_1, c_2\} \cdot \max\{\|x_m\|^2, 1-\|x_m\|^2\} \\
&\geq \frac{\min\{c_1, c_2\}}{2}
\end{aligned} \tag{7.56}
$$

where we used the fact that $\max_{0\leq y\leq 1}\{y, 1-y\} \geq 1/2$ because $0 \leq \|x_m\|^2 \leq 1$. Since $\min\{c_1, c_2\}/2 > 0$ is a constant that is independent of w, result (7.49) holds.

References

[1] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

[2] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 674–690, Aug. 2011.

[3] A. Bertrand and M. Moonen, "Low-complexity distributed total least squares estimation in ad hoc sensor networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4321–4333, Aug. 2012.

[4] O. N. Gharehshiran, V. Krishnamurthy, and G. Yin, "Distributed energy-aware diffusion least mean squares: Game-theoretic learning," *IEEE J. Sel. Top. Signal Process.*, vol. 7, no. 5, pp. 1–16, Oct. 2013.

[5] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, July 2008.

[6] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[7] L. Xiao, S. Boyd, and S. Lall, "A space-time diffusion scheme for peer-to-peer least-squares estimation," in *Proc. ACM/IEEE Int. Conf. Inform. Process. Sensor Networks*, Nashville, TN, Apr. 2006, pp. 168–176.

[8] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[9] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "A collaborative training algorithm for distributed learning," *IEEE Trans. Inf. Theory*, vol. 55, no. 4, pp. 1856–1871, Apr. 2009.

[10] S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections: A unifying framework for linear and nonlinear classification and regression tasks," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 97–123, Jan. 2011.

[11] S. Chouvardas, K. Slavakis, and S. Theodoridis, "Adaptive robust distributed learning in diffusion sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4692–4707, Oct. 2011.

[12] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

[13] D. Gesbert, S. G. Kiani, A. Gjendemsjo, and G. E. Oien, "Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks," *Proc. IEEE*, vol. 95, no. 12, pp. 2393–2409, Dec. 2007.

[14] P. Di Lorenzo and S. Barbarossa, "A bio-inspired swarming algorithm for decentralized access in cognitive radio," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6160–6174, Dec. 2011.

[15] K. M. Passino, "Biomimicry of bacterial foraging for distributed optimization and control," *IEEE Control Syst. Mag.*, vol. 22, no. 3, pp. 52–67, June 2002.

[16] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: Algorithms and theory," *IEEE Trans. Autom. Control*, vol. 51, no. 3, pp. 401–420, Mar. 2006.

[17] S. Barbarossa and G. Scutari, "Bio-inspired sensor network design," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 26–35, May 2007.

[18] F. S. Cattivelli and A. H. Sayed, "Modeling bird flight formations using diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2038–2051, May 2011.

[19] S-Y. Tu and A. H. Sayed, "Mobile adaptive networks," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 649–664, Aug. 2011.

[20] J. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," *IEEE Trans. Autom. Control*, vol. 29, no. 1, pp. 42–50, Jan. 1984.

[21] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, Sept. 1986.

[22] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *System Control Lett.*, vol. 53, no. 9, pp. 65–78, Sept. 2004.

[23] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, June 2006.

[24] P. Braca, S. Marano, and V. Matta, "Running consensus in wireless sensor networks," in *Proc. Int. Conf. Inform. Fusion (FUSION)*, Cologne, Germany, June - July 2008, pp. 1–6.

[25] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[26] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, Jan. 2009.

[27] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 772–790, Aug. 2011.

[28] O. Hlinka, O. Sluciak, F. Hlawatsch, and P. M. Djuric, "Likelihood consensus and its application to distributed particle filtering," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4334–4349, Aug. 2012.

[29] D. P. Bertsekas, "A new class of incremental gradient methods for least squares problems," *SIAM J. Optim.*, vol. 7, no. 4, pp. 913–926, 1997.

[30] A. Nedic and D. P. Bertsekas, "Incremental subgradient methods for non-differentiable optimization," *SIAM J. Optim.*, vol. 12, no. 1, pp. 109–138, 2001.

[31] M. G. Rabbat and R. D. Nowak, "Quantized incremental algorithms for distributed optimization," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 798–808, Apr. 2005.

[32] D. Blatt, A. Hero, and H. Gauchman, "A convergent incremental gradient method with constant step size," *SIAM J. Optim.*, vol. 18, no. 1, pp. 29–51, Feb. 2007.

[33] C. G. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 223–229, Aug. 2007.

[34] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. Towfic, "Diffusion strategies for adaptation and learning over networks," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 155–171, May 2013.

[35] A. H. Sayed, "Diffusion adaptation over networks," in *Academic Press Library in Signal Processing*, R. Chellapa and S. Theodoridis, Eds., vol. 3, pp. 323–454. Academic Press, Elsevier, 2014. Also available as arXiv:1205.4220v2, May 2012.

[36] X. Zhao and A. H. Sayed, "Performance limits for distributed estimation over LMS adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5107–5124, Oct 2012.

[37] S. Kar and J. M. F. Moura, "Sensor networks with random links: Topology design for distributed consensus," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3315–3326, July 2008.

[38] T. C. Aysal, A. D. Sarwate, and A. G. Dimakis, "Reaching consensus in wireless networks with probabilistic broadcast," in *Proc. Allerton Conf. Commun., Control, Comput.*, Allerton House, UIUC, IL, Sept. and Oct. 2009, pp. 732–739.

[39] T. C. Aysal, M. E. Yildiz, and A. Scaglione, "Broadcast Gossip algorithms for consensus," *IEEE Trans. Signal Process.*, vol. 57, pp. 2748–2761, 2009.

[40] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1383–1400, Mar. 2010.

[41] D. Jakovetic, J. Xavier, and J. M. F. Moura, "Weight optimization for consensus algorithms with correlated switching topology," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3788–3801, July 2010.

[42] D. Jakovetic, J. Xavier, and J. M. F. Moura, "Cooperative convex optimization in networked systems: augmented Lagranian algorithms with directed Gossip communication," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3889–3902, Aug. 2011.

[43] C. G. Lopes and A. H. Sayed, "Diffusion adaptive networks with changing topology," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, Mar./Apr. 2008, pp. 3285–3288.

[44] N. Takahashi and I. Yamada, "Link probability control for probabilistic diffusion least-mean squares over resource-constrained networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, Mar. 2010, pp. 3518–3521.

[45] X. Zhao and A. H. Sayed, "Attaining optimal batch performance via distributed processing over networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vancouver, Canada, May 2013, pp. 1–5.

[46] A. H. Sayed, *Adaptive Filters*, Wiley, NJ, 2008.

[47] R. Abdolee and B. Champagne, "Diffusion LMS algorithms for sensor networks over non-ideal inter-sensor wireless channels," in *Proc. IEEE Int. Conf. Dist. Comput. Sensor Systems (DCOSS)*, Barcelona, Spain, June 2011, pp. 1–6.

[48] A. Khalili, M. A. Tinati, A. Rastegarnia, and J. A. Chambers, "Steady-state analysis of diffusion LMS adaptive networks with noisy links," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 974–979, Feb. 2012.

[49] S-Y. Tu and A. H. Sayed, "Adaptive networks with noisy links," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Houston, TX, Dec. 2011, pp. 1–5.

[50] G. Mateos, I. D. Schizas, and G. B. Giannakis, "Performance analysis of the consensus-based distributed LMS algorithm," *EURASIP J. Advances Signal Process.*, vol. 2009, pp. 1–19, 2009, Article ID 981030, doi:10.1155/2009/981030.

[51] X. Zhao and A. H. Sayed, "Clustering via diffusion adaptation over networks," in *Proc. Int. Workshop Cognit. Inform. Process. (CIP)*, Baiona, Spain, May 2012, pp. 1–6.

[52] S.-Y. Tu and A. H. Sayed, "Distributed decision-making over adaptive networks," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1054–1069, Mar. 2014.

[53] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over netowrks," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4129–4144, Aug. 2014.

[54] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS over multitask networks," *submitted for publication*, Apr. 2014. Also available at arXiv:1404.6813v1 [cs.SY].

[55] J. Liu, M. Chu, and J. E. Reich, "Multitarget tracking in distributed sensor networks," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 36–46, May 2007.

[56] X. Zhang, "Adaptive control and reconfiguration of mobile wireless sensor networks for dynamic multi-target tracking," *IEEE Trans. Autom. Control*, vol. 56, no. 10, pp. 2429–2444, Oct. 2011.

[57] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, New York, 2001.

[58] I. Francis and S. Chatterjee, "Classification and estimation of several multiple regressions," *The Annals of Statistics*, vol. 2, no. 3, pp. 558–561, 1974.

[59] X.-R. Li and Y. Bar-Shalom, "Multiple-model estimation with variable structure," *IEEE Trans. Autom. Control*, vol. 41, no. 4, pp. 478–493, Apr. 1996.

[60] V. Cherkassky and Y. Ma, "Multiple model regression estimation," *IEEE Trans. Neural Netw.*, vol. 16, no. 4, pp. 785–798, July 2005.

[61] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, Oxford, UK, 4th edition, 2009.

[62] L. Jacob, F. Bach, and J.-P. Vert, "Clustered multi-task learning: A convex formulation," in *Proc. Neural Inform. Process. Systems. (NIPS)*, Vancouver, Canada, Dec. 2008, pp. 1–8.

[63] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks — Part I: Sequential node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5277–5291, Oct. 2010.

[64] N. Bogdanovic, J. Plata-Chaves, and K. Berberidis, "Distributed diffusion-based LMS for node-specific parameter estimation over adaptive netowrks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 7223–7227.

[65] J. Chen and A. H. Sayed, "Distributed Pareto optimization via diffusion strategies," *IEEE J. Sel. Top. Signal Process.*, vol. 7, no. 2, pp. 205–220, Apr. 2013.

[66] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, July 2014.

[67] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks — Part I: Transient analysis," *submitted for publication*, Dec. 2013. Also available on arXiv:1312.7581v2 [cs.MA].

[68] X. Zhao and A. H. Sayed, "Asynchronous adaptation and learning over networks — Part I: Modeling and stability analysis," *submitted for publication*, Dec. 2013. Also available on arXiv:1312.5434v2 [cs.SY].

[69] K. Kreutz-Delgado, "The complex gradient operator and the CR-calculus," 2009, arXiv:0906.4835 [math.OC].

[70] Z. J. Towfic, J. Chen, and A. H. Sayed, "On distributed online classification in the midst of concept drifts," *Neurocomputing*, vol. 112, pp. 138–152, July 2013.

[71] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, Cambirdge, UK, 2004.

[72] B. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.

[73] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM J. Optim.*, vol. 10, no. 3, pp. 627–642, 2000.

[74] M. H. DeGroot and M. J. Schervish, *Probability and Statistics*, Addison-Wesley, Boston, MA, 4 edition, 2011.

[75] T. Adali, P. J. Schreier, and L. L. Scharf, "Complex-valued signal processing: The proper way to deal with impropriety," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5101–5125, Nov. 2011.

[76] A. van den Bos, "Complex gradient and Hessian," *IEE Proc.-Vis. Image Signal Process.*, vol. 141, no. 6, pp. 380–383, Dec. 1994.

[77] X. Zhao and A. H. Sayed, "Asynchronous adaptation and learning over networks — Part II: Performance analysis," *submitted for publication*, Dec. 2013. Also available on arXiv:1312.5438v2 [cs.SY].

[78] X. Zhao, S.-Y. Tu, and A. H. Sayed, "Diffusion adaptation over networks under imperfect information exchange and non-stationary data," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3460–3475, July 2012.

[79] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, PA, 1994.

[80] J. A. Bondy and U. S. R. Murty, *Graph Theory*, Springer, 2008.

[81] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *J. Mach. Learn. Res.*, vol. 11, pp. 985–1042, Sept. 2010.

[82] S. U. Pillai, T. Suel, and S. Cha, "The Perron-Frobenius theorem: Some of its applications," *IEEE Signal Process. Mag.*, vol. 22, no. 2, pp. 62–75, Mar. 2005.

[83] R. H. Koning, H. Neudecker, and T. Wansbeek, "Block Kronecker products and the vecb operator," *Linear Algebra Appl.*, vol. 149, pp. 165–184, Apr. 1991.

[84] J. W. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circuits Syst.*, vol. 25, no. 9, pp. 772–781, Sept. 1978.

[85] H. V. Henderson and S. R. Searle, "The vec-permutation matrix, the vec operator and Kronecker products: A review," *Linear Multilinear Algebra*, vol. 9, no. 4, pp. 271–288, 1981.

[86] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

[87] A. J. Laub, *Matrix Analysis for Scientists and Engineers*, SIAM, PA, 2005.

[88] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, MA, 1990.

[89] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge Univ. Press, Cambridge, UK, 1991.

[90] X. Zhao and A. H. Sayed, "Asynchronous adaptation and learning over networks — Part III: Comparison analysis," *submitted for publication*, Dec. 2013. Also available on arXiv:1312.5439v2 [cs.SY].

[91] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous Multivariate Distributions Vol. 1: Models and Applications*, Wiley, New York, 2nd edition, 2000.

[92] R. J. Connor and J. E. Mosimann, "Concepts of independence for proportions with a Generalization of the Dirichlet distribution," *J. Am. Stat. Assoc.*, vol. 64, no. 325, pp. 194–206, Mar. 1969.

[93] J. Aitchison, "A general class of distributions on the simplex," *J. R. Statist. Soc. B*, vol. 47, no. 1, pp. 136–146, 1985.

[94] O. E. Barndorff-Nielsen and B. Jorgensen, "Some parametric models on the simplex," *J. Multivariate Anal.*, vol. 39, no. 1, pp. 106–116, Oct. 1991.

[95] T.-T. Wong, "Generalized Dirichlet distribution in Bayesian analysis," *Appl. Math. Comput.*, vol. 97, no. 2-3, pp. 165–181, Dec. 1998.

[96] R. K. S. Hankin, "A generalization of the Dirichlet distribution," *J. Stat. Soft.*, vol. 33, no. 11, pp. 1–18, Feb. 2010.

[97] W.-Y. Chang, R. D. Gupta, and D. St. P. Richards, "Structural properties of the generalized Dirichlet distributions," *Contemp. Math.*, vol. 516, pp. 109–124, 2010.

[98] T.-T. Wong, "Parameter estimation for generalized Dirichlet distributions from the sample estimates of the first and the second moments of random variables," *Comput. Stat. Data Anal.*, vol. 54, no. 7, pp. 1756–1765, July 2010.

[99] S. Favaro, G. Hadjicharalambous, and I. Prunster, "On a class of distributions on the simplex," *J. Stat. Plan Infer.*, vol. 141, no. 9, pp. 2987–3004, Sept. 2011.

[100] J. Aitchison and S. M. Shen, "Logistic-Normal distributions: Some properties and uses," *Biometrika*, vol. 67, no. 2, pp. 261–272, Aug. 1980.

[101] P. J. Lenk, "The Logistic Normal distribution for Bayesian, nonparametric, predictive densities," *J. Am. Stat. Assoc.*, vol. 83, no. 402, pp. 509–516, June 1988.

[102] V. Seshadri, "General exponential models on the unit simplex and related multivariate inverse Gaussian distributions," *Stat. Probabil. Lett.*, vol. 14, no. 5, pp. 385–391, July 1992.

[103] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, Jan. 2003.

[104] A. Gelman, "Method of moments using Monte Carlo simulation," *J. Comput. Graph Stat.*, vol. 4, no. 1, pp. 36–54, Feb. 1995.

[105] A. Khalili, M. A. Tinati, A. Rastegarnia, and J. A. Chambers, "Transient analysis of diffusion least-mean squares adaptive networks with noisy channels," *Int. J. Adapt. Control Signal Process.*, Sept. 2011, doi: 10.1002/acs.1279.

[106] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[107] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *Proc. ACM/IEEE Int. Conf. Inform. Process. Sensor Networks*, Los Angeles, CA, Apr. 2005, pp. 63–70.

[108] N. Takahashi, I. Yamada, and A. H. Sayed, "Diffusion least-mean squares with adaptive combiners: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4795–4810, Sept. 2010.

[109] J. Arenas-Garcia, V. Gomez-Verdejo, and A. R. Figueiras-Vidal, "New algorithms for improved adaptive convex combination of LMS transversal filters," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 6, pp. 2239–2249, Dec. 2005.

[110] D. Mandic, P. Vayanos, C. Boukis, B. Jelfs, S. L. Goh, T. Gautama, and T. Rutkowski, "Collaborative adaptive learning using hybrid filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Honolulu, HI, Apr. 2007, pp. 921–924.

[111] M. T. M. Silva and V. H. Nascimento, "Improving the tracking capability of adaptive filters via convex combination," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3137–3149, July 2008.

[112] S. S. Kozat and A. C. Singer, "A performance-weighted mixture of LMS filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Apr. 2009, pp. 3101–3104.

[113] R. Candido, M. T. M. Silva, and V. H. Nascimento, "Transient and steady-state analysis of the affine combination of two adaptive filters," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4064–4078, Aug. 2010.

[114] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion recursive least-squares for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1865–1877, May 2008.

[115] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, Cambridge, UK, 1985.

[116] T. Y. Al-Naffouri and A. H. Sayed, "Transient analysis of data-normalized adaptive filters," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 639–652, Mar. 2003.

[117] S-Y. Tu and A. H. Sayed, "Optimal combination rules for adaptation and learning over netowrks," in *Proc. IEEE Int. Workshop Comput. Advances Multi-Sensor Adapt. Process. (CAMSAP)*, San Juan, Puerto Rico, Dec. 2011, pp. 317–320.

[118] F. S. Cattivelli and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 2069–2084, Sept. 2010.

[119] X. Zhao and A. H. Sayed, "Distributed clustering and learning over networks," *submitted for publication*, Sept. 2014.

[120] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer-Verlag, New York, 2003.

[121] A. N. Shiryaev, *Probability*, Springer, Nauka, Moscow, 1980.

[122] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958.

[123] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer, New York, 2nd edition, 1998.

[124] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 2, Wiley, New York, 2nd edition, 1995.

[125] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 1, Wiley, New York, 2nd edition, 1995.

[126] P. G. Moschopoulos, "The distribution of the sum of independent Gamma random variables," *Ann. Inst. Statist. Math.*, vol. 37, no. A, pp. 541–544, 1985.

[127] G. K. Karagiannidis, N. C. Sagias, and T. A. Tsiftsis, "Closed-form statistics for the sum of squared nakagami-m variates and its applications," *IEEE Trans. Commun.*, vol. 54, no. 8, pp. 1353–1359, Aug. 2006.

[128] J. P. Imhof, "Computing the distribution of quadratic forms in normal variables," *Biometrika*, vol. 48, no. 3-4, pp. 419–426, Dec. 1961.

[129] J. Sheil and I. O'Muircheartaigh, "The distribution of non-negative quadratic forms in normal variables," *J. Roy. Stat. Soc. C-App.*, vol. 26, no. 1, pp. 92–98, 1977.

[130] H. Liu, Y. Tang., and H. H. Zhang, "A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables," *Comput. Stat. Data An.*, vol. 53, no. 4, pp. 853–856, Feb. 2009.

[131] P. Duchesne and P. L. De Micheaux, "Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods," *Comput. Stat. Data An.*, vol. 54, no. 4, pp. 858–862, Apr. 2010.

[132] H.-T. Ha and S. B. Provost, "An accurate approximation to the distribution of a linear combination of non-central chi-square random variables," *REVSTAT Stat. J.*, vol. 11, no. 3, pp. 231–254, Nov. 2013.

[133] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 2nd edition, 2006.

[134] P. Li, T. J. Hastie, and K. W. Church, "Nonlinear estimators and tail bounds for dimension reduction in ℓ_1 using Cauchy random projections," *J. Mach. Learn. Res.*, vol. 8, pp. 2497–2532, Oct. 2007.

[135] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 2 edition, 1995.

[136] Z. J. Towfic and A. H. Sayed, "Adaptive penalty-based distributed stochastic convex optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3924–3938, Aug. 2014.

[137] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," Biometrika, vol. 57, no. 1, pp. 97–109, Apr. 1970.

[138] D. P. Bertsekas, "Incremental proximal methods for large scale convex optimization," *Math. Program.*, vol. 129, no. 2, pp. 163–195, Oct. 2011.

[139] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Incremental stochastic subgradient algorithms for convex optimization," *SIAM J. Optim.*, vol. 20, no. 2, pp. 691–717, 2009.