Inference and Optimization over Networks: Communication Efficiency and Optimality

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Anit Kumar Sahu

B.Tech., Electronics and Electrical Communication Engineering, IIT Kharagpur
M. Tech., Telecommunication Systems Engineering, IIT Kharagpur

Carnegie Mellon University
Pittsburgh, PA

December 2018

Dedicated to you, Mama for being my dream weaver and for sacrificing your academic career after my birth. Here is the result of your sacrifices and for making a researcher out of me.....

Abstract

We study distributed inference, learning and optimization in scenarios which involve networked entities in time-varying and random networks, which are ad-hoc in nature. In this thesis, we propose distributed recursive algorithms where the networked entities simultaneously incorporate locally sensed information and information obtained from the neighborhood. The class of distributed algorithms proposed in this thesis encompasses distributed estimation, distributed composite hypothesis testing and distributed optimization. The central theme of the scenarios considered involve systems constrained by limited on board batteries and hence constrained by limited sensing, computation and extremely limited communication resources. A typical example of such resource constrained scenarios being distributed data-parallel machine learning systems, in which the individual entities are commodity devices such as cellphones.

Due to the inherent ad-hoc nature of the aforementioned setups, in conjunction with random environments render these setups central coordinator-less. Keeping in mind the resource constrained nature of such setups, we propose distributed inference and optimization algorithms which characterize the interplay between communication, computation and optimality, while allowing for heterogeneity among clients in terms of objectives, data collection and statistical dependencies.

With massive data, models for learning and optimization have been getting more and more complex to the extent of being almost analytically intractable. In such models, obtaining gradients for the associated loss function is very expensive and potentially intractable due to the lack of a closed form for the loss function. A major thrust of this thesis is gradient free zeroth order optimization which encompasses distributed setups which exhibit data parallelism and also potentially analytically intractable loss functions. On top of gradient free optimization, in this thesis we also study projection free zeroth order methods for constrained optimization.

The techniques developed in this thesis are generic and are of independent interest in classical fields such as stochastic approximation, statistical decision theory and optimization.

Acknowledgements

First and foremost, I would like to extend my gratitude to my advisor Soummya Kar. Since the day I joined him as an intern as a naive undergrad in summer of 2012, he has shaped me as a researcher in countless ways. I feel extremely privileged and lucky to be his *full*[1] student. Soummya's penchant for rigor and the *devil lies in the details* quotes have fundamentally changed the way how I approach proofs. His enthusiasm, innate curiosity, endless efforts and encouragement at every stage made this entire experience a very rewarding one. I am especially grateful to him, for how approachable he was throughout to discuss the silliest of ideas I had, for trusting me through my darkest days of depression, for letting me explore directions of my choice and for being my unrelenting advocate.

I would like to thank my committee members Jose Moura, Michael Rabbat and David Woodruff for their insightful comments, valuable criticism and suggestions. I would like to thank Jack Schaeffer and Larry Wasserman for being fantastic instructors to my favorite courses that I have taken at CMU. I would like to thank Jelena Kovacevic for her mental support during my days of depression.

I would like to extend my thanks to my collaborators Jose Moura, Vincent Poor, Dusan Jakovetic, Dragana Bajovic and Manzil Zaheer. I started this journey reading and getting inspired by Dusan and Dragana's papers. It was a dream come true for me to be able to collaborate and exchange ideas with them. Manzil's ability to focus on the story while keeping the details in sight made it a really fruitful experience collaborating with him.

I would also like to thank the ECE staff members - Carol Patterson, Claire Baurele, Allison Ervin, Samantha Goldstein and Nathan Snizaski for helping me in various ways many times and handling all my conference travels. A special thanks to Chuck and Al from the CMU shuttle and escort service for getting me home in really late hours all these years.

My special thanks to my "PH-B level" mates over the years- Sergio Pequito, Brian Swenson, Javad Mohammadi, Yuan Chen, Yaoqing Yang, Hadi Amini, Kyri Baker, June Zheng, Nipun Popli, Joya Deri, Nicola Forti, Walter Lucia, Jonathan Donadee, JY Joo, Kyle Anderson, Nikos Arechiga, Milos Cvetkovic, Subhro Das, Stephen Kruzick, Liangyan Gui, Joel Harley, Jonathan Mei, Evgeny Toropov, Satwik Kottur, Mark Cheung, Shanghang Zhang, Steven Aday, Vishwanath S R V, Joao Saude, Jian Wang, Vinay Prabhu, Andrew Hsu, Sean Weerakkody, John Costanzo, Carmel Fiscko, Paul Griffioen, Vincent Mornardo, Andrew Hsu and so many more for helping me stay sane and keep my chin up through this journey.

I would like to thank Christine Asenjo, Julian Asenjo and Ian Asenjo whom I met through the host family program for making me feel as if home had suddenly moved 10000 miles closer for me. Thanks for making me feel at home by making me a part of many events.

Thanks to the Crazy Mocha on Ellsworth Avenue and CTR on Walnut Street for providing me with the much needed caffeine at weird times to keep my nocturnal *owl* self going strong. Thanks to the TV show *The Mentalist* for making it seemingly okay to work/sleep/eat from a couch. Thanks to Grubhub, for letting me access to food at the weirdest times.

Thanks to Jon Francis aka *St. Jon* for being a sounding board for me since I have known him and for the much needed Friday night getaways. Thanks to Manzil aka *Sir*, for all the interesting technical discussions on just about any topic, for all the impeccable real-life advice, for all the restaurant suggestions and for cheering me up all the times when I was down. Thanks to Abhijit aka *senior doctor Abhijit* and Subhro aka *God Das*, for making me feel at home in Pittsburgh with all the advice and dinners. To Yaoqing aka *Prof. X*, thanks for all the discussions regarding proofs. Thanks to Jonathan aka *JMei*, for all the fun times, for all the support in the tough times.

I would like to thank my father, Banabihari Sahu, for his impeccable trust in me, for letting me dream and for supporting me at every step.

Last but not the least, I would like to thank my mother, Kabita Sahu from the bottom of my heart, for imbibing in me the essence of hard work, for making me a fighter to the core and for giving wind to my dreams and aspirations. Mama, you sacrificed your career in academia for me. I hope this thesis and your unlimited contributions and struggles to mold me into what I am today makes you live your would have been doctoral experience.

This work was partially supported by NSF through grants ECCS-1306128, ECCS-1408222 and CCF-1513936.

[1] This was coined by Sergio so as to refer to me as I was the first student of Soummya to be advised just by him.

Contents

Bibliography **185**

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Distributed data processing techniques have been increasingly employed to solve problems pertaining to optimization and statistical inference. With massive computing resources that are available at scale, and ever growing sizes of data sets, it becomes highly desirable, if not necessary, to distribute the task among multiple machines or multiple cores. The benefits of splitting the task into smaller subtasks are multi-pronged, namely, it makes the problem at hand, scalable, parallelized and fast. In the context of distribution stochastic optimization, several methods (see, for example Zhang et al. (2013b,a); Heinze et al. (2016); Ma et al. (2015); Recht et al. (2011)) have been proposed which exhibit impressive performance in platforms such as Mapreduce and Spark. The aforementioned methods, though highly scalable, are designed for master-worker or similar types of architectures. That is, they require the presence of a master node, i.e., a central coordinator which is tasked with splitting the dataset by data points (batches) or by features among worker nodes and enabling the read/write operations of the iterates of the worker nodes so as to ensure information fusion across the worker nodes. However, with several emerging applications, master-worker type architectures may not be feasible or desirable due to physical constraints.

In this thesis, we are interested in systems and applications where the entire data is not available at a central/master node, is sensed in a streaming fashion and is intrinsically distributed across the worker nodes[1]. Such scenarios arise, e.g., in systems which involve Internet of Things (IoT). For example, a smart campus with sensors of various kinds, a smart building or monitoring a large scale industrial plant. Therein, a network of large number of heterogeneous entities (usually, geographically spread) connected in a arbitrary network structure individually perform sensing for data arriving in a streaming fashion. The sensing devices have limited communication capabilities owing to on board power constraints and harsh environments. A typical IoT framework is characterized by a heterogeneous network of entities without a central coordinator, where entities have localized knowledge and can exchange information among each other through an arbitrary pre-specified communication graph. Furthermore, the data samples arrive in a streaming fashion. The ad-hoc nature of the IoT framework necessitates the information exchange in a crafted manner.

The goal of the thesis is to study learning, inference and optimization in the context of aforementioned ad-hoc networked systems deployed in random environments with the networked entities being heterogeneous in nature and have limited resources in terms of communication, computation and sensing. We start by highlighting the key aspects of the inference and optimization algorithms studied in this thesis.

[1] We use worker nodes, agents, entities and nodes interchangeably through out this thesis.

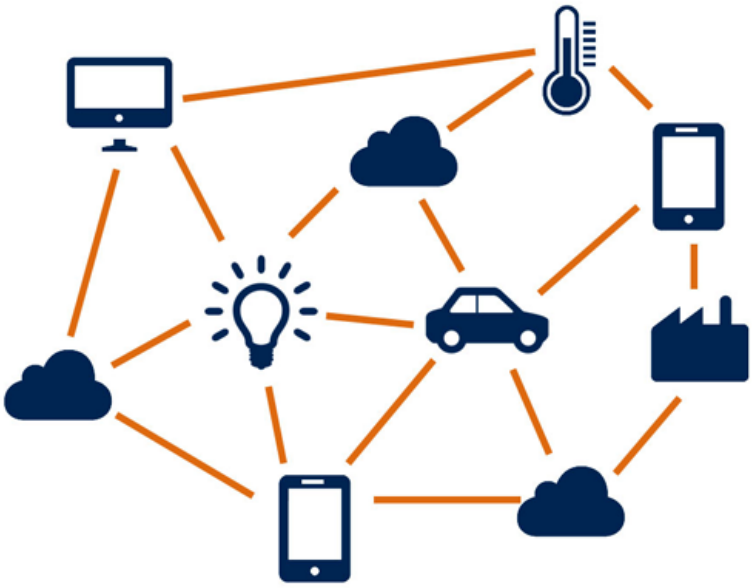


Figure 1.1: A typical Internet of Things setup

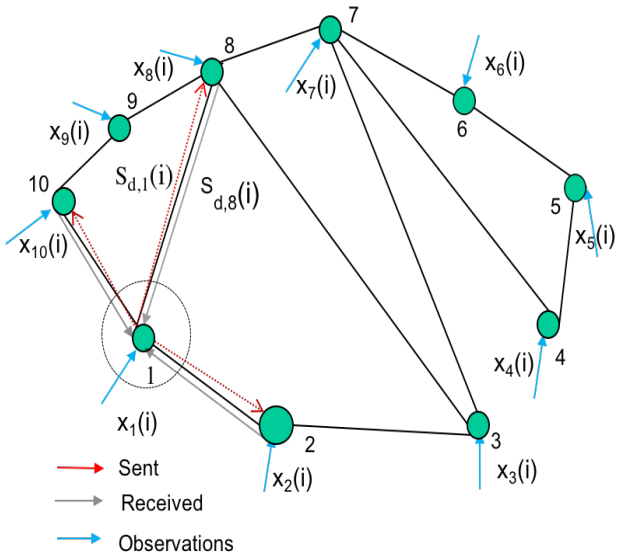


Figure 1.2: Distributed Architecture without a central node

Distributed Information Processing A major focus of this thesis is to develop and characterize algorithms for distributed information processing in architectures involving entities which are inter connected without a central coordinator. Figure 1.2 illustrates the main characteristics of the distributed setup that we consider in this thesis. In particular, our setup allows for both stored data and streaming data at the networked entities. The information exchange in the distributed setup as depicted in figure 1.2 is restricted to neighborhoods of the networked entities. In addition to the local neighborhood constrained information exchange, the information exchanged is limited to a sufficient statistic of the data and raw data is never exchanged.

Heterogeneous Entities The algorithms that we develop and characterize in this thesis are built around distributed information processing setups which involve entities that depict varying levels of heterogeneity. A setup, i.e., Internet of Things where heterogeneity is exhibited in almost all instances of its deployment is depicted in Figure 1.1. A common trait in terms of heterogeneity which ties all our algorithms is accounting for different ways in which the entities collect their data, i.e., the sensing models and the associated statistical characteristics. In addition to the heterogeneous sensing models, the distributed algorithms developed in this thesis provide for heterogeneous objectives and therein, heterogeneous local communication schemes, i.e., allowance for messages of different dimension exchanged among the entities.

Recursive Algorithms The ad-hoc nature of the distributed architecture in addition to the decentralized data configuration considered in this thesis necessitates collaborations among the entities so as to accomplish a task for the network as a whole. In light of the collaboration among the agents on top of the streaming data setting, the algorithms developed in this thesis are recursive in nature. At the heart of the distributed algorithms developed in this thesis, lies the interplay of the information obtained by each entity through its local data collection and the information obtained from its neighborhood and a carefully tuned mixing of the sensed information and the information obtained from the neighborhood. The global behavior emerging from such mixing leads to non-trivial analytical characterization. The techniques developed in this thesis to characterize the arising global behavior provides intuition and understanding towards tuning of local entity level interactions to accomplish a global objective.

1.2 Contributions

In the following, we first highlight the key thematic contributions of this thesis.

- **Ad-hoc Deployment and Communication** In most applications of interest, the first challenge comes from the ad-hoc nature of the networked entities. By ad-hoc in terms of deployment we mean, the sudden inclusion of an entity into the network or the sudden shut down of an entity in the network. By ad-hoc communication, we mean the uncertainty in the availability of communication links in the network. Moreover, the quality of a communication link connecting two entities is contingent on the available on board battery power of the sending entity. It is well known that the connectivity of a network is crucial to guarantee any reasonable performance of a distributed algorithm. In order to encapsulate the ad-hoc nature of the networked setup and the successive degradation of unreliable communication links over epochs, the communication protocols around which our proposed algorithms are built allow for non-identically distributed communication graphs. Technically speaking, the communication protocols employed in this work goes beyond typical independent and identically distributed (i.i.d.) construction of communication topologies. The protocols employed in this thesis not only incorporate epoch wise degradation of communication links but also allow for the communication graphs to be disconnected at all times with the requirement that they are connected only on average.

- **Communication Efficiency** Distributed algorithms for statistical inference and optimization in the resource constrained networked frameworks are characterized by a central coordinator-less recursive procedures, where each entity in the network maintains its own estimate or optimizer for the problem at hand. Resource constrained networked frameworks such as IoT, involve entities which sense information, perform on board computation and send/receive information from neighboring entities.

It is noteworthy that the battery power required for communication is a couple of orders higher than the power required for computation and sensing. Over the course of a particular distributed task, the quality of communication also degrades over iterations. Thus, so as to replicate real life systems, the communication protocol employed in any distributed scheme should be able to reflect the degrading quality of communication. Hence, it is of particular interest to construct distributed schemes which are frugal in terms of communication without compromising on the performance of the associated algorithm. However, distributed recursive algorithms are plagued by the need to communicate at each time and requiring apriori knowledge of the dimensions of the parameter. In this thesis, we focus on communication efficiency while keeping optimality in mind. In other words, we explore the paradigm as to how much of redundancy can be removed as far as the messages exchanged are concerned without losing in terms of the associated performance metrics.

- **Black Box Optimization** A distributed architecture when deployed out of a data center is characterized by entities which depict *plug and play* behavior. Technically speaking, entities join and leave the networked setup at unaccounted times. With the aforementioned *plug and play* behavior of entities, it is resource wise prohibitive to communicate global model information among the entities, which in conjunction with cold starts for a new entity in the network makes model based optimization practically difficult to implement. Hence, schemes which require minimal knowledge of the model ,i.e., black box schemes where only evaluations of the model are available are particularly attractive avenues in such setting. In this thesis, we study and develop distributed zeroth order optimization schemes which takes into account computation and communication efficiency. In particular, we establish convergence rates of such stochastic optimization algorithms through non-asymptotic analysis and thus explicitly characterizing the performance through the algorithm parameters.

The contributions of this thesis can be further categorized into three technical areas, namely, distributed detection, distributed estimation distributed optimization. We now summarize contributions from the aforementioned technical areas.

Sequential Testing In most applications built around networked setups involving resource constrained entities, it is of particular interest to tune the algorithm so as to achieve reasonable performance using minimal resources. Thus, it is of particular interest to develop algorithms which are sample efficient. For example, classical fixed sample size simple hypothesis testing uses 50% more number of samples than a testing scheme where the number of samples are not decided apriori. Hence, extending sequential testing schemes to networked setups is of particular interest for sample efficiency. In this thesis, we developed and analyzed distributed sequential detection algorithms for which we characterized the algorithm parameters and the associated stopping time distribution. In that, we characterized the performance of the proposed algorithm in terms of the connectivity of the graph induced by the inter-agent communication topology.

Recursive Composite Hypothesis Testing In the problem of testing a set of hypotheses involving composite hypothesis in a networked setup the onus is on achieving reasonable detection performance by utilizing as fewer resources as possible, which includes data samples, communication and sensing energy. The major challenge in composite testing is to balance the search for the approximate parameterization of the hypothesis which is in force, while at the same time deciding as to which hypothesis is true. Given the resource constrained nature of setups considered in this thesis, classical composite hypothesis testing procedures which first exploit the samples to find the approximate parameterization of the hypothesis in

force followed by framing a detection procedure are prohibitive. The resource constraints pose a challenging question in the lines of whether a detection procedure can be formulated which estimates the underlying parameter while deciding as to which hypothesis is in force parallely. In this thesis, we answer this question positively by developing an online recursive detection procedure for which we characterize the algorithm parameters for which exponentially decaying probabilities of errors can be obtained in the large sample limit.

Communication Efficient Inference and Optimization In resource constrained networked frameworks characterized by a lack of central coordinator, distributed algorithms tend to be recursive. In such recursive schemes typically each entity in the network maintains its own estimate, decision statistic or optimizer depending on the problem at hand. Given the resource constraints in terms of sensing, computation and communication, it is of particular interest to develop distributed schemes which are frugal in terms of communication without suffering a loss in terms of performance. However, distributed recursive algorithms are plagued by the need to communicate at each time and requiring apriori knowledge of the dimensions of the parameter and thus trying to maintain an estimate of the size of a possibly high-dimensional parameter. In one hand, we develop and analyze the performance of an algorithm which caters to the problem where the entities have no knowledge about the dimension of the parameter and have heterogeneous objectives in terms of reconstructing only a few entries of the parameter. We specifically design an algorithm, where agents exchange lower dimensional messages and characterize the performance of the algorithm in terms of the interest sets of the entities. On the other hand, we explore a paradigm for estimation and optimization where we construct a communication scheme which continually sparsifies message exchanges between the agents while still achieving the optimal error performance.

Finite Sample Characterization In resource constrained networked setups, the entities involved have limited capacity so as to process sensed information. Moreover, the number of samples processed for accomplishing any distributed inference and optimization task is essentially finite. While asymptotic characterization in terms of samples highlights the limits of performance of an algorithm, finite sample guarantees takes into account the transient behavior of the evolution of the algorithm and hence providing for better tuning for the algorithm parameters. Especially in the context of optimization, typical assumptions involve knowledge of parameters characterizing smoothness and landscape of the loss function which is unfortunately available when dealing with a real-life problem. Hence, asymptotic rates emanating from an analysis involving aforementioned assumptions are potentially violated when employed for real data sets. Thus, non-asymptotic finite sample characterization of such schemes in terms of convergence rates provides the much needed understanding regarding the contribution of the different algorithm parameters which then allows for simplistic debugging. In the context of distributed optimization, we study non-asymptotic rates for optimization schemes which focus on communication efficiency and optimality. In particular, we provide intuition as to how the averaged connectivity of the network and the nature of the stochastic oracle being queried affects the convergence rate.

Gradient Free Optimization With the usage of machine learning algorithms in different research areas, stochastic optimization has become ubiquitous. In settings based around datacenters, where the computation and communication resources are abundant, first order stochastic schemes (Stochastic Gradient Descent type schemes), i.e., algorithms involving direct computation of gradients are ideal candidates. However, in many engineering applications, where the closed form of loss functions are not available or not analytic, gra-

dient computations are no longer feasible. On top of non-analytic loss functions, in the context of resource constrained networked setups, complicated gradient computations are prohibitive. Hence, a potentially attractive alternate is to employ gradient free optimization schemes, i.e., zeroth order optimization. Unlike Bayesian optimization which is extremely sample efficient but are extremely expensive to implements for dimensions beyond 10, zeroth order optimization is applicable to high-dimensional optimization. In this thesis, we study and develop distributed zeroth order optimization schemes which takes into account computation and communication efficiency. In particular, we establish convergence rates of such stochastic optimization algorithms through non-asymptotic analysis and thus explicitly characterizing the performance through the algorithm parameters for general classes of convex functions.
We summarize by detailing the organization of the thesis and list the contributions of each chapter broadly.

Distributed Detection Technical contributions pertaining to distributed detection in this thesis can be summarized in terms of development and analysis of detection schemes addressing sequential testing, recursive composite hypothesis testing and communication efficient simple hypothesis testing. For detailed exposition and concise statements, we refer the reader to the introductory sections of the chapters 2-4.

- **Chapter 2:** Under rather generic assumptions on the agent observation models, it is well-known that in a (hypothetical) centralized scenario or one in which inter-agent communication is all-to-all corresponding to a complete communication graph, the sequential probability ratio test (SPRT) (Wald et al. (1945)) turns out to be the optimal procedure for sequential testing of binary hypotheses; specifically, the SPRT minimizes the expected detection time (and hence the number of agent observation samples that need to be processed) while achieving requisite error performance in terms of specified probability of false alarm (α) and probability of miss (β) tolerances. The SPRT and its variants have been applied in various contexts, see, for example, spectrum sensing in cognitive radio networks (Choi et al. (2009); Jayaprakasam and Sharma (2009); Chaudhari et al. (2009)), target tracking Blostein and Richardson (1994), to name a few. However, the key ingredient of the SPRT algorithm, the thresholds of the algorithm do not extend to the ad-hoc networked setups. In this thesis, we developed and analyzed a distributed sequential detection algorithms for which we characterized the algorithm parameters and the associated stopping time distribution. In that, we characterized the performance of the proposed algorithm in terms of the connectivity of the graph induced by the inter-agent communication topology.

- **Chapter 3:** The problem of composite hypothesis testing is relevant to many practical applications, including cooperative spectrum sensing Zarrin and Lim (2009); Font-Segura and Wang (2010); Zou et al. (2010) and MIMO radars Tajer et al. (2010), where the onus is also on achieving reasonable detection performance by utilizing as fewer resources as possible, which includes data samples, communication and sensing energy. In classical composite hypothesis testing procedures such as the Generalized Likelihood Ratio Test (GLRT) Zeitouni et al. (1992), the detection procedure which uses the underlying parameter estimate based on all the collected samples as a plug-in estimate may not be initiated until a reasonably accurate parameter estimate, typically the maximum likelihood estimate of the underlying parameter (state) is obtained. Usually in setups which employ the classical (centralized) generalized likelihood ratio tests, the data collection phase precedes the parameter estimation and detection statistic update phase which makes the procedure essentially an *offline* batch procedure. In contrast to the fully centralized setup, we focus on a fully distributed setup where the communication between the agents is restricted to a pre-assigned possibly sparse communication graph and propose two algorithms namely, *consensus* + *innovations* GLRT Non-Linear ($\mathcal{CIGLRT}-\mathcal{NL}$) and *consensus* +

innovations GLRT Linear ($\mathcal{CIGLRT}-\mathcal{L}$), which are of the *consensus* + *innovations* form and are based on fully distributed setups. in spite of being a recursive algorithm and hence suboptimal, the proposed algorithms guarantee asymptotically decaying probabilities of false alarm and miss under minimal conditions of global observability and connectivity of the inter-agent communication graph.

- **Chapter 4:** In this chapter, we derive the large deviation rate for convergence in probability of products of independent but not identically distributed stochastic matrices arising in time-varying distributed consensus-type networks. More precisely, we consider the model in which there exists a baseline topology that describes all possible communications and nodes are activated sparsely. At any given time, a node is active with a certain time-dependent probability, and any two nodes communicate if they are both active at that time. Under this model, we compute the exact rate for exponential decay of probabilities that the matrix products stay bounded away from their limiting matrix. We show that the rate is given by the minimal vertex cut of the baseline topology, where the node costs are defined by their limiting activation probabilities. The computed rate has many potential applications in distributed inference with intermittent communications. We provide an application in the context of *consensus+innovations* distributed detection. Therein, we show that optimal error exponent is achievable under a very general model of sparsified activations, thus effectively constructing asymptotically optimal detectors with significant communications savings.

Distributed Estimation Technical contributions pertaining to distributed estimation in this thesis can be summarized in terms of development and analysis of communication efficient distributed estimation in terms of frugal communication cost. The frugality in terms of the communication cost is explored both in terms of low-dimensional messages and low communication cost for general heterogeneous sensing models. For detailed exposition and concise statements, we refer the reader to the introductory sections of the chapters 5-7.

- **Chapter 5:** In this chapter, we explore a paradigm where we construct a communication scheme which continually sparsifies message exchanges between the agents while still achieving the optimal error performance. We propose **C**ommunication efficient **RE**cursive **D**istributed estimati**O**n algorithm, $\mathcal{CREDO}$ for networked multi-worker setups without a central master node. $\mathcal{CREDO}$ is designed for scenarios in which the worker nodes aim to collaboratively estimate a vector parameter of interest using distributed online sampled data at the individual worker nodes. The individual worker nodes at each iteration, update their estimate of the parameter by assimilating their latest sensed information and estimates from their time-varying neighborhood worker nodes over a (possibly sparse) communication graph. The underlying inter-worker communication protocol is randomized and makes communications be increasingly (probabilistically) sparse. Under minimal conditions on the inter-worker information exchange network and the sensing models, almost sure convergence of the estimate sequence to the true parameter is established. Further, we characterize the performance of $\mathcal{CREDO}$ in terms of asymptotic covariance of the estimate sequences and specifically establishes the achievability of optimal asymptotic covariance. The analysis reveals an interesting interplay between the algorithm's communication cost $\mathcal{C}_t$ and the asymptotic covariance. Most notably, it is shown that $\mathcal{CREDO}$ may be designed to achieve a $\Theta\left(\mathcal{C}_t^{-2+\zeta}\right)$ decay of the mean square error ($\zeta > 0$, arbitrarily small) at each worker node, which significantly improves over the existing $\Theta\left(\mathcal{C}_t^{-1}\right)$ rates. On real datasets $\mathcal{CREDO}$ requires on an average up to $3\times$ less communications to obtain reasonable mean square error as compared to benchmark schemes.

- **Chapter 6:** In this chapter we address the design and analysis of communication efficient distributed algorithms for solving weighted non-linear least squares problems in multi-agent networks. *Communication efficiency* is highly relevant in modern applications like cyber-physical systems and internet of things, where a significant portion of the involved devices have energy constraints in terms of limited battery power. Furthermore, *non-linear models* arise frequently in such systems, like, e.g., with power grid state estimation. We first propose a distributed recursive estimator of the *consensus+innovations* type, namely $\mathcal{CIWNLS}$, in which the agents update their parameter estimates at each observation sampling epoch in a collaborative way by simultaneously processing the latest locally sensed information (*innovations*) and the parameter estimates from other agents (*consensus*) in the local neighborhood conforming to a pre-specified inter-agent communication topology. Under rather weak conditions on the connectivity of the inter-agent communication and a *global observability* criterion, it is shown that, at every network agent, $\mathcal{CIWNLS}$ leads to consistent parameter estimates. Furthermore, we develop and analyze a non-linear communication efficient distributed algorithm dubbed $\mathcal{CREDO}-\mathcal{NL}$ (non-linear $\mathcal{CREDO}$). $\mathcal{CREDO}-\mathcal{NL}$ generalizes the recently proposed linear method $\mathcal{CREDO}$ to non-linear models. We establish for a broad class of non-linear least squares problems and generic underlying multi-agent network topologies $\mathcal{CREDO}-\mathcal{NL}$'s strong consistency. Furthermore, we demonstrate communication efficiency of the method, both theoretically and by simulation examples. For the former, we rigorously prove that $\mathcal{CREDO}-\mathcal{NL}$ achieves significantly faster mean squared error rates in terms of the elapsed communication cost over existing alternatives. For the latter, the considered simulation experiments show communication savings by at least an order of magnitude.

- **Chapter 7:** In this chapter, we present a communication efficient distributed algorithm, $\mathcal{CIRFE}$ of the *consensus+innovations* type, to estimate a high-dimensional parameter in a multi-agent network, in which each agent is interested in reconstructing only a few components of the parameter. This problem arises for example when monitoring the high-dimensional distributed state of a large-scale infrastructure with a network of limited capability sensors and where each sensor is tasked with estimating some local components of the state. This chapter explores the paradigm of communication efficiency which involves reduction in the dimension of messages exchanged among agents. Under minimal conditions on the inter-agent communication network and the sensing models, almost sure convergence of the estimate sequence at each agent to the components of the true parameter in its interest set is established. Furthermore, the chapter establishes the performance of $\mathcal{CIRFE}$ in terms of asymptotic covariance of the estimate sequences and specifically characterizes the dependencies of the component wise asymptotic covariance in terms of the number of agents tasked with estimating it. Finally, simulation experiments demonstrate the efficacy of $\mathcal{CIRFE}$.

Distributed Optimization Technical contributions pertaining to distributed optimization in this thesis can be summarized in terms of development and analysis of communication efficient distributed optimization in terms of frugal communication cost. The frugality in terms of the communication cost is studied in conjunction with exploration of stochastic oracles for optimization, while keeping the resource constrained setups in mind. Furthermore, a gradient free and projection free approach for stochastic optimization is also explored. For detailed exposition and concise statements, we refer the reader to the introductory sections of the chapters 8-10.

- **Chapter 8:** In this chapter, we establish the $O(\frac{1}{t})$ convergence rate for distributed stochastic gradient

methods that operate over strongly convex costs and random networks. The considered class of methods is standard each node performs a weighted average of its own and its neighbors solution estimates (consensus), and takes a negative step with respect to a noisy version of its local functions gradient (innovation). The underlying communication network is modeled through a sequence of temporally independent identically distributed (i.i.d.) Laplacian matrices connected on average, while the local gradient noises are also i.i.d. in time, have finite second moment, and possibly unbounded support. We show that, after a careful setting of the consensus and innovations potentials (weights), the distributed stochastic gradient method achieves a (order-optimal) $O(\frac{1}{t})$ convergence rate in the mean square distance from the solution. This is the first order-optimal convergence rate result on distributed strongly convex stochastic optimization when the network is random and/or the gradient noises have unbounded support. Furthermore, we examine fundamental tradeoffs in distributed SGD in multi-agent networks in terms of *communication cost* (number of per-node transmissions) and *computational cost*, measured by the number of per-node noisy function evaluations with zeroth order methods. Under standard assumptions on the cost functions and the noise statistics, we establish with the proposed method the $O(1/(C_{\text{comm}})^{4/3-\zeta})$ mean square error convergence rate, for distributed SGD, where C_{comm} is the expected number of network communications and $\zeta > 0$ is arbitrarily small. The method is shown to achieve order-optimal convergence rates in terms of computational cost C_{comp}, $O(1/(C_{\text{comp}}))$ while achieving the order-optimal convergence rates in terms of iterations. Experiments on real-life datasets illustrate the efficacy of the proposed algorithms.

- **Chapter 9:** In the context of distributed optimization in resource constrained networked setups, due to computational bottlenecks the networked entities in harsher environments, the access is limited to stochastic zeroth order oracles. In conjunction with the inherent randomness in the network, it makes the optimization problem at hand challenging and poses the question as to how close is the performance of the distributed algorithm with respect to its centralized counterpart. For the described random networks based optimization setting with access to a stochastic zeroth order oracle, we develop a distributed stochastic approximation method of the Kiefer-Wolfowitz type. Furthermore, under standard smoothness and strong convexity assumptions on the local costs, we establish the $O(1/t^{1/2})$ (in terms of iteration t) mean square convergence rate for the method – the rate that matches that of the method's centralized counterpart under equivalent conditions. Furthermore, we examine fundamental tradeoffs in iterative distributed zeroth order stochastic optimization in multi-agent networks in terms of *communication cost* (number of per-node transmissions) and *computational cost*, measured by the number of per-node noisy function evaluations with zeroth order methods. Specifically, we develop novel distributed stochastic optimization methods for zeroth order strongly convex optimization by utilizing a probabilistic inter-agent communication protocol that increasingly sparsifies communications among agents as time progresses. Under standard assumptions on the cost functions and the noise statistics, we establish with the proposed method the $O(1/(C_{\text{comm}})^{8/9-\zeta})$ mean square error convergence rate, for zeroth order optimization, where C_{comm} is the expected number of network communications and $\zeta > 0$ is arbitrarily small. The method is shown to achieve order-optimal convergence rates in terms of computational cost C_{comp}, $O(1/(C_{\text{comp}})^{2/3})$ while achieving the order-optimal convergence rates in terms of iterations. Experiments on real-life datasets illustrate the efficacy of the proposed algorithms.

- **Chapter 10:** In this chapter, we focus on the problem of *constrained stochastic* optimization. A zeroth order Frank-Wolfe algorithm is proposed, which in addition to the projection-free nature of the

vanilla Frank-Wolfe algorithm makes it gradient free. Under convexity and smoothness assumption, we show that the proposed algorithm converges to the optimal objective function at a rate $O\left(1/T^{1/3}\right)$, where T denotes the iteration count. In particular, the primal sub-optimality gap is shown to have a dimension dependence of $O\left(d^{1/3}\right)$, which is the best known dimension dependence among all zeroth order optimization algorithms with one directional derivative per iteration. For non-convex functions, we obtain the *Frank-Wolfe* gap to be $O\left(d^{1/3}T^{-1/4}\right)$. The proposed algorithm does not depend on hard to estimates like Lipschitz constants and thus is easy to deploy in practice. Experiments on black-box optimization setups demonstrate the efficacy of the proposed algorithm.

Part I

Distributed Detection

Chapter 2

Distributed Sequential Detection

2.1 Introduction

The motivation behind studying sequential as opposed to fixed sample size testing is that in most practical agent networking scenarios, especially in applications that are time-sensitive and/or resource constrained, the priority is to achieve inference as quickly as possible by expending the minimal amount of resources (data samples, sensing energy and communication). By sequential we mean, instead of considering fixed sample size hypothesis tests in which the objective is to minimize the probabilities of decision error (the false alarm and the miss) based on a given deterministic number of samples or observation data collected by the network agents, we are interested in the design of testing procedures that in the *quickest* time or using the *minimal* amount of sensed data samples at the agents can distinguish between the two hypotheses with guaranteed accuracy given in terms of pre-specified tolerances on false alarm and miss probabilities. We focus on distributed application environments which are devoid of fusion centers and in which inter-agent collaboration or information exchange is limited to a pre-assigned, possibly sparse, communication structure.

The focus of this chapter is on sequential simple hypothesis testing in multi-agent networks in which the goal is to detect the (binary) state of the environment based on observations at the agents. By sequential we mean, instead of considering fixed sample size hypothesis tests in which the objective is to minimize the probabilities of decision error (the false alarm and the miss) based on a given deterministic number of samples or observation data collected by the network agents, we are interested in the design of testing procedures that in the *quickest* time or using the *minimal* amount of sensed data samples at the agents can distinguish between the two hypotheses with guaranteed accuracy given in terms of pre-specified tolerances on false alarm and miss probabilities. The motivation behind studying sequential as opposed to fixed sample size testing is that in most practical agent networking scenarios, especially in applications that are time-sensitive and/or resource constrained, the priority is to achieve inference as quickly as possible by expending the minimal amount of resources (data samples, sensing energy and communication). Furthermore, we focus on distributed application environments which are devoid of fusion centers[1] and in which inter-agent collaboration or information exchange is limited to a pre-assigned, possibly sparse, communication structure.

Under rather generic assumptions on the agent observation models, it is well-known that in a (hypothetical) centralized scenario or one in which inter-agent communication is all-to-all corresponding to a complete

[1]By fusion center or center, we mean a hypothetical decision-making architecture in which a (central) entity has access to all agent observations at all times and/or is responsible for decision-making on behalf of the agents.

communication graph, the sequential probability ratio test (SPRT) (Wald et al. (1945)) turns out to be the optimal procedure for sequential testing of binary hypotheses; specifically, the SPRT minimizes the expected detection time (and hence the number of agent observation samples that need to be processed) while achieving requisite error performance in terms of specified probability of false alarm (α) and probability of miss (β) tolerances. The SPRT and its variants have been applied in various contexts, see, for example, spectrum sensing in cognitive radio networks (Choi et al. (2009); Jayaprakasam and Sharma (2009); Chaudhari et al. (2009)), target tracking Blostein and Richardson (1994), to name a few. However, the SPRT, in the current multi-agent context, would require computing a (centralized) decision statistic at all times, which, in turn, would either require all-to-all communication among the agents or access to the entire network data at all times at a fusion center. In contrast, restricted by a pre-assigned possibly sparse collaboration structure among the agents, in this paper we present and characterize a distributed sequential detection algorithm, the $\mathcal{CISPRT}$, based on the *consensus+innovations* approach (see, for example Kar and Moura (2008b); Kar et al. (2012)). Specifically, focusing on a setting in which the agent observations over time are conditionally Gaussian and independent and identically distributed (i.i.d.), we study the $\mathcal{CISPRT}$ sequential detection procedure in which each network agent maintains a local (scalar) test statistic which is updated over time by simultaneously assimilating the test statistics of neighboring agents at the previous time instant (a consensus potential) and the most recent observations (innovations) obtained by the agent and its neighbors. Also, similar in spirit to the (centralized) SPRT, each agent chooses two (local) threshold parameters (design choices) and the test termination at an agent (and subsequent agent decision on the hypotheses) is determined by whether the local test statistic at the agent lies in the interval defined by the thresholds or not. This justifies the nomenclature that the $\mathcal{CISPRT}$ is a distributed SPRT type algorithm of the *consensus+innovations* form.

2.2 Related Work

Detection schemes in multi-agent networks which involve fusion centers, where all agents in the network transmit their local measurements, local decisions or local likelihood ratios to a fusion agent which subsequently makes the final decision (see, for example, Chamberland and Veeravalli (2003); Tsitsiklis et al. (1993); Blum et al. (1997); Veeravalli et al. (1993)) have been well studied. Consensus-based approaches for fully distributed but single snapshot processing, i.e., in which the agents first collect their observations possibly over a long time horizon and then deploy a consensus-type protocol Jadbabaie et al. (2003); Olfati-Saber et al. (2007); Dimakis et al. (2010) to obtain distributed information fusion and decision-making have also been explored, see, for instance, Kar et al. (2008); Kar and Moura (2007). Generalizations and variants of this framework have been developed, see for instance Zhang and Blum (2014) which proposes truncated versions of optimal testing procedures to facilitate efficient distributed computation using consensus; scenarios involving distributed information processing where some of the agents might be faulty or there is imperfect model information (see, for example, Zhou et al. (2011, 2012)) have also been studied. More relevant to the current context are distributed detection techniques that like the $\mathcal{CISPRT}$ procedure perform simultaneous assimilation of neighborhood decision-statistics and local agent observations in the same time step, see, in particular, the running consensus approach Braca et al. (2008, 2010), the diffusion approach Cattivelli and Sayed (2009a,b, 2011b) and the consensus+innovations approach Bajovic et al. (2011); Jakovetic et al. (2012); Kar et al. (2011). These works address important questions in fixed (but possibly large) sample size distributed hypothesis testing, including asymptotic characterization of detection errors Braca et al. (2010); Cattivelli and Sayed (2011b), fundamental performance limits as characterized by large deviations decay of

detection error probabilities in generic nonlinear observation models and random networks Bajovic et al. (2011); Jakovetic et al. (2012), and detection with noisy communication links Kar et al. (2011).

2.3 Problem Formulation

2.3.1 System Model

The N agents deployed in the network decide on either of the two hypothesis H_0 and H_1. Each agent i at (discrete) time t makes a scalar observation $y_i(t)$ of the form

$$Under \quad H_\theta : y_i(t) = \mu_\theta + n_i(t), \quad \theta = 0, 1. \tag{2.1}$$

For the rest of the chapter we consider $\mu_1 = \mu$ and $\mu_0 = -\mu$, and assume that, the agent observation noise processes are independent and identically distributed (i.i.d.) Gaussian processes under both hypotheses formalized as follows:

Assumption 2.3.1. *For each agent i the noise sequence $\{n_i(t)\}$ is i.i.d. Gaussian with mean zero and variance σ^2 under both H_0 and H_1. The noise sequences are also spatially uncorrelated, i.e., $\mathbb{E}_\theta[n_i(t)n_j(t)] = \mathbf{0}$ for all $i \neq j$ and $\theta \in \{0, 1\}$.*

Collect the $y_i(t)$'s, $i = 1, 2, \cdots N$ into the $N \times 1$ vector $\mathbf{y}(t) = (y_1(t), \cdots, y_N(t))^\top$ and the $n_i(t)$'s, $i = 1, 2, \cdots N$ into the $N \times 1$ vector $\mathbf{n}(t) = (n_1(t), \cdots, n_N(t))^\top$.
The log-likelihood ratio at the i-th sensor at time index t is calculated as follows:-

$$\eta_i(t) = \frac{f_1(y_i(t))}{f_0(y_i(t))} = \frac{2\mu y_i(t)}{\sigma^2}, \tag{2.2}$$

where $f_0(\cdot)$ and $f_1(\cdot)$ denote the probability distribution functions (p.d.f.s) of $y_i(t)$ under H_0 and H_1 respectively.
We note that,

$$\eta_i(t) \sim \begin{cases} \mathcal{N}(m, 2m), & H = H_1 \\ \mathcal{N}(-m, 2m), & H = H_0, \end{cases} \tag{2.3}$$

where $\mathcal{N}(\cdot)$ denotes the Gaussian p.d.f. and $m = \frac{2\mu^2}{\sigma^2}$. The Kullback-Leibler divergence at each agent is given by

$$KL = m. \tag{2.4}$$

2.3.2 Sequential Hypothesis Testing – Centralized or All-To-All Communication Scenario

In a fully connected network scenario, each agent behaves like a (hypothetical) center and the information available at any agent n at time t is the sum-total of network observations till t, formalized by the σ-algebra Jacod and Shiryaev (1987)

$$\mathcal{G}_c(t) = \sigma\left\{y_i(s),\ \forall i = 1, 2, \cdots N \text{ and } \forall 1 \leq s \leq t\right\}. \tag{2.5}$$

An admissible test D_c consists of a stopping criteria, where at each time t the decision to stop or continue taking observations is adapted to $\mathcal{G}_c(t)$. Denote by T_{D_c} the termination time of D_c, a random time taking values in $\mathbb{Z}_+ \cup \{\infty\}$. Let $\mathbb{P}_{FA}^{D_c}$ and $\mathbb{P}_{M}^{D_c}$ denote the associated probabilities of false alarm and miss respectively, i.e.,

$$\mathbb{P}_{FA}^{D_c} = \mathbb{P}_0\left(\widehat{H}_{D_c} = 1\right) \quad \text{and} \quad \mathbb{P}_{M}^{D_c} = \mathbb{P}_1\left(\widehat{H}_{D_c} = 0\right). \tag{2.6}$$

Now, denoting by $\mathcal{D}_c$ the class of all such (centralized) admissible tests, the goal in sequential hypothesis testing is to obtain a test in $\mathcal{D}_c$ that minimizes the expected stopping time subject to attaining specified error constraints. It turns out that Wald's SPRT Wald et al. (1948) can be designed to minimize each of the above criteria. Hence, without loss of generality, we adopt $\mathbb{E}_1[T_{D_c}]$ as our test design objective.

$$\begin{aligned} &\min_{D_c \in \mathcal{D}_c} \mathbb{E}_1[T_{D_c}], \\ &\text{s.t. } \mathbb{P}_{FA}^{D_c} \leq \alpha, \mathbb{P}_{M}^{D_c} \leq \beta, \end{aligned} \tag{2.7}$$

for specified α and β. Before proceeding further, we make the following assumption:

Assumption 2.3.2. *The pre-specified error metrics, i.e., α and β, satisfy $\alpha, \beta \in (0, 1/2)$.*

Noting that the (centralized) Kullback-Leibler divergence, i.e., the divergence between the probability distributions induced on the joint observation space $\mathbf{y}(t)$ by the hypotheses H_1 and H_0, is Nm where m is defined in (2.3), we obtain (see Wald et al. (1948)) for each $D_c \in \mathcal{D}_c$ that attains $\mathbb{P}_{FA}^{D_c} \leq \alpha$ and $\mathbb{P}_{M}^{D_c} \leq \beta$,

$$\mathbb{E}_1[T_{D_c}] \geq \mathcal{M}(\alpha, \beta), \tag{2.8}$$

where the universal lower bound $\mathcal{M}(\alpha, \beta)$ is given by

$$\mathcal{M}(\alpha, \beta) = \frac{(1-\beta)\log(\frac{1-\beta}{\alpha}) + \beta \log(\frac{\beta}{1-\alpha})}{Nm}. \tag{2.9}$$

Formally, the stopping time of the SPRT is defined as follows: Denote by $S_c(t)$ (the centralized) test statistic

$$S_c(t) = \sum_{s=1}^{t} \frac{\mathbf{1}^\top}{N} \eta(s), \tag{2.10}$$

where $\eta(s)$ denotes the vector of log-likelihood ratios $\eta_i(s)$'s at the agents and the stopping time is given by,

$$T_c = \inf\{t \mid S_c(t) \notin [\gamma_c^l, \gamma_c^h]\}. \tag{2.11}$$

At T_c the following decision rule is followed:

$$H = \begin{cases} H_0, & S_c(T_c) \leq \gamma_c^l \\ H_1, & S_c(T_c) \geq \gamma_c^h. \end{cases} \tag{2.12}$$

The optimality of the SPRT w.r.t. the formulation (2.7) is well-studied; in particular, in Wald et al. (1948) it was shown that, for any specified α and β, there exist choices of thresholds (γ_c^l, γ_c^h) such that the SPRT (2.11)-(2.12) achieves the minimum in (2.7) among all possible admissible tests D_c in $\mathcal{D}_c$. For given α and β, exact analytical expressions of the optimal thresholds are intractable in general. A commonly used

choice of thresholds, see Wald et al. (1945), is given by

$$\gamma_c^h = \log\big(\frac{1-\beta}{\alpha}\big)$$
$$\gamma_c^l = \log\big(\frac{\beta}{1-\alpha}\big), \tag{2.13}$$

which, although not strictly optimal in general, ensures that $\mathbb{P}_{FA}^c \leq \alpha$ and $\mathbb{P}_M^c \leq \beta$. The SPRT with thresholds given by (2.13) guarantees that (see Chernoff (1972))

$$\lim_{\epsilon\to 0} \frac{\mathbb{E}_1[T_c]}{\mathcal{M}(\epsilon,\epsilon)} = 1, \tag{2.14}$$

where $\mathcal{M}(\cdot)$ is defined in (2.9). In the sequel, given a testing procedure $D_c \in \mathcal{D}_c$ and assuming $\alpha = \beta = \epsilon$, we will study the quantity $\limsup_{\epsilon\to 0}(\mathbb{E}_1[T_{D_c}]/\mathcal{M}(\epsilon,\epsilon))$ as a measure of its efficiency.

2.3.3 Subclass of Distributed Tests

The SPRT (2.11)-(2.12) requires computation of the statistic $S_c(t)$ (see (2.10)) at all times, which, in turn, requires access to all agent observations at all times. Given a graph $G = (V, E)$, possibly sparse, modeling inter-agent communication, we consider scenarios in which inter-agent cooperation is limited to a single round of message exchanges among neighboring agents per observation sampling epoch. The information set $\mathcal{G}_{d,i}(t)$ includes the observations sampled by agent i and the messages received from its neighbors till time t, and is formally given by the σ-algebra

$$\mathcal{G}_{d,i}(t) = \sigma\left\{y_i(s), m_{i,j}(s), \quad \forall 1 \leq s \leq t, \forall j \in \Omega_i\right\}. \tag{2.15}$$

The quantity $m_{i,j}(s)$ denotes the message received by i from its neighbor $j \in \Omega_i$ at time s. Based on the information content $\mathcal{G}_{d,i}(t)$ at time t, an agent decides on whether to continue taking observations or to stop in the case of which, it decides on one of the hypothesis H_0 or H_1. Intuitively, and formally by (2.15), we have

$$\mathcal{G}_{d,i}(t) \subset \mathcal{G}_c(t) \quad \forall i, t. \tag{2.16}$$

Formally, this implies that the class of distributed tests $\mathcal{D}_d$ is a subset of the class of centralized or all-possible tests $\mathcal{D}_c$. We are interested in characterizing the distributed test that conforms to the communication restrictions above and is optimal in the following sense:

$$\min_{D_d\in\mathcal{D}_d} \max_{i=1,2,\cdots,N} \mathbb{E}_1[T_{D_d,i}],$$
$$\text{s.t. } \mathbb{P}_{FA}^{D_d,i} \leq \alpha, \mathbb{P}_M^{D_d,i} \leq \beta, \forall i = 1, 2, \cdots, N. \tag{2.17}$$

In the above, $T_{D_d,i}$ denotes the termination (stopping) time at an agent i and $\mathbb{P}_{FA}^{D_d,i}$, $\mathbb{P}_M^{D_d,i}$, the respective false alarm and miss probabilities at i. Rather than solving (2.17), we propose a distributed testing procedure of the consensus+innovations type which is efficiently implementable and analyze its performance w.r.t. the optimal centralized testing procedure. Our results clearly demonstrate the benefits of collaboration (even over a sparse communication network) as, in contrast, in the non-collaboration case (i.e., each agent relies on its own observations only) each agent would require N times the expected number of observations to achieve prescribed α and β as compared to the optimal centralized scenario.

2.4 $\mathcal{CISPRT}$: A Distributed Sequential Detector

We propose a distributed sequential detection scheme where network communication is restricted to a more *localized* agent-to-agent interaction scenario. Before discussing the details of our algorithm, we state an assumption on the inter-agent communication graph.

Assumption 2.4.1. *The inter-agent communication graph is connected, i.e.* $\lambda_2(\mathbf{L}) > 0$, *where* $\mathbf{L}$ *denotes the associated graph Laplacian matrix.*

Decision Statistic Update. In the proposed distributed algorithm, each agent i maintains a test statistic $P_{d,i}(t)$, which is updated recursively in a distributed fashion as follows :

$$P_{d,i}(t+1) = \frac{t}{t+1}\left(w_{ii}P_{d,i}(t) + \sum_{j\in\Omega_i} w_{ij}P_{d,j}(t)\right) + \frac{1}{t+1}\left(w_{ii}\eta_i(t+1) + \sum_{j\in\Omega_i} w_{ij}\eta_j(t+1)\right), \tag{2.18}$$

where Ω_i denotes the communication neighborhood of agent i and the w_{ij}'s denote appropriately chosen combination weights (to be specified later).
Now we state some design assumptions on the weight matrix $\mathbf{W}$.

Assumption 2.4.2. *We design the weights* w_{ij}*'s in* (2.18) *such that the matrix* $\mathbf{W}$ *is non-negative, symmetric, irreducible and stochastic, i.e., each row of* $\mathbf{W}$ *sums to one.*

Note that, by the stochasticity of $\mathbf{W}$, the quantity r satisfies

$$r = ||\mathbf{W} - \mathbf{J}||. \tag{2.19}$$

For connected graphs, a simple way to design $\mathbf{W}$ is to assign equal combination weights, in which case we have,

$$\mathbf{W} = \mathbf{I} - \delta\mathbf{L}, \tag{2.20}$$

where δ is a suitably chosen constant. The smallest value of r is obtained by setting δ to be equal to $2/(\lambda_2(\mathbf{L}) + \lambda_N(\mathbf{L}))$.
Stopping Criterion for the Decision Update. Let $S_{d,i}(t)$ denote the quantity $tP_{d,i}(t)$, and let $\gamma^h_{d,i}$ and $\gamma^l_{d,i}$ be thresholds at an agent i (to be determined later) such that agent i stops and makes a decision only when,

$$S_{d,i}(t) \notin [\gamma^l_{d,i}, \gamma^h_{d,i}] \tag{2.21}$$

for the first time. The stopping time for reaching a decision at an agent i is then defined as,

$$T_{d,i} = \inf\{t \ \ | S_{d,i}(t) \notin [\gamma^l_{d,i}, \gamma^h_{d,i}]\}, \tag{2.22}$$

and the following decision rule is adopted at $T_{d,i}$:

$$H = \begin{cases} H_0 & S_{d,i}(T_{d,i}) \leq \gamma_{d,i}^l \\ H_1 & S_{d,i}(T_{d,i}) \geq \gamma_{d,i}^h. \end{cases} \tag{2.23}$$

We refer to this distributed scheme (2.18), (2.22) and (2.23) as the *consensus+innovations* SPRT ($\mathcal{CISPRT}$) hence forth.

Remark 2.4.1. *It is to be noted that the decision statistic update rule is distributed and recursive, in that, to realize* (2.18) *each agent needs to communicate its current statistic and a scalar function of its latest sensed observation to its neighbors only; furthermore, the local update rule* (2.18) *is a combination of a consensus term reflecting the weighted combination of neighbors' statistics and a local innovation term reflecting the new sensed information of itself and its neighbors. Note that the stopping times $T_{d,i}$'s are random and generally take different values for different agents. It is to be noted that the $T_{d,i}$'s are in fact stopping times with respect to the respective agent information filtrations $\mathcal{G}_{d,i}(t)$'s as defined in* (2.15)*. For subsequent analysis we refer to the stopping time of an agent as the stopping time for reaching a decision at an agent.*

We end this section by providing some elementary properties of the distributed test statistics.

Proposition 2.4.2. *Let the Assumptions 2.3.1, 2.4.1 and 2.4.2 hold. For each t and i, the statistic $S_{d,i}(t)$, defined in* (C.53)-(C.44)*, is Gaussian under both H_0 and H_1. In particular, we have*

$$\mathbb{E}_0[S_{d,i}(t)] = -mt \quad \textit{and} \quad \mathbb{E}_1[S_{d,i}(t)] = mt, \tag{2.24}$$

and

$$\mathbb{E}_0\left[(S_{d,i}(t) + mt)^2\right] = \mathbb{E}_1\left[(S_{d,i}(t) - mt)^2\right] \leq \frac{2mt}{N} + \frac{2mr^2(1 - r^{2t})}{1 - r^2}. \tag{2.25}$$

2.5 Main Results

We formally state the main results in this section. Most of the proofs pertaining to the main results are relegated to Appendix B.

2.5.1 Thresholds for the $\mathcal{CISPRT}$

In this section we derive thresholds for the $\mathcal{CISPRT}$, see (C.53)-(C.44), in order to ensure that the procedure terminates in finite time a.s. at each agent and the agents achieve specified error probability requirements. For the proposed $\mathcal{CISPRT}$, we intend to derive thresholds which guarantee the error performance in terms of the error probability requirements α and β, i.e., such that $\mathbb{P}_{FA}^{d,i} \leq \alpha$ and $\mathbb{P}_{M}^{d,i} \leq \beta$, $\forall i = 1, 2, \ldots, N$, where $\mathbb{P}_{FA}^{d,i}$ and $\mathbb{P}_{M}^{d,i}$ represent the probability of false alarm and the probability of miss for the ith agent defined as

$$\begin{aligned} \mathbb{P}_{FA}^{d,i} &= \mathbb{P}_0(S_{d,i}(T_{d,i}) \geq \gamma_{d,i}^h) \\ \mathbb{P}_{M}^{d,i} &= \mathbb{P}_1(S_{d,i}(T_{d,i}) \leq \gamma_{d,i}^l), \end{aligned} \tag{2.26}$$

with $T_{d,i}$ as defined in (C.43).

Theorem 2.5.1. *Let the Assumptions 2.3.1-2.4.2 hold.*
1) Then, for each α and β there exist $\gamma_{d,i}^h$ and $\gamma_{d,i}^l$, $\forall i = 1, 2, \cdots, N$, such that $\mathbb{P}_{FA}^{d,i} \leq \alpha$ and $\mathbb{P}_M^{d,i} \leq \beta$ and the test concludes in finite time a.s. i.e.

$$\mathbb{P}_1(T_{d,i} < \infty) = 1, \forall i = 1, 2, \cdots, N, \tag{2.27}$$

where $T_{d,i}$ is the stopping time for reaching a decision at agent i.
2) In particular, for given α and β, any choice of thresholds $\gamma_{d,i}^h$ and $\gamma_{d,i}^l$ satisfying

$$\gamma_{d,i}^h \geq \frac{8(k+1)}{7N}\left(\log\left(\frac{2}{\alpha}\right) - \log(1 - e^{\frac{-Nm}{4(k+1)}})\right) = \gamma_d^{h,0} \tag{2.28}$$

$$\gamma_{d,i}^l \leq \frac{8(k+1)}{7N}\left(\log\left(\frac{\beta}{2}\right) + \log(1 - e^{\frac{-Nm}{4(k+1)}})\right) = \gamma_d^{l,0}, \tag{2.29}$$

where m is defined in (2.4) and k is defined by

$$Nr^2 = k, \tag{2.30}$$

with r as in (2.19), achieves a.s. finite stopping at an agent i while ensuring that $\mathbb{P}_{FA}^{d,i} \leq \alpha$ and $\mathbb{P}_M^{d,i} \leq \beta$.

Proof. Let $\hat{A} = e^{\gamma_{d,i}^l}$ and $\hat{B} = e^{\gamma_{d,i}^h}$ where $\gamma_{d,i}^h$ and $\gamma_{d,i}^l \in \mathbb{R}$ are thresholds (to be designed) for the $\mathcal{CISPRT}$. In the following derivation, for a given random variable z and an event A, we use the notation $\mathbb{E}[z; A]$ to denote the expectation $\mathbb{E}[z\mathbb{I}_A]$. Let T denote the random time which can take values in $\overline{\mathbb{Z}}_+$ given by

$$T = \inf\left\{t | S_{d,i}(t) \notin \left[\gamma_{d,i}^l, \gamma_{d,i}^h\right]\right\}. \tag{2.31}$$

First, we show that for any $\gamma_{d,i}^h$ and $\gamma_{d,i}^l \in \mathbb{R}$,

$$\mathbb{P}_0\left(T < \infty\right) = \mathbb{P}_1\left(T < \infty\right) = 1, \tag{2.32}$$

i.e., the random time T defined in (2.31) is a.s. finite under both the hypotheses. Indeed, we have,

$$\begin{aligned}
&\mathbb{P}_1\left(T > t\right) \leq \mathbb{Q}\left(\frac{-\gamma_{d,i}^h + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\right) \\
&\Rightarrow \lim_{t\to\infty} \mathbb{P}_1\left(T > t\right) = 0 \\
&\Rightarrow \mathbb{P}_1\left(T < \infty\right) = 1.
\end{aligned} \tag{2.33}$$

The proof for H_0 follows in a similar way.
Now, since (2.32) holds, the quantity $S_{d,i}(T)$ is well-defined a.s. under H_0. Now, noting that, under H_0, for any t, the quantity $S_{d,i}(t)$ is Gaussian with mean $-mt$ and variance upper bounded by $\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}$

(see Proposition 2.4.2), we have,

$$\begin{aligned}\mathbb{P}_{FA}^{d,i} &= \mathbb{P}_0(S_{d,i}(T) \geq \log \hat{B}) = \sum_{t=1}^{\infty} \mathbb{P}_0(T=t, S_{d,i}(t) \geq \log \hat{B}) \\ &\leq \sum_{t=1}^{\infty} \mathbb{P}_0(S_{d,i}(t) \geq \log \hat{B}) \\ &\leq \sum_{t=1}^{\infty} \mathbb{Q}\Big(\frac{\log \hat{B} + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\Big). \end{aligned} \tag{2.34}$$

To obtain a condition for $\gamma_{d,i}^h$ in the $\mathcal{CISPRT}$ such that $\mathbb{P}_{FA}^{d,i} \leq \alpha$, let's define $k > 0$ such that $k = Nr^2$. Now, note that k thus defined satisfies

$$\frac{2mr^2(1-r^{2t})}{1-r^2} \leq \frac{2mkt}{N}, \quad \forall t. \tag{2.35}$$

Then we have, by (2.34)-(2.35),

$$\begin{aligned}\mathbb{P}_{FA}^{d,i} &\leq \sum_{t=1}^{\infty} \mathbb{Q}\Big(\frac{\log \hat{B} + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\Big) \leq \sum_{t=1}^{\infty} \mathbb{Q}\Big(\frac{\log \hat{B} + mt}{\sqrt{\frac{2mt(k+1)}{N}}}\Big) \leq \frac{1}{2}\sum_{t=1}^{\infty} e^{\frac{-(\gamma_{d,i}^h)^2 - m^2t^2 - 2\gamma_{d,i}^h mt}{\frac{4mt(k+1)}{N}}} \\ &= \frac{e^{-\frac{N\gamma_{d,i}^h}{2(k+1)}}}{2}\Big(\sum_{t=1}^{\lfloor \frac{\gamma_{d,i}^h}{2m} \rfloor} e^{\frac{-N(\gamma_{d,i}^h)^2 - Nm^2t^2}{4mt(k+1)}} + \sum_{t=\lfloor \frac{\gamma_{d,i}^h}{2m} \rfloor + 1}^{\lfloor \frac{\gamma_{d,i}^h}{m} \rfloor} e^{\frac{-N(\gamma_{d,i}^h)^2 - Nm^2t^2}{4mt(k+1)}} \\ &+ \sum_{t=\lfloor \frac{\gamma_{d,i}^h}{m} \rfloor + 1}^{\lfloor \frac{2\gamma_{d,i}^h}{m} \rfloor} e^{\frac{-N(\gamma_{d,i}^h)^2 - Nm^2t^2}{4mt(k+1)}} + \sum_{t=\lfloor \frac{2\gamma_{d,i}^h}{m} \rfloor + 1}^{\infty} e^{\frac{-N(\gamma_{d,i}^h)^2 - Nm^2t^2}{4mt(k+1)}} \Big) \\ &\leq \frac{e^{-\frac{N\gamma_{d,i}^h}{2(k+1)}}}{2}\Big(e^{-\frac{N\gamma_{d,i}^h}{2(k+1)}} \underbrace{\sum_{t=1}^{\lfloor \frac{\gamma_{d,i}^h}{2m} \rfloor} e^{\frac{-Nmt}{4(k+1)}}}_{(1)} + e^{-\frac{N\gamma_{d,i}^h}{4(k+1)}} \underbrace{\sum_{t=\lfloor \frac{\gamma_{d,i}^h}{2m} \rfloor + 1}^{\lfloor \frac{\gamma_{d,i}^h}{m} \rfloor} e^{\frac{-Nmt}{4(k+1)}}}_{(2)} \\ &+ e^{-\frac{N\gamma_{d,i}^h}{8(k+1)}} \underbrace{\sum_{t=\lfloor \frac{\gamma_{d,i}^h}{m} \rfloor + 1}^{\lfloor \frac{2\gamma_{d,i}^h}{m} \rfloor} e^{\frac{-Nmt}{4(k+1)}}}_{(3)} + \underbrace{\sum_{t=\lfloor \frac{2\gamma_{d,i}^h}{m} \rfloor + 1}^{\infty} e^{\frac{-Nmt}{4(k+1)}}}_{(4)} \Big) \\ &\leq \frac{e^{-\frac{N\gamma_{d,i}^h}{2(k+1)}}}{2(1-e^{-\frac{Nm}{4(k+1)}})}\Big(e^{-\frac{N\gamma_{d,i}^h}{2(k+1)}} + e^{-\frac{N\gamma_{d,i}^h}{4(k+1)}} e^{-\frac{N\gamma_{d,i}^h}{8(k+1)}} + e^{-\frac{N\gamma_{d,i}^h}{8(k+1)}} e^{-\frac{N\gamma_{d,i}^h}{4(k+1)}} + e^{-\frac{N\gamma_{d,i}^h}{2(k+1)}} \Big) \\ &\leq \frac{2e^{-\frac{7N\gamma_{d,i}^h}{8(k+1)}}}{1-e^{-\frac{Nm}{4(k+1)}}}. \end{aligned} \tag{2.36}$$

In the above set of equations we use the fact that $\mathbb{Q}(x)$ is a non-increasing function, the inequality $\mathbb{Q}(x) \leq \frac{1}{2}e^{\frac{-x^2}{2}}$, and we upper bound (1) – (4) by their infinite geometric sums.

We now note that, a sufficient condition for $\mathbb{P}_{FA}^{d,i} \leq \alpha$ to hold is the following:

$$\frac{2e^{-\frac{7N\gamma_{d,i}^{h}}{8(k+1)}}}{1-e^{-\frac{Nm}{4(k+1)}}} \leq \alpha. \tag{2.37}$$

Solving (2.37), we have that, any $\gamma_{d,i}^{h}$ that satisfies

$$\gamma_{d,i}^{h} \geq \gamma_{d}^{h,0} = \frac{8(k+1)}{7N}\left(\log\left(\frac{2}{\alpha}\right) - \log(1-e^{-\frac{Nm}{4(k+1)}})\right), \tag{2.38}$$

achieves $\mathbb{P}_{FA}^{d,i} \leq \alpha$ in the $\mathcal{CI}SPRT$. Proceeding as in (2.34) and (2.36) we have that, any $\gamma_{d,i}^{l}$ that satisfies

$$\gamma_{d,i}^{l} \leq \gamma_{d}^{l,0} \doteq \frac{8(k+1)}{7N}\left(\log\left(\frac{\beta}{2}\right) + \log(1-e^{-\frac{Nm}{4(k+1)}})\right), \tag{2.39}$$

achieves $\mathbb{P}_{M}^{d,i} \leq \beta$ in the $\mathcal{CI}SPRT$.
Clearly, by the above, any pair $(\gamma_{d,i}^{h}, \gamma_{d,i}^{l})$ satisfying $\gamma_{d,i}^{h} \in [\gamma_{d}^{h,0}, \infty)$ and $\gamma_{d,i}^{l} \in (-\infty, \gamma_{d}^{l,0}]$ (see (2.38) and (2.39)) ensures that $\mathbb{P}_{FA}^{d,i} \leq \alpha$ and $\mathbb{P}_{M}^{d,i} \leq \beta$. The a.s. finiteness of the corresponding stopping time $T_{d,i}$ (see (C.43)) under both H_0 and H_1 follows readily by arguments as in (2.32). □

Remark 2.5.2. *It is to be noted that the derived thresholds are sufficient conditions only. The approximations (see (1) − (4) in (2.36)) made in the steps of deriving the expressions of the thresholds were done so as to get a tractable expression of the range. By solving the following set of equations*

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{\infty} e^{\frac{-N(\gamma_{d,i}^{l})^2 - Nm^2t^2 + 2N\gamma_{d,i}^{l}mt}{4mt(k+1)}} &\leq \beta \\
\frac{1}{2}\sum_{t=1}^{\infty} e^{\frac{-N(\gamma_{d,i}^{h})^2 - Nm^2t^2 - 2N\gamma_{d,i}^{h}mt}{4mt(k+1)}} &\leq \alpha
\end{aligned} \tag{2.40}$$

numerically, tighter thresholds can be obtained.

The first assertion ensures that for any set of pre-specified error metrics α and β (satisfying Assumption 2.3.2), the $\mathcal{CI}SPRT$ can be designed to achieve the error requirements while ensuring finite stopping a.s. The factor k in the closed form expressions of the thresholds in (C.36) and (C.40) relates the value of the thresholds to the rate of flow of information r and, hence, in turn, can be related to the degree of connectivity of the inter-agent communication graph under consideration, see (2.19)-(2.20) and the accompanying discussion. From Assumption 2.4.2, we have that $r < 1$. As r goes smaller, which intuitively means increased rate of flow of information in the inter-agent network, the value of thresholds needed to achieve the pre-specified error metrics become smaller i.e. the interval $[\gamma_{d,i}^{l}, \gamma_{d,i}^{h}]$ shrinks for all $i = 1, 2, \cdots, N$.

Remark 2.5.3. *We remark the following: 1) We have shown that the $\mathcal{CI}SPRT$ algorithm can be designed so as to achieve the pre-specified error metrics at every agent i. This, in turn, implies that the probability of not reaching decision consensus among the agents can be upper bounded by $N\beta$ when conditioned on H_1 and $N\alpha$ when conditioned on H_0. It is to be noted that with $\alpha \to 0$ and $\beta \to 0$, the probability of not reaching decision consensus conditioned on either of the hypothesis goes to 0 as well; 2) The factor k in the closed form expressions of the thresholds in (C.36) and (C.40) relates the value of the thresholds to the rate of flow of information r and, hence, in turn, can be related to the degree of connectivity of the inter-*

agent communication graph under consideration, see (2.19)-(2.20) *and the accompanying discussion. From Assumption 2.4.2, we have that* $r < 1$. *As* r *goes smaller, which intuitively means increased rate of flow of information in the inter-agent network, the value of thresholds needed to achieve the pre-specified error metrics become smaller i.e. the interval* $[\gamma_{d,i}^l, \gamma_{d,i}^h]$ *shrinks for all* $i = 1, 2, \ldots, N$.

2.5.2 Probability Distribution of $T_{d,i}$ and T_c

We first characterize the stopping time distributions for the centralized SPRT detector (see Section 2.3.2) and those of the distributed $\mathcal{CI}SPRT$. Subsequently, we compare the centralized and distributed stopping times by studying their respective large deviation tail probability decay rates.

Theorem 2.5.4. *(Darling and Siegert (1953); Hieber and Scherer (2012)) Let the Assumptions 2.3.1 and 2.3.2 hold and given the SPRT for the centralized setup in* (2.10)-(2.12)*, we have*

$$\mathbb{P}_1(T_c > t) \geq \exp\left(\frac{N\mu\gamma_c^l}{\sigma^2}\right) K_t^{\infty}\left(\gamma_c^h\right) - \exp\left(\frac{N\mu\gamma_c^h}{\sigma^2}\right) K_t^{\infty}\left(\gamma_c^l\right), \tag{2.41}$$

where

$$K_t^S(a) = \frac{\sigma^2\pi}{N(\gamma_c^h - \gamma_c^l)^2} \sum_{s=1}^{S} \frac{l(-1)^{l+1}}{\frac{Nm}{4} + \frac{\sigma^2 s^2\pi^2}{2N(\gamma_c^h - \gamma_c^l)^2}} \exp\left(-\left(\frac{Nm}{4} + \frac{\sigma^2 s^2\pi^2}{2N(\gamma_c^h - \gamma_c^l)^2}\right) t\right) \sin\left(\frac{s\pi a}{\gamma_c^h - \gamma_c^l}\right), \tag{2.42}$$

whereas, T_c *is defined in* (2.11) *and* γ_c^h *and* γ_c^l *are the associated SPRT thresholds chosen to achieve specified error requirements* α *and* β.

The above characterization of the stopping distribution of Wald's SPRT was obtained in Darling and Siegert (1953); Hieber and Scherer (2012). In particular, this was derived by studying the first passage time distribution of an associated continuous time Wiener process with a constant drift; intuitively, the continuous time approximation of the discrete time SPRT consists of replacing the discrete time likelihood increments by a Wiener process accompanied by a constant drift that reflects the mean of the hypothesis in place. This way, the sequence obtained by sampling the continuous time process at integer time instants is equivalent in distribution to the (discrete time) Wald's SPRT. The term on the R.H.S. of (2.41) is exactly equal to the probability that the first passage time of the continuous time Wiener process with left and right boundaries γ_c^l and γ_c^h respectively is greater than t, whereas, is, in general, a lower bound for the discrete time SPRT (as given in Theorem 2.5.4) as increments in the latter happen at discrete (integer) time instants only.
We now provide a characterization of the stopping time distributions of the $\mathcal{CI}SPRT$ algorithm.

Lemma 2.5.5. *Let the assumptions 2.3.1-2.4.2 hold. Consider the* $\mathcal{CI}SPRT$ *algorithm given in* (2.18)*,* (C.43) *and* (C.44) *and suppose that, for specified* α *and* β*, the thresholds* $\gamma_{d,i}^h$ *and* $\gamma_{d,i}^l$*,* $i = 1, \cdots, N$*, are chosen to satisfy the condtions derived in* (C.36) *and* (C.40)*. We then have,*

$$\mathbb{P}_1(T_{d,i} > t) \leq \mathbb{Q}\left(\frac{-\gamma_{d,i}^h + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\right), \quad \forall i = 1, 2, \ldots, N, \tag{2.43}$$

where $T_{d,i}$ *is the stopping time of the* i*-th agent to reach a decision as defined in* (C.43).

2.5.3 Comparison of stopping times of the distributed and centralized detectors

In this section we compare the stopping times T_c and $T_{d,i}$ by studying their respective large deviation tail probability decay rates. We utilize the bounds derived in Theorem 2.5.4 and Lemma 2.5.5 to this end.

Corollary 2.5.6. *Let the hypotheses of Lemma 2.5.4 hold. Then we have the following large deviation characterization for the tail probabilities of T_c:*

$$\liminf_{t\to\infty} \frac{1}{t}\log(\mathbb{P}_1(T_c > t)) \geq -\frac{Nm}{4} - \frac{\sigma^2\pi^2}{2N(\gamma_c^h - \gamma^l)^2}. \tag{2.44}$$

It is to be noted that the exponent is a function of the thresholds γ_c^h and γ_c^l and with the decrease in the error constraints α and β, $\frac{Nm}{4} + \frac{\sigma^2\pi^2}{2N(\gamma_c^h-\gamma^l)^2} \approx \frac{Nm}{4}$.

Theorem 2.5.7. *Let the hypotheses of Lemma 2.5.5 hold. Then we have the following large deviation characterization for the tail probabilities of the $T_{d,i}$'s:*

$$\limsup_{t\to\infty} \frac{1}{t}\log(\mathbb{P}_1(T_{d,i} > t)) \leq -\frac{Nm}{4} \quad ,\forall i = 1, 2, \ldots, N. \tag{2.45}$$

Importantly, the upper bound for the large deviation exponent of the $\mathcal{CISPRT}$ in Theorem 2.5.7 is independent of the inter-agent communication topology as long as the connectivity conditions Assumptions 2.4.1-2.4.2 hold. Finally, in the asymptotic regime, i.e., as N goes to ∞, since $\frac{\sigma^2\pi^2}{2N(\gamma_c^h-\gamma^l)^2} = o(Nm)$, we have that the performance of the distributed $\mathcal{CISPRT}$ approaches that of the centralized SPRT, in the sense of stopping time tail exponents, as N tends to ∞.

2.5.4 Comparison of the expected stopping times of the centralized and distributed detectors

In this section we compare the expected stopping times of the centralized SPRT detector and the proposed $\mathcal{CISPRT}$ detector. Recall that $\mathbb{E}_j[T_{d,i}]$ and $\mathbb{E}_j[T_c]$ represent the expected stopping times for reaching a decision for the $\mathcal{CISPRT}$ (at an agent i) and its centralized counterpart respectively, where $j \in \{0, 1\}$ denotes the hypothesis on which the expectations are conditioned on. Without loss of generality we compare the expectations conditioned on Hypothesis H_1, similar conclusions (with obvious modifications) hold when the expectations are conditioned on H_0 (see also Section 2.3.2).
Also, for the sake of mathematical brevity and clarity, we approximate $\alpha = \beta = \epsilon$ in this subsection.
Recall from Section 2.3.2 and note that, at any instant of time t, the information σ-algebra $\mathcal{G}_{d,i}(t)$ at any agent i is a subset of $\mathcal{G}_c(t)$, the information σ-algebra of a (hypothetical) center, which has access to the data of all agents at all times. This implies that any distributed procedure (in particular the $\mathcal{CISPRT}$) can be implemented in the centralized setting, and, since $M(\epsilon)$ (see (2.9)) constitutes a lower bound on the expected stopping time of any sequential test achieving error probabilities $\alpha = \beta = \epsilon$, we have that

$$\frac{\mathbb{E}_1[T_{d,i}]}{\mathcal{M}(\epsilon)} \geq 1, \quad \forall \quad i = 1, 2, \ldots, N, \tag{2.46}$$

for all $\epsilon \in (0, 1/2)$. In order to provide an upper bound on the ratio $\mathbb{E}_1[T_{d,i}]/\mathcal{M}(\epsilon)$ and, hence, compare the performance of the proposed $\mathcal{CISPRT}$ detector with the optimal centralized detector, we first obtain a characterization of $\mathbb{E}_1[T_{d,i}]$ in terms of the algorithm thresholds as follows.

Theorem 2.5.8. *Let the assumptions 2.3.1-2.4.2 hold and let $\alpha = \beta = \epsilon$. Suppose that the thresholds of the $\mathcal{CISPRT}$ be chosen as $\gamma_{d,i}^h = \gamma_d^{h,0}$ and $\gamma_{d,i}^l = \gamma_d^{l,0}$ for all $i = 1, \cdots, N$, where $\gamma_d^{h,0}$ and $\gamma_d^{l,0}$ are defined in (C.36)-(C.40). Then, the stopping time $T_{d,i}$ of the $\mathcal{CISPRT}$ at an agent i satisfies*

$$\frac{(1-2\epsilon)\gamma_{d,i}^h}{m} - \frac{c}{m} \leq \mathbb{E}_1[T_{d,i}] \leq \frac{5\gamma_{d,i}^h}{4m} + \frac{1}{1 - e^{\frac{-Nm}{4(k+1)}}}, \tag{2.47}$$

where $k = Nr^2$, r is as defined in (2.19), and $c > 0$ is a constant that may be chosen to be independent of the thresholds and the ϵ.

It is to be noted that, when $\alpha = \beta = \epsilon$, then $\gamma_{d,i}^h = -\gamma_{d,i}^l$ from (C.36) and (C.40). The upper bound derived in the above assertion might be loose, owing to the approximations related to the non-elementary $\mathbb{Q}$-function. We use the derived upper bound for comparing the performance of the $\mathcal{CISPRT}$ algorithm with that of its centralized counterpart. The constant $c > 0$ in the lower bound is independent of the thresholds $\gamma_{d,i}^l$ and $\gamma_{d,i}^h$ (and hence, also independent of the error tolerance ϵ) and is a function of the network topology and the Gaussian model statistics only. Explicit expressions and bounds on c may be obtained by refining the various estimates in the proofs of Theorem 2.5.8, see Section 2.5. However, for the current purposes, it is important to note that $c = o(\gamma_d^{h,0})$, i.e., as ϵ goes to zero or equivalently in the limit of large thresholds $c/\gamma_d^{h,0} \to 0$. Hence, as $\epsilon \to 0$, the more readily computable quantity $\frac{(1-2\epsilon)\gamma_{d,i}^h}{m}$ may be viewed as a reasonably good approximation to the lower bound in Theorem 2.5.8.

Theorem 2.5.9. *Let the hypotheses of Theorem 2.5.1 hold. Then, we have the following characterization of the ratio of the expected stopping times of the $\mathcal{CISPRT}$ and the centralized detector in asymptotics of the ϵ,*

$$1 \leq \limsup_{\epsilon \to 0} \frac{\mathbb{E}_1[T_{d,i}]}{\mathcal{M}(\epsilon)} \leq \frac{10(k+1)}{7}, \quad \forall i = 1, 2, \cdots, N, \tag{2.48}$$

where $k = Nr^2$ and r is as defined in (2.19).

In Liu and Mei (2017), the constants 10/7 and 8/7 have been improved to 1 and 1 respectively.
Theorem 2.5.9 shows that the $\mathcal{CISPRT}$ algorithm can be designed in such a way that with pre-specified error metrics α and β going to 0 , the ratio of the expected stopping time for the $\mathcal{CISPRT}$ algorithm and its centralized counterpart are bounded above by $\frac{10(k+1)}{7}$ where the quantity k depends on r which essentially quantifies the dependence of the $\mathcal{CISPRT}$ algorithm on the network connectivity.

Remark 2.5.10. *It is to be noted that the derived upper bound for the ratio of the expected stopping times of the $\mathcal{CISPRT}$ algorithm and its centralized counterpart may not be a tight upper bound. The looseness in the upper bound is due to the fact that the set of thresholds chosen are oriented to be sufficient conditions and not necessary. As pointed out in Remark 2.5.3 there might exist possibly better choice of thresholds for which the pre-specified error metrics are satisfied. Hence, given a set of pre-specified error metrics and a network topology the upper bound of the derived assertion above can be minimized by choosing the optimal weights for $\mathbf{W}$ as shown in Xiao and Boyd (2004). It can be seen that the ratio of expected stopping times of the isolated SPRT based detector case, i.e., the non-collaboration case, and the centralized SPRT based detector is N (see Section 2.3.2). So, for the $\mathcal{CISPRT}$ case in order to make savings as far as the stopping time is concerned with respect to the isolated SPRT based detector, $\frac{10(k+1)}{7} \leq N$ should be satisfied. Hence, we have that $r \leq \sqrt{\frac{7N-10}{10N}}$ is a sufficient condition for the same.*

2.6 Dependence of the $\mathcal{CISPRT}$ on Network Connectivity: Illustration

In this section, we illustrate the dependence of the $\mathcal{CISPRT}$ algorithm on the network connectivity, by considering a class of graphs. Recall from section 2.4 that the quantity r quantifies the *rate of information flow* in the network, and in general, the smaller the r the faster is the convergence of information dissemination algorithms (such as the consensus or gossip protocol (Dimakis et al. (2010); Kar and Moura (2008c, 2009)) on the graph and the optimal design of symmetric weight matrices $\mathbf{W}$ for a given network topology that minimizes the value r can be cast as a semi-definite optimization problem Xiao and Boyd (2004).
To quantify the dependence of the $\mathcal{CISPRT}$ algorithm on the graph topology, we note that the limit derived in (2.48) is a function of $\mathbf{W}$ and can be re-written as follows :

$$\limsup_{\epsilon\to 0}\frac{\mathbb{E}_1[T_{d,i}]}{\mathcal{M}(\epsilon)}\leq\frac{10(Nr^2+1)}{7}\doteq\mathcal{R}(\mathbf{W}), \tag{2.49}$$

i.e., the derived upper bound $\mathcal{R}(\mathbf{W})$ is a function of the chosen weight matrix W. Based on (2.49), naturally, a weight design guideline would be to design $\mathbf{W}$ (under the network topological constraints) so as to minimize $\mathcal{R}(\mathbf{W})$, which, by (2.49) and as discussed earlier corresponds to minimizing $r=\|\mathbf{W}-\mathbf{J}\|$. This leads to the following upper bound on the achievable performance of the $\mathcal{CISPRT}$:

$$\limsup_{\epsilon\to 0}\frac{\mathbb{E}_1[T_{d,i}]}{\mathcal{M}(\epsilon)}\leq\min_{\mathbf{W}}\mathcal{R}(\mathbf{W}). \tag{2.50}$$

By restricting attention to constant link weights, i.e., $\mathbf{W}$'s of the form $(\mathbf{I}-\delta\mathbf{L})$ and noting that

$$\min_{\delta}\|\mathbf{I}-\delta\mathbf{L}-\mathbf{J}\|=\frac{(\lambda_N(\mathbf{L})-\lambda_2(\mathbf{L}))}{(\lambda_2(\mathbf{L})+\lambda_N(\mathbf{L}))}, \tag{2.51}$$

we further obtain

$$\limsup_{\epsilon\to 0}\frac{\mathbb{E}_1[T_{d,i}]}{\mathcal{M}(\epsilon)}\leq\min_{\mathbf{W}}\mathcal{R}(\mathbf{W})\leq\min_{\delta}\mathcal{R}(\mathbf{I}-\delta\mathbf{L})=\frac{10}{7}+\frac{10N(\lambda_N(\mathbf{L})-\lambda_2(\mathbf{L}))^2}{7(\lambda_2(\mathbf{L})+\lambda_N(\mathbf{L}))^2}. \tag{2.52}$$

The final bound obtained in (2.52) might not be tight, being an upper bound (there may exist $\mathbf{W}$ matrices not of the form $\mathbf{I}-\delta\mathbf{L}$ with smaller r) to a possibly loose upper bound derived in (2.48), but, nonetheless, directly relates the performance of the $\mathcal{CISPRT}$ to the spectra of the graph Laplacian and hence the graph topology. From (2.52) we may further conclude that networks with smaller value of the ratio $\lambda_2(\mathbf{L})/\lambda_N(\mathbf{L})$ tend to achieve better performance. This leads to an interesting graph design question: given resource constraints, specifically, say a restriction on the number of edges of the graph, how to design inter-agent communication networks that tend to minimize the eigen-ratio $\lambda_2(\mathbf{L})/\lambda_N(\mathbf{L})$ so as to achieve improved $\mathcal{CISPRT}$ performance. To an extent, such graph design questions have been studied in prior work, see Kar et al. (2008), which, for instance, shows that expander graphs tend to achieve smaller $\lambda_2(\mathbf{L})/\lambda_N(\mathbf{L})$ ratios given a constraint on the total number of network edges.

2.7 Probability Distribution of T_c

Though the tails of the stopping time distributions for two sided random walks, i.e., the stopping time distribution of the centralized SPRT have been studied, as highlighted in Wald et al. (1945), the exact distribution is not well known. We provide a numerical way so as to calculate the stopping time distribution

of the centralized SPRT procedure. Define the event C_i as $-\gamma_c \leq S_c(i) \leq \gamma_c$.
Define the probability $\mathbb{P}(T_c > t)$ as follows:

$$\mathbb{P}(T_c > t) = \mathbb{P}\left(-\gamma_c \leq S_c(i) \leq \gamma_c, i = 1, 2, \ldots, t\right). \tag{2.53}$$

We note that $\mathbb{P}(C_i) = \mathbb{Q}\left(\frac{-\gamma_c - mi}{\sqrt{2mi/N}}\right) - \mathbb{Q}\left(\frac{\gamma_c - mi}{\sqrt{2mi/N}}\right)$ decreases with increasing i.
Observing that $S_c(i)$ form a Markov chain, (2.53) can be rewritten as follows:

$$\begin{aligned}\mathbb{P}(T_c > t) &= \mathbb{P}(C_1)\prod_{i=2}^{t}\mathbb{P}(C_i|C_{i-1}) \\ &= \frac{\prod_{i=2}^{t}\mathbb{P}(C_i \cap C_{i-1})}{\prod_{i=2}^{t-1}\mathbb{P}(C_i)}\end{aligned} \tag{2.54}$$

Denote $G'(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ and $G(x) = \int_{-\infty}^{x} e^{-t^2/2}dt$. Let $x_i = \mathbf{1}^T/N = S_c(i) - S_c(i-1)$.
Then, we have,

$$\begin{aligned}\mathbb{P}(C_i \cap C_{i-1}) &= \int_{-\gamma_c}^{\gamma_c} \frac{1}{\sqrt{4\pi m(i-1)/N}} e^{-N\frac{(y-m(i-1))^2}{4m(i-1)}} \int_{-\gamma_c - y}^{\gamma_c - y} \frac{1}{\sqrt{4\pi m/N}} e^{-N\frac{(x_i - m)^2}{4m}} dx_i dy \\ &= \int_{-\gamma_c}^{\gamma_c} \frac{1}{\sqrt{4\pi m(i-1)/N}} e^{-N\frac{(y-m(i-1))^2}{4m(i-1)}} \left(G\left(\frac{\gamma_c - y - m}{\sqrt{2m/N}}\right) - G\left(\frac{-\gamma_c - y - m}{\sqrt{2m/N}}\right)\right) dy \\ &= \int_{\frac{-\gamma_c - m(i-1)}{\sqrt{2m(i-1)/N}}}^{\frac{\gamma_c - m(i-1)}{\sqrt{2m(i-1)/N}}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \left(G\left(\frac{\gamma_c - t\sqrt{\frac{2m(i-1)}{N}} - mi}{\sqrt{2m/N}}\right) - G\left(\frac{-\gamma_c - t\sqrt{\frac{2m(i-1)}{N}} - mi}{\sqrt{2m/N}}\right)\right) dt \\ &= \int_{\gamma_{c,i}^l}^{\gamma_{c,i}^h} G'(t)(G(t\sqrt{i-1} + t_i^h) - G(t\sqrt{i-1} + t_i^l))dt\end{aligned} \tag{2.55}$$

where $\gamma_{c,i}^l = \frac{m(i-1)+\gamma_c}{\sqrt{2m(i-1)/N}}$, $\gamma_{c,i}^h = \frac{m(i-1)-\gamma_c}{\sqrt{2m(i-1)/N}}$, $t_i^h = \frac{\gamma_c - mi}{\sqrt{2m/N}}$ and $t_i^l = \frac{-\gamma_c - mi}{\sqrt{2m/N}}$.
We use the following identity from Owen (1956)

$$\begin{aligned}\int G'(x)G(a+bx)dt = T(x, \tfrac{a}{x\sqrt{1+b^2}}) + T(\tfrac{a}{\sqrt{1+b^2}}, \tfrac{x\sqrt{1+b^2}}{a}) - T(x, \tfrac{a+bx}{x}) \\ -T(\tfrac{a}{\sqrt{1+b^2}}, \tfrac{ab+x(1+b^2)}{a}) + G(x)G(\tfrac{a}{\sqrt{1+b^2}})\end{aligned} \tag{2.56}$$

where $T(h, a)$ is the Owen's T function which gives the probability of the event $\{X > h\} \cap \{0 < Y < aX\}$ where X and Y are independent random variables.
Using the identity in equation (2.56), (2.55) can be rewritten as follows:

$$\begin{aligned}\mathbb{P}(C_i \cap C_{i-1}) &= \int_{\gamma_{c,i}^l}^{\gamma_{c,i}^h} G'(t)(G(t\sqrt{i-1} + t_i^h) - G(t\sqrt{i-1} + t_i^l))dt \\ &= T\left(\gamma_{c,i}^h, \frac{t_i^h}{\gamma_{c,i}^h\sqrt{i}}\right) + T\left(\frac{t_i^h}{\sqrt{i}}, \frac{\gamma_{c,i}^h\sqrt{i}}{t_i^h}\right) - T\left(\gamma_{c,i}^h, \frac{t_i^h + \sqrt{i-1}\gamma_{c,i}^h}{\gamma_{c,i}^h}\right) \\ &- T\left(\frac{t_i^h}{\sqrt{i}}, \frac{t_i^h\sqrt{i-1} + \gamma_{c,i}^h i}{t_i^h}\right) + G\left(\gamma_{c,i}^h\right) G\left(\frac{t_i^h}{\sqrt{i}}\right)\end{aligned}$$

$$
\begin{aligned}
&- T\left(\gamma_{c,i}^{l}, \frac{t_i^h}{\gamma_{c,i}^{l}\sqrt{i}}\right) - T\left(\frac{t_i^h}{\sqrt{i}}, \frac{\gamma_{c,i}^{l}\sqrt{i}}{t_i^h}\right) + T\left(\gamma_{c,i}^{l}, \frac{t_i^h + \sqrt{i-1}\gamma_{c,i}^{l}}{\gamma_{c,i}^{l}}\right) \\
&+ T\left(\frac{t_i^l}{\sqrt{i}}, \frac{t_i^l\sqrt{i-1} + \gamma_{c,i}^{l} i}{t_i^h}\right) - G\left(\gamma_{c,i}^{l}\right) G\left(\frac{t_i^h}{\sqrt{i}}\right) \\
&- T\left(\gamma_{c,i}^{h}, \frac{t_i^l}{\gamma_{c,i}^{h}\sqrt{i}}\right) - T\left(\frac{t_i^l}{\sqrt{i}}, \frac{\gamma_{c,i}^{h}\sqrt{i}}{t_i^l}\right) + T\left(\gamma_{c,i}^{h}, \frac{t_i^l + \sqrt{i-1}\gamma_{c,i}^{h}}{\gamma_{c,i}^{h}}\right) \\
&+ T\left(\frac{t_i^l}{\sqrt{i}}, \frac{t_i^l\sqrt{i-1} + \gamma_{c,i}^{h} i}{t_i^l}\right) - G\left(\gamma_{c,i}^{h}\right) G\left(\frac{t_i^l}{\sqrt{i}}\right) \\
&+ T\left(\gamma_{c,i}^{l}, \frac{t_i^l}{\gamma_{c,i}^{l}\sqrt{i}}\right) + T\left(\frac{t_i^l}{\sqrt{i}}, \frac{\gamma_{c,i}^{l}\sqrt{i}}{t_i^l}\right) - T\left(\gamma_{c,i}^{l}, \frac{t_i^l + \sqrt{i-1}\gamma_{c,i}^{l}}{\gamma_{c,i}^{l}}\right) \\
&- T\left(\frac{t_i^l}{\sqrt{i}}, \frac{t_i^l\sqrt{i-1} + \gamma_{c,i}^{l} i}{t_i^l}\right) + G\left(\gamma_{c,i}^{l}\right) G\left(\frac{t_i^l}{\sqrt{i}}\right). \qquad (2.57)
\end{aligned}
$$

Though the stopping time distribution obtained above is rendered analytically intractable, various algorithms exist so as to evaluate Owen's T Function numerically.

2.8 Simulations

We generate planar random geometric networks of 30, 300 and 1000 agents. The x coordinates and the y coordinates of the agents are sampled from an uniform distribution on the open interval $(0, 1)$. We link two vertices by an edge if the distance between them is less than or equal to g. We go on re-iterating this procedure until we get a connected graph. We construct the geometric network for each of $N = 30, 300$ and 1000 cases with three different values of g i.e. $g = 0.3, 0.6$ and 0.9. The values of r obtained in each case is specified in Table 2.1. We consider two cases, the $\mathcal{CI}SPRT$ case and the non-collaborative case. We consider $\alpha = \beta = \epsilon$

r	g=0.3	g=0.6	g=0.9
N=30	0.8241	0.5580	0.2891
N=300	0.7989	0.6014	0.2166
N=1000	0.7689	0.5940	0.2297

Table 2.1: SPRT Comparison: Values of r

and ranging from 10^{-8} to 10^{-4} in steps of 10^{-6}. For each such ϵ, we conduct 2000 simulation runs to empirically estimate the stopping time distribution $\mathbb{P}_1(T > t)$ of a randomly chosen agent (with uniform selection probability) for each of the cases. From these empirical probability distributions of the stopping times, we estimate the corresponding expected stopping times. Figure 2.4 shows the instantaneous behavior of the test statistics in the case of $N = 300$ with $\epsilon = 10^{-10}$. In Figures 2.1, 2.2 and 2.3 it is demonstrated that the ratio of the expected stopping time of the $\mathcal{CI}SPRT$ algorithm and the universal lower bound $\mathcal{M}(\epsilon)$ is less than that of the ratio of the expected stopping times of the isolated (non-collaborative) case and $\mathcal{M}(\epsilon)$. The ratio of the theoretical lower bound of the expected stopping time of the $\mathcal{CI}SPRT$ derived in Theorem 2.5.8 and $\mathcal{M}(\epsilon)$ was also studied. More precisely, we compared the experimental ratio of the expected stopping times of the $\mathcal{CI}SPRT$ and $\mathcal{M}(\epsilon)$ with the ratio of the quantity $\frac{(1-2\epsilon)\gamma_{d,i}^{h}}{m}$ (the small ϵ approximation of the theoretical lower bound given in Theorem 2.5.8, see also the discussion provided in Section 2.5 after the statement of Theorem 2.5.8) and $\mathcal{M}(\epsilon)$. It can be seen that the experimental ratio

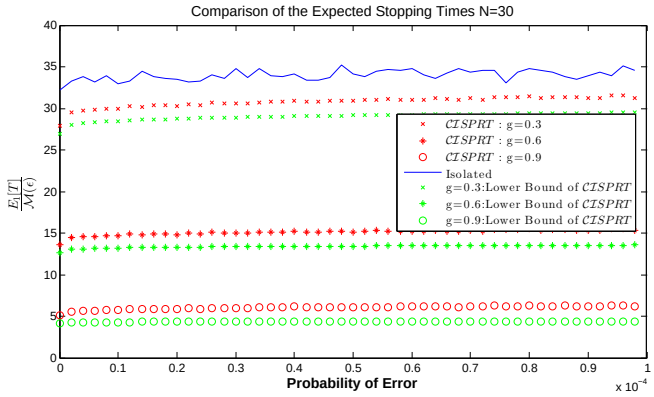


Figure 2.1: Comparison of Stopping Time Distributions for N=30

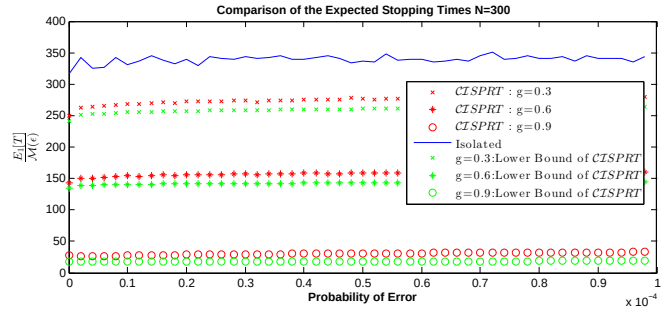


Figure 2.2: Comparison of Stopping Time Distributions for N=300

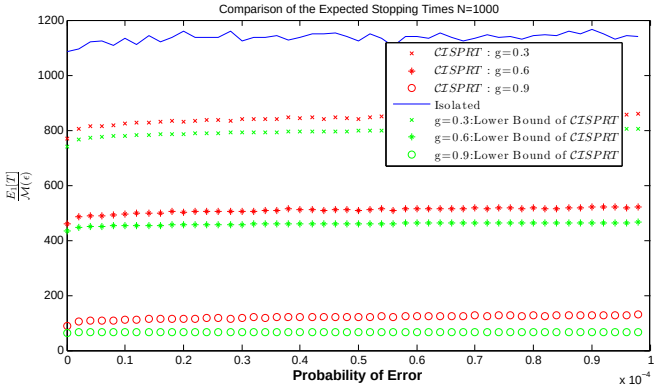


Figure 2.3: Comparison of Stopping Time Distributions for N=1000

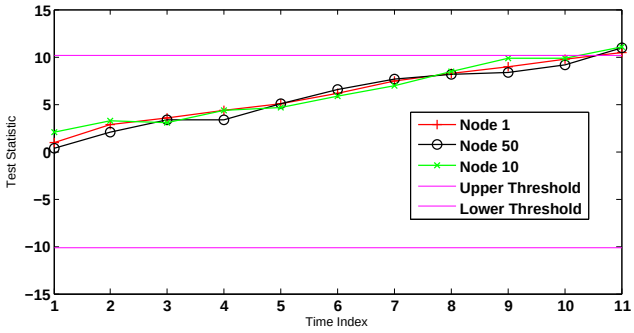


Figure 2.4: Instantaneous behavior of $S_{d,i}(t)$

of the expected stopping times of the $\mathcal{CISPRT}$ and $\mathcal{M}(\epsilon)$ is very close to the ratio of the (approximate) theoretical lower bound of expected stopping time of the $\mathcal{CISPRT}$ and $\mathcal{M}(\epsilon)$, which shows that the lower bound derived in Theorem 2.5.8 is reasonable. Figure 2.4 is an example of a single run of the algorithm which shows the instantaneous behavior of the distributed test statistic $S_{d,i}(t)$ for $N = 300$, where we have plotted three randomly chosen agents i.e. $i = 1$, $i = 10$ and $i = 50$.

2.9 Summary of Contributions

- **Finite Stopping Property:** We show that, given any value of probability of false alarm α and probability of miss β, the $\mathcal{CISPRT}$ algorithm can be designed such that each agent achieves the specified error performance metrics and the test procedure terminates in finite time almost surely (a.s.) at each agent. We derive closed form expressions for the local threshold parameters at the agents as functions of α and β which ensures that the $\mathcal{CISPRT}$ achieves the above property.

- **Asymptotic Characterization:** In the asymptotics of vanishing error metrics (i.e., as $\alpha, \beta \to 0$), we quantify the ratio of the expected stopping time $T_{d,i}(\alpha, \beta)$ for reaching a decision at an agent i through the $\mathcal{CISPRT}$ algorithm and the expected stopping time $T_c(\alpha, \beta)$ for reaching a decision by the optimal centralized (SPRT) procedure, i.e., the quantity $\frac{\mathbb{E}[T_{d,i}(\alpha,\beta)]}{\mathbb{E}[T_c(\alpha,\beta)]}$, which in turn is a metric of efficiency of the proposed algorithm as a function of the network connectivity.

2.10 Conclusion and Future Directions

In this chapter we have considered sequential detection of Gaussian binary hypothesis observed by a sparsely interconnected network of agents. The $\mathcal{CISPRT}$ algorithm we proposed combines two terms : a *consensus* term that updates at each sensor its test statistic with the test statistics provided by agents in its one-hop neighborhood and an *innovation* term that updates the current agent test statistic with the new local sensed information. We have shown that the $\mathcal{CISPRT}$ can be designed to achieve a.s. finite stopping at each network agent with guaranteed error performance. We have provided explicit characterization of its expected stopping time as a function of the network connectivity. The performance of the $\mathcal{CISPRT}$ was further benchmarked w.r.t. the optimal centralized sequential detector, the SPRT. An interesting future direction would be to consider networks with random time-varying topology. We also intend to develop extensions of the $\mathcal{CISPRT}$ for setups with correlated and non-linear non-Gaussian observation models.

Chapter 3

Distributed Composite Hypothesis Testing

3.1 Introduction

The focus of this chapter is on distributed composite hypothesis testing in multi-agent networks in which the goal is not only to estimate the state (possibly high dimensional) of the environment but also detect as to which hypothesis is in force based on the sensed information across all the agents at all times. To be specific, we are interested in the design of recursive detection algorithms to decide between a simple null hypothesis and a composite alternative parameterized by a continuous vector parameter, which exploit available sensing resources to the maximum and obtain reasonable detection performance, i.e., have asymptotically (in the large sample limit) decaying probabilities of errors. Technically speaking, we are interested in the study of algorithms which can process sensed information as and when they are sensed and not wait till the end until all the sensed data has been collected. To be specific, the sensed data refers to the observations made across all the agents at all times. The problem of composite hypothesis testing is relevant to many practical applications, including cooperative spectrum sensing Zarrin and Lim (2009); Font-Segura and Wang (2010); Zou et al. (2010) and MIMO radars Tajer et al. (2010), where the onus is also on achieving reasonable detection performance by utilizing as fewer resources as possible, which includes data samples, communication and sensing energy. In classical composite hypothesis testing procedures such as the Generalized Likelihood Ratio Test (GLRT) Zeitouni et al. (1992), the detection procedure which uses the underlying parameter estimate based on all the collected samples as a plug-in estimate may not be initiated until a reasonably accurate parameter estimate, typically the maximum likelihood estimate of the underlying parameter (state) is obtained. Usually in setups which employ the classical (centralized) generalized likelihood ratio tests, the data collection phase precedes the parameter estimation and detection statistic update phase which makes the procedure essentially an *offline* batch procedure. By *offline* batch procedures, we mean algorithms where the sensing phase precedes any kind of information processing and the entire data is processed in batches.[1] The motivation behind studying recursive online detection algorithms in contrast to offline batch processing based detection algorithms is that in most multi-agent networked scenarios, which are typically energy constrained, the priority is to obtain reasonable inference performance by expending fewer amount of resources. Moreover, in centralized scenarios, where the communication graph is all-to-all, the implementation suffers from high communication overheads, synchronization issues and high energy requirements. Motivated by

[1] We emphasize that, by offline, we strictly refer to the classical implementation of the GLRT. Recursive variants of GLRT type approaches have been developed for a variety of testing problems including sequential composite hypothesis testing and change detection (see, for example, Siegmund and Venkatraman (1995); Willsky and Jones (1976); Chang and Dunn (1979)), although in centralized processing scenarios.

requirements such as the latter, we propose distributed recursive composite hypothesis testing algorithms, where the inter-agent collaboration is restricted to a pre-assigned, possibly sparse communication graph and the detection and estimation schemes run in a parallel fashion with a view to reduce energy and resource consumption while achieving reasonable detection performance.

In the domain of hypothesis testing, when one of the hypotheses is composite, i.e., the hypothesis is parameterized by a continuous vector parameter and the underlying parameter is unknown apriori, one of the most well-known algorithms is the Generalized Likelihood Ratio Testing (GLRT). The GLRT has an estimation procedure built into it, where the underlying parameter estimate is used as a plug-in estimate for the decision statistic. In a centralized setting or in a scenario where the inter-agent communication graph is all-to-all, the fusion center has access to all the sensed information and the parameter estimates across all the agents at all times. The procedure of obtaining the underlying parameter estimate, which in turn employs a maximization, achieves reasonable performance in general, but has a huge communication overhead which makes it infeasible to be implemented in practice, especially in networked environments. In contrast to the fully centralized setup, we focus on a fully distributed setup where the communication between the agents is restricted to a pre-assigned possibly sparse communication graph. In this chapter, we propose two algorithms namely, *consensus + innovations* GLRT Non-Linear ($\mathcal{CIGLRT}-\mathcal{NL}$) and *consensus + innovations* GLRT Linear ($\mathcal{CIGLRT}-\mathcal{L}$), which are of the *consensus + innovations* form and are based on fully distributed setups. We specifically focus on a setting in which the agents obtain conditionally Gaussian and independent and identically distributed observations and update their parameter estimates and decision statistics by simultaneous assimilation of the information obtained from the neighboring agents (*consensus*) and the latest locally sensed information (*innovation*). Also similar, to the classical GLRT, both of our algorithms involve a parameter estimation scheme and a detection algorithm. This justifies the names $\mathcal{CIGLRT}-\mathcal{L}$ and $\mathcal{CIGLRT}-\mathcal{NL}$ which are distributed GLRT type algorithms of the *consensus + innovations* form. In this chapter, so as to replicate typical practical sensing environments accurately, we model the underlying vector parameter as a static parameter, whose dimension is M (possibly large) and every agent's observations, say for agent n, is M_n dimensional, where $M_n << M$, thus rendering the parameter locally unobservable at each agent. We show that, under a minimal global observability condition imposed on the collective observation model and connectedness of the communication graph, the parameter estimate sequences are consistent and the detection schemes achieve asymptotically decaying probabilities of errors in the large sample limit.

3.2 Related Work

Existing work in the literature on distributed detectors can be broadly classified into three classes. The first class includes architectures which are characterized by presence of a fusion center and all the agents transmit their decision or local measurements or test statistics or its quantized version to the fusion center (see, for example Blum et al. (1997); Tsitsiklis et al. (1993)) and subsequently the estimation and detection schemes are conducted by the fusion center. The second class consists of consensus schemes (see, for example Kar and Moura (2007); Olfati-Saber et al. (2006)) with no fusion center and in which in the first phase the agents collect information over a long period of time from the environment followed by the second phase, in which agents exchange information (through consensus or gossip type procedures Kar and Moura (2007); Olfati-Saber et al. (2007); Jadbabaie et al. (2003)) in their respective neighborhoods which are in turn specified by a pre-assigned communication graph or a sequence of possibly sparse time-varying communication graphs satisfying appropriate connectivity conditions. The third class consists of schemes which perform simultaneous assimilation of information obtained from sensing and communication (see,

for example Bajovic et al. (2011); Kar et al. (2011); Cattivelli and Sayed (2011b)). Distributed detection algorithms from the third class of detectors described above, can be further sub-categorized to three classes, namely the running consensus approach Braca et al. (2008, 2010), the diffusion approach Cattivelli and Sayed (2009a,b, 2011b); Zou et al. (2010) and the *consensus+innovations* approach Bajovic et al. (2011); Jakovetic et al. (2012); Kar et al. (2011). These works address important questions pertaining to binary simple hypothesis and also characterize the fundamental limits of the detection scheme through large deviations analysis. Other relevant recent work include Lalitha et al. (2014, 2015); Nedić et al. (2014). However, as compared to the aforementioned works, the objective of the detection scheme in this chapter is to decide between a simple null hypothesis and a composite alternative which is parameterized by a vector parameter which can take values in a continuous space. In the context of distributed composite hypothesis testing, the proposed algorithm involves a recursive parameter estimation update and a decision statistic update running in parallel. The proposed algorithms, $\mathcal{CIGLRT}-\mathcal{NL}$ and $\mathcal{CIGLRT}-\mathcal{L}$ involve parameter estimate updates in a non-linear observation model and a linear observation model setting respectively. It is to be noted that in the context of centralized detection literature, the weak convergence of the decision statistics under the null and the alternate hypothesis (see, Wilks (1938) for example) is usually not enough to establish the decay rates of the probability of errors. To be specific, in this chapter we extend Wilks' theorem to the distributed recursive setup and characterize the asymptotic normality of the decision statistic sequence. However, the statistical dependencies exhibited in the decision static update due to the parameter estimation scheme and the decision statistic update running in a parallel fashion warrants the development of technical machinery so as to address concentration of measure for sums of non i.i.d random variables which in turn helps characterize the decay exponent of the probability of errors which we develop in this chapter.

3.3 Problem Formulation

3.3.1 System Model and Preliminaries

There are N agents deployed in the network. Every agent n at time index t makes a noisy observation $\mathbf{y}_n(t)$. The observation $\mathbf{y}_n(t)$ is a M_n-dimensional vector, a noisy nonlinear function of θ^* which is a M-dimensional parameter, i.e., $\theta^* \in \mathbb{R}^M$. The observation $\mathbf{y}_n(t)$ comes from a probability distribution $\mathbb{P}_0$ under the hypothesis $\mathcal{H}_0$, whereas, under the composite alternative $\mathcal{H}_1$, the observation is sampled from a probability distribution which is a member of a parametric family $\{\mathbb{P}_{\theta^*}\}$. We emphasize here that the parameter θ^* is deterministic but unknown. Formally,

$$\begin{aligned}\mathcal{H}_1 &: \mathbf{y}_n(t) = \mathbf{h}_n(\theta^*) + \gamma_n(t)\\ \mathcal{H}_0 &: \mathbf{y}_n(t) = \gamma_n(t),\end{aligned} \tag{3.1}$$

where $\mathbf{h}_n(.)$ is, in general, non-linear function, $\{\mathbf{y}_n(t)\}$ is a $\mathbb{R}^{M_n}$-valued observation sequence for the n-th agent, where typically $M_n << M$ and $\{\gamma_n(t)\}$ is a zero-mean temporally i.i.d Gaussian noise sequence at the n-th agent with nonsingular covariance matrix $\mathbf{\Sigma}_n$, where $\mathbf{\Sigma}_n \in \mathbb{R}^{M_n \times M_n}$. Moreover, the noise sequences at two agents n, l with $n \neq l$ are independent.

By taking $\mathbf{h}_n(\mathbf{0}) = \mathbf{0}$, $\forall n$ and certain other identifiability and regularity conditions outlined below, in the above formulation the null hypothesis corresponds to $\theta^* = 0$ and the composite alternative to the case $\theta^* \neq \mathbf{0}$.

Since, the sources of randomness in our formulation are the observations $\mathbf{y}_n(t)$'s made by the agents in the

network, we define the natural filtration $\{\mathcal{F}_t\}$ generated by the random observations, i.e.,

$$\mathcal{F}_t = \sigma\left(\{\{\mathbf{y}_n(s)\}_{n=1}^N\}_{s=0}^{t-1}\right), \tag{3.2}$$

which is the sequence of σ-algebras induced by the observation processes, in order to model the overall available network information at all times. Finally, a stochastic process $\{\mathbf{x}(t)\}$ is said to be $\{\mathcal{F}_t\}$-adapted if the σ-algebra $\sigma(\mathbf{x}(t))$ is a subset of $\mathcal{F}_t$ at each t.

3.3.2 Preliminaries : Generalized Likelihood Ratio Tests

Consider, for instance, a generalized target detection problem in which the absence of target is modeled by a simple hypothesis $\mathcal{H}_0$, whereas, its presence corresponds to a composite alternative $\mathcal{H}_1$, as it is parametrized by a continuous vector parameter (perhaps modeling its location and other attributes) which is unknown apriori. Let $\mathbf{y}(t)$ denote the collection of the data from the agents, i.e., $\mathbf{y}(t) = \left[\mathbf{y}_1^\top(t)\cdots\mathbf{y}_N^\top(t)\right]^\top$, at time t, which is $\sum_{n=1}^N M_n$ dimensional. Specifically, the GLRT decision procedure decides on the hypothesis[2] as follows:

$$\mathcal{H} = \begin{cases} \mathcal{H}_1, & \text{if} \quad \max_\theta \sum_{t=0}^T \log \frac{f_\theta(\mathbf{y}(t))}{f_0(\mathbf{y}(t))} > \eta, \\ \mathcal{H}_0, & \text{otherwise,} \end{cases} \tag{3.3}$$

where η is a predefined threshold, T denotes the number of sensed observations and assuming that the data from the agents are conditionally independent $f_\theta(\mathbf{y}(t)) = f_\theta^1(\mathbf{y}_1(t))\cdots f_\theta^N(\mathbf{y}_N(t))$ denotes the likelihood of observing $\mathbf{y}(t)$ under $\mathcal{H}_1$ and realization θ of the parameter and $f_\theta^n(\mathbf{y}_n(t))$ denotes the likelihood of observing $\mathbf{y}_n(t)$ at the n-th agent under $\mathcal{H}_1$ and realization θ of the parameter ; similarly, $f_0(\mathbf{y}(t)) = f_0^1(\mathbf{y}_1(t))\cdots f_0^N(\mathbf{y}_N(t))$ denotes the likelihood of observing $\mathbf{y}(t)$ under $\mathcal{H}_0$ and $f_0^N(\mathbf{y}_N(t))$ denotes the likelihood of observing $\mathbf{y}_n(t)$ at the n-th agent under $\mathcal{H}_0$. The key bottleneck in the implementation of the classical GLRT as formulated in (3.3) is the maximization

$$\max_\theta \sum_{t=0}^T \log \frac{f_\theta(\mathbf{y}(t))}{f_0(\mathbf{y}(t))} = \max_\theta \sum_{t=0}^T \sum_{n=1}^N \log \frac{f_\theta^n(\mathbf{y}_n(t))}{f_0^n(\mathbf{y}_n(t))} \tag{3.4}$$

which involves the computation of the generalized log-likelihood ratio, i.e., the decision statistic. In general, a maximizer of (3.4) is not known beforehand as it depends on the entire sensed data collected across all the agents at all times, and hence as far as communication complexity in the GLRT implementation is concerned, the maximization step incurs the major overhead – in fact, a direct implementation of the maximization (3.4) requires access to the entire raw data $\mathbf{y}(t)$ at all times t at the fusion center.

3.4 Distributed Generalized Likelihood Ratio Testing

To mitigate the communication overhead, we present distributed message passing schemes in which agents, instead of forwarding raw data to a fusion center, participate in a collaborative iterative process to obtain a maximizing θ. The agents also maintain a copy of their local decision statistic, where the decision statistic is updated by assimilating local decision statistics from the neighborhood and the latest sensed information. In order to obtain reasonable decision performance with such localized communication, we propose a distributed

[2]It is important to note that the considered setup does not admit uniquely most powerful tests Scharf (1991).

detector of the *consensus + innovations* type. To this end, we propose two algorithms, namely
1) $\mathcal{CIGLRT}-\mathcal{NL}$, which is a general algorithm based on a non-linear observation model with additive Gaussian noise. We specifically show that the decision errors go to zero asymptotically as time $t \to \infty$ or equivalently, in the large sample limit, if the thresholds are chosen appropriately, and
2) $\mathcal{CIGLRT}-\mathcal{L}$, where we specifically consider a linear observation model. In the case of $\mathcal{CIGLRT}-\mathcal{L}$, we not only show that the probabilities of errors go to zero asymptotically, but also, we characterize the large deviations exponent upper bounds for the probabilities of errors arising from the decision scheme under minimal assumptions of global observability and connectedness of the communication graph.
The algorithms $\mathcal{CIGLRT}-\mathcal{NL}$ and $\mathcal{CIGLRT}-\mathcal{L}$ are motivated from the bottlenecks that one would encounter from centralized batch processing. To be specific, if a hypothetical fusion center which had access to all the agents' observations at all times were to conduct the parameter estimation in a recursive way, it would do so in the following way:

$$\begin{aligned}\theta_c(t+1) &= \theta_c(t) \\ &+ \underbrace{\alpha_t \sum_{n=1}^{N} \nabla \mathbf{h}_n\left(\theta_c(t)\right) \mathbf{\Sigma}_n^{-1}\left(\mathbf{y}_n(t) - \mathbf{h}_n\left(\theta_c(t)\right)\right)}_{\text{Global Innovation}},\end{aligned}$$

where $\{\theta_c(t)\}$ represents the centralized estimate sequence. Similarly, the centralized decision statistic update can be represented as follows:

$$z_c(t+1) = \frac{t}{t+1} z_c(t) + \underbrace{\frac{1}{t+1} \sum_{n=1}^{N} \log \frac{f_{\theta_c(t)}(y_n(t))}{f_0(y_n(t))}}_{\text{Global Innovation}},$$

where $\{z_c(t)\}$ represents the centralized decision statistic sequence. It is to be noted that the centralized scheme may not be implementable in our distributed multi-agent setting with sparse inter-agent interaction primarily due to the fact that the desired global innovation computation requires instantaneous access to the entire set of network sensed data at all times at a central computing resource. If in the case of a distributed setup, an agent n in the network were to replicate the centralized update by replacing the global innovation in accordance with its local innovation, the updates for the parameter estimate and the decision statistic would be as follows:

$$\widehat{\theta}_n(t+1) = \widehat{\theta}_n(t) + \underbrace{\alpha_t \nabla \mathbf{h}_n\left(\widehat{\theta}_n(t)\right) \mathbf{\Sigma}_n^{-1}\left(\mathbf{y}_n(t) - \mathbf{h}_n\left(\widehat{\theta}_n(t)\right)\right)}_{\text{Local Innovation}},$$

where $\left\{\widehat{\theta}_n(t)\right\}$ represents the estimate sequence at agent n. Similarly, the decision statistic update at agent n would have been:

$$\widehat{z}_n(t+1) = \frac{t}{t+1} \widehat{z}_n(t) + \underbrace{\frac{1}{t+1} \log \frac{f_{\widehat{\theta}_n(t)}(y_n(t))}{f_0(y_n(t))}}_{\text{Local Innovation}},$$

where $\{\hat{z}_n(t)\}$ represents the decision statistic sequence at agent n. The above correspond to purely decentralized local processing with no inter-agent collaboration whatsoever. However, note that in absence of

local observability both the parameter estimates and decision statistics would be erroneous and sub-optimal. Hence, as a surrogate to the global innovation in the centralized recursions, the local estimators compute a local innovation based on the locally sensed data as an agent has access to the information in its neighborhood. However, they intend to compensate for the resulting information loss by incorporating an agreement or consensus potential into their updates in which the individual estimators.
We, first present the algorithm $\mathcal{CIGLRT}-\mathcal{NL}$.

3.5 Non-linear Observation Models : Algorithm $\mathcal{CIGLRT}-\mathcal{NL}$

We start by making some identifiability assumptions on our sensing model before stating the algorithm.

Assumption 3.5.1. *The sensing model is globally observable, i.e., any two distinct values of θ and θ^* in the parameter space $\mathbb{R}^M$ satisfy*

$$\sum_{n=1}^{N} \|\mathbf{h}_n(\theta) - \mathbf{h}_n(\theta^*)\|^2 = 0 \tag{3.5}$$

if and only if $\theta = \theta^*$.

We propose a distributed detector of the *consensus+innovations* form for the scenario outlined in (3.1). Before discussing the details of our algorithm, we state an assumption on the inter-agent communication graph.

Assumption 3.5.2. *The inter-agent communication graph is connected, i.e., $\lambda_2(\mathbf{L}) > 0$, where $\mathbf{L}$ denotes the associated graph Laplacian matrix.*

We now present the distributed $\mathcal{CIGLRT}-\mathcal{NL}$ algorithm. The sequential decision procedure consists of three interacting recursive processes operating in parallel, namely, a parameter estimate update process, a decision statistic update process, and a detection decision formation rule, as described below. We state an assumption on the sensing functions before stating the algorithm.

Assumption 3.5.3. *For each agent n, $\forall \theta \neq \theta_1$, the sensing functions $\mathbf{h}_n$ are continuously differentiable on $\mathbb{R}^M$ and Lipschitz continuous with constants $k_n > 0$, i.e.,*

$$\|\mathbf{h}_n(\theta) - \mathbf{h}_n(\theta_1)\| \leq k_n \|\theta - \theta_1\|. \tag{3.6}$$

Parameter Estimate Update. The algorithm $\mathcal{CIGLRT}-\mathcal{NL}$ generates a sequence $\{\theta_n(t)\} \in \mathbb{R}^M$ of estimates of the parameter θ^* at the n-th agent according to the distributed recursive scheme

$$\begin{aligned}\theta_n(t+1) = \theta_n(t) &- \underbrace{\beta_t \sum_{l\in\Omega_n} (\theta_n(t) - \theta_l(t))}_{\text{neighborhood consensus}} \\ &+ \underbrace{\alpha_t \nabla \mathbf{h}_n(\theta_n(t))\, \mathbf{\Sigma}_n^{-1} (\mathbf{y}_n(t) - \mathbf{h}_n(\theta_n(t)))}_{\text{local innovation}},\end{aligned} \tag{3.7}$$

where Ω_n denotes the communication neighborhood of agent n and $\nabla \mathbf{h}_n(.)$ denotes the gradient of $\mathbf{h}_n$, which is a matrix of dimension $\mathbf{M} \times \mathbf{M}_n$, with the (i,j)-th entry given by $\frac{\partial [\mathbf{h}_n(\theta_n(t))]_j}{\partial [\theta_n(t)]_i}$.

The update in (3.7) can be written in a compact manner as follows:

$$\begin{aligned}\theta(t+1) &= \theta(t) - \beta_t \left(\mathbf{L} \otimes \mathbf{I}_M\right) \theta(t) \\ &+ \alpha_t \mathbf{G}(\theta(t)) \mathbf{\Sigma}^{-1} \left(\mathbf{y}(t) - \mathbf{h}\left(\theta(t)\right)\right),\end{aligned} \tag{3.8}$$

where $\theta(t) = \left[\theta_1^\top(t) \cdots \theta_N^\top(t)\right]^\top$, $\mathbf{h}(\theta(t)) = \left[\mathbf{h}_1^\top(\theta_1(t)) \cdots \mathbf{h}_N^\top(\theta_N(t))\right]^\top$, $\mathbf{\Sigma}^{-1} = diag\left[\mathbf{\Sigma}_1^{-1}, \cdots, \mathbf{\Sigma}_N^{-1}\right]$ and $\mathbf{G}\left(\theta(t)\right) = diag\left[\nabla \mathbf{h}_1\left(\theta_1(t)\right), \cdots, \nabla \mathbf{h}_N\left(\theta_N(t)\right)\right]$.

Assumption 3.5.4. *There exists a constant $c_1 > 0$ for each pair of θ and $\acute{\theta}$ with $\theta \neq \acute{\theta}$ such that the following aggregate strict monotonicity condition holds*

$$\sum_{n=1}^{N} \left(\theta - \acute{\theta}\right)^\top \left(\nabla \mathbf{h}_n\left(\theta\right)\right) \mathbf{\Sigma}_n^{-1} \left(\mathbf{h}_n\left(\theta\right) - \mathbf{h}_n\left(\acute{\theta}\right)\right) \geq c_1 \left\|\theta - \acute{\theta}\right\|^2. \tag{3.9}$$

For example, in assumption 3.5.4, if $h_n(.)$'s are linear, i.e., $h_n(\theta^*) = \mathbf{H}_n\theta^*$, where $\mathbf{H}_n \in \mathbb{R}^{M_n \times M}$ the monotonicity condition is trivially satisfied by the positive definiteness of the matrix $\sum_{n=1}^{N} \mathbf{H}_n^\top \mathbf{\Sigma}_n^{-1} \mathbf{H}_n$.
We make the following assumption on the weight sequences $\{\alpha_t\}$ and $\{\beta_t\}$:

Assumption 3.5.5. *The weight sequences $\{\alpha_t\}_{t\geq 0}$ and $\{\beta_t\}_{t\geq 0}$ are given by*

$$\alpha_t = \frac{a}{(t+1)} \;\; \beta_t = \frac{b}{(t+1)^{\tau_2}}, \tag{3.10}$$

where $ac_1 \geq 1$ with c_1 is as defined in Assumption 3.5.4.

where $0 < \tau_2 < 1/2, b > 0$.

Decision Statistic Update. The algorithm $\mathcal{CIGLRT} - \mathcal{L}$ generates a scalar-valued decision statistic sequence $\{z_n(t)\}$ at the n-th agent according to the distributed recursive scheme

$$\begin{aligned}z_n(t+1) = \frac{t}{t+1} &\left(z_n(t) - \underbrace{\delta \sum_{l \in \Omega_n} \left(z_n(t) - z_l(t)\right)}_{\text{neighborhood consensus}} \right) \\ &+ \underbrace{\frac{1}{t+1} \log \frac{f_{\theta_n(t)}(y_n(t))}{f_0(y_n(t))}}_{\text{local innovation}},\end{aligned} \tag{3.11}$$

where $f_\theta(.)$ and $f_0(.)$ represent the likelihoods under $\mathcal{H}_1$ and $\mathcal{H}_0$ respectively,

$$\delta \in \left(0, \frac{2}{\lambda_N\left(\mathbf{L}\right)}\right). \tag{3.12}$$

However, we specifically choose $\delta = \frac{2}{\lambda_2(\mathbf{L}) + \lambda_N(\mathbf{L})}$ for subsequent analysis.

Decision Rule. The following decision rule is adopted at all times t at all agents n :

$$\mathcal{H}_n(t) = \begin{cases} \mathcal{H}_0 & z_n(t) \leq \eta \\ \mathcal{H}_1 & z_n(t) > \eta, \end{cases} \tag{3.13}$$

where $\mathcal{H}_n(t)$ denotes the local selection (decision) at agent n at time t. Under the aegis of such a decision rule, the associated probability of errors are as follows:

$$\begin{aligned}\mathbb{P}_{M,\theta^*}(t) &= \mathbb{P}_{1,\theta^*}\left(z_n(t) \leq \eta\right)\\ \mathbb{P}_{FA}(t) &= \mathbb{P}_0\left(z_n(t) > \eta\right),\end{aligned} \tag{3.14}$$

where $\mathbb{P}_{M,\theta^*}$ and $\mathbb{P}_{FA}$ refer to probability of miss and probability of false alarm respectively. We refer to the parameter estimate update, the decision statistic update and the decision rule in (3.8), (3.11) and (3.13) respectively, as the $\mathcal{CIGLRT}-\mathcal{NL}$ algorithm.

3.6 Linear Observation Models : Algorithm $\mathcal{CIGLRT}-\mathcal{L}$

In this section, we develop the algorithm $\mathcal{CIGLRT}-\mathcal{L}$ for linear observation models which lets us specifically characterize the large deviations exponent upper bounds for probability of miss and probability of false alarm. There are N agents deployed in the network. Every agent n at time index t makes a noisy observation $\mathbf{y}_n(t)$, a noisy function of θ^* which is a M-dimensional parameter. Formally the observation model for the n-th agent is given by,

$$\mathbf{y}_n(t) = \mathbf{H}_n\theta^* + \gamma_n(t), \tag{3.15}$$

where $\{\mathbf{y}_n(t)\} \in \mathbb{R}^{M_n}$ is the observation sequence for the n-th agent and $\{\gamma_n(t)\}$ is a zero mean temporally i.i.d Gaussian noise sequence at the n-th agent with nonsingular covariance $\mathbf{\Sigma}_n$, where $\mathbf{\Sigma}_n \in \mathbb{R}^{M_n \times M_n}$. The noise processes are independent across different agents. If M is large, in practical applications each agent's observations may only correspond to a subset of the components of θ^*, with $M_n << M$, which basically renders the parameter of interest θ^* locally unobservable at each agent. Under local unobservability, in isolation, an agent cannot estimate the entire parameter. However under appropriate observability conditions, it may be possible for each agent to get a consistent estimate of θ^*. Moreover, depending on as to which hypothesis is in force, the observation model is formalized as follows:

$$\begin{aligned}\mathcal{H}_1 &: \mathbf{y}_n(t) = \mathbf{H}_n\theta^* + \gamma_n(t)\\ \mathcal{H}_0 &: \mathbf{y}_n(t) = \gamma_n(t).\end{aligned} \tag{3.16}$$

We formalize the assumptions on the inter-agent communication graph and global observability.

Assumption 3.6.1. *We require the following global observability condition. The matrix* $\mathbf{G}$

$$\mathbf{G} = \sum_{n=1}^{N} \mathbf{H}_n^\top \mathbf{\Sigma}_n^{-1} \mathbf{H}_n \tag{3.17}$$

is full rank.

Remark 3.6.1. *It is to be noted that Assumption 3.5.1 reduces to Assumption 3.6.1 for linear models, i.e., by taking* $\mathbf{h}_n(\theta^*) = \mathbf{H}_n\theta^*$.

Algorithm $\mathcal{CIGLRT}-\mathcal{L}$

The algorithm $\mathcal{CIGLRT}-\mathcal{L}$ consists of three parts, namely, parameter estimate update, decision statistic update and the decision rule.

Parameter Estimate Update. The algorithm $\mathcal{CIGLRT}-\mathcal{L}$ generates a sequence $\{\theta_n(t)\}\in\mathbb{R}^M$ which are estimates of θ^* at the n-th agent according to the following recursive scheme

$$\theta_n(t+1)=\theta_n(t)-\underbrace{\beta_t\sum_{l\in\Omega_n}(\theta_n(t)-\theta_l(t))}_{\text{neighborhood consensus}}+\underbrace{\alpha_t\nabla_\theta\log\frac{f_{\theta_n(t)}(\mathbf{y}_n(t))}{f_0(\mathbf{y}_n(t))}}_{\text{local innovation}}, \tag{3.18}$$

where Ω_n denotes the communication neighborhood of agent n, $\nabla(.)$ denotes the gradient and $\{\beta_t\}$ and $\{\alpha_t\}$ are consensus and innovation weight sequences respectively (to be specified shortly) and

$$\log\frac{f_{\theta_n(t)}(\mathbf{y}_n(t))}{f_0(\mathbf{y}_n(t))}=\theta_n(t)^\top\mathbf{H}_n^\top\mathbf{\Sigma}_n^{-1}\mathbf{y}_n(t)-\frac{\theta_n(t)^\top\mathbf{H}_n^\top\mathbf{\Sigma}_n^{-1}\mathbf{H}_n\theta_n(t)}{2}. \tag{3.19}$$

It is to be noted that the parameter estimation update of the $\mathcal{CIGLRT}-\mathcal{L}$ algorithm is a special case of the $\mathcal{CIGLRT}-\mathcal{NL}$ algorithm with $\mathbf{h}_n(\theta^*)=\mathbf{H}_n\theta^*$.

The update in (3.18) can be written in a compact manner as follows:

$$\theta(t+1)=\theta(t)-\beta_t(\mathbf{L}\otimes\mathbf{I}_M)\theta(t)+\alpha_t\mathbf{G}_H\mathbf{\Sigma}^{-1}\left(\mathbf{y}(t)-\mathbf{G}_H^\top\theta(t)\right), \tag{3.20}$$

where $\theta(t)=[\theta_1^\top(t)\ \theta_2^\top(t)\cdots\theta_N^\top(t)]^\top$, $\mathbf{G}_H=diag[\mathbf{H}_1^\top,\mathbf{H}_2^\top,\cdots,\mathbf{H}_N^\top]$, $\mathbf{y}(t)=[\mathbf{y}_1^\top(t)\ \mathbf{y}_2^\top(t)\cdots\mathbf{y}_N^\top(t)]^\top$ and $\mathbf{\Sigma}=diag\left[\mathbf{\Sigma}_1,\cdots,\mathbf{\Sigma}_N\right]$.

We make the following assumptions on the weight sequences $\{\alpha_t\}$ and $\{\beta_t\}$.

Assumption 3.6.2. *The weight sequences $\{\alpha_t\}$ and $\{\beta_t\}$ are of the form*

$$\alpha_t=\frac{a}{(t+1)}\ \beta_t=\frac{a}{(t+1)^{\delta_2}}, \tag{3.21}$$

where $a\geq 1$ and $0<\delta_2\leq 1$.

Decision Statistic Update. The algorithm $\mathcal{CIGLRT}-\mathcal{L}$ generates a decision statistic sequence $\{z_n(t)\}$ at the n-th agent according to the distributed recursive scheme

$$\hat{z}_n(kt-k+1)=\theta_n(k(t-1))^\top\mathbf{H}_n^\top\mathbf{\Sigma}_n^{-1}\left(\mathbf{s}_n(k(t-1))-\frac{\mathbf{H}_n\theta_n(k(t-1))}{2}\right), \tag{3.22}$$

where $\mathbf{s}_n(k(t-1))=\sum_{i=0}^{k(t-1)}\frac{\mathbf{y}_n(i)}{k(t-1)+1}$, i.e., the time averaged sum of local observations at agent n, and the underlying parameter estimate used in the test statistic is the estimate at time $k(t-1)$. In other words, at every time instant $k(t-1)+1$ (times which are one modulo k), where k is a pre-determined positive integer (k

to be specified shortly), an agent n, incorporates its local observations made in the past k time instants, in the above mentioned manner in (3.22). It is to be noted that, independent of the decision statistic update, $\mathbf{s}_n(k(t-1))$ is updated as and when a new observation is made at agent n. After incorporating the local observations, every agent n undergoes $k-1$ rounds of consensus, which can be expressed in a compact form as follows:

$$\hat{\mathbf{z}}(kt) = \mathbf{W}^{k-1}\mathbf{G}_\theta(k(t-1))\mathbf{\Sigma}^{-1} \times \left(\mathbf{s}(k(t-1)) - \frac{\mathbf{G}_H^\top \theta(k(t-1))}{2}\right), \tag{3.23}$$

where $\mathbf{z}(t) = [z_1(t)\cdots z_N(t)]$, $\mathbf{G}_\theta(t) = diag\left[\theta_1^\top(t)\mathbf{H}_1^\top, \theta_2^\top(t)\mathbf{H}_2^\top, \cdots, \theta_N^\top(t)\mathbf{H}_N^\top\right]$ and $\mathbf{s}(t) = \left[\mathbf{s}_1^\top(t)\ \mathbf{s}_2^\top(t)\cdots\mathbf{s}_n^\top(t)\right]^\top$, where $\mathbf{W}$ is a $N\times N$ weight matrix, where we assign $w_{ij}=0$, if $(i,j)\notin E$. The sequence $\{\hat{z}_n(t)\}$ is an auxiliary sequence and the decision statistic sequence $\{z_n(t)\}$ is generated from the auxiliary sequence in the following way:

$$z_n(kt) = \hat{z}_n(kt), \forall t, \tag{3.24}$$

where as in the interval $[k(t-1), kt-1]$, the value of the decision statistic stays constant corresponding to its value at $z_n(kt-k),\ \forall t$.

Remark 3.6.2. *The $\mathcal{CIGLRT}-\mathcal{NL}$ algorithm is for general non-linear observations with a very intuitive decision statistic update which seeks to track the time average of the log-likelihood ratios over time. The $\mathcal{CIGLRT}-\mathcal{NL}$ algorithm can be extended to general linear models but the linear model version of the $\mathcal{CIGLRT}-\mathcal{NL}$ algorithm is different from the $\mathcal{CIGLRT}-\mathcal{L}$. In particular, the results of this chapter are inconclusive as to whether the linear model version of $\mathcal{CIGLRT}-\mathcal{NL}$ algorithm is able to achieve exponential decay in terms of the probability of miss or not. This is why we introduce the $\mathcal{CIGLRT}-\mathcal{L}$ algorithm for the specific linear case for which we are able to establish exponential decays for both the probabilities of miss and false alarm. The algorithms are different. Intuitively speaking, the performance of the $\mathcal{CIGLRT}-\mathcal{NL}$ algorithm in terms of decay of the probability of miss is affected as there is no mechanism in the decision statistic update so as to get rid of the initial bad parameter estimates, which are weighed equally as the later more accurate parameter estimates. The decision statistics update for the $\mathcal{CIGLRT}-\mathcal{L}$ algorithm however ensures that at any time after kt, the previous parameter estimates i.e., $\theta(ks), s=1,\cdots,t-1$ do not contribute to the decision statistic. Due to the nature of the decision statistic update, for the $\mathcal{CIGLRT}-\mathcal{L}$ it can be shown that $t\left\|\theta(t)-\theta_N^*\right\|^2 = \frac{t\mathbf{P}_t}{2}\gamma_{G,t}\gamma_{G,t}^\top$, where*

$$\gamma_{G,t} = [\gamma_G^\top(0)\ \gamma_G^\top(1)\ \cdots\ \gamma_G^\top(t-1)]^\top,$$

where $\gamma_G(t) = \mathbf{G}_H\mathbf{\Sigma}^{-1}\gamma(t)$ and $\mathbf{P}_t$ is a block matrix of dimension $NMt\times NMt$, whose (i,j)-th block $i,j=0,\cdots,t-1$ is given as follows:

$$[\mathbf{P}_t]_{ij} = \alpha_i\alpha_j \prod_{u=0}^{t-2-i}\mathbf{A}(t-1-u)\prod_{v=j+1}^{t-1}\mathbf{A}(v),$$

where $\mathbf{A}(t) = \mathbf{I}_{NM} - \beta_t\left(\mathbf{L}\otimes\mathbf{I}_M\right) - \alpha_t\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top$. The matrix $t\left\|\mathbf{P}_t\right\|$ can be shown to be bounded. For the $\mathcal{CIGLRT}-\mathcal{NL}$ though, instead of $t\left\|\theta(t)-\theta_N^*\right\|^2$, we have to deal with $\sum_{s=1}^t\left\|\theta(s)-\theta_N^*\right\|^2$ which in turn does

not stay bounded as $t \to \infty$. To sum it up, $\mathcal{CIGLRT} - \mathcal{L}$ is indeed a different algorithm from $\mathcal{CIGLRT} - \mathcal{NL}$ with a carefully designed decision statistic update so as to be able to characterize the exponential decay of both the probabilities of error.

Now we state some design assumptions on the weight matrix $\mathbf{W}$.

Assumption 3.6.3. *The entries in the weight matrix $\mathbf{W}$ are designed in such a way that $\mathbf{W}$ is non-negative, symmetric, irreducible and stochastic, i.e., each row of $\mathbf{W}$ sums to one.*

We remark that, if Assumption 3.6.3 is satisfied, then the second largest eigenvalue in magnitude of $\mathbf{W}$, denoted by r, turns out to be strictly less than one, see for example Dimakis et al. (2010). Note that, by the stochasticity of $\mathbf{W}$, the quantity r satisfies

$$r = ||\mathbf{W} - \mathbf{J}||, \tag{3.25}$$

where $\mathbf{J} = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$.
A intuitive way to design $\mathbf{W}$ is to assign equal combination weights, in which case we have,

$$\mathbf{W} = \mathbf{I}_N - \delta\mathbf{L}, \tag{3.26}$$

where $\delta \in \left(0, \frac{2}{\lambda_N(\mathbf{L})}\right)$. For subsequent analysis, we specifically choose $\delta = \frac{2}{\lambda_2(\mathbf{L})+\lambda_N(\mathbf{L})}$.
Decision Rule. The following decision rule is adopted at all times t :

$$\mathcal{H}_n(t) = \begin{cases} \mathcal{H}_0 & z_n(t) \leq \eta \\ \mathcal{H}_1 & z_n(t) > \eta, \end{cases} \tag{3.27}$$

where $\mathcal{H}_n(t)$ is the local decision at time t at agent n.

Remark 3.6.3. *For both the proposed algorithms the agents reach asymptotic agreement or consensus in terms of the parameter estimate and the decision statistic. The decisions in the initial few time steps might be different. But, with subsequent cooperation among the agents, each agent gets to the same local decision eventually with probability one. Hence, the overall decision then is the decision of any local agent. It is to be noted that, to reach decision consensus the agents need to reach consensus on the indicator function with respect to the threshold. However, the indicator function $\mathbb{I}_{\{z_n(t)>\eta\}}$ is discontinuous at the threshold. But, from Theorems 4.2 and 4.3, we have that the decision statistics not only reach consensus but converge to $\frac{\mathbf{h}^\top(\theta_N^*)\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{2N}$ in expectation under $\mathcal{H}_1$ and 0 under $\mathcal{H}_0$ and hence the threshold is chosen in such a way that the decision statistics of different agents reach consensus to a value strictly different from the threshold so that the indicator function $\mathbb{I}_{\{z_n(t)>\eta\}}$ is continuous at the limiting consensus value. This ensures in turn that the binary decisions at the agents also reach consensus.*

Under the aegis of such a decision rule, the associated probability of errors are as follows:

$$\begin{aligned} \mathbb{P}_{M,\theta^*}(t) &= \mathbb{P}_{1,\theta^*}\left(z_n(t) \leq \eta\right) \\ \mathbb{P}_{FA}(t) &= \mathbb{P}_0\left(z_n(t) > \eta\right), \end{aligned} \tag{3.28}$$

where $\mathbb{P}_{M,\theta^*}$ and $\mathbb{P}_{FA}$ refer to probability of miss and probability of false alarm respectively.

Remark 3.6.4. *Note that, the decision statistic update requires the agents to store a copy of the running time-average of their observations. The additional memory requirement to store the running average stays constant, as the average $\mathbf{s}_n(t)$, say for agent n, can be updated recursively. It is to be noted that the decision statistic update in (3.23) has time-delayed parameter estimates and observations, i.e., delayed in the sense, in the ideal case the decision statistic update at a particular time instant, say t, would be using the parameter estimate at time t, but owing to the k rounds of consensus, the algorithm uses parameter estimates which are delayed by k time steps. Whenever, the k rounds of consensus are done with, the algorithm incorporates its latest estimates and observations into decision statistics at respective agents. After the k rounds of consensus, it is ensured that with inter-agent collaboration, the decision statistic at each agent attains more accuracy. Hence, there is an inherent trade-off between the performance (number of rounds of consensus) and the time delay. If the number of rounds of consensus is increased, the algorithm attains better detection performance asymptotically (the error probabilities have larger exponents), but at the same time the time lag in incorporating the latest sensed information into the decision statistic increases affecting possibly transient characteristics and vice-versa.*

We make an assumption on k which concerns with the number of rounds of consensus in the decision statistic update of $\mathcal{CIGLRT}-\mathcal{L}$.

Assumption 3.6.4. *Recall r as defined in (3.25). The number of rounds k of consensus between two updates of agent decision statistics satisfies*

$$k \geq 1 + \left\lfloor \frac{-3\log N}{2\log r} \right\rfloor. \tag{3.29}$$

We make an assumption on a, which is in turn defined in (3.21).

Assumption 3.6.5. Recall a as defined in Assumption 3.6.2. We assume that a satisfies

$$a \geq \frac{1}{2c_1} + 2, \tag{3.30}$$

where c_1[3] is defined as

$$\begin{aligned} c_1 &= \min_{\|\mathbf{x}\|=1} \mathbf{x}^\top \left(\mathbf{L}\otimes\mathbf{I}_M + \mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)\mathbf{x} \\ &= \lambda_{\min}\left(\mathbf{L}\otimes\mathbf{I}_M + \mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right). \end{aligned} \tag{3.31}$$

3.7 Consistency and Exponential Decay of Errors

In this section, we provide the main results concerning the algorithms $\mathcal{CIGLRT}-\mathcal{NL}$ and $\mathcal{CIGLRT}-\mathcal{L}$. The proofs are relegated to Appendix C.

Theorem 3.7.1. *Consider the $\mathcal{CIGLRT}-\mathcal{NL}$ algorithm under Assumptions 3.5.1-3.5.5 with the additional that $ac_1 \geq 1$ with c_1 as defined in Assumption 3.5.4, and the sequence $\{\theta(t)\}_{t\geq 0}$ generated according to (3.8). We then have*

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty}(t+1)^\tau \|\theta_n(t)-\theta^*\| = 0, \forall 1\leq n\leq N\right) = 1, \tag{3.32}$$

[3]We will later show that c_1 is strictly greater than zero.

for all $\tau \in [0, 1/2)$.

To be specific, the estimate sequence $\{\theta_n(t)\}_{t\geq 0}$ at agent n is strongly consistent. The next result concerns with the characterization of thresholds which ensures the probability of miss and probability of false alarm as defined in (3.14) go to zero asymptotically.

Theorem 3.7.2. *Let the hypotheses of Theorem 3.7.1 hold. Consider the decision rule defined in (3.13). For all* θ^* *which satisfy*

$$\frac{\mathbf{h}^\top\left(\theta_N^*\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N} > \frac{\left(\frac{1}{N}+\sqrt{N}r\right)\sum_{n=1}^N M_n}{2}, \tag{3.33}$$

we have the following choice of the thresholds

$$\frac{\left(\frac{1}{N}+\sqrt{N}r\right)\sum_{n=1}^N M_n}{2} < \eta < \frac{\mathbf{h}^\top\left(\theta_N^*\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N}, \tag{3.34}$$

for which we have that $\mathbb{P}_{M,\theta^*}(t) \to 0$ *and* $\mathbb{P}_{FA}(t) \to 0$ *as* $t \to \infty$. *Specifically,* $\mathbb{P}_{FA}(t)$ *decays to zero exponentially with the following large deviations exponent upper bound*

$$\limsup_{t\to\infty}\frac{1}{t}\log\left(\mathbb{P}_0\left(z_n(t) > \eta\right)\right) \leq -LE\left(\min\{\lambda^*, 1\}\right), \tag{3.35}$$

where $\theta_N^* = \mathbf{1}_N \otimes \theta^*$, $LE(.)$ *and* λ^* *are given by*

$$\begin{aligned} LE(\lambda) &= \frac{\eta\lambda}{\frac{1}{N}+\sqrt{N}} + \left(\frac{\sum_{n=1}^N M_n}{2}\right)\log\left(1 - \frac{\lambda\left(\frac{1}{N}+\sqrt{N}r\right)}{\frac{1}{N}+\sqrt{N}}\right), \\ \lambda^* &= \frac{\frac{1}{N}+\sqrt{N}}{\frac{1}{N}+\sqrt{N}r} - \frac{\left(\frac{1}{N}+\sqrt{N}\right)\sum_{n=1}^N M_n}{2\eta}. \end{aligned} \tag{3.36}$$

It is to be noted that as the observation parameters, i.e., M_n, N and the connectivity of the communication graph, i.e., r are known apriori, the threshold can be chosen to be $\frac{\left(\frac{1}{N}+\sqrt{N}r^{k-1}\right)\sum_{n=1}^N M_n}{2} + \epsilon$, where ϵ can be chosen to be arbitrarily small. The next theorem characterizes the large deviations exponent upper bound for the probability of miss and probability of false alarm related to the decision statistic sequence $\{z_n(t)\}$ generated at agent n, by the decision statistic update part of the $\mathcal{CIGLRT}-\mathcal{L}$ algorithm. We define the following quantities which will play a crucial role in stating the next theorem: let c_4 and c_4^* be given by

$$c_4 = \frac{1}{\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|\left(\sum_{v=0}^{t_1-1}\alpha_v^2\prod_{u=v+1}^{t_1-1}\left\|\mathbf{I}_{NM}-\beta_u\left(\mathbf{L}\otimes\mathbf{I}_M\right)-\alpha_u\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|\frac{(t_1+1)^{2c_1\alpha_0}}{kt_1^{2c_1\alpha_0-1}}+\frac{\alpha_0^2}{kt_1}+\frac{\alpha_0^2}{2c_1\alpha_0-1}\right)}, \tag{3.37}$$

and

$$c_4^* = \frac{2c_1\alpha_0-1}{\alpha_0^2\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|} - \frac{NM}{2\eta_2} \tag{3.38}$$

respectively, where η_2 is given by

$$\eta_2 = \frac{-2N\eta + (\theta^*)^\top \mathbf{G}\theta^* \left(1 - N\sqrt{N}r^{k-1}\right)}{4\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|\left(1 + N\sqrt{N}r^{k-1}\right)}, \tag{3.39}$$

and t_1 defined as

$$t_1 = \max\{t_2, t_3\}, \tag{3.40}$$

where t_3 is such that, $\forall t \geq t_3$,

$$\lambda_{\min}\left(\mathbf{L}\otimes\mathbf{I}_M + \mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)\alpha_t < 1, \tag{3.41}$$

and t_2 is such that[4], $\forall t \geq t_2$,

$$\beta_t\lambda_N(\mathbf{L}) + \alpha_t\lambda_{\max}\left(\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right) < 1. \tag{3.42}$$

Theorem 3.7.3. *Let the hypotheses of theorem 3.7.1 hold. Consider, the decision statistic update of the $\mathcal{CIGLRT}-\mathcal{L}$ algorithm in (3.22). For all θ^*, which satisfy the following condition*

$$\frac{(\theta^*)^\top \mathbf{G}\theta^*\left(1 - N\sqrt{N}r^{k-1}\right)}{2N} > \frac{\left(\frac{1}{N} + \sqrt{N}r^{k-1}\right)\sum_{n=1}^N M_n}{2} + \frac{M\alpha_0^2\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|^2\left(1 + N\sqrt{N}r^{k-1}\right)}{2c_1\alpha_0 - 1}, \tag{3.43}$$

we have the following range of feasible thresholds,

$$\frac{\left(\frac{1}{N} + \sqrt{N}r^{k-1}\right)\sum_{n=1}^N M_n}{2} < \eta < \frac{(\theta^*)^\top \mathbf{G}\theta^*\left(1 - N\sqrt{N}r^{k-1}\right)}{2N} - \frac{M\alpha_0^2\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|^2\left(1 + N\sqrt{N}r^{k-1}\right)}{2c_1\alpha_0 - 1}, \tag{3.44}$$

for which we have the following large deviations upper bound characterization for the probability of false alarm $\mathbb{P}_{FA}$:

$$\begin{aligned}\limsup_{t\to\infty}\frac{1}{t}\log\left(\mathbb{P}_0\left(z_n(t) > \eta\right)\right) &\leq -\frac{\eta}{\frac{1}{N} + \sqrt{N}r^{k-1}} \\ &- \frac{\sum_{n=1}^N M_n}{2}\left(1 + \log\frac{2\eta}{\left(\frac{1}{N} + \sqrt{N}r^{k-1}\right)\sum_{n=1}^N M_n}\right) \\ &= LD_0(\eta),\end{aligned} \tag{3.45}$$

and the following large deviations upper bound characterization for the probability of miss $\mathbb{P}_M$:

$$\limsup_{t\to\infty}\frac{1}{t}\log\left(\mathbb{P}_{1,\theta^*}\left(z_n(t) < \eta\right)\right)$$

[4]It is to be noted that such t_2 and t_3 exist as $\alpha_t, \beta_t \to 0$ as $t \to \infty$.

$$\leq \max \left\{ -\frac{\left(-\frac{\eta}{4}+\frac{(\theta^*)^\top \mathbf{G}\theta^*\left(\frac{1}{N}-\sqrt{N}r^{k-1}\right)}{8}\right)^2}{2\sum_{j=1}^{N}(\theta^*)^\top \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1}\mathbf{H}_j\theta^*\left(\frac{1}{N}+\sqrt{N}r^{k-1}\right)^2}, \right.$$

$$\left. -LD\left(\min\left\{c_4, c_4^*\right\}\right)\right\} = LD_1\left(\eta\right), \tag{3.46}$$

where,

$$LD(\lambda) = \lambda\eta_2 + \frac{NM}{2}\log\left(1-\frac{\lambda\alpha_0^2\left\|\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right\|}{2c_1\alpha_0-1}\right), \tag{3.47}$$

and c_4, c_4^*, η_2, t_1, t_2 and t_3 are parameters which are functions of N, r and the sensing model.

The bounds derived for the range of parameter θ^* for which exponential decay of error probabilities can be ensured, for both the distributed $\mathcal{CIGLRT}-\mathcal{L}$ detector and the centralized detector are conservative and hence might not be tight. With better network connectivity, the upper bounds of the large deviations exponent of the distributed detector approach the upper bounds of the large deviations exponents of that of the centralized detector. The range of θ^*'s for which the distributed detector ensures exponential decay of error probabilities becomes bigger with better network connectivity[5], i.e., with smaller r. For our analysis of probability of errors and their respective decay rate characterizations, we considered a uniform (i.e., agent and time independent) choice of thresholds. As far as the agent dependent thresholds are concerned, the thresholds for each agent could be functions of an individual agent's connectivity, thus allowing for more degrees of freedom in the design. Intuitively, an agent with *better connectivity* would have a wider range of thresholds to choose from so as to do a more flexible trade-off between the probability of false alarm and the probability of miss. Though an individual agent dependent bounding would be ideal, but it makes the analysis seemingly intractable. For example, while bounding $\sum_{j=1}^{N}\phi_{n,j}(k-1)\frac{(\theta^*)^\top \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1}\mathbf{H}_j\theta^*}{2}$, where $\phi_{n,j}(k-1) = \left[\mathbf{W}^{k-1}\right]_{n,j}$, we use the same bound for each of $\phi_{n,j}(k-1)$ which is in turn given by $\phi_{n,j}(k-1) \geq \frac{1}{N}-\sqrt{N}r^{k-1}$. In such a setting, tracking the row corresponding to a particular agent in the weight matrix is seemingly intractable. Having said that, if individual agent dependent thresholds could have been used, the extent to which an agent can distinguish between different hypotheses would be different owing to different range of thresholds available to an agent to choose from.

The use of time dependent thresholds however does not seem to affect the zone of indifference. It is to be noted that the zone of the indifference is characterized in terms of the range of θ^*'s under the alternate hypothesis for which exponential decay of both the probability of errors can be ensured. To be specific, in the analysis of the probability of the miss the crucial part of the analysis is the bound for the term $t\left\|\mathbf{P}_t\right\|$ which is given by

$$t\left\|\mathbf{P}_t\right\| \leq c_3\frac{(t_1+1)^{2c_1\alpha_0}}{t^{2c_1\alpha_0-1}} + \frac{\alpha_0^2}{t} + \frac{\alpha_0^2}{2c_1\alpha_0-1}, \quad \forall t \geq t_1.$$

If the thresholds are adjusted, so as to take into account the time-decaying terms $c_3\frac{(t_1+1)^{2c_1\alpha_0}}{t^{2c_1\alpha_0-1}}$ and $\frac{\alpha_0^2}{t}$, both the terms decay to zero as $t\to\infty$ thereby ensuring that the large deviations exponent upper bound stays the same. However, it is to be noted that at any finite time, time dependent thresholds would give tighter

[5]Intuitively, r indicates how well a network is connected. For e.g. if a network is fully connected, i.e., has an all-to-all connected communication graph and hence $\mathbf{W}=\mathbf{J}$, $r=0$. In the absence of communication, $\mathbf{W}=\mathbf{I}$ and $r=1$. Hence, a lower value of r indicates better connectivity of the graph.

probability of error bounds, thus improving transients of the approach.

The assumption regarding the inter-agent communication network which is instrumental in obtaining the results of this chapter is the connectivity of the graph i.e., there exists a path between every two nodes. As long as the graph is connected, the results continue to hold. For example, for a network with N agents, there needs to be at least $N-1$ links for the graph to be connected, which is very sparsely connected as out of the possible $N(N-1)/2$ possible links only $N-1$ links are present. Hence, the algorithms apply to very sparse (but connected) networks. However, it is important to note that the large deviations upper bounds might get worse with increasing sparsity of the inter-agent communication graph. The sparsity of the inter-agent communication graph is in turn reflected by the quantity[6] $r = \|\mathbf{W} - \mathbf{J}\|$, where $\mathbf{W}$ and $\mathbf{J}$ represent the weight matrices associated with the inter-agent communication graph under consideration and a completely connected graph respectively. When $r = 0$ it basically points to the case that $\mathbf{W} = \mathbf{J}$, i.e., the network under consideration is completely connected. The other extreme case $r = 1$ is the case when $\mathbf{W} = \mathbf{I}$, i.e., when there are no links in the network. To sum it up, small values of $1-r$ reflect the sparseness of the network under consideration.

Furthermore, note that with increasing k, i.e., the time lag or equivalently the number of rounds of consensus between incorporating latest estimates (see (3.23)), the range of parameter θ^* for which exponential decay of error probabilities can be ensured increases, the large deviations upper bounds for the probabilities of miss and false alarm also increase. However, k cannot be made arbitrarily large just based on improvement of the large deviations upper bounds, as large deviations analysis is essentially an asymptotic characterization and at the same time with increase in k the inherent time delay in incorporating new estimates into the decision statistic also increases, and hence affecting the transient performance of the procedure. Recall from the decision statistic update in (3.23), that the decision statistic update takes the value $z_n(kt-k)$ at all times $t \in [kt-k, kt-1]$. Thus, only at time instants which are of the form kt, the decision statistic has the minimum time-lag k with respect to the latest information available in the multi-agent network which also makes the analysis more tractable. Moreover, from the perspective of a faulty agent, low k would result in particularly bad detection performance as the dynamics of an accurate detection procedure at a faulty agent depends on the information it receives from its neighbors, which shows the necessity of inter-agent collaboration. In absence of a distributed mechanism characterized by a communication graph, a defective agent would fail to come up with a reasonable decision at all times, as the local sensed data at a defective agent is completely non-informative. Finally, no inference procedure is free of the *curse of dimensionality*. It is to be noted that with increasing M, i.e., dimension of the underlying parameter θ^*, the range of θ^* for which exponential decay of probabilities of errors can be ensured shrinks, the feasible range of thresholds also shrinks and finally the large deviations exponent upper bound for the probability of miss also decreases.

3.8 $\mathcal{CIGLRT}$: Imperfect Communication

In this section, in addition to the setup for $\mathcal{CIGLRT} - \mathcal{NL}$, we consider noisy communication. In particular, we assume that the inter-agent communication is imperfect, i.e., noisy. To be specific, we assume that an agent pair (i, j) exchange information over a vector additive zero-mean noise channel. Formally speaking, if agent i transmits a data vector $\mathbf{z} \in \mathbb{R}^k$ to agent j, the information received at agent j is given by

$$\tilde{\mathbf{z}} = \mathbf{z} + \psi_{i,j}, \tag{3.48}$$

[6] It is to be noted that as long as the graph is connected, $r < 1$.

where the noise vector $\psi_{i,j}$ is Gaussian with zero mean and has finite variance $\mathbf{\Sigma}_{ij}$. Furthermore, we assume the transmission channel noises are independent over transmissions and across the graph links.

Parameter Estimate Update. The algorithm generates the sequence $\{\theta_n(t)\} \in \mathbb{R}^M$ at the n-th agent according to the following recursive scheme

$$\begin{aligned}\theta_n(t+1) &= \theta_n(t) \underbrace{- b\alpha_t \sum_{l\in\Omega_n} (\theta_n(t) - \theta_l(t) - \psi_{n,l}(t))}_{\text{neighborhood consensus}} \\ &+ \underbrace{\alpha_t \nabla \mathbf{h}_n(\theta_n(t)) \mathbf{\Sigma}_n^{-1} (\mathbf{y}_n(t) - \mathbf{h}_n(\theta_n(t)))}_{\text{local innovation}} \\ &= \theta_n(t) - b\alpha_t \sum_{l\in\Omega_n} (\theta_n(t) - \theta_l(t)) \\ &+ \alpha_t \nabla \mathbf{h}_n(\theta_n(t)) \mathbf{\Sigma}_n^{-1} (\mathbf{y}_n(t) - \mathbf{h}_n(\theta_n(t))) + b\alpha_t \sum_{l\in\Omega_n} \psi_{n,l}(t), \end{aligned} \tag{3.49}$$

where Ω_n denotes the communication neighborhood of agent n, b is a positive constant, $\psi_{n,l}(t)$ is the communication noise in the link between n and l, $\nabla \mathbf{h}_n(.)$ denotes the gradient of $\mathbf{h}_n$, which is a matrix of dimension $\mathbf{M} \times \mathbf{M}_n$, with the (i,j)-th entry given by $\frac{\partial [\mathbf{h}_n(\theta_n(t))]_j}{\partial [\theta_n(t)]_i}$ and $\{\alpha_t\}$ is the innovation weight sequence (to be specified shortly). Note that, in (3.49), each agent $l \in \Omega_n$ intends to send it's exact estimate to agent n, but agent n receives a noisy version of estimates from agents in its neighborhood as the inter-agent communication is over noisy links. The update in (3.18) can be written in a compact manner as follows:

$$\begin{aligned}\theta(t+1) &= \theta(t) - b\alpha_t (\mathbf{L} \otimes \mathbf{I}_M) \theta(t) \\ &+ \alpha_t \mathbf{G}(\theta(t)) \mathbf{\Sigma}^{-1} (\mathbf{y}(t) - \mathbf{h}(\theta(t))) + b\alpha_t \mathbf{\Psi}(t), \end{aligned} \tag{3.50}$$

where $\theta(t)^\top = [\theta_1(t)^\top \cdots \theta_N(t)^\top]$, $\mathbf{h}(\theta(t)) = \left[\mathbf{h}_1^\top(\theta_1(t)) \cdots \mathbf{h}_N^\top(\theta_N(t))\right]^\top$, $\mathbf{y}(t)^\top = [y_1(t)^\top \cdots y_N(t)^\top]^\top$, $\mathbf{G}(\theta(t)) = diag[\nabla \mathbf{h}_1(\theta_1(t)), \cdots, \nabla \mathbf{h}_N(\theta_N(t))]$, $\mathbf{\Sigma} = diag[\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_N]$ and $\mathbf{\Psi}^\top(t) = \left[\left(\sum_{l\in\Omega_1} \psi_{1,l}(t)\right)^\top \cdots \left(\sum_{l\in\Omega_N} \psi_{N,l}(t)\right)^\top\right]^\top$.

We make the following assumption on the weight sequence $\{\alpha_t\}$.

Assumption 3.8.1. *The weight sequence $\{\alpha_t\}$ is of the form $\alpha_t = (t+1)^{-1}$ and the positive constant b is such that $b < \frac{1}{\lambda_N(\mathbf{L})}$.*

3.9 Main Results: $\mathcal{CIGLRT}$ Imperfect Communication

We define the following quantities which will be crucial for stating the next theorem : let $\mathbf{\Sigma}_{c,1}^*$ and $\mathbf{\Sigma}_{c,0}^*$ be given by

$$\begin{aligned}\mathbf{\Sigma}_{c,1}^* &= \mathbf{V}_L \mathbf{M}_1 \mathbf{V}_L^\top \\ \mathbf{\Sigma}_{c,0}^* &= \mathbf{V}_L \mathbf{M}_0 \mathbf{V}_L^\top \end{aligned} \tag{3.51}$$

respectively, where $\mathbf{M}_1$ and $\mathbf{M}_0$ are given by

$$[\mathbf{M}_1]_{ij} = \left[\mathbf{V}_L^\top \mathbf{\Sigma}_1^* \mathbf{V}_L\right]_{ij} (b[\mathbf{D}_L]_{ii} + b[\mathbf{D}_L]_{jj} + 1)^{-1}$$

$$[\mathbf{M}_0]_{ij} = \left[\mathbf{V}_L^\top \boldsymbol{\Sigma}_0^* \mathbf{V}_L\right]_{ij} \left(b\,[\mathbf{D}_L]_{ii} + b\,[\mathbf{D}_L]_{jj} + 1\right)^{-1}, \tag{3.52}$$

respectively, and $\boldsymbol{\Sigma}_1^*$ and $\boldsymbol{\Sigma}_0^*$ are given by

$$\begin{aligned}\boldsymbol{\Sigma}_1^* &= \mathbf{h}^*(\mathbf{1}_N \otimes \theta^*)\boldsymbol{\Sigma}^{-1}(\mathbf{h}^*(\mathbf{1}_N \otimes \theta^*))^\top + b^2\boldsymbol{\Sigma}_c \\ \boldsymbol{\Sigma}_0^* &= b^2\boldsymbol{\Sigma}_c,\end{aligned} \tag{3.53}$$

respectively, whereas $\mathbf{V}_L$ and $\mathbf{D}_L$ represent the matrix of eigenvectors and eigenvalues of $\mathbf{L}$ respectively, i.e.,

$$\mathbf{L} = \mathbf{V}_L^\top \mathbf{D}_L \mathbf{V}_L, \tag{3.54}$$

and $\boldsymbol{\Sigma}_c$ denotes the covariance matrix of the channel noise encountered in the test statistic exchange among agents given by the process $\{\zeta(t)\}$.

Theorem 3.9.1. *Consider the algorithm under Assumptions 3.5.1-3.5.5, and the sequence $\{\mathbf{z}(t)\}$. We then have under $\mathbb{P}_{\theta^*}$*

$$\begin{aligned}&\sqrt{t+1}\left(\mathbf{z}(t) - (b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N \otimes \theta^*)\boldsymbol{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N \otimes \theta^*)}{2}\right) \\ &\overset{\mathcal{D}}{\Longrightarrow} \mathcal{N}\left(0, \boldsymbol{\Sigma}_{c,1}^*\right)\end{aligned} \tag{3.55}$$

$\forall n$, and under $\mathbb{P}_0$

$$\sqrt{t+1}\left(\mathbf{z}(t)\right) \overset{\mathcal{D}}{\Longrightarrow} \mathcal{N}\left(0, \boldsymbol{\Sigma}_{c,0}^*\right), \ \forall n, \tag{3.56}$$

where $\overset{\mathcal{D}}{\Longrightarrow}$ denotes convergence in distribution (weak convergence).

Theorem 3.9.1 asserts the asymptotic normality of the test statistic $\{z_n(t)\}$, $\forall n$. It is to be noted that the asymptotic mean of $\mathbf{z}(t)$ which is given by $(b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N\otimes\theta^*)\boldsymbol{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N\otimes\theta^*)}{2}$ has all of its entries positive, as $(b\mathbf{L}+\mathbf{I})$ is a M-matrix (see, Poole and Boullion (1974)) and hence its inverse has all of its entries non-negative, i.e., $\left[(b\mathbf{L}+\mathbf{I})^{-1}\right]_{ij} \geq 0$, $\forall i,j = 1,\cdots,N$. The next result concerns with the characterization of thresholds which ensures the probability of miss and probability of false alarm as defined in (3.28) decay to zero asymptotically.

Theorem 3.9.2. *Let the hypotheses of Theorem 3.9.1 hold. Consider the decision rule defined in (3.13). For agent n, all θ^* which satisfy*

$$\left[(b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N \otimes \theta^*)\boldsymbol{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N \otimes \theta^*)}{2}\right]_n > \frac{2\sum_{n=1}^N M_n}{N}, \tag{3.57}$$

we have the following choice of the thresholds

$$\frac{2\sum_{n=1}^N M_n}{N} < \eta_n < \left[(b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N \otimes \theta^*)\boldsymbol{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N \otimes \theta^*)}{2}\right]_n, \tag{3.58}$$

for which we have that $\mathbb{P}_{M,\theta^*}(t) \to 0$ and $\mathbb{P}_{FA}(t) \to 0$ as $t \to \infty$. Specifically, $\mathbb{P}_{FA}(t)$ decays to zero

exponentially with the following large deviations exponent[7]

$$\limsup_{t\to\infty} \frac{1}{t} \log\left(\mathbb{P}_0\left(z_n(t) > \eta_n\right)\right) \leq \max\left\{-\frac{\eta_n^2}{8b^2\left\|\mathbf{\Sigma}_c\right\|}, -LE(\lambda^*)\right\}, \tag{3.59}$$

where $LE(.)$ *and* λ^* *are given by*

$$LE(\lambda) = \frac{N\eta_n\lambda}{4} + \left(\frac{\sum_{n=1}^N M_n}{2}\right)\log\left(1-\lambda\right), \quad \lambda^* = \frac{2\sum_{n=1}^N M_n}{N\eta_n} \tag{3.60}$$

respectively.

It is to be noted that the thresholds across agents can be chosen to be different owing to the unequal asymptotic mean at different agents and hence the large deviations upper bound across different agents may be different. We discuss how the above result can be used in practice to identify thresholds that lead to asymptotic decay of the probabilities of error. It is to be noted that as the observation parameters, i.e., M_n, N are known apriori, the threshold can be chosen to be $\frac{2\sum_{n=1}^N M_n}{N} + \epsilon$, where ϵ can be chosen to be arbitrarily small. Further, from the feasible range of thresholds in (3.34), a range on the θ^*s' can be obtained in terms of $\|\mathbf{h}\left(\mathbf{1}_N \otimes \theta^*\right)\|$ such that under $\mathcal{H}_1$, as long as the true value θ^* of the parameter belongs to this range, the probability of miss is guaranteed to decay to zero asymptotically. It is important to note in this context that there exists some weak signals, i.e., signals with low $\|\mathbf{h}\left(\mathbf{1}_N \otimes \theta^*\right)\|$ (but non-zero), for which there may not exist a choice of thresholds to ensure asymptotically decaying probability of miss. The signals for which Theorem 3.9.1 is rendered to be inconclusive in the manner described above, can be categorized in terms of θ^*.

3.10 Simulations

3.10.1 $\mathcal{CIGLRT} - \mathcal{NL}$

We generate a random geometric network of 10 agents. The location of agents are generated by sampling the x coordinates and the y coordinates from a uniform distribution on the interval $[0, 1]$. We link two vertices by an edge if the distance between them is less than or equal to $g = 0.4$. We go on iterating this procedure until we get a connected graph. The network connectivity expressed in terms of $r = \|\mathbf{W} - \mathbf{J}\|$ is given by $r = 0.3904$. We consider the underlying parameter to be a 5-dimensional parameter, i.e., $M = 5$ and $\theta^* = [\pi/6 \ -\pi/4 \ \pi/4 \ -\pi/5 \ \pi/6]$. For the nonlinear observation model, we consider trigonometric sensing functions which are given by, $\mathbf{f}_1(\boldsymbol{\theta}) = 5\sin(\theta_1+\theta_2), \mathbf{f}_2(\boldsymbol{\theta}) = 5\sin(\theta_3+\theta_2), \mathbf{f}_3(\boldsymbol{\theta}) = 5\sin(\theta_3+\theta_4), \mathbf{f}_4(\boldsymbol{\theta}) = 5\sin(\theta_4+\theta_5), \mathbf{f}_5(\boldsymbol{\theta}) = 5\sin(\theta_1+\theta_5), \mathbf{f}_6(\boldsymbol{\theta}) = 5\sin(\theta_1+\theta_3), \mathbf{f}_7(\boldsymbol{\theta}) = 5\sin(\theta_4+\theta_2), \mathbf{f}_8(\boldsymbol{\theta}) = 5\sin(\theta_3+\theta_5), \mathbf{f}_9(\boldsymbol{\theta}) = 5\sin(\theta_1+\theta_4)$ and $\mathbf{f}_{10}(\boldsymbol{\theta}) = 5\sin(\theta_1+\theta_5)$, where the underlying parameter is 5 dimensional, $\boldsymbol{\theta} = [\theta_1,\ \theta_2,\ \theta_3,\ \theta_4,\ \theta_5]$. However, we restrict the values of $\boldsymbol{\theta}$ to the set $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^5 \in \mathbb{R}^5$. Note that, in spite of restricting the parameter to a set, the alternate hypothesis is still parameterized by vector parameters from a continuous set. The local sensing models are unobservable, but collectively they are globally observable since, $\sin(\cdot)$ is one-to-one on

[7] By large deviations exponent, we mean the large deviations upper bound.

the set $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^5$ and the set of linear combinations of the $\boldsymbol{\theta}$ components corresponding to the arguments of the $\sin(\cdot)$'s constitute a full-rank system for $\boldsymbol{\theta}$. The agents make noisy scalar observations where the observation noise process is Gaussian and the noise covariance is given by $\mathbf{R} = 2\mathbf{I}_{10}$. It is readily seen that the sensing model with the restriction that the parameter can take values from the set $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^5$ satisfy Assumptions 3.5.1-3.5.4. We carry out 2000 Monte-Carlo simulations for analyzing the convergence of the parameter estimate sequences. The estimates are initialized to be 0, i.e., $\boldsymbol{\theta}_n(0) = \mathbf{0}$ for $n = 1, \cdots, 10$. The normalized error for the n-th agent at time t is given by the quantity $\|\boldsymbol{\theta}_n(t) - \boldsymbol{\theta}^*\|$. Figure 3.1 shows the estimation error at every agent against the time index t. For the analysis of the probability of miss, we

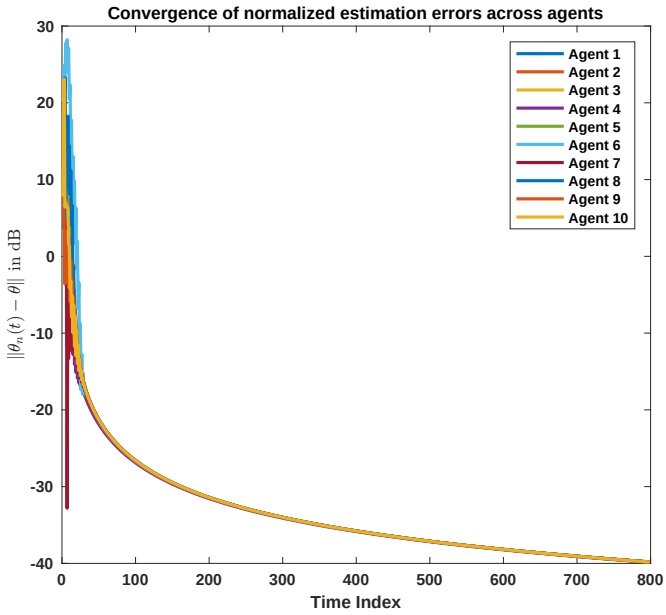


Figure 3.1: $\mathcal{CIGLRT} - \mathcal{NL}$: Convergence of estimation error at each agent

run the algorithm for 2000 sample paths so as to empirically estimate the probability of miss. Figure 3.2 verifies the assertion in Theorem 3.7.2. The threshold is chosen to be equal to 7. It is to be noted that, from Figure 3.2 the probability of miss starts decaying even before the parameter estimates get reasonably close to the true underlying parameter, which further indicates the *online* nature of the proposed algorithm $\mathcal{CIGLRT} - \mathcal{NL}$. The decay of the probability of the miss can be attributed to the fact that, in order to reach the correct decision, the decision statistics of the agents need to cross the threshold which is achieved even before the agents' parameter estimates reach close to the true underlying parameter.

3.10.2 $\mathcal{CIGLRT} - \mathcal{L}$

We generate a planar ring network of 10 agents, where every agent has exactly two neighbors. We consider the underlying parameter to be a 5-dimensional parameter, i.e., $M = 5$ and $\theta^* = \theta^* = [1\ 0.9\ 1.2\ 1.1\ 1.5]$. The observation matrices for the agents are of the dimension 5×1, i.e., $M_n = 1$, $\forall n$. Specifically the $\mathbf{H}_n$'s are given by $H_1 = [1\ 1\ 0\ 0\ 0], H_2 = [0\ 1\ 1\ 0\ 0], H_3 = [0\ 0\ 1\ 1\ 0], H_4 = [0\ 0\ 0\ 1\ 1], H_5 = [1\ 0\ 0\ 0\ 1], H_6 = [1\ 0\ 1\ 0\ 0], H_7 = [0\ 1\ 0\ 1\ 0], H_8 = [0\ 0\ 1\ 0\ 1], H_9 = [1\ 0\ 0\ 1\ 0], H_{10} = [0\ 1\ 0\ 0\ 1]$. The noise covariance matrix $\boldsymbol{\Sigma}$ is taken to be $3\mathbf{I}_{10}$. We emphasize that the above design ensures global observability (in the sense of Assumption 3.6.1), as the matrix G is invertible, but at the same time the parameter of interest is locally

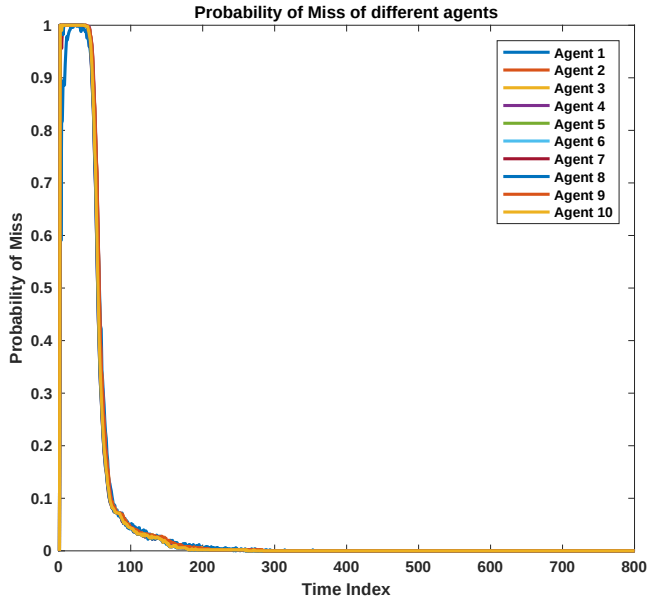


Figure 3.2: $\mathcal{CIGLRT} - \mathcal{NL}$: Probability of miss of the agents

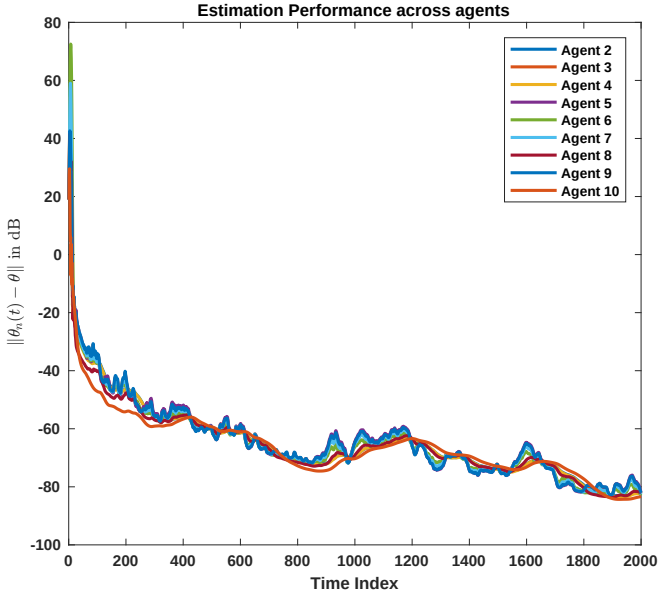


Figure 3.3: $\mathcal{CIGLRT} - \mathcal{L}$: Convergence of estimation error at each agent

unobservable at all agents. The network is poorly connected which in turn is reflected by the quantity $r = \|\mathbf{W} - \mathbf{J}\|$ and is given by 0.8404. In particular, for the parameter estimation algorithm, $a = 9.1$ and $\delta_2 = 0.4$, where a, δ_2 are as defined in Assumption 3.6.2. The time-lag k is taken to be $k = 20$. Figure 3.3 shows the convergence of the parameter estimates of the agents to the underlying parameter in different dimensions which in turn demonstrates the consistency of the parameter estimation scheme.

For the analysis of the probability of miss, we run the algorithm for 2000 sample paths. The threshold is chosen as $\eta = \frac{\left(\frac{1}{N} + \sqrt{N} r^{k-1}\right) \sum_{n=1}^{N} M_n}{2} + 0.01 = 0.8280$. The evolution of the test statistic can be closely seen in

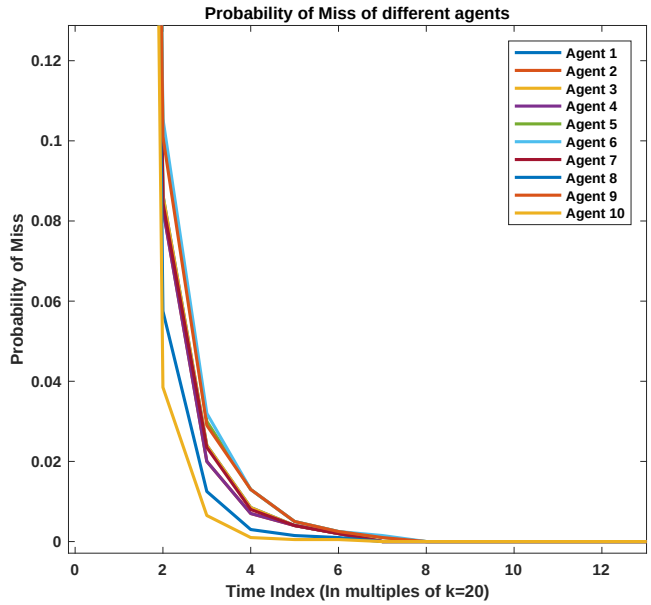


Figure 3.4: $\mathcal{CIGLRT} - \mathcal{L}$: Probability of Miss at each agent

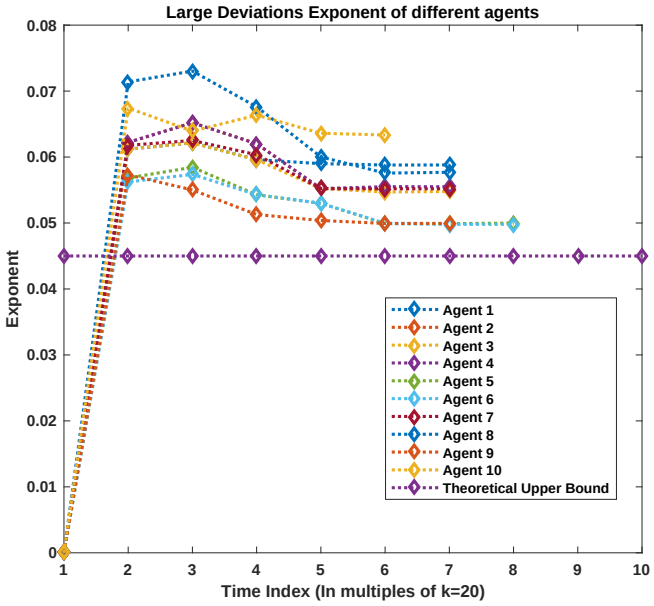


Figure 3.5: $\mathcal{CIGLRT} - \mathcal{L}$: Large Deviations Exponent Upper bounds

Figure 3.4 with the probability of miss decaying exponentially and thus verifying the assertion in Theorem 3.7.3. It is to be noted that, from Figure 3.4 the probability of miss starts decaying even before the parameter estimates get reasonably close to the true underlying parameter, which further indicates the *online* nature of the proposed algorithm $\mathcal{CIGLRT}-\mathcal{L}$. The large deviations exponent across different agents are plotted in Figure 3.5. The theoretical upper bound in (3.46) is found to be 0.045. The simulation results verify that the empirically estimated large deviations exponent upper bound for different agents is upper bounded by the theoretically derived upper bound[8]

3.11 Summary of Contributions

- **Distributed Composite Hypothesis Testing Algorithms:** We propose two distributed recursive composite hypothesis testing algorithms, where the composite alternative concerning the state of the field is modeled as a vector of (continuous) unknown parameters determining the parametric family of probability measures induced on the agents' observation spaces under the hypotheses. Due to the local unobservability, agents can not conduct hypothesis testing just based on their own samples. We explicitly characterized a distributed decision making scheme, where agents collaborate in their neighborhood over a possibly sparse communication graph.

- **Recursive Detection Algorithm with decaying probabilities of errors:** We show that in spite of being a recursive algorithm (hence suboptimal[9]), the proposed algorithm $\mathcal{CIGLRT}-\mathcal{NL}$, which is based on general non-linear observation models, guarantees asymptotically decaying probabilities of false alarm and miss under minimal conditions of global observability and connectivity of the inter-agent communication graph. We also characterize the feasible choice of thresholds and other algorithm design parameters for which such an asymptotic decay of probabilities of errors in the large sample (time) limit can be guaranteed. Through algorithm $\mathcal{CIGLRT}-\mathcal{L}$, we focus on a linear observation setup, where we not only characterize thresholds and other algorithm parameters which ensure exponentially decaying probabilities of error, but also analyze the upper bounds of the associated large deviations exponent of the probabilities of error under global observability as functions of the network and model parameters.

- **Extension of Wilks' Theorem:** In the context of centralized detection literature, the weak convergence of the decision statistics under the null and the alternate hypothesis (see, Wilks (1938) for example) is usually not enough to establish the decay rates of the probability of errors. To be specific, in this chapter we extend Wilks' theorem to the distributed recursive setup and characterize the asymptotic normality of the decision statistic sequence. However, the statistical dependencies exhibited in the decision static update due to the parameter estimation scheme and the decision statistic update running in a parallel fashion warrants the development of technical machinery so as to address concentration of measure for sums of non i.i.d random variables which in turn helps characterize the decay exponent of the probability of errors.

[8] It is an upper bound if the quantity of interest is $\frac{1}{t}\log\left(\mathbb{P}(.)\right)$. It is a lower bound if the quantity of interest is the positive exponent, i.e., $-\frac{1}{t}\log\left(\mathbb{P}(.)\right)$.

[9] The sub-optimality with respect to GLRT is due to inaccurate parameter estimates being incorporated into the decision statistic in the proposed algorithm in contrast to the *optimal* parameter estimate incorporated into the decision statistic in case of the classical GLRT.

3.12 Conclusion and Future Directions

In this chapter, we have considered the problem of a recursive composite hypothesis testing in a network of sparsely interconnected agents where the objective is to test a simple null hypothesis against a composite alternative concerning the state of the field, modeled as a vector of (continuous) unknown parameters determining the parametric family of probability measures induced on the agents' observation spaces under the hypotheses. We have proposed two *consensus+innovations* type recursive *online* algorithms, $\mathcal{CIGLRT}-\mathcal{NL}$ and $\mathcal{CIGLRT}-\mathcal{L}$, in which every agent updates its parameter estimate and decision statistic by simultaneous processing of neighborhood information and local newly sensed information and in which the inter-agent collaboration is restricted to a possibly sparse but connected communication graph. We have established the consistency of the parameter estimate sequences and characterized the large deviations exponent upper bounds of the probabilities of errors pertaining to the detection scheme for the algorithms. A natural direction for future research consists of considering models with non-Gaussian noise. We also intend to develop extensions of the $\mathcal{CIGLRT}-\mathcal{NL}$ in which the parameter domain is restricted to constrained domains such as convex subsets of the Euclidean space or manifolds.

Chapter 4

Communication Efficient Distributed Detection

4.1 Introduction

In this chapter, we study convergence in probability of products of random, independent, but not identically distributed stochastic and symmetric matrices $\mathbf{W}_t$, where the topologies that underline the matrices have time-varying distributions. Specifically, we consider the model in which there exists a baseline graph describing all feasible communication links; nodes randomly activate over time, independently one from another, such that each node is active with a certain time-dependent probability, and two nodes communicate only if they are active at the same time. A major motivation for studying the products of stochastic matrices that underlie the described randomized time varying communication protocol is the recent work Sahu et al. (2018e); therein, it is shown that incorporating the described protocol into consensus+innovations distributed estimation significantly improves the estimator's communication efficiency.

In this chapter, our goal is to characterize for the described model of the $\mathbf{W}_t$'s the speed at which the probabilities that the product of the $\mathbf{W}_t$'s stays bounded away from its limiting matrix. More precisely, we are interested in computing

$$\mathcal{R} = -\lim_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(\|\mathbf{W}_t \cdot \ldots \cdot \mathbf{W}_1 - \mathbf{J}\| \geq \epsilon\right), \tag{4.1}$$

for an arbitrary $\epsilon \in (0, 1]$, provided that the limit in (4.1) exists[1]. Here, N is the number of network nodes, and the limiting matrix $\mathbf{J} = \mathbf{1}\mathbf{1}^\top/N$.While prior work derives rate (4.1) for products of independent, identically distributed (i.i.d.) matrices, the non-i.i.d. case of independent matrices with time-varying distributions has not been studied before. Quantity (4.1) is an important metric that has many potential applications in consensus+innovations-based distributed inference. For example, reference Bajovic et al. (2011) (see also Bajović et al. (2013)) studies error exponents for Bayes error probability of consensus+innovations distributed detection under the i.i.d. matrices model. The reference shows that rate $\mathcal{R}$ critically determines detection error exponent.

It is well-known that, if the random (doubly stochastic) matrices $\mathbf{W}_t$ are i.i.d., then the product $\mathbf{W}_t \cdot \ldots \cdot \mathbf{W}_1$ converges almost surely to the consensus matrix $\mathbf{J}$ Bajović et al. (2013), and, since almost sure convergence implies convergence in probability, we thus have that the probabilities in (4.1) decay to zero. In our previous

[1] As we show later, the limit in (4.1) exists and is independent of ϵ.

work Bajović et al. (2013), we show that this convergence is *exponentially fast*, and we computed the exact rate $\mathcal{R}$ for the i.i.d. matrices. In this chapter, we consider the time varying randomized activation model described above for which the weight matrices are no longer identically distributed. We show that the limit in (4.1) continues to exist, and moreover we compute exactly the limit $\mathcal{R}$. Specifically, we show that $\mathcal{R}$ is given by the minimal vertex cut of the baseline graph, where the nodes' associated cut costs are defined by the nodes' limiting activation probabilities.

We demonstrate the significance of the studied non-i.i.d. matrix model and the derived rate $\mathcal{R}$ in the context of *consensus+innovations* distributed detection. More precisely, we consider a distributed detector with a randomized and time-varying sparsified communication protocol, where neighborhood communications are probabilistically sparsified in a time varying fashion with the goal of reducing the detector's communication cost. By utilizing result (4.1), we first show theoretically that the detector with time-varying and sparsified protocol can be designed to achieve asymptotic optimality at all signal-to-noise ratio (SNR) regimes; this is achieved when the activation probabilities converge to unity, possibly at a very slow rate, e.g., as $1 - \Omega(1/\log(t))$. In contrast, the previously studied i.i.d.-based detectors achieve asymptotic optimality only if the SNR exceeds a threshold. Therefore, effectively, we construct a "universally" asymptotically optimal detector that makes significant communication savings. Intuitively, by slowly increasing the node activation probabilities, we achieve the "optimality range" of the static protocol, but the rate of increase of probabilities is crafted carefully to achieve a reduced communication cost. Interestingly, while in Sahu et al. (2018e) it is possible to decrease the activation probabilities over time and achieve an order optimal $O(1/t)$ estimation mean square error (MSE) decay, here the probabilities need to increase in time (though possibly very slowly). The intuition for this difference is that the "baseline process" here – the decay of error probability – is much faster (it behaves like $e^{-c\,t}$, $c > 0$ a constant) than the "baseline process" in Sahu et al. (2018e) – the rate of MSE decay, which behaves like $1/t$.

Despite being very challenging to prove, rate $\mathcal{R}$ has a clear intuitive interpretation. The time instants when the matrix product gets a step closer to the limiting matrix J are the time instants when the union graph of the topologies gets connected. It is then easy to see that the event (4.1) is feasible only if the number of these improvement times is sufficiently small. We show that the event in (4.1) is therefore equivalent to the event that the number of improvements is sublinear in t. The latter effectively corresponds to the scenario that the activated nodes fail to form a connected graph. The most likely way in which this can happen is given by the vertex cut with participating nodes which are the "easiest" to disconnect, i.e., with smallest limiting occurrence probabilities.

4.2 Related Work

Products of stochastic matrices have been studied for a long time, e.g., DeGroot (1974), and the problem receives a continued interest, e.g., Olfati-Saber and Murray (2004); Tahbaz-Salehi and Jadbabaie (2010); Touri and Nedic (2014). The problem of computing the exact large deviations rate in (4.1), arising, e.g., in the analysis of distributed detection Bajovic et al. (2011), see also Braca et al. (2010); Cattivelli and Sayed (2011a); Nedić et al. (2014); Shahrampour et al. (2014); Nedić et al. (2016), has been studied before in Bajović et al. (2013) for the i.i.d. matrices and in Bajović et al. (2012) for temporally dependent matrices, where the temporal dependence is modeled through a Markov chain. This chapter complements the prior work by establishing the limit (4.1) for a class of time varying matrix distributions that have not been studied before. As explained above, the newly studied class has a significant relevance for distributed detection.

4.3 Model and preliminaries

Random nodes' activation and random matrix model. The network of nodes is modeled as an undirected graph $\overline{G} = (\overline{V}, \overline{E})$, where $\overline{V} = \{1, 2, ..., N\}$ is the set of nodes, and $\overline{E} \subseteq \binom{\overline{V}}{2}$ is the set of all communication links between nodes. We assume that $\overline{G}$ is connected. During network operation, network nodes activate at random with certain probabilities that we assume are different for different nodes. To each node $i \in \overline{V}$, we associate, for each time $t = 1, 2, ...$, a Bernoulli random variable $\xi_{i,t}$, which is equal to 1 if i is active at time t, and otherwise is 0. Let $p_{i,t} = \mathbb{P}(\xi_{i,t} = 1) \in (0, 1)$ denote the probability that i is active at time t and let V_t collect all the nodes in $\overline{V}$ that are active at time t. For arbitrary two nodes $i, j \in \overline{V}$ to communicate at time t, it is necessary and sufficient that i and j are both active at that time.
Let G_t denote the graph obtained by collecting all the nodes that are active at time t, together with their induced communication links. More precisely, $G_t = (V_t, E_t)$, where the set of active edges at time t is given by

$$E_t = \left\{\{i, j\} \in \overline{E} : i, j \in V_t\right\}. \tag{4.2}$$

We make the following assumption on the nodes' activations and on the weight matrices $\mathbf{W}_t$.

Assumption 4.3.1.

1. For each $i \in \overline{V}$, $\xi_{i,t}$ and $\xi_{i,s}$ are independent for $t \neq s$, for any $t, s \geq 1$.

2. For each node i, $p_{i,t}$ increases monotonically with t according to the following formula:

$$p_{i,t} = p_i(1 - \alpha_t), \tag{4.3}$$

where $\alpha_t \in (0, 1]$ is a monotonically decreasing sequence converging to 0, equal for all nodes, and $p_i \in (0, 1]$ is the limiting activation probability of node i.

It is easy to see from Assumption 1 that the topologies G_t, $t \geq 1$, are independent. We further make the following assumptions on the weight matrices $\mathbf{W}_t$.

Assumption 4.3.2. 1. The weight matrices $\mathbf{W}_t$, $t \geq 1$, are independent.

2. For each t, each realization of $\mathbf{W}_t$ is symmetric, stochastic and has positive diagonals, and it conforms to the structure of G_t, i.e., $[\mathbf{W}_t]_{ij} = 0$ if $\{i, j\} \notin E_t$, $i \neq j$.
3. There exists $\delta > 0$ such that, for each t, $[\mathbf{W}_t]_{ij} > \delta$ whenever $[\mathbf{W}_t]_{ij} > 0$.

The rest of the section gives preliminaries needed to state and prove the main result on rate $\mathcal{R}$ calculation.

Union graph of topologies. We denote by $\Gamma(t, s)$ the random graph that collects the edges from all the graphs G_r that appeared from time $r = s + 1$ to $r = t$, $s < t$, i.e.,

$$\Gamma(t, s) := \Gamma(\{G_t, G_{t-1}, \ldots, G_{s+1}\}).$$

Similarly as with Γ, for any $s < t$ we analogously define the product matrix over time window $r = s + 1$ to $r = t$,

$$\mathbf{\Phi}(t, s) = \mathbf{W}_t \mathbf{W}_{t-1} \cdot \ldots \mathbf{W}_{s+1}. \tag{4.4}$$

To simplify the notation, it is also of interest to introduce the error matrix $\widetilde{\mathbf{\Phi}}(t, s) = \mathbf{\Phi}(t, s) - \mathbf{J}$, a norm of which quantifies how close the product is to its limit $\mathbf{J}$.

Using the notion of the union graph Γ, we define the sequence of times T_i, $i = 1, 2, \ldots$, that mark times when Γ gets connected:

$$T_i = \min\{t \geq T_{i-1} + 1 : \ \Gamma(t, T_{i-1}) \text{ is connected}\}, \text{ for } i \geq 1, \tag{4.5}$$

with $T_0 = 0$. It is well-known that for every time window $(s, t]$ over which the occurred edges accumulate to a connected graph, the spectral norm of the error matrix $\widetilde{\mathbf{\Phi}}(t, s)$ drops below one (see, e.g., Lemma 11 in Bajović et al. (2013)). Hence, the sequence of times $\{T_i\}_{i\geq 1}$ therefore defines the times when the averaging process makes an improvement and gets closer to matrix J. Finally, for any fixed $t \geq 1$, we introduce the number of improvements until time t, denoted by M_t, $M_t = \max\{i \geq 0 : T_i \leq t\}$.

Vertex cut. For an arbitrary graph $G = (V, E)$ the vertex cut is defined as any subset of graph nodes C such that the remaining graph $G_{\setminus C} := \left(V \setminus C, E_{\setminus C}\right)$ is not connected, where $E_{\setminus C} := \{\{i, j\} \in E : i, j \in V \setminus C\}$; or, in words, $G_{\setminus C}$ is the graph obtained from the initial graph G by removing from G all the vertices that belong to C and all the edges connected to these vertices. We denote the set of all vertex cuts of G by $\mathcal{C}(G)$. If each node $i \in V$ is assigned a cost $c_i \in \mathbb{R}$, then, the minimal vertex cut is defined as a vertex cut $C \subseteq V$ such that the sum of costs of nodes in C is minimal among all vertex cuts $C \in \mathcal{C}(G)$. We denote the associated cost by

$$VC\left(G, \{c_i\}_{i\in V}\right) = \min_{C\in\mathcal{C}(G)} \left\{\sum_{i\in C} c_i\right\}. \tag{4.6}$$

4.4 Main result

We now state the main result of the chapter– existence and characterization of the rate $\mathcal{R}$ via vertex cut.

Theorem 4.4.1. *Let Assumptions 4.3.1 and 4.3.2 hold. The rate of consensus $\mathcal{R}$ in (1) is then, for any $\epsilon \in (0, 1]$, given by*

$$\mathcal{R} = VC\left(G, \{-\log q_i\}_{i\in V}\right), \tag{4.7}$$

where $q_i = (1 - p_i)$.

For simplicity, we will present a proof for the case when each p_i is strictly less than one, but the result can be extended to allow for the unit values of some or all of the p_i's. Also, note that, when $p_i = 1 - q_i = 1$, for $i = 1, ..., N$, then $\mathcal{R} = +\infty$.

Proof. We start with the following result from Bajović et al. (2013), which asserts that, if the number of improvements until time t scale linearly with t, for $t \geq 1$, then, starting from some finite time t, the events in (4.1) have zero probabilities (see Lemma 14, part 1 in Bajović et al. (2013)).

Lemma 4.4.2. *Consider the sequence of events $\{M_t \geq \beta t\}$, where $\beta \in (0, 1]$, $t = 1, 2, \ldots$. For every $\beta, \epsilon \in (0, 1]$, there exists sufficiently large $t_0 = t_0(\beta, \epsilon)$ such that*

$$\mathbb{P}\left(\left\|\widetilde{\mathbf{\Phi}}(t, 0)\right\| \geq \epsilon, \ M_t \geq \beta t\right) = 0, \quad \forall t \geq t_0(\beta, \epsilon). \tag{4.8}$$

Using the preceding result, it is easy to see that, for any fixed $\beta \in (0, 1)$, a necessary condition for $\left\|\widetilde{\mathbf{\Phi}}(t, 0)\right\| \geq \epsilon$ is that $M_t \leq \beta t$ (as otherwise the probability of this event is 0, which is asserted by Lemma 4.4.2). On the other hand, it is easy to see that the sufficient condition for this event to occur is that $M_t = 0$ (as in

this case $\left\|\widetilde{\mathbf{\Phi}}(t,0)\right\| = 1$, see Lemma 11 in Bajović et al. (2013)). Thus, we have that for each $\beta \in (0,1)$, all $t \geq t_0(\beta,\epsilon)$,

$$\mathbb{P}\left(M_t = 0\right) \leq \mathbb{P}\left(\left\|\widetilde{\mathbf{\Phi}}(t,0)\right\| \geq \epsilon\right) \leq \mathbb{P}\left(M_t < \beta t\right). \tag{4.9}$$

We prove the theorem by proving first that the left hand-side has an exponential decay rate equal to $\mathcal{R}$, as given in (4.7). We then show that the right hand-side probability in (4.9) decays with an β-dependent rate that gets closer to $\mathcal{R}$ as β decreases to 0, and in the limit equals $\mathcal{R}$.

We start by noting that the event $M_t = 0$ is equivalent to the event that $\Gamma(t,0)$ is disconnected. Let $C^\star$ denote the minimal vertex cut of $\overline{G}$, where the node costs are assigned as in the claim of the theorem. It is easy to see that a sufficient condition for $\Gamma(t,0)$ to be disconnected is that each of the nodes in the set $C^\star$ was inactive over the time interval from time 1 to time t. Thus,

$$\begin{aligned}\mathbb{P}\left(\Gamma(t,0)\,;\text{not connected}\right) &= \mathbb{P}\left(V_t \cap C^\star = \emptyset, k = 1,...,t\right)\\ &= \prod_{k=1}^{t}\prod_{i\in C^\star}(1-p_{i,k})\\ &\geq \prod_{i\in C^\star}(1-p_i)^t,\end{aligned} \tag{4.10}$$

where the second equality follows by the Assumption 4.3.1.1 and last inequality follows by the monotonic increase of the probabilities of nodes' activations, Assumption 4.3.1.2. Computing the logarithm, dividing by t and computing the limit $t \to +\infty$, we obtain from (4.10) and (4.9)

$$\liminf_{t\to+\infty}\frac{1}{t}\log\mathbb{P}\left(\left\|\widetilde{\mathbf{\Phi}}(t,0)\right\| \geq \epsilon\right) \geq -\mathcal{R}. \tag{4.11}$$

We next turn to computing the exponential rate of the right hand-side in the inequality (4.9). We start by noting that

$$\begin{aligned}\mathbb{P}\left(M_t < \beta t\right) &= \sum_{m=0}^{\lceil\beta t\rceil-1}\mathbb{P}\left(M_t = m\right)\\ &= \sum_{m=0}^{\lceil\beta t\rceil-1}\sum_{1\leq t_1\leq\ldots\leq t_m\leq k}\mathbb{P}\left(T_l = t_l, \text{for } 1\leq l\leq m,\, T_{m+1} > t\right),\end{aligned} \tag{4.12}$$

where in the second equality we consider all possible realizations of T_l, $l \leq m$. We focus on one arbitrary allocation of times $T_l = t_l$, $1 \leq l \leq m$, $T_{m+1} > t$ and the respective probability.

By the construction of the sequence T_l, for each $l \leq m$, we have that supergraph $\Gamma(t_l - 1, t_{l-1})$ is not connected, for $l \leq m$. Also, the condition $T_{m+1} > k$ implies that $\Gamma(t, T_m)$ is not connected. Denoting $t_{m+1} = t + 1$ for compact representation, we have

$$\begin{aligned}&\mathbb{P}\left(T_l = t_l, \text{for}\, l \leq m,\, T_{m+1} > t\right) &(4.13)\\ &\leq \mathbb{P}\left(\Gamma(t_l - 1, t_{l-1})\,\text{not connected, for}\, l \leq m+1\right)\\ &= \prod_{l=1}^{m+1}\mathbb{P}(\Gamma(t_l - 1, t_{l-1})\,\text{not connected}) &(4.14)\end{aligned}$$

where the last equality follows by the independence of the graph realizations. Note that, for arbitrary $t > s$,

the event that the supergraph $\Gamma(t,s)$ is not connected can be represented as the union of events that an arbitrary vertex cut of $\overline{G}$ was absent from the random graphs G_k over time window $s < k \leq t$, i.e.,

$$\{\Gamma(t,s) \text{ not connected}\} = \cup_{C \in \mathcal{C}(G)} \{V_k \cap C = \emptyset,\ s < k \leq t\}. \tag{4.15}$$

Applying (4.15) to each of the intervals $t_{l-1} < k \leq t_l - 1$ and computing the probabilities, we get by the union bound

$$\begin{aligned}
&\mathbb{P}\left(\{\Gamma(t_l - 1, t_{l-1}) \text{ not connected}\}\right) \\
&\leq \sum_{C \in \mathcal{C}(\overline{G})} \mathbb{P}\left(V_k \cap C = \emptyset,\ t_{l-1} \leq k \leq t_l - 1\right) \\
&= \sum_{C \in \mathcal{C}(\overline{G})} \prod_{t_{l-1} \leq k \leq t_l - 1} \prod_{i \in C} (1 - p_{i,k}).
\end{aligned} \tag{4.16}$$

Expressing $1 - p_{i,k} = (1-p_i)\left(1 + \frac{p_i}{1-p_i}\alpha_k\right) \leq (1-p_i)(1+\kappa\alpha_k)$, where $\kappa = \max_{i \in V} p_i/(1-p_i)$, and using the fact that $1 + x \leq e^x$, we obtain from (4.16):

$$\begin{aligned}
&\mathbb{P}\left(\{\Gamma(t_l - 1, t_{l-1}) \text{ not connected}\}\right) \\
&\leq \sum_{C \in \mathcal{C}(\overline{G})} e^{N\kappa \sum_{t_{l-1} < k \leq t_l - 1} \alpha_k} \prod_{i \in C} (1-p_i)^{(t_l - 1 - t_{l-1})} \\
&\leq \left|\mathcal{C}(\overline{G})\right| e^{N\kappa \sum_{t_{l-1} < k \leq t_l - 1} \alpha_k} e^{-(t_l - 1 - t_{l-1}) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)}.
\end{aligned} \tag{4.17}$$

Applying the preceding inequality for each time interval $t_{l-1} < k \leq t_l - 1$ yields in (4.13)

$$\begin{aligned}
&\mathbb{P}\left(T_l = t_l, \text{for } l \leq m,\ T_{m+1} > t\right) \\
&\leq \left|\mathcal{C}(\overline{G})\right|^m e^{N\kappa \sum_{1 \leq k \leq t,\ k \neq t_i} \alpha_k} e^{-(k-m) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)} \\
&\leq \left|\mathcal{C}(\overline{G})\right|^m e^{N\kappa \sum_{k=1}^{t} \alpha_k} e^{-(k-m) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)}.
\end{aligned} \tag{4.18}$$

The preceding bound holds for each of the terms in (4.12) that correspond to a fixed number of improvements $M_t = m$, and since there are $\binom{t}{m}$ possible allocations of times $T_1, ..., T_m$, we obtain

$$\begin{aligned}
&\mathbb{P}\left(M_t = m\right) \\
&\leq \binom{t}{m} \left|\mathcal{C}(\overline{G})\right|^m e^{-(k-m) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)} e^{N\kappa \sum_{k=1}^{t} \alpha_k} \\
&\leq \left(\frac{te}{m}\right)^m \left|\mathcal{C}(\overline{G})\right|^m e^{-(k-m) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)} e^{N\kappa \sum_{k=1}^{t} \alpha_k}.
\end{aligned} \tag{4.19}$$

It is easy to see that, for any $\beta < 1/2$, the preceding bound is maximal for $m = \beta t$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(M_t < \beta t\right) \leq \beta t \left(\frac{te}{\beta t}\right)^{\beta t} &\left|\mathcal{C}(\overline{G})\right|^{\beta t} \\
&\times e^{-(k - \beta t) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)} e^{N\kappa \sum_{k=1}^{t} \alpha_k}.
\end{aligned} \tag{4.20}$$

Computing the logarithm, dividing by t, and taking the limit $t \to +\infty$ yields

$$\limsup_{t\to+\infty} \mathbb{P}\left(M_t < \beta t\right) \leq \beta \log \frac{e}{\beta} + \beta \log |\mathcal{C}(\overline{G})| \\ - (1-\beta) VC\left(\overline{G}, \{-\log q_i\}_{i\in V}\right) + \kappa \lim_{t\to+\infty} \frac{\sum_{k=1}^{t} \alpha_k}{t}. \tag{4.21}$$

By Assumption 4.3.1.2, the last term vanishes, and taking the infimum with respect to $\beta > 0$, we finally obtain

$$\limsup_{t\to+\infty} \mathbb{P}\left(M_t < \beta t\right) \leq -VC\left(\overline{G}, \{-\log q_i\}_{i\in V}\right). \tag{4.22}$$

Combining with (4.11), the claim of Theorem 4.4.1 follows. □

4.5 Application to distributed detection

We now demonstrate the usefulness of Theorem 4.4.1 by applying it to consensus+innovations distributed detection; see also Bajovic et al. (2011); Sahu and Kar (2017). The detection problem is as follows. Sensors in a N-node network cooperate to decide on a binary hypothesis, H_1 versus H_0. Each sensor i, at each time step t, $t = 1, 2, ...$, performs a measurement $Y_{i,t}$; the measurements are i.i.d., both in time and across sensors, where under hypothesis H_l, $Y_{i,t}$ has the density function f_l, $l = 0, 1$, for $i = 1, \ldots, N$ and $t = 1, 2, \ldots$. In the aforementioned detector, each sensor i maintains its (scalar) local decision statistic $x_{i,t}$ and compares it with the zero threshold; if $x_{i,t} > 0$, sensor i accepts H_1; otherwise, it accepts H_0. At each time t, a sensor i updates its decision statistic $x_{i,t}$ by exchanging its decision statistic in the neighborhood and by assimilation of decision statistics from its neighborhood and its latest sensed information through a log-likelihood ratio $L_{i,t} = \log \frac{f_1(Y_{i,t})}{f_0(Y_{i,t})}$:

$$x_{i,t} = \sum_{j\in O_{i,t}} [\mathbf{W}_t]_{ij} \left(\frac{t-1}{t} x_{j,t-1} + \frac{1}{t} L_{j,t}\right), \tag{4.23}$$

with $x_{i,0} = 0$. Here $O_{i,t}$ is the (random) neighborhood of sensor i at time t (including i), and $[\mathbf{W}_t]_{ij}$ is the (random) averaging weight that sensor i assigns to sensor j at time t. We let the $N \times N$ matrix $\mathbf{W}_t$ that collects the weights $[\mathbf{W}_t]_{ij}$ in (4.23) adhere to the model in Assumption 4.3.2. In other words, sensors utilize a randomized communication protocol as described in Assumptions 4.3.1 and 4.3.2 and the preceding paragraphs. We additionally assume that sensors' observations are independent from the activation protocol, i.e., from matrices $\mathbf{W}_t$. To assess communication-wise benefits of the sparsified communication protocol, we benchmark detector (4.23) against the detector with the same algorithmic form as in (4.23), except that the weight matrix is replaced by a constant doubly stochastic matrix $\mathbf{W}$. Intuitively, the benchmark utilizes communications across all links at all times, and it is hence natural to expect that it has a better performance with respect to time, i.e., with respect to the number of measurements processed. However, as shown ahead, the detector with sparsified communications practically matches the benchmark's performance time-wise while achieving a better performance communication-wise.

For detector (4.23), rate of consensus $\mathcal{R}$ plays a major role in its asymptotic performance, as measured by the worst-sensor error exponent of the Bayes error probability: $\min_{i=1,\ldots,N} \left\{-\frac{1}{t} \log\left(P_{i,t}^{\mathrm{e}}\right)\right\}$, where $P_{i,t}^{\mathrm{e}}$ is the Bayes error probability for sensor i at time t. While prior work Bajovic et al. (2011) characterized asymptotic performance of detectors of form (4.23) when the weight matrices are deterministic or randomly varying in an i.i.d. fashion, Theorem 4.4.1 gives us the opportunity to characterize here the detection performance

under the assumed sparsified time-varying protocol. Namely, it can be shown that the results in Bajovic et al. (2011) can be extended to the matrix model here to show the following: if the rate of consensus $\mathcal{R}$ in (4.1) satisfies: $\mathcal{R} \geq (N-1)C_{\text{tot}}$, where $C_{\text{tot}} = C_{\text{tot}}(N, f_1, f_0)$ is the exponential decay rate of the error probability of the best centralized detector (Chernoff information), then distributed detector (4.23) with sparsified time varying communications as defined in Assumptions 1 and 2 is asymptotically optimal, i.e., it achieves the best possible error exponent.

We now comment on the achieved result. It is known from Bajovic et al. (2011) that a detector of the form (4.23) is asymptotically optimal under either deterministic or under an i.i.d. weight matrix model, provided that $\mathcal{R}$ exceeds $(N-1)C_{\text{tot}}$. Here, we show that asymptotic optimality is still achievable under the time-varying weight model satisfying Assumption 4.3.2 with slowly increasing node activation probabilities. This result has, for example, the following implication on improving communication efficiency in distributed detection: a detector of form (4.23) with the assumed time-varying randomized communication protocol – wherein the activation probabilities are slowly increasing to unity (and so $\mathcal{R} = +\infty$) – is asymptotically optimal for any value of SNR (any value of C_{tot}) and is equivalent to the detector with a constant weight matrix $\mathbf{W}$. Hence, as the detector with the randomized protocol has a lower communication cost while essentially equivalent performance time-wise, one can expect that it becomes more communication efficient. We next present a simulation example that confirms such improvements in communication efficiency.

We consider a geometric network with $N = 20$ sensors. We place the sensors uniformly over a unit square, and connect those sensors whose distance d_{ij} is less than a radius. The total number of (undirected) links is 63. For the sensors' measurements, we use the Gaussian distribution $f_1 \sim \mathcal{N}(m, \sigma^2)$, $f_0 \sim \mathcal{N}(0, \sigma^2)$, with $m = 0.01$, and $\sigma^2 = 0.2$. We consider two different detectors of form (4.23). The first one is the benchmark for which each link is online at all times. The second detector utilizes at each node i activation probability $p_t = 1 - 1/\log(t+2)$, $t = 1, 2, \ldots$. For the averaging weights of the benchmark, we use for each link $\{i, j\}$ a constant weight $\mathbf{W}_{ij} = 0.1$. To compensate for random activations, the second detector assigns weight $0.1/p_t$ whenever a link is online. Figure 1 (top) plots the simulated Bayes error probability versus per node communication cost, averaged across nodes and across 20,000 Monte Carlo algorithm runs. We can see that the detector with time-varying sparsified communications (solid line) achieves significant communication savings with respect to the benchmark, while at the same time practically matches the benchmark's performance with respect to time (see Figure 1, bottom).

4.6 Contributions

- **Large Deviations Characterization for product of doubly stochastic weight matrices**: For product of doubly stochastic weight matrices which are not identically distributed, we characterized the exponential rate of convergence and quantified the large deviations exponent to be given by the minimal vertex cut of the baseline graph, where the nodes' associated cut costs are defined by the nodes' limiting activation probabilities.
- **Communication Efficient Distributed Detector**: We theoretically established that the detector with time-varying and sparsified protocol can be designed to achieve asymptotic optimality at all signal-to-noise ratio (SNR) regimes; this is achieved when the activation probabilities converge to unity, possibly at a very slow rate, e.g., as $1 - \Omega(1/\log(t))$. Therefore, effectively, we constructed a "universally" asymptotically optimal detector that makes significant communication savings.

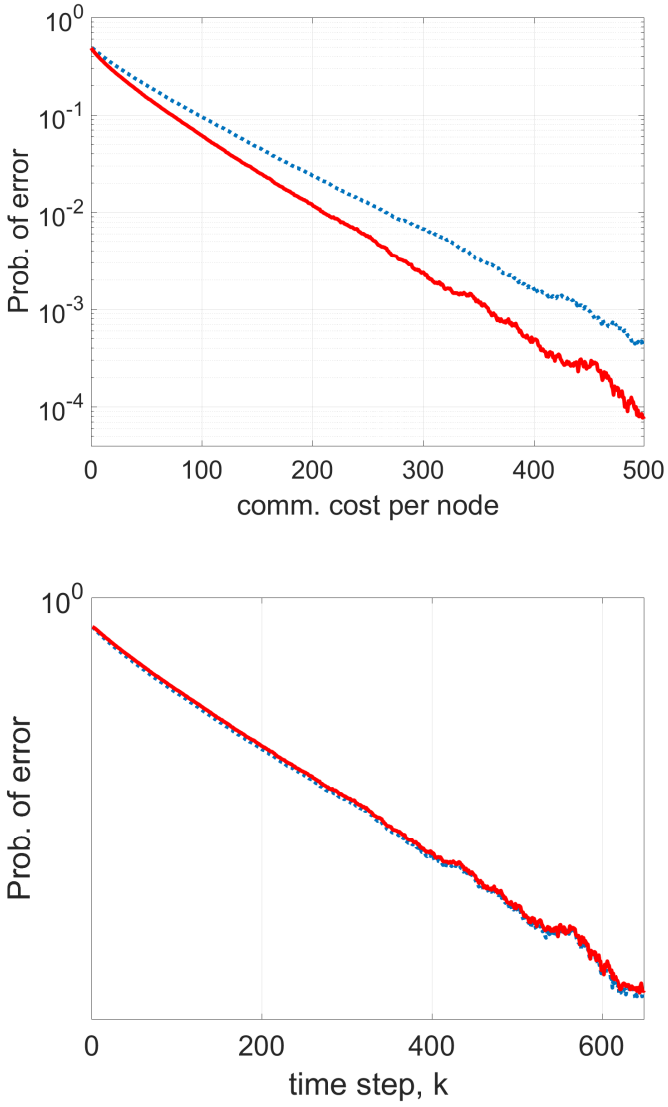


Figure 4.1: Estimated probability of error (in the log scale) versus per node communication cost (top) and versus time (bottom) for the benchmark detector (blue dotted line) and the detector with sparsifying communications (red solid line).

4.7 Conclusion and Future Directions

We have derived the exact large deviations rate for products of a class of non-i.i.d. random stochastic and symmetric matrices that arise with distributed inference under randomized communication protocols. We applied the results to consensus+innovations distributed detection to derive universally asymptotically optimal detectors with significantly reduced communication cost. Future directions involve characterizing large deviations rate for products of matrices with Markovianity and extending the communication efficient protocol to composite hypothesis testing.

Part II

Distributed Estimation

Chapter 5

Communication Efficient Distributed Linear Estimation: *CREDO*

5.1 Introduction

Distributed data processing techniques have been increasingly employed to solve problems pertaining to optimization and statistical inference. With massive computing resources that are available at scale, and ever growing sizes of data sets, it becomes highly desirable, if not necessary, to distribute the task among multiple machines or multiple cores. The benefits of splitting the task into smaller subtasks are multi-pronged, namely, it makes the problem at hand, scalable, parallelized and fast. In the context of distribution stochastic optimization, several methods (see, for example Zhang et al. (2013b,a); Heinze et al. (2016); Ma et al. (2015); Recht et al. (2011)) have been proposed which exhibit impressive performance in platforms such as Mapreduce and Spark. The aforementioned methods, though highly scalable, are designed for master-worker or similar types of architectures. That is, they require the presence of a master node, i.e., a central coordinator which is tasked with splitting the dataset by data points (batches) or by features among worker nodes and enabling the read/write operations of the iterates of the worker nodes so as to ensure information fusion across the worker nodes. However, with several emerging applications, master-worker type architectures may not be feasible or desirable due to physical constraints. Specifically, we are interested in systems and applications where the entire data is not available at a central/master node, is sensed in a streaming fashion and is intrinsically distributed across the worker nodes. Such scenarios arise, e.g., in systems which involve Internet of Things (IoT). For example, a smart campus with sensors of various kinds, a smart building or monitoring a large scale industrial plant. Therein, a network of large number of heterogeneous entities (usually, geographically spread) connected in a arbitrary network structure individually perform sensing for data arriving in a streaming fashion. The sensing devices have limited communication capabilities owing to on board power constraints and harsh environments. A typical IoT framework is characterized by a heterogeneous network of entities without a central coordinator, where entities have localized knowledge and can exchange information among each other through an arbitrary pre-specified communication graph. Furthermore, the data samples arrive in a streaming fashion. The ad-hoc nature of the IoT framework necessitates the information exchange in a crafted manner, rather than just a single or few rounds of communication at the end as in Zhang et al. (2013b,a); Heinze et al. (2016); Ma et al. (2015).

Distributed algorithms for statistical inference and optimization in the aforementioned frameworks are char-

acterized by a central coordinator-less recursive procedures, where each entity in the network maintains its own estimate or optimizer for the problem at hand. Also, due to heterogeneity of the entities and lack of global model information, the information exchange is limited to the iterates and not the raw data. This additionally enhances privacy as far as the data samples are concerned. In particular, the diffusion and *consensus+innovations* schemes have been extensively used for various distributed inference problems in the aforementioned frameworks, which include distributed parameter estimation, distributed detection and multi-task learning, to name a few (see, for example Kar and Moura (2011); Cattivelli and Sayed (2010); Bajović et al. (2015); Lopes and Sayed (2008); Sahu and Kar (2016); Jakovetic et al. (2011); Chen et al. (2014)). Other variants of distributed recursive algorithms of such kinds have generated a lot of interest of lately (see, for example Nedić et al. (2014); Ram et al. (2010a); Braca et al. (2008); Ram et al. (2010b, 2009); Nedic and Ozdaglar (2009)).

An entity or node in an IoT setup is usually equipped with on board communication and computation units. However, finite battery power calls for frugal communication protocols as the power used in communication tends to beat the power required for on board computation. Thus, *communication efficiency* is highly relevant and sought for in such scenarios. The previously studied distributed algorithms mentioned above, have the mean square error decay as $\Theta(\mathcal{C}_t^{-1})$, in terms of the communication cost $\mathcal{C}_t$. In this chapter, we present a distributed recursive algorithm, **C**ommunication Efficient **RE**cursive **D**istributed Estimati**O**n ($\mathcal{CREDO}$) characterized by a frugal communication protocol while guaranteeing provably reasonable performance, which improves the MSE communication rate dependence to $\Theta\left(\mathcal{C}_t^{-2+\zeta}\right)$. Specifically, this chapter focuses on the above described class of *distributed, recursive* algorithms for estimation of an unknown vector parameter $\boldsymbol{\theta}$, where each worker continuously observes noisy measurements of low-dimensional linear transformations of $\boldsymbol{\theta}$. For this problem,we improve the communication efficiency of existing distributed recursive estimation methods primarily in the *consensus+innovations* and the diffusion frameworks Kar and Moura (2011); Cattivelli and Sayed (2010); Bajović et al. (2015); Lopes and Sayed (2008); Sahu and Kar (2016); Jakovetic et al. (2011); Chen et al. (2014), which in turn improves the communication efficiency of variants such as Nedić et al. (2014); Ram et al. (2010a); Braca et al. (2008); Ram et al. (2010b, 2009); Nedic and Ozdaglar (2009).

5.1.1 Contributions

Our contributions are as follows:

We propose a scheme, namely $\mathcal{CREDO}$, where each node at time t communicates only with a certain probability that decays sub linearly to zero in t. That is, communications are increasingly sparse, so that communication cost scales as $\Theta(t^{\delta})$, where the growing rate δ is a tunable parameter strictly less than one that can go down to 0.5.

We show that, despite significantly lower communication cost, the proposed method achieves the best possible $\Theta(1/t)$ rate of MSE decay in time t (t also equals to per-worker number of data samples). Importantly, this result translates into significant improvements in the rate at which MSE decays with communication cost $\mathcal{C}_t$ – namely from $\Theta(1/\mathcal{C}_t)$ with existing methods to $\Theta(1/\mathcal{C}_t^{2-\zeta})$ with the proposed method, where $\zeta > 0$ arbitrarily small.

We further study asymptotic normality and the corresponding asymptotic variance of the proposed method (that in a sense relates to the constant in the $\Theta(1/t)$ MSE decay rate). We characterize and quantify interesting trade-offs between the communication cost and the asymptotic variance of the method. In particular, we explicitly quantify the regime (the range of communication cost growing rate δ) where the

asymptotic variance is network-independent and, at the same time, communication cost is strictly sub linear ($\delta < 1$). Numerical examples both on synthetic and real data sets confirm the significantly improved communication efficiency of the proposed method.

A key insight behind $\mathcal{CREDO}$ is that it recognizes that inter-node communications can be made (probabilistically) increasingly sparse without sacrificing estimation performance. It is known from stochastic approximation that the weights that each node assigns to its neighboring nodes can be made to decrease with time while keeping the estimator strongly consistent. $\mathcal{CREDO}$ replaces such a deterministic weight $w(t)$ (t being time) with a Bernoulli random variable that equals one with probability $w(t) < 1$. Thus, $\mathcal{CREDO}$ is much cheaper to implement as communication takes place only with probability $w(t)$, with $w(t)$ decaying to zero. Despite the adaptive weighting being very different, existence of broad regimes of algorithm parameters are shown where $\mathcal{CREDO}$'s estimation performance matches closely the benchmarks iteration-wise. However, as $\mathcal{CREDO}$ has much fewer communications per iteration, it becomes more communication efficient.

To achieve the results above, we developed several technical innovations. Specifically, the studied setup requires analysis of *mixed time-scale stochastic approximation* algorithms with *three different time scales*. This setup stands in contrast with the classical single time-scale stochastic approximation, the properties of which are well known. It is also very different from the more commonly studied two time-scale stochastic approximation (see, for instance Borkar (2008)) in which a fast process is coupled with a slower dynamical system. We develop here new technical tools that allow us to handle the case of number of operating time-scales to be three instead of two as in Kar et al. (2013a) for mixed time-scale stochastic approximation (described in details later).

5.2 Related Work

We now briefly review the literature on distributed inference and motivate our algorithm $\mathcal{CREDO}$. Distributed inference algorithms can be broadly divided into two classes. The first class of distributed inference algorithms proposed in Liu and Ihler (2014); Ma et al. (2015); Ma and Takáč (2015); Heinze et al. (2016); Zhang et al. (2013b) require a central master node so as to coordinate as far as assigning sub-tasks to the worker nodes is concerned. There are two reasons as to why such methods do not apply in our setting. Firstly, in them, in order for the central node to be able to assign sub-tasks, such a setup requires the central node to have access to the entire dataset. However, in the setup considered in this chapter, where the data samples are intrinsically distributed among the worker nodes and rather ad-hoc, the presence of a central master node is highly impractical. Even in the case when the data is distributed among nodes to start with, the local data samples collected via (3.15) are not sufficient to uniquely reconstruct the global parameter of interest. In particular, the sensing matrix $\mathbf{H}_n$ at an agent n is rank deficient, i.e., $rank(\mathbf{H}_n) = M_n < M$, in general. We refer to this phenomenon as *local unobservability*. With communication being the most power hungry aspect for an ad-hoc sensing entity, communicating raw data back to a central node so as to re-assign the data among worker nodes is prohibitive. Thus in such an ad-hoc and distributed setup, a communication protocol should involve information fusion via exchange of the latest estimates among worker nodes enables each worker node to aggregate information about all the entries of the parameter.

Secondly, they do not apply to the model (3.15) being considered here. For example, if $\mathbf{H}_n = h\mathbf{I}$, it reduces to the case, where each worker can work independently to obtain a reasonably good estimate of $\boldsymbol{\theta}$ and algorithms such as $CoCoA^+$ (Ma et al. (2015)) and $Dual - LOCO$ (Heinze et al. (2016)) can then address the problem efficiently through data splitting across samples and features respectively. However,

if $\mathbf{H}_n = \mathbf{e}_n^\top$, where $\mathbf{e}_n$ is the n-th canonical basis vector of $\mathbb{R}^M$, a random splitting across samples would lead to estimates with a high mean square error, while a feature wise splitting is still possible. But, in the case when, $\mathbf{H}_n = (\mathbf{e}_n + \mathbf{e}_{n-1})^\top$, neither sample splitting nor feature splitting is possible and such a setup necessitates more rounds of communication as opposed to just one round of communication at the end as in the case of $CoCoA^+$ (Ma et al. (2015)) and $Dual - LOCO$ (Heinze et al. (2016)).

The second class of distributed inference algorithms involve setups, which are characterized by the absence of a master node. Communication efficient distributed recursive algorithms in the context of distributed optimization with no central node, where data is available apriori and is not collected in a streaming fashion has been addressed in Tsianos et al. (2012, 2013); Jakovetic et al. (2016) through increasingly sparse communication, adaptive communication scheme and selective activation of nodes respectively. However, the explicit characterization of the performance metric for instance MSE, in terms of the communication cost has not been addressed in the aforementioned references.

The very well studied class of distributed estimation algorithms in the *consensus+innovations* framework Kar and Moura (2011); Kar et al. (2013a) characterize the algorithm parameters, under which estimate sequences optimal in the sense of asymptotic covariance can be obtained. However, the inter-agent message passing and the associated communication cost is not taken into account in the aforementioned algorithms. The lack of exploration into the dimension of communication cost in the context of distributed estimation algorithms in the *consensus+innovations* framework motivated us to develop a stochastic communication protocol in this chapter, which exploited the redundancy in inter-agent message passing while not compromising on the optimality aspect of the estimate sequence. Hence, in order to test the efficacy of our stochastic message-passing protocol, we take the distributed estimation algorithm proposed in Kar and Moura (2011); Kar et al. (2013a) as the benchmark.

5.3 Problem Setup: Motivation and Preliminaries

There are N workers deployed in the network. Every worker n at time index t makes a noisy observation $\mathbf{y}_n(t)$, a noisy function of $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^M$. Formally the observation model for the n-th worker is given by,

$$\mathbf{y}_n(t) = \mathbf{H}_n\boldsymbol{\theta} + \gamma_n(t), \tag{5.1}$$

where $\mathbf{H}_n \in \mathbb{R}^{M_n \times M}$ is the sensing matrix, where $M_n < M$, $\{\mathbf{y}_n(t)\} \in \mathbb{R}^{M_n}$ is the observation sequence for the n-th worker and $\{\gamma_n(t)\}$ is a zero mean temporally independent and identically distributed (i.i.d.) noise sequence at the n-th worker with nonsingular covariance $\boldsymbol{\Sigma}_n$, where $\boldsymbol{\Sigma}_n \in \mathbb{R}^{M_n \times M_n}$. The noise processes are independent across different workers. We state an assumption on the noise processes before proceeding further. The linear observation model assumed here can be extended to nonlinear observation models when the nonlinear functions are regular and sufficiently smooth.

Assumption 5.3.1. *There exists $\epsilon_1 > 0$, such that, for all n, $\mathbb{E}_{\boldsymbol{\theta}}\left[\|\gamma_n(t)\|^{2+\epsilon_1}\right] < \infty$.*

The above assumption encompasses a general class of noise distributions in the setup. The heterogeneity of the setup is exhibited in terms of the sensing matrix and the noise covariances at the worker nodes. Each worker node is interested in reconstructing the true underlying parameter $\boldsymbol{\theta}$. We assume a worker node is aware only of its local observation model and hence does not know about the observation matrix and noise processes of other worker nodes. In this chapter, we are interested in estimators that, at each node n, continuously produce estimates of $\boldsymbol{\theta}$ at each time t, i.e., after each new sample $\mathbf{y}(t)$ is acquired.

5.3.1 Motivation and Related Work

We now briefly review the literature on distributed inference and motivate our algorithm $\mathcal{CREDO}$. Distributed inference algorithms can be broadly divided into two classes. The first class of distributed inference algorithms proposed in Liu and Ihler (2014); Ma et al. (2015); Ma and Takáč (2015); Heinze et al. (2016); Zhang et al. (2013b) require a central master node so as to coordinate as far as assigning sub-tasks to the worker nodes is concerned. There are two reasons as to why such methods do not apply in our setting. Firstly, in them, in order for the central node to be able to assign sub-tasks, such a setup requires the central node to have access to the entire dataset. However, in the setup considered in this chapter, where the data samples are intrinsically distributed among the worker nodes and rather ad-hoc, the presence of a central master node is highly impractical. Even in the case when the data is distributed among nodes to start with, the local data samples collected via (5.1) are not sufficient to uniquely reconstruct the global parameter of interest. In particular, the sensing matrix $\mathbf{H}_n$ at an agent n is rank deficient, i.e., $rank(\mathbf{H}_n) = M_n < M$, in general. We refer to this phenomenon as *local unobservability*. With communication being the most power hungry aspect for an ad-hoc sensing entity, communicating raw data back to a central node so as to re-assign the data among worker nodes is prohibitive. Thus in such an ad-hoc and distributed setup, a communication protocol should involve information fusion via exchange of the latest estimates among worker nodes enables each worker node to aggregate information about all the entries of the parameter.

Secondly, they do not apply to the model (5.1) being considered here. For example, if $\mathbf{H}_n = h\mathbf{I}$, it reduces to the case, where each worker can work independently to obtain a reasonably good estimate of $\boldsymbol{\theta}$ and algorithms such as $CoCoA^+$ (Ma et al. (2015)) and $Dual - LOCO$ (Heinze et al. (2016)) can then address the problem efficiently through data splitting across samples and features respectively. However, if $\mathbf{H}_n = \mathbf{e}_n^\top$, where $\mathbf{e}_n$ is the n-th canonical basis vector of $\mathbb{R}^M$, a random splitting across samples would lead to estimates with a high mean square error, while a feature wise splitting is still possible. But, in the case when, $\mathbf{H}_n = (\mathbf{e}_n + \mathbf{e}_{n-1})^\top$, neither sample splitting nor feature splitting is possible and such a setup necessitates more rounds of communication as opposed to just one round of communication at the end as in the case of $CoCoA^+$ (Ma et al. (2015)) and $Dual - LOCO$ (Heinze et al. (2016)).

The second class of distributed inference algorithms involve setups, which are characterized by the absence of a master node. Communication efficient distributed recursive algorithms in the context of distributed optimization with no central node, where data is available apriori and is not collected in a streaming fashion has been addressed in Tsianos et al. (2012, 2013); Jakovetic et al. (2016) through increasingly sparse communication, adaptive communication scheme and selective activation of nodes respectively. However, the explicit characterization of the performance metric for instance MSE, in terms of the communication cost has not been addressed in the aforementioned references.

The very well studied class of distributed estimation algorithms in the *consensus+innovations* framework Kar and Moura (2011); Kar et al. (2013a) characterize the algorithm parameters, under which estimate sequences optimal in the sense of asymptotic covariance can be obtained. However, the inter-agent message passing and the associated communication cost is not taken into account in the aforementioned algorithms. The lack of exploration into the dimension of communication cost in the context of distributed estimation algorithms in the *consensus+innovations* framework motivated us to develop a stochastic communication protocol in this chapter, which exploited the redundancy in inter-agent message passing while not compromising on the optimality aspect of the estimate sequence. Hence, in order to test the efficacy of our stochastic message-passing protocol, we take the distributed estimation algorithm proposed in Kar and Moura (2011); Kar et al. (2013a) as the benchmark.

5.4 $\mathcal{CREDO}$: Communication efficient REcursive Distributed estimatiOn

We now present the proposed $\mathcal{CREDO}$ estimator. $\mathcal{CREDO}$ is based on a specifically handcrafted time decaying communication rate protocol. Intuitively, we basically exploit the idea that, once the information flow starts in the graph and a worker node is able to accumulate sufficient information about the parameter of interest, the need to communicate with its neighboring nodes goes down. Technically speaking, for each node n, at every time t, we introduce a binary random variable $\psi_{n,t}$, where

$$\psi_{n,t} = \begin{cases} \rho_t & \textit{with probability } \zeta_t \\ 0 & \textit{else}, \end{cases} \tag{5.2}$$

where $\psi_{i,t}$'s are independent both across time and the nodes, i.e., across t and n respectively. The random variable $\psi_{n,t}$ abstracts out the decision of the node n at time t whether to participate in the neighborhood information exchange or not. We specifically take ρ_t and ζ_t of the form

$$\rho_t = \frac{\rho_0}{(t+1)^{\epsilon/2}}, \zeta_t = \frac{\zeta_0}{(t+1)^{(\tau_1/2-\epsilon/2)}}, \tag{5.3}$$

where $0 < \epsilon < \tau_1$ and $0 < \tau_1 \leq 1$. Furthermore, define β_t to be

$$\beta_t = (\rho_t \zeta_t)^2 = \frac{\beta_0}{(t+1)^{\tau_1}}. \tag{5.4}$$

With the above development in place, we define the random time-varying Laplacian $\mathbf{L}(t)$, where $\mathbf{L}(t) \in \mathbb{R}^{N \times N}$ which abstracts the inter-node information exchange as follows:

$$\mathbf{L}_{i,j}(t) = \begin{cases} -\psi_{i,t}\psi_{j,t} & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l \neq i} \psi_{i,t}\psi_{l,t} & i = j. \end{cases} \tag{5.5}$$

The above communication protocol allows two nodes to communicate only when the link is established in a bi-directional fashion and hence avoids directed graphs. The design of the communication protocol as depicted in (5.2)-(5.5) not only decays the weight assigned to the links over time but also decays the probability of the existence of a link. Such a design is consistent with frameworks where the working nodes have finite power and hence not only the number of communications, but also, the quality of the communication decays over time. We have, for $\{i,j\} \in E$:

$$\begin{aligned} \mathbb{E}\left[\mathbf{L}_{i,j}(t)\right] &= -\left(\rho_t \zeta_t\right)^2 = -\beta_t = -\frac{c_3}{(t+1)^{\tau_1}} \\ \mathbb{E}\left[\mathbf{L}_{i,j}^2(t)\right] &= \left(\rho_t^2 \zeta_t\right)^2 = \frac{c_4}{(t+1)^{\tau_1+\epsilon}}. \end{aligned} \tag{5.6}$$

Thus, we have that, the variance of $\mathbf{L}_{i,j}(t)$ is given by,

$$Var\left(\mathbf{L}_{i,j}(t)\right) = \frac{\beta_0 \rho_0^2}{(t+1)^{\tau_1+\epsilon}} - \frac{a^2}{(t+1)^{2\tau_1}}. \tag{5.7}$$

Define, the mean of the random time-varying Laplacian sequence $\{\mathbf{L}(t)\}$ as $\overline{\mathbf{L}}(t) = \mathbb{E}[\mathbf{L}(t)]$ and $\widetilde{\mathbf{L}}(t) = \mathbf{L}(t) - \overline{\mathbf{L}}(t)$. Note that, $\mathbb{E}\left[\widetilde{\mathbf{L}}(t)\right] = \mathbf{0}$, and

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{L}}(t)\right\|^2\right] \leq N^2\mathbb{E}\left[\widetilde{\mathbf{L}}_{i,j}^2(t)\right] = \frac{N^2\beta_0\rho_0^2}{(t+1)^{\tau_1+\epsilon}} - \frac{N^2a^2}{(t+1)^{2\tau_1}}, \tag{5.8}$$

where $\|\cdot\|$ denotes the L_2 norm. The above equation follows from equivalence of the L_2 and Frobenius norms. We also have that, $\overline{\mathbf{L}}(t) = \beta_t\overline{\mathbf{L}}$, where

$$\overline{\mathbf{L}}_{i,j} = \begin{cases} -1 & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l\neq i} L_{i,l} & i = j. \end{cases} \tag{5.9}$$

We formalize the assumptions on the inter-worker communication graph and global observability.

Assumption 5.4.1. *We require the following global observability condition. The matrix* $\mathbf{G} = \sum_{n=1}^{N} \mathbf{H}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{H}_n$ *is full rank.*

Assumption 5.4.1 is crucial for our distributed setup. This notion of rendering the parameter locally unobservable while it being globally observable in the context of distributed inference was introduced in Kar and Moura (2011), and has been subsequently used in Lalitha et al. (2014); Sahu and Kar (2017). It is to be noted that such an assumption is needed for even a setup with a centralized node which has access to all the data samples at each of the worker nodes at each time. Assumption 5.4.1 ensures that if a node could stack all the data samples together at any time t, it would have sufficient information about the parameter of interest so as to be able to estimate the parameter of interest without any communication. Hence, the requirement for this assumption naturally extends to our distributed setup. We formalize an assumption on the connectivity of the inter-agent communication graph before proceeding further.

Assumption 5.4.2. *The inter-agent communication graph is connected on average, i.e.,* $\lambda_2(\overline{\mathbf{L}}) > 0$*, which implies* $\lambda_2(\overline{\mathbf{L}}(t)) > 0$*, where* $\overline{\mathbf{L}}(t)$ *denotes the mean of the Laplacian matrix* $\mathbf{L}(t)$ *and* $\lambda_2(\cdot)$ *denotes the second smallest eigenvalue.*

Assumption 5.4.2 ensures consistent information flow among the worker nodes. Technically speaking, the communication graph modeled here as a random undirected graph need not be connected at all times. Hence, at any given time, only a few of the possible links could be active. The connectedness in average basically ensures that over time, the information from each worker node in the graph reaches other worker nodes over time in a symmetric fashion and thus ensuring information flow. It is to be noted that assumption 5.4.2 ensures that $\overline{\mathbf{L}}(t)$ is connected at all times as $\overline{\mathbf{L}}(t) = \beta_t\overline{\mathbf{L}}$. With the communication protocol established, we propose an update, where every node n generates an estimate sequence $\{\mathbf{x}_n(t)\}$, where $\mathbf{x}_n(t) \in \mathbb{R}^M$ in the following way:

$$\begin{aligned} \mathbf{x}_n(t+1) = \mathbf{x}_n(t) &- \underbrace{\sum_{l\in\Omega_n} \psi_{n,t}\psi_{l,t}\left(\mathbf{x}_n(t) - \mathbf{x}_l(t)\right)}_{\text{Neighborhood Consensus}} \\ &+ \underbrace{\alpha_t\mathbf{H}_n^\top\mathbf{\Sigma}_n^{-1}\left(\mathbf{y}_n(t) - \mathbf{H}_n\mathbf{x}_n(t)\right)}_{\text{Local Innovation}}, \end{aligned} \tag{5.10}$$

where Ω_n denotes the neighborhood of node n with respect to the network encapsulated by $\overline{\mathbf{L}}$ and α_t is the innovation gain sequence which is given by $\alpha_t = a/(t+1)$. It is to be noted that a node n can send and receive information in its neighborhood at time t, when $\psi_{n,t} \neq 0$. At the same time, when $\psi_{n,t} = 0$, node n neither transmits nor receives information. The link between node n and node l gets assigned a weight of ρ_t^2 if and only if $\psi_{n,t} \neq 0$ and $\psi_{l,t} \neq 0$.

Remark 5.4.1. *The stochastic approximation procedure, employed here is a mixed time-scale stochastic approximation as opposed to the classical single time-scale stochastic approximation, the properties of which are well known. Note, the above notion of mixed time-scale is very different from the more commonly studied two time-scale stochastic approximation (see, for instance Borkar (2008)) in which a fast process is coupled with a slower dynamical system. More relevant to our study are the mixed time-scale dynamics encountered in Gelfand and Mitter (1991) and Kar et al. (2013a) in which a single update procedure is influenced by multiple potentials with different time-decaying weights. However, as opposed to the innovations term being a martingale difference sequence in the context of mixed time-scale stochastic approximation as proposed in Gelfand and Mitter (1991), the mixed time-scale stochastic approximation employed in this chapter does not have an innovation term which is a martingale difference sequence and hence is of sufficient technical interest. The addition of the residual Laplacian $\widetilde{L}(t)$ sequence in the update further complicates the update in the context of this chapter, by making the number of operating time-scales to be three instead of two as in Kar et al. (2013a) for which we had to develop new technical machinery.*

The above update can be written in a compact form as follows:

$$\begin{aligned}\mathbf{x}(t+1) &= \left(\mathbf{I}_{NM} - \mathbf{L}(t) \otimes \mathbf{I}_M\right)\mathbf{x}(t) \\ &+ \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1}\left(\mathbf{y}(t) - \mathbf{G}_H^\top \mathbf{x}(t)\right),\end{aligned} \tag{5.11}$$

where $\alpha_t = \frac{a}{t+1}$, $\mathbf{x}(t) = [\mathbf{x}_1^\top(t)\, \mathbf{x}_2^\top(t) \cdots \mathbf{x}_N^\top(t)]^\top$, $\mathbf{G}_H = diag[\mathbf{H}_1^\top, \mathbf{H}_2^\top, \cdots, \mathbf{H}_N^\top]$, $\mathbf{y}(t) = [\mathbf{y}_1^\top(t)\, \mathbf{y}_2^\top(t) \cdots \mathbf{y}_N^\top(t)]^\top$ and $\mathbf{\Sigma} = diag\left[\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_N\right]$.

Remark 5.4.2. *The Laplacian sequence that plays a key role in the analysis, takes the form $L(t) = \beta_t \overline{L} + \widetilde{L}(t)$, where $\widetilde{L}(t)$ the residual Laplacian sequence does not scale with β_t owing to the fact that the communication rate is chosen adaptively makes the analysis significantly different from Kar et al. (2013a). Thus, unlike Kar et al. (2013a), the Laplacian matrix sequence is not identically distributed; the sequence of effective Laplacians have a decaying mean, thus adding another time-scale in the already mixed time-scale dynamics which necessitates the development of new technical tools which lets us prove the order optimal convergence of the estimate sequence.*

We formalize an assumption on the innovation gain sequence $\{\alpha_t\}$ before proceeding further.

Assumption 5.4.3. *Let $\lambda_{min}(\cdot)$ denote the smallest eigenvalue. We require that a satisfies*[1]

$$a \min\{\lambda_{min}(\mathbf{\Gamma}), \lambda_{min}\left(\overline{\mathbf{L}} \otimes \mathbf{I}_M + \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right), \beta_0^{-1}\} \geq 1,$$

where $\otimes$ denotes the Kronecker product.

The communication cost per node for the proposed algorithm is given by $\mathcal{C}_t = \sum_{s=0}^{t-1} \zeta_s = \Theta\left(t^{1+(\epsilon-\tau_1)/2}\right)$, which in turn is strictly sub-linear as $\epsilon < \tau_1$.

[1] Note that, $\mathbf{\Gamma}$ and $\overline{\mathbf{L}} \otimes \mathbf{I}_M + \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top$ are positive definite matrices.

5.4.1 Consistency of $\mathcal{CREDO}$

Theorem 5.4.3. *Let assumptions 5.3.1-5.4.3 hold and let τ_1 in the consensus potential in (5.4) be such that $0 < \tau_1 \leq 1$. Consider the sequence $\{\mathbf{x}_n(t)\}$ generated by (5.10) at each worker n. Then, for each n, we have*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty} \mathbf{x}_n(t) = \boldsymbol{\theta}\right) = 1. \tag{5.12}$$

In particular, if τ_1 satisfies $0 < \tau_1 \leq 0.5 - (2+\epsilon_1)^{-1}$, we have that for all $\tau \in [0, 1/2)$,

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty} (t+1)^{\tau}\|\mathbf{x}_n(t) - \boldsymbol{\theta}\| = 0\right) = 1.$$

At this point, the estimate sequence generated by $\mathcal{CREDO}$ at any worker n is strongly consistent, i.e., $\mathbf{x}_n(t) \to \boldsymbol{\theta}$ almost surely (a.s.) as $t \to \infty$. Furthermore, the above characterization for $0 < \tau_1 \leq 0.5 - (2+\epsilon_1)^{-1}$ yields order-optimal convergence, i.e., from results in classical estimation theory, it is known that there exists no $\tau \geq 1/2$ such that a estimator $\{\boldsymbol{\theta}_c(t)\}$ satisfies $(t+1)^{\tau}\|\boldsymbol{\theta}_c(t) - \boldsymbol{\theta}\| \to 0$ a.s. as $t \to \infty$. The proof is relegated to Appendix D. We now state a main result which establishes the MSE communication rate for the proposed algorithm $\mathcal{CREDO}$.

Communication Efficiency of $\mathcal{CREDO}$

Theorem 5.4.4. *Let the hypothesis of Theorem 5.4.3 hold. Then, we have,*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2\right] = \Theta\left(\mathcal{C}_t^{-\frac{2}{\epsilon - \tau_1 + 2}}\right), \tag{5.13}$$

where $\epsilon < \tau_1$ and is as defined in (5.3).

The version of the $\mathcal{CREDO}$ algorithm, with $\beta_t = a(t+1)^{-1}$, achieves a communication cost of $\mathcal{C}_t = \Theta\left(t^{0.5(1+\epsilon)}\right)$. Hence, the MSE as a function of $\mathcal{C}_t$ in the case of $\tau_1 = 1$ is given by $\text{MSE} = \Theta(\mathcal{C}_t^{-2/(1+\epsilon)})$. However, it can be shown from standard arguments in stochastic approximation that updates with $\beta_t = a(t+1)^{-1-\delta}$ with $\delta > 0$, though results in a communication cost of $\mathcal{C}_t = \Theta(t^{0.5(1+\epsilon-\delta)})$, it does not generate estimate sequences which converge to $\boldsymbol{\theta}$. The proof is relegated to Appendix D.

With the above development in place, we state a result which allows us to benchmark the asymptotic efficiency of the proposed algorithm and the instantiations of it in terms of τ_1. To be specific, the next result establishes the asymptotic normality of the parameter estimate sequence $\{\mathbf{x}_n(t)\}$ and characterizes the asymptotic covariance of the estimate sequence, while the proof is relegated to Appendix D.

Theorem 5.4.5. *Let the hypotheses of Theorem 5.4.3 hold and in addition let $0 < \tau_1 \leq 0.5 - (2+\epsilon_1)^{-1}$. Then, we have,*

$$\sqrt{t+1}\,(\mathbf{x}_n(t) - \boldsymbol{\theta}) \overset{\mathcal{D}}{\Longrightarrow} \mathcal{N}\left(0, \frac{a\mathbf{I}}{2N} + \frac{\left(\boldsymbol{\Gamma} - \frac{\mathbf{I}}{2a}\right)^{-1}}{4N}\right), \tag{5.14}$$

where $\boldsymbol{\Gamma} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{H}_n^{\top}\boldsymbol{\Sigma}_n^{-1}\mathbf{H}_n$.

The asymptotic covariance as established in (5.14) is independent of the network. Technically speaking, as long as the averaged Laplacian $\overline{\mathbf{L}}$ is connected, and the consensus and the innovation potentials, i.e., β_t and α_t respectively are chosen appropriately, the asymptotic covariance is independent of the network

connectivity, i.e., it is independent of the network instantiations across all times and is just a function of the sensing model parameters and the noise covariance. It is to be noted that the optimal asymptotic covariance achieved by the oracle estimator is given by $N\mathbf{\Gamma}$. Such an asymptotic covariance can be achieved by a distributed setup where every worker node is aware of every other worker node's sensing model. To be particular, if a gain matrix $G = \sum_{n=1}^{N} N^{-1}\mathbf{H}_n^\top \mathbf{\Sigma}_n^{-1}\mathbf{H}_n$ is multiplied to the innovation term of the update in (5.10), the optimal asymptotic covariance is achievable (see, for example Kar et al. (2013a)). However, such an update would need global model information available at each worker node.

We now discuss the interesting trade-offs between the communication cost and the asymptotic covariance that follow from Theorem 5.4.5 and some existing results Kar et al. (2012, 2013a) (see Table 5.1). At this juncture, we consider the setup, where the τ_1 in the consensus potential β_t in (5.4) is taken to be $1/2 - (2+\epsilon_1)^{-1} \leq \tau_1 \leq 1$. We specifically consider the case where $\tau_1 = 1$. It has been established in prior work (see, for example Kar et al. (2012)) that in this case the asymptotic covariance depends on the network instantiation. To be specific, the averaged Laplacian $\overline{\mathbf{L}}$ which abstracts out the time-averaged information flow among the worker nodes has a key role in the asymptotic covariance in such a case. However, such a scheme, i.e., a single time scale variant of the proposed algorithm (in general for $1/2 - (2+\epsilon_1)^{-1} \leq \tau_1 \leq 1$) enjoys a lower communication rate. Technically speaking, for the case when $\tau_1 = 1$, the communication rate is given by $\mathcal{C}_t = \Theta\left(t^{0.5(1+\epsilon)}\right)$. Hence, there is an intrinsic trade-off between the communication rate and the achievable asymptotic variance.

Intuitively, the algorithm exhibits a threshold behavior in terms of the consensus potential τ_1. The threshold behavior is summarized in table 5.1. In the case when, $\tau_1 < 1/2 - \frac{1}{2+\epsilon_1}$, the algorithm achieves a network

Table 5.1: $\mathcal{CREDO}$: Trade-off between Communication cost and Asymptotic Covariance

Trade-Odd	**Convergence**	**Asymptotic Covariance**	**Comm. Cost.**
$0 < \tau_1 < \frac{1}{2} - \frac{1}{2+\epsilon_1}$	**Consistent**	**Network Independent**	$\Theta\left(t^{\frac{3}{4}+\frac{\epsilon}{2}}\right)$
$\frac{1}{2} - \frac{1}{2+\epsilon_1} \leq \tau_1 \leq 1$	**Consistent**	Network Dependent	$\Theta\left(t^{\frac{1+\epsilon}{2}}\right)$
$\tau_1 > 1$	Does not converge	Diverges	$\Theta(1)$

independent asymptotic covariance while ensuring the communication rate to be strictly sub linear. However, in the case when $1/2 - \frac{1}{2+\epsilon_1} \leq \tau_1 \leq 1$, the algorithm has a communication rate which is lower than the previous regime, but then achieves asymptotic covariance which depends on the network explicitly. Finally, in the case when $\tau_1 > 1$, the algorithm does not even converge to the true underlying parameter.

5.5 Directed $\mathcal{CREDO}$

Till now, with the communication scheme we have forced each instance of the communication graph to be undirected. Technically speaking, an agent can send or receive information when it is awake. We motivate a communication scheme next, where we allow each instance of the communication graph to be directed as long as they are undirected on average. Intuitively speaking, it encapsulates a setup where agents at each epoch do not receive information from their neighbors, but over time the number of messages sent and received from neighbors is roughly the same. We present directed $\mathcal{CREDO}$ next. For each node n, at every

time t, we introduce a binary random variable $\psi_{n,t}$, where

$$\psi_{n,t} = \begin{cases} \rho_t & \textit{with probability } \zeta_t \\ 0 & \textit{else}, \end{cases} \tag{5.15}$$

where $\psi_{i,t}$'s are independent both across time and the nodes, i.e., across t and n respectively. The random variable $\psi_{n,t}$ abstracts out the decision of the node n at time t whether to transmit its statistic to the neighborhood or not. We specifically take ρ_t and ζ_t of the form

$$\rho_t = \frac{\rho_0}{(t+1)^{\epsilon}}, \zeta_t = \frac{\zeta_0}{(t+1)^{(\tau_1-\epsilon)}}, \tag{5.16}$$

where $0 < \epsilon < \tau_1$ and $0 < \tau_1 \leq 1$. Furthermore, define β_t to be

$$\beta_t = \rho_t \zeta_t = \frac{\beta_0}{(t+1)^{\tau_1}}. \tag{5.17}$$

With the above development in place, we define the random time-varying Laplacian $\mathbf{L}(t)$, where $\mathbf{L}(t) \in \mathbb{R}^{N \times N}$ which abstracts the inter-node information exchange as follows:

$$\mathbf{L}_{i,j}(t) = \begin{cases} -\psi_{j,t} & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ \sum_{l \neq i} \psi_{l,t} & i = j. \end{cases} \tag{5.18}$$

We have, for $\{i,j\} \in E$:

$$\begin{aligned} \mathbb{E}\left[\mathbf{L}_{i,j}(t)\right] &= -\rho_t \zeta_t = -\beta_t = -\frac{\beta_0}{(t+1)^{\tau_1}} \\ \mathbb{E}\left[\mathbf{L}_{i,j}^2(t)\right] &= \rho_t^2 \zeta_t = \frac{\rho_0 \beta_0}{(t+1)^{\tau_1+\epsilon}}. \end{aligned} \tag{5.19}$$

Thus, we have that, the variance of $\mathbf{L}_{i,j}(t)$ is given by,

$$Var\left(\mathbf{L}_{i,j}(t)\right) = \frac{\beta_0 \rho_0}{(t+1)^{\tau_1+\epsilon}} - \frac{\beta_0^2}{(t+1)^{2\tau_1}}. \tag{5.20}$$

Define, the mean of the random time-varying Laplacian sequence $\{\mathbf{L}(t)\}$ as $\overline{\mathbf{L}}(t) = \mathbb{E}\left[\mathbf{L}(t)\right]$ and $\widetilde{\mathbf{L}}(t) = \mathbf{L}(t) - \overline{\mathbf{L}}(t)$. Note that, in spite of the sequence $\{\mathbf{L}(t)\}$ being a directed graph sequence, $\overline{\mathbf{L}}(t)$ is undirected. Note that, $\mathbb{E}\left[\widetilde{\mathbf{L}}(t)\right] = \mathbf{0}$, and

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{L}}(t)\right\|^2\right] \leq N^2 \mathbb{E}\left[\widetilde{\mathbf{L}}_{i,j}^2(t)\right] = \frac{N^2 \beta_0 \rho_0}{(t+1)^{\tau_1+\epsilon}} - \frac{N^2 \beta_0^2}{(t+1)^{2\tau_1}}, \tag{5.21}$$

where $\|\cdot\|$ denotes the L_2 norm. The above equation follows from equivalence of the L_2 and Frobenius norms. We also have that, $\overline{\mathbf{L}}(t) = \beta_t \overline{\mathbf{L}}$, where

$$\overline{\mathbf{L}}_{i,j} = \begin{cases} -1 & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l \neq i} L_{i,l} & i = j. \end{cases} \tag{5.22}$$

We formalize the assumption regarding the average connectivity and the undirectedness of the sequence of the communication graphs.

Assumption 5.5.1. *The inter-agent communication graph is connected and undirected on average, i.e., $\lambda_2(\overline{\mathbf{L}}) > 0$, which implies $\lambda_2(\overline{\mathbf{L}}(t)) > 0$, where $\overline{\mathbf{L}}(t)$ denotes the mean of the Laplacian matrix $\mathbf{L}(t)$ and $\lambda_2(\cdot)$ denotes the second smallest eigenvalue.*

With the communication protocol established, we propose an update, where every node n generates an estimate sequence $\{\mathbf{x}_n(t)\}$, where $\mathbf{x}_n(t) \in \mathbb{R}^M$ in the following way:

$$\begin{aligned} \mathbf{x}_n(t+1) = \mathbf{x}_n(t) - \underbrace{\sum_{l \in \Omega_n} \psi_{l,t} \left(\mathbf{x}_n(t) - \mathbf{x}_l(t)\right)}_{\text{Neighborhood Consensus}} \\ + \underbrace{\alpha_t \mathbf{H}_n^\top \mathbf{\Sigma}_n^{-1} \left(\mathbf{y}_n(t) - \mathbf{H}_n \mathbf{x}_n(t)\right)}_{\text{Local Innovation}}, \end{aligned} \tag{5.23}$$

where Ω_n denotes the neighborhood of node n with respect to the network encapsulated by $\overline{\mathbf{L}}$ and α_t is the innovation gain sequence which is given by $\alpha_t = a/(t+1)$.
The above update can be written in a compact form as follows:

$$\begin{aligned} \mathbf{x}(t+1) = \left(\mathbf{I}_{NM} - \mathbf{L}(t) \otimes \mathbf{I}_M\right) \mathbf{x}(t) \\ + \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \left(\mathbf{y}(t) - \mathbf{G}_H^\top \mathbf{x}(t)\right), \end{aligned} \tag{5.24}$$

where $\alpha_t = \frac{a}{t+1}$, $\mathbf{x}(t) = [\mathbf{x}_1^\top(t)\, \mathbf{x}_2^\top(t) \cdots \mathbf{x}_N^\top(t)]^\top$, $\mathbf{G}_H = diag[\mathbf{H}_1^\top, \mathbf{H}_2^\top, \cdots, \mathbf{H}_N^\top]$, $\mathbf{y}(t) = [\mathbf{y}_1^\top(t)\, \mathbf{y}_2^\top(t) \cdots \mathbf{y}_N^\top(t)]^\top$ and $\mathbf{\Sigma} = diag\left[\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_N\right]$.

Lemma 5.5.1. *For each n, the process $\{\mathbf{x}_n(t)\}$ satisfies*

$$\mathbb{P}_\theta \left(\sup_{t \geq 0} \|\mathbf{x}(t)\| < \infty \right) = 1. \tag{5.25}$$

Proof. We first note that,

$$\mathbf{L}(t) = \beta_t \overline{\mathbf{L}} + \widetilde{\mathbf{L}}(t), \tag{5.26}$$

where $\mathbb{E}\left[\widetilde{\mathbf{L}}(t)\right] = \mathbf{0}$ and $\mathbb{E}\left[\widetilde{\mathbf{L}}_{i,j}^2(t)\right] = \frac{c_4}{(t+1)^{\tau_1+\epsilon}} - \frac{c_3^2}{(t+1)^{2\tau_1}}$.
Define, $\mathbf{z}(t) = \mathbf{x}(t) - \mathbf{1}_N \otimes \boldsymbol{\theta}^*$ and $V(t) = \|\mathbf{z}(t)\|^2$. By conditional independence, we have that,

$$\begin{aligned} \mathbb{E}\left[V(t+1)|\mathcal{F}_t\right] = V(t) \\ + \mathbf{z}^\top(t) \left(\mathbf{I}_{NM} - \beta_t \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right) - \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right)^2 \mathbf{z}(t) \end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}_{\boldsymbol{\theta}^*}\left[\left\|\left(\widetilde{\mathbf{L}}(t)\otimes\mathbf{I}_M\right)\mathbf{z}(t)\right\|^2\right] \\
&+ \alpha^2(t)\mathbb{E}_{\boldsymbol{\theta}^*}\left[\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}(t)-\mathbf{G}_H^\top\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right\|^2\right] \\
&- 2\mathbf{z}^\top(t)\left(\beta_t\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)+\alpha_t\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)\mathbf{z}(t),
\end{aligned} \tag{5.27}
$$

where the filtration $\{\mathcal{F}_t\}$ may be taken to be the natural filtration generated by the random observations, the random Laplacians i.e.,

$$
\mathcal{F}_t = \sigma\left(\left\{\{\mathbf{y}_n(s)\}_{n=1}^N, \{\mathbf{L}(s)\}\right\}_{s=0}^{t-1}\right), \tag{5.28}
$$

which is the σ-algebra induced by the observation processes. For $t \geq t_1$, it can be shown that,

$$
\begin{aligned}
&\mathbf{z}^\top(t)\left(\mathbf{I}_{NM}-\beta_t\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)-\alpha_t\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)^2\mathbf{z}(t) \\
&\leq \left(1-c_4\alpha_t\right)^2\left\|\mathbf{z}(t)\right\|^2.
\end{aligned} \tag{5.29}
$$

Note, that $\widetilde{\mathbf{L}}(t)$ is row stochastic and hence $\left(\widetilde{\mathbf{L}}(t)\otimes\mathbf{I}_M\right)\mathbf{z}(t) = \mathbf{z}_{\mathcal{C}^\perp}(t)$.

$$
\mathbb{E}_{\boldsymbol{\theta}^*}\left[\left\|\left(\widetilde{\mathbf{L}}(t)\otimes\mathbf{I}_M\right)\mathbf{z}(t)\right\|^2\right] \leq \frac{c_5\left\|\mathbf{z}_{\mathcal{C}^\perp}(t)\right\|^2}{(t+1)^{\tau_1+\epsilon}} \tag{5.30}
$$

We use the following inequalities so as to analyze the recursion in (5.27).

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}^*}\left[\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}(t)-\mathbf{G}_H^\top\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right\|^2\right] \leq c_6 \\
&\mathbf{z}^\top(t)\left(\beta_t\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)+\alpha_t\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)\mathbf{z}(t) \\
&\geq \beta_t\lambda_2\left(\overline{\mathbf{L}}\right)\left\|\mathbf{z}_{\mathcal{C}^\perp}(t)\right\|^2 + c_7\alpha_t\left\|\mathbf{z}(t)\right\|^2.
\end{aligned} \tag{5.31}
$$

Using the inequalities derived in (5.31), we have,

$$
\begin{aligned}
&\mathbb{E}\left[V(t+1)|\mathcal{F}_t\right] \leq (1+c_8\alpha^2(t))V(t) \\
&- c_9\left(\beta_t - \frac{c_5}{(t+1)^{\tau_1+\epsilon}}\right)\left\|\mathbf{z}_{\mathcal{C}^\perp}(t)\right\|^2 + c_6\alpha^2(t).
\end{aligned} \tag{5.32}
$$

As $\frac{c_5}{(t+1)^{\tau_1+\epsilon}}$ goes to zero faster than β_t, $\exists t_2$ such that $\forall t \geq t_2$, $\beta_t \geq \frac{c_5}{(t+1)^{\tau_1+\epsilon}}$. By the above construction we obtain $\forall t \geq t_2$,

$$
\mathbb{E}_{\theta^*}[V(t+1)|\mathcal{F}_t] \leq (1+\alpha^2(t))V(t) + \widehat{\alpha}_t^2, \tag{5.33}
$$

where $\widehat{\alpha}(t) = \sqrt{c_6}\alpha_t$. The product $\prod_{s=t}^{\infty}(1+\alpha_s^2)$ exists for all t. Now let $\{W(t)\}$ be such that

$$
W(t) = \left(\prod_{s=t}^{\infty}(1+\alpha_s^2)\right)V_2(t) + \sum_{s=t}^{\infty}\widehat{\alpha}_s^2,\ \forall t \geq t_2. \tag{5.34}
$$

By (5.34), it can be shown that $\{W(t)\}$ satisfies,

$$
\mathbb{E}_{\theta^*}[W(t+1)|\mathcal{F}_t] \leq W(t). \tag{5.35}
$$

Hence, $\{W(t)\}$ is a non-negative super martingale and converges a.s. to a bounded random variable W^* as $t \to \infty$. It then follows from (C.163) that $V(t) \to W^*$ as $t \to \infty$. Thus, we conclude that the sequences $\{\mathbf{x}_n(t)\}$ are bounded for all n. □

Lemma 5.5.2. *Let the hypothesis of Lemma 5.5.1 hold. Then, we have,*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty} \mathbf{x}_n(t) = \boldsymbol{\theta}\right) = 1. \tag{5.36}$$

Proof of Lemma 5.5.2. Following as in the proof of Lemma 5.5.1, for t large enough

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)|\mathcal{F}_t] &\leq \left(1 - 2c_4\alpha_t + c_7\alpha_t^2\right) V(t) + c_6\alpha_t^2 \\ &\leq V(t) + c_6\alpha_t^2,\end{aligned} \tag{5.37}$$

as for t large enough, $-2c_4\alpha_t + c_7\alpha_t^2 < 0$. Now, consider the $\{\mathcal{F}_t\}$-adapted process $\{V_1(t)\}$ defined as follows

$$\begin{aligned}V_1(t) &= V(t) + c_6 \sum_{s=t}^{\infty} \alpha_s^2 \\ &= V(t) + c_8 \sum_{s=t}^{\infty} (t+1)^{-2},\end{aligned} \tag{5.38}$$

for appropriately chosen positive constant c_8.Since, $\{(t+1)^{-2}\}$ is summable, the process $\{V_1(t)\}$ is bounded from above. Moreover, it also follows that $\{V_1(t)\}_{t\geq t_1}$ is a supermartingale and hence converges a.s. to a finite random variable. By definition from (5.38), we also have that $\{V(t)\}$ converges to a non-negative finite random variable V^*. Finally, from (5.37), we have that,

$$\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)] \leq (1 - c_7\alpha_t)\, \mathbb{E}_{\boldsymbol{\theta}}[V(t)] + c_9(t+1)^{-2}, \tag{5.39}$$

for $t \geq t_1$. The sequence $\{V(t)\}$ then falls under the purview of Lemma C.3.1, and we have $\mathbb{E}_{\boldsymbol{\theta}}[V(t)] \to 0$ as $t \to \infty$. Finally, by Fatou's Lemma, where we use the non-negativity of the sequence $\{V(t)\}$, we conclude that

$$0 \leq \mathbb{E}_{\boldsymbol{\theta}}[V^*] \leq \liminf_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}}[V(t)] = 0, \tag{5.40}$$

which thus implies that $V^* = 0$ a.s. Hence, $\|\mathbf{z}(t)\| \to 0$ as $t \to \infty$ and the desired assertion follows. □

Theorem 5.5.3. *Let the hypothesis of Theorem 5.5.1 hold. Then, we have,*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2\right] = \Theta\left(\mathcal{C}_t^{-\frac{1}{\epsilon - \tau_1 + 1}}\right), \tag{5.41}$$

where $\epsilon < \tau_1$ and is as defined in (5.3).

It is to be noted that, in particular if $\tau_1 = 1$, we have that $\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2\right] = \Theta\left(\mathcal{C}_t^{-\frac{1}{\epsilon}}\right)$ for all $\epsilon > 0$.

Proof of Theorem 5.5.3. Proceeding as in proof of Lemma 5.5.2, we have, for t large enough

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)|\mathcal{F}_t] &\leq \left(1 - 2c_4\alpha_t + c_7\alpha_t^2\right) V(t) + c_6\alpha_t^2 \\ &\leq V(t) + c_6\alpha_t^2,\end{aligned} \tag{5.42}$$

as for t large enough, $-c_4\alpha_t + c_7\alpha_t^2 < 0$. Before proceeding further, we note that, from (5.29),

$$\begin{aligned}
&\mathbf{x}^\top \left(\beta_t \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right) + \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right) \mathbf{x} \\
&= \alpha_t \mathbf{x}^\top \left(\frac{\beta_t}{\alpha_t} \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right) + \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right) \mathbf{x} \\
&\geq \alpha_t \mathbf{x}^\top \left(\left(\mathbf{L} \otimes \mathbf{I}_M\right) + \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right) \mathbf{x} \geq c_4 \alpha_t,
\end{aligned} \tag{5.43}$$

where

$$c_4 = \lambda_{\min} \left(\left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right) + \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right). \tag{5.44}$$

Thus, we have that

$$\left\| \mathbf{I}_{NM} - \beta_t \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right) - \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top \right\| \leq 1 - c_4 \alpha_t, \tag{5.45}$$

for all $t \geq t_1$, where t_1 is chosen to be appropriately large. Now, consider the $\{\mathcal{F}_t\}$-adapted process $\{V_1(t)\}$ defined as follows

$$\begin{aligned}
V_1(t) &= V(t) + c_6 \sum_{s=t}^{\infty} \alpha_s^2 \\
&= V(t) + c_8 \sum_{s=t}^{\infty} (t+1)^{-2},
\end{aligned} \tag{5.46}$$

for appropriately chosen positive constant c_8.Since, $\{(t+1)^{-2}\}$ is summable, the process $\{V_1(t)\}$ is bounded from above. Moreover, it also follows that $\{V_1(t)\}_{t \geq t_1}$ is a supermartingale and hence converges a.s. to a finite random variable. By definition from (5.38), we also have that $\{V(t)\}$ converges to a non-negative finite random variable V^*. Finally, from (5.42), we have that,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)] \leq (1 - c_4\alpha_t)\, \mathbb{E}_{\boldsymbol{\theta}}[V(t)] + c_8 (t+1)^{-2} \\
&\Rightarrow E_{\boldsymbol{\theta}}[V(t+1)] \leq (1 - c_4\alpha_t)\, \mathbb{E}_{\boldsymbol{\theta}}[V(t)] + c_{10} \alpha_t (t+1)^{-1}
\end{aligned} \tag{5.47}$$

for $t \geq t_1$. The summability of $\{\alpha_t\}$ in conjunction with assumption **??** ensures that the sequence $\{V(t)\}$ then falls under the purview of Lemma C.1.3, and we have

$$\begin{aligned}
&\limsup_{t \to \infty} (t+1) \mathbb{E}_{\boldsymbol{\theta}}[V(t+1)] < \infty \\
&\Rightarrow \mathbb{E}_{\boldsymbol{\theta}}[V(t)] = O\left(\frac{1}{t}\right).
\end{aligned} \tag{5.48}$$

It is to be noted that the communication cost $\mathcal{C}_t$ for the proposed $\mathcal{CREDO}$ algorithm, is given by $\mathcal{C}_t = \Theta\left(t^{1+\epsilon-\tau_1}\right)$ and thus the assertion follows. □

5.6 Simulation Results

This section corroborates our theoretical findings through simulation examples and demonstrates on both synthetic and real data sets communication efficiency of $\mathcal{CREDO}$. Subsection 5.6.1 considers synthetic data, while Subsection 5.6.2 presents simulation results on real data sets.

5.6.1 Synthetic Data

Specifically, we compare the proposed communication-efficient distributed estimator, $\mathcal{CREDO}$, with the benchmark distributed recursive estimator which utilizes all inter-neighbor communications at all times, i.e., has a linear communication cost. The example demonstrates that the proposed communication-efficient estimator matches the MSE rate of the benchmark estimator. The simulation also shows that the proposed estimator improves the MSE *communication rate* with respect to the benchmark. The simulation setup is as follows. We consider three instances of undirected graphs with $N = 20$ nodes, with relative degrees[2] of nodes slated at 0.3736, 0.5157 and 0.6578. The graphs were generated as connected graph instances of the random geometric graph model with radius $r = \sqrt{\ln(N)/N}$. We set $M = 10$ and $M_n = 1$, for all $n = 1, ..., N$; i.e., the unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^{10}$, while each node makes a scalar observation at each time t. The noises $\gamma_n(t)$ are Gaussian and are i.i.d. both in time and across nodes and have the covariance matrix equal to $0.25 \times I$. The sampling matrices $\mathbf{H}_n$'s are chosen to be 2-sparse, i.e., every nodes observes a linear combination of two arbitrary entries of the vector parameter. The non-zero entries of the $\mathbf{H}_n$'s are sampled from a standard normal distribution. The sampling matrices $\mathbf{H}_n$'s at the same time satisfy Assumption 5.4.1. The parameters of the benchmark and the proposed estimator are as follows. The benchmark estimator's consensus weight is set to $0.1(t+1)^{-0.49}$. With the proposed estimator, we study the first two regimes as illustrated in Table 5.1,i.e., $0 < \tau_1 < \frac{1}{2}$ and $\frac{1}{2} \leq \tau_1 \leq 1$. For the second regime, we study two different cases. We set $\rho_t = 0.1(t+1)^{-0.01}$ for both the regimes. We set $\zeta_t = (t+1)^{-0.235}$, $\zeta_t = (t+1)^{-0.315}$ and $\zeta_t = (t+1)^{-0.49}$ for the above mentioned first and two cases of the second regime respectively; that is, with the proposed estimator, we set $\epsilon = 0.01$, $\tau_1 = 0.49$, $\epsilon = 0.01$, $\tau_1 = 0.65$ and $\epsilon = 0.01$, $\tau_1 = 1$ for the first and two cases of the second regime respectively. Note that the Laplacian matrix associated with the benchmark estimator and the expected Laplacian matrix associated with the proposed estimator, $\mathcal{CREDO}$ are equal in each of the three generated networks, i.e., $\overline{\mathbf{L}} = \mathbf{L}$. With all the three estimators, the innovation weight is set to $\alpha_t = (3.68(t+20))^{-1}$. Note that all the theoretical results hold unchanged for the "time-shifted" α_t used here. The purpose of the shift in the innovation potential is to avoid large innovation weights in the initial iterations. As a performance metric, we use the relative MSE estimate averaged across nodes:

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\|\mathbf{x}_n(t)-\boldsymbol{\theta}\|^2}{\|\mathbf{x}_n(0)-\boldsymbol{\theta}\|^2},$$

further averaged across 50 independent runs of the three estimators. Here, $\mathbf{x}_n(0)$ is node n's initial estimate. With both estimators, at each run, at all nodes, we set $\mathbf{x}_n(0) = 0$. Figure 5.1 plots the estimated relative MSE versus time t in log-log scale for the three networks. From figure 5.1, we can see that the MSE decay of the proposed estimator coincides with that of the benchmark estimator, especially in the $\tau_1 = 0.49$ regime across all the three networks, inspite of having lower communication costs. $\mathcal{CREDO}$ with $\tau_1 = 0.65$ and $\tau_1 = 1$, has higher convergence constants[3] with respect to the MSE decay rates as compared to the benchmark estimator, though with far lower communication costs. We can also see that, for network 1 and network 2, with relative degree slated at 0.3136 and 0.5157 respectively, the MSE in the case of $\tau_1 = 0.65$ and $\tau_1 = 1$ shifts further away from the MSE curve of network 3 and thus illustrating the network dependent convergence constant in the regime $1/2 \leq \tau_1 \leq 1$. At the same time, from Figure 5.1 it can be seen that with $\tau_1 = 0.49$, the convergence is practically independent of the network similar to the convergence of

[2]Relative degree is the ratio of the number of links in the graph to the number of possible links in the graph.

[3]It basically points to the fact that, though the MSE in all the cases have a t^{-1} scaling, but the variance of the $\tau_1 = 0.65$ and $\tau_1 = 1$ case involves bigger constants and thus larger variance.

the benchmark estimator, as predicted by Theorem 5.4.5. Figure 5.2 plots the estimated relative MSE versus average per-node communication cost C_t. We can see that the proposed scheme has an improved communication rate with respect to the benchmark, as predicted by the theory. In spite of higher convergence constants with respect to the MSE decay rates, in the case of $\tau_1 = 0.65$ and $\tau_1 = 1$, the MSE decay rate in terms of the communication cost is still faster than the benchmark estimator. Also, in the case of $\tau_1 = 0.49$, there is a close to $10\times$ reduction in the communication cost for the same achievable relative MSE of 0.005 as compared to the benchmark estimator. Figure 5.2, illustrates the trade-off between the MSE decay rate and the communication cost, there in, the lowest communication cost enjoyed by $\mathcal{CREDO}$ results in higher convergence constant with respect to the MSE decay, while the lowest convergence constant with respect to the MSE decay rate enjoyed by the benchmark estimator results in the highest communication cost.

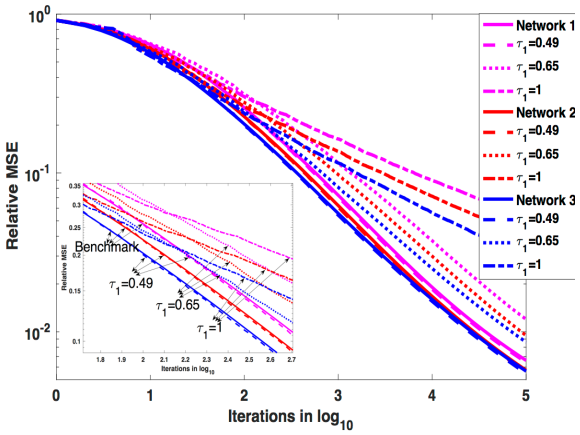


Figure 5.1: *Comparison of the proposed and benchmark estimators in terms of relative MSE: Number of Iterations.* The solid lines represent the benchmark, the three different colors indicate the three different networks, while the three regimes are represented by the dotted lines.

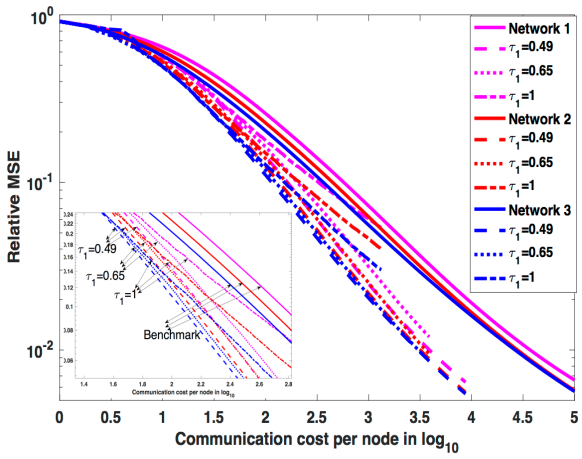
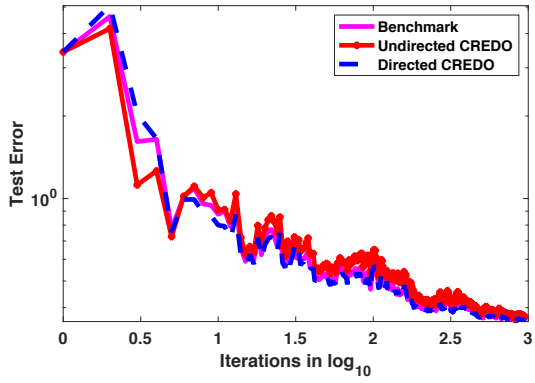


Figure 5.2: *Comparison of the proposed and benchmark estimators in terms of relative MSE: Communication cost per node.* The solid lines represent the benchmark, the three different colors indicate the three different networks, while the three regimes are represented by the dotted lines.

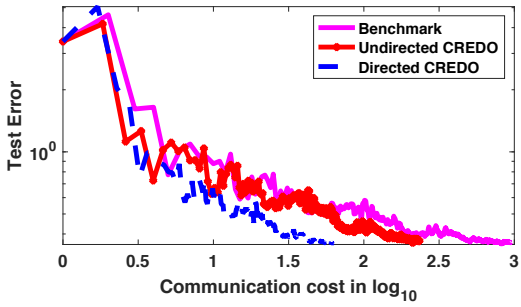
5.6.2 Real Datasets

In order to evaluate the performance of $\mathcal{CREDO}$, we ran experiments on three real-world datasets, namely cadata (Lib), Abalone (Lichman (2013)) and bank (Del).

For the cadata dataset (20640 data points, 8 features), we divided the samples into 20 equal parts of 900 data points each, after keeping 2640 data points as the test set. For the 20 node network, we constructed a random geometric graph. For the Abalone dataset (4177 data points, 8 features), we divided the samples into 10 equal parts of 360 points each, after keeping 577 data points as the test set. For the 10 node network, we constructed a random geometric graph. For the bank dataset (8192 data points, 9 features), we divided the samples into 20 equal parts equal parts of 350 points each, after keeping 1192 data points as the test set. For the 20 node network, we constructed a random geometric graph. We added Gaussian noise to the dependent variables, i.e., housing price, the age of Abalone and fraction of rejecting customers respectively. The training datasets, with respect to the sensing model (5.1), have dynamic regressors (a regressor here corresponds to a feature vector of one data point), i.e, time-varying $\mathbf{H}_n$'s for each agent n. Thus, we perform a pre-processing step where we average the training data points' regressors at each node to obtain an averaged $\overline{\mathbf{H}}_n$, which is then subsequently used at every iteration t in the update (5.10). For each experiment (each dataset), a consistency check is done by ensuring that $\sum_{n=1} \overline{\mathbf{H}}_n^\top \mathbf{\Sigma}_n^{-1} \overline{\mathbf{H}}_n$ is invertible and thus global observability holds. As the number of data points at each node are the same, we sample along iterations t data points at each node without replacement, and thus the total number of iterations t we run the algorithms equals the number of data points at each node. In other words, the algorithm passes through each data point exactly once. We summarize the comparison of the number of communications needed by directed and undirected $\mathcal{CREDO}$ and the benchmark algorithm at the test error obtained after the total number of iterations in Table 5.2. In particular, the test errors obtained in the cadata, abalone and the bank dataset are 0.015, 0.03 and 0.007 of the initial test error, respectively. In figures 5.3, 5.4 and 5.5, we plot the evolution of the test error for each of the datasets as a function of the number of iterations

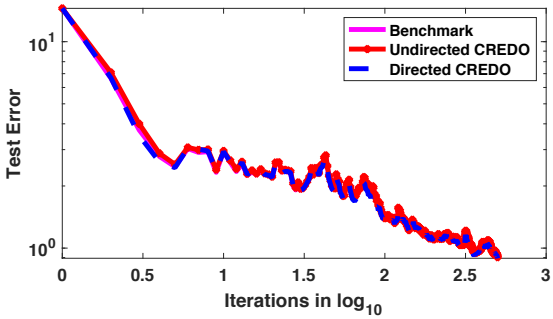


(a) Comparison of Test Error: Number of Iterations

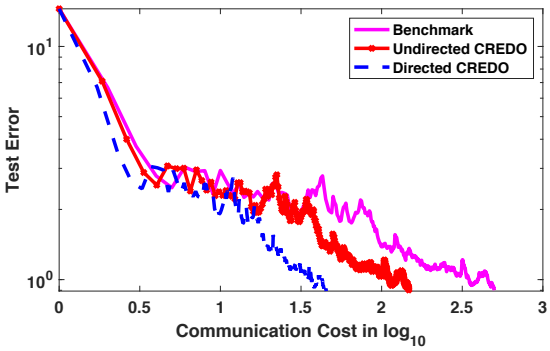


(b) Comparison of Test Error: Communication cost per node

Figure 5.3: CADATA Dataset: Comparison of the $\mathcal{CREDO}$ and benchmark estimators



(a) Comparison of Test Error: Number of Iterations

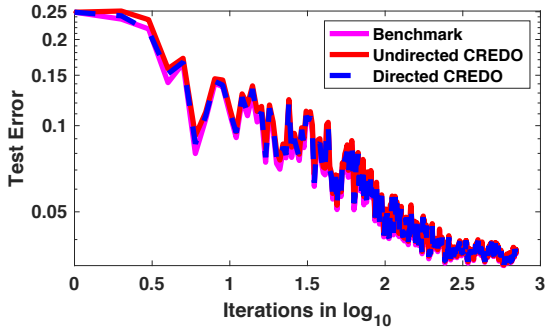


(b) Comparison of Test Error: Communication cost per node

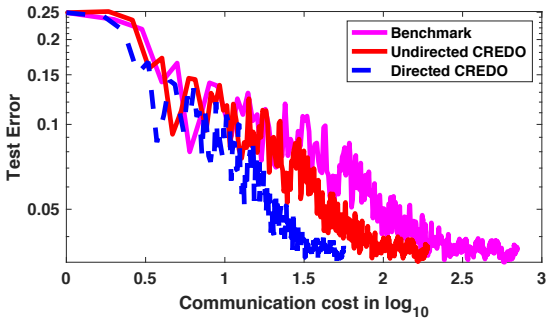
Figure 5.4: Abalone Dataset: Comparison of the $\mathcal{CREDO}$ and benchmark estimators

and the communication cost. It can be seen that while undirected $\mathcal{CREDO}$ matches the final test error of that of the benchmark algorithm, it requires on average thrice as less number of communications. Directed $\mathcal{CREDO}$ reduces the number of communications even further by reducing the number of communications by three times while maintaining the same test error.

Note that the theoretical setup rigorously establishes results pertaining to observation models with static regressors, i.e., static sensing matrices. However, the simulations on the real world datasets show that in spite of the time-varying regressors, the algorithm continues to demonstrate its improved communication efficiency over the benchmark. Moreover, as the sampling at each node is without replacement, the transients as far the performance is concerned can be improved by making the weight sequences decay after a few iterations instead of every iteration. Such a decay, while ensuring that the algorithm requirements are satisfied, would ensure for faster assimilation of new data points.



(a) Comparison of Test Error: Number of Iterations



(b) Comparison of Test Error: Communication cost per node

Figure 5.5: Bank Dataset: Comparison of the $\mathcal{CREDO}$ and benchmark estimators

Table 5.2: $\mathcal{CREDO}$: Communication cost across three datasets

Dataset	Test Error	Network size	Directed $\mathcal{CREDO}$	Undirected $\mathcal{CREDO}$	Benchmark
CADATA	2.15	20	**86**	261	722
ABALONE	0.95	10	**61**	192	558
BANK	0.015	20	**98**	252	696

5.7 Summary of Contributions

- **Improved MSE-Communication Rate Tradeoff:** We show that, despite significantly lower communication cost, $\mathcal{CREDO}$ achieves the best possible $O(1/t)$ rate of MSE decay in time t (t also equals to per-worker number of data samples). Importantly, this result translates into significant improvements in the rate at which MSE decays with communication cost $\mathcal{C}_t$ – namely from $O(1/\mathcal{C}_t)$ with existing methods to $O(1/\mathcal{C}_t^{2-\zeta})$ with the proposed method, where $\zeta > 0$ arbitrarily small. With the directed $\mathcal{CREDO}$ development, we demonstrate that the MSE communication cost dependence can be further improved to $O(1/\mathcal{C}_t^{-1/\zeta})$. In particular, the algorithm $\mathcal{CREDO}$ points to the fact that several communications in typical distributed estimation setup are redundant and there is no loss in the convergence rate when such redundant communications are dropped.
- **Three time-scale stochastic approximation:** To achieve the results above, we developed several technical innovations. Specifically, the studied setup requires analysis of *mixed time-scale stochastic approximation* algorithms with *three different time scales*. This setup stands in contrast with the classical single time-scale stochastic approximation, the properties of which are well known. It is also very different from the more commonly studied two time-scale stochastic approximation (see, for instance Borkar (2008)) in which a fast process is coupled with a slower dynamical system. We develop here new technical tools that allow us to handle the case of number of operating time-scales to be three instead of two as in Kar et al. (2013a) for mixed time-scale stochastic approximation.

5.8 Conclusion and Future Directions

In this chapter, we have proposed a communication efficient distributed recursive estimation schemes $\mathcal{CREDO}$, for which we established strong consistency of the estimation sequence and characterized the asymptotic co-

variance of the estimate sequence in terms of the sensing model and the noise covariance. The communication efficiency of $\mathcal{CREDO}$ was characterized in terms of the dependence of the MSE on the communication cost. To be specific, we have established that the MSE of $\mathcal{CREDO}$ can be as good as $\Theta\left(\mathcal{C}_t^{-2+\zeta}\right)$, where $\zeta > 0$ and ζ is arbitrarily small. Future research directions include the development of communication schemes, which are adaptive in terms of the connectivity of a node, and local decision making in terms of whether to communicate or not based on neighborhood information. The algorithm presented in this chapter can be thought of as a distributed method to solve a stochastic optimization problem with a stochastic least squares-type cost function. A natural direction is hence to extend the proposed ideas to stochastic distributed optimization.

Chapter 6

Distributed Weighted Non-linear Least Squares: $\mathcal{CIWNLS}$

6.1 Introduction

The chapter focuses on distributed nonlinear least squares estimation in distributed information settings. Each agent in the network senses sequentially over time independent and identically distributed (i.i.d) time-series that are (nonlinear) functions of the underlying vector parameter of interest corrupted by noise. To be specific, we are interested in the design of recursive estimation algorithms to estimate a vector parameter of interest that are consistent and order-optimal in the sense of pathwise convergence rate and such that their asymptotic error covariances are comparable with that of the centralized weighted nonlinear least squares estimator[1]. The estimation algorithms we design are recursive – they process the agents' observations at all times as and when they are sensed, rather than batch processing. This contrasts with centralized setups, where a fusion center has access to all the observations across different agents at all times, i.e., the inter-agent communication topology is all-to-all or all-to-one. Centralized estimators are burdened by high communication overheads, synchronization issues, and high energy requirements. Moreover, there is the requirement of global model information, i.e., the fusion center requiring information about the local models of all agents. All these make centralized estimation algorithms difficult to implement in multi-agent distributed setups of the type considered in this chapter, motivating us to revisit the problem of distributed *sequential* parameter estimation. To accommodate energy constraints in many practical networked and wireless settings, the inter-agent collaboration is limited to a pre-assigned possibly sparse communication graph. Moreover, due to limited computation and storage capabilities of individual agents in a typical multi-agent networked setting, we restrict to scenarios where individual agents are only aware of their local model information; hence, we allow for heterogeneity among agents, with different agents possibly having different local sensing models and noise statistics. This chapter proposes a distributed recursive algorithm, namely, the $\mathcal{CIWNLS}$ (*Consensus* + *innovations* Weighted Nonlinear Least Squares), which is of the *consensus* + *innovations* form Kar et al. (2012). We specifically focus on a setting in which the agents make i.i.d observations sequentially over time, only possess local model information, and update their parameter estimates by simultaneous assimilation of the information obtained from their neighboring agents (*consensus*) and current locally sensed information (*innovation*). This justifies the name $\mathcal{CIWNLS}$,

[1] A centralized estimator has access to all agent data at all times and has sufficient computing ability to implement the classical weighted nonlinear least squares estimator Jennrich (1969); Wu (1981) at all times.

which is a distributed weighted nonlinear least squares (WNLS) type algorithm of the *consensus+innovations* form. To replicate practical sensing environments accurately, we model the underlying vector parameter as a static parameter, that takes values in a parameter set $\Theta \subseteq \mathbb{R}^M$ (possibly a strict subset of $\mathbb{R}^M$). The dimension M is possibly large, but the observation of any agent n is M_n dimensional with typically $M_n \ll M$ in most applications; this renders the parameter locally unobservable at each agent. The key assumptions concerning the sensing functions in this chapter are required to hold only on the parameter set Θ and not on the entire space[2] $\mathbb{R}^M$. The distributed sequential estimation approach of the *consensus + innovations* form that we present accomplishes the following:

Consistency under global observability: We assume *global observability*[3] and certain *monotonicity* properties of the multi-agent sensing model, as well as the connectedness of the inter-agent communication graph. We show that our recursive distributed estimator generates parameter estimate sequences that are strongly consistent at each agent. Global observability is a minimal requirement for consistency; in fact, it is necessary for consistency of centralized estimators as well.

Optimal pathwise convergence rate[4]: We show that the proposed distributed estimation algorithm $\mathcal{CIWNLS}$ yields order-optimal pathwise convergence rate under certain smoothness conditions on the sensing model. These conditions are standard in the recursive estimation literature and we require them to hold only on the parameter set Θ. Even though recursive, our distributed estimation approach guarantees that the parameter estimates are feasible at all times, i.e., they belong to the parameter set Θ. Further, the parameter estimates at each local agent n are as good as the optimal centralized estimator as far as pathwise convergence rate is concerned. The key point to note here is that, for the above order optimality to hold we need to only assume that the inter-agent communication graph is connected irrespective of how sparse the link realizations are.

Asymptotic Normality: Under standard smoothness conditions on the sensing model, the proposed distributed estimation algorithm $\mathcal{CIWNLS}$ is shown to yield asymptotically normal[5] parameter estimate sequences. Distributed estimation does pay a price. The asymptotic covariance of the proposed distributed estimator is not as efficient as that of the centralized estimator; nonetheless, it shows the benefits of inter-agent collaboration. In absence of inter-agent collaboration, the parameter of interest most likely is unobservable at each individual agent, and hence non-collaborative or purely decentralized procedures will lead to divergence under the usual asymptotic normality scaling at the individual network agents.

6.2 Related Work

Distributed inference approaches addressing problems related to distributed estimation, parallel computing, and optimization in multi-agent environments through interacting stochastic gradient and stochastic approximation algorithms have been developed extensively in the literature –see, for example, early work Tsitsiklis (1984); Tsitsiklis et al. (1986); Bertsekas et al. (1984); Kushner and Yin (1987). Existing distributed estima-

[2] By taking the parameter set $\Theta = \mathbb{R}^M$, the unconstrained parameter estimation problem can be addressed, and thus the setup in this chapter enables a richer class of formulations.

[3] Global observability corresponds to the centralized setting, where an estimator has access to the observations of all sensors at all times. The assumption of global observability does not mean that each sensor is observable; rather, if there was a centralized estimator with simultaneous access to all the sensor measurements, this centralized estimator would be able to reasonably estimate the underlying parameter. A more precise definition is provided later in Assumption 6.3.2.

[4] By *optimal pathwise convergence rate*, we mean the pathwise convergence rate of the centralized estimator to the true underlying parameter with noisy observations.

[5] An estimate sequence is asymptotically normal if its $\sqrt{t}$ scaled error process, i.e., the difference between the sequence and the true parameter converges in distribution to a normal random variable, where t refers to (discrete) time or equivalently the number of sampling epochs.

tion schemes in the literature can be broadly divided into three classes. The first class includes architectures that are characterized by the presence of a fusion center (see, for example Aysal and Barner (2008); Luo (2005)) that receives the estimates or local measurements or their quantized versions from the network agents and performs estimation. The second class involves single snapshot data collection (see, for example Das and Mesbahi (2006); Schizas et al. (2008c)) followed by distributed consensus or optimization protocols to fuse the initial estimates. In contrast to these classes, the third class involves agents making observations sequentially over time and where inter-agent communication, limited to arbitrary pre-assigned possibly sparse topologies, occurs at the same rate as sensing (see, for example Braca et al. (2010, 2008); Kar et al. (2012); Lopes and Sayed (2008)). Two representative schemes from the third class are *consensus+innovations* type Kar et al. (2012); Kar and Moura (2008a) and diffusion type algorithms Lopes and Sayed (2008); Chen and Sayed (2012); Matta et al. (2016); Cattivelli and Sayed (2011b, 2010). Broadly speaking, these algorithms simultaneously assimilate a single round of neighborhood information, consensus like in Bertsekas et al. (1984); Olfati-Saber et al. (2007); Dimakis et al. (2010); Jadbabaie et al. (2003), with the locally sensed latest information, the local innovation; see for example *consensus+innovation* approaches for nonlinear distributed estimation Kar et al. (2012, 2013b) and detection Bajovic et al. (2011); Jakovetic et al. (2012); Sahu and Kar (2017). A key difference between the diffusion algorithms discussed above and the *consensus+innovations* algorithms presented in this chapter is the nature of the innovation gains (the new information fusion weights). In the diffusion framework, the innovation gains are taken to be constant, whereas, in the *consensus+innovations* schemes these are made to decay over time in a controlled fashion. The constant innovation gains in the diffusion approaches facilitate adaptation in dynamic parameter environments, but at the same time lead to non-zero residual estimation error at the agents (see, for example, Lopes and Sayed (2008)), whereas the time-varying innovation weights in the *consensus+innovations* approach ensure consistent parameter estimates at the agents. In Kar and Moura (2014), strong consistency of the parameter estimate sequence is established, and it is shown that the proposed algorithm is asymptotically efficient, i.e., its asymptotic covariance is the same as that of the optimal centralized estimator. However, in Kar and Moura (2014), the smoothness assumptions on the sensing functions need to hold on the entire parameter space, i.e., $\mathbb{R}^M$. In contrast, we consider here a setup where the parameter belongs to a constrained set and the smoothness conditions on the sensing functions need to hold only on the constrained parameter set; this allows the algorithm proposed in this chapter, namely $\mathcal{CIWNLS}$, to be applicable to other types of application scenarios. Moreover, in Kar and Moura (2014) the problem setup needs more detailed knowledge of the statistics of the noise processes involved, as it aims to obtain asymptotically efficient (in that the agent estimates are asymptotically normal with covariance equal to the inverse of the associated Fisher information rate) estimates for general statistical exponential families. In particular, to achieve asymptotic efficiency, Kar and Moura (2014) develops a *consensus+innovations* type distributed recursive variant of the maximum likelihood estimator (MLE) that requires knowledge of the detailed observation statistics. In contrast, in this chapter, our setup only needs knowledge of the noise covariances and the sensing functions. Technically speaking, for additive noisy observation models, the weighted nonlinear squares estimation, the distributed version of which is proposed in this chapter, applies to fairly generic estimation scenarios, i.e., where observation noise statistics are unknown.

6.3 Sensing Model and Preliminaries

Let $\boldsymbol{\theta} \in \Theta$ (to be specified shortly) be an M-dimensional (vector) parameter that is to be estimated by a network of N agents. We specifically consider a discrete time system. Each agent n at time t makes a noisy

observation $\mathbf{y}_n(t)$ that is a noisy function (nonlinear) of the parameter. Formally, the observation model for the n-th agent is given by

$$\mathbf{y}_n(t) = \mathbf{f}_n(\boldsymbol{\theta}) + \zeta_n(t), \tag{6.1}$$

where $\mathbf{f}_n(\cdot)$ is, in general, a non-linear function, $\{\mathbf{y}_n(t)\}$ is a $\mathbb{R}^{M_n}$-valued observation sequence for the n-th agent and for each n, $\{\zeta_n(t)\}$ is a zero-mean temporally independent and identically distributed (i.i.d.) noise sequence with nonsingular covariance matrix $\mathbf{R}_n$, such that, $\zeta_n(t)$ is $\mathcal{F}_{t+1}$-adapted and independent of $\mathcal{F}_t$. In typical application scenarios, the observation at each agent is low-dimensional, i.e., $M_n \ll M$, and usually a function of only a subset of the M components of $\boldsymbol{\theta}$, i.e., agent n observes a function of K_n components of $\boldsymbol{\theta}$ with $K_n \ll M$, which most likely entails that the parameter of interest $\boldsymbol{\theta}$ is locally unobservable at the individual agents. Hence, to achieve a reasonable estimate of the parameter $\boldsymbol{\theta}$, it is necessary for the agents to collaborate through inter-agent message passing schemes.

Since, the sources of randomness in our formulation are the observations $\mathbf{y}_n(t)$'s by the agents, the filtration $\{\mathcal{F}_t\}$ may be taken to be the natural filtration generated by the random observations, i.e.,

$$\mathcal{F}_t = \sigma\left(\left\{\{\mathbf{y}_n(s)\}_{n=1}^{N}\right\}_{s=0}^{t-1}\right), \tag{6.2}$$

which is the σ-algebra induced by the observation processes.

To motivate our distributed estimation approach (presented in Section 6.4) and benchmark its performance with respect to the optimal centralized estimator, we now review some concepts from centralized estimation theory.

Centralized weighted nonlinear least-squares (WNLS) estimation. Consider a network of agents where a hypothetical fusion center has access to the observations made by all the agents at all times and then conducts the estimation scheme. In such scenarios, one of the most widely used estimation approaches is the weighted nonlinear least squares (WNLS) (see, for example, Jennrich (1969)). The WNLS is applicable to fairly generic estimation scenarios, for instance, even when the observation noise statistics are unknown, which precludes other classical estimation approaches such as the maximum likelihood estimation. We discuss below some useful theoretical properties of WNLS like, in case the observation noise is Gaussian, it coincides with the (asymptotically) efficient maximum likelihood estimator. To formalize, for each t, define the cost function

$$\mathcal{Q}_t(\mathbf{z}) = \sum_{s=0}^{t}\sum_{n=1}^{N} \left(\mathbf{y}_n(s) - \mathbf{f}_n(\mathbf{z})\right)^{\top} \mathbf{R}_n^{-1} \left(\mathbf{y}_n(s) - \mathbf{f}_n(\mathbf{z})\right), \tag{6.3}$$

where $\mathbf{R}_n$ denotes the positive definite covariance of the measurement noise $\zeta_n(t)$. The WNLS estimate $\widehat{\boldsymbol{\theta}}_t$ of $\boldsymbol{\theta}$ at each time t is obtained by minimizing the cost functional $\mathcal{Q}_t(\cdot)$,

$$\widehat{\boldsymbol{\theta}}_t \in \operatorname*{argmin}_{\mathbf{z}\in\Theta} \mathcal{Q}_t(\mathbf{z}). \tag{6.4}$$

Under rather weak assumptions on the sensing model (stated below), the existence and asymptotic behavior of WNLS estimates have been analyzed in the literature.

Assumption 6.3.1. *The set Θ is a closed convex subset of $\mathbb{R}^M$ with non-empty interior $\operatorname{int}(\Theta)$ and the true (but unknown) parameter $\boldsymbol{\theta} \in \operatorname{int}(\Theta)$.*

Assumption 6.3.2. *The sensing model is globally observable, i.e., any pair $\boldsymbol{\theta}, \acute{\boldsymbol{\theta}}$ of possible parameter instances in Θ satisfies*

$$\sum_{n=1}^{N} \left\| \mathbf{f}_n(\boldsymbol{\theta}) - \mathbf{f}_n(\acute{\boldsymbol{\theta}}) \right\|^2 = 0 \tag{6.5}$$

if and only if $\boldsymbol{\theta} = \acute{\boldsymbol{\theta}}$.

Assumption 6.3.3. *The sensing function $\mathbf{f}_n(.)$ for each n is continuously differentiable in the interior $\text{int}(\Theta)$ of the set Θ. For each $\boldsymbol{\theta}$ in the set Θ, the matrix $\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$ that is given by*

$$\boldsymbol{\Gamma}_{\boldsymbol{\theta}} = \frac{1}{N} \sum_{n=1}^{N} \nabla \mathbf{f}_n(\boldsymbol{\theta}) \mathbf{R}_n^{-1} \nabla \mathbf{f}_n^{\top}(\boldsymbol{\theta}), \tag{6.6}$$

where $\nabla \mathbf{f}$ denotes the gradient of $\mathbf{f}(\cdot)$, is invertible.

Smoothness conditions on the sensing functions, such as the one imposed by assumption 6.3.3 is common in the literature addressing statistical inference algorithms in non-linear settings. Note that the matrix $\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$ is well defined at the true value of the parameter $\boldsymbol{\theta}$ as $\boldsymbol{\theta} \in \text{int}(\Theta)$ and the continuous differentiability of the sensing functions hold for all $\boldsymbol{\theta} \in \text{int}(\Theta)$.

Assumption 6.3.4. *There exists $\epsilon_1 > 0$, such that, for all n, $\mathbb{E}_{\boldsymbol{\theta}} \left[\|\zeta_n(t)\|^{2+\epsilon_1} \right] < \infty$.*

The following classical result characterizes the asymptotic properties of (centralized) WNLS estimators.

Proposition 6.3.1. (Jennrich (1969)) *Let the parameter set Θ be compact and the sensing function $f_n(\cdot)$ be continuous on Θ for each n. Then, a WNLS estimator of $\boldsymbol{\theta}$ exists, i.e., there exists an $\{\mathcal{F}_t\}$-adapted process $\{\widehat{\boldsymbol{\theta}}_t\}$ such that*

$$\widehat{\boldsymbol{\theta}}_t \in argmin_{\mathbf{z} \in \Theta} \mathcal{Q}_t(\mathbf{z}), \ \forall t. \tag{6.7}$$

Moreover, if the model is globally observable, i.e., Assumption 6.3.2 holds, the WNLS estimate sequence $\{\widehat{\boldsymbol{\theta}}_t\}$ is consistent, i.e.,

$$\mathbb{P}_{\boldsymbol{\theta}} \left(\lim_{t \to \infty} \widehat{\boldsymbol{\theta}}_t = \boldsymbol{\theta} \right) = 1. \tag{6.8}$$

Additionally, if Assumption 6.3.3 holds, the parameter estimate sequence is asymptotically normal, i.e.,

$$\sqrt{t+1} \left(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta} \right) \overset{D}{\Longrightarrow} \mathcal{N}(0, \boldsymbol{\Sigma}_c), \tag{6.9}$$

where

$$\boldsymbol{\Sigma}_c = (N \boldsymbol{\Gamma}_{\boldsymbol{\theta}})^{-1}, \tag{6.10}$$

$\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$ is as given by (6.6) and $\overset{\mathcal{D}}{\Longrightarrow}$ refers to convergence in distribution (weak convergence).

The WNLS estimator, apart from needing a fusion center that has access to the observations across all agents at all times, also incorporates a batch data processing as implemented in (6.4). To mitigate the enormous communication overhead incurred in (6.4), much work in the literature has focused on the development

of sequential albeit centralized estimators that process the observations $\mathbf{y}(t)$ across agents in a recursive manner. Under additional smoothness assumptions on the local observation functions $\mathbf{f}_n(\cdot)$'s, recursive centralized estimators of the stochastic approximation type have been developed by several authors, see, for example, Sakrison (1965); Has'minskij (1974); Pfanzagl (1973); Stone (1975); Fabian (1978). Such centralized estimators require a fusion center that process the observed data in batch mode or recursive form. The fusion center receives the entire set of agents' observations, $\{\mathbf{y}_n(t)\}$, $n = 1, 2, 3, \cdots, N$, at all times t. Moreover, both in the batch and the recursive processing form, the fusion center needs global model information in the form of the local observation functions $\mathbf{f}_n(\cdot)$'s and the observation noise statistics, i.e., the noise covariances $\mathbf{R}_n$'s across all agents. In contrast, this chapter develops collaborative distributed estimators of $\boldsymbol{\theta}$ at each agent n of the network, where each agent n has access to its local sensed data $\mathbf{y}_n(t)$ only and local model information, i.e., its own local sensing function $\mathbf{f}_n(\cdot)$ and noise covariance $\mathbf{R}_n$. To mitigate the communication overhead, we present distributed message passing schemes in which agents, instead of forwarding raw data to a fusion center, participate in a collaborative iterative process to estimate the underlying parameter $\boldsymbol{\theta}$. The agents also maintain a copy of their local parameter estimate that is updated by simultaneously processing local parameter estimates from their neighbors and the latest sensed information. To obtain a good parameter estimate with such localized communication, we propose a distributed estimator that incorporates neighborhood information mixing and local data processing simultaneously (at the same rate). Such estimators are referred to as *consensus + innovations* estimators, see Kar et al. (2012), for example.

Example: Distributed Static Phase Estimation in Smart Grids

Many applications within cyber physical systems and internet of things can be modeled as non-linear distributed estimation problems of type (1). Such class of models arises, e.g., with state estimation in power systems; therein, a phasorial representation of voltages and currents is usually utilized, wherein non-linearity in general emerges from power-flow equations. Here, we focus on the specific problem within the class, namely distributed static phase estimation in smart grids. We describe the model briefly and refer to, e.g., Ilic' and Zaborszky (2000); Kar et al. (2012) for more details. Here, graph G corresponds to a power grid network of $n = 1, ..., N$ generators and loads (here a single generator or a single load is a node in the graph), while the edge set E corresponds to the set of transmission lines or interconnections. (For simplicity, even though not necessary, we assume that the physical interconnection network matches the inter-node communication network.) Assume that G is connected. The state of a node n is described by $(\mathcal{V}_n, \phi_n)$, where $\mathcal{V}_n$ is the voltage magnitude and ϕ_n is the phase angle. As commonly assumed, e.g., Ilic' and Zaborszky (2000), we let the voltages $\mathcal{V}_n$ be known constants; on the other hand, angles ϕ_n are unknown ant are to be estimated. Following a standard approximation path, the real power flow across the transmission line between nodes n and l can be expressed as, e.g., Ilic' and Zaborszky (2000):

$$\mathcal{P}_{nl}(\phi) = \mathcal{V}_n \, \mathcal{V}_l \, b_{nl} \, \sin(\phi_{nl}), \tag{6.11}$$

where ϕ is the vector that collects the unknown phase angles ϕ_n across all nodes, b_{nl} is line (n, l)'s admittance, and $\phi_{nl} = \phi_n - \phi_l$. Denote by $E_m \subset E$ the set of lines equipped with power flow measuring devices. The power flow measurement at line (n, l) is then given by:

$$y_{nl}(t) = \mathcal{P}_{nl}(\phi) + \gamma_{nl}(t) = \mathcal{V}_n \, \mathcal{V}_l \, b_{nl} \sin(\phi_{nl}) + \gamma_{nl}(t), \tag{6.12}$$

where $\{\gamma_{nl}(t)\}$ is the zero mean i.i.d. measurement noise with finite moment $\mathbb{E}[|\gamma_{nl}(t)|^{2+\epsilon_1}]$, for some $\epsilon_1 > 0$. Assume that each measurement $y_{nl}(t)$ is assigned to one of its incident nodes n or l. Further,

let Ω'_n denote the set of all indexes l such that measurements $y_{nl}(t)$ are available at node n. Then, it becomes clear that the angle estimation problem is a special case of model (1), with the measurement vectors $\mathbf{y}_n(t) = [y_{nl}(t),\ l \in \Omega'_n]^\top,\ n = 1, ..., N$, noise vectors $\gamma_n(t) = [\gamma_{nl}(t),\ l \in \Omega'_n]^\top,\ n = 1, ..., N$, and sensing functions $\mathbf{f}_n(\boldsymbol{\phi}) = [\mathcal{V}_n\, \mathcal{V}_l\, b_{nl}\, \sin(\phi_{nl}),\ l \in \Omega'_n]^\top.\ n = 1, ..., N$. It can be shown that under reasonable assumptions on noise angle ranges (that correspond to the admissible parameter set $\boldsymbol{\Theta}$) and the smart grid network and admittances structure, the assumptions we make on the sensing model are satisfied,[6] and hence $\mathcal{CREDO} - \mathcal{NL}$ can be effectively applied; we refer to Ilic' and Zaborszky (2000); Kar et al. (2012) for details.

6.4 A Distributed Estimator : $\mathcal{CIWNLS}$

We state formally assumptions pertaining to the inter-agent communication and additional smoothness conditions on the sensing functions required in the distributed setting.

Assumption 6.4.1. *The inter-agent communication graph is connected, i.e., $\lambda_2(\mathbf{L}) > 0$, where $\mathbf{L}$ denotes the associated graph Laplacian matrix.*

Assumption 6.4.2. *For each n, the sensing function $\mathbf{f}_n(\cdot)$ is Lipschitz continuous on Θ, i.e., for each agent n, there exists a constant $k_n > 0$ such that*

$$\|\mathbf{f}_n(\boldsymbol{\theta}) - \mathbf{f}_n(\boldsymbol{\theta}^*)\| \leq k_n \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|, \tag{6.13}$$

for all $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta$.

Distributed algorithm. In the proposed implementation, each agent n updates at each time t its estimate sequence $\{\mathbf{x}_n(t)\}$ and an auxiliary sequence $\{\widehat{\mathbf{x}}_n(t)\}$ using a two-step collaborative procedure; specifically, 1) $\widehat{\mathbf{x}}_n(t)$ is updated by a *consensus+innovations* rule and, subsequently, 2) a local projection to the feasible parameter set Θ updates $\mathbf{x}_n(t)$. Formally, the overall update rule at an agent n corresponds to

$$\begin{aligned}\widehat{\mathbf{x}}_n(t+1) = \mathbf{x}_n(t) &- \underbrace{\beta_t \sum_{l \in \Omega_n} (\mathbf{x}_n(t) - \mathbf{x}_l(t))}_{\text{neighborhood consensus}} \\ &- \underbrace{\alpha_t \left(\nabla \mathbf{f}_n(\mathbf{x}_n(t))\right) \mathbf{R}_n^{-1} \left(\mathbf{f}_n(\mathbf{x}_n(t)) - \mathbf{y}_n(t)\right)}_{\text{local innovation}}\end{aligned} \tag{6.14}$$

and

$$\mathbf{x}_n(t+1) = \mathcal{P}_\Theta[\widehat{\mathbf{x}}_n(t+1)], \tag{6.15}$$

where : Ω_n is the communication neighborhood of agent n (determined by the Laplacian $\mathbf{L}$); $\nabla f_n(\cdot)$ is the gradient of $\mathbf{f}_n$, which is a matrix of dimension $\mathbf{M} \times \mathbf{M}_n$, with the (i,j)-th entry given by $\frac{\partial [\mathbf{f}_n(\mathbf{x}_n(t))]_j}{\partial [\mathbf{x}_n(t)]_i}$; $\mathcal{P}_\Theta[\cdot]$ the projection operator corresponding to projecting[7] on Θ; and $\{\beta_t\}$ and $\{\alpha_t\}$ are consensus and innovation

[6] To see this, note that the dependence of the measurements on the state is through sinusoidal functions, which are everywhere differentiable and thus the gradient of $f(\cdot)$ within the domain Θ exists everywhere. Moreover, as the derivatives of $\sin(\cdot)$ and $\cos(\cdot)$ are bounded, the norm of gradient is bounded in the example considered to motivate the formulation. Finally, regarding assumption 6.3.2, it can be shown that the assumption is satisfied if: (1) graph G is connected; (2) the set of admissible phase angle values, i.e., the parameter constraint set $\boldsymbol{\Theta}$, is chosen appropriately; (3) the real power flow between nodes n and l is non-zero if and only if there exists a physical transmission line connecting the nodes; and (4) voltage magnitude $\mathcal{V}_n \neq 0$, for all nodes n. Please see Proposition 27 in Kar et al. (2012).

[7] The projection on Θ is unique under assumption 6.3.1.

weight sequences given by

$$\beta_t = \frac{b}{(t+1)^{\delta_1}}, \alpha_t = \frac{a}{t+1}, \tag{6.16}$$

where $a, b > 0, 0 < \delta_1 < 1/2 - 1/(2+\epsilon_1)$ and ϵ_1 was defined in Assumption 6.3.4.
The update in (6.14) can be written in a compact manner as follows:

$$\begin{aligned}\widehat{\mathbf{x}}(t+1) &= \mathbf{x}(t) - \beta_t \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{x}(t) \\ &+ \alpha_t \mathbf{G}(\mathbf{x}(t)) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{x}(t)\right)\right),\end{aligned} \tag{6.17}$$

where: $\mathbf{x}(t)^\top = [\mathbf{x}_1(t)^\top \cdots \mathbf{x}_N(t)^\top]$; $\widehat{\mathbf{x}}(t)^\top = [\widehat{\mathbf{x}}_1(t)^\top \cdots \widehat{\mathbf{x}}_N(t)^\top]$, $\mathbf{f}(\mathbf{x}(t)) = \left[\mathbf{f}_1(\mathbf{x}_1(t))^\top \cdots \mathbf{f}_N(\mathbf{x}_N(t))^\top\right]^\top$; $\mathbf{R}^{-1} = \text{diag}\left[\mathbf{R}_1^{-1}, \cdots, \mathbf{R}_N^{-1}\right]$; and $\mathbf{G}\left(\mathbf{x}(t)\right) = \text{diag}\left[\nabla \mathbf{f}_1\left(\mathbf{x}_1(t)\right), \cdots, \nabla \mathbf{f}_N\left(\mathbf{x}_N(t)\right)\right]$. We refer to the parameter estimate update in (6.15) and the projection in (6.16) as the $\mathcal{CIWNLS}$ (*Consensus+ innovations* Weighted Nonlinear Least Squares) algorithm.

Remark 6.4.1. *The parameter update is recursive and distributed in nature and hence is an* online *algorithm. Moreover, the projection step in* (6.15) *ensures that the parameter estimate sequence* $\{\mathbf{x}_n(t)\}$ *is feasible and belongs to the parameter set* Θ *at all times* t.

Methods for analyzing the convergence of distributed stochastic algorithms of the form (6.14)-(6.17) and variants were developed in Kar et al. (2012); Kar and Moura (2011); Kar et al. (2013a); Kar and Moura (2014). The key is to obtain conditions that ensure the existence of appropriate stochastic Lyapunov functions. To enable this, we propose a condition on the sensing functions (standard in the literature of general recursive procedures) that guarantees the existence of such Lyapunov functions and, hence, the convergence of the distributed estimation procedure.

Assumption 6.4.3. *The following aggregate strict monotonicity condition holds: there exists a constant* $c_1 > 0$ *such that for each pair* $\boldsymbol{\theta}, \acute{\boldsymbol{\theta}}$ *in* Θ *we have that*

$$\sum_{n=1}^{N} \left(\boldsymbol{\theta} - \acute{\boldsymbol{\theta}}\right)^\top \left(\nabla f_n(\boldsymbol{\theta})\right) \mathbf{R}_n^{-1} \left(f_n(\boldsymbol{\theta}) - f_n(\acute{\boldsymbol{\theta}})\right) \geq c_1 \left\|\boldsymbol{\theta} - \acute{\boldsymbol{\theta}}\right\|^2. \tag{6.18}$$

We assume that the noise covariances are known apriori. However, in scenarios where the noise covariances are not known apriori, in order to verify Assumption 6.4.3, only the gradient $\nabla \mathbf{f}_n(\cdot)$ needs to be computed. In case of unknown noise distribution, i.e., unknown noise covariance, the first few observations can be used to estimate the noise covariance so as to get a reasonable estimate of the inverse noise covariance. The estimated noise covariance can then be used to verify the assumption.

Remark 6.4.2. *We comment on the Assumption 6.3.1-6.4.3. Assumptions 6.3.1-6.3.4 are classical with respect to the WNLS convergence. Assumption 6.4.2 specifies some smoothness conditions of the non-linear sensing functions. The smoothness conditions aid in establishing the consistency of the recursive* $\mathcal{CIWNLS}$ *algorithm. The classical WNLS is usually posed in a non-recursive manner, while the distributed algorithm we propose is recursive and hence, to ensure convergence we need Lyapunov type conditions, which in turn is specified by Assumption 6.4.3. Moreover, Assumptions 6.4.2-6.4.3 are only sufficient conditions. The key assumptions to establish our main results, Assumptions 6.3.1, 6.3.2, 6.4.2, and 6.4.3 are required to hold only in the parameter set* Θ *and need not hold globally in the entire space* $\mathbb{R}^M$. *This allows our approach*

to apply to very general nonlinear sensing functions. For example, for functions of the trigonometric type (see Section 2.8 for an illustration), properties such as the strict monotonicity condition in 6.4.3 hold in the fundamental period, but not globally. As another specific instance, if the $\mathbf{f}_n(\cdot)$'s are linear[8]*, condition (6.5) in Assumption 6.3.2, reduces to $\sum_{n=1}^{N} \mathbf{F}_n^\top \mathbf{R}_n^{-1} \mathbf{F}_n$ being full rank (and hence positive definite). The monotonicity condition in Assumption 6.4.3 in this context coincides with Assumption 6.3.2, i.e., it is trivially satisfied by the positive definiteness of the matrix $\sum_{n=1}^{N} \mathbf{F}_n^\top \mathbf{R}_n^{-1} \mathbf{F}_n$. Asymptotically efficient distributed parameter estimation schemes for the general linear model have been developed in Kar and Moura (2011); Kar et al. (2013a).*

6.5 Main Results: $\mathcal{CIWNLS}$

This section states the main results, while the proofs are relegated to Appendix E. The first concerns the consistency of the estimate sequence in the $\mathcal{CIWNLS}$ algorithm.

Theorem 6.5.1. *Let assumptions 6.3.1-6.3.2 and 6.3.4-6.4.3 hold. Furthermore, assume that the constant a in (6.16) satisfies*

$$ac_1 \geq 1, \tag{6.19}$$

where c_1 is defined in Assumption 6.4.3. Consider the sequence $\{\mathbf{x}_n(t)\}$ generated by (6.15)-(6.16) at each agent n. Then, for each n, we have

$$\mathbb{P}_{\boldsymbol{\theta}} \left(\lim_{t\to\infty} (t+1)^\tau \|\mathbf{x}_n(t) - \boldsymbol{\theta}\| = 0 \right) = 1 \tag{6.20}$$

for all $\tau \in [0, 1/2)$. In particular, the estimate sequence generated by the distributed algorithm (6.14)-(6.17) at any agent n is consistent, i.e., $\mathbf{x}_n(t) \to \boldsymbol{\theta}$ a.s. as $t \to \infty$.

At this point, we note that the convergence in Theorem 6.5.1 is order-optimal, in that standard arguments in (centralized) estimation theory show that in general there exists no $\tau \geq 1/2$ such that a centralized WNLS estimator $\{\widehat{\boldsymbol{\theta}}_t\}$ satisfies $(t+1)^\tau \|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\| \to 0$ a.s. as $t \to \infty$.
The next result establishes the asymptotic normality of the parameter estimate sequence $\{\mathbf{x}_n(t)\}$ and characterizes the asymptotic covariance of the proposed $\mathcal{CIWNLS}$ estimator. This can be benchmarked with the asymptotic covariance of the centralized WNLS estimator.

Theorem 6.5.2. *Let the assumptions 6.3.1-6.4.3 hold. Assume that in addition to assumption 6.3.1, the parameter set Θ is a bounded set. Furthermore, let a defined in (6.16) satisfy*

$$a > \max\left\{ \frac{1}{c_1}, \frac{1}{2\inf_{\boldsymbol{\theta}\in\Theta} \boldsymbol{\Lambda}_{\boldsymbol{\theta},min}} \right\}, \tag{6.21}$$

where c_1 is defined in Assumption 6.4.3, $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{\theta},min}$ denote respectively the diagonal matrix of eigenvalues and the minimum eigenvalue of $\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$, with $\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$ defined in (6.6). Then, for each n, the parameter estimate sequence at agent n, $\{\mathbf{x}_n(t)\}$, under $\mathbb{P}_{\boldsymbol{\theta}}$ satisfies the following asymptotic normality condition,

$$\sqrt{t+1}\,(\mathbf{x}_n(t) - \boldsymbol{\theta}) \xRightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{\Sigma}_d), \tag{6.22}$$

[8] To be specific, $\mathbf{f}_n(\boldsymbol{\theta})$ is then given by $\mathbf{F}_n\boldsymbol{\theta}$, where $\mathbf{F}_n$ is the sensing matrix with dimensions $M_n \times M$.

where

$$\mathbf{\Sigma}_d = \frac{\mathbf{aI}}{2N} + \frac{\left(N\mathbf{\Gamma}_{\boldsymbol{\theta}} - \frac{N\mathbf{I}}{2a}\right)^{-1}}{4}, \tag{6.23}$$

and $\xRightarrow{\mathcal{D}}$ *refers to convergence in distribution (weak convergence).*

As the parameter set Θ is a bounded set in Theorem 6.5.2, in addition to being a closed set as in Assumption 6.3.1, we have that Θ is compact and hence $\inf_{\boldsymbol{\theta}\in\Theta} \mathbf{\Lambda}_{\boldsymbol{\theta},\min} = \min_{\boldsymbol{\theta}\in\Theta} \mathbf{\Lambda}_{\boldsymbol{\theta},\min}$, i.e., the infimum is attained. Moreover, from Assumption 6.3.3, we have that the matrix $\mathbf{\Gamma}_{\boldsymbol{\theta}}$ is invertible $\forall \boldsymbol{\theta} \in \Theta$ and hence $\inf_{\boldsymbol{\theta}\in\Theta} \mathbf{\Lambda}_{\boldsymbol{\theta},\min} > 0$. Further, under the assumption that $a > \frac{1}{2\inf_{\boldsymbol{\theta}\in\Theta} \mathbf{\Lambda}_{\boldsymbol{\theta},\min}}$, the difference of the asymptotic covariance of the distributed estimator and that of the centralized estimator, i.e., the matrix $\frac{a\mathbf{I}}{2N} + \frac{\left(\mathbf{\Gamma}_{\boldsymbol{\theta}} - \frac{\mathbf{I}}{2a}\right)^{-1}}{4N} - \left(N\mathbf{\Gamma}_{\boldsymbol{\theta}}\right)^{-1}$, is positive semidefinite. The above claim can be established by comparing the i-th eigenvalue of the asymptotic covariance of the distributed estimator with that of the centralized estimator, as both the covariance matrices are simultaneously diagonalizable, i.e., have the same set of eigenvectors. To be specific,

$$\begin{aligned} & a^2\mathbf{\Lambda}^2_{\boldsymbol{\theta},ii} - 2a\mathbf{\Lambda}_{\boldsymbol{\theta},ii} + 1 \geq 0 \\ \Rightarrow & \frac{1}{\mathbf{\Lambda}_{\boldsymbol{\theta},ii}} \leq \frac{a^2\mathbf{\Lambda}_{\boldsymbol{\theta},ii}}{2a\mathbf{\Lambda}_{\boldsymbol{\theta},ii} - 1} \\ \Rightarrow & \frac{1}{N\mathbf{\Lambda}_{\boldsymbol{\theta},ii}} \leq \frac{a^2\mathbf{\Lambda}_{\boldsymbol{\theta},ii}}{2aN\mathbf{\Lambda}_{\boldsymbol{\theta},ii} - N}, \end{aligned} \tag{6.24}$$

which holds for all $i = 1, 2, \cdots, N$.

We now benchmark the asymptotic covariance of the proposed estimator $\mathcal{CIWNLS}$ with that of the optimal centralized estimator. From Assumption 6.4.2, we have for all $\boldsymbol{\theta} \in \Theta$

$$\|\mathbf{\Gamma}_{\boldsymbol{\theta}}\| \leq \max_{n=1,\cdots,N} k_n^2 \left\|\mathbf{R}_n^{-1}\right\| = k^*_{\max}, \tag{6.25}$$

where k_n is defined in Assumption 6.4.2. Moreover, from the hypotheses of Theorem 6.5.2 we have that $\mathbf{\Lambda}_{\boldsymbol{\theta},\min} > \frac{1}{2a}$, for all $\boldsymbol{\theta} \in \Theta$. Thus, we have the following characterization of the eigenvalues for the matrix $\mathbf{\Gamma}_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta} \in \Theta$,

$$\frac{1}{2a} < \mathbf{\Lambda}_{\boldsymbol{\theta},ii} \leq k^*_{\max}, \tag{6.26}$$

for all i. The difference of the i-th eigenvalue of the asymptotic covariance of the distributed estimator and the centralized estimator $\mathbf{\Lambda}_{d,i}$, is given by $\mathbf{\Lambda}_{d,i} = \frac{(a\mathbf{\Lambda}_{\boldsymbol{\theta},ii}-1)^2}{N\mathbf{\Lambda}_{\boldsymbol{\theta},ii}(2a\mathbf{\Lambda}_{\boldsymbol{\theta},ii}-1)}$. Now, we consider two cases. Specifically, if the condition

$$k^*_{\max} > \max\left\{\frac{1}{c_1}, \frac{1}{2\inf_{\boldsymbol{\theta}\in\Theta} \mathbf{\Lambda}_{\boldsymbol{\theta},\min}}\right\}, \tag{6.27}$$

is satisfied, then a can be chosen to be $a < 1/k^*_{\min}$, and then we have,

$$\frac{1}{2a} < \mathbf{\Lambda}_{\boldsymbol{\theta},ii} \leq \frac{1}{a}. \tag{6.28}$$

It is to be noted that the function $h(x) = \frac{(ax-1)^2}{Nx(2ax-1)}$ is non-increasing in the interval $\left(\frac{1}{2a}, \frac{1}{a}\right)$. Hence, we

have that

$$\|\mathbf{\Sigma}_d - \mathbf{\Sigma}_c\| = \frac{\left(a\mathbf{\Lambda}_{\boldsymbol{\theta},\min} - 1\right)^2}{N\mathbf{\Lambda}_{\boldsymbol{\theta},\min}\left(2a\mathbf{\Lambda}_{\boldsymbol{\theta},\min} - 1\right)}, \tag{6.29}$$

where,

$$\max\left\{\frac{1}{c_1}, \frac{1}{2\inf_{\boldsymbol{\theta}\in\Theta}\mathbf{\Lambda}_{\boldsymbol{\theta},\min}}\right\} < a < \frac{1}{k^*_{\max}}. \tag{6.30}$$

In the case, when the condition in (6.27) is violated, we have that,

$$\begin{aligned}
&\|\mathbf{\Sigma}_d - \mathbf{\Sigma}_c\| \\
&= \max\left\{\frac{\left(a\mathbf{\Lambda}_{\boldsymbol{\theta},\min} - 1\right)^2}{N\mathbf{\Lambda}_{\boldsymbol{\theta},\min}\left(2a\mathbf{\Lambda}_{\boldsymbol{\theta},\min} - 1\right)}, \frac{\left(a\mathbf{\Lambda}_{\boldsymbol{\theta},\max} - 1\right)^2}{N\mathbf{\Lambda}_{\boldsymbol{\theta},\max}\left(2a\mathbf{\Lambda}_{\boldsymbol{\theta},\max} - 1\right)}\right\} \\
&\leq \max\left\{\frac{\left(a\mathbf{\Lambda}_{\boldsymbol{\theta},\min} - 1\right)^2}{N\mathbf{\Lambda}_{\boldsymbol{\theta},\min}\left(2a\mathbf{\Lambda}_{\boldsymbol{\theta},\min} - 1\right)}, \frac{\left(ak^*_{\max} - 1\right)^2}{Nk^*_{\max}\left(2ak^*_{\max} - 1\right)}\right\},
\end{aligned} \tag{6.31}$$

where $\mathbf{\Lambda}_{\boldsymbol{\theta},\max}$ denotes the largest eigenvalue of $\mathbf{\Gamma}_{\boldsymbol{\theta}}$. Note that the proposition in (6.31) is equivalent to (6.29), when the condition in (6.27) is satisfied. Hence, for all feasible choices of a, which are in turn given by (6.21), the characterization in (6.31) holds.
The above mentioned findings can be precisely stated in the form of the following corollary:

Corollary 6.5.3. *Let the hypothesis of Theorem 6.5.2 hold. Then, we have,*

$$\begin{aligned}
&\|\mathbf{\Sigma}_d - \mathbf{\Sigma}_c\| \\
&\leq \max\left\{\frac{\left(a\mathbf{\Lambda}_{\boldsymbol{\theta},min} - 1\right)^2}{N\mathbf{\Lambda}_{\boldsymbol{\theta},min}\left(2a\mathbf{\Lambda}_{\boldsymbol{\theta},min} - 1\right)}, \frac{\left(ak^*_{max} - 1\right)^2}{Nk^*_{max}\left(2ak^*_{max} - 1\right)}\right\},
\end{aligned} \tag{6.32}$$

where $\mathbf{\Sigma}_d$ and $\mathbf{\Sigma}_c$ are defined in (6.23) and (6.10), respectively.

Furthermore, as noted above that the difference of the asymptotic covariance of the distributed estimator and that of the centralized estimator is positive semi-definite. (This is intuitively expected, as a distributed procedure may not outperform a well-designed centralized procedure.) The inefficiency of the distributed estimator with respect to the centralized WNLS estimator, as far as the asymptotic covariance is concerned, is due to the use of suboptimal innovation gains (see, for example (6.14)) used in the parameter estimate update. An optimal innovation gain sequence would require the knowledge of the global model information, i.e., the sensing functions and the noise covariances across all the agents. (See Remark 4.3 below for a detailed discussion.) Though the distributed estimation scheme is suboptimal with respect to the centralized estimator as far as the asymptotic covariance is concerned, its performance is significantly better than the non-collaborative case, i.e., in which agents perform estimation in a fully decentralized or isolated manner. In particular, since the agent sensing models are likely to be locally unobservable for $\boldsymbol{\theta}$, the asymptotic covariances in the non-collaborative scenario may diverge due to non-degeneracy.

Remark 6.5.4. *In this context, we briefly review the methodology adopted in Kar and Moura (2014) to achieve asymptotically efficient distributed estimators for general standard statistical exponential families. The asymptotic efficiency of the estimator proposed in Kar and Moura (2014) is a result of a certainty-*

equivalence type distributed optimal gain sequence generated through an auxiliary consistent parameter estimate sequence (see, for example Section III.A. in Kar and Moura (2014)). The generation of the auxiliary estimate sequence comes at the cost of more communication and computation complexity as two other parallel recursions run in addition to the parameter estimate recursion. To be specific, the adaptive gain refinement, which is the key for achieving asymptotic efficiency, involves communication of the gain matrices that belong to the space $\mathbb{R}^{M\times M}$. Moreover, in a setting like the one considered in this chapter, where the parameter belongs to a closed convex subset $\Theta \in \mathbb{R}^M$, the parameter estimate at all times provided by the algorithm in Kar and Moura (2014) might not be feasible. In contrast, the communication and computation complexity in $\mathcal{CIWNLS}$ is significantly lower than that of the algorithm proposed in Kar and Moura (2014). The price paid by $\mathcal{CIWNLS}$ is the lower asymptotic performance as measured in terms of the asymptotic covariance as we discuss next.

We compare the computational and communication overhead of $\mathcal{CIWNLS}$ and of the algorithm in Kar and Moura (2014). For simplicity, we consider a d-regular communication graph with every agent connected to d other agents. We compute the computation and communication overhead agent wise. In one sampling epoch, an agent in $\mathcal{CIWNLS}$ communicates M-dimensional parameter estimates to its neighbors, i.e., the communication overhead is Md. In the algorithm proposed in Kar and Moura (2014) (see, (8)-(11) in Section $III.A$), an agent not only communicates its auxiliary and optimal parameter estimates but also its gain matrix, to its neighbors, with the communication overhead $2Md + M^2d$. With respect to the computational overhead, in every sampling epoch the number of computations in $\mathcal{CIWNLS}$ at agent n is given by $O\left(M_nM + M(d+1)\right)$. The maximum computational overhead across all agents is thus given by $\max_{n=1,\cdots,N} O\left(M_nM + M(d+1)\right)$. In comparison, the number of computations at any agent in the algorithm proposed in Kar and Moura (2014) is given by $O\left(M^3 + 2M^2(d+1) + 2M(d+1)\right)$. Thus, the communication and computational complexity of the proposed algorithm is much lower than that of the algorithm proposed in Kar and Moura (2014), at the cost of suboptimal asymptotic estimation error covariances.

6.6 Communication Efficient $\mathcal{CIWNLS}$: $\mathcal{CREDO}-\mathcal{NL}$

Following along the lines of the sparsifying communication protocol proposed in Chapter 4, we propose a communication efficient version of the algorithm $\mathcal{CIWNLS}$. In particular, following the notation as in 4, we propose an update, where every node n generates an estimate sequence $\{\mathbf{x}_n(t)\}$, where $\mathbf{x}_n(t) \in \mathbb{R}^M$ in the following way:

$$\widehat{\mathbf{x}}_n(t+1) = \mathbf{x}_n(t) \underbrace{- \beta_t \sum_{l\in\Omega_n} \psi_{n,t}\psi_{l,t}\left(\mathbf{x}_n(t) - \mathbf{x}_l(t)\right)}_{\text{neighborhood consensus}}$$
$$\underbrace{- \alpha_t \left(\nabla \mathbf{f}_n(\mathbf{x}_n(t))\right) \mathbf{R}_n^{-1}\left(\mathbf{f}_n(\mathbf{x}_n(t)) - \mathbf{y}_n(t)\right)}_{\text{local innovation}} \tag{6.33}$$

and

$$\mathbf{x}_n(t+1) = \mathcal{P}_\Theta[\widehat{\mathbf{x}}_n(t+1)], \tag{6.34}$$

where Ω_n denotes the neighborhood of node n with respect to the network represented by $\overline{\mathbf{L}}$, α_t is the innovation gain sequence which is given by $\alpha_t = \alpha_0/(t+1)$, $\alpha_0 > 0$, and $\mathcal{P}_\Theta[\cdot]$ the projection operator corresponding to projecting on Θ. The random variable $\psi_{n,t}$ determines the activation state of a node n. By activation we mean, if $\psi_{n,t} \neq 0$ then node n can send and receive information in its neighborhood at

time t. However, when $\psi_{n,t} = 0$, node n neither transmits nor receives information. The link between node n and node l gets assigned a weight of ρ_t^2 if and only if $\psi_{n,t} \neq 0$ and $\psi_{l,t} \neq 0$.
The update in (6.33) can be written in a compact manner as follows:

$$\begin{aligned}\widehat{\mathbf{x}}(t+1) &= \mathbf{x}(t) - (\mathbf{L}(t) \otimes \mathbf{I}_M)\, \mathbf{x}(t) \\ &+ \alpha_t \mathbf{G}(\mathbf{x}(t)) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{x}(t)\right)\right). \end{aligned} \tag{6.35}$$

Here, $\otimes$ denotes the Kronecker product, and:

$$\begin{aligned}\mathbf{x}(t)^\top &= [\mathbf{x}_1(t)^\top \cdots \mathbf{x}_N(t)^\top] \\ \widehat{\mathbf{x}}(t)^\top &= [\widehat{\mathbf{x}}_1(t)^\top \cdots \widehat{\mathbf{x}}_N(t)^\top] \\ \mathbf{f}(\mathbf{x}(t)) &= \left[\mathbf{f}_1(\mathbf{x}_1(t))^\top \cdots \mathbf{f}_N(\mathbf{x}_N(t))^\top\right]^\top \\ \mathbf{R}^{-1} &= \operatorname{diag}\left[\mathbf{R}_1^{-1}, \cdots, \mathbf{R}_N^{-1}\right] \\ \mathbf{G}\left(\mathbf{x}(t)\right) &= \operatorname{diag}\left[\nabla \mathbf{f}_1\left(\mathbf{x}_1(t)\right), \cdots, \nabla \mathbf{f}_N\left(\mathbf{x}_N(t)\right)\right]. \end{aligned}$$

We refer to the parameter estimate update in (6.33) and the projection in (6.34) in conjunction with the randomized communication protocol as the $\mathcal{CREDO} - \mathcal{NL}$ algorithm.

6.7 Main Results: $\mathcal{CREDO} - \mathcal{NL}$

In this section, we present the main results of the proposed algorithm $\mathcal{CREDO} - \mathcal{NL}$, while the proofs are relegated to Appendix E.

Theorem 6.7.1. *Let assumptions 6.3.1-6.3.2 and 6.3.4-6.4.3 hold. Consider the sequence $\{\mathbf{x}_n(t)\}$ generated by algorithm (6.33) at each agent n, with the parameters set to $\rho_t = \frac{\rho_0}{(t+1)^{\epsilon/2}}$, $\zeta_t = \frac{\zeta_0}{(t+1)^{(1/2-\epsilon/2)}}$, and $\alpha_t = \alpha_0/(t+1)$, where $\rho_0, \zeta_0, \alpha_0$ are arbitrary positive numbers. Then, for each n, we have*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty} \mathbf{x}_n(t) = \boldsymbol{\theta}\right) = 1. \tag{6.36}$$

Theorem 6.7.1 verifies that the estimate sequence generated by $\mathcal{CREDO} - \mathcal{NL}$ at any agent n is strongly consistent, i.e., $\mathbf{x}_n(t) \to \boldsymbol{\theta}$ almost surely (a.s.) as $t \to \infty$.
We now state a main result which establishes the MSE communication rate for the proposed algorithm $\mathcal{CREDO} - \mathcal{NL}$.

Theorem 6.7.2. *Let the hypothesis of Theorem 6.7.1 hold. Then, we have,*

$$\mathbb{E}\left[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2\right] = \Theta\left(\frac{1}{t}\right). \tag{6.37}$$

Furthermore, we have:

$$\mathbb{E}\left[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2\right] = \Theta\left(\mathcal{C}_t^{-\frac{2}{\epsilon+1}}\right), \tag{6.38}$$

where $0 < \epsilon < 1$ and is as defined in (5.3).

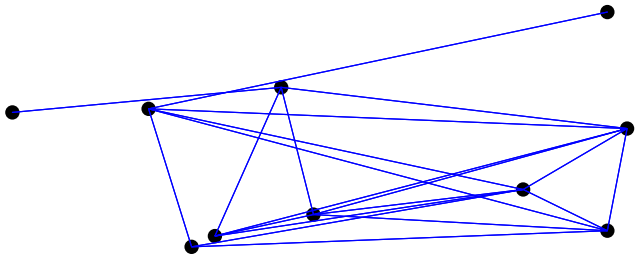


Figure 6.1: $\mathcal{CIWNLS}$: Network Deployment of 10 agents

Remark 6.7.3. *Observe that $\mathcal{CREDO}-\mathcal{NL}$ algorithm, with $\beta_t = \beta_0\,(t+1)^{-1}$ has communication cost of $\mathcal{C}_t = \Theta\left(t^{0.5(1+\epsilon)}\right)$. From this, we can see that MSE as a function of $\mathcal{C}_t$ is given by MSE $= \Theta(\mathcal{C}_t^{-2/(1+\epsilon)})$. Of course, with β_t that decays faster than $1/t$, communication cost reduces further. However, it can be shown that in this case the algorithm no longer produces good estimates. Namely, from standard arguments in stochastic approximation, it can be shown that for $\beta_t = \beta_0\,(t+1)^{-1-\delta}$, with $\delta > 0$, $\mathcal{CREDO}-\mathcal{NL}$'s estimate sequence may not converge to $\boldsymbol{\theta}$.*

6.8 Simulations

6.8.1 $\mathcal{CIWNLS}$

We generate a random geometric network of 10 agents, shown in Figure 2.1. The x coordinates and the y coordinates of the agents are sampled from a uniform distribution on the interval $[0, 1]$. We link two vertices by an edge if the distance between them is less than or equal to $g = 0.4$. We go on re-iterating this procedure until we get a connected graph. We choose the parameter set Θ to be $\Theta = \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^5 \in \mathbb{R}^5$. This choice of Θ conforms with Assumption 6.3.1. The sensing functions are chosen to be certain trigonometric functions as described below. The underlying parameter is 5 dimensional, $\boldsymbol{\theta} = [\theta_1,\ \theta_2,\ \theta_3,\ \theta_4,\ \theta_5]$. The sensing functions across different agents are given by, $\mathbf{f}_1(\boldsymbol{\theta}) = \sin(\theta_1+\theta_2), \mathbf{f}_2(\boldsymbol{\theta}) = \sin(\theta_3+\theta_2), \mathbf{f}_3(\boldsymbol{\theta}) = \sin(\theta_3+\theta_4), \mathbf{f}_4(\boldsymbol{\theta}) = \sin(\theta_4+\theta_5), \mathbf{f}_5(\boldsymbol{\theta}) = \sin(\theta_1+\theta_5), \mathbf{f}_6(\boldsymbol{\theta}) = \sin(\theta_1+\theta_3), \mathbf{f}_7(\boldsymbol{\theta}) = \sin(\theta_4+\theta_2), \mathbf{f}_8(\boldsymbol{\theta}) = \sin(\theta_3+\theta_5), \mathbf{f}_9(\boldsymbol{\theta}) = \sin(\theta_1+\theta_4)$ and $\mathbf{f}_{10}(\boldsymbol{\theta}) = \sin(\theta_1+\theta_5)$. Clearly, the local sensing models are unobservable, but collectively they are globally observable since, in the parameter set Θ under consideration, $\sin(\cdot)$ is one-to-one and the set of linear combinations of the $\boldsymbol{\theta}$ components corresponding to the arguments of the $\sin(\cdot)$'s constitute a full-rank system for $\boldsymbol{\theta}$. Hence, the sensing model conforms to Assumption 6.3.2. The agents make noisy scalar observations where the observation noise process is Gaussian and the noise covariance is given by $\mathbf{R} = 2\mathbf{I}_{10}$. The true (but unknown) value of the parameter is taken to be $\boldsymbol{\theta} = [\pi/6,\ -\pi/7,\ \pi/12,\ -\pi/5,\ \pi/16]$. It is readily verified that this sensing model and the parameter set $\Theta = \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^5$ satisfy Assumptions 6.3.4-6.4.3. The projection operator $\mathcal{P}_\Theta$ onto the set Θ defined in (6.15) is given by,

$$[\mathbf{x}_n(t)]_i = \begin{cases} \frac{\pi}{4} & [\widehat{\mathbf{x}}_n(t)]_i \geq \frac{\pi}{4} \\ [\widehat{\mathbf{x}}_n(t)]_i & \frac{-\pi}{4} < [\widehat{\mathbf{x}}_n(t)]_i < \frac{\pi}{4} \\ \frac{-\pi}{4} & [\widehat{\mathbf{x}}_n(t)]_i < \frac{-\pi}{4}, \end{cases} \tag{6.39}$$

for all $i = 1, \cdots, M$.

The sensing model is motivated by distributed static phase estimation in smartgrids. For a more complete

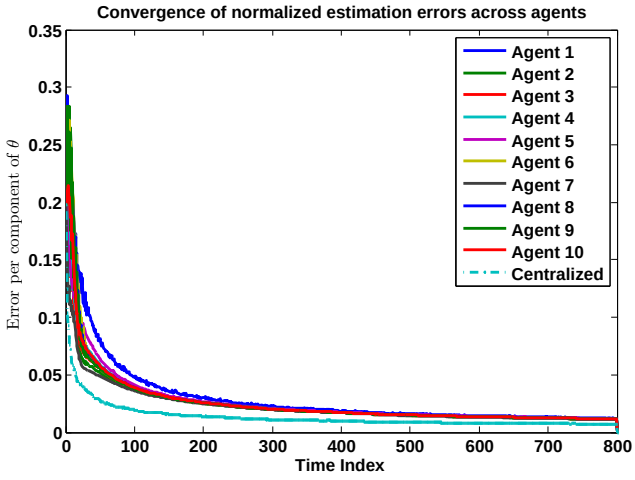


Figure 6.2: $\mathcal{CIWNLS}$: Convergence of normalized estimation error at each agent

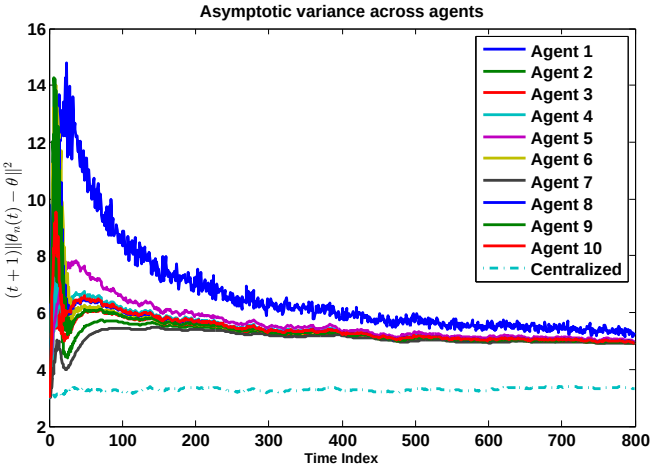


Figure 6.3: $\mathcal{CIWNLS}$: Asymptotic variance at each agent

treatment of the classical problem of static phase estimation in power grids, we direct the reader to Ilic' and Zaborszky (2000). Coming back to the current context, to be specific, the physical grid can be modeled as a network with the loads and generators being the nodes (vertices), while the transmission lines being the edges, and the sensing model reflects the power flow equations. The goal of distributed static phase estimation is to estimate the vector of phases from line flow data. The interested reader is directed to Section *IV.D* of Kar et al. (2012) for a detailed treatment of distributed static phase estimation.

We carry out 250 Monte-Carlo simulations for analyzing the convergence of the parameter estimates and their asymptotic covariances. The estimates are initialized to be 0, i.e., $\mathbf{x}_n(0) = \mathbf{0}$ for $n = 1, \cdots, 5$. The normalized error for the n-th agent at time t is given by the quantity $\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|/5$. Figure 6.2 shows the normalized error at every agent against the time index t. We compare it with the normalized error of the centralized estimator in Figure 2.2. We note that the errors converge to zero as established in Theorem 6.7.1. The decrease in error is rapid in the beginning and slows down with increasing t; this is a consequence of the decreasing weight sequences $\{\alpha_t\}$ and $\{\beta_t\}$. Finally, in Fig. 6.3 we compare the asymptotic variances

of our scheme and that of the centralized WNLS estimator. For the distributed estimator $\mathcal{CIWNLS}$ and for each n, Fig. 6.3 plots the quantities $(t+1)\|\mathbf{x}_n(t)-\boldsymbol{\theta}\|^2$ averaged over the Monte-Carlo trials. By Theorem 4.2, this quantity is expected to converge to the trace of the asymptotic covariance $\boldsymbol{\Sigma}_d$ of the $\mathcal{CIWNLS}$, i.e., $\mathrm{tr}(\boldsymbol{\Sigma}_d)$, the same for all n. We also simulate the centralized WNLS and plot the scaled error $(t+1)\left\|\widehat{\boldsymbol{\theta}}(t)-\boldsymbol{\theta}\right\|^2$ averaged over the Monte-Carlo trials. Similarly, from Proposition 6.3.1 we have that $\mathbb{E}_{\boldsymbol{\theta}}\left[(t+1)\left\|\widehat{\boldsymbol{\theta}}(t)-\boldsymbol{\theta}\right\|^2\right] \to \mathrm{tr}\left(\boldsymbol{\Sigma}_c\right)$. In this simulation setup, $\mathrm{tr}\left(\boldsymbol{\Sigma}_c\right)$ and $\mathrm{tr}\left(\boldsymbol{\Sigma}_d\right)$ are evaluated to be 3.6361 and 5.4517, respectively, a loss of about 1.76 dB. From the simulation experiment conducted above, the experimental values of $\mathrm{tr}\left(\boldsymbol{\Sigma}_c\right)$ and $\mathrm{tr}\left(\boldsymbol{\Sigma}_d\right)$ are found to be 3.9554 and 5.6790 respectively.

6.8.2 $\mathcal{CREDO}-\mathcal{NL}$

This section corroborates our theoretical findings through simulation examples and demonstrates the communication efficiency of $\mathcal{CREDO}-\mathcal{NL}$.

Specifically, we compare the proposed communication efficient distributed estimator, $\mathcal{CREDO}-\mathcal{NL}$, with the benchmark distributed recursive estimator in (6.14) and the diffusion algorithm as in Towfic et al. (2016)[9], which both utilize all inter-neighbor communications at all times, i.e., they have a linear communication cost. The example demonstrates that the proposed communication efficient estimator has a similar MSE iteration-wise rate as the two benchmark estimators. The simulation also shows that the proposed estimator improves the MSE *communication rate* with respect to the two benchmarks.

The parameters of the two benchmarks and of the proposed estimator are as follows. The benchmark estimator in (6.14) has the consensus weight set to $0.48(t+1)^{-1}$. For the proposed estimator, we set $\rho_t = 0.45(t+1)^{-0.01}$ and $\zeta_t = (t+1)^{-0.49}$. The step size sequence for the benchmark estimator proposed in Towfic et al. (2016) is set to $\mu_t = (0.3(t+20))^{-1}$. It is to be noted that the Laplacian matrix considered for the benchmark estimator and the expected Laplacian matrix for the proposed estimator, $\mathcal{CREDO}-\mathcal{NL}$ are equal, i.e., $\overline{\mathbf{L}} = \mathbf{L}$. The innovation weight is set to $\alpha_t = (0.3(t+20))^{-1}$. It is to be noted that with the time shifted innovation potential, the theoretical results continue to hold. As a performance metric, we use the relative MSE estimate averaged across nodes:

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\|\mathbf{x}_n(t)-\boldsymbol{\theta}\|^2}{\|\mathbf{x}_n(0)-\boldsymbol{\theta}\|^2},$$

further averaged across 100 independent runs of the estimators. In the above equation, $\mathbf{x}_n(0)$ refers to the initial estimates at each node, which is set as $\mathbf{x}_n(0)=0$. Figure 6.4 plots the relative MSE decay in terms of the number of iterations or the number of samples. It can be seen that the MSE decay of the two benchmark estimators and the MSE decay of the proposed estimator $\mathcal{CREDO}-\mathcal{NL}$ are very similar with respect to the iteration count. Figure 6.5 plots the MSE decay of the three estimators in terms of the communication cost per node. It can be seen for example that, at a relative MSE level of 10^{-1}, the proposed estimator requires 20x and $18x$ less communications as compared to the estimator in (6.14) and the algorithm in Towfic et al.

[9] Applied to our setting and in our notation, the diffusion method as in Towfic et al. (2016) takes the following form:

$$\mathbf{x}'_n(t+1) = \mathbf{x}_n(t) - \mu_t\left(\nabla\mathbf{f}_n(\mathbf{x}_n(t))\right)\mathbf{R}_n^{-1}\left(\mathbf{f}_n(\mathbf{x}_n(t)) - \mathbf{y}_n(t)\right)$$
$$\mathbf{x}_n(t+1) = \sum_{l\in\Omega_n\cup\{n\}} a_{ln}\,\mathbf{x}'_l(t+1).$$

Here, $\mathbf{x}_n(t)$ is the solution estimate at agent n, $\mathbf{x}'_n(t)$ is an auxiliary sequence at agent n, μ_t is the step-size, and the a_{ln}'s are combination weights that constitute together a $N\times N$ column-stochastic matrix.

(2016). One can also notice a faster MSE decay in terms of the communication cost for $\mathcal{CREDO}-\mathcal{NL}$ as compared to the benchmark (6.14), thus confirming our theory.

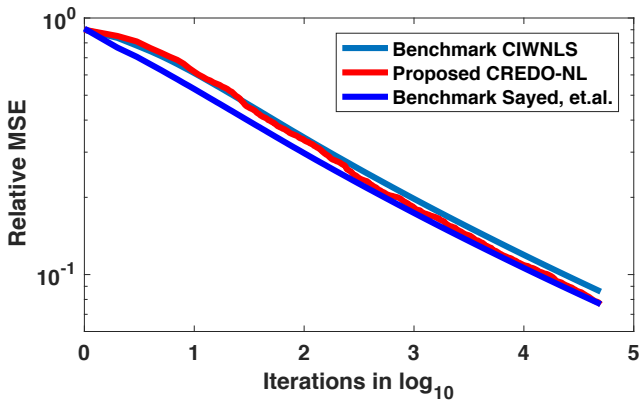


Figure 6.4: *Comparison of the proposed and benchmark estimators in terms of relative MSE: Number of Iterations.* The light blue line represents the $\mathcal{CIWNLS}$ algorithm, the dark blue line represents the diffusion based algorithm proposed in Towfic et al. (2016) and the red line represents the proposed estimator.

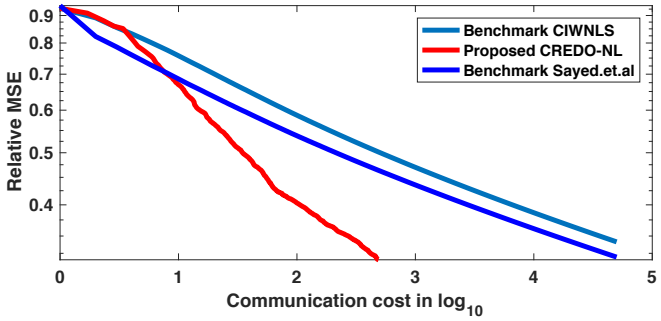


Figure 6.5: *Comparison of the proposed and benchmark estimators in terms of relative MSE: Communication Cost Per Node.* The light blue line represents the $\mathcal{CIWNLS}$ algorithm, the dark blue line represents the diffusion based algorithm proposed in Towfic et al. (2016) and the red line represents the proposed estimator.

Discussion

In the context of existing work on non-linear distributed methods, e.g., Kar and Moura (2014); Sahu et al. (2016); Jennrich (1969); Ram et al. (2010a,b, 2009); Nedic and Ozdaglar (2009). this chapter contributes by developing a method with a strictly faster communication rate of $\Theta(1/\mathcal{C}_t^{2-\zeta})(\zeta > 0$ arbitrarily small) with respect to existing $\Theta(1/\mathcal{C}_t)$ rates. Further, with respect to existing works that develop methods designed to achieve communication efficiency, e.g., Tsianos et al. (2012, 2013); Jakovetic et al. (2016); Lan et al. (2017); Wang et al. (2016), we develop here a different scheme with *randomized increasingly sparse communications*. Finally, this chapter is a continuation of works Sahu et al. (2018e,a) but, in contrast with Sahu et al. (2018e,a), it considers non-linear observation models. It would be interesting to apply the proposed method on real data sets, e.g., in the context of IoT or power systems applications, in addition to synthetic data tests considered here.

6.9 Summary of Contributions

- **Asymptotic characterization of WNLS**: We show the intrinsic trade-off between communication and optimality. In particular, by restricting the exchange of model information among the agents, by not exchanging gain matrices as in Kar and Moura (2014), we quantify the loss in the estimation accuracy in terms of the asymptotic covariance. Distributed estimation does pay a price. The asymptotic covariance of the proposed distributed estimator is not as efficient as that of the centralized estimator; nonetheless, it shows the benefits of inter-agent collaboration. In absence of inter-agent collaboration, the parameter of interest most likely is unobservable at each individual agent, and hence non-collaborative or purely decentralized procedures will lead to divergence under the usual asymptotic normality scaling at the individual network agents.
- **Communication Efficient** $\mathcal{CIWNLS}$: Despite dropping communications and the presence of non-linearities in the sensing model, we show that the proposed algorithm achieves the optimal $\Theta(1/t)$ rate of the mean square error (MSE) decay. The achievability of the optimal MSE decay in terms of time t translates into significant improvements in the rate at which MSE scales with respect to the per-agent average communication cost $\mathcal{C}_t$ up to time t – namely from $\Theta(1/\mathcal{C}_t)$ with existing methods, e.g., Kar and Moura (2014); Sahu et al. (2016); Jennrich (1969); Ram et al. (2010a,b, 2009); Nedic and Ozdaglar (2009), to $\Theta(1/\mathcal{C}_t^{2-\zeta})$ with the proposed method, where $\zeta > 0$ is arbitrarily small. We also establish strong consistency of the estimate sequence at each agent, showing that each agent's local estimator converges almost surely to the true parameter $\boldsymbol{\theta}$.

6.10 Conclusion and Future Directions

In this chapter, we have considered the problem of distributed recursive parameter estimation in a network of sparsely interconnected agents. We have proposed a *consensus* + *innovations* nonlinear least squares type algorithm, $\mathcal{CIWNLS}$, in which every agent updates its parameter estimate at every observation sampling epoch by simultaneous processing of neighborhood information and locally sensed new information and in which the inter-agent collaboration is restricted to a possibly sparse but connected communication graph. Under rather weak conditions, connectivity of the inter-agent communication and a *global observability* criterion, we have shown that the proposed algorithm leads to consistent parameter estimates at each agent. Furthermore, under standard smoothness assumptions on the sensing nonlinearities, we have established order-optimal pathwise convergence rates and the asymptotic normality of the parameter estimate sequences generated by the proposed distributed estimator $\mathcal{CIWNLS}$. In the context of $\mathcal{CREDO}-\mathcal{NL}$ which is a communication efficient distributed estimation scheme for non-linear observation models, we established strong consistency of the estimate sequence at each agent and characterized the MSE decay in terms of the per-agent communication cost $\mathcal{C}_t$. $\mathcal{CREDO}-\mathcal{NL}$ achieves the MSE decay rate $\Theta\left(\mathcal{C}_t^{-2+\zeta}\right)$, where $\zeta > 0$ and ζ is arbitrarily small. A natural direction for future research consists of obtaining techniques and conditions to obtain innovation gains so as to reduce the gap between the agent asymptotic covariances and that of the centralized WNLS estimator. Methods developed in Kar and Moura (2014) may be employed and extended to obtain such characterization. Future research directions also include extending the proposed algorithm to a mixed-time scale stochastic approximation type algorithm, so as to achieve an asymptotic covariance independent of the network, as well as to extend the presented ideas to distributed stochastic optimization.

Chapter 7

Communication Efficient Distributed Estimation: Random Fields Estimation

7.1 Introduction

In this chapter, we are interested in distributed inference of the state of large-scale cyber physical systems (CPS) like sensor networks monitoring a spatially distributed field tracking environmental modalities, or CPS where physical entities with sensing capabilities are deployed over large areas. An important example of the systems of interest is the smart grid–a large network of generators and loads instrumented with, for example, phasor measurement units (PMUs). Our goal is to reconstruct the physical field or state of the CPS that is represented by a parameter or a random field. The structure of the physical layer is reflected through the coupling among the observation sequences across different nodes. Suppose, for the purpose of illustration, corresponding to each field location, there is a low-power inexpensive sensor monitoring the location. The noisy sensor measurement at a location in the field is not only influenced by the parameter or field component at that location but possibly is a function of *neighboring* field components. As an example, in the smart grid context, a sensor at a node (location) may obtain a measurement of the power flowing into that node, which in turn is a function of the field components (e.g., voltages, angles) at that node and neighboring nodes. This coupling among parameter components in the measurements will be referred to as the *physical coupling* in the sequel. However, since the local measurement at a location is influenced by multiple field parameters, due to possible lack of observability or identifiability, in order to come up with a provably consistent estimates of the parameter components of interest, each agent exchanges information with its neighborhood which conforms to a pre-assigned inter-agent communication graph. The inter-agent communication graph forms the cyber layer of the system and is different from that of the physical layer, i.e., the coupling structure among the parameter components induced by the distributed measurement model. Due to the large-scale of the CPS and the high-dimensionality of the field, reconstructing the entire field at each agent may be too taxing and beyond the capability of the agents, and hence, agents may only be interested in estimating certain components of the parameter field locally; furthermore, the components of interest at a given agent, referred to as the *interest set* of the agent, varies from agent to agent.
We propose a scheme, namely $\mathcal{CIRFE}$, where each entity reconstructs only a subset of the components of the state modeled by a vector parameter, and thereby also reducing the dimension of messages being communicated among the agents. Under mild conditions of the connectivity of the network, we establish consistency of the estimate sequence at each agent with respect to the components of the parameters in its interest set.

The proposed scheme allows heterogeneity in terms of agents' objectives, while still allowing for inter-agent collaboration. Through $\mathcal{CIRFE}$, we address communication efficiency for the class of distributed inference algorithm of the *consensus+innovations* form by reducing the dimension of vectors exchanged among the agents.

7.2 Related Work

Existing central coordinator less distributed estimation schemes, such as in Kar and Moura (2011); Das and Mesbahi (2006); Schizas et al. (2008a); Lopes and Sayed (2008); Stankovic et al. (2007); Schizas et al. (2008b); Ram et al. (2010b); Sahu and Kar (2016) aim to reconstruct the entire parameter at each node of the networked setup, thus conforming to a homogeneous objective across all nodes. Distributed recursive inference schemes addressing estimation of a possibly high-dimensional parameter vector (see, for example Kar and Moura (2011); Das and Mesbahi (2006); Schizas et al. (2008a); Lopes and Sayed (2008); Stankovic et al. (2007); Schizas et al. (2008b); Ram et al. (2010b); Sahu and Kar (2016)), tend to communicate at each (discrete or slotted) time step in its neighborhood and exchange estimates of the entire parameter vector at each time instant. Owing to the high-dimensionality of the state vector and limited storage and processing capabilities in the individual entities of a large-scale CPS, exchanging high-dimensional estimates may be undesirable. The aforementioned setups subsume the knowledge of the dimension of the state vector to be estimated and hence adapt the storage requirements at each agent to cater to the exact dimension of the state vector. A random field estimation scheme in a fully distributed setup with arbitrary connected inter-agent communication topology where agents reconstruct only a subset of the physical field which, in turn, is coupled with the sensing field was also proposed in Kar (2010) (Chapter 3). The current work is inspired by Kar (2010) and generalizes the development in Kar (2010) in several fronts to achieve better estimate performance.

7.3 Problem Formulation

Consider N physical agents monitoring a field over a large physical area. Each agent n is associated with a scalar state θ_n^*, which represents the field intensity parameter at its location. The agents are equipped with sensing capabilities. We assume each agent observes a time-series of measurements, given by noisy linear functions of its state and the states of *neighboring* agents. Due to this coupling in the observations, an agent should cooperate with neighbors to reconstruct its own state. For simplicity, we assume that the individual agent states are scalars. Our results can be generalized to vector valued states, though at the cost of extra notation. The observation at each agent is of the form:

$$\mathbf{y}_n(t) = \mathbf{H}_n\boldsymbol{\theta}^* + \gamma_n(t), \tag{7.1}$$

where $\mathbf{H}_n \in \mathbb{R}^{M_n \times N}$ is a sparsifying (to be clarified soon) sensing matrix, $\{\mathbf{y}_n(t)\}$ is a $\mathbb{R}^{M_n}$-valued observation sequence for the n-th agent and for each n where possibly $M_n \ll N$, $\{\gamma_n(t)\}$ is a zero-mean temporally independent and identically distributed (i.i.d.) noise sequence with nonsingular covariance matrix $\mathbf{R}_n$. It is to be noted that the assumption that the dimension of the parameter $\boldsymbol{\theta}^*$ is equal to the number of agents, N, is simply made for clarity of presentation. In particular, all our proofs and assertions will continue to hold with appropriate modifications if the dimension of the global parameter is different from N.

Assumption 7.2.1. *There exists $\epsilon_1 > 0$, such that, for all n, $\mathbb{E}_{\boldsymbol{\theta}}\left[\|\gamma_n(t)\|^{2+\epsilon_1}\right] < \infty$.*

The above assumption encompasses a broad class of noise distributions in the setup. The heterogeneity of the setup is exhibited in terms of the sensing matrix and the noise covariances at the agents. We now formalize an assumption on global model observability.

Assumption 7.2.2. *The matrix* $\mathbf{G} = \sum_{n=1}^{N} \mathbf{H}_n^\top \mathbf{R}_n^{-1} \mathbf{H}_n$ *is full rank.*

Assumption 7.2.2 is crucial for our distributed setup. It is to be noted that such an assumption is needed for even a setup with a centralized node which has access to all the data samples at each of the agent nodes at each time. Assumption 2.3.2 ensures that if a hypothetical fusion center could stack all the data samples together at any time t, it would have sufficient information so as to be able to unambiguously estimate the parameter of interest. Hence, the requirement for this assumption naturally extends to our distributed setup. As far as reconstructing the parameter $\boldsymbol{\theta}$ is concerned, there is an inherent scalability issue as the dimension of the parameter scales with the size of the network. Owing to the ad-hoc nature of setups as described above and observations being made at different agents in a sequential manner, one has to resort to recursive message-passing schemes while conforming to a communication protocol specified by a inter-agent communication graph. Given the possibly high-dimensional state of the field, it is not desirable and communication-wise feasible to exchange the high-dimensional data in the form of parameter estimates and for each agent to estimate the entire vector. Before, going over specifics of our algorithm, we next review recursive estimation both in the centralized and distributed setups.

7.3.1 Preliminaries

In this section, we go over the preliminaries of classical distributed estimation.

Distributed Estimation:

In the setup described above in (7.1), if a hypothetical fusion center having access to the data samples at all nodes at all times were to conduct the parameter estimation in a recursive manner, a (centralized) recursive least-squares type approach could be employed as follows:

$$\mathbf{x}_c(t+1) = \mathbf{x}_c(t) + \underbrace{\frac{a}{t+1} \sum_{n=1}^{N} \mathbf{H}_n^\top \boldsymbol{\Sigma}_n^{-1} \left(\mathbf{y}_n(t) - \mathbf{H}_n \mathbf{x}_c(t)\right)}_{\text{Global Innovation}},$$

where a is a positive constant such that $a > N / \left(\lambda_{min}\left(\sum_{n=1}^{N} \mathbf{H}_n^\top \mathbf{R}_n^{-1} \mathbf{H}_n\right)\right)$. However, such a fusion center based scheme may not be implementable in our distributed multi-agent setting with time-varying sparse inter-agent interaction primarily due to the fact that the desired global innovation computation requires instantaneous access to the entire set of network sensed data at all times at the fusion center. Moreover, the fusion center intends to reconstruct the entire high-dimensional state and thus, maintains a N-dimensional estimate at all times. If in the case of a distributed setup, an agent n in the network were to replicate the centralized update by replacing the global innovation in accordance with its local innovation, the update for the parameter estimate becomes

$$\widehat{\mathbf{x}}_n(t+1) = \widehat{\mathbf{x}}_n(t)$$

$$+\underbrace{\frac{a}{t+1}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\left(\mathbf{y}_n(t)-\mathbf{H}_n\widehat{\mathbf{x}}_n(t)\right)}_{\text{Local Innovation}},$$

where $\{\widehat{\mathbf{x}}_n(t)\}$ represents the estimate sequence at agent n. The above update involves purely decentralized and independent local processing with no collaboration among the agents whatsoever. However, note that in the case when the data samples obtained at each agent lacks information about all the features, the parameter estimates would be erroneous and sub-optimal. As in the case of the fusion center based approach outlined above, each agent maintains a N-dimensional estimate at all times and hence the messages exchanged in the neighborhood are N-dimensional and could be very large depending on the size of the network. Hence, as a surrogate to the global innovation in the centralized recursions, the local estimators compute a local innovation based on the locally sensed data as an agent has access to information only in its neighborhood. The information loss at a node is compensated by incorporating an agreement or consensus potential into their updates which is then incorporated (see, for example Kar and Moura (2011); Kar et al. (2013a); Sahu et al. (2016)) as follows:

$$\mathbf{x}_n(t+1)=\mathbf{x}_n(t)-\underbrace{\frac{b}{(t+1)^{\delta_1}}\sum_{l\in\Omega_n(t)}\left(\mathbf{x}_n(t)-\mathbf{x}_l(t)\right)}_{\text{Neighborhood Consensus}}$$
$$+\underbrace{\frac{a}{t+1}\boldsymbol{\Gamma}^{-1}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\left(\mathbf{y}_n(t)-\mathbf{H}_n\mathbf{x}_n(t)\right)}_{\text{Local Innovation}}, \tag{7.2}$$

where $0<\delta_1<1$, $\Omega_n(t)$ represents the neighborhood of agent n at time t and a,b are appropriately chosen positive constants. In the above scheme, the information exchange among agent nodes is limited to the parameter estimates. It has been shown in previous work that under appropriate conditions (see, for example Kar and Moura (2011)), the estimate sequence $\{\mathbf{x}_n(t)\}$ converges to $\boldsymbol{\theta}^*$ and is asymptotically normal, i.e.,

$$\sqrt{t+1}\left(\mathbf{x}_n(t)-\boldsymbol{\theta}\right)\stackrel{\mathcal{D}}{\Longrightarrow}\mathcal{N}\left(0,\left(N\boldsymbol{\Gamma}\right)^{-1}\right),$$

where $\boldsymbol{\Gamma}=\frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\mathbf{R}_n^{-1}\mathbf{H}_n$ and $\stackrel{\mathcal{D}}{\Longrightarrow}$ denotes convergence in distribution. The above established asymptotic normality also points to the conclusion that the MSE decays as $\Theta(1/t)$. For future reference, we will refer to the distributed estimation approach in (7.2) as the classical consensus+innovations approach. The aforementioned scheme, though optimal in terms of the asymptotic covariance entails the availability of global model information at each agent and exchange of the entire parameter estimate which in turn is N-dimensional among agents. Furthermore, due to the inherent spatial coupling in the observation sequence at each node with other nodes in its neighborhood, the availability of a particular entry of the state vector is localized to a small area. Hence, a large-scale deployment of such a system, would incorporate a significant delay for an agent to assimilate information about a particular entry of the state vector which is not local with respect to its neighborhood. Moreover, such a scheme requires the knowledge of the dimension of the state vector at each agent and storage of a high-dimensional local estimates, same as the size of the entire state vector. Such prior knowledge about attributes of the parameter such as dimension in conjunction with requirement for large memory at each agent might be practically infeasible owing to the ad-hoc nature and limited sensing, computation and storage capabilities of agents in a networked setup.

Thus, in both of the schemes above, specifically in the case which involves estimating a high-dimensional pa-

rameter, it might not be practical to estimate the entire parameter at each agent. In such a high-dimensional parameter estimation scheme, it is highly favorable to estimate only a few entries of the parameter based on the requirements of each agent, which could potentially reduce the dimensions of messages being exchanged in the network thereby reducing the implementation complexity considerably.

7.3.2 Connections with Distributed Optimization

In principle, distributed stochastic optimization, with each node interested in a few entries of the optimization variable, is more general than the distributed estimation/random fields setup studied here. Indeed, one recovers the setup here with specializing the cost functions to be quadratic. However, this is true only for a very generic formulation of distributed stochastic optimization, where no strong convexity is assumed, each node is interested in a subset of the variable of interest, and the gradient (first order) information is subject to noise, and the underlying network is random. However, to the best of our knowledge, there is no present work that simultaneously addresses all of these aspects. For example, in Mota et al. (2015), the setup involves a static network connected at all times with each agent having access to an incremental first order oracle, i.e., access to exact gradient information; the chapter establishes convergence the iterate sequences to the optimizer, however, rates of convergence are not provided. In Alghunaim and Sayed (2017), the authors consider coupled distributed stochastic optimization setups where the coupling is induced by interest sets of different agents over static networks. The setup in Alghunaim and Sayed (2017) encompasses estimation setups, given that *global observability*[1] holds for each entry of the parameter in the respective clusters, which in turn is subsumed in the setup. Technically speaking, typical distributed optimization setups rely on *local observability*[2] without assuming *local correctness*[3] at each agent. However, in the case of distributed estimation, the agents lack *local observability* but preserve *local correctness*. Moreover, the study of the mean square error in Alghunaim and Sayed (2017) reflects errors in terms of the step sizes only and does not reflect explicit dependence in terms of the number of agents collaborating to estimate a particular entry of the parameter. In comparison with Mota et al. (2015); Alghunaim and Sayed (2017), we consider a distributed estimation setup over time-varying networks connected only on average and provide asymptotic characterization of the estimator as time goes to ∞. Furthermore, we specifically characterize the scaling of the asymptotic variance of each entry of the parameter in terms of the number of agents interested in reconstructing the particular entry in question. We also characterize the fundamental condition so as to generate consistent estimates of each entry of the parameter and show that connectivity of the network and global observability is not enough to ensure consistency of the estimates. We direct the reader to assumption 7.4.3 and the discussion after assumption 7.4.4 for a detailed illustration. In particular, we establish that connectivity of the subgraphs induced by the interest sets is a sufficient condition to enforce assumption 7.4.3. It is an open question as to what is a necessary condition (in terms of the network structure, sensing structure, and the interest sets' structure) so as to enforce assumption 7.4.3.

[1] Global Observability refers to the condition, when the parameter can be reconstructed by stacking the samples collected from all the agents.

[2] Local observability refers to the condition, where an agent can reconstruct its own state based on its own observation sequence.

[3] Local correctness refers to the condition, where the set of local optimizers for the agent's local cost function includes the optimizer of the global objective.

7.4 $\mathcal{CIRFE}$: Distributed Random Fields Estimation

In this section, we develop the algorithm $\mathcal{CIRFE}$. The parameter to be reconstructed which is the vector of states accumulated over the entire network is $\boldsymbol{\theta}^* \in \mathbb{R}^N$. The sparsifying nature of $\mathbf{H}_n$ in (7.1) is related to the coupling induced by the measurements in the field. To be specific, let us define $\widetilde{\mathcal{I}}_n$ as the set of agents whose states influence the measurement $\mathbf{y}_n(t)$ at agent n, i.e., $\widetilde{\mathcal{I}}_n$ collects the agents for which the corresponding columns of matrix $\mathbf{H}_n$ is non-zero. In what follows, we say an agent n is physically coupled to an agent l if the observation at agent n is influenced by the state component θ_l^*. Typically, $\widetilde{\mathcal{I}}_n$ is a small subset of the total number of agents N. Technically speaking, the above mentioned coupling induced by the measurements can be expressed in terms of an adjacency matrix, $\hat{\mathbf{A}}$, where $\hat{\mathbf{A}}_{nl} = 1$ if $l \in \widetilde{\mathcal{I}}_n$ and 0 otherwise. Now, that we have abstracted out the physical coupling (physical layer) in the networked system under consideration, we discuss about the communication layer (cyber layer), i.e., the inter-agent communication network and the associated communication protocol. Before getting into the communication protocol, we introduce *interest sets* of agents' around which the communication protocol is built.

Assumption 7.4.1. *The set of agents physically coupled with agent n is a subset of the interest set of agent n, i.e., $\widetilde{\mathcal{I}}_n \subset \mathcal{I}_n$.*

Furthermore, we assume that the interest set of every agent n is non-empty.
We number the nodes (equivalently, components of $\boldsymbol{\theta}$) in the interest sets of agents in increasing order. Thus, the interest set $\mathcal{I}_n$ at an agent n can be considered to be a vector with dimension $|\mathcal{I}_n|$. For example, $\mathcal{I}_n(r) = p$ indicates that agent p is the r-th agent in increasing order in the interest set $\mathcal{I}_n$. We also have that $\mathcal{I}_n^{-1}(p) = r$. Moreover, as each agent n is only interested in reconstructing the states of agents in its interest set, the estimate at agent n, $\mathbf{x}_n(t) \in \mathbb{R}^{|\mathcal{I}_n|}$, $\forall\ t$. At every time instant t, an agent n simultaneously fuses information received from the neighbors and the latest sensed information to update its parameter estimate. However, as the interest set of agents in the neighborhood might not be the same as that of the agent itself, the information received from the neighbors needs censoring. Let the message received from agent l at time t be denoted by $\mathbf{x}_l(t) \in \mathbb{R}^{|\mathcal{I}_l|}$, where $l \in \Omega_n$. The censored message processed by agent n, $\mathbf{x}_{l,n}^r(t) \in \mathbb{R}^{|\mathcal{I}_n|}$ is generated as follows:

$$\mathbf{e}_j^\top \mathbf{x}_{l,n}^r(t) = \begin{cases} \mathbf{e}_{\mathcal{I}_l^{-1}(\mathcal{I}_n(j))}^\top \mathbf{x}_l(t) & \mathcal{I}_n(j) \in \mathcal{I}_l \\ 0 & \text{otherwise,} \end{cases} \tag{7.3}$$

where $\mathbf{e}_j$ and $\mathbf{e}_{\mathcal{I}_l^{-1}(\mathcal{I}_n(j))}$ are canonical vectors with $\mathbf{e}_j \in \mathbb{R}^{|\mathcal{I}_n|}$ and $\mathbf{e}_{\mathcal{I}_l^{-1}(\mathcal{I}_n(j))} \in \mathbb{R}^{|\mathcal{I}_l|}$. Agent n only wants to use estimates of those states from an agent in its neighborhood which are common to their interest sets. Formally, with agent l, agent n only wants to use estimates of the states in the set $\mathcal{I}_n \cap \mathcal{I}_l$. Similarly, while using the obtained estimate states from the neighbors, only those states in the set $\mathcal{I}_n \cap \mathcal{I}_l$ are updated. We also define the transformed estimate $\mathbf{x}_{l,n}^s(t) \in \mathbb{R}^{|\mathcal{I}_n|}$ at agent n, for each $l \in \Omega_n(t)$ as follows:

$$\mathbf{e}_j^\top \mathbf{x}_{l,n}^s(t) = \begin{cases} \mathbf{e}_j^\top \mathbf{x}_n(t) & \mathcal{I}_n(j) \in \mathcal{I}_l \\ 0 & \text{otherwise.} \end{cases} \tag{7.4}$$

where $j \in \{1, \cdots, |\mathcal{I}_n|\}$. The agent n also incorporates the latest sensed information $\mathbf{y}_n(t)$ while updating the parameter estimate at each sampling epoch and only retains the components of interest, i.e., those in

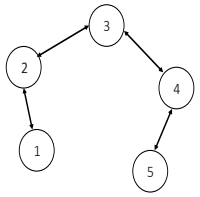


Figure 7.1: A network example emphasizing the notion of structural observability.

$\mathcal{I}_n$. For a given vector $\mathbf{z} \in \mathbb{R}^{|\mathcal{I}_n|}$, let $\mathbf{z}^{\mathcal{P}_{\mathcal{I}_n}} \in \mathbb{R}^N$ be the vector whose j-th component is given by

$$\mathbf{e}_j^\top \mathbf{z}^{\mathcal{P}_{\mathcal{I}_n}} = \begin{cases} \mathbf{e}_{\mathcal{I}_n^{-1}(j)}^\top \mathbf{z} & j \in \mathcal{I}_n \\ 0 & \text{otherwise.} \end{cases} \tag{7.5}$$

Finally, for a given vector $\mathbf{z} \in \mathbb{R}^N$, $\mathbf{z}_{\mathcal{I}_n}$ denotes the vector in $\mathbb{R}^{|\mathcal{I}_n|}$, where $\mathbf{e}_j^\top \mathbf{z}_{\mathcal{I}_n} = \mathbf{e}_{\mathcal{I}_n(j)}^\top \mathbf{z}$.
We now introduce the algorithm $\mathcal{CIRFE}$ for distributed parameter estimation:

$$\begin{aligned} \mathbf{x}_n(t+1) = \mathbf{x}_n(t) - \underbrace{\sum_{l \in \Omega_n(t)} \beta_t \left(\mathbf{x}_{l,n}^s(t) - \mathbf{x}_{l,n}^r(t) \right)}_{\text{Neighborhood Consensus}} \\ + \underbrace{\alpha_t \mathbf{H}_n^\top \mathbf{R}_n^{-1} \left(\mathbf{y}_n(t) - \mathbf{H}_n \mathbf{x}_n^{\mathcal{P}_{\mathcal{I}_n}}(t) \right)_{\mathcal{I}_n}}_{\text{Local Innovation}}, \end{aligned} \tag{7.6}$$

where $\Omega_n(t)$ represents the neighborhood of agent n at time t; and $\{\beta_t\}$ and $\{\alpha_t\}$ are the consensus and innovation weight sequences given by

$$\beta_t = \frac{\beta_0}{(t+1)^{\delta_1}}, \alpha_t = \frac{a}{t+1}, \tag{7.7}$$

where $a, b > 0$ and $0 < \delta_1 < 1/2 - 1/(2+\epsilon_1)$ and ϵ_1 was as defined in Assumption 2.3.1. It is to be noted that with the interest set of each agent being $\mathcal{I}_n = \{1, 2, \cdots, N\}$, we have that the update in (7.6) reduces to the classical consensus+innovations update for linear parameter estimation schemes (see, Kar et al. (2013a) for example). Thus, the classical consensus+innovations parameter estimation scheme, is strictly a special case of the update in (7.6).
We now illustrate the introduced setup and algorithm (7.6) with a 5 agents network example in Fig. 7.1. Each node n corresponds to a physical component θ_n^*. Thus, $\boldsymbol{\theta}^* \in \mathbb{R}^5$. The solid lines connecting the nodes correspond to the inter-node communication pattern. Each node observes a noisy scalar functional. In particular, we assume

$$\begin{aligned} y_3(t) &= \frac{1}{3}\left(\theta_2^* + \theta_3^* + \theta_4^*\right) + \gamma_3(t) \\ y_n(t) &= \theta_n^* + \gamma_n(t), n = 1, 2, 4, 5. \end{aligned} \tag{7.8}$$

Note that then the noise covariance matrix $\mathbf{R}_n$ is a positive scalar, $n = 1, 2, ..., 5$. Also, for $n \neq 5$, $\mathbf{H}_n$ is a 5-dimensional (row) vector with all entries equal to zero except the n-th entry which equals one. On the other hand, $\mathbf{H}_3 = [\,0,\, 1/3,\, 1/3,\, 1/3,\, 0\,]$. we have that $\widetilde{\mathcal{I}_n} = \{n\}$ for $n = 1, 2, 4, 5$, and $\widetilde{\mathcal{I}_3} = \{2, 3, 4\}$. Let

us also assume that the agents' interest sets are given by $\mathcal{I}_n = \widetilde{\mathcal{I}_n}$, for each $n = 1, 2, ..., 5$. For notational simplicity, we omit time index t when writing the agents' estimates; that is, we write $\mathbf{x}_n$ in place of $\mathbf{x}_n(t)$. Also, we denote by $[\mathbf{x}_n]_i$ the i-th entry of $\mathbf{x}_n$. Then, agent 3's estimate $\mathbf{x}_3$ is a 3×1 vector, with $[\mathbf{x}_3]_1$ being an estimate of θ_2^*, $[\mathbf{x}_3]_2$ being an estimate of θ_3^*, and $[\mathbf{x}_3]_3$ being an estimate of θ_4^*. Regarding the remaining agents $n \neq 3$, we have that $\mathbf{x}_n$ is a scalar, with $\mathbf{x}_n$ being an estimate of θ_n^*. Next, consider agent 3 and its interaction with agent 2. The censored quantity $\mathbf{x}_{32}^r$ at agent 3 based on the received message from agent 2 equals $\mathbf{x}_{32}^r = [\,\mathbf{x}_2,\, 0,\, 0\,]^\top$. Further, the agent 3's own censored estimate, adapted so that it can be combined with $\mathbf{x}_{32}^r$, equals $\mathbf{x}_{32}^s = [\,[\mathbf{x}_3]_1,\, 0,\, 0\,]^\top$. Note that the first entry in both $\mathbf{x}_{32}^r$ and $\mathbf{x}_{32}^s$ corresponds to an estimate of θ_2^*, the second entry of both $\mathbf{x}_{32}^r$ and $\mathbf{x}_{32}^s$ corresponds to an estimate of θ_3^*, and the third entry of both $\mathbf{x}_{32}^r$ and $\mathbf{x}_{32}^s$ corresponds to an estimate of θ_4^*. The second and third entry in both $\mathbf{x}_{32}^r$ and $\mathbf{x}_{32}^s$ is zero, because the intersection of the agents' 2 and 3 interest sets $\mathcal{I}_1 \cap \mathcal{I}_2 = \{2\}$, i.e., it does not include the interest for θ_3^* nor for θ_4^*. Further, we have that $\mathbf{x}_{23}^r = [\mathbf{x}_3]_1$ and $\mathbf{x}_{23}^s = \mathbf{x}_2$. The remaining pairs of quantities $\mathbf{x}_{nl}^r$ and and $\mathbf{x}_{nl}^s$ are defined analogously. Next, agent 3's estimate "lifted" to the $N = 5$-dimensional space equals $\widetilde{\mathbf{x}}_3 = [0,\ [\mathbf{x}_3]_1,\, [\mathbf{x}_3]_2,\, [\mathbf{x}_3]_3,\, 0\,]^\top$. Note that the first and fifth entries in $\widetilde{\mathbf{x}}_3$ are zero, because agent 3 does not have interest in θ_1^* nor in θ_5^*. Similarly, we have that $\widetilde{x}_2 = [\,0,\, \mathbf{x}_2,\, 0,\, 0,\, 0\,]^\top$. We next specialize the update rule (7.6) for the example considered here and agent 3; we have:

$$\underbrace{\begin{bmatrix}[\mathbf{x}_3]_1(t+1)\\ [\mathbf{x}_3]_2(t+1)\\ [\mathbf{x}_3]_3(t+1)\end{bmatrix}}_{\mathbf{x}_3(t+1)} = \underbrace{\begin{bmatrix}[\mathbf{x}_3]_1(t)\\ [\mathbf{x}_3]_2(t)\\ [\mathbf{x}_3]_3(t)\end{bmatrix}}_{\mathbf{x}_3(t)} + \beta_t \underbrace{\left(\begin{bmatrix}\mathbf{x}_2(t)\\ 0\\ 0\end{bmatrix} - \begin{bmatrix}[\mathbf{x}_3]_1(t)\\ 0\\ 0\end{bmatrix}\right)}_{\mathbf{x}_{32}^r(t) - \mathbf{x}_{32}^s(t)}$$
$$+ \beta_t \underbrace{\left(\begin{bmatrix}0\\ 0\\ \mathbf{x}_4(t)\end{bmatrix} - \begin{bmatrix}0\\ 0\\ [\mathbf{x}_3]_3(t)\end{bmatrix}\right)}_{\mathbf{x}_{34}^r(t) - \mathbf{x}_{34}^s(t)}$$
$$+ \alpha_t \underbrace{\begin{bmatrix}1/3\\ 1/3\\ 1/3\end{bmatrix}}_{(\mathbf{H}_3^\top)_{\mathcal{I}_3}} R_3^{-1} \left(y_3(t) - \frac{1}{3}([\mathbf{x}_3]_1(t) + [\mathbf{x}_3]_2(t) + [\mathbf{x}_3]_3(t))\right). \tag{7.9}$$

We formalize an assumption on the connectivity of the inter-agent communication graph before proceeding further.

Assumption 7.4.2. *The inter-agent communication graph is connected on average, i.e., $\lambda_2(\overline{\mathbf{L}}) > 0$, where $\overline{\mathbf{L}}$ denotes the mean of the sequence of identically and independently distributed (i.i.d) graph Laplacian sequence $\{\mathbf{L}(t)\}$.*

Remark 7.4.1. *In the parameter estimation scheme in (7.6), an agent n uses only those components of its neighbor l's estimate $\mathbf{x}_l(t)$, which belong to its interest set $\mathcal{I}_n$. Thus, agents n and l combine components linearly which belong to $\mathcal{I}_n \cap \mathcal{I}_l$ and reject the rest of the components. From an implementation viewpoint, it is desirable for an agent l to only transmit those components to agent n which belong to $\mathcal{I}_n \cap \mathcal{I}_l$ instead of transmitting the entire $\mathbf{x}_l(t)$ to agent n as the one which involves exchanging only those components which are common to the agents has lower communication overhead. In the former case, the receiving agent n will zero out the components it does not require, so both the transmission strategies would lead to the same update.*

Moreover, in the innovation term, where an agent n uses its own previous state to compute the innovation, an agent subsequently retains only the components of interest so as to keep the update economical in terms of size. We also emphasize here that the inter-agent communication graphs $\{L(t)\}$ and the physical adjacency matrix $\hat{A}$ induced by the measurement coupling may be structurally different.

We now present a more compact representation of the $\mathcal{CIRFE}$ algorithm so as to be able to establish its asymptotic convergence properties. Let $\mathcal{I}$ denote a subset of $\{1, 2, \cdots, N\}$. Define the diagonal matrix $P_{\mathcal{I}}$ which selects the corresponding non-zero components of $\mathcal{I}$ from a $\mathbb{R}^{N^2}$ dimensional vector. In particular, $P_{\mathcal{I}} = \text{diag}\,[P_{\mathcal{I}_1}, \cdots, P_{\mathcal{I}_n}]$, where each $P_{\mathcal{I}_n} \in \mathbb{R}^{N\times N}$ and is a diagonal matrix such $[P_{\mathcal{I}_n}]_{i,i} = 1$ if $i \in \mathcal{I}_n$ or 0 otherwise.

For the 5-agent network example associated with Figure 1, we have for $n \neq 3$ that $P_{\mathcal{I}_n}$ is the 5×5 matrix with all the entries equal to zero, except the (n, n)-th entry which equals one. The matrix $P_{\mathcal{I}_3}$ has all the entries equal to zero, except the $(2, 2)$-th, $(3, 3)$-th, and $(4, 4)$-th entries, which al equal to one.

For the estimate sequence $\{\mathbf{x}_n(t)\}$ at agent n, let $\{\widetilde{\mathbf{x}}_n(t)\} \in \mathbb{R}^N$ denote the auxiliary estimate sequence, where $\widetilde{\mathbf{x}}_n(t) = \mathbf{x}_n(t)^{\mathcal{P}_{\mathcal{I}_n}}$. With the above development in place, it is easy to see that, for $\mathbf{y} \in \mathbb{R}^N$, $\left(\mathbf{x}^r_{l,n}(t)\right)^{\mathcal{P}_{\mathcal{I}_n}} = P_{\mathcal{I}_n} P_{\mathcal{I}_l} \widetilde{\mathbf{x}}_l(t)$, $\left(\mathbf{x}^s_{l,n}(t)\right)^{\mathcal{P}_{\mathcal{I}_n}} = P_{\mathcal{I}_n} P_{\mathcal{I}_l} \widetilde{\mathbf{x}}_n(t)$ and $(\mathbf{x}_n(t))^{\mathcal{P}_{\mathcal{I}_n}} = P_{\mathcal{I}_n} \widetilde{\mathbf{x}}_n(t)$. The $\mathcal{CIRFE}$ update in (7.6) can then be written in terms of the auxiliary processes as follows:

$$\begin{aligned}\widetilde{\mathbf{x}}_n(t+1) = \widetilde{\mathbf{x}}_n(t) - \sum_{l\in\Omega_n(t)} \beta_t P_{\mathcal{I}_n} P_{\mathcal{I}_l} \left(\widetilde{\mathbf{x}}_n(t) - \widetilde{\mathbf{x}}_l(t)\right) \\ + \alpha_t P_{\mathcal{I}_n} \mathbf{H}_n^\top \mathbf{R}_n^{-1} \left(\mathbf{y}_n(t) - \mathbf{H}_n P_{\mathcal{I}_n} \widetilde{\mathbf{x}}_n(t)\right). \end{aligned} \tag{7.10}$$

We introduce the matrix $\mathbf{L}_{\mathcal{P}} \in \mathbb{R}^{N^2\times N^2}$ so as to make the above representation more compact.

$$[\mathbf{L}_{\mathcal{P}}(t)]_{nl} = \begin{cases} -P_{\mathcal{I}_n} \sum_{r=1:r\neq n}^{N} \mathbf{L}_{nr}(t) P_{\mathcal{I}_r} & \text{if } n = l \\ \mathbf{L}_{nl}(t) P_{\mathcal{I}_l} P_{\mathcal{I}_n} & \text{otherwise,} \end{cases} \tag{7.11}$$

where $[\mathbf{L}_{\mathcal{P}}(t)]_{nl} \in \mathbb{R}^{N\times N}$ denotes the (n, l)-th sub-block of the block matrix $\mathbf{L}_{\mathcal{P}}$. It follows by elementary matrix multiplication properties that $\mathcal{P}\mathbf{L}_{\mathcal{P}}(t) = \mathbf{L}_{\mathcal{P}}(t)$. It is also to be noted that $\mathbf{L}_{\mathcal{P}}$ is a symmetric matrix. The matrix $\mathbf{L}_{\mathcal{P}}(t)$ at each time step t can be decomposed as follows:

$$\mathbf{L}_{\mathcal{P}}(t) = \overline{\mathbf{L}_{\mathcal{P}}} + \widetilde{\mathbf{L}_{\mathcal{P}}}(t), \tag{7.12}$$

where $\{\mathbf{L}_{\mathcal{P}}(t)\}$ is an i.i.d. sequence with mean $\overline{\mathbf{L}_{\mathcal{P}}}$ and $\widetilde{\mathbf{L}_{\mathcal{P}}}(t) = \mathbf{L}_{\mathcal{P}}(t) - \mathbb{E}\left[\mathbf{L}_{\mathcal{P}}(t)\right]$. Thus, we have that the residual sequence $\{\widetilde{\mathbf{L}_{\mathcal{P}}}(t)\}$ satisfies $\mathbb{E}\left[\widetilde{\mathbf{L}_{\mathcal{P}}}(t)\right] = \mathbf{0}$.

With the above development in place, the update in (7.10) can be written in a compact form as follows:

$$\widetilde{\mathbf{x}}(t+1) = \widetilde{\mathbf{x}}(t) - \beta_t \mathbf{L}_{\mathcal{P}}(t)\widetilde{\mathbf{x}}(t) + \alpha_t \mathcal{P}\mathbf{G}_H \mathbf{R}^{-1}\left(\mathbf{y}(t) - \mathbf{G}_H^\top \mathcal{P}\widetilde{\mathbf{x}}(t)\right), \tag{7.13}$$

where $\widetilde{\mathbf{x}}^\top(t) = \left[\widetilde{\mathbf{x}}_1^\top(t), \cdots, \widetilde{\mathbf{x}}_N^\top(t)\right]^\top$, $\mathbf{y}(t)^\top = [y_1(t)^\top \cdots y_N(t)^\top]^\top$, $\mathbf{R} = diag\,[\mathbf{R}_1, \cdots, \mathbf{R}_N]$, $\mathcal{P} = diag\,[\mathcal{P}_{\mathcal{I}_1}, \cdots, \mathcal{P}_{\mathcal{I}_N}]$, and $\mathbf{G}_H = diag[\mathbf{H}_1^\top, \mathbf{H}_2^\top, \cdots, \mathbf{H}_N^\top]$.

Remark 7.4.2. *In the case when the noise covariance is not known apriori, a recursive estimator of the inverse noise covariance can be used so as to be used as a plugin estimate for $\mathbf{R}_n^{-1}$. A plugin estimate for*

$\mathbf{R}_n^{-1}$ *at time* $t+1$, *denoted by* $\widehat{\mathbf{R}}_n^{-1}(t+1)$ *can be generated as follows:*

$$\mathbf{Q}_n(t+1) = \frac{1}{t}\sum_{s=0}^{t}\mathbf{y}_n(s)\mathbf{y}_n^\top(s) - \left(\frac{1}{t}\sum_{s=0}^{t}\mathbf{y}_n(s)\right)\left(\frac{1}{t}\sum_{s=0}^{t}\mathbf{y}_n(s)\right)^\top$$
$$\widehat{\mathbf{R}}_n^{-1}(t+1) = \left(\mathbf{Q}_n(t+1) + \gamma_t\mathbf{I}_{M_n}\right)^{-1},$$

where γ_t *is a time-decaying sequence such that* $\gamma_t \to 0$ *as* $t \to \infty$.
Also, given the sensing model and the assumption that the dimension of the observations at each agent n, *given by* M_n *is* $M_n \ll N$, *inverting a low-dimensional matrix is not particularly computationally taxing. In particular,* M_n *can be equal to* 1 *for instance in which the inverse noise covariance matrix can be estimated seamlessly. Furthermore, it is to be noted that the update can be adapted to be of the following form, where* $\mathbf{R}^{-1}$ *is replaced by* $\mathbf{I}$

$$\widetilde{\mathbf{x}}(t+1) = \widetilde{\mathbf{x}}(t) - \beta_t\mathbf{L}_{\mathcal{P}}(t)\widetilde{\mathbf{x}}(t) + \alpha_t\mathcal{P}\mathbf{G}_H\left(\mathbf{y}(t) - \mathbf{G}_H^\top\mathcal{P}\widetilde{\mathbf{x}}(t)\right),$$

which does not require the inverse noise covariance. We remark that with the above update, the algorithm still retains the property concerning the almost sure convergence of the parameter estimate at each agent to the entries of the parameter corresponding to its interest set. Thus, the computational cost can be reduced drastically with an update of the following form as defined above, which does not involve any matrix inversions. Thus, when knowledge or calculation of $\mathbf{R}^{-1}$ *is an issue, algorithm* (7.13) *can be replaced with the update above, retaining consistency but possibly with a loss in terms of the asymptotic covariance.*

Remark 7.4.3. *The recursive update in* (7.13) *is of the stochastic approximation type. The stochastic approximation procedure, employed here is a mixed time-scale stochastic approximation as opposed to the classical single time-scale stochastic approximation (see, for example Nevelson and Khasminskiĭ (1973)). The above notion of mixed time-scale is very different from the more commonly studied two time-scale stochastic approximation (see, for instance Borkar (2008)) in which a fast process is coupled with a slower dynamical system. The approach employed here is similar to the ones in Gelfand and Mitter (1991) and Kar et al. (2013a) in which a single update procedure is influenced by multiple potentials with different time-decaying weights. Now, suppose that the interest set of each agent consists of all components of* $\boldsymbol{\theta}^*$, *i.e., the update in* (7.13) *reduces to the classical consensus+innovations update in (2). A key technical step employed in the analysis of classical consensus+innovations procedures of the type in* (7.2) *(see, for example, Kar et al. (2013a)) consists of an approximation of the update in* (7.2) *to a single time-scale stochastic approximation procedure that is* asymptotically equivalent *to the former, in particular, that converges to the original iterate sequence at a rate faster than* $(t+1)^{0.5}$. *Typically, in the context of* (7.2) *the approximating single time-scale procedure is the network-averaged estimate sequence,* $\widetilde{\mathbf{x}}_{avg}(t) = \left(\frac{\mathbf{1}_N^\top}{N}\otimes\mathbf{I}_N\right)\widetilde{\mathbf{x}}(t)$, *and the analysis in Kar et al. (2013a) uses the fact that the Laplacian* $\mathbf{L}(t)$ *in* (7.2) *has a left eigen vector of* $\mathbf{1}_{N^2}$ *and that every agent is interested in estimating the entire parameter vector. However, in the context of the update in* (7.13), *every agent is interested in only a few entries of the parameter which makes the characterization of asymptotic properties of the estimate sequences highly non-trivial and substantially different from prior work on consensus+innovations type estimation procedures Kar et al. (2013a) in which agents share the common objective of estimating all components of the parameter. However, in contrast to prior work on consensus+innovations type estimation procedures (see, for example Kar et al. (2013a)) in which agents share the common objective of estimating all components of the parameter, the analysis with heterogeneous*

agent objectives in (7.13), in that each agent is interested in a different subset of components, requires new technical machinery. In particular, to obtain asymptotic properties of (7.13), we develop a more generalized approximation of the mixed time-scale procedure to an appropriate single time-scale procedure that takes into account of the heterogeneity in agent objectives; this approximation and subsequent analysis require new technical tools that we develop in this chapter.

Define the subspace $\mathcal{S}_P \in \mathbb{R}^{N^2}$ by $\mathcal{S}_P = \left\{\mathbf{y} \in \mathbb{R}^{N^2} | \mathbf{y} = \mathcal{P}\mathbf{w}, \text{for some } \mathbf{w} \in \mathbb{R}^{N^2}\right\}$. We now formalize a key assumption relating the interest sets $\mathcal{I}_n$ to the network connectivity and global observability.

Assumption 7.4.3. *There exists a constant $c_1 > 0$ such that,*

$$\mathbf{y}^\top \left(\frac{\beta_0}{\alpha_0}\overline{\mathbf{L}_{\mathcal{P}}} + \mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\mathbf{G}_H^\top\mathcal{P}\right)\mathbf{y} \geq c_1 \left\|\mathbf{y}\right\|^2, \forall\ \mathbf{y} \in \mathcal{S}_P, \tag{7.14}$$

where $\mathcal{P} = diag\left[\mathcal{P}_{\mathcal{I}_1}, \cdots, \mathcal{P}_{\mathcal{I}_N}\right]$.

We formalize an assumption on the innovation gain sequence $\{\alpha_t\}$ before proceeding further.

Assumption 7.4.4. *Let $\lambda_{min}(\cdot)$ denote the smallest eigenvalue. We require that a satisfies,*

$$a \min\{\lambda_{min}\left(\sum_{n=1}^{N} \mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}_n^{-1}\mathbf{H}_n\mathcal{P}_{\mathcal{I}_n}\right), c_1, \beta_0^{-1}\} \geq 1,$$

where $\otimes$ denotes the Kronecker product and c_1 is defined in (7.14).

It is to be noted that in Assumption 7.4.3, if $\mathcal{P} = \mathbf{I}_{N^2}$, then the subspace $\mathcal{S}_P$ reduces to $\mathbb{R}^{N^2}$ and the condition in (7.14) reduces to a commonly employed Lyapunov condition in classical consensus+innovations type inference procedures (see, for example, Lemma 6 in Kar and Moura (2011)) which, in turn, can be enforced by global observability and the mean connectivity of the network under consideration. However, in the case when $\mathcal{P} \neq \mathbf{I}_{N^2}$, the case considered in this chapter, global observability and connectivity of the network is not sufficient to obtain the condition in (7.14). The insufficiency of global observability and connectivity of the network in order to enforce (7.14) can be attributed to heterogeneous objectives of the agents and censoring of messages at agents leading to an inherent information loss. Intuitively, such a condition calls for existence of information pathways between agents who share a particular component in their interest sets and the particular component in question to be observable at this set of agents collectively. As we show in the following (Lemma 7.4.4), a sufficient condition for Assumption 7.4.3 is that in addition to the global observability and the mean network connectedness, the induced subgraph for every entry of the vector $\boldsymbol{\theta}^*$ needs to be connected. The induced subgraph for the r-th entry is the set of agents and their associated links which have the r-th entry of $\boldsymbol{\theta}^*$ in their interest sets.

In the following, we will establish consistency of the $\mathcal{CIRFE}$ under Assumption 7.4.3. We now show by a simple example that, in general, Assumption 7.4.3 is stronger than mean connectivity and global observability. To this end, consider again the simple network consisting of 5 nodes in Fig. 7.1 and (7.8). Clearly, in this case, $G = \sum_{n=1}^{5} \mathbf{H}_n^\top\mathbf{H}_n$ is invertible and, as shown, the communication network is connected. In case, every node wants to estimate the entire $\boldsymbol{\theta}^*$, then the above inference task reduces to the inference setup considered in Kar and Moura (2011); Sahu et al. (2016). Consider the case where $\mathcal{I}_n = \widetilde{\mathcal{I}}_n$ for $n = 1, 2, 3, 4$, i.e., these nodes are interested in reconstructing only their own states and those who influence their observations. However, let $\mathcal{I}_5 = \{5, 1\}$, i.e., node 5 is interested in the state of node 1. This problem falls under the purview of $\mathcal{CIRFE}$. Clearly, Assumption 7.4.2 is satisfied. However, it can be shown by calculating the

various terms, that assumption 7.4.3 is not satisfied and hence, convergence of $\mathcal{CIRFE}$ to desired values is not guaranteed. This shows that mean connectivity and global observability is not sufficient for assumption 7.4.3 in general. We provide an intuitive explanation, why the $\mathcal{CIRFE}$ is not expected to yield accurate estimates in this case and why the Lyapunov type requirement in assumption 7.4.3 is sufficient for $\mathcal{CIRFE}$'s desired convergence.. Looking at Fig. 7.1, we note that the only node that observes (at least partially) the component θ_1^* is node 1, i.e., the influence of the state θ_1^* only affects the observations at node 1. Clearly, for node 5 to be able to reconstruct θ_1^*, it should be able to access information about θ_1^* from the allowed communication graph. Moreover, there is a path connecting node 1 to node 5. However, the other nodes in the path are not interested in reconstructing θ_1^*, so they do not participate in the exchange of information regarding θ_1^*. For example, node 2 ignores the estimate of θ_1^* at node 1 and similarly the others. As a result, the information about θ_1^* never reaches node 5, although the communication network is connected. Note that the induced subgraph of component 1 of $\boldsymbol{\theta}^*$ is disconnected, and it involves only nodes 1 and 5 and no links.

At the same time, it is easy to see that this problem is resolved if an extra communication link is added between nodes 1 and 5. Thus, we see that connectivity of the subgraph formed by those nodes interested in reconstructing θ_1 seems to facilitate proper information flow necessary for the desired convergence of $\mathcal{CIRFE}$. Based on this intuition, we formulate a general structural connectivity condition (see Kar (2010)) that guarantees the satisfaction of 7.4.3 which, in turn, will be used subsequently to derive the convergence of $\mathcal{CIRFE}$. We direct the reader to Lemma 3.4.1 in Kar (2010) for a proof.

Lemma 7.4.4 (Lemma 3.4.1 in Kar (2010))**.** *Let assumption 7.4.2 be satisfied and the global observability condition hold. For each component r of θ^*, define the subset $\mathcal{I}^r \subset [1, \cdots, N]$ by*

$$\mathcal{I}^r = \{n \in [1, \cdots, N] \mid r \in \mathcal{I}_n\} \tag{7.15}$$

Let $\overline{\mathcal{G}}$ denote the network graph corresponding to the mean Laplacian $\overline{\mathbf{L}}$, i.e., there is an edge between nodes n and l in $\overline{G}$ iff *the (n,l)-th entry in $\overline{\mathbf{L}}$ is non-zero. For each $1 \leq r \leq N$, denote the induced subgraph $\overline{\mathcal{G}}_r$ of $\overline{\mathcal{G}}$ with node set $\mathcal{I}^r$. Then, condition 7.4.3 is satisfied if $\overline{\mathcal{G}}_r$ is connected for all r.*

Technically speaking, the average connectedness of the induced subgraphs in conjunction with the global observability of the entry of the parameter relevant to the subgraphs is enough to ensure consistency of the estimate sequence of the entry of the parameter. The combinatorial perspective brought about by the preceding observation being, can one relax the connectivity of the induced subgraph. For example, consider the r-th entry of the parameter. Let the number of agents interested to estimate the entry is N_r out of which M agents (referred to as $\mathcal{O}$-agents) have the entry incorporated into their observations. In the case, when one can split N_r agents into disconnected components where each component consists of non-zero number of agents which observe the entry and the entry is rendered globally observable with respect to those $\mathcal{O}$-agents in that component, would ensure the estimates of that entry being consistent at each agent which is interested to reconstruct that agent. However, as the subgraphs induced by interest sets are coupled in lieu of the interest sets, it might not be possible to ensure such a construction as the one described before for each entry of the parameter.

7.5 $\mathcal{CIRFE}$: Main Results

In this section we formally state the main results concerning the distributed parameter estimation $\mathcal{CIRFE}$ algorithm, while the proofs are relegated to F. The first result concerns with the consistency of the parameter estimate sequence at each agent n.

Theorem 7.5.1. *Consider the parameter estimate sequence $\{\widetilde{\mathbf{x}}(t)\}$ generated by the $\mathcal{CIRFE}$ algorithm according to (7.6). Then, we have,*

$$\mathbb{P}_{\boldsymbol{\theta}^*}\left(\lim_{t\to\infty}\widetilde{\mathbf{x}}(t)=\mathcal{P}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right)=1. \tag{7.16}$$

At this point, we note that the estimate sequence generated by $\mathcal{CIRFE}$ at any agent n is strongly consistent, i.e., $\mathbf{x}_n(t)\to\boldsymbol{\theta}^*_{\mathcal{I}_n}$ almost surely (a.s.) as $t\to\infty$. It is also to be noted that, owing to the heterogeneous objectives of the agents, the consensus in terms of the estimates sequences across any pair of agents is only limited to the common components of the parameter in their interest sets.

Theorem 7.5.2. *Let the hypothesis of theorem 7.5.1 hold. Then, we have,*

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{x}}(t)-\mathcal{P}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right\|^2\right]=O\left(\frac{1}{t}\right) \tag{7.17}$$

Thus, we note that the mean square error of the estimate sequence with respect to the components of the parameter $\boldsymbol{\theta}^*$ decays as $1/t$.

Theorem 7.5.3. *Let the hypotheses of Theorem 7.5.1 hold. Then, the time-scaled sequence $\sqrt{t+1}\left(\widetilde{\mathbf{x}}(t)-\mathcal{P}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right)$ is asymptotically normal, i.e.,*

$$\sqrt{t+1}\left(\widetilde{\mathbf{x}}(t)-\mathcal{P}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right)\xRightarrow{\mathcal{D}}\mathcal{N}\left(\mathbf{0},\mathbf{S}_R\right), \tag{7.18}$$

where

$$\begin{aligned}\mathbf{S}_R&=\mathbf{P}\mathbf{M}\mathbf{P}^\top\\ [\mathbf{M}]_{ij}&=\left[\mathbf{P}\mathbf{Q}\left(\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}_n^{-1}\mathbf{H}_n\mathcal{P}_{\mathcal{I}_n}\right)\mathbf{Q}\mathbf{P}\right]_{ij}\\ &\times\left([\boldsymbol{\Lambda}]_{ii}+[\boldsymbol{\Lambda}]_{jj}-1\right)^{-1},\end{aligned} \tag{7.19}$$

and $\mathbf{P}$ and $\boldsymbol{\Lambda}$ are orthonormal and diagonal matrices such that $\mathbf{P}^\top\mathbf{Q}\left(\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}_n^{-1}\mathbf{H}_n\mathcal{P}_{\mathcal{I}_n}\right)\mathbf{P}=\boldsymbol{\Lambda}$, in which, $\mathbf{Q}=diag\left[\frac{1}{Q_1},\frac{1}{Q_2},\cdots,\frac{1}{Q_N}\right]$, with Q_i denoting the number of agents interested in the i-th entry of $\boldsymbol{\theta}^*$.

noindent Theorem 7.5.3 establishes the asymptotic normality of the time-scaled (auxilliary) estimate sequence. Noting that the estimate sequence $\{\mathbf{x}_n(t)\}$ is a linear transformation of the auxiliary estimate sequence, we conclude that $\sqrt{t+1}\left(\mathbf{x}_n(t)-\boldsymbol{\theta}^*_{\mathcal{I}_n}\right)$ is also asymptotically normal. It is also to be noted that, when the interest sets of each agent is the identity matrix, i.e., every agent is interested to reconstruct the entire parameter, the matrix $\mathbf{Q}$ reduces to $\frac{\mathbf{I}}{N}$ and the asymptotic covariance reduces to that of the classical consensus+innovations linear parameter estimation case (see Kar and Moura (2011) and the corresponding

update in (7.2)). In this sense, the classical linear parameter estimation case is a special case of the problem being addressed here. It is to be noted that the case in which $\mathbf{Q}$ reduces to $\frac{1}{\widetilde{Q}}\mathbf{I}$ for some $\widetilde{Q} < N$ ($\widetilde{Q} < N$ agents interested in each entry of $\boldsymbol{\theta}^*$), the asymptotic covariance reduces to,

$$\mathbf{S}_R = \frac{a\mathbf{I}}{2\widetilde{Q}} + \frac{\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\mathbf{R}_n^{-1}\mathbf{H}_n + \frac{\mathbf{I}}{2a}\right)^{-1}}{\widetilde{Q}}.$$

The asymptotic covariance as derived in Theorem 7.5.3 explicitly showcases the heterogeneity in the scaling with respect to different components of the parameter through $\mathbf{Q}$, as different components have different cardinalities of interest sets.

The convergence rate is unaffected by the communication of low-dimensional estimates, i.e., the mean square error of the proposed scheme decays as $1/t$ as characterized by Theorem 7.5.2. However, by communicating low dimensional estimates which is due to the interest sets being strict subsets of $\{1, 2, \cdots, N\}$, the variance of the estimation scheme is affected in terms of scaling by the number of agents. In particular, as demonstrated by Theorem 7.5.3, the variance of the estimate sequence scales inversely with the number of agents interested to reconstruct the particular entry. Thus, larger the size of the communicated estimates lower is the variance. For instance, the variance scaling as $1/N$ is obtained if every agent is interested to reconstruct the entire parameter. Intuitively speaking, the difference in scaling can be attributed to averaging by a smaller number of agents against averaging by the entire network. However, note that the scaling is only with respect to the asymptotic covariance and as we will demonstrate later on line graphs, the finite time variance of the error estimates can be lower for the proposed algorithm with respect to agents which directly do not observe the component of the parameter being estimated.

7.6 Simulation Results

In this section, we demonstrate the efficiency of the proposed algorithm $\mathcal{CIRFE}$ through simulation experiments on a synthetic dataset. In particular, we construct a 10 node ring network, where every agent has exactly two nodes in its communication neighborhood. We number the nodes from 1 to 10. The neighbors for the i-th node in the communication graph are the nodes $(i-1)mod\ 10$ and $(i+1)mod\ 10$.

The physical coupling which affects each agent's observations is assumed to be an agent's 2-hop neighborhood. For instance, node 1's observations are affected by the value of the field at nodes 9, 10, 2 and 3. Thus, $\widetilde{\mathcal{I}}_1 = \{9, 10, 2, 3\}$. The interest set of each agent is taken to be all the field values which affects its observation. For instance, $\mathcal{I}_1 = \{9, 10, 1, 2, 3\}$. We resort to a static Laplacian in the simulation setup here. We also note that in this case the inter-agent communication network is sparser than the physical network induced by measurement coupling. Each agent makes a scalar observation at each time. Hence, the observation matrix for each agent is given by a 5-sparse 10-dimensional row vector. To be specific, the observation matrices used in the simulation setup are given by $\mathbf{H}_1 = [1.0, 1.2, 1.3, 0, 0, 0, 0, 0, 1.4, 1.5]$, $\mathbf{H}_2 = [1.5, 1.0, 1.2, 1.3, 0, 0, 0, 0, 0, 1.4]$, $\mathbf{H}_3 = [1.4, 1.5, 1.0, 1.2, 1.3, 0, 0, 0, 0, 0]$, $\mathbf{H}_4 = [0, 1.4, 1.5, 1.0, 1.2, 1.3, 0, 0, 0, 0]$, $\mathbf{H}_5 = [0, 0, 1.4, 1.5, 1.0, 1.2, 1.3, 0, 0, 0]$, $\mathbf{H}_6 = [0, 0, 0, 1.4, 1.5, 1.0, 1.2, 1.3, 0, 0]$, $\mathbf{H}_7 = [0, 0, 0, 0, 1.4, 1.5, 1.0, 1.2, 1.3, 0]$, $\mathbf{H}_8 = [0, 0, 0, 0, 0, 1.4, 1.5, 1.0, 1.2, 1.3]$, $\mathbf{H}_9 = [1.3, 0, 0, 0, 0, 0, 1.4, 1.5, 1.0, 1.2]$ and $\mathbf{H}_{10} = [1.2, 1.3, 0, 0, 0, 0, 0, 1.4, 1.5, 1.0]$. The noise covariance $\mathbf{R}$ is taken to be $\mathbf{I}_{10}$. The parameter capturing the field values is taken to be $\boldsymbol{\theta} = [1.2, 1.3, 1.4, 0.8, 0.7, 1.1, 0.9, 1.0, 1.8, 0.6]$. It can be seen that Assumption 7.4.3 is satisfied, by verifying Lemma 7.4.4 for the third parameter component θ_3^*.

We carry out 500 Monte-Carlo simulations for analyzing the convergence of the parameter estimates. The

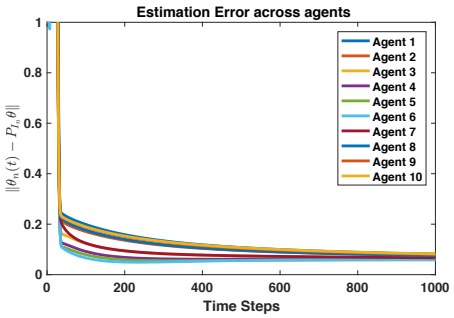


Figure 7.2: $\mathcal{CIRFE}$: Convergence of normalized estimation error at each agent

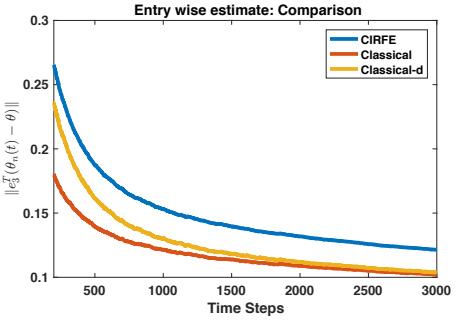


Figure 7.3: $\mathcal{CIRFE}$: Comparison of $e_3^\top \boldsymbol{\theta}^*$ estimation error

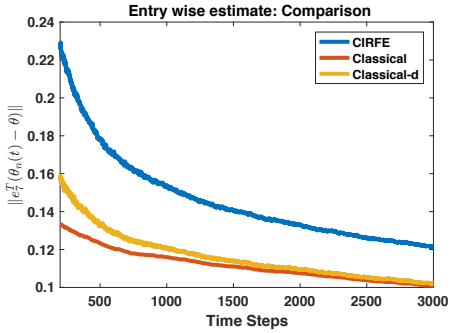


Figure 7.4: $\mathcal{CIRFE}$: Comparison of $e_7^\top \boldsymbol{\theta}^*$ estimation error

estimates are initialized as $\mathbf{x}_n(0) = \mathbf{0}$ for $n = 1, \cdots, 10$. The normalized error for the n-th agent at time t is given by the quantity $\|\mathbf{x}_n(t) - \mathcal{P}_{\mathcal{I}_n}\boldsymbol{\theta}\|/5$, as each agent's interest set has the cardinality of 5. Figure 7.2 shows the normalized error at every agent against the time index t. In Figures 7.3 and 7.4 we compare the performance of $\mathcal{CIRFE}$ to the classical distributed estimator in Kar and Moura (2011) (see (7.2) for the corresponding update), where each agent is interested in reconstructing the entire state or the parameter vector. We refer to the estimates of the distributed estimator in Kar and Moura (2011) as "classical" and "classical-d" (to be specified shortly) in the sequel. In Figures 7.3 and 7.4, "Classical-d" represents the case in the algorithm in Kar and Moura (2011), where an agent does not observe the entry to be estimated and entirely depends on the neighborhood communication to estimate the quantity of interest. We specifically study the estimation performance of the agents in the "Classical-d" case, as these are the agents that tend to increase the communication overhead considerably by being interested in estimates of components that they do not directly observe, relying on other agents possibly far off to obtain the desired information. Note

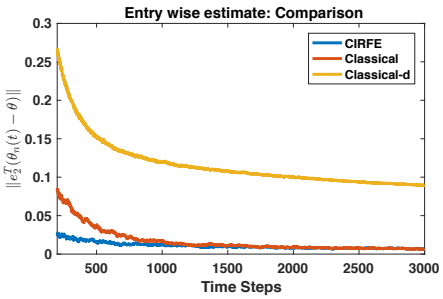


Figure 7.5: $\mathcal{CIRFE}$: Comparison of $e_2^\top \boldsymbol{\theta}^*$ estimation error

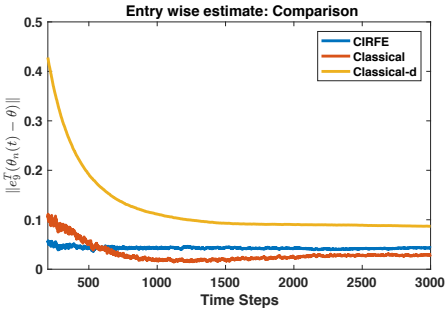


Figure 7.6: $\mathcal{CIRFE}$: Comparison of $e_9^\top \boldsymbol{\theta}^*$ estimation error

that, in the current simulation setup, such class of agents do not exist for the proposed $\mathcal{CIRFE}$ algorithm. It can be observed from figures 7.3 and 7.4 that the estimation error in $\mathcal{CIRFE}$ is higher than that of the classical distributed estimator but at the same time exchanging 5-dimensional or even smaller dimensional messages as opposed to 10-dimensional messages in the case of the classical consensus+innovations estimator in Kar and Moura (2011). This analysis brings about an inherent trade-off between estimation error and the dimension of the messages exchanged between agents. It is also to be noted that the agents in case of $\mathcal{CIRFE}$ store 5-dimensional vectors at each time step as opposed to 10-dimensional vectors in the case of the classical. An intuitive way to interpret the higher estimation error is noting the fact that, effective for the algorithm $\mathcal{CIRFE}$, the estimation procedure for each entry of the parameter $\boldsymbol{\theta}^*$ effectively happens over a line graph, whereas for the Classical and "Classical-d" procedures the communication graph to which the estimation procedure conforms to is a ring graph. In order to demonstrate the effectiveness of the algorithm $\mathcal{CIRFE}$, we consider a line graph, where the agents have the same sensing model as in the previous case except for the two edges of the line graph. Thus, agent 1 and 10's observations are dependent on agent 2 and agent 9's state. Furthermore, we assume that each agent's observation is physically coupled with the states of the agents' in its one-hop neighborhood. The interest set for the 1st and 10th agents are taken to be $\{1, 2\}$ and $\{9, 10\}$ respectively. All the other agents, have interest sets of cardinality three, i.e, itself and its one-hop neighborhood. In Figures 7.5 and 7.6 we compare the performance of $\mathcal{CIRFE}$ to the classical distributed estimator in Kar and Moura (2011) (see (7.2) for the corresponding update), with the aforementioned line graph setup. For the "classical-d" case, the agent selected was the farthest end of the graph. It is well known that under a line graph, the performance of a distributed protocol is affected due to poor connectivity. It can be seen from figures 7.5 and 7.6 that the performance of $\mathcal{CIRFE}$ closely resembles that of the classical benchmark algorithm with respect to an agent which observes the particular entry. However, for agents far away from the agent which observes the particular entry, $\mathcal{CIRFE}$ outperforms them. Intuitively speaking,

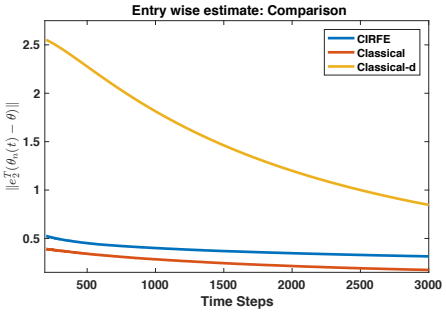


Figure 7.7: $\mathcal{CIRFE}$: Comparison of $e_2^\top \boldsymbol{\theta}^*$ estimation error

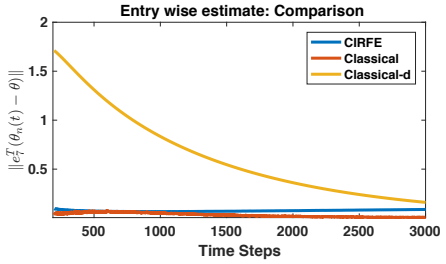


Figure 7.8: $\mathcal{CIRFE}$: Comparison of $e_7^\top \boldsymbol{\theta}^*$ estimation error

while in this case, the communication protocol for each entry of the parameter in $\mathcal{CIRFE}$ conforms to a line graph, where the maximum number of vertices in each is 3, for the benchmark the line graph consists of 10 agents. In order to reinforce the effectiveness of $\mathcal{CIRFE}$, we ran experiments on a 30 node line graph, where each agent except the nodes numbered 1, 2, 29 and 30, have an interest set of cardinality 5. The nodes numbered 1, 2, 29 and 30 are assumed to have interest sets of cardinality 3, 4, 4 and 3 respectively. For instance the interest sets of agents 1 and 2 are given by $\{1, 2, 3\}$ and $\{1, 2, 3, 4\}$ respectively. We assume that the physical coupling which affects each agent's observation is limited to its two-hop neighborhood. In Figures 7.7 and 7.8 we compare the performance of $\mathcal{CIRFE}$ to the classical distributed estimator in Kar and Moura (2011) (see (7.2) for the corresponding update), with the aforementioned line graph setup. For the "classical-d" case, the agent selected was the farthest end of the graph as in the previous case. It can be seen from figures 7.7 and 7.7 that the performance of $\mathcal{CIRFE}$ closely resembles that of the classical benchmark algorithm with respect to an agent which observes the particular entry. However, for agents far away from the agent which observes the particular entry, $\mathcal{CIRFE}$ outperforms them.

Technically speaking, in the classical case, an agent which is diameter number of steps away from a particular agent requires diameter number of time steps to fuse information from the other agent for an entry which it does not observe. In contrast with the classical case, the estimation of a particular entry of the parameter effectively happens over the induced subgraph with respect to the particular entry which typically will have smaller diameter as compared to the original graph. In conclusion, forcing an agent to obtain estimates of all parameter components may actually slow down the overall process in many scenarios of interest (especially situations involving large graphs with poor connectivity), as some of these components are only observed at agents geographically distant from the agent under consideration.

7.7 Summary of Contributions

- **Generalized *consensus+innovations*:** Through $\mathcal{CIRFE}$, we illustrated that the classical *consensus+innovations* framework is a special case of $\mathcal{CIRFE}$. In particular, we extended the idea of consensus to a heterogeneous version which exhibits consensus to subspaces which is common to a few agents. Such a construction, also threw light on the necessity of looking beyond global observability and connectivity of the graph for establishing consistency of the estimate sequence.
- **Lower Dimensional Message Exchange:** We propose a scheme, namely $\mathcal{CIRFE}$, where each entity reconstructs only a subset of the components of the state modeled by a vector parameter, and thereby also reducing the dimension of messages being communicated among the agents. Under mild conditions of the connectivity of the network, we establish consistency of the estimate sequence at each agent with respect to the components of the parameters in its interest set. The proposed scheme allows heterogeneity in terms of agents' objectives, while still allowing for inter-agent collaboration. Technically speaking, the heterogeneity in terms of agents' objectives seeks for consensus in agents' estimates only in terms of the components in their interest sets rather than the entire high-dimensional parameter vector as in the case of other distributed estimation algorithms proposed in Kar and Moura (2011); Das and Mesbahi (2006); Schizas et al. (2008a); Lopes and Sayed (2008); Stankovic et al. (2007); Schizas et al. (2008b); Ram et al. (2010b); Sahu and Kar (2016).

7.8 Conclusion and Future Directions

In this chapter, we have proposed a *consensus + innovations* type algorithm, $\mathcal{CIRFE}$, for estimating a high-dimensional parameter or field that exhibits a cyber-physical flavor. In the proposed algorithm, every agent updates its estimate of a few components of the high-dimensional parameter vector by simultaneous processing of neighborhood information and local newly sensed information and in which the inter-agent collaboration is restricted to a possibly sparse communication graph. Under rather generic assumptions we establish the consistency of the parameter estimate sequence and characterize the asymptotic variance of the proposed estimator. A natural direction for future research consists of considering models with non-linear observation functions and extension of the proposed algorithm $\mathcal{CIRFE}$ to quantized communication schemes in the lines of Kar et al. (2012) and Zhang et al. (2017).

Part III

Distributed Optimization

Chapter 8

Communication Efficient Stochastic Optimization: First Order

8.1 Introduction

Distributed optimization and learning algorithms attract a great interest in recent years, thanks to their widespread applications including distributed estimation in networked systems, e.g., Kar and Moura (2011), distributed control, e.g., Bullo et al. (209), and big data analytics, e.g., Daneshmand et al. (2015).

In this chapter, we study communication efficient distributed stochastic optimization algorithms that operate over random networks and minimize smooth strongly convex costs. We consider standard distributed stochastic gradient methods where at each time step, each node makes a weighted average of its own and its neighbors' solution estimates, and performs a step in the negative direction of its noisy local gradient. Before going to the communication efficient version, we first study distributed stochastic optimization over random networks. The underlying network is allowed to be *randomly varying*, similarly to, e.g., the models in Lobel and Ozdaglar (2011); Lobel et al. (2011); Jakovetic et al. (2014b). More specifically, the network is modeled through a sequence of independent identically distributed (i.i.d.) graph Laplacian matrices, where the network is assumed to be connected on average. (This translates into the requirement that the algebraic connectivity of the mean Laplacian matrix is strictly positive.) Random network models are highly relevant in, e.g., internet of things (IoT) and cyber physical systems (CPS) applications, like, e.g., predictive maintenance and monitoring in industrial manufacturing systems, monitoring smart buildings, etc. Therein, networked nodes often communicate through unreliable/intermittent wireless links, due to, e.g., low-power transmissions or harsh environments.

We show that, by carefully designing the consensus and the gradient weights (potentials), the considered distributed stochastic gradient algorithm achieves the order-optimal $O(1/k)$ rate of decay of the mean squared distance from the solution (mean squared error – MSE). This is achieved for twice continuously differentiable strongly convex local costs, assuming also that the noisy gradients are unbiased estimates of the true gradients and that the noise in gradients has bounded second moment. To the best of our knowledge, this is the first time an order-optimal convergence rate for distributed strongly convex stochastic optimization has been established for random networks. For the communication efficient first order distributed stochastic optimization, we propose a novel method that is shown to achieve the $O(1/(C_{\text{comm}})^{4/3-\zeta})$ MSE communication rate. At the same time, the proposed method retains the order-optimal $O(1/(C_{\text{comp}}))$ MSE rate in terms of the computational cost, the best achievable rate in the corresponding centralized setting.

8.2 Related Work

We now briefly review the literature. In the context of the extensive literature on distributed optimization, the most relevant to our work are the references on: 1) distributed strongly convex stochastic (sub)gradient methods; and 2) distributed (sub)gradient methods over random networks (both deterministic and stochastic methods). For the former thread of works, several papers give explicit convergence rates under different assumptions. Regarding the underlying network, references Tsianos and Rabbat (2012); Towfic et al. (2016) consider static networks, while the works Yuan et al. (2018); Vanli et al. (2017); Nedic and Olshevsky (2016) consider deterministic time-varying networks.

References Tsianos and Rabbat (2012); Towfic et al. (2016) consider distributed strongly convex optimization for static networks, assuming that the data distributions that underlie each node's local cost function are equal (reference Tsianos and Rabbat (2012) considers empirical risks while reference Towfic et al. (2016) considers risk functions in the form of expectation); this essentially corresponds to each nodes' local function having the same minimizer. References Yuan et al. (2018); Vanli et al. (2017); Nedic and Olshevsky (2016) consider deterministically varying networks, assuming that the "union graph" over finite windows of iterations is connected. The papers Tsianos and Rabbat (2012); Towfic et al. (2016); Yuan et al. (2018); Vanli et al. (2017) assume undirected networks, while Nedic and Olshevsky (2016) allows for directed networks but assumes a bounded support for the gradient noise. The works Tsianos and Rabbat (2012); Yuan et al. (2018); Vanli et al. (2017); Nedic and Olshevsky (2016) allow the local costs to be non-smooth, while Towfic et al. (2016) assumes smooth costs, as we do here. With respect to these works, we consider random networks, undirected networks, smooth costs, and allow the noise to have unbounded support.

Distributed optimization over random networks has been studied in Lobel and Ozdaglar (2011); Lobel et al. (2011); Jakovetic et al. (2014b). References Lobel and Ozdaglar (2011); Lobel et al. (2011) consider non-differentiable convex costs and no (sub)gradient noise, while reference Jakovetic et al. (2014b) considers differentiable costs with Lipschitz continuous and bounded gradients, and it also does not allow for gradient noise, i.e., it considers methods with exact (deterministic) gradients.

Finally, we review the class of works that are concerned with designing distributed methods that achieve communication efficiency, e.g., Tsianos et al. (2012, 2013); Jakovetic et al. (2016); Lan et al. (2017); Wang et al. (2016); Sahu et al. (2018e,d). In Wang et al. (2016), a data censoring method is employed in the context of distributed least squares estimation to reduce computational and communication costs. However, the communication savings in Wang et al. (2016) are a constant proportion with respect to a method which utilizes all communications at all times, thereby not improving the order of the convergence rate. References Tsianos et al. (2012, 2013); Jakovetic et al. (2016) also consider a different setup than we do here, namely they study distributed optimization where the data is available a priori (i.e., it is not streamed). This corresponds to an intrinsically different setting with respect to the one studied here, where actually geometric MSE convergence rates are attainable with stochastic-type methods, e.g., Mokhtari and Ribeiro (2016). In terms of the strategy to save communications, references Tsianos et al. (2012, 2013); Jakovetic et al. (2016); Lan et al. (2017) consider, respectively, deterministically increasingly sparse communication, an adaptive communication scheme, and selective activation of agents. These strategies are different from ours; we utilize randomized, increasingly sparse communications in general. In references Sahu et al. (2018e,d), we study distributed estimation problems and develop communication-efficient distributed estimators. The problems studied in Sahu et al. (2018e,d) have a major difference with respect to the current chapter in that, in Sahu et al. (2018e,d), the assumed setting yields individual nodes' local gradients to evaluate to zero at the global solution. In contrast, the model assumed here does not feature such property, and hence it is

more challenging.

8.3 Problem Setup

The network of N agents in our setup collaboratively aim to solve the following unconstrained problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{N} f_i(\mathbf{x}), \tag{8.1}$$

where $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex function available to node i, $i = 1, ..., N$. We make the following assumption on the functions $f_i(\cdot)$:

Assumption 8.2.1. For all $i = 1, ..., N$, function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is twice continuously differentiable with Lipschitz continuous gradients. In particular, there exist constants $L, \mu > 0$ such that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\mu\, \mathbf{I} \preceq \nabla^2 f_i(\mathbf{x}) \preceq L\mathbf{I}.$$

From Assumption 8.2.1 we have that each f_i, $i = 1, \cdots, N$, is strongly convex with modulus μ. Using standard properties of convex functions, we have for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \nabla f_i(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2,$$
$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|.$$

We consider distributed stochastic zeroth order optimization to solve (8.1) over random networks. Inter-agent communication is modeled by a sequence of independent and identically distributed (i.i.d.) undirected random networks: at each time instant $k = 0, 1, ...$, the underlying inter-agent communication network is denoted by $\mathcal{G}(k) = (V, \mathbf{E}(k))$, with $V = \{1, ..., N\}$ being the set of nodes and $\mathbf{E}(k)$ being the random set of undirected edges. The edge connecting node i and j is denoted as $\{i, j\}$. The time-varying random neighborhood of node i at time k (excluding node i) is represented as $\Omega_i(k) = \{j \in V : \{i, j\} \in \mathbf{E}(k)\}$. The graph Laplacian of the random graph $\mathcal{G}(k)$ at time k is given by $\mathbf{L}(k) \in \mathbb{R}^{N\times N}$, where $\mathbf{L}(k)$ is given by $\mathbf{L}_{ij}(k) = -1$, if $\{i, j\} \in \mathbf{E}(k)$, $i \neq j$; $\mathbf{L}_{ij}(k) = 0$, if $\{i, j\} \notin \mathbf{E}(k)$, $i \neq j$; and $\mathbf{L}_{ii}(k) = -\sum_{j\neq i} \mathbf{L}_{ij}(k)$. It is to be noted that the Laplacian at each time instant is symmetric and a positive semidefinite matrix. As the considered graph sequence is i.i.d., we have that $\mathbb{E}[\,\mathbf{L}(k)\,] = \overline{\mathbf{L}}$. Let the graph corresponding to $\overline{\mathbf{L}}$ be given by $\overline{\mathcal{G}} = (V, \overline{\mathbf{E}})$.
We make the following assumption on $\overline{\mathcal{G}}$.

Assumption 8.2.2. The inter-agent communication graph is connected on average, i.e., $\overline{\mathcal{G}}$ is connected. In other words, $\lambda_2(\overline{\mathcal{L}}) > 0$.

We denote by $|\mathcal{L}|$ the cardinality of a set of Laplacians chosen from the total number of possible Laplacians (necessarily finite) so as to ensure $\underline{p} = \inf_{\mathbf{L}\in\mathcal{L}} \mathbb{P}\,(\mathbf{L}(t) = \mathbf{L}) > 0$.

8.3.1 Gradient noise model and the algorithm

We consider the following distributed stochastic gradient method to solve (8.1). Each node i, $i = 1, ..., N$, maintains over time steps (iterations) $k = 0, 1, ...$, its solution estimate $\mathbf{x}_i(k) \in \mathbb{R}^m$. Specifically, for arbitrary

deterministic initial points $\mathbf{x}_i(0) \in \mathbb{R}^m$, $i = 1, ..., N$, the update rule at node i and $k = 0, 1, ...$, is as follows:

$$\begin{aligned} \mathbf{x}_i(k+1) &= \mathbf{x}_i(k) - \beta_k \sum_{j \in \Omega_i(k)} (\mathbf{x}_i(k) - \mathbf{x}_j(k)) \\ &- \alpha_k \left(\nabla f_i(\mathbf{x}_i(k)) + \mathbf{v}_i(k) \right). \end{aligned} \tag{8.2}$$

The update (8.2) is performed in parallel by all nodes $i = 1, ..., N$. The algorithm iteration is realized as follows. First, each node i broadcasts $\mathbf{x}_i(k)$ to all its available neighbors $j \in \Omega_i(k)$, and receives $\mathbf{x}_j(k)$ from all $j \in \Omega_i(k)$. Subsequently, each node i, $i = 1, ..., N$ makes update (8.2), which completes an iteration. In (8.2), α_k is the step-size that we set to $\alpha_k = \alpha_0/(k+1)$, $k = 0, 1, ...$, with $\alpha_0 > 0$; and β_k is the (possibly) time-varying weight that each node assigns to all its neighbors. We set $\beta_k = \beta_0/(k+1)^\nu$, $k = 0, 1, ...$, with $\nu \in [0, 1/2]$. Here, $\beta_0 > 0$ is a constant that should be taken to be sufficiently small; e.g., one can set $\beta_0 = 1/(1+\theta)$, where θ is the maximal degree (number of neighbors of a node) across network. Finally, $\mathbf{v}_i(k)$ is noise in the calculation of the f_i's gradient at iteration k.
For future reference, we also present algorithm (8.2) in matrix format. Denote by $\mathbf{x}(k) = \left[\mathbf{x}_1^\top(k), \cdots, \mathbf{x}_N^\top(k)\right]^\top \in \mathbb{R}^{Nm}$ the vector that stacks the solution estimates of all nodes. Also, define function $F : \mathbb{R}^{Nm} \mapsto \mathbb{R}$, by $F(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x}_i)$, with $\mathbf{x} = \left[\mathbf{x}_1^\top, \cdots, \mathbf{x}_N^\top\right]^\top \in \mathbb{R}^{Nm}$. Finally, let $\mathbf{W}_k = (\mathbf{I} - \mathbf{L}_k) \otimes \mathbf{I}_m$, where $\mathbf{L}_k = \beta_k \boldsymbol{\mathcal{L}}(k)$. Then, for $k = 0, 1, ...$, algorithm (8.2) can be compactly written as follows:

$$\mathbf{x}(k+1) = \mathbf{W}_k \mathbf{x}(k) - \alpha_k \left(\nabla F(\mathbf{x}(k)) + \mathbf{v}(k) \right). \tag{8.3}$$

We make the following standard assumption on the gradient noises. First, denote by $\mathcal{F}_k$ the history of algorithm (8.2) up to time k; that is, $\mathcal{F}_k$, $k = 1, 2, ...$, is an increasing sequence of sigma algebras, where $\mathcal{F}_k$ is the sigma algebra generated by the collection of random variables $\{ \boldsymbol{\mathcal{L}}(s), \mathbf{v}_i(t) \}$, $i = 1, ..., N$, $s = 0, ..., k-1$, $t = 0, ..., k-1$.

Assumption 8.3.2. For each $i = 1, ..., N$, the sequence of measurement noises $\{\widehat{\mathbf{v}}_i(k)\}$ satisfies for all $k = 0, 1, ...$:

$$\begin{aligned} &\mathbb{E}[\, \widehat{\mathbf{v}}_i(k) \,|\, \mathcal{F}_k \,] = 0, \text{ almost surely (a.s.)} \\ &\mathbb{E}[\, \|\widehat{\mathbf{v}}_i(k)\|^2 \,|\, \mathcal{F}_k \,] \leq c_f \|\mathbf{x}_i(k)\|^2 + \sigma^2, \text{ a.s.,} \end{aligned} \tag{8.4}$$

where c_f and σ^2 are nonnegative constants.

It is to be noted that assumption 8.3.2 is trivially satisfied, when $\{\mathbf{v}_i(k)\}$ is an i.i.d. zero-mean, finite second moment, noise sequence such that $\mathbf{v}_i(k)$ is also independent of the history $\mathcal{F}_k$. However, the assumption allows the noise to be dependent on the current iterate at all times.

8.3.2 A machine learning motivation

The optimization-algorithmic model defined by Assumptions 8.2.1 and 8.3.2 subsumes, e.g., important machine learning applications. Consider the scenario where f_i corresponds to the risk function associated with the node i's local data, i.e.,

$$f_i(\mathbf{x}) = \mathbb{E}_{\mathbf{d}_i \sim P_i} \left[\ell_i(\mathbf{x}; \mathbf{d}_i) \right] + \Psi_i(\mathbf{x}). \tag{8.5}$$

Here, P_i is node i's local distribution according to which its data samples $\mathbf{d}_i \in \mathbb{R}^q$ are generated; $\ell_i(\cdot;\cdot)$ is a loss function that is convex in its first argument for any fixed value of its second argument; and $\Psi : \mathbb{R}^m \to \mathbb{R}$

is a strongly convex regularizer. Similarly, f_i can be an empirical risk function:

$$f_i(\mathbf{x}) = \frac{1}{n_i}\left(\sum_{j=1}^{n_i} \ell_i\left(\mathbf{x}; \mathbf{d}_{i,j}\right)\right) + \Psi_i(\mathbf{x}), \tag{8.6}$$

where $\mathbf{d}_{i,j}$, $j = 1, ..., n_i$, is the set of training examples at node i. Examples for the loss $\ell_i(\cdot;\cdot)$ include the following:

$$\begin{aligned} \ell_i(\mathbf{x}; \mathbf{a_i}, b_i) &= \frac{1}{2}\left(\mathbf{a}_i^\top \mathbf{x} - b_i\right)^2 \quad \text{(quadratic loss)} \\ \ell_i(\mathbf{x}; \mathbf{a_i}, b_i) &= \ln\left(1 + \exp(-b_i(\mathbf{a}_i^\top \mathbf{x}))\right) \quad \text{(logistic loss)} \end{aligned} \tag{8.7}$$

For the quadratic loss above, a data sample $\mathbf{d}_i = (\mathbf{a}_i, b_i)$, where $\mathbf{a}_i$ is a regressor vector and b_i is a response variable; for the logistic loss, $\mathbf{a}_i$ is a feature vector and $b_i \in \{-1, +1\}$ is its class label. Clearly, both the risk (8.5) and the empirical risk (8.6) satisfy Assumption 8.2.1 for the losses in (8.7).

We next discuss the search directions in (8.2) and Assumption 8.3.2 for the gradient noise. A common search direction in machine learning algorithms is the gradient of the loss with respect to a single data point[1]:

$$g_i(\mathbf{x}) = \nabla \ell_i\left(\mathbf{x}; \mathbf{d}_i\right) + \nabla \Psi_i(\mathbf{x}).$$

In case of the risk function (8.7), $\mathbf{d}_i$ is drawn from distribution P_i; in case of the empirical risk (8.6), $\mathbf{d}_i$ can be, e.g., drawn uniformly at random from the set of data points $\mathbf{d}_{i,j}$, $j = 1, ..., n_i$, with repetition along iterations. In both cases, gradient noise $\mathbf{v}_i = g_i(\mathbf{x}) - \nabla f_i(\mathbf{x})$ clearly satisfies assumption (8.3.2). To see this, consider, for example, the risk function (8.5), and let us fix iteration k and node i's estimate $\mathbf{x}_i(k) = \mathbf{x}_i$. Then,

$$\begin{aligned} & \mathbb{E}\left[\mathbf{v}_i(k) \,|\, \mathcal{F}_k\right] = \mathbb{E}\left[\mathbf{g}_i(k) - \nabla f_i(\mathbf{x}_i(k)) \,|\, \mathbf{x}_i(k) = \mathbf{x}_i\right] \\ = \;& \mathbb{E}[\nabla \ell_i\left(\mathbf{x}_i(k); \mathbf{d}_i\right) \,|\, \mathbf{x}_i(k) = \mathbf{x}_i] + \nabla \Psi_i(\mathbf{x}_i) \\ - \;& (\nabla f_i(\mathbf{x}_i) + \nabla \Psi_i(\mathbf{x})) \\ = \;& \mathbb{E}_{\mathbf{d}_i \sim P_i}[\nabla \ell_i\left(\mathbf{x}; \mathbf{d}_i\right)] + \nabla \Psi_i(\mathbf{x}_i) \\ - \;& (\mathbb{E}_{\mathbf{d}_i \sim P_i}[\nabla \ell_i\left(\mathbf{x}; \mathbf{d}_i\right)] + \nabla \Psi_i(\mathbf{x}_i)) = 0. \end{aligned}$$

Further, for the empirical risk, assumption (8.3.2) holds trivially. For the risk function (8.5), assumption (8.3.2) holds for a sufficiently "regular" distribution P_i. For instance, it is easy to show that the assumption holds for the logistic loss in (8.7) when P_i has finite second moment, while it holds for the square loss in (8.7) when P_i has finite fourth moment.

Note that our setting allows that the data generated at different nodes be generated through different distributions P_i, as well as that the nodes utilize different losses ℓ_i's and regularizers Ψ_i's. Mathematically, this means that $\nabla f_i(x^\star) \neq 0$, in general. In words, if a node i relies only on its local data $\mathbf{d}_i$, it cannot recover the true solution $x^\star$. Nodes then engage in a collaborative algorithm (8.2) through which, as shown ahead, they can recover the global solution $x^\star$.

[1] Similar considerations hold for a loss with respect to a mini-batch of data points; this discussion is abstracted for simplicity.

8.4 Performance Analysis

8.4.1 Statement of main results and auxiliary lemmas

We are now ready to state our main result.

Theorem 8.4.1. *Consider algorithm (8.2) with step-sizes* $\alpha_k = \frac{\alpha_0}{k+1}$ *and* $\beta_k = \frac{\beta_0}{(k+1)^\nu}$, *where* $\beta_0 > 0$, $\alpha_0 > 2\,N/\mu$, *and* $\nu \in [0, 1/2]$. *Then, for each node* i*'s solution estimate* $\mathbf{x}_i(k)$ *and the solution* $\mathbf{x}^\star$ *of problem (8.1), there holds:*

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O(1/k).$$

We remark that the condition $\alpha_0 > 2\,N/\mu$ can be relaxed to require only a positive α_0, in which case the rate becomes $O(\ln(k)/k)$, instead of $O(1/k)$.[2] Also, to avoid large step-sizes at initial iterations for a large α_0, step-size α_k can be modified to $\alpha_k = \alpha_0/(k + k_0)$, for arbitrary positive constant k_0, and Theorem 8.4.1 continues to hold. Theorem 8.4.1 establishes the $O(1/k)$ MSE rate of convergence of algorithm (8.2); due to the assumed f_i's strong convexity, the theorem also implies that $\mathbb{E}\left[f(\mathbf{x}_i(k)) - f(\mathbf{x}^\star)\right] = O(1/k)$. Note that the expectation in Theorem 8.4.1 is both with respect to randomness in gradient noises and with respect to the randomness in the underlying network. The $O(1/k)$ rate does not depend on the statistics of the underlying random network, as long as the network is connected on average (i.e., satisfies Assumption 8.2.2.) The hidden constant depends on the underlying network statistics, but simulation examples suggest that the dependence is usually not strong (see Section 4).

Proof strategy and auxiliary lemmas. Our strategy for proving Theorem 8.4.1 is as follows. We first establish the mean square boundedness (uniform in k) of the iterates $\mathbf{x}_i(k)$, which also implies the uniform mean square boundedness of the gradients $\nabla f_i(\mathbf{x}_i(k))$ (Subsection 3-B). We then bound, in the mean square sense, the disagreements of different nodes' estimates, i.e., quantities $(\mathbf{x}_i(k) - \mathbf{x}_j(k))$, showing that $\mathbb{E}[\,\|\mathbf{x}_i(k) - \mathbf{x}_j(k)\|^2\,] = O(1/k)$ (Subsection 3-C). This allows us to show that the (hypothetical) global average of the nodes' solution estimates $\overline{\mathbf{x}}(k) := \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i(k)$ evolves according to a stochastic gradient method with the gradient estimates that have a sufficiently small bias and finite second moment. This allows us to show the $O(1/k)$ rate on the mean square error at the global average, which in turn allows to derive a similar bound at the individual nodes' estimates (Subsection 3-D).

In completing the strategy above, we make use of the following Lemma; the Lemma is a minor modification of Lemmas 4 and 5 in Kar and Moura (2011).

Lemma 8.4.2. *Let* $z(k)$ *be a nonnegative (deterministic) sequence satisfying:*

$$z(k+1) \leq (1 - r_1(k))\, z_1(k) + r_2(k),$$

where $\{r_1(k)\}$ *and* $\{r_2(k)\}$ *are deterministic sequences with*

$$\frac{a_1}{(k+1)^{\delta_1}} \leq r_1(k) \leq 1 \;\text{ and }\; r_2(k) \leq \frac{a_2}{(k+1)^{\delta_2}},$$

with $a_1, a_2, \delta_1, \delta_2 > 0$. *Then, (a) if* $\delta_1 = \delta_2 = 1$, *there holds:* $z(k) = O(1)$; *(b) if* $\delta_1 = 1/2$ *and* $\delta_2 = 3/2$, *then* $z(k) = O(1/k)$; *and (c) if* $\delta_1 = 1$, $\delta_2 = 2$, *and* $a_1 > 1$, *then* $z(k) = O(1/k)$.

[2] This subtlety comes from equation (32) ahead and the requirement that $c_{20} > 1$. If $c_{20} \leq 1$, it can be shown that in (8.32) the right hand side modifies to a $O(\ln(k)/k)$ quantity.

Subsequent analysis in Subsections 3-b until 3-d restricts to the case when $\nu = 1/2$, i.e., when consensus weights equal $\beta_k = \frac{\beta_0}{(k+1)^{1/2}}$. That is, for simplicity of presentation, we prove Theorem 8.4.1 for case $\nu = 1/2$. As it can be verified in subsequent analysis, the proof of Theorem 8.4.1 extends to a generic $\mu \in [0, 1/2)$ as well. As another step in simplifying notations, throughout Subsections 3-b and 3-c, we let $m = 1$ to avoid extensive usage of Kronecker products; again, the proofs extend to a generic $m > 1$.

8.4.2 Mean square boundedness of the iterates

This Subsection shows the uniform mean square boundedness of the algorithm iterates and the gradients evaluated at the algorithm iterates.

Lemma 8.4.3. *Consider algorithm (8.2), and let Assumptions 1-3 hold. Then, there exist nonnegative constants c_x and $c_{\partial f}$ such that, for all $k = 0, 1, ...$, there holds:*

$$\mathbb{E}[\,\|\mathbf{x}(k)\|^2\,] \leq c_x \quad \text{and} \quad \mathbb{E}[\,\|\nabla F(\mathbf{x}(k))\|^2\,] \leq c_{\partial f}.$$

Proof.
Denote by $\mathbf{x}^o = x^* \mathbf{1}_N$ and recall (8.3). Then, we have:

$$\begin{aligned} \mathbf{x}(k+1) - \mathbf{x}^o &= \mathbf{W}_k(\mathbf{x}(k) - \mathbf{x}^o) \\ &- \alpha_k \left(\nabla F(\mathbf{x}(k)) - \nabla F(\mathbf{x}^o)\right) \\ &- \alpha_k \mathbf{v}(k) - \alpha_k \nabla F(\mathbf{x}^o). \end{aligned} \tag{8.8}$$

By mean value theorem, we have:

$$\begin{aligned} & \nabla F(\mathbf{x}(k)) - \nabla F(\mathbf{x}^o) \\ = & \left[\int_{s=0}^{1} \nabla^2 F\left(\mathbf{x}^o + s(\mathbf{x}(k) - \mathbf{x}^o)\right) \, d\, s\right] (\mathbf{x}(k) - \mathbf{x}^o) \\ = & \mathbf{H}_k \left(\mathbf{x}(k) - \mathbf{x}^o\right). \end{aligned} \tag{8.9}$$

Note that $L\mathbf{I} \succcurlyeq \mathbf{H}_k \succcurlyeq \mu\mathbf{I}$. Using (8.9) in (8.8) we have:

$$\begin{aligned} & \mathbf{x}(k+1) - \mathbf{x}^o = \left(\mathbf{W}_k - \alpha_k \mathbf{H}_k\right) (\mathbf{x}(k) - \mathbf{x}^o) \\ - & \alpha_k \mathbf{v}(k) - \alpha_k \nabla F(\mathbf{x}^o). \end{aligned} \tag{8.10}$$

Denote by $\boldsymbol{\zeta}(k) = \mathbf{x}(k) - \mathbf{x}^o$ and by $\boldsymbol{\xi}(k) = \left(\mathbf{W}_k - \alpha_k \mathbf{H}_k\right) (\mathbf{x}(k) - \mathbf{x}^o) - \alpha_k \, \nabla F(\mathbf{x}^o)$. Then, there holds:

$$\begin{aligned} & \mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2 \,|\, \mathcal{F}_k\,] \leq \|\boldsymbol{\xi}(k)\|^2 \\ - & 2\alpha_k \, \boldsymbol{\xi}(k)^\top \mathbb{E}[\,\mathbf{v}(k) \,|\, \mathcal{F}_k\,] + \alpha_k^2 \, \mathbb{E}[\,\|\mathbf{v}(k)\|^2 \,|\, \mathcal{F}_k\,] \\ \leq & \|\boldsymbol{\xi}(k)\|^2 + N \, \alpha_k^2 \, (c_v \, \|\mathbf{x}(k)\|^2 + c_v'), \text{ a.s.}, \end{aligned} \tag{8.11}$$

where we used Assumption 8.3.2 and the fact that $\boldsymbol{\xi}(k)$ is measurable with respect to $\mathcal{F}_k$. We next bound $\|\boldsymbol{\xi}(k)\|^2$. Note that $\|\mathbf{W}_k - \alpha_k \, \boldsymbol{H}_k\| \leq 1 - \mu \, \alpha_k$ for sufficiently large k. Therefore, we have for sufficiently large k:

$$\|\boldsymbol{\xi}(k)\| \leq (1 - \mu \, \alpha_k) \, \|\boldsymbol{\zeta}(k)\| + \alpha_k \, \|\nabla F(\mathbf{x}^o)\|. \tag{8.12}$$

We now use the following inequality:

$$(a+b)^2 \leq (1+\theta)\, a^2 + \left(1+\frac{1}{\theta}\right) b^2, \tag{8.13}$$

for any $a, b \in \mathbb{R}$ and $\theta > 0$. We set $\theta = \frac{c_0}{k+1}$, with $c_0 > 0$. Using the inequality (8.13) in (8.12), we have:

$$\begin{aligned} \|\boldsymbol{\xi}(k)\|^2 \leq & \left(1+\frac{c_0}{k+1}\right)(1-\alpha_k\mu)^2 \\ \times \quad & \|\boldsymbol{\zeta}(k)\|^2 + \left(1+\frac{k+1}{c_0}\right)\alpha_k^2 \|\nabla F(\mathbf{x}^o)\|^2. \end{aligned}$$

Next, for $c_0 < \alpha_0\mu$, the last inequality implies:

$$\begin{aligned} \|\boldsymbol{\xi}(k)\|^2 \leq & \left(1-\frac{c_1}{k+1}\right)\|\boldsymbol{\zeta}(k)\|^2 \\ & + \frac{c_2}{k+1}\|\nabla F(\mathbf{x}^o)\|^2, \end{aligned} \tag{8.14}$$

for some constants $c_1, c_2 > 0$. Combining (8.14) and (8.11), we get:

$$\begin{aligned} \mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2 \,|\, \mathcal{F}_k\,] \quad \leq \quad & \left(1-\frac{c_1'}{k+1}\right)\|\boldsymbol{\zeta}(k)\|^2 \\ + \quad & \frac{c_2'}{k+1}, \end{aligned} \tag{8.15}$$

for some $c_1', c_2' > 0$. Taking expectation in (8.15) and applying Lemma 8.4.2, it follows that $\mathbb{E}[\,\|\boldsymbol{\zeta}(k)\|^2\,] = \mathbb{E}[\,\|\mathbf{x}(k)-\mathbf{x}^o\|^2\,]$ is uniformly (in k) bounded from above by a positive constant. It is easy to see that the latter implies that $\mathbb{E}[\,\|\mathbf{x}(k)\|^2\,]$ is also uniformly bounded. Using the Lipschitz continuity of ∇F, we finally also have that $\mathbb{E}[\,\|\nabla F(\mathbf{x}(k))\|^2\,]$ is also uniformly bounded. The proof of Lemma 8.4.3 is now complete.

8.4.3 Disagreement bounds

Recall the (hypothetically available) global average of nodes' estimates $\overline{\mathbf{x}}(k) = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i(k)$, and denote by $\widetilde{\mathbf{x}}_i(k) = \mathbf{x}_i(k) - \overline{\mathbf{x}}(k)$ the quantity that measures how far apart is node i's solution estimate from the global average. Introduce also vector $\widetilde{\mathbf{x}}(k) = (\,\widetilde{\mathbf{x}}_1(k), ..., \widetilde{\mathbf{x}}_N(k)\,)^\top$, and note that it can be represented as $\widetilde{\mathbf{x}}(k) = (\mathbf{I}-\mathbf{J})\,\mathbf{x}(k)$, where we recall $\mathbf{J} = \frac{1}{N}\mathbf{1}\mathbf{1}^\top$. We have the following Lemma.

Lemma 8.4.4. *Consider algorithm (8.2) under Assumptions 1–3. Then, there holds:*

$$\mathbb{E}[\,\|\widetilde{\mathbf{x}}(k)\|^2\,] = O(1/k).$$

As detailed in the next Subsection, Lemma 8.4.4 is important as it allows to sufficiently tightly bound the bias in the gradient estimates according to which the global average $\overline{\mathbf{x}}(k)$ evolves.

Proof. It is easy to show that the process $\{\widetilde{\mathbf{x}}(k)\}$ follows the recursion:

$$\widetilde{\mathbf{x}}(k+1) = \widetilde{\mathbf{W}}(k)\widetilde{\mathbf{x}}(k) - \alpha_k\,(\mathbf{I}-\mathbf{J})\,\underbrace{(\nabla F(\mathbf{x}(k)) + \mathbf{v}(k))}_{\mathbf{w}(k)}, \tag{8.16}$$

where $\widetilde{\mathbf{W}}(k) = \mathbf{W}(k) - \mathbf{J} = \mathbf{I} - \mathbf{L}(k) - \mathbf{J}$. Note that, $\mathbb{E}\left[\|\mathbf{w}(k)\|^2\right] \leq c_7 < \infty$, which follows due to the mean square boundedness of $\mathbf{x}(k)$ and $\nabla F(\mathbf{x}(k))$. Then, we have:

$$\|\widetilde{\mathbf{x}}(k+1)\| \leq \left\|\widetilde{\mathbf{W}}(k)\right\| \|\widetilde{\mathbf{x}}(k)\| + \alpha_k \|\mathbf{w}(k)\|.$$

We now invoke Lemma 4.4 in Kar et al. (2013b) to note that, after an appropriately chosen k_1, we have for $\forall k \geq k_1$,

$$\|\widetilde{\mathbf{x}}(k+1)\| \leq (1 - r(k)) \|\widetilde{\mathbf{x}}(k)\| + \alpha_k \|\mathbf{w}(k)\|, \tag{8.17}$$

with $r(k)$ being a $\mathcal{F}_k$-adapted process that satisfies $r(k) \in [0,1]$, a.s., and:

$$\mathbb{E}\left[r(k)|\mathcal{F}_k\right] \geq c_8\beta_k = \frac{c_9}{(k+1)^{\frac{1}{2}}} \ a.s., \tag{8.18}$$

for some constants $c_8, c_9 > 0$. Using (8.13) in (8.17), we have:

$$\begin{aligned} \|\widetilde{\mathbf{x}}(k+1)\|^2 &\leq (1+\theta_k)(1-r(k))^2 \|\widetilde{\mathbf{x}}(k)\|^2 \\ + &\left(1 + \frac{1}{\theta_k}\right) \alpha_k^2 \|\mathbf{w}(k)\|^2, \end{aligned}$$

for $\theta_k = \frac{c_{10}}{(k+1)^{\frac{1}{2}}}$. Then, we have:

$$\begin{aligned} &\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2 |\mathcal{F}_k\right] \leq (1+\theta_k)\left(1 - \frac{c_9}{(k+1)^{\frac{1}{2}}}\right)^2 \|\widetilde{\mathbf{x}}(k)\|^2 \\ &+ \left(1 + \frac{1}{\theta_k}\right) \alpha_k^2 \mathbb{E}[\,\|\mathbf{w}(k)\|^2 \,|\, \mathcal{F}_k\,], \text{ a.s.} \end{aligned}$$

Next, for $c_{10} < c_9$ (c_{10} can be chosen freely), we have:

$$\begin{aligned} \mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] &\leq \left(1 - \frac{c_{11}}{(k+1)^{\frac{1}{2}}}\right) \mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right] \\ &+ \frac{c_{12}}{(k+1)^{\frac{3}{2}}} \end{aligned} \tag{8.19}$$

Utilizing Lemma 8.4.2, inequality (8.19) finally yields $\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] = O\left(\frac{1}{k}\right)$. The proof of the Lemma is complete.

8.4.4 Proof of Theorem 8.4.1

We are now ready to prove Theorem 8.4.1.

Proof.

Consider global average $\overline{\mathbf{x}}(k) = \frac{1}{N}\sum_{n=1} \mathbf{x}_i(k)$. From (8.16), we have:

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \alpha_k \left[\frac{1}{N}\sum_{i=1}^{N} \nabla f_i\left(\mathbf{x}_i(k)\right) + \underbrace{\frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i(k)}_{\overline{\mathbf{v}}(k)}\right]$$

which implies:

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{N}\left[\sum_{i=1}^{N} \nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right) + \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right] - \alpha_k \overline{\mathbf{v}}(k).$$

Recall $f(\cdot) = \sum_{i=1}^{N} f_i(\cdot)$. Then, we have:

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{N}\nabla f\left(\overline{\mathbf{x}}(k)\right) - \frac{\alpha_k}{N}\left[\sum_{i=1}^{N} \nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right] - \alpha_k \overline{\mathbf{v}}(k), \tag{8.20}$$

which implies:

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{N}\left[\nabla f\left(\overline{\mathbf{x}}(k)\right) + \mathbf{e}(k)\right], \tag{8.21}$$

where

$$\mathbf{e}(k) = N\overline{\mathbf{v}}(k) + \underbrace{\sum_{i=1}^{N}\left(\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right)}_{\boldsymbol{\epsilon}(k)}. \tag{8.22}$$

Note that, $\left\|\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right\| \leq L\left\|\mathbf{x}_i(k) - \overline{\mathbf{x}}(k)\right\| = L\left\|\widetilde{\mathbf{x}}_i(k)\right\|$. Thus, we can conclude for

$$\boldsymbol{\epsilon}(k) = \sum_{i=1}^{N}\left(\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right)$$

the following:

$$\mathbb{E}\left[\left\|\boldsymbol{\epsilon}(k)\right\|^2\right] \leq \frac{c_{15}}{(k+1)}. \tag{8.23}$$

Note here that (8.21) is an inexact gradient method for minimizing f with step size α_k/N and the random gradient error $\mathbf{e}(k) = N\overline{\mathbf{v}}(k) + \boldsymbol{\epsilon}(k)$. The term $N\overline{\mathbf{v}}(k)$ is zero-mean, while the gradient estimate bias is induced by $\boldsymbol{\epsilon}(k)$; as per (8.23), the bias is at most $O(1/k)$ in the mean square sense.
With the above development in place, we rewrite (8.20) as follows:

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{N}\nabla f\left(\overline{\mathbf{x}}(k)\right) - \frac{\alpha_k}{N}\boldsymbol{\epsilon}(k) - \alpha_k\overline{\mathbf{v}}(k). \tag{8.24}$$

This implies, recalling that $\mathbf{x}^\star$ is the solution to (8.1):

$$\overline{\mathbf{x}}(k+1) - \mathbf{x}^\star = \overline{\mathbf{x}}(k) - \mathbf{x}^\star \tag{8.25}$$

$$- \frac{\alpha_k}{N}\left[\nabla f\left(\overline{\mathbf{x}}(k)\right) - \underbrace{\nabla f\left(\mathbf{x}^\star\right)}_{=0}\right] - \frac{\alpha_k}{N}\boldsymbol{\epsilon}(k) - \alpha_k\overline{\mathbf{v}}(k). \tag{8.26}$$

By the mean value theorem, we have:

$$\nabla f\left(\overline{\mathbf{x}}(k)\right) - \nabla f\left(\mathbf{x}^\star\right) = \underbrace{\left[\int_{s=0}^{1} \nabla^2 f\left(\mathbf{x}^\star + s\left(\overline{\mathbf{x}}(k) - \mathbf{x}^\star\right)\right)\right] d\, s}_{\overline{\mathbf{H}}_k}$$
$$\times \left(\overline{\mathbf{x}}(k) - \mathbf{x}^\star\right), \tag{8.27}$$

where it is to be noted that $NL \succcurlyeq \overline{\mathbf{H}}_k \succcurlyeq N\mu$. Using (8.27) in (8.24), we have:

$$\left(\overline{\mathbf{x}}(k+1) - \mathbf{x}^\star\right) = \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right]\left(\overline{\mathbf{x}}(k) - \mathbf{x}^\star\right) \tag{8.28}$$
$$- \frac{\alpha_k}{N}\boldsymbol{\epsilon}(k) - \alpha_k \overline{\mathbf{v}}(k).$$

Denote by $\mathbf{m}(k) = \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right]\left(\overline{\mathbf{x}}(k) - \mathbf{x}^\star\right) - \frac{\alpha_k}{N}\boldsymbol{\epsilon}(k)$. Then, (8.28) is rewritten as:

$$\left(\overline{\mathbf{x}}(k+1) - \mathbf{x}^\star\right) = \mathbf{m}(k) - \alpha_k \overline{\mathbf{v}}(k), \tag{8.29}$$

and so:

$$\left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^\star\right\|^2 \leq \left\|\mathbf{m}(k)\right\|^2 - 2\alpha_k \mathbf{m}(k)^\top \overline{\mathbf{v}}(k)$$
$$+ \alpha_k^2 \left\|\overline{\mathbf{v}}(k)\right\|^2.$$

The latter inequality implies:

$$\mathbb{E}[\left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^\star\right\|^2 \mid \mathcal{F}_k] \leq \left\|\mathbf{m}(k)\right\|^2$$
$$- 2\alpha_k \mathbf{m}(k)^\top \mathbb{E}[\,\overline{\mathbf{v}}(k) \mid \mathcal{F}_k] + \alpha_k^2 \mathbb{E}[\left\|\overline{\mathbf{v}}(k)\right\|^2 \mid \mathcal{F}_k], \text{ a.s.}$$

Taking expectation, using the fact that $\mathbb{E}[\,\overline{\mathbf{v}}(k) \mid \mathcal{F}_k] = 0$, Assumption 8.3.2, and Lemma 8.4.3, we obtain:

$$\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^\star\right\|^2\right] \leq \mathbb{E}\left[\left\|\mathbf{m}(k)\right\|^2\right] + \frac{c_{17}}{(k+1)^2}, \tag{8.30}$$

for some constant $c_{17} > 0$. Next, using (8.13), we have for $\mathbf{m}(k)$ the following:

$$\left\|\mathbf{m}(k)\right\|^2 \leq (1+\theta_k)\left\|\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right\|^2 \left\|\overline{\mathbf{x}}(k) - \mathbf{x}^\star\right\|^2$$
$$+ \left(1 + \frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2}\left\|\boldsymbol{\epsilon}(k)\right\|^2$$
$$\leq (1+\theta_k)\left(1 - c_{18}\alpha_k\right)^2 \left\|\overline{\mathbf{x}}(k) - \mathbf{x}^\star\right\|^2$$
$$+ \left(1 + \frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2}\left\|\boldsymbol{\epsilon}(k)\right\|^2,$$

with $c_{18} = \mu/N$, because $\mu\,\mathbf{I} \preceq \overline{\mathbf{H}}_k \preceq L\,\mathbf{I}$. After choosing $\theta_k = \frac{c_{19}}{(k+1)}$ such that $c_{19} < \alpha_0\, c_{18}/2 = \alpha_0\,\mu/(2N)$ and after taking expectation, we obtain:

$$\mathbb{E}[\left\|\mathbf{m}(k)\right\|^2] \leq \left(1 - \frac{c_{20}}{k+1}\right)\mathbb{E}[\left\|\overline{\mathbf{x}}(k) - \mathbf{x}^\star\right\|^2] + \frac{c_{21}}{(k+1)^2}, \tag{8.31}$$

where $c_{20} > \alpha_0 \mu/(2N) > 1$ (because $\alpha_0 > 2N/\mu$) and c_{21} is a positive constant. Combining (8.31) and (8.30), we get:

$$\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^\star\|^2\right] \leq \left(1 - \frac{c_{20}}{k+1}\right) \|\overline{\mathbf{x}}(k) - \mathbf{x}^\star\|^2 + \frac{c_{21}}{(k+1)^2} + \frac{c_{17}}{(k+1)^2}.$$

Invoking Lemma 8.4.2, the latter inequality implies:

$$\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^\star\|^2\right] \leq \frac{c_{22}}{(k+1)}, \tag{8.32}$$

for some constant $c_{22} > 0$. Therefore, for the global average $\overline{\mathbf{x}}(k)$, we have obtained the mean square rate $O\left(\frac{1}{k}\right)$. Finally, we note that,

$$\|\mathbf{x}_i(k) - \mathbf{x}^\star\| \leq \|\overline{\mathbf{x}}(k) - \mathbf{x}^\star\| + \left\|\underbrace{\mathbf{x}_i(k) - \overline{\mathbf{x}}(k)}_{\widetilde{\mathbf{x}}_i(k)}\right\|. \tag{8.33}$$

After using:

$$\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2 \leq 2\|\widetilde{\mathbf{x}}_i(k)\|^2 + 2\|\overline{\mathbf{x}}(k) - \mathbf{x}^\star\|^2,$$

and taking expectation, it follows that $\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O\left(\frac{1}{k}\right)$, for all $i = 1, ..., N$. The proof is complete.

8.5 Communication Efficient Distributed Stochastic Optimization

In this section, we develop the communication efficient distributed stochastic optimization. We consider distributed stochastic gradient methods to solve (8.1). That is, we study algorithms of the following form:

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k) - \sum_{j \in \Omega_i(k)} \gamma_{i,j}(k) \left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right) - \alpha_k \widehat{\mathbf{g}}_i(\mathbf{x}_i(k)), \tag{8.34}$$

where the weight assigned to an incoming message $\gamma_{i,j}(k)$ and the neighborhood of an agent $\Omega_i(k)$ are determined by the specific instance of the designated communication protocol. The approximated gradient $\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))$ is specific to the optimization, i.e., whether it is a zeroth order optimization or a first order optimization scheme. Technically speaking, a first order optimization scheme approximates the gradient as an unbiased estimate of the gradient. In the case of first order optimization, the agents query a stochastic first order oracle ($\mathcal{SFO}$) and receive unbiased estimates of the gradient. In subsequent sections, we will explore the gradient approximations in greater detail. Before stating the algorithms, we first discuss the communication scheme. Specifically, we adopt the following model.

8.5.1 Communication Scheme

The inter-node communication network to which the information exchange between nodes conforms to is modeled as an *undirected* simple connected graph $G = (V, E)$, with $V = [1 \cdots N]$ and E denoting the set of nodes and communication links. The neighborhood of node n is given by $\Omega_n = \{l \in V \,|\, (n, l) \in E\}$.

The node n has degree $d_n = |\Omega_n|$. The structure of the graph is described by the $N \times N$ adjacency matrix, $\mathbf{A} = \mathbf{A}^\top = [\mathbf{A}_{nl}]$, $\mathbf{A}_{nl} = 1$, if $(n,l) \in E$, $\mathbf{A}_{nl} = 0$, otherwise. The graph Laplacian $\overline{\mathbf{L}} = \mathbf{D} - \mathbf{A}$ is positive semidefinite, with eigenvalues ordered as $0 = \lambda_1(\overline{\mathbf{L}}) \leq \lambda_2(\overline{\mathbf{L}}) \leq \cdots \leq \lambda_N(\overline{\mathbf{L}})$, where $\mathbf{D}$ is given by $\mathbf{D} = \text{diag}\,(d_1 \cdots d_N)$. Thus, $\overline{\mathbf{L}}$ corresponds to the maximal graph, i.e., the graph of all *allowable* communications. We now describe our randomized communication protocol that selects a (random) subset of the allowable links at each time instant for information exchange.
For each node i, at every time k, we introduce a binary random variable $\psi_{i,k}$, where

$$\psi_{i,k} = \begin{cases} \rho_k & \textit{with probability } \zeta_k \\ 0 & \textit{otherwise}, \end{cases} \tag{8.35}$$

where $\psi_{i,k}$'s are independent both across time and the nodes, i.e., across k and i respectively. The random variable $\psi_{i,k}$ abstracts out the decision of the node i at time k whether to participate in the neighborhood information exchange or not. We specifically take ρ_k and ζ_k of the form

$$\rho_k = \frac{\rho_0}{(k+1)^{\epsilon/2}}, \ \zeta_k = \frac{\zeta_0}{(k+1)^{(\tau/2-\epsilon/2)}}, \tag{8.36}$$

where $0 < \tau \leq \frac{1}{2}$ and $0 < \epsilon < \tau$. Furthermore, define β_k to be

$$\beta_k = (\rho_k \zeta_k)^2 = \frac{\beta_0}{(k+1)^\tau}, \tag{8.37}$$

where $\beta_0 = \rho_0^2 \zeta_0^2$. With the above development in place, we define the random time-varying Laplacian $\mathbf{L}(k)$, where $\mathbf{L}(k) \in \mathbb{R}^{N \times N}$ abstracts the inter-node information exchange as follows:

$$\mathbf{L}_{i,j}(k) = \begin{cases} -\psi_{i,k}\psi_{j,k} & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ \sum_{l \neq i} \psi_{i,k}\psi_{l,k} & i = j. \end{cases} \tag{8.38}$$

The above communication protocol allows two nodes to communicate only when the link is established in a bi-directional fashion and hence avoids directed graphs. The design of the communication protocol as depicted in (8.35)-(8.38) not only decays the weight assigned to the links over time but also decays the probability of the existence of a link. Such a design is consistent with frameworks where the agents have finite power and hence not only the number of communications, but also, the quality of the communication decays over time. We have, for $\{i,j\} \in E$ and $i \neq j$:

$$\begin{aligned} \mathbb{E}\left[\mathbf{L}_{i,j}(k)\right] &= -\left(\rho_k \zeta_k\right)^2 = -\beta_k = -\frac{\beta_0}{(k+1)^\tau} \\ \mathbb{E}\left[\mathbf{L}_{i,j}^2(k)\right] &= \left(\rho_k^2 \zeta_k\right)^2 = \frac{\rho_0^2 \beta_0}{(k+1)^{\tau+\epsilon}}. \end{aligned} \tag{8.39}$$

Thus, we have that, the variance of $\mathbf{L}_{i,j}(k)$ is given by,

$$Var\left(\mathbf{L}_{i,j}(k)\right) = \frac{\beta_0 \rho_0^2}{(k+1)^{\tau+\epsilon}} - \frac{\beta_0^2}{(k+1)^{2\tau}}. \tag{8.40}$$

Define, the mean of the random time-varying Laplacian sequence $\{\mathbf{L}(k)\}$ as $\overline{\mathbf{L}}_k = \mathbb{E}\left[\mathbf{L}(k)\right]$ and $\widetilde{\mathbf{L}}(k) = \mathbf{L}(k) - \overline{\mathbf{L}}_k$. Note that, $\mathbb{E}\left[\widetilde{\mathbf{L}}(k)\right] = \mathbf{0}$, and

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{L}}(k)\right\|^2\right] \leq 4N^2 \mathbb{E}\left[\widetilde{\mathbf{L}}_{i,j}^2(k)\right] = \frac{4N^2 \beta_0 \rho_0^2}{(k+1)^{\tau+\epsilon}} - \frac{4N^2 \beta_0^2}{(k+1)^{2\tau}}, \tag{8.41}$$

where $\|\cdot\|$ denotes the $\mathcal{L}_2$ norm. The above equation follows by relating the $\mathcal{L}_2$ and Frobenius norms. We also have that, $\overline{\mathbf{L}}_k = \beta_k \overline{\mathbf{L}}$, where

$$\overline{\mathbf{L}}_{i,j} = \begin{cases} -1 & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l \neq i} \overline{\mathbf{L}}_{i,l} & i = j. \end{cases} \tag{8.42}$$

Technically speaking, the communication graph at each time k encapsulated as $\mathbf{L}(k)$ need not be connected at all times, although the graph of allowable links G is connected.. In fact, at any given time k, only a few of the possible links could be active. However, since $\overline{\mathbf{L}}_k = \beta_k \overline{\mathbf{L}}$, we note that, by Assumption 8.2.2, the instantaneous Laplacian $\mathbf{L}(k)$ is connected on average.The connectedness in average basically ensures that over time, the information from each agent in the graph reaches other agents over time in a symmetric fashion and thus ensuring information flow, while providing the leeway for the instantaneous communication graphs at different times to be not connected.
We employ a primal algorithm for solving the optimization problem in (8.1). In particular, the update in (8.34) can then be written in a vector form as follows:

$$\mathbf{x}(k+1) = \mathbf{W}_k \mathbf{x}(k) - \alpha_k \widehat{\mathbf{G}}(\mathbf{x}(k)), \tag{8.43}$$

where $\mathbf{x}(k) = \left[\mathbf{x}_1^\top(k), \cdots, \mathbf{x}_N^\top(k)\right]^\top \in \mathbb{R}^{Nd}$, $F(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x}_i)$, $\mathbf{x} = \left[\mathbf{x}_1^\top, \cdots, \mathbf{x}_N^\top\right]^\top \in \mathbb{R}^{Nd}$, $\widehat{\mathbf{G}}(\mathbf{x}(k)) = [\widehat{\mathbf{g}}_i^\top(\mathbf{x}_i(k)), \cdots, \widehat{\mathbf{g}}^\top(\mathbf{x}_N(k))]^\top$ and $\mathbf{W}_k = (\mathbf{I} - \mathbf{L}(k)) \otimes \mathbf{I}_d$. We state an assumption on the weight sequences before proceeding further.

Assumption 8.5.1. The weight sequence α_k is given by $\alpha_0/(k+1)$, where $\alpha_0 > 1/\mu$. For the sequence ρ_k as defined in (8.36), it is chosen in such a way that,

$$\rho_0^2 \leq \frac{4N^2}{\lambda_2\left(\overline{\mathbf{L}}\right)}. \tag{8.44}$$

Communication Cost Define the communication cost $\mathcal{C}_k$ to be the expected per-node number of transmissions up to iteration k, i.e.,

$$\mathcal{C}_k = \mathbb{E}\left[\sum_{s=0}^{k-1} \mathbb{I}_{\{node\ C\ transmits\ at\ s\}}\right], \tag{8.45}$$

where $\mathbb{I}_A$ represents the indicator of event A. Note that the per-node communication cost in (8.45) is the same as the network average of communication costs across all nodes, as the activation probabilities are homogeneous across nodes. We now proceed to the main results pertaining to the proposed communication efficient first order optimization schemes.

8.6 Convergence rates: Statement of main results and interpretations

We state the main result concerning the mean square error at each agent i next, while the proof is relegated to Appendix G.

Theorem 8.6.1. *Consider algorithm (8.2) with step-sizes $\alpha_k = \frac{\alpha_0}{k+1}$ and $\beta_k = \frac{\beta_0}{(k+1)^{1/2}}$, where $\beta_0 > 0$ and $\alpha_0 > 2/\mu$.*

1) Then, for each node i's solution estimate $\mathbf{x}_i(k)$ and the solution $\mathbf{x}^\star$ of problem (8.1), , $\forall k \geq 0$ there holds:

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] \leq 2M_k + \frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)} + 2Q_k + \frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)}, \tag{8.46}$$

where, $\Delta_{1,\infty} = 2\|\nabla F(\mathbf{x}(k))\|^2 + 4c_u q_\infty(N,\alpha_0) + 4N\sigma_1^2$ and $q_\infty(N,\alpha_0) = \mathbb{E}\left[\|\mathbf{x}(k_2) - \mathbf{x}^o\|^2\right] + \frac{\pi^2}{6}\alpha_0^2\left(2c_u N\|\mathbf{x}^o\|^2 + N\sigma_u^2\right) + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}$, $k_2 = \max\{k_0, k_1\}$, $k_0 = \inf\{k|\mu^2\alpha_k^2 < 1\}$ and $k_1 = \inf\left\{k|\frac{\mu}{2} > 2c_u\alpha_k\right\}$, with M_k and Q_k decaying faster than the rest of the terms.
2) The communication cost is given by,

$$\mathbb{E}\left[\sum_{t=1}^{k}\zeta_t\right] = O\left(k^{\frac{3}{4}+\frac{\epsilon}{2}}\right),$$

leading to the following MSE-communication rate:

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O\left(\mathcal{C}_k^{-\frac{4}{3}+\zeta}\right), \tag{8.47}$$

where ζ can be arbitrarily small.

We remark that the condition $\alpha_0 > 2/\mu$ can be relaxed to require only a positive α_0, in which case the rate becomes $O(\ln(k)/k)$, instead of $O(1/k)$. Also, to avoid large step-sizes at initial iterations for a large α_0, step-size α_k can be modified to $\alpha_k = \alpha_0/(k+k_0)$, for arbitrary positive constant k_0, and Theorem 8.6.1 continues to hold. Theorem 8.6.1 establishes the $O(1/k)$ MSE rate of convergence of algorithm (8.2); due to the assumed f_i's strong convexity, the theorem also implies that $\mathbb{E}\left[f(\mathbf{x}_i(k)) - f(\mathbf{x}^\star)\right] = O(1/k)$.

8.7 Simulation example

We provide a simulation example on ℓ_2-regularized logistic losses and random networks where links fail independently over iterations and across different links, with probability p_{fail}. The simulation corroborates the derived $O(1/k)$ rate of algorithm (8.2) over random networks and shows that deterioration due to increase of p_{fail} is small.

We consider empirical risk minimization (8.6) with the logistic loss in (8.7) and the regularization functions set to $\Psi_i(\mathbf{x}) = \frac{\kappa}{2}\|\mathbf{x}\|^2$, $i = 1, ..., N$, where $\kappa > 0$ is the regularization parameter that is set to $\kappa = 0.5$.

The number of data points per node is $n_i = 10$. We generate the "true" classification vector $x' = ((\mathbf{x}_1')^\top, x_0')^\top$ by drawing its entries independently from standard normal distribution. Then, the class labels are generated as $b_{ij} = \text{sign}\left((\mathbf{x}_1')^\top\mathbf{a}_{i,j} + x_0' + \epsilon_{ij}\right)$, where ϵ_{ij}'s are drawn independently from normal distribution with zero mean and standard deviation 2. The feature vectors $\mathbf{a}_{i,j}$, $j = 1, ..., n_i$, at node i are generated as follows: each entry of each vector is a sum of a standard normal random variable and a uniform random variable with support $[0,\, 5\,i]$. Different entries within a feature vector are drawn independently, and also different vectors are drawn independently, both intra node and inter nodes. Note that the feature vectors at different nodes are drawn from different distributions.

The algorithm parameters are set as follows. We let $\beta_k = \frac{1}{\theta\,(k+1)^{1/2}}$, $\alpha_k = \frac{1}{k+1}$, $k = 0, 1, ...$ Here, θ is the maximal degree across all nodes in the network and here equals $\theta = 6$. Algorithm (8.2) is initialized with $\mathbf{x}_i(0) = 0$, for all $i = 1, ..., N$.

We consider a connected network $\mathcal{G}$ with $N = 10$ nodes and 23 links, generated as a random geometric graph: nodes are placed randomly (uniformly) on a unit square, and the node pairs whose distance is less than a radius are connected by an edge. We consider the random network model where each (undirected) link in network $\mathcal{G}$ fails independently across iterations and independently from other links with probability p_{fail}. We consider the cases $p_{\text{fail}} \in \{0; 0.5; 0.9\}$. Note that the case $p_{\text{fail}} = 0$ corresponds to network $\mathcal{G}$ with all its links always online, more precisely, with links failing with zero probability. Algorithm (8.2) is then run on each of the described network models, i.e., for each $p_{\text{fail}} \in \{0; 0.5; 0.9\}$. This allows us to assess how much the algorithm performance degrades with the increase of p_{fail}. We also include a comparison with the following centralized stochastic gradient method:

$$\mathbf{y}(k+1) = \mathbf{y}(k) - \frac{1}{N(k+1)} \sum_{i=1}^{N} \nabla \ell \left(\mathbf{y}(k); \mathbf{a}_i(k), b_i(k) \right), \tag{8.48}$$

where $(\mathbf{a}_i(k), b_i(k))$ is drawn uniformly from the set $(\mathbf{a}_{i,j}, b_{i,j})$, $j = 1, ..., n_i$. Note that algorithm (8.48) makes an unbiased estimate of $\sum_{i=1}^{N} \nabla f_i(\mathbf{y}(k))$ by drawing a sample uniformly at random from each node's data set. Algorithm (8.48) is an idealization of (8.2): it shows how (8.2) would be implemented if there existed a fusion node that had access to all nodes' data. Hence, the comparison with (8.48) allows us to examine how much the performance of (8.2) degrades due to lack of global information, i.e., due to the distributed nature of the considered problem. Note that step-size in (8.48) is set to $1/N(k+1)$ for a meaningful comparison with (8.2), as this is the step-size effectively utilized by the hypothetical global average of the nodes' iterates with (8.2). As an error metric, we use the mean square error (MSE) estimate averaged across nodes: $\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2$.

Figure 1 plots the estimated MSE, averaged across 100 algorithm runs, versus iteration number k for different values of parameter p_{fail} in $\log_{10}$-$\log_{10}$ scale. Note that here the slope of the plot curve corresponds to the sublinear rate of the method; e.g., the -1 slope corresponds to a $1/k$ rate. First, note from the Figure that, for any value of p_{fail}, algorithm (8.2) achieves on this example (at least) the $1/k$ rate, thus corroborating our theory. Next, note that the increase of the link failure probability only increases the constant in the MSE but does not affect the rate. (The curves that correspond to different values of p_{fail} are "vertically shifted.") Interestingly, the loss due to the increase of p_{fail} is small; e.g., the curves that correspond to $p_{\text{fail}} = 0.5$ and $p_{\text{fail}} = 0$ (no link failures) practically match. Figure 1 also shows the performance of the centralized method (8.48). We can see that, except for the initial few iterations, the distributed method (8.2) is very close in performance to the centralized method. In Figure 8.2, the test error of the communication efficient first order optimization scheme is compared with the test error of the benchmark scheme which refers to the optimization scheme with the communication graph abstracted by a static Laplacian in terms of iterations or equivalently the number of queries per agent to the stochastic first order oracle, i.e., gradient evaluations. Figure 8.3 demonstrates the superiority of the proposed communication efficient first order optimization scheme in terms of the test error versus communication cost as compared to the benchmark as predicted by Theorem 8.6.1. For example, at the same relative test error level, the proposed algorithm uses up to 3x less number of transmissions as compared to the benchmark scheme.

8.8 Contributions

- $O(1/k)$ **rate of decay**: We showed that, by carefully designing the consensus and the gradient weights (potentials), the considered distributed stochastic gradient algorithm achieves the order-optimal $O(1/k)$

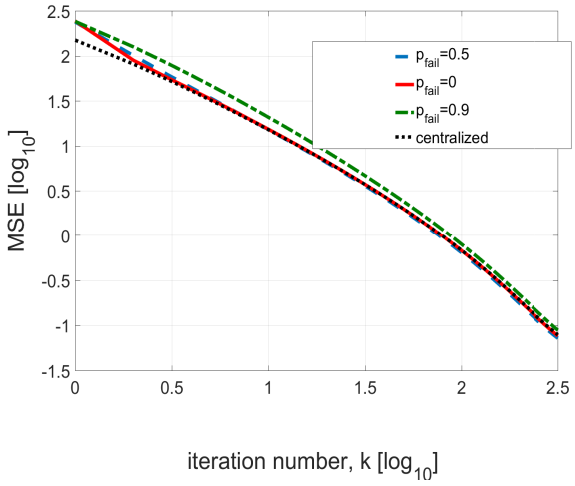


Figure 8.1: Estimated MSE versus iteration number k for algorithm (8.2) with link failure probability $p_{\text{fail}} = 0$ (red, solid line); 0.5 (blue, dashed line); and 0.9 (green, dash-dot line). The Figure also shows the performance of the centralized stochastic gradient method in (8.48) (black, dotted line).

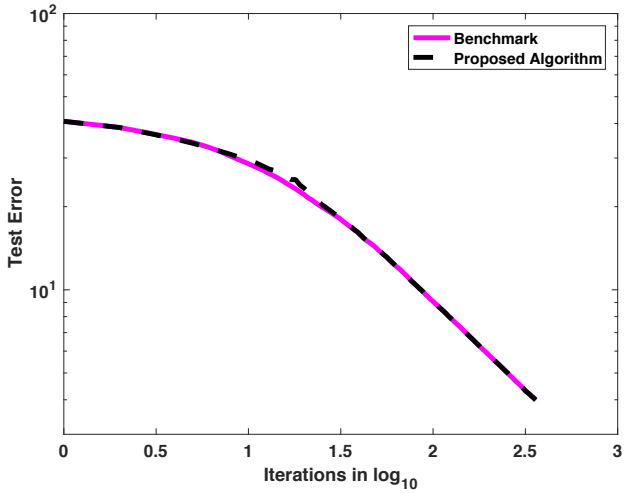


Figure 8.2: Communication Efficient 1st order Optimization: Test Error vs Iteration

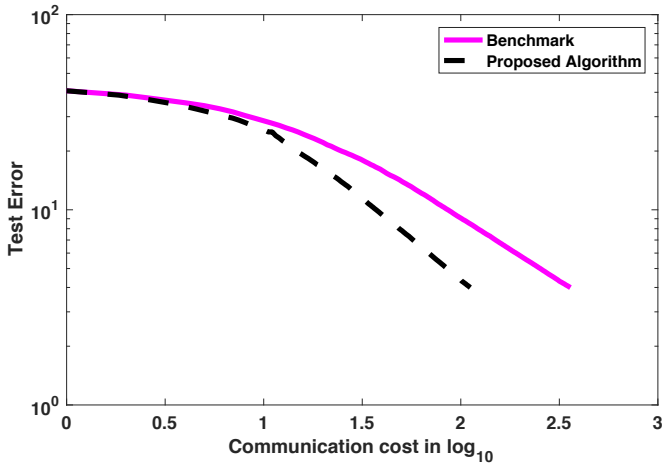


Figure 8.3: Communication Efficient 1st order Optimization: Test Error vs Communication Cost

rate of decay of the mean squared distance from the solution (mean squared error – MSE). This is achieved for twice continuously differentiable strongly convex local costs, assuming also that the noisy gradients are unbiased estimates of the true gradients and that the noise in gradients has bounded second moment. To the best of our knowledge, this is the first time an order-optimal convergence rate for distributed strongly convex stochastic optimization has been established for random networks.

- $O(1/(C_{\text{comm}})^{4/3-\zeta})$ **MSE communication rate**: We developed novel methods for first order distributed stochastic optimization, based on a probabilistic inter-agent communication protocol that increasingly sparsifies agent communications over time. For the first order distributed stochastic optimization, we propose a novel method that is shown to achieve the $O(1/(C_{\text{comm}})^{4/3-\zeta})$ MSE communication rate. At the same time, the proposed method retains the order-optimal $O(1/(C_{\text{comp}}))$ MSE rate in terms of the computational cost, the best achievable rate in the corresponding centralized setting.

8.9 Conclusion

In this chapter, we considered a distributed stochastic gradient method for smooth strongly convex optimization. Through the analysis of the considered method, we established for the first time the order optimal $O(1/k)$ MSE convergence rate for the assumed optimization setting when the underlying network is randomly varying. Furthermore, we have developed and analyzed a novel class of methods for distributed stochastic optimization of the first order that are based on increasingly sparse randomized communication protocols. We have established for the proposed first order method explicit mean square error (MSE) convergence rates with respect to (appropriately defined) computational cost C_{comp} and communication cost C_{comm}. The proposed first order method achieves the $O(1/(C_{\text{comm}})^{4/3-\zeta})$ MSE communication rate, while maintaining the order-optimal $O(1/(C_{\text{comp}}))$ MSE computational rate. Numerical examples on real data demonstrate the communication efficiency of the proposed methods.

Chapter 9

Communication-Efficient Stochastic Optimization: Zeroth Order

9.1 Introduction

Stochastic optimization has taken a central role in problems of learning and inference making over large data sets. Many practical setups are inherently distributed in which, due to sheer data size, it may not be feasible to store data in a single machine or agent. Further, due to the complexity of the objective functions (often, loss functions in the context of learning and inference problems), explicit computation of gradients or exactly evaluating the objective at desired arguments could be computationally prohibitive. The class of stochastic optimization problems of interest can be formalized in the following way:

$$\min f(\mathbf{x}) = \min \mathbb{E}_{\boldsymbol{\xi}\sim P}\left[F(\mathbf{x};\boldsymbol{\xi})\right],$$

where the information available to implement an optimization scheme usually involves gradients, i.e., $\nabla F(\mathbf{x};\boldsymbol{\xi})$ or function values of $F(\mathbf{x};\boldsymbol{\xi})$ itself. However, both the gradients and the function values are only unbiased estimates of the gradients and the function values of the desired objective $f(\mathbf{x})$. Moreover, due to huge data sizes and distributed applications, the data is often split across different agents, in which case the (global) objective reduces to the sum of N local objectives, $F(\mathbf{x};\boldsymbol{\xi}) = \sum_{i=1}^{N} F_i(\mathbf{x};\boldsymbol{\xi})$, where N denotes the number of agents. Such kind of scenarios are frequently encountered in setups such as empirical risk minimization in statistical learning Vapnik (1998). In order to address the aforementioned problem setup, we study zeroth distributed stochastic strongly convex optimization over networks.

There are N networked nodes, interconnected through a preassigned possibly sparse communication graph, that collaboratively aim to minimize the sum of their locally known strongly convex costs. We focus on zeroth and first order distributed stochastic optimization methods, where at each time instant (iteration) k, each node queries a stochastic zeroth order oracle ($\mathcal{SZO}$) for a noisy estimate of its local function's value at the current iterate (zeroth order optimization). In the proposed stochastic optimization methods, an agent updates its iterate at each iteration by simultaneously assimilating information obtained from the neighborhood (consensus) and the queried information from the relevant oracle (innovations). In the light of the aforementioned distributed protocol, our focus is then on examining the tradeoffs between the *communication cost*, measured by the number of per-node transmissions to their neighboring nodes in the network; and *computational cost*, measured by the number of queries made to $\mathcal{SZO}$ (zeroth order optimization).

Contributions. Our main contributions are as follows. We first analyze a distributed zeroth order optimiza-

tion scheme for strongly convex functions utilizing Kiefer Wolfowitz stochastic approximation. Furthermore, we develop novel methods for zeroth order distributed stochastic optimization, based on a probabilistic inter-agent communication protocol that increasingly sparsifies agent communications over time. For the proposed zeroth order method, we establish the $O(1/(C_{\text{comm}})^{8/9-\zeta})$ mean square error (MSE) convergence rate in terms of communication cost C_{comm}, where $\zeta > 0$ is arbitrarily small. At the same time, the method achieves the order-optimal $O(1/(C_{\text{comp}})^{2/3})$ MSE rate in terms of computational cost C_{comp} in the context of strongly convex functions with second order smoothness.

The achieved results reveal an interesting relation between the zeroth and first order distributed stochastic optimization. Namely, as we show here, the zeroth order method achieves a slower MSE communication rate than the first order method due to the (unavoidable) presence of bias in nodes' local functions' gradient estimation. Interestingly, increasing the degree of smoothness[1] p in cost functions coupled with a fine-tuned gradient estimation scheme, adapted to the smoothness degree, effectively reduces the bias and enables the zeroth order optimization mean square error to scale as $O(1/(C_{\text{comp}})^{(p-1)/p})$. Thus, with increased smoothness and appropriate gradient estimation schemes, the zeroth order optimization scheme gets increasingly close in mean square error of its first order counterpart. In a sense, we demonstrate that the first order (bias-free) stochastic optimization corresponds to the limiting case of the zeroth order stochastic optimization when $p \to \infty$.

In more detail, the proposed distributed communication efficient stochastic methods work as follows. They utilize an increasingly sparse communication protocol that we recently proposed in the context of distributed estimation problems Sahu et al. (2018e). Therein, at each time step (iteration) k, each node participates in the communication protocol with its immediate neighbors with a time-decreasing probability p_k. The probabilities of communicating are equal across all nodes, while the nodes' decisions whether to communicate or not are independent of the past and of the other nodes. Upon the transmission stage, if active, each node makes a weighted average of its own solution estimate and the solution estimates received from all of its communication-active (transmitting) neighbors, assigning to each neighbor a time-varying weight β_k. In conjunction with the averaging step, the nodes in parallel assimilate the obtained neighborhood information and the local information through a local gradient approximation step – based on the noisy functions estimates only – with step-size α_k.

By structure, the proposed distributed zeroth and first order stochastic methods are of a similar nature, expect for the fact that rather than approximating local gradients based on the noisy functions estimates in the zeroth order case, the first order setup assumes noisy gradient estimates are directly available.

Brief literature review. We now briefly review the literature. In the context of the extensive literature on distributed optimization, the most relevant to our work are the references that fall within the following three classes of works: 1) distributed strongly convex stochastic optimization; 2) distributed optimization over random networks (both deterministic and stochastic methods); and 3) distributed optimization methods that aim to improve communication efficiency.

9.2 Related Work

While we pursue stochastic optimization in this chapter, the case of deterministic noiseless distributed optimization has seen much progress (Boyd et al. (2011); Shi et al. (2015); Yuan et al. (2016, 2017)) and more recently accelerated methods (Jakovetic et al. (2014a); Xi and Khan (2017)). For the first class of works,

[1] Degree of smoothness p refers to the function under consideration being p-times continuously differentiable with the p-th order derivative being Lipschitz continuous.

several papers give explicit convergence rates in terms of the iteration counter k, that here translates into computational cost C_{comp} or equivalently number of queries to $\mathcal{SZO}$ or $\mathcal{SFO}$, under different assumptions. Regarding the underlying network, references Tsianos and Rabbat (2012); Towfic et al. (2016) consider static networks, while the works Yuan et al. (2018); Vanli et al. (2017); Nedic and Olshevsky (2016) consider deterministic time-varying networks. They all consider *first order* optimization.

References Tsianos and Rabbat (2012); Towfic et al. (2016) consider distributed first order strongly convex optimization for static networks, assuming that the data distributions that underlie each node's local cost function are equal (reference Tsianos and Rabbat (2012) considers empirical risks while reference Towfic et al. (2016) considers risk functions in the form of expectation); this essentially corresponds to each nodes' local function having the same minimizer. References Yuan et al. (2018); Vanli et al. (2017); Nedic and Olshevsky (2016) consider deterministically varying networks, assuming that the "union graph" over finite windows of iterations is connected. The papers Tsianos and Rabbat (2012); Towfic et al. (2016); Yuan et al. (2018); Vanli et al. (2017) assume undirected networks, while Nedic and Olshevsky (2016) allows for directed networks and assumes a bounded support for the gradient noise. The works Tsianos and Rabbat (2012); Yuan et al. (2018); Vanli et al. (2017); Nedic and Olshevsky (2016) allow the local costs to be non-smooth, while Towfic et al. (2016) assumes smooth costs, as we do here. With respect to these works, we consider random networks (that are undirected and connected on average), smooth costs, and allow the noise to have unbounded support. The authors of Hajinezhad et al. (2017) propose a distributed zeroth optimization algorithm for non-convex minimization with a static graph, where a random directions-random smoothing approach was employed.

For the second class of works, distributed optimization over random networks has been studied in Lobel and Ozdaglar (2011); Lobel et al. (2011); Jakovetic et al. (2014b). References Lobel and Ozdaglar (2011); Lobel et al. (2011) consider non-differentiable convex costs, first order methods, and no (sub)gradient noise, while reference Jakovetic et al. (2014b) considers differentiable costs with Lipschitz continuous and bounded gradients, first order methods, and it also does not allow for gradient noise, i.e., it considers methods with exact (deterministic) gradients. Reference Jakovetic et al. (2018) considers distributed stochastic first order methods and establishes the method's $O(1/k)$ convergence rate. References Sahu et al. (2018b) considers a zeroth order distributed stochastic approximation method, which queries the $\mathcal{SZO}$ $2d$ times at each iteration where d is the dimension of the optimizer and establishes the method's $O(1/k^{1/2})$ convergence rate in terms of the number of iterations under first order smoothness.

In summary, each of the references in the two classes above is not primarily concerned with studying communication rates of distributed stochastic methods. Prior work achieves order-optimal rates in terms of computational cost (that translates here into the number of iterations k), both for the zeroth order, e.g., Sahu et al. (2018b), and for the first order, e.g., Jakovetic et al. (2018), distributed strongly convex optimization.[2] In contrast, we establish here communication rates as well. This chapter and our prior works Sahu et al. (2018b,c) distinguish further from other works on distributed zeroth order optimization, e.g., Hajinezhad et al. (2017); Duchi et al. (2015), in that, not only the gradient is approximated through function values due to the absence of first order information, but also the function values themselves are subject to noise. Reference Sahu et al. (2018c) considers a communication efficient zeroth order approximation scheme, where the convergence rate is established to be $O(1/k^{1/2})$ and the MSE-communication is improved to $O(1/(C_{\text{comm}})^{2/3-\zeta})$. In contrast to Sahu et al. (2018c), with additional smoothness assumptions we improve the convergence rate to $O(1/k^{2/3})$ and the MSE-communication is further improved to $O(1/(C_{\text{comm}})^{8/9-\zeta})$.

[2] The works in the first two classes above utilize a non-diminishing amount of communications across iterations, and hence they achieve at best the $O(1/(C_{\text{comm}}))$ (first order optimization) and $O(1/(C_{\text{comm}})^{1/2})$ communication rates.

Finally, we review the class of works that are concerned with designing distributed methods that achieve communication efficiency, e.g., Tsianos et al. (2012, 2013); Jakovetic et al. (2016); Lan et al. (2017); Wang et al. (2016); Sahu et al. (2018e,d). In Wang et al. (2016), a data censoring method is employed in the context of distributed least squares estimation to reduce computational and communication costs. However, the communication savings in Wang et al. (2016) are a constant proportion with respect to a method which utilizes all communications at all times, thereby not improving the order of the convergence rate. References Tsianos et al. (2012, 2013); Jakovetic et al. (2016) also consider a different setup than we do here, namely they study distributed optimization where the data is available a priori (i.e., it is not streamed). This corresponds to an intrinsically different setting with respect to the one studied here, where actually geometric MSE convergence rates are attainable with stochastic-type methods, e.g., Mokhtari and Ribeiro (2016). In terms of the strategy to save communications, references Tsianos et al. (2012, 2013); Jakovetic et al. (2016); Lan et al. (2017) consider, respectively, deterministically increasingly sparse communication, an adaptive communication scheme, and selective activation of agents. These strategies are different from ours; we utilize randomized, increasingly sparse communications in general. In references Sahu et al. (2018e,d), we study distributed estimation problems and develop communication-efficient distributed estimators. The problems studied in Sahu et al. (2018e,d) have a major difference with respect to the current setup in that, in Sahu et al. (2018e,d), the assumed setting yields individual nodes' local gradients to evaluate to zero at the global solution. In contrast, the model assumed here does not feature such property, and hence it is more challenging.

Finally, we comment on the recent paper Lan et al. (2017) that develops communication-efficient distributed methods for both non-stochastic and stochastic distributed first order optimization, both in the presence and in the absence of the strong convexity assumption. For the stochastic, strongly convex first order optimization, Lan et al. (2017) shows that the method therein gets ϵ-close to the solution in $O(1/\sqrt{\epsilon})$ communications and with an $O(1/\epsilon)$ computational cost. The current setup has several differences with respect to Lan et al. (2017). First, reference Lan et al. (2017) does not study zeroth order optimization. Second, this work assumes for the gradient noise to be independent of the algorithm iterates. This is a strong assumption that may be not satisfied, e.g., with many machine learning applications. Third, while we assume here twice differentiable costs, this assumption is not imposed in Lan et al. (2017). Finally, the method in Lan et al. (2017) is considerably more complex than the one proposed here, with two layers of iterations (inner and outer iterations). In particular, the inner iterations involve solving an exact minimization problem which necessarily points to the usage of an off-the-shelf solver, the computation cost of which is not factored into the computation cost in Lan et al. (2017).

9.3 Model and the proposed algorithms

The network of N agents in our setup collaboratively aim to solve the following unconstrained problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{N} f_i(\mathbf{x}), \tag{9.1}$$

where $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is a strongly convex function available to node i, $i = 1, ..., N$. We make the following assumption on the functions $f_i(\cdot)$:

Assumption 9.3.1. For all $i = 1, ..., N$, function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is twice continuously differentiable with

Lipschitz continuous gradients. In particular, there exist constants $L, \mu > 0$ such that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\mu \, \mathbf{I} \preceq \nabla^2 f_i(\mathbf{x}) \preceq L \, \mathbf{I}.$$

From Assumption 9.3.1 we have that each f_i, $i = 1, \cdots, N$, is μ-strongly convex. Using standard properties of strongly convex functions, we have for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \nabla f_i(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2,$$
$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \, \|\mathbf{x} - \mathbf{y}\|.$$

We also have that from assumption 9.3.1, the optimization problem in (9.1) has a unique solution, which we denote by $\mathbf{x}^* \in \mathbb{R}^d$. Throughout the chapter, we use the sum function which is defined as $f : \mathbb{R}^d \to \mathbb{R}$, $f(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x})$. We consider distributed stochastic gradient methods to solve (9.1). That is, we study algorithms of the following form:

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k) - \sum_{j \in \Omega_i(k)} \gamma_{i,j}(k) \left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right)$$
$$- \alpha_k \widehat{\mathbf{g}}_i(\mathbf{x}_i(k)), \tag{9.2}$$

where the weight assigned to an incoming message $\gamma_{i,j}(k)$ and the neighborhood of an agent $\Omega_i(k)$ are determined by the specific instance of the designated communication protocol. The approximated gradient $\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))$ is specific to the optimization, i.e., whether it is a zeroth order optimization or a first order optimization scheme. Technically speaking, as we will see later, a zeroth order optimization scheme approximates the gradient as a biased estimate of the gradient while a first order optimization scheme approximates the gradient as an unbiased estimate of the gradient. The variation in the gradient approximation across first order and zeroth order methods can be attributed to the fact that the oracles from which the agents query for information pertaining to the loss function differ. For instance, in the case of the zeroth order optimization, the agents query a stochastic zeroth order oracle ($\mathcal{SZO}$) and in turn receive noisy function values (unbiased estimates) for the queried point. However, in the case of first order optimization, the agents query a stochastic first order oracle ($\mathcal{SFO}$) and receive unbiased estimates of the gradient. In subsequent sections, we will explore the gradient approximations in greater detail. We first focus on a distributed zeroth order scheme for random networks. In the sequel, we will turn our attention to communication efficient zeroth order schemes. We consider distributed stochastic zeroth order optimization to solve (9.1) over random networks. Inter-agent communication is modeled by a sequence of independent and identically distributed (i.i.d.) undirected random networks: at each time instant $k = 0, 1, ...$, the underlying inter-agent communication network is denoted by $\mathcal{G}(k) = (V, \mathbf{E}(k))$, with $V = \{1, ..., N\}$ being the set of nodes and $\mathbf{E}(k)$ being the random set of undirected edges. The edge connecting node i and j is denoted as $\{i, j\}$. The time-varying random neighborhood of node i at time k (excluding node i) is represented as $\Omega_i(k) = \{j \in V : \{i, j\} \in \mathbf{E}(k)\}$. The graph Laplacian of the random graph $\mathcal{G}(k)$ at time k is given by $\mathbf{L}(k) \in \mathbb{R}^{N \times N}$, where $\mathbf{L}(k)$ is given by $\mathbf{L}_{ij}(k) = -1$, if $\{i, j\} \in \mathbf{E}(k)$, $i \neq j$; $\mathbf{L}_{ij}(k) = 0$, if $\{i, j\} \notin \mathbf{E}(k)$, $i \neq j$; and $\mathbf{L}_{ii}(k) = -\sum_{j \neq i} \mathbf{L}_{ij}(k)$. It is to be noted that the Laplacian at each time instant is symmetric and a positive semidefinite matrix. As the considered graph sequence is i.i.d., we have that $\mathbb{E}[\,\mathbf{L}(k)\,] = \overline{\mathbf{L}}$. Let the graph corresponding to $\overline{\mathbf{L}}$ be given by $\overline{\mathcal{G}} = (V, \overline{\mathbf{E}})$.
We make the following assumption on $\overline{\mathcal{G}}$.

Assumption 9.3.2. The inter-agent communication graph is connected on average, i.e., $\overline{\mathcal{G}}$ is connected. In other words, $\lambda_2(\overline{\mathcal{L}}) > 0$.

9.4 Distributed Kiefer Wolfowitz type Optimization

We employ a distributed Kiefer Wolfowitz stochastic approximation (KWSA) type method to solve (9.1). Each node i, $i = 1, ..., N$, in our setup maintains a local copy of its local estimate of the optimizer $\mathbf{x}_i(k) \in \mathbb{R}^d$ at all times. In order to carry out the optimization, each agent i makes queries to a stochastic zeroth order oracle at time k, from which the agent obtains noisy function values of $f_i(\mathbf{x}_i(k))$. Denote the noisy value of $f_i(\cdot)$ as $\widehat{f}_i(\cdot)$ where,

$$\widehat{f}_i(\mathbf{x}_i(k)) = f_i(\mathbf{x}_i(k)) + \widehat{v}_i(k). \tag{9.3}$$

Due to the unavailability of the analytic form of the functionals, the gradient can not be evaluated and hence, we resort to a gradient approximation. In order to approximate the gradient, each agent makes two calls to the stochastic zeroth order oracle corresponding to each dimension. For instance, for dimension $j \in \{1, \cdots, d\}$ agent i queries for $f_i(\mathbf{x}_i(k) + c_k\mathbf{e}_j)$ and $f_i(\mathbf{x}_i(k) - c_k\mathbf{e}_j)$ at time k and obtains $\widehat{f}_i(\mathbf{x}_i(k) + c_k\mathbf{e}_j)$ and $\widehat{f}_i(\mathbf{x}_i(k) - c_k\mathbf{e}_i)$ respectively, where c_k is a carefully chosen time-decaying potential (to be specified soon). Denote by $\mathbf{g}_i(\mathbf{x}_i(k))$ the approximated gradient, obtained as for each $j \in \{1, \cdots, d\}$:

$$\begin{aligned}
&\mathbf{e}_j^\top \mathbf{g}_i(\mathbf{x}_i(k)) = \frac{\widehat{f}_i\left(\mathbf{x}_i(k) + c_k\mathbf{e}_j\right) - \widehat{f}_i\left(\mathbf{x}_i(k) - c_k\mathbf{e}_j\right)}{2c_k} \\
&\Rightarrow \mathbf{e}_j^\top \mathbf{g}_i(\mathbf{x}_i(k)) = \frac{f_i\left(\mathbf{x}_i(k) + c_k\mathbf{e}_j\right)}{2c_k} \\
&- \frac{f_i\left(\mathbf{x}_i(k) - c_k\mathbf{e}_j\right)}{2c_k} + \frac{\hat{v}_{i,j}^+(k) - \hat{v}_{i,j}^-(k)}{2c_k},
\end{aligned} \tag{9.4}$$

where $\hat{v}_{i,j}^+(k)$ and $\hat{v}_{i,j}^-(k)$ denote the measurement noise corresponding to the measurements $\widehat{f}_i\left(\mathbf{x}_i(k) + c_k\mathbf{e}_j\right)$ and $\widehat{f}_i\left(\mathbf{x}_i(k) - c_k\mathbf{e}_j\right)$ respectively. The vectors $\widehat{\mathbf{v}}_i^+(k) \in \mathbb{R}^d$ and $\widehat{\mathbf{v}}_i^-(k) \in \mathbb{R}^d$ stack all the component wise measurement noise at a node i and are given by $\widehat{\mathbf{v}}_i^+(k) = \left[\hat{v}_{i,1}^+(k), \cdots, \hat{v}_{i,N}^+(k)\right]$ and $\widehat{\mathbf{v}}_i^-(k) = \left[\hat{v}_{i,1}^-(k), \cdots, \hat{v}_{i,N}^-(k)\right]$ respectively. For the rest of this section, we define $\mathbf{v}_i(k) \doteq \left(\widehat{\mathbf{v}}_i^+(k) - \widehat{\mathbf{v}}_i^-(k)\right)/2$. Using the mean value theorem, we have,

$$\mathbf{g}_i(\mathbf{x}_i(k)) = \nabla f(\mathbf{x}_i(k)) + c_k\mathbf{P}_i(\mathbf{x}_i(k)) + \frac{\mathbf{v}_i(k)}{c_k}, \tag{9.5}$$

where

$$\begin{aligned}
\mathbf{e}_j^\top \mathbf{P}_i(\mathbf{x}_i(k)) &= \frac{\mathbf{e}_j^\top \nabla^2 f(\mathbf{x}_i(k) + c_k\alpha_{i,j}^+\mathbf{e}_j)\mathbf{e}_j}{2} \\
&- \frac{\mathbf{e}_j^\top \nabla^2 f(\mathbf{x}_i(k) - c_k\alpha_{i,j}^-\mathbf{e}_j)\mathbf{e}_j}{2},
\end{aligned}$$

where $0 \leq \alpha_{i,j}^+, \alpha_{i,j}^- \leq 1$. Finally, for arbitrary deterministic initializations $\mathbf{x}_i(0) \in \mathbb{R}^d$, $i = 1, ..., N$, the optimizer update rule at node i and $k = 0, 1, ...,$ is given as follows:

$$\begin{aligned}
\mathbf{x}_i(k+1) &= \mathbf{x}_i(k) - \beta_k \sum_{j \in \Omega_i(k)} \left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right) \\
&- \alpha_k \mathbf{g}_i(\mathbf{x}_i(k)).
\end{aligned} \tag{9.6}$$

It is to be noted that unlike first order stochastic gradient methods, where the algorithm has access to unbiased estimates of the gradient. The local gradient estimates $\mathbf{g}_i(\cdot)$ used in (9.6) are biased (see (9.5)) due to the unavailability of the exact gradient functions and their approximations using the zeroth order scheme in (9.54). The update is carried on in all agents parallely in a synchronous fashion. The weight sequences $\{\alpha_k\}$, $\{c_k\}$ and $\{\beta_k\}$ are given by $\alpha_k = \alpha_0/(k+1)$, $c_k = c_0/(k+1)^\delta$ and $\beta_k = \beta_0/(k+1)^\tau$ respectively, where $\alpha_0, c_0, \beta_0 > 0$. We state an assumption on the weight sequences before proceeding further.

Assumption 9.4.1. The constants α_0, $\delta > 0$ and $\tau \in (0,1)$ are chosen such that,

$$\sum_{k=1}^{\infty} \frac{\alpha_k^2}{c_k^2} < \infty. \tag{9.7}$$

Denote by $\mathbf{x}(k) = \left[\mathbf{x}_1^\top(k), \cdots, \mathbf{x}_N^\top(k)\right]^\top \in \mathbb{R}^{Nd}$, $\mathbf{P}(\mathbf{x}(k)) = \left[\mathbf{P}_1^\top(\mathbf{x}_1(k)), \cdots, \mathbf{P}_N^\top(\mathbf{x}_N(k))\right]^\top \in \mathbb{R}^{Nd}$ the vectors that stacks the local optimizers and the gradient bias terms (see (9.5))of all nodes. Also, define function $F : \mathbb{R}^{Nd} \mapsto \mathbb{R}$, by $F(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x}_i)$, with $\mathbf{x} = \left[\mathbf{x}_1^\top, \cdots, \mathbf{x}_N^\top\right]^\top \in \mathbb{R}^{Nd}$. Finally, let $\mathbf{W}_k = (\mathbf{I} - \mathbf{L}_k) \otimes \mathbf{I}_d$, where $\mathbf{L}_k = \beta_k \mathbf{L}(k)$. Then the update in (9.58) can be written as:

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{W}_k \mathbf{x}(k) \\ &- \alpha_k \left(\nabla F(\mathbf{x}(k)) + c_k \mathbf{P}(\mathbf{x}(k)) + \frac{\mathbf{v}(k)}{c_k} \right). \end{aligned} \tag{9.8}$$

Let $\mathcal{F}_k$ denote the history of the proposed algorithm up to time k. Given that the sources of randomness in our algorithm are the noise sequence $\{\mathbf{v}(k)\}$ and the random network sequence $\{\mathbf{L}_k\}$, $\mathcal{F}_k$ is given by the σ-algebra generated by the collection of random variables $\{\mathbf{L}(s), \mathbf{v}_i(s)\}$, $i = 1, ..., N$, $s = 0, ..., k-1$.

Assumption 9.4.2. For each $i = 1, ..., N$, the sequence of measurement noises $\{\widehat{\mathbf{v}}_i(k)\}$ satisfies for all $k = 0, 1, ...$:

$$\begin{aligned} &\mathbb{E}[\,\widehat{\mathbf{v}}_i(k) \,|\, \mathcal{F}_k\,] = 0, \text{ almost surely (a.s.)} \\ &\mathbb{E}[\,\|\widehat{\mathbf{v}}_i(k)\|^2 \,|\, \mathcal{F}_k\,] \leq c_f \|\mathbf{x}_i(k)\|^2 + \sigma^2, \text{ a.s.,} \end{aligned} \tag{9.9}$$

where c_f and σ^2 are nonnegative constants.

It is to be noted that assumption 9.4.2 is trivially satisfied, when $\{\mathbf{v}_i(k)\}$ is an i.i.d. zero-mean, finite second moment, noise sequence such that $\mathbf{v}_i(k)$ is also independent of the history $\mathcal{F}_k$. However, the assumption allows the noise to be dependent on the current iterate at all times.

9.5 Performance Analysis: Distributed KWSA

9.5.1 Main Result and Auxiliary Lemmas

We state the main result concerning the mean square error at each agent i next.

Theorem 9.5.1. *1) Consider the optimizer estimate sequence $\{\mathbf{x}(k)\}$ generated by the algorithm (9.6). Let assumptions 9.3.1-9.4.2 hold. Then, for each node i's optimizer estimate $\mathbf{x}_i(k)$ and the solution $\mathbf{x}^\star$ of*

problem (9.1), $\forall k \geq k_2$ *there holds:*

$$\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right] \leq 2R_k + \frac{64N\Delta_{1,\infty}\alpha_0^2}{\mu c_0^2 p_{\mathcal{L}}^2 \beta_0^2 (k+1)^{2-2\tau-2\delta}} \\ \frac{4(L-\mu)^2N^2dc_0^2}{\mu(k+1)^{2\delta}} + \frac{8\Delta_{1,\infty}\alpha_0^2}{p_{\mathcal{L}}^2\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}} \\ + \frac{4N\alpha_0\left(c_f q_\infty(N,d,\alpha_0,c_0)+N\sigma_1^2\right)}{c_0^2\mu(k+1)^{1-2\delta}}, \tag{9.10}$$

where, $k_2 = \max\left\{k_0, (4|E|\rho_0^2)^{1/\epsilon}-1\right\}$, $\Delta_{1,\infty} = 6c_f q_\infty(N,d,\alpha_0,c_0)+6N\sigma_1^2$ *and* $q_\infty(N,d,\alpha_0,c_0) = \mathbb{E}\left[\|\mathbf{x}(k_0)-\mathbf{x}^o\|^2\right] + \frac{\sqrt{Nd}(L-\mu)\alpha_0c_0}{\delta} + \frac{Nd(L-\mu)^2\alpha_0^2c_0^2}{1+2\delta} + \frac{\alpha_0^2\left(2c_fN\|\mathbf{x}^o\|^2+N\sigma^2\right)}{c_0^2(1-2\delta)} + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} + \frac{2\alpha_0^2c_0\sqrt{Nd}(L-\mu)\|\nabla F(\mathbf{x}^o)\|}{1+\delta}$. *In the latter* k_0 *is given by* $k_0 = \inf\left\{k \;:\; \frac{\mu}{2} > (L-\mu)\sqrt{Nd}c_k + \frac{2c_f\alpha_k}{c_k^2}\right\}$. R_k *is a term which decay faster than the rest of the terms.*

2) In particular, the rate of decay of the RHS of (9.67) *is given by* $(k+1)^{-\delta_1}$, *where* $\delta_1 = \min\{1-2\delta, 2-2\tau-2\delta, 2\delta\}$. *By, optimizing over* τ *and* δ, *we obtain that for* $\tau = 1/2$ *and* $\delta = 1/4$ *and hence,*

$$\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right] \leq 2R_k + \frac{64N\Delta_{1,\infty}\alpha_0^2}{\mu c_0^2 p_{\mathcal{L}}^2 \beta_0^2 (k+1)^{0.5}} \\ \frac{4(L-\mu)^2N^2dc_0^2}{\mu(k+1)^{0.5}} + \frac{8\Delta_{1,\infty}\alpha_0^2}{p_{\mathcal{L}}^2\beta_0^2c_0^2(k+1)^{0.5}} \\ + \frac{4N\alpha_0\left(c_f q_\infty(N,d,\alpha_0,c_0)+N\sigma_1^2\right)}{c_0^2\mu(k+1)^{0.5}} = O\left(\frac{1}{k^{\frac{1}{2}}}\right), \quad \forall i.$$

Theorem 9.5.1 establishes the $O(1/k^{1/2})$ MSE rate of convergence of the algorithm (9.6); due to the assumed f_i's strong convexity, the theorem also implies that $\mathbb{E}\left[f(\mathbf{x}_i(k))-f(\mathbf{x}^\star)\right] = O(1/k^{1/2})$. Note that the expectation in Theorem 9.5.1 is both with respect to randomness in gradient noises and with respect to the randomness in the underlying network. The $O(1/k^{1/2})$ rate is independent of the statistics of the underlying random network, as long as the network is connected on average.

From (9.10), it might seem that the dependence of the upper bound is linear in terms of d. However, on tuning the constants $\alpha_0 \asymp d^{-1/5}$, $\beta_0 \asymp d^{-1/10}$ and $c_0 \asymp d^{-3/10}$, the dependence of $\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]$ can be reduced to $d^{2/5}$. It is to be noted that the upper bound derived in (9.10) matches with that of the minimax bound for (centralized) zeroth order optimization with twice continuously differentiable cost functions as derived in Duchi et al. (2015). The sublinear rate of convergence of zeroth order optimization algorithms in the context of KWSA can be attributed to the biased gradients. For better finite time convergence rates, bias-reduction techniques such as the "twicing trick" and finite difference interpolation techniques can be used.

Proof strategy and auxiliary lemmas. Establishing the main result in Theorem 9.5.1 involves three crucial steps which are outlined in the subsections 9.5.2, 9.5.3 and 9.5.4. Subsection 9.5.2 concerns with the mean square boundedness of the iterates $\mathbf{x}_i(k)$, which also implies the mean square boundedness of the gradients $\nabla f_i(\mathbf{x}_i(k))$. In subsection 9.5.3, the mean square error of the disagreements of a node's optimizer estimate with respect to the network averaged optimizer estimate ,i.e., $\overline{\mathbf{x}}(k) := \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i(k)$, is characterized in terms of k and the algorithm parameters. Finally, subsection 9.5.4 characterizes the optimality gap of the networked average optimizer estimate sequence with respect to the optimizer of (9.1) and on combining the result from subsection 9.5.3, the result follows.

9.5.2 Mean square boundedness of the iterate sequence

This subsection shows the mean square boundedness of the algorithm iterates.

Lemma 9.5.2. *Let the hypotheses of Theorem 9.5.1 hold. In addition assume that, $\|\nabla F(\mathbf{1}_N \otimes \mathbf{x}^*)\|$ is bounded. Then, we have,*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}(k)-\mathbf{x}^o\right\|^2\right] \leq q_{k_0}(N,d,\alpha_0,c_0)\\
&+\frac{\sqrt{Nd}(L-\mu)\alpha_0 c_0}{\delta}+\frac{Nd(L-\mu)^2\alpha_0^2 c_0^2}{1+2\delta}\\
&+\frac{\alpha_0^2\left(2c_f N\left\|\mathbf{x}^o\right\|^2+N\sigma^2\right)}{c_0^2(1-2\delta)}+4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}\\
&\doteq q_\infty(N,d,\alpha_0,c_0),
\end{aligned}$$

where $\mathbb{E}\left[\left\|\mathbf{x}(k_0)-\mathbf{x}^o\right\|^2\right] \leq q_{k_0}(N,d,\alpha_0,c_0)$ *and* $k_0=\inf\left\{k \;:\; \frac{\mu}{2} > (L-\mu)\sqrt{Nd}c_k+\frac{2c_f\alpha_k}{c_k^2}\right\}$.

Proof.

$$\begin{aligned}
&\mathbf{x}(k+1)=\mathbf{W}_k\mathbf{x}(k)\\
&-\frac{\alpha_k}{c_k}\left(c_k\nabla F(\mathbf{x}(k))+c_k^2\mathbf{P}(\mathbf{x}(k))+\mathbf{v}(k)\right).
\end{aligned} \tag{9.11}$$

Denote $\mathbf{x}^o=\mathbf{1}_N\otimes x^*$. Then, we have,

$$\begin{aligned}
&\mathbf{x}(k+1)-\mathbf{x}^o=\mathbf{W}_k(\mathbf{x}(k)-\mathbf{x}^o)\\
&-\alpha_k\left(\nabla F(\mathbf{x}(k))-\nabla F(\mathbf{x}^o)\right)\\
&-\frac{\alpha_k}{c_k}\mathbf{v}(k)-\alpha_k\nabla F(\mathbf{x}^o)-\alpha_k c_k\mathbf{P}(\mathbf{x}(k)).
\end{aligned} \tag{9.12}$$

By Leibnitz rule, we have,

$$\begin{aligned}
&\nabla F(\mathbf{x}(k))-\nabla F(\mathbf{x}^o)\\
&=\left[\int_{s=0}^1\nabla^2 F\left(\mathbf{x}^o+s(\mathbf{x}(k)-\mathbf{x}^o)\right)ds\right](\mathbf{x}(k)-\mathbf{x}^o)\\
&=\mathbf{H}_k\left(\mathbf{x}(k)-\mathbf{x}^o\right).
\end{aligned} \tag{9.13}$$

By Lipschitz continuity of the gradients and strong convexity of $f(\cdot)$, we have that $L\mathbf{I}\succcurlyeq\mathbf{H}_k\succcurlyeq\mu\mathbf{I}$. Denote by $\boldsymbol{\zeta}(k)=\mathbf{x}(k)-\mathbf{x}^o$ and by $\boldsymbol{\xi}(k)=\left(\mathbf{W}_k-\alpha_k\mathbf{H}_k\right)\left(\mathbf{x}(k)-\mathbf{x}^o\right)-\alpha_k\nabla F(\mathbf{x}^o)$. Then, there holds:

$$\begin{aligned}
&\mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2\,|\,\mathcal{F}_k\,]\leq\mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2|\,\mathcal{F}_k\,\right]\\
&-2\alpha_k\mathbb{E}\left[\boldsymbol{\xi}(k)^\top|\,\mathcal{F}_k\right]\mathbb{E}[\,\mathbf{v}(k)\,|\,\mathcal{F}_k\,]+\alpha_k^2\,\mathbb{E}[\,\|\mathbf{v}(k)\|^2\,|\,\mathcal{F}_k\,]\\
&+\alpha_k^2c_k^2\mathbf{P}^\top(\mathbf{x}(k))\mathbf{P}(\mathbf{x}(k))-2\alpha_k c_k\mathbf{P}^\top(\mathbf{x}(k))\mathbb{E}\left[\boldsymbol{\xi}(k)|\mathcal{F}_k\right]\\
&+\mathbf{P}\left(\mathbf{x}(k)\right)^\top\mathbb{E}\left[\mathbf{v}(k)|\mathcal{F}_k\right].
\end{aligned} \tag{9.14}$$

We use the following inequalities:

$$-2\alpha_k c_k\mathbf{P}^\top(\mathbf{x}(k))\left(\mathbf{I}-\beta_k\overline{\mathbf{L}}-\alpha_k\mathbf{H}_k\right)(\mathbf{x}(k)-\mathbf{x}^o)$$

$$\begin{aligned}
&\leq 2\alpha_k c_k \left\|\mathbf{P}(\mathbf{x}(k))\right\| \left\|\mathbf{I} - \beta_k \overline{\mathbf{L}} - \alpha_k \mathbf{H}_k\right\| \left\|\mathbf{x}(k) - \mathbf{x}^o\right\| \\
&\leq \sqrt{Nd}(L-\mu)\alpha_k c_k \left(1 - \mu\alpha_k\right) \left(1 + \left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2\right) \\
&\leq \sqrt{Nd}(L-\mu)\alpha_k c_k + \sqrt{Nd}(L-\mu)\alpha_k c_k \left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2, \qquad (9.15)
\end{aligned}$$

$$\alpha_k^2 c_k^2 \mathbf{P}^\top(\mathbf{x}(k))\mathbf{P}(\mathbf{x}(k)) \leq Nd\left(L-\mu\right)^2 \alpha_k^2 c_k^2, \qquad (9.16)$$

and

$$\begin{aligned}
&\frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\left\|\mathbf{v}(k)\right\|^2 | \mathcal{F}_k\right] \leq \frac{\alpha_k^2}{c_k^2} c_f N \left\|\mathbf{x}(k)\right\|^2 + \frac{\alpha_k^2}{c_k^2} N\sigma^2 \\
&\leq 2\frac{\alpha_k^2}{c_k^2} c_f \left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2 + \frac{\alpha_k^2}{c_k^2}\left(2c_f \left\|\mathbf{x}^o\right\|^2 + N\sigma^2\right). \qquad (9.17)
\end{aligned}$$

Then from (9.14), we have,

$$\begin{aligned}
&\mathbb{E}[\left\|\boldsymbol{\zeta}(k+1)\right\|^2 | \mathcal{F}_k] \leq \mathbb{E}\left[\left\|\boldsymbol{\xi}(k)\right\|^2 | \mathcal{F}_k\right] \\
&+ \sqrt{Nd}(L-\mu)\alpha_k c_k \|\boldsymbol{\zeta}(k)\|^2 + 2\frac{\alpha_k^2}{c_k^2} c_f \|\boldsymbol{\zeta}(k)\|^2 \\
&+ \frac{\alpha_k^2}{c_k^2}\left(2c_f \left\|\mathbf{x}^o\right\|^2 + N\sigma^2\right) + \sqrt{Nd}(L-\mu)\alpha_k c_k \\
&+ Nd\left(L-\mu\right)^2 \alpha_k^2 c_k^2 + 2\alpha_k^2 c_k \sqrt{Nd}(L-\mu)\left\|\nabla F(\mathbf{x}^o)\right\|. \qquad (9.18)
\end{aligned}$$

We next bound $\mathbb{E}\left[\left\|\boldsymbol{\xi}(k)\right\|^2 | \mathcal{F}_k\right]$. Note that $\left\|\mathbf{W}_k - \alpha_k \boldsymbol{H}_k\right\| \leq 1 - \mu\,\alpha_k$. Therefore, we have:

$$\left\|\boldsymbol{\xi}(k)\right\| \leq \left(1 - \mu\,\alpha_k\right)\left\|\boldsymbol{\zeta}(k)\right\| + \alpha_k \left\|\nabla F(\mathbf{x}^o)\right\|. \qquad (9.19)$$

We now use the following inequality:

$$(a+b)^2 \leq (1+\theta)\,a^2 + \left(1 + \frac{1}{\theta}\right) b^2, \qquad (9.20)$$

for any $a, b \in \mathbb{R}$ and $\theta > 0$. We set $\theta = \mu\alpha_k$. Using the inequality (9.20) in (9.19), we have:

$$\begin{aligned}
&\left\|\boldsymbol{\xi}(k)\right\|^2 \leq \left(1 + \mu\alpha_k\right)\left(1 - \alpha_k\mu\right)^2 \left\|\boldsymbol{\zeta}(k)\right\|^2 \\
&+ \left(1 + \frac{1}{\mu\alpha_k}\right)\alpha_k^2 \|\nabla F(\mathbf{x}^o)\|^2 \\
&\leq \left(1 - \alpha_k\mu\right)\left\|\boldsymbol{\zeta}(k)\right\|^2 + 2\frac{\alpha_k}{\mu}\|\nabla F(\mathbf{x}^o)\|^2. \qquad (9.21)
\end{aligned}$$

Using (9.21) in (9.18), we have,

$$\begin{aligned}
&\mathbb{E}[\left\|\boldsymbol{\zeta}(k+1)\right\|^2 | \mathcal{F}_k] \leq 2\frac{\alpha_k}{\mu}\|\nabla F(\mathbf{x}^o)\|^2 + \left\|\boldsymbol{\zeta}(k)\right\|^2 \\
&\times \left(1 - \alpha_k\mu + \sqrt{Nd}(L-\mu)\alpha_k c_k + 2\frac{\alpha_k^2}{c_k^2} c_f\right) \\
&+ \frac{\alpha_k^2}{c_k^2}\left(2c_f \left\|\mathbf{x}^o\right\|^2 + N\sigma^2\right) + \sqrt{Nd}(L-\mu)\alpha_k c_k \\
&+ Nd\left(L-\mu\right)^2 \alpha_k^2 c_k^2 + 2\alpha_k^2 c_k \sqrt{Nd}(L-\mu)\left\|\nabla F(\mathbf{x}^o)\right\|. \qquad (9.22)
\end{aligned}$$

Define k_0 as follows:

$$k_0 = \inf\left\{k \ : \ \frac{\mu}{2} > (L-\mu)\sqrt{Nd}c_k + \frac{2c_f\alpha_k}{c_k^2}\right\}.$$

It is to be noted that k_0 is necessarily finite as $c_k \to 0$ and $\alpha_k c_k^{-2} \to 0$ as $k \to \infty$.

Proposition 9.5.3. *Let the hypotheses of Theorem 9.5.1 hold. Then, we have $\forall k \geq k_0$,*

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{\zeta}(k+1)\|^2\right] \leq q_{k_0}(N,d,\alpha_0,c_0) + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} \\
&+ \frac{\sqrt{Nd}(L-\mu)\alpha_0 c_0}{\delta} + \frac{Nd(L-\mu)^2\alpha_0^2c_0^2}{1+2\delta} \\
&+ \frac{\alpha_0^2\left(2c_fN\|\mathbf{x}^o\|^2 + N\sigma^2\right)}{c_0^2(1-2\delta)} + \frac{2\sqrt{Nd}(L-\mu)\|\nabla F(\mathbf{x}^o)\|}{1+\delta} \\
&\doteq q_\infty(N,d,\alpha_0,c_0)
\end{aligned}$$

□

With the above development in place, we can bound the variance of the noise process $\{\mathbf{v}(k)\}$ as follows:

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{v}(k)\|^2|\mathcal{F}_k\right] \leq 0.5\mathbb{E}\left[\left\|\widehat{\mathbf{v}}^+(k)\right\|^2|\mathcal{F}_k\right] \\
&+ 0.5\mathbb{E}\left[\left\|\widehat{\mathbf{v}}^-(k)\right\|^2|\mathcal{F}_k\right] \\
&\leq 2c_fq_\infty(N,d,\alpha_0,c_0) + 2N\underbrace{\left(\sigma^2 + \|\mathbf{x}^*\|^2\right)}_{\sigma_1^2}.
\end{aligned} \tag{9.23}$$

9.5.3 Disagreement Bounds

We now study the disagreement of the optimizer sequence $\{\mathbf{x}_i(k)\}$ at a node i with respect to the (hypothetically available) network averaged optimizer sequence, i.e., $\overline{\mathbf{x}}(k) = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i(k)$. Define the disagreement at the i-th node as $\widetilde{\mathbf{x}}_i(k) = \mathbf{x}_i(k) - \overline{\mathbf{x}}(k)$. The vectorized version of the disagreements $\widetilde{\mathbf{x}}_i(k)$, $i = 1, \cdots, N$, can then be written as $\widetilde{\mathbf{x}}(k) = (\mathbf{I} - \mathbf{J})\,\mathbf{x}(k)$, where $\mathbf{J} = \frac{1}{N}\left(\mathbf{1}_N \otimes \mathbf{I}_d\right)\left(\mathbf{1}_N \otimes \mathbf{I}_d\right)^\top = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top \otimes \mathbf{I}_d$. We have the following Lemma:

Lemma 9.5.4. *Let the hypotheses of Theorem 9.5.1 hold. Then, we have $\forall k \geq k_1$*

$$\begin{aligned}
&\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] \leq Q_k + \frac{4\Delta_{1,\infty}\alpha_0^2}{p_{\mathcal{L}}^2\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}} \\
&= O\left(\frac{1}{k^{2-2\delta-2\tau}}\right),
\end{aligned}$$

where Q_k is a term which decays faster than $(k+1)^{-2+2\tau+2\delta}$ and $k_1 = (4|E|\rho_0^2)^{1/\epsilon} - 1$.

As detailed in the next Subsection, Lemma 9.5.4 plays a crucial role in providing a tight bound for the bias in the gradient estimates according to which the global average $\overline{\mathbf{x}}(k)$ evolves.

Proof. The process $\{\widetilde{\mathbf{x}}(k)\}$ follows the recursion:

$$\begin{aligned}\widetilde{\mathbf{x}}(k+1) &= \widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\\ &- \frac{\alpha_k}{c_k}(\mathbf{I}-\mathbf{J})\underbrace{\left(c_k\nabla F(\mathbf{x}(k)) + \mathbf{v}(k) + c_k^2\mathbf{P}(\mathbf{x}(k))\right)}_{\mathbf{w}(k)},\end{aligned} \tag{9.24}$$

where $\widetilde{\mathbf{W}}_k = \mathbf{W}_k - \mathbf{J} = (\mathbf{I}-\mathbf{L}_k)\otimes\mathbf{I}_d - \mathbf{J}$. Using (9.20) in (9.24), we have,

$$\begin{aligned}\|\widetilde{\mathbf{x}}(k+1)\|^2 &\le (1+\theta_k)\left\|\widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\right\|^2\\ &+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\|\widetilde{\mathbf{w}}(k)\|^2.\end{aligned} \tag{9.25}$$

We, now bound the term $\mathbb{E}\left[\left\|\widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\right\|^2|\mathcal{F}_k\right]$.

$$\begin{aligned}&\mathbb{E}\left[\left\|\widetilde{\mathbf{W}}(k)\widetilde{\mathbf{x}}(k)\right\|^2|\mathcal{F}_k\right] = \widetilde{\mathbf{x}}^\top(k)\mathbb{E}\left[\widetilde{\mathbf{W}}^2(k)-\mathbf{J}|\mathcal{F}_k\right]\widetilde{\mathbf{x}}(k)\\ &= \widetilde{\mathbf{x}}^\top(k)\left(\mathbf{I}-2\beta_k\overline{\mathbf{L}}+\beta_k^2\overline{\mathbf{L}}^2+\widetilde{\mathbf{L}}(k)^2-\mathbf{J}\right)\widetilde{\mathbf{x}}(k)\\ &\le \left(1-2\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)+\beta_k^2\lambda_N^2\left(\overline{\mathbf{L}}\right)\right.\\ &\left.+\frac{4|E|\beta_0\rho_0^2}{(k+1)^{1/2+\epsilon}}-4\beta_k^2|E|\right)\|\widetilde{\mathbf{x}}(k)\|^2\\ &\le \left(1-2\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)+\frac{4|E|\beta_0\rho_0^2}{(k+1)^{1/2+\epsilon}}\right)\|\widetilde{\mathbf{x}}(k)\|^2\\ &\le \left(1-\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2,\end{aligned} \tag{9.26}$$

where the last inequality holds for $k \ge (4|E|\rho_0^2)^{1/\epsilon} - 1 \doteq k_1$. Then, we have, $\forall k \ge k_1$,

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2|\mathcal{F}_k\right] &\le (1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2\\ &+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{F}_k\right],\end{aligned} \tag{9.27}$$

where

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{F}_k\right] \le 3c_k^2\|\nabla F(\mathbf{x}(k))\|^2 + 3\mathbb{E}\left[\|\mathbf{v}(k)\|^2|\mathcal{F}_k\right]\\ &+3c_k^2\|P(\mathbf{x}(k))\|^2\\ &\le 3c_k^2\|\nabla F(\mathbf{x}(k))\|^2 + 3c_k^2Nd(L-\mu)^2\\ &+6c_fq_\infty(N,d,\alpha_0,c_0)+6N\sigma_1^2\\ &\Rightarrow \mathbb{E}\left[\|\mathbf{w}(k)\|^2\right] \le 3\left(2c_f+c_k^2L^2\right)q_\infty(N,d,\alpha_0,c_0)\\ &+3c_k^2Nd(L-\mu)^2+6N\sigma_1^2\\ &=\underbrace{6c_fq_\infty(N,d,\alpha_0,c_0)+6N\sigma_1^2}_{\Delta_{1,\infty}}\\ &+\underbrace{3c_k^2Nd(L-\mu)^2+3c_k^2L^2q_\infty(N,d,\alpha_0,c_0)}_{c_k^2\Delta_{2,\infty}} \doteq \Delta_k\end{aligned}$$

$$\Rightarrow \mathbb{E}\left[\|\mathbf{w}(k)\|^2\right] < \infty. \tag{9.28}$$

With the above development in place, we then have,

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] &\leq (1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2 \\ &+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\Delta_k.\end{aligned} \tag{9.29}$$

In particular, we choose $\theta(k)=\frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)$.

Proposition 9.5.5. *Let the hypotheses of Theorem 9.5.1 hold. Then, we have $k \geq k_2 = \max\{k_0, k_1\}$ where $k_1 = (4|E|\rho_0^2)^{1/\epsilon} - 1$,*

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] \leq Q_k + \frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2 c_0^2 (k+1)^{2-2\tau-2\delta}}.$$

Hence, we have the disagreement given by,

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] = O\left(\frac{1}{k^{2-2\delta-2\tau}}\right).$$

□

9.5.4 Proof of Theorem 9.5.1

In this subsection, we complete the proof of Theorem 9.5.1 by characterizing the optimality gap of the network averaged optimizer estimate sequence and then combining it with the result obtained in Lemma 9.5.4.

Denote $\overline{\mathbf{x}}(k)=\frac{1}{N}\sum_{n=1}\mathbf{x}_i(k)$. From (9.24), we have,

$$\begin{aligned}&\overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k) \\ &-\frac{\alpha_k}{c_k}\left[\frac{c_k}{N}\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right)+\underbrace{\frac{c_k^2}{N}\sum_{i=1}^{N}\mathbf{P}_i\left(\mathbf{x}_i(k)\right)}_{\overline{\mathbf{P}}\left(\mathbf{x}(k)\right)}+\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i(k)}_{\overline{\mathbf{v}}(k)}\right] \\ &\Rightarrow \overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k) \\ &-\frac{\alpha_k}{Nc_k}\left[\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right)-\nabla f_i\left(\overline{\mathbf{x}}(k)\right)+\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right] \\ &-\frac{\alpha_k}{c_k}\left(\overline{\mathbf{v}}(k)+\overline{\mathbf{P}}\left(\mathbf{x}(k)\right)\right).\end{aligned} \tag{9.30}$$

Recall that $f(\cdot)=\sum_{i=1}^{N}f_i(\cdot)$. Then, we have,

$$\begin{aligned}\overline{\mathbf{x}}(k+1) &= \overline{\mathbf{x}}(k)-\frac{\alpha_k}{N}\nabla f\left(\overline{\mathbf{x}}(k)\right) \\ &-\frac{\alpha_k}{N}\left[\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right)-\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right] \\ &-\frac{\alpha_k}{c_k}\left(\overline{\mathbf{v}}(k)+\overline{\mathbf{P}}\left(\mathbf{x}(k)\right)\right)\end{aligned}$$

$$\Rightarrow \overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{Nc_k}\left[c_k \nabla f\left(\overline{\mathbf{x}}(k)\right) + \mathbf{e}(k)\right], \tag{9.31}$$

where

$$\begin{aligned}\mathbf{e}(k) &= N\overline{\mathbf{v}}(k) \\ &+ \underbrace{N\overline{P}\left(\mathbf{x}(k)\right) + c_k \sum_{i=1}^{N}\left(\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right)}_{\boldsymbol{\epsilon}(k)}.\end{aligned} \tag{9.32}$$

Note that, $c_k \left\|\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right\| \leq c_k L \left\|\mathbf{x}_i(k) - \overline{\mathbf{x}}(k)\right\| = c_k L \left\|\widetilde{\mathbf{x}}_i(k)\right\|$. We also have that, $\left\|\overline{\mathbf{P}}\left(\mathbf{x}(k)\right)\right\| \leq (L-\mu)\sqrt{d}c_k^2$. Thus, we can conclude that, $\forall k \geq k_2$

$$\begin{aligned}&\boldsymbol{\epsilon}(k) = c_k \sum_{i=1}^{N}\left(\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right) + N\overline{P}\left(\mathbf{x}(k)\right) \\ &\Rightarrow \left\|\boldsymbol{\epsilon}(k)\right\|^2 \leq 2NL^2c_k^2\left\|\widetilde{\mathbf{x}}(k)\right\|^2 + 2(L-\mu)^2N^2dc_k^4 \\ &\Rightarrow \mathbb{E}\left[\left\|\boldsymbol{\epsilon}(k)\right\|^2\right] \leq \frac{8NL^2\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{2-2\tau}} + \frac{2(L-\mu)^2N^2dc_0^4}{(k+1)^{4\delta}} \\ &+ \frac{4NL^2Q_kc_0^2}{(k+1)^{2\delta}}.\end{aligned} \tag{9.33}$$

With the above development in place, we rewrite (9.31) as follows:

$$\begin{aligned}&\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{N}\nabla f\left(\overline{\mathbf{x}}(k)\right) - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{v}}(k) \\ &\Rightarrow \overline{\mathbf{x}}(k+1) - \mathbf{x}^* = \overline{\mathbf{x}}(k) - \mathbf{x}^* - \frac{\alpha_k}{N}\left[\nabla f\left(\overline{\mathbf{x}}(k)\right) - \underbrace{\nabla f\left(\mathbf{x}^*\right)}_{=0}\right] \\ &- \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{v}}(k).\end{aligned} \tag{9.34}$$

By Leibnitz rule, we have,

$$\begin{aligned}&\nabla f\left(\overline{\mathbf{x}}(k)\right) - \nabla f\left(\mathbf{x}^*\right) \\ &= \underbrace{\left[\int_{s=0}^{1}\nabla^2 f\left(\mathbf{x}^* + s\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right)\right)ds\right]}_{\overline{\mathbf{H}}_k}\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right),\end{aligned} \tag{9.35}$$

where it is to be noted that $NL \succcurlyeq \overline{\mathbf{H}}_k \succcurlyeq N\mu$. Using (9.35) in (9.34), we have,

$$\begin{aligned}&\left(\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right) = \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right]\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right) \\ &- \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{v}}(k).\end{aligned} \tag{9.36}$$

Denote by $\mathbf{m}(k) = \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right]\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right) - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k)$ and note that $\mathbf{m}(k)$ is conditionally independent from $\overline{\mathbf{v}}(k)$ given the history $\mathcal{F}_k$. Then (9.36) can be rewritten as:

$$\begin{aligned}&\left(\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right) = \mathbf{m}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{v}}(k) \\ &\Rightarrow \left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right\|^2 \leq \left\|\mathbf{m}(k)\right\|^2 \\ &- 2\frac{\alpha_k}{c_k}\mathbf{m}(k)^\top\overline{\mathbf{v}}(k) + \frac{\alpha_k^2}{c_k^2}\left\|\overline{\mathbf{v}}(k)\right\|^2.\end{aligned} \tag{9.37}$$

Using the properties of conditional expectation and noting that $\mathbb{E}\left[\mathbf{v}(k)|\mathcal{F}_k\right] = \mathbf{0}$, we have,

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\|^2 |\mathcal{F}_k\right] \leq \|\mathbf{m}(k)\|^2 + \frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\|v(k)\|^2 |\mathcal{F}_k\right] \\
&\Rightarrow \mathbb{E}\left[\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\|^2\right] \leq \mathbb{E}\left[\|\mathbf{m}(k)\|^2\right] \\
&+ \frac{2\alpha_k^2\left(c_f q_\infty(N,d,\alpha_0,c_0) + N\sigma_1^2\right)}{c_k^2}.
\end{aligned} \tag{9.38}$$

Using (9.20), we have for $\mathbf{m}(k)$,

$$\begin{aligned}
&\|\mathbf{m}(k)\|^2 \leq (1+\theta_k)\left\|\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right\|^2 \|\overline{\mathbf{x}}(k) - \mathbf{x}^*\|^2 \\
&+ \left(1 + \frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2 c_k^2}\|\boldsymbol{\epsilon}(k)\|^2 \\
&\leq (1+\theta_k)\left(1 - \frac{\mu\alpha_0}{k+1}\right)^2 \|\overline{\mathbf{x}}(k) - \mathbf{x}^*\|^2 \\
&+ \left(1 + \frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2 c_k^2}\|\boldsymbol{\epsilon}(k)\|^2.
\end{aligned} \tag{9.39}$$

On choosing $\theta_k = \frac{\mu\alpha_0}{k+1}$, we have for all $k \geq k_2$,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{m}(k)\|^2\right] \leq \left(1 - \frac{\mu\alpha_0}{N(k+1)}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k) - \mathbf{x}^*\|^2\right] \\
&+ \frac{16L^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}} + \frac{4(L-\mu)^2 N d\alpha_0 c_0^2}{\mu(k+1)^{1+2\delta}} \\
&+ \frac{8Q_k c_0^2}{\mu(k+1)^{2\delta}} \\
&\Rightarrow \mathbb{E}\left[\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\|^2\right] \leq \left(1 - \frac{\mu\alpha_0}{N(k+1)}\right) \\
&\times \mathbb{E}\left[\|\overline{\mathbf{x}}(k) - \mathbf{x}^*\|^2\right] \\
&+ \frac{16L^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}} + \frac{4(L-\mu)^2 N d\alpha_0 c_0^2}{\mu(k+1)^{1+2\delta}} \\
&+ \frac{8Q_k c_0^2}{\mu(k+1)^{2\delta}} + \frac{2\alpha_0^2\left(c_f q_\infty(N,d,\alpha_0,c_0) + N\sigma_1^2\right)}{c_0^2(k+1)^{2-2\delta}}.
\end{aligned} \tag{9.40}$$

Then, we have $\forall k \geq k_2$

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\|^2\right] \\
&\leq \underbrace{\exp\left(-\frac{\mu}{N}\sum_{l=k_5}^{k}\alpha_l\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k) - \mathbf{x}^*\|^2\right]}_{t_6} \\
&\underbrace{+ \exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{16L^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}}_{t_7} \\
&\underbrace{+ \exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{4(L-\mu)^2 N d\alpha_0 c_0^2}{\mu(k+1)^{1+2\delta}}}_{t_8}
\end{aligned}$$

$$
\begin{aligned}
&+ \underbrace{\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{8Q_k c_0^2}{\mu(k+1)^{2\delta}}}_{t_9}\\
&+ \underbrace{\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{2\alpha_0^2 c_f q_\infty(N,d,\alpha_0,c_0)}{c_0^2(k+1)^{2-2\delta}}}_{t_{10}}\\
&+ \underbrace{\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{2\alpha_0^2 N\sigma_1^2}{c_0^2(k+1)^{2-2\delta}}}_{t_{11}}\\
&+ \underbrace{\frac{32L^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}}_{t_{12}}\\
&+ \underbrace{\frac{4(L-\mu)^2N^2dc_0^2}{\mu(k+1)^{2\delta}}}_{t_{13}} + \underbrace{\frac{2Nc_0^2Q_k}{\mu\alpha_0(k+1)^{2\delta-1}}}_{t_{14}}\\
&+ \underbrace{\frac{4N\alpha_0\left(c_f q_\infty(N,d,\alpha_0,c_0)+N\sigma_1^2\right)}{c_0^2\mu(k+1)^{1-2\delta}}}_{t_{15}}. \qquad (9.41)
\end{aligned}
$$

It is to be noted that the term t_6 decays exponentially. The terms t_7, t_8, t_9, t_{10} and t_{11} decay faster than its counterparts in the terms t_{12}, t_{13}, t_{14} and t_{15} respectively. We note that Q_l also decays faster. Hence, the rate of decay of $\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\right\|^2\right]$ is determined by the terms t_{12}, t_{13} and t_{15}. Thus, we have that, $\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\right\|^2\right] = O\left(k^{-\delta_1}\right)$, where $\delta_1 = \min\{1-2\delta, 2-2\tau-2\delta, 2\delta\}$. For notational ease, we refer to $t_6+t_7+t_8+t_9+t_{10}+t_{11}+t_{14} = M_k$ from now on. Finally, we note that,

$$
\begin{aligned}
&\|\mathbf{x}_i(k)-\mathbf{x}^*\| \leq \|\overline{\mathbf{x}}(k)-\mathbf{x}^*\| + \left\|\underbrace{\mathbf{x}_i(k)-\overline{\mathbf{x}}(k)}_{\widetilde{\mathbf{x}}_i(k)}\right\|\\
&\Rightarrow \|\mathbf{x}_i(k)-\mathbf{x}^*\|^2 \leq 2\|\widetilde{\mathbf{x}}_i(k)\|^2 + 2\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\\
&\Rightarrow \mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right] \leq 2R_k + \frac{64N\Delta_{1,\infty}\alpha_0^2}{\mu c_0^2 p_{\mathcal{L}}^2\beta_0^2(k+1)^{2-2\tau-2\delta}}\\
&\frac{4(L-\mu)^2N^2dc_0^2}{\mu(k+1)^{2\delta}} + \frac{8\Delta_{1,\infty}\alpha_0^2}{p_{\mathcal{L}}^2\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}}\\
&+ \frac{4N\alpha_0\left(c_f q_\infty(N,d,\alpha_0,c_0)+N\sigma_1^2\right)}{c_0^2\mu(k+1)^{1-2\delta}}\\
&\Rightarrow \mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right] = O\left(\frac{1}{k^{\delta_1}}\right), \quad \forall i, \qquad (9.42)
\end{aligned}
$$

where $\delta_1 = \min\{1-2\delta, 2-2\tau-2\delta, 2\delta\}$ and $R_k = M_k + Q_k$. By, optimizing over τ and δ, we obtain that for $\tau = 1/2$ and $\delta = 1/4$,

$$
\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right] = O\left(\frac{1}{k^{\frac{1}{2}}}\right), \quad \forall i.
$$

Before stating the communication efficient algorithm, we first discuss the communication scheme. Specifi-

cally, we adopt the following model.

Communication Scheme

The inter-node communication network to which the information exchange between nodes conforms to is modeled as an *undirected* simple connected graph $G = (V, E)$, with $V = [1 \cdots N]$ and E denoting the set of nodes and communication links. The neighborhood of node n is given by $\Omega_n = \{l \in V \,|\, (n, l) \in E\}$. The node n has degree $d_n = |\Omega_n|$. The structure of the graph is described by the $N \times N$ adjacency matrix, $\mathbf{A} = \mathbf{A}^\top = [\mathbf{A}_{nl}]$, $\mathbf{A}_{nl} = 1$, if $(n, l) \in E$, $\mathbf{A}_{nl} = 0$, otherwise. The graph Laplacian $\overline{\mathbf{L}} = \mathbf{D} - \mathbf{A}$ is positive semidefinite, with eigenvalues ordered as $0 = \lambda_1(\overline{\mathbf{L}}) \leq \lambda_2(\overline{\mathbf{L}}) \leq \cdots \leq \lambda_N(\overline{\mathbf{L}})$, where $\mathbf{D}$ is given by $\mathbf{D} = \text{diag}\,(d_1 \cdots d_N)$. We make the following assumption on $\overline{\mathbf{L}}$.

Assumption 9.5.1. The inter-agent communication graph is connected on average, i.e., $\overline{\mathbf{L}}$ is connected. In other words, $\lambda_2(\overline{\mathbf{L}}) > 0$.

Thus, $\overline{\mathbf{L}}$ corresponds to the maximal graph, i.e., the graph of all *allowable* communications. We now describe our randomized communication protocol that selects a (random) subset of the allowable links at each time instant for information exchange.
For each node i, at every time k, we introduce a binary random variable $\psi_{i,k}$, where

$$\psi_{i,k} = \begin{cases} \rho_k & \textit{with probability } \zeta_k \\ 0 & \textit{otherwise}, \end{cases} \tag{9.43}$$

where $\psi_{i,k}$'s are independent both across time and the nodes, i.e., across k and i respectively. The random variable $\psi_{i,k}$ abstracts out the decision of the node i at time k whether to participate in the neighborhood information exchange or not. We specifically take ρ_k and ζ_k of the form

$$\rho_k = \frac{\rho_0}{(k+1)^{\epsilon/2}}, \;\; \zeta_k = \frac{\zeta_0}{(k+1)^{(\tau/2-\epsilon/2)}}, \tag{9.44}$$

where $0 < \tau \leq \frac{1}{2}$ and $0 < \epsilon < \tau$. Furthermore, define β_k to be

$$\beta_k = (\rho_k \zeta_k)^2 = \frac{\beta_0}{(k+1)^\tau}, \tag{9.45}$$

where $\beta_0 = \rho_0^2 \zeta_0^2$. With the above development in place, we define the random time-varying Laplacian $\mathbf{L}(k)$, where $\mathbf{L}(k) \in \mathbb{R}^{N \times N}$ abstracts the inter-node information exchange as follows:

$$\mathbf{L}_{i,j}(k) = \begin{cases} -\psi_{i,k}\psi_{j,k} & \{i, j\} \in E, i \neq j \\ 0 & i \neq j, \{i, j\} \notin E \\ \sum_{l \neq i} \psi_{i,k}\psi_{l,k} & i = j. \end{cases} \tag{9.46}$$

The above communication protocol allows two nodes to communicate only when the link is established in a bi-directional fashion and hence avoids directed graphs. The design of the communication protocol as depicted in (3.27)-(9.46) not only decays the weight assigned to the links over time but also decays the probability of the existence of a link. Such a design is consistent with frameworks where the agents have finite power and hence not only the number of communications, but also, the quality of the communication decays over time. We have, for $\{i, j\} \in E$ and $i \neq j$:

$$\mathbb{E}\left[\mathbf{L}_{i,j}(k)\right] = -\left(\rho_k \zeta_k\right)^2 = -\beta_k = -\frac{\beta_0}{(k+1)^\tau}$$

$$\mathbb{E}\left[\mathbf{L}_{i,j}^2(k)\right] = \left(\rho_k^2\zeta_k\right)^2 = \frac{\rho_0^2\beta_0}{(k+1)^{\tau+\epsilon}}. \tag{9.47}$$

Thus, we have that, the variance of $\mathbf{L}_{i,j}(k)$ is given by,

$$Var\left(\mathbf{L}_{i,j}(k)\right) = \frac{\beta_0\rho_0^2}{(k+1)^{\tau+\epsilon}} - \frac{\beta_0^2}{(k+1)^{2\tau}}. \tag{9.48}$$

Define, the mean of the random time-varying Laplacian sequence $\{\mathbf{L}(k)\}$ as $\overline{\mathbf{L}}_k = \mathbb{E}\left[\mathbf{L}(k)\right]$ and $\widetilde{\mathbf{L}}(k) = \mathbf{L}(k) - \overline{\mathbf{L}}_k$. Note that, $\mathbb{E}\left[\widetilde{\mathbf{L}}(k)\right] = \mathbf{0}$, and

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{L}}(k)\right\|^2\right] \leq 4N^2\mathbb{E}\left[\widetilde{\mathbf{L}}_{i,j}^2(k)\right] = \frac{4N^2\beta_0\rho_0^2}{(k+1)^{\tau+\epsilon}} - \frac{4N^2\beta_0^2}{(k+1)^{2\tau}}, \tag{9.49}$$

where $\|\cdot\|$ denotes the $\mathcal{L}_2$ norm. The above equation follows by relating the $\mathcal{L}_2$ and Frobenius norms.
We also have that, $\overline{\mathbf{L}}_k = \beta_k\overline{\mathbf{L}}$, where

$$\overline{\mathbf{L}}_{i,j} = \begin{cases} -1 & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l\neq i}\overline{\mathbf{L}}_{i,l} & i = j. \end{cases} \tag{9.50}$$

Technically speaking, the communication graph at each time k encapsulated as $\mathbf{L}(k)$ need not be connected at all times, although the graph of allowable links G is connected.. In fact, at any given time k, only a few of the possible links could be active. However, since $\overline{\mathbf{L}}_k = \beta_k\overline{\mathbf{L}}$, we note that, by Assumption 9.5.1, the instantaneous Laplacian $\mathbf{L}(k)$ is connected on average.The connectedness in average basically ensures that over time, the information from each agent in the graph reaches other agents over time in a symmetric fashion and thus ensuring information flow, while providing the leeway for the instantaneous communication graphs at different times to be not connected.
We employ a primal algorithm for solving the optimization problem in (9.1). In particular, the update in (9.2) can then be written in a vector form as follows:

$$\mathbf{x}(k+1) = \mathbf{W}_k\mathbf{x}(k) - \alpha_k\widehat{\mathbf{G}}(\mathbf{x}(k)), \tag{9.51}$$

where $\mathbf{x}(k) = \left[\mathbf{x}_1^\top(k), \cdots, \mathbf{x}_N^\top(k)\right]^\top \in \mathbb{R}^{Nd}$, $F(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x}_i)$, $\mathbf{x} = \left[\mathbf{x}_1^\top, \cdots, \mathbf{x}_N^\top\right]^\top \in \mathbb{R}^{Nd}$, $\widehat{\mathbf{G}}(\mathbf{x}(k)) = [\widehat{\mathbf{g}}_i^\top(\mathbf{x}_i(k)), \cdots, \widehat{\mathbf{g}}^\top(\mathbf{x}_N(k))]^\top$ and $\mathbf{W}_k = (\mathbf{I} - \mathbf{L}(k)) \otimes \mathbf{I}_d$. We state an assumption on the weight sequences before proceeding further.

Assumption 9.5.2. The weight sequence α_k is given by $\alpha_0/(k+1)$, where $\alpha_0 > 1/\mu$. For the sequence ρ_k as defined in (5.3), it is chosen in such a way that,

$$\rho_0^2 \leq \frac{4N^2}{\lambda_2\left(\overline{\mathbf{L}}\right)}. \tag{9.52}$$

In the following sections, we propose two different approaches to solve the optimization problem in (9.1). The first approach involves zeroth order optimization, while the second approach involves a first order optimization. We first study the zeroth order approach to the problem in (9.1).

9.6 Communication Efficient Zeroth Order: RDSA

We employ a random directions stochastic approximation (RDSA) type method from Nesterov and Spokoiny (2011) adapted to our distributed setup to solve (9.1). Each node i, $i = 1, ..., N$, in our setup maintains a local

copy of its local estimate of the optimizer $\mathbf{x}_i(k) \in \mathbb{R}^d$ at all times. In addition to the smoothness assumption in 2.3.1, we define additional smoothness assumptions in the context of zeroth order optimization.

Assumption 9.6.1. For all $i = 1, ..., N$, the functions $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ have their Hessian to be M-Lipschitz, i.e.,

$$\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{y})\| \leq M \|\mathbf{x} - \mathbf{y}\|, \forall i = 1, \cdots, N.$$

In order to carry out the optimization, each agent i makes queries to the $\mathcal{SZO}$ at time k, from which the agent obtains noisy function values of $f_i(\mathbf{x}_i(k))$. Denote the noisy value of $f_i(\cdot)$ as $\widehat{f}_i(\cdot)$ where,

$$\widehat{f}_i(\mathbf{x}_i(k)) = f_i(\mathbf{x}_i(k)) + \widehat{v}_i(k; \mathbf{x}_i(k)), \tag{9.53}$$

where the first argument in $\widehat{v}_i(k; \mathbf{x}_i(k))$ is the iteration number, and the second argument is the point at which the $\mathcal{SZO}$ oracle is queried. The properties of the noise $\widehat{v}_i(k; \mathbf{x}_i(k))$ are discussed further ahead. Typically due to the unavailability of the analytic form of the functionals in zeroth order methods, the gradient cannot be explicitly evaluated and hence, we resort to a gradient approximation. In order to approximate the gradient, each agent makes three calls to the stochastic zeroth order oracle. For instance, agent i queries for $f_i(\mathbf{x}_i(k) + c_k \mathbf{z}_{i,k})$, $f_i(\mathbf{x}_i(k) + c_k \mathbf{z}_{i,k}/2)$ and $f_i(\mathbf{x}_i(k))$ at time k and obtains $\widehat{f}_i(\mathbf{x}_i(k) + c_k \mathbf{z}_{i,k})$, $\widehat{f}_i(\mathbf{x}_i(k) + c_k \mathbf{z}_{i,k}/2)$ and $\widehat{f}_i(\mathbf{x}_i(k))$ respectively, where c_k is a carefully chosen time-decaying constant and $\mathbf{z}_{i,k}$ is a random vector (to be specified soon) such that $\mathbb{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$.
Denote by $\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))$ the approximated gradient which is given by:

$$\begin{aligned}
\widehat{\mathbf{g}}_i(\mathbf{x}_i(k)) &\doteq 2\widetilde{\mathbf{g}}_i\left(\mathbf{x}_i(k), \frac{c_k}{2}\right) - \widetilde{\mathbf{g}}_i\left(\mathbf{x}_i(k), c_k\right) \\
&= \frac{4\widehat{f}_i\left(\mathbf{x}_i(k) + \frac{c_k}{2}\mathbf{z}_{i,k}\right) - 4\widehat{f}_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k} \\
&- \frac{\widehat{f}_i\left(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}\right) - \widehat{f}_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k},
\end{aligned} \tag{9.54}$$

where $\widetilde{\mathbf{g}}_i(\cdot, \cdot)$ represents a first order finite difference operation and $\theta_1, \theta_2 \in [0, 1]$. Note that, the gradient approximation derived in (9.54) involves the noise in the retrieved function value from the $\mathcal{SZO}$ differently from other RDSA approaches such as in Duchi et al. (2015); Nesterov and Spokoiny (2011). The finite difference technique used in (9.54) resembles, *the twicing trick* commonly used in Kernel density estimation which is employed so as to reduce bias and approximately eliminate the effect of the second degree term from the bias. It is also to be noted that the number of queries made to the $\mathcal{SZO}$ at every gradient approximation is 3. Thus, we can write,

$$\begin{aligned}
\widehat{\mathbf{g}}_i(\mathbf{x}_i(k)) &= \nabla f_i\left(\mathbf{x}_i(k)\right) + \underbrace{\mathbb{E}\left[\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))|\mathcal{F}_k\right] - \nabla f_i\left(\mathbf{x}_i(k)\right)}_{c_k \mathbf{b}_i(\mathbf{x}_i(k))} \\
&+ \underbrace{\mathbf{g}_i(\mathbf{x}_i(k)) - \mathbb{E}\left[\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))|\mathcal{F}_k\right] + \frac{v_i(k; \mathbf{x}_i(k))\mathbf{z}_{i,k}}{c_k}}_{\mathbf{h}_i(\mathbf{x}_i(k))},
\end{aligned} \tag{9.55}$$

where

$$\begin{aligned}
\mathbf{g}_i(\mathbf{x}_i(k)) &= \frac{4f_i\left(\mathbf{x}_i(k) + \frac{c_k}{2}\mathbf{z}_{i,k}\right) - 4f_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k} \\
&- \frac{f_i\left(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}\right) - f_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k},
\end{aligned} \tag{9.56}$$

$$v_i(k;\mathbf{x}_i(k)) = 4\left(\widehat{f}_i\left(\mathbf{x}_i(k)+\frac{c_k}{2}\mathbf{z}_{i,k}\right) - f_i\left(\mathbf{x}_i(k)+\frac{c_k}{2}\mathbf{z}_{i,k}\right)\right)$$
$$- 3(\widehat{f}_i(\mathbf{x}_i(k)) - f_i(\mathbf{x}_i(k))) - (\widehat{f}_i\left(\mathbf{x}_i(k)+c_k\mathbf{z}_{i,k}\right)$$
$$- f_i\left(\mathbf{x}_i(k)+c_k\mathbf{z}_{i,k}\right)), \tag{9.57}$$

and, $\mathcal{F}_k$ denotes the history of the proposed algorithm up to time k. Given that the sources of randomness in our algorithm are the noise sequence $\{\mathbf{v}(k;\mathbf{x}(k))\}$, the random network sequence $\{\mathbf{L}(k)\}$ and the random vectors for directional derivatives $\{\mathbf{z}_k\}$, $\mathcal{F}_k$ is given by the σ-algebra generated by the collection of random variables $\{\mathbf{L}(s),\ \mathbf{v}(k;\mathbf{x}(k)),\ \mathbf{z}_{i,s}\}$, $i = 1,...,N,\ s = 0,...,k-1$.

In general, the higher order smoothness imposed by Assumption 2.4.1 allows us to use a higher order finite difference approximation for estimating the gradient. Due to assumption 9.5.2, the bias in the gradient estimate by employing a second order finite difference approximation of the gradient is of the order $O(c_k^2)$. Instead, a first order finite difference approximation of the gradient would have yielded a bias of $O(c_k)$. More generally, an assumption involving p-th order smoothness of the loss functions would have enabled usage of a p-th degree finite difference approximation of the gradient thus leading to a bias of $O(c_k^p)$.

Assumption 9.6.2. The $z_{i,k}$'s are drawn from a distribution P such that $\mathbb{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$, $s_1(P) = \mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right]$ and $s_2(P) = \mathbb{E}\left[\|\mathbf{z}_{i,k}\|^6\right]$ are finite.

We provide two examples of two such distributions. If $\mathbf{z}_{i,k}$'s are drawn from $\mathcal{N}(0,\mathbf{I}_d)$, then $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right] = d(d+2)$ and $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^6\right] = d(d+2)(d+4)$. If $\mathbf{z}_{i,k}$'s are drawn uniformly from the l_2-ball of radius $\sqrt{d}$, then we have, $\|\mathbf{z}_{i,k}\| = \sqrt{d}$, $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right] = d^2$ and $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right] = d^3$. For the rest of the chapter, we assume that $\mathbf{z}_{i,k}$'s are sampled from a normal distribution with $\mathbf{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$ or uniformly from the surface of the l_2-ball of radius $\sqrt{d}$.

Remark 9.6.1. *The RDSA scheme (see, for example Nesterov and Spokoiny (2011)) used here is similar to the simultaneous perturbation stochastic approximation scheme (SPSA) as proposed in Spall (1992). In SPSA, each dimension i of the optimization iterate is perturbed by a random variable Δ_i. However, instead of RDSA where the directional derivative is taken along the sampled vector $\mathbf{z}$, the directional derivative in case of SPSA is along the direction $[1/\Delta_1,\cdots,1/\Delta_d]$ which thus needs boundedness of the inverse moments of the random variable Δ_i. The particular choice for Δ_i's is taken to be the Bernoulli distribution with Δ_i's taking values 1 and -1 with probability 0.5. It is to be noted that at each iteration, both RDSA and SPSA approximate the gradient by making two calls to the stochastic zeroth order oracle as opposed to d calls in the case of Kiefer Wolfowitz Stochastic Approximation (KWSA) (see, Kiefer and Wolfowitz (1952) for example).*

For arbitrary deterministic initializations $\mathbf{x}_i(0) \in \mathbb{R}^d$, $i = 1,...,N$, the optimizer update rule at node i and $k = 0,1,...,$ is given as follows:

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k) - \sum_{j\in\Omega_i(k)} \psi_{i,k}\psi_{j,k}\left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right)$$
$$- \alpha_k\widehat{\mathbf{g}}_i(\mathbf{x}_i(k)), \tag{9.58}$$

where $\widehat{\mathbf{g}}_i(\cdot)$ is as defined in (9.55). Comparing to the general update in (9.2), the time-varying weight $\gamma_{i,j}(k)$ at agent i to the incoming message from agent j is given by $\psi_{j,k}$.

Remark 9.6.2. *The main intuition behind the randomized activation albeit in a controlled manner for both the zeroth order and first order optimization methods is the fact that in expectation both the updates exactly reduce to the update where the communication graph between agents is realized by the expected Laplacian.*

It is to be noted that unlike first order stochastic gradient methods, where the algorithm has access to unbiased estimates of the gradient, the local gradient estimates $\mathbf{g}_i(\cdot)$ used in (9.58) are biased (see (9.55)) due to the unavailability of the exact gradient functions and their approximations using the zeroth order scheme in (9.54). The update is carried on in all agents parallely in a synchronous fashion. The weight sequences $\{\alpha_k\}$, $\{c_k\}$ and $\{\beta_k\}$ are given by $\alpha_k = \alpha_0/(k+1)$, $c_k = c_0/(k+1)^\delta$ and $\beta_k = \beta_0/(k+1)^\tau$ respectively, where $\alpha_0, c_0, \beta_0 > 0$. We state an assumption on the weight sequences before proceeding further.

Assumption 9.6.3. The sequence c_k is given by:

$$c_k = \frac{1}{s_1(P)(k+1)^\delta}, \tag{9.59}$$

where $\delta > 0$. The constant $\delta > 0$ is chosen in such a way that,

$$\sum_{k=1}^{\infty} \frac{\alpha_k^2}{c_k^2} < \infty \tag{9.60}$$

The update in (9.58) can be written as:

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{W}_k \mathbf{x}(k) - \alpha_k \nabla F(\mathbf{x}(k)) - \alpha_k c_k \mathbf{b}(\mathbf{x}(k)) \\ &- \alpha_k \mathbf{h}(\mathbf{x}(k)), \end{aligned} \tag{9.61}$$

where $\mathbf{b}(\mathbf{x}(k)) = \left[\mathbf{b}_1^\top(\mathbf{x}_1(k)), \cdots, \mathbf{b}_N^\top(\mathbf{x}_N(k))\right]^\top \in \mathbb{R}^{Nd}$ and $\mathbf{h}(\mathbf{x}(k)) = \left[\mathbf{h}_1^\top(\mathbf{x}_1(k)), \cdots, \mathbf{h}_N^\top(\mathbf{x}_N(k))\right]^\top \in \mathbb{R}^{Nd}$. We state an assumption on the measurement noises next.

Assumption 9.6.4. For each $i = 1, ..., N$, the sequence of measurement noises $\{v_i(k; \mathbf{x}_i(k))\}$ satisfies for all $k = 0, 1, ...$:

$$\begin{aligned} &\mathbb{E}[\, v_i(k; \mathbf{x}_i(k)) \,|\, \mathcal{F}_k, \mathbf{z}_{i,k}] = 0, \text{ almost surely (a.s.)} \\ &\mathbb{E}[\, v_i(k; \mathbf{x}_i(k))^2 \,|\, \mathcal{F}_k, \mathbf{z}_{i,k}] \leq c_v \|\mathbf{x}_i(k)\|^2 + \sigma_v^2, \text{ a.s.}, \end{aligned} \tag{9.62}$$

where c_v and σ_v^2 are nonnegative constants.

Assumption 9.6.4 is standard in the analysis of stochastic optimization methods, e.g., Towfic et al. (2016). It is stated in terms of noise $\mathbf{v}_i(k; \mathbf{x}_i(k))$ in (9.57) rather then directly in terms of the $\mathcal{SZO}$ noises in equation (9.53), for notational simplicity. An equivalent statement can be made in terms of the noises in (9.53). The assumption about the conditional independence between the random directions $\mathbf{z}_{i,k}$ and the function noise $v_i(k; \mathbf{x}_i(k))$ is mild. It merely formalizes the model that we consider, namely that, given history $\mathcal{F}_k$, drawing a random direction sample $\mathbf{z}_{i,k}$ and querying function values from the $\mathcal{SZO}$ are performed in a statistically independent manner.
We remark that by Assumption 9.6.4,

$$\begin{aligned} &\mathbb{E}\left[v_i(k; \mathbf{x}_i(k)) \mathbf{z}_{i,k} | \mathcal{F}_k\right] = \mathbb{E}\left[\mathbf{z}_{i,k} \mathbb{E}\left[v_i(k; \mathbf{x}_i(k)) | \mathcal{F}_k, \mathbf{z}_{i,k}\right] \mid \mathcal{F}_k\right] \\ &\Rightarrow \mathbb{E}\left[\mathbf{v}_{\mathbf{z}}(k; \mathbf{x}(k)) \mid \mathcal{F}_k\right] = \mathbf{0}. \end{aligned} \tag{9.63}$$

and,

$$\begin{aligned}&\mathbb{E}\left[\|v_i(k;\mathbf{x}_i(k))\mathbf{z}_{i,k}\|^2\,|\mathcal{F}_k\right]\\&=\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^2\,\mathbb{E}\left[v_i^2(k;\mathbf{x}_i(k))|\mathcal{F}_k,\mathbf{z}_{i,k}\right]|\mathcal{F}_k\right]\\&\leq\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^2\right]\left(c_v\|\mathbf{x}_i(k)\|^2+\sigma_v^2\right),\end{aligned}\tag{9.64}$$

where if $\mathbf{z}_{i,k}$'s are sampled from a normal distribution with $\mathbf{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right]=\mathbf{I}_d$ or uniformly from the surface of the l_2-ball of radius $\sqrt{d}$, then we have,

$$\mathbb{E}\left[\|v_i(k;\mathbf{x}_i(k))\mathbf{z}_{i,k}\|^2\,|\mathcal{F}_k\right]\leq d\left(c_v\|\mathbf{x}_i(k)\|^2+\sigma_v^2\right).\tag{9.65}$$

Communication Cost Define the communication cost $\mathcal{C}_k$ to be the expected per-node number of transmissions up to iteration k, i.e.,

$$\mathcal{C}_k=\mathbb{E}\left[\sum_{s=0}^{k-1}\mathbb{I}_{\{node\ C\ transmits\ at\ s\}}\right],\tag{9.66}$$

where $\mathbb{I}_A$ represents the indicator of event A. Note that the per-node communication cost in (9.66) is the same as the network average of communication costs across all nodes, as the activation probabilities are homogeneous across nodes. We now proceed to the main results pertaining to the proposed zeroth order optimization scheme.

9.7 Convergence rates: Statement of main results and interpretations

In this section, we state the main results while the proofs are relegated to Appendix H.

9.7.1 Main Results: RDSA

We state the main result concerning the mean square error at each agent i next.

Theorem 9.7.1. *1) Consider the optimizer estimate sequence $\{\mathbf{x}(k)\}$ generated by the algorithm (9.58). Let assumptions 9.3.1-9.5.2 and 9.6.1-9.6.4 hold. Then, for each node i's optimizer estimate $\mathbf{x}_i(k)$ and the solution $\mathbf{x}^\star$ of problem (9.1), $\forall k\geq 0$ there holds:*

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]&\leq 2M_k+\frac{64NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{2-2\tau-2\delta}}\\&\frac{16NM^2d^2(P)c_0^4}{\mu^2(k+1)^{4\delta}}+2Q_k+\frac{8\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}}\\&+\frac{4N\alpha_0\left(dc_vq_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{1-2\delta}},\end{aligned}\tag{9.67}$$

where, $\Delta_{1,\infty}=6dc_vq_\infty(N,d,\alpha_0,c_0)+6dN\sigma_1^2$ and $q_\infty(N,d,\alpha_0,c_0)=\mathbb{E}\left[\|\mathbf{x}(k_2)-\mathbf{x}^o\|^2\right]+4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}+\frac{\sqrt{N}s_1(P)M\alpha_0c_0^2}{8\delta}+\frac{Ns_1^2(P)M^2\alpha_0^2c_0^4}{16(1+4\delta)}+\frac{d\alpha_0^2\left(2c_vN\|\mathbf{x}^o\|^2+N\sigma_v^2\right)}{c_0^2(1-2\delta)}+\frac{\alpha_0^2c_0^2\sqrt{N}s_1(P)M\|\nabla F(\mathbf{x}^o)\|}{1+2\delta}+\frac{2N\alpha_0^2c_0^4s_2(P)}{1+4\delta}+\frac{4\alpha_0^2c_0^2Ns_1(P)}{1+2\delta}\|\nabla F\left(\mathbf{x}^o\right)\|^2$, $k_2=\max\{k_0,k_1\}$, $k_0=\inf\{k|\mu^2\alpha_k^2<1\}$ and $k_1=\inf\left\{k|\frac{\mu}{2}>\frac{\sqrt{N}}{4}s_1(P)Mc_k^2+\frac{2dc_v\alpha_k}{c_k^2}+4\alpha_kc_k^2Ns_1(P)L^2\right\}$, with M_k and Q_k decaying faster than the rest of the terms.

2) In particular, the rate of decay of the RHS of (9.67) is given by $(k+1)^{-\delta_1}$, where $\delta_1=\min\{1-2\delta,2-2\tau-2\delta,4\delta\}$.

By, optimizing over τ and δ, we obtain that for $\tau = 1/2$ and $\delta = 1/6$,

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] \leq 2M_k + \frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{2/3}}$$
$$\frac{16NM^2d^2(P)c_0^2}{\mu^2(k+1)^{2/3}} + 2Q_k + \frac{8\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2c_0^2(k+1)^{2/3}}$$
$$+ \frac{4N\alpha_0\left(dc_v q_\infty(N,d,\alpha_0,c_0) + dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{2/3}} = O\left(\frac{1}{k^{\frac{2}{3}}}\right), \quad \forall i.$$

3) The communication cost is given by,

$$\mathbb{E}\left[\sum_{t=1}^{k}\zeta_t\right] = O\left(k^{\frac{3}{4}+\frac{\epsilon}{2}}\right).$$

and the MSE-communication rate is given by,

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O\left(\mathcal{C}_k^{-8/9+\zeta}\right), \tag{9.68}$$

where ζ can be arbitrarily small.

Theorem 9.7.1 asserts an $O\left(\mathcal{C}_k^{-8/9+\zeta}\right)$ MSE-communication rate can be achieved while keeping the MSE decay rate at $O\left(k^{-\frac{2}{3}}\right)$. The performance of the zeroth order optimization scheme depends explicitly on the connectivity of the expected Laplacian through the terms $\frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2(\overline{\mathbf{L}})c_0^2\beta_0^2(k+1)^{0.5}}$ and $\frac{8\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2(\overline{\mathbf{L}})\beta_0^2c_0^2(k+1)^{0.5}}$. In particular, communication graphs which are well connected, i.e., have higher values of $\lambda_2\left(\overline{\mathbf{L}}\right)$ will have lower MSE as compared to a counterpart with lower values of $\lambda_2\left(\overline{\mathbf{L}}\right)$.

If higher order smoothness assumptions are made, i.e., a p-th order smoothness assumption is made which is then exploited by means of a p-th degree finite difference gradient approximation, then by repeating the same proof arguments, the rate in terms of iteration count can be shown to improve to $O\left(k^{-\frac{p}{p+1}}\right)$. The improvement can be attributed to a better bias-variance tradeoff as illustrated by the terms $\frac{8M^2d^2(P)c_0^4}{\mu^2(k+1)^{2p\delta}}$ and $\frac{4N\alpha_0\left(dc_v q_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{1-2\delta}}$. The corresponding MSE-communication rate improves to $O\left(\mathcal{C}_k^{-\frac{4p}{3(p+1)}+\zeta}\right)$.

9.8 Simulations

9.8.1 Distributed KWSA

We provide a simulation example pertaining to ℓ_2-regularized logistic losses in random network characterized by link failures independent across iteration and links with probability p_{fail}. To be specific, we consider ℓ_2-regularized empirical risk minimization with logistic loss, where the regularization function is given by $\Psi_i(\mathbf{x}) = \frac{\kappa}{2}\|\mathbf{x}\|^2$, $i = 1, ..., N$, with $\kappa = 0.3$. In our simulation setup, each node has access to $n_i = 10$ data points. The class labels and the classification vector given by $b_{ij} = \text{sign}\left((\mathbf{x}_1')^\top\mathbf{a}_{i,j} + x_0' + \epsilon_{ij}\right)$ and $x' = ((\mathbf{x}_1')^\top, x_0')^\top$ respectively have ϵ_{ij}s and the entries of x' drawn independently from standard normal distribution. The feature vectors $\mathbf{a}_{i,j}$, $j = 1, ..., n_i$, across different nodes $i = 1, \cdots, N$ and across different entries are drawn independently from different distributions. To be specific, at node i, $\mathbf{a}_{i,j}$, $j = 1, ..., n_i$ is generated by adding a standard normal random variable and an uniform random variable with support $[0,\, 5\,i]$.

We set $\beta_k = \frac{1}{\theta\,(k+1)^{1/2}}$, $\alpha_k = \frac{1}{k+1}$, $c_k = \frac{1}{(k+1)^{1/4}}$, $k = 0, 1, ...$, where $\theta = 7$ is the maximum degree across nodes. The optimizer estimate at each node is initialized as $\mathbf{x}_i(0) = 0$, $\forall\ i = 1, ..., N$.

We consider a connected network $\mathcal{G}$ with $N = 10$ nodes and 23 links, generated as an instance of a random geometric graph. The random network model assumes link failures independent across iterations and links with probability p_{fail}, where $p_{\text{fail}} \in \{0;\ 0.5;\ 0.7\}$. The case $p_{\text{fail}} = 0$ corresponds to the case where none of the links fail. We also include a comparison with the centralized zeroth order KWSA based optimization method:

$$\mathbf{y}(k+1) = \mathbf{y}(k) - \frac{1}{N(k+1)} \sum_{i=1}^{N} \nabla g_i\left(\mathbf{y}(k);\ \mathbf{a}_i(k), b_i(k)\right), \tag{9.69}$$

where $(\mathbf{a}_i(k), b_i(k))$ is drawn uniformly from the set $(\mathbf{a}_{i,j}, b_{i,j})$, $j = 1, ..., n_i$. Algorithm (9.69) shows how (9.58) would be implemented if there existed a fusion node with access to all nodes' data. Hence, the comparison with (9.69) allows us to study the degradation of (9.6) due to lack of global model information. The step size for (9.69) is set to $1/N(k+1)$. As an error metric, we use the mean square error (MSE) estimate averaged across nodes: $\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2$.

Figure 9.1 plots the estimated MSE, averaged across 100 algorithm runs, versus iteration number k for $p_{\text{fail}} \in \{0;\ 0.5;\ 0.7\}$ in $\log_{10}$-$\log_{10}$ scale. The slope of the plot curve corresponds to the sublinear rate of the method; e.g., the $-1/2$ slope corresponds to a $1/k^{0.5}$ rate. It is to be noted that for all values of p_{fail}, the algorithm (9.58) achieves on this example (at least) the $1/k^{0.5}$ rate, thus corroborating our theory. The increase of the link failure probability only increases the constant in the MSE but does not affect the rate but the curves are only vertically shifted. Interestingly, the loss due to the increase of p_{fail} is small; e.g., the curves that correspond to $p_{\text{fail}} = 0.5$ and $p_{\text{fail}} = 0$ (no link failures) practically match. Figure 9.1 also shows the performance of the centralized method (9.69). We can see that, the distributed method (9.6) is very close in performance to the centralized method.

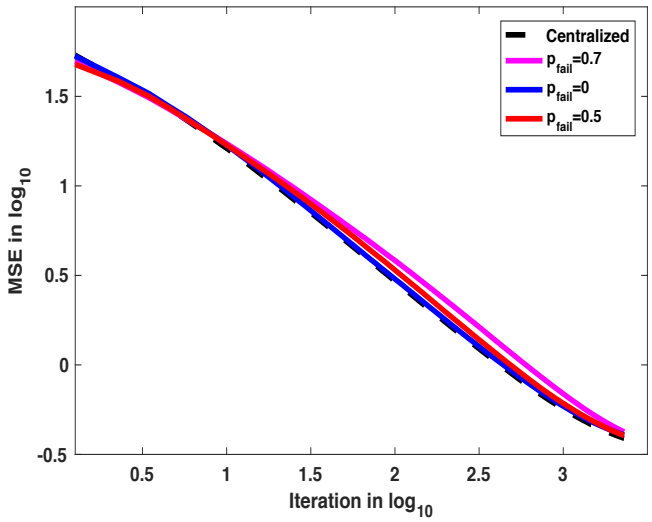


Figure 9.1: Estimated MSE versus iteration number k for algorithm (9.58) with link failure probability $p_{\text{fail}} = 0$ (blue, solid line); 0.5 (red, solid line); and 0.7 (pink, solid line). The Figure also shows the performance of the centralized stochastic gradient method in (9.69) (black, dashed line).

9.8.2 Communication Efficient RDSA

In this section, we provide evaluations of the proposed communication efficient zeroth order optimization algorithm on the Abalone dataset (Lib). To be specific, we consider ℓ_2-regularized empirical risk minimization

for the Abalone dataset, where the regularization function is given by $\Psi_i(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$. We consider a 10 node network for both the zeroth and first order optimization schemes. The Abalone dataset has 4177 data points out of which 577 data points are kept aside as the test set and the other 3600 is divided equally among the 10 nodes resulting in each node having 360 data points. For the zeroth order optimization, we compare the proposed undirected sequence of Laplacian constructions based optimization scheme and the static Laplacian (Benchmark) based optimization schemes. The benchmark scheme is characterized by the communication graph being static and thereby resulting agents connected through a link to exchange messages at all times. The data points at each node are sampled without replacement in a contiguous manner. The vectors $\mathbf{z}_{i,k}$s for evaluating directional derivatives were sampled from a normal distribution with identity covariance. Figure 9.2 compares the test error for the three aforementioned schemes, where it can be clearly observed that the test error is indistinguishable in terms of the number of iterations or equivalently in terms of the number of queries to the stochastic zeroth oracle. Figure 9.3 demonstrates the superiority the proposed algorithm in terms of the test error versus communication cost as compared to the benchmark as predicted by Theorem 9.7.1. For example, at the same relative test error level, the proposed algorithm uses up to 3x less number of transmissions as compared to the benchmark scheme.

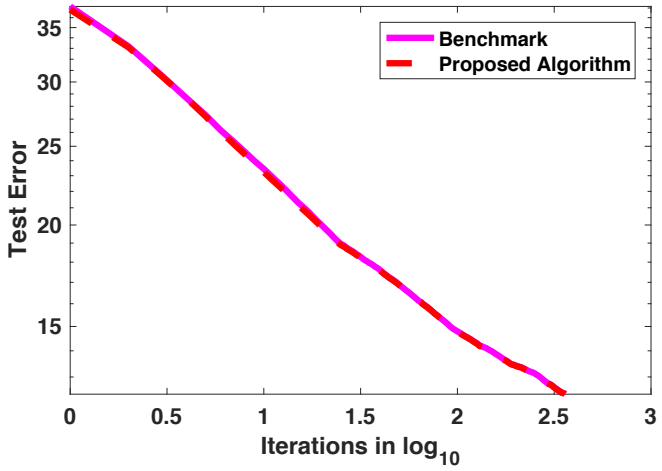


Figure 9.2: Communication Efficient RDSA: Test Error vs Iterations

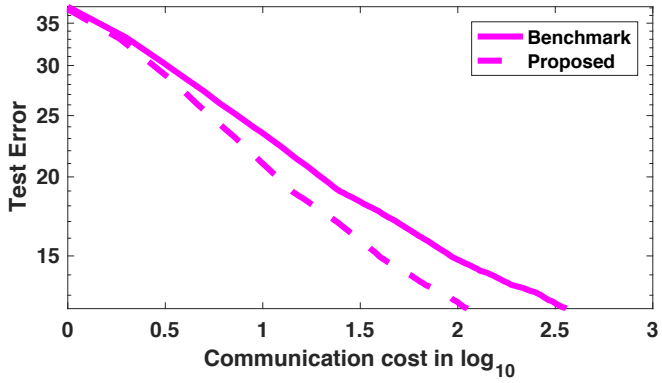


Figure 9.3: Communication Efficient RDSA: Test Error vs Communication Cost

9.9 Contributions

- **Non-Asymptotic Rates for Distributed KWSA**: Through the distributed KWSA algorithm, we specifically characterized non-asymptotic rates of the algorithm in terms of the different algorithm parameters and the network connectivity. While asymptotic properties of KWSA in terms of almost sure convergence and asymptotic normality of the optimizer sequence has been studied in Kiefer and Wolfowitz (1952), to the best of our knowledge this is the first time a non-asymptotic characterization of a distributed implementation of KWSA has been obtained.
- **Bias-Reduced Communication Efficient RDSA**: We proposed a communication efficient distributed zeroth order scheme akin to the RDSA scheme albeit using three function evaluation at each epoch which in spirit draws from the twicing trick in kernel density estimation. In addition to the twicing trick, we have established for the proposed zeroth order method explicit mean square error (MSE) convergence rates with respect to (appropriately defined) computational cost C_{comp} and communication cost C_{comm}. Specifically, the proposed zeroth order method achieves the $O(1/(C_{\text{comm}})^{8/9-\zeta})$ MSE communication rate, which significantly improves over the rates of existing methods, while maintaining the order-optimal $O(1/(C_{\text{comp}})^{2/3})$ MSE computational rate.

9.10 Conclusion and Future Directions

In this chapter, we have developed and analyzed a novel class of methods for distributed stochastic optimization of the zeroth and first order that are based on increasingly sparse randomized communication protocols. We have established for the proposed zeroth order method explicit mean square error (MSE) convergence rates with respect to (appropriately defined) computational cost C_{comp} and communication cost C_{comm}. Specifically, the proposed zeroth order method achieves the $O(1/(C_{\text{comm}})^{8/9-\zeta})$ MSE communication rate, which significantly improves over the rates of existing methods, while maintaining the order-optimal $O(1/(C_{\text{comp}})^{2/3})$ MSE computational rate. Numerical examples on real data demonstrate the communication efficiency of the proposed methods. Future directions include extending the communication efficient scheme to non-convex and non-smooth functions.

Chapter 10

Zeroth Order Frank Wolfe

10.1 Introduction

In this chapter, we aim to solve the following stochastic optimization problem:

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) = \min_{x\in\mathcal{C}} \mathbb{E}_{\mathbf{y}\sim\mathcal{P}}\left[F(\mathbf{x};\mathbf{y})\right], \tag{10.1}$$

where $\mathcal{C} \in \mathbb{R}^d$ is a closed convex set. This problem of stochastic constrained optimization has been a focus of immense interest in the context of convex functions Bubeck et al. (2015) and non-convex functions especially in the context of deep learning Goodfellow et al. (2016). Solutions to the problem (10.1) can be broadly classified into two classes: algorithms which require a projection at each step, for example, projected gradient descent Bubeck et al. (2015) and projection free methods such as the Frank-Wolfe algorithm Jaggi (2013). Furthermore, algorithms designed to solve the above optimization problem access various kinds of oracles, i.e., first order oracle (gradient queries) and zeroth order oracle (function queries). In this chapter, we focus on a stochastic version of projection free method, namely Frank-Wolfe algorithm, with access to a zeroth order oracle.

Derivative free optimization or zeroth order optimization is motivated by settings where the analytical form of the function is not available or when the gradient evaluation is computationally prohibitive. Developments in zeroth order optimization has been fueled by various applications ranging from problems in medical science, material science and chemistry Gray et al. (2004); Marsden et al. (2008); Gray et al. (2004); Deming et al. (1978); Marsden et al. (2007). In the context of machine learning, zeroth order methods have been applied to attacks on deep neural networks using black box models Chen et al. (2017), scalable policy optimization for reinforcement learning Choromanski et al. (2018) and optimization with bandit feedback Bubeck et al. (2012). For the problem in (10.1), it is well known that the primal sub-optimality gap of first order schemes are dimension independent. However, algorithms which involve a projection operator might be expensive in practice depending on the structure of $\mathcal{C}$. Noting the potentially expensive projection operators, projection free methods such as Frank-Wolfe Jaggi (2013) have had a resurgence. Frank-Wolfe avoids the projection step, and only requires access to a linear minimization oracle, which can be implemented efficiently and needs to be solved to a certain degree of exactness. Stochastic versions of Frank-Wolfe have been studied in both the convex Hazan and Kale (2012); Hazan and Luo (2016); Mokhtari et al. (2018) and non-convex Reddi et al. (2016) setting with access to stochastic first order oracles (SFO). However, convergence of stochastic Frank-Wolfe with access to only stochastic zeroth order oracle (SZO) remains unexplored.

Reference	Setting	Memory	Primal Rate	Oracle
Jaggi (2013)	Det. Convex	-	$O(1/t)$	SFO
Hazan and Kale (2012)	Stoch. Convex	$O(t)$	$O(1/t^{1/2})$	SFO
Mokhtari et al. (2018)	Stoch. Convex	$O(1)$	$O(1/t^{1/3})$	SFO
Lacoste-Julien (2016)	Det. Non-convex	-	$O(1/t^{1/2})$	SFO
Reddi et al. (2016)	Stoch. Non-convex	$O(\sqrt{t})$	$O(1/t^{1/4})$	SFO
RDSA [Theorem 10.5.2(1)]	Stoch. Convex	1	$O(d^{1/3}/t^{1/3})$	SZO
I-RDSA [Theorem 10.5.2(2)]	Stoch. Convex	m	$O((d/m)^{1/3}/t^{1/3})$	SZO
KWSA [Theorem 10.5.2(3)]	Stoch. Convex	d	$O(1/t^{1/3})$	SZO
I-RDSA [Theorem 10.5.3]	Stoch. Non-convex	m	$O((d/m)^{1/3}/t^{1/4})$	SZO

Table 10.1: Convergence of Frank-Wolfe: Det. refers to deterministic while stoch. refers to stochastic. Memory indicates the number of samples at which the gradients needs to be tracked in the first order case. In the zeroth order case, it indicates the number of directional derivatives being evaluated at one sample. The rates correspond to the rate of decay of $\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)]$ in the convex setting and the Frank-Wolfe duality gap in context of the non-convex setting.

10.2 Related Work

Algorithms for convex optimization with access to a SZO have been studied in Wang et al. (2018); Duchi et al. (2015); Liu et al. (2018); Sahu et al. (2018b), where in Liu et al. (2018) to address constrained optimization a projection step was considered. In the context of projection free methods, Frank and Wolfe (1956) studied the Frank-Wolfe algorithm for smooth convex functions with line search which was extended to encompass inexact linear minimization step in Jaggi (2013). Subsequently with additional assumptions, the rates for classical Frank-Wolfe was improved in Lacoste-Julien and Jaggi (2015); Garber and Hazan (2015). Stochastic versions of Frank-Wolfe for convex optimization with number of calls to SFO at each iteration dependent on the number of iterations with additional smoothness assumptions have been studied in Hazan and Kale (2012); Hazan and Luo (2016) so as to obtain faster rates, while Mokhtari et al. (2018) studied the version with a mini-batch size of 1. In the context of non-convex optimization, a deterministic Frank-Wolfe algorithm was studied in Lacoste-Julien (2016), while Reddi et al. (2016) addressed the stochastic version of Frank-Wolfe and further improved the rates by using variance reduction techniques. Table 10.1 gives a summary of the rates of various algorithms. For the sake of comparison, we do not compare our rates with those of variance reduced versions of stochastic Frank-Wolfe in Reddi et al. (2016); Hazan and Luo (2016), as our proposed algorithm does not employ variance reduction techniques which tend to incorporate multiple restarts and extra memory in order to achieve better rates. However, note that our algorithm can be extended so as to incorporate variance reduction techniques.

In this chapter, we study a setting of the stochastic Frank-Wolfe where a small batch-size (independent of dimension or the number of iterations) is sampled at each epoch while having access to a zeroth order oracle. Unlike, the first order oracle based stochastic Frank-Wolfe, the zeroth order counterpart is only able to generate biased gradient estimates. We focus on three different zeroth order gradient approximation schemes, namely, the classical Kiefer Wolfowitz stochastic approximation (KWSA) Kiefer and Wolfowitz (1952), random directions stochastic approximation (RDSA) Nesterov and Spokoiny (2011); Duchi et al. (2015), and an improvized RDSA (I-RDSA). KWSA samples directional derivatives along the canonical basis directions at each iteration, while RDSA samples one directional derivative at each iteration, and I-

RDSA samples $m < d$ directional derivatives at each iteration. Naïve usage of the biased gradient estimates in the linear minimization step, in addition to the stochasticity of the function evaluations, can lead to potentially diverging iterate sequences.

To circumvent the potential divergence issue due to non-decaying gradient noise and bias, we use a gradient averaging technique used in Yang et al. (2016); Ruszczyński (2008); Mokhtari et al. (2018) to get a surrogate gradient estimate which reduces the noise and the associated bias. The gradient averaging technique intuitively reduces the linear minimization step to that of an inexact minimization if the exact gradient was available. For each of the zeroth order optimization schemes, i.e., KWSA, RDSA, and I-RDSA, we derive primal sub-optimality bounds and Frank-Wolfe duality gap bounds and quantify the dependence in terms of the dimension and the number of epochs. We show that the primal sub-optimality gap to be of the order $O(d^{1/3}/T^{1/3})$ for RDSA, which improves to $O((d/m)^{1/3}/T^{1/3})$ for I-RDSA, and $O(1/T^{1/3})$ for KWSA at the cost of additional directional derivatives. The dimension dependence in zeroth order optimization is unavoidable due to the inherent bias-variance trade-off but nonetheless, the dependence on the number of iterations matches that of its first order counterpart in Mokhtari et al. (2018). Furthermore, we also derive rates for non-convex functions and show the Frank-Wolfe duality gap to be $O(d^{1/3}/T^{1/4})$, where the dependence on the number of iterations matches that of its first order counterpart in Reddi et al. (2016). To complement the theoretical results, we also demonstrate the efficacy of our algorithm through empirical evaluations on datasets. In particular, we perform experiments on a dataset concerning constrained black box non-convex optimization, where generic first order methods are rendered unusable and show that our proposed algorithm converges to a first order stationary point.

10.3 Frank-Wolfe: First to Zeroth Order

In this paper, the objective is to solve the following optimization problem:

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) = \min_{x\in\mathcal{C}} \mathbb{E}_{\mathbf{y}\sim\mathcal{P}}\left[F(\mathbf{x};\mathbf{y})\right], \tag{10.2}$$

where $\mathcal{C} \in \mathbb{R}^d$ is a closed convex set, the loss functions and the expected loss functions, $F(\cdot;\mathbf{y})$ and $f(\cdot)$ respectively are possibly non-convex. However, in the context of the optimization problem posed in (10.2), we assume that we have access to a stochastic zeroth order oracle (SZO). On querying a SZO at the iterate $\mathbf{x}_t$, yields an unbiased estimate of the loss function $f(\cdot)$ in the form of $F(\mathbf{x}_t;\mathbf{y}_t)$. Before proceeding to the algorithm and the subsequent results, we revisit preliminaries concerning the Frank-Wolfe algorithm and zeroth order optimization.

10.3.1 Background: Frank-Wolfe Algorithm

The celebrated Frank-Wolfe algorithm is based around approximating the objective by a first-order Taylor approximation. In the case, when exact first order information is available, i.e., one has access to an incremental first order oracle (IFO), a deterministic Frank-Wolfe method involves the following steps:

$$\begin{aligned}\mathbf{v}_t &= \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}} \langle \nabla f(\mathbf{x}_t), \mathbf{v}\rangle \\ \mathbf{x}_{t+1} &= (1-\gamma_{t+1})\,\mathbf{x}_t + \gamma_{t+1}\mathbf{v}_t,\end{aligned} \tag{10.3}$$

where $\gamma_t = \frac{2}{t+2}$. A linear minimization oracle (LMO) is queried at every epoch. Note that, the exact minimization in (3.1) is a linear program[1] and can be performed efficiently without much computational overload. It is worth nothing that the exact minimization in (3.1) can be replaced by an inexact minimization of the following form, where a $\mathbf{v} \in \mathcal{C}$ is chosen to satisfy,

$$\langle \nabla f(\mathbf{x}_t), \mathbf{v} \rangle \leq \underset{\mathbf{v} \in \mathcal{C}}{\operatorname{argmin}} \langle \nabla f(\mathbf{x}_t), \mathbf{v} \rangle + \gamma_t C_1,$$

and the algorithm can be shown to retain the same convergence rate (see, for example Jaggi (2013)).

10.3.2 Background: Zeroth Order Optimization

The crux of zeroth order optimization consists of gradient approximation schemes from appropriately sampled values of the objective function. We briefly describe the few well known zeroth order gradient approximation schemes. The Kiefer-Wolfowitz stochastic approximation (KWSA, see Kiefer and Wolfowitz (1952)) scheme approximates the gradient by sampling the objective function along the canonical basis vectors. Formally, gradient estimate can be expressed as:

$$\mathbf{g}(\mathbf{x}_t; \mathbf{y}) = \sum_{i=1}^{d} \frac{F(\mathbf{x}_t + c_t \mathbf{e}_i; \mathbf{y}) - F(\mathbf{x}_t; \mathbf{y})}{c_t} \mathbf{e}_i, \tag{10.4}$$

where c_t is a carefully chosen time-decaying sequence. KWSA requires d samples at each step to evaluate the gradient. However, in order to avoid sampling the objective function d times, random directions based gradient estimators have been proposed recently (see, for example Duchi et al. (2015); Nesterov and Spokoiny (2011)). The random directions gradient estimator (RDSA) involves estimating the directional derivative along a randomly sampled direction from an appropriate probability distribution. Formally, the random directions gradient estimator is given by,

$$\mathbf{g}(\mathbf{x}_t; \mathbf{y}, \mathbf{z}_t) = \frac{F(\mathbf{x}_t + c_t \mathbf{z}_t; \mathbf{y}) - F(\mathbf{x}_t; \mathbf{y})}{c_t} \mathbf{z}_t, \tag{10.5}$$

where $\mathbf{z}_t \in \mathbb{R}^d$ is a random vector sampled from a probability distribution such that $\mathbb{E}\left[\mathbf{z}_t \mathbf{z}_t^\top\right] = \mathbf{I}_d$ and c_t is a carefully chosen time-decaying sequence. With $c_t \to 0$, both the gradient estimators in (10.4) and (10.5) turn out to be unbiased estimators of the gradient $\nabla f(\mathbf{x}_t)$.

10.4 Zeroth Order Stochastic Frank-Wolfe: Algorithm & Analysis

In this section, we start by stating assumptions which are required for our analysis.

Assumption 10.4.1. In problem (10.2), the set $\mathcal{C}$ is bounded with finite diameter R.

Assumption 10.4.2. F is convex and Lipschitz continuous with $\sqrt{\mathbb{E}\left[\|\nabla_x F(\mathbf{x}; \cdot)\|^2\right]} \leq L_1$ for all $\mathbf{x} \in \mathcal{C}$.

Assumption 10.4.3. The expected function $f(\cdot)$ is convex. Moreover, its gradient ∇f is L-Lipschitz continuous over the set $\mathcal{C}$, i.e., for all $x, y \in \mathcal{C}$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|. \tag{10.6}$$

[1] Technically speaking, when $\mathcal{C}$ is given by linear constraints.

Algorithm 1 Deterministic Zeroth Order Frank Wolfe

Require: Input, Loss Function $F(x)$, L (Lipschitz constant for the gradients), Convex Set $\mathcal{C}$, Sequences $\gamma_t = \frac{2}{t+1}, c_t = \frac{L\gamma_t}{d}$.
Output: : $\mathbf{x}_T$ or $\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$.
1: Initialize $\mathbf{x}_0 \in \mathcal{C}$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Compute $\mathbf{g}(\mathbf{x}_t) = \sum_{i=1}^{d} \frac{F(\mathbf{x}_t + c_t\mathbf{e}_i) - F(\mathbf{x}_t)}{c_t}\mathbf{e}_i$,
4: Compute $\mathbf{v}_t = \operatorname{argmin}_{\mathbf{s}\in\mathcal{C}}\langle \mathbf{s}, \mathbf{g}(\mathbf{x}_t)\rangle$,
5: Compute $\mathbf{x}_{t+1} = (1-\gamma_t)\,\mathbf{x}_t + \gamma_t\mathbf{v}_t$.
6: **end for**

Assumption 10.4.4. The $\mathbf{z}_t$'s are drawn from a distribution μ such that $M(\mu) = \mathbb{E}\left[\|\mathbf{z}_t\|^6\right]$ is finite, and for any vector $\mathbf{g} \in \mathbb{R}^d$, there exists a function $s(d) : \mathbb{N} \mapsto \mathbb{R}_+$ such that,

$$\mathbb{E}\left[\|\langle \mathbf{g}, \mathbf{z}_t\rangle \mathbf{z}_t\|^2\right] \leq s(d)\,\|\mathbf{g}\|^2 .$$

Assumption 10.4.5. The unbiased gradient estimates, $\nabla F(\mathbf{x}; \mathbf{y})$ of $\nabla f(\mathbf{x})$, i.e., $\mathbb{E}_{\mathbf{y}\sim\mathcal{P}}\left[\nabla F(\mathbf{x}; \mathbf{y})\right] = \nabla f(\mathbf{x})$ satisfy

$$\mathbb{E}\left[\|\nabla F(\mathbf{x}, \mathbf{y}) - \nabla f(\mathbf{x})\|^2\right] \leq \sigma^2 \tag{10.7}$$

We note that Assumptions 10.4.1-10.4.3 and 10.4.5 are standard in the context of stochastic optimization. Assumption 2.4.2 provides for the requisite moment conditions for the sampling distribution of the directions utilized for finding directional derivatives so as to be able to derive concentration bounds. In particular, if μ is taken to be uniform on the surface of the $\mathbb{R}^d$ Euclidean ball with radius $\sqrt{d}$, then we have that $M(\mu) = d^3$ and $s(d) = d$. Moreover, if μ is taken to be $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, then $M(\mu) = d(d+2)(d+4) \approx d^3$ and $s(d) = d$. For the rest of the paper, we take μ to be either uniform on the surface of the $\mathbb{R}^d$ Euclidean ball with radius $\sqrt{d}$ or $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Before getting into the stochastic case, we demonstrate how a typical zeroth order Frank-Wolfe framework corresponds to an inexact classical Frank-Wolfe optimization in the deterministic setting.

10.4.1 Deterministic Zeroth Order Frank-Wolfe

The deterministic version of the optimization in (10.2) can be re-stated as follows:

$$\min_{\mathbf{x}\in\mathcal{C}} F(\mathbf{x}) . \tag{10.8}$$

In order to elucidate the equivalence of a typical zeroth order Frank-Wolfe framework corresponds to an inexact classical Frank-Wolfe optimization, we restrict our attention to the Kiefer-Wolfowitz stochastic approximation (KWSA) for gradient estimation. In particular, the KWSA gradient estimator in (10.4) can be expressed as follows:

$$\begin{aligned}\mathbf{g}(\mathbf{x}_t) &= \sum_{i=1}^{d} \frac{F(\mathbf{x}_t + c_t\mathbf{e}_i) - F(\mathbf{x}_t)}{c_t}\mathbf{e}_i \\ &= \nabla F(\mathbf{x}_t) + \sum_{i=1}^{d} \frac{c_t}{2}\langle \mathbf{e}_i, \nabla^2 F(\mathbf{x}_t + \lambda_t c_t \mathbf{e}_i)\mathbf{e}_i\rangle \mathbf{e}_i,\end{aligned} \tag{10.9}$$

where $\lambda \in [0,1]$. The linear optimization step with the current gradient approximation reduces to:

$$\begin{aligned}
&\langle \mathbf{v}, \mathbf{g}(\mathbf{x}_t)\rangle = \langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle \\
&+ \frac{c_t}{2}\sum_{i=1}^{d}\langle \mathbf{e}_i, \nabla^2 F(\mathbf{x}_t + \lambda_t c_t \mathbf{e}_i)\mathbf{e}_i\rangle\langle \mathbf{v}, \mathbf{e}_i\rangle \\
&\Rightarrow \min_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \mathbf{g}(\mathbf{x}_t)\rangle \leq \min_{\mathbf{s}\in\mathcal{C}}\langle \mathbf{s}, \nabla F(\mathbf{x}_t)\rangle + \frac{c_t LRd}{2}.
\end{aligned} \tag{10.10}$$

In particular, if c_t is chosen to be $c_t = \frac{\gamma_t}{d}$ and $\gamma_t = \frac{2}{t+1}$, we obtain the following bound characterizing the primal gap:

Theorem 10.4.1. *Given the zeroth order Frank-Wolfe algorithm in Algorithm 1, we obtain the following bound:*

$$F(\mathbf{x}_t) - F(\mathbf{x}^*) = \frac{Q_{ns}}{t+2}, \tag{10.11}$$

where $Q_{ns} = \max\{2(F(\mathbf{x}_0) - F(\mathbf{x}^*)), 4LR^2\}$.

The proof of the above theorem is relegated to Appendix I. Theorem 10.4.1 asserts that with appropriate scaling of c_t, i.e., the smoothing parameter for the zeroth order gradient estimator, the iteration dependence of the primal gap matches that of the classical Frank-Wolfe scheme. In particular, for a primal gap of ϵ, the number of iterations needed for the zeroth order scheme in algorithm 1 is $O\left(\frac{1}{\epsilon}\right)$, while the number of calls to the linear minimization oracle and zeroth order oracle are given by $O\left(\frac{1}{\epsilon}\right)$ and $O\left(\frac{d}{\epsilon}\right)$ respectively.
In summary, Theorem 10.4.1 shows that the deterministic zeroth order Frank-Wolfe algorithm reduces to the inexact classical Frank-Wolfe algorithm with the corresponding primal being dimension independent. However, the dimension independence comes at the cost of querying the zeroth order oracle d times at each iteration. In the sequel, we will focus on the random directions gradient estimator in (10.5) for the stochastic zeroth order Frank-Wolfe algorithm.

10.4.2 Zeroth Order Stochastic Frank-Wolfe

In this section, we formally introduce our proposed zeroth order stochastic Frank-Wolfe algorithm. A naive replacement of $\nabla f(\mathbf{x}_k)$ by its stochastic counterpart, i.e., $\nabla F(\mathbf{x}_k; \mathbf{y}_k)$ would make the algorithm potentially divergent due to non-vanishing variance of gradient approximations. Moreover, the naive replacement would lead to the linear minimization constraint to hold only in expectation and thereby potentially also making the algorithm divergent. We use a well known averaging trick to counter this problem which is as follows:

$$\mathbf{d}_t = (1-\rho_t)\,\mathbf{d}_{t-1} + \rho_t g\left(\mathbf{x}_t, \mathbf{y}_t\right), \tag{10.12}$$

where $g\left(\mathbf{x}_t, \mathbf{y}_t\right)$ is a gradient approximation, $\mathbf{d}_0 = \mathbf{0}$ and ρ_t is a time-decaying sequence. Technically speaking, such a scheme allows for $\mathbb{E}\left[\left\|\mathbf{d}_t - \nabla f\left(\mathbf{x}_t\right)\right\|^2\right]$ to go to zero asymptotically. With the above averaging scheme, we replace the linear minimization and the subsequent steps as follows:

$$\begin{aligned}
\mathbf{d}_t &= (1-\rho_t)\,\mathbf{d}_{t-1} + \rho_t g\left(\mathbf{x}_t, \mathbf{y}_t\right) \\
\mathbf{v}_t &= \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{d}_t, \mathbf{v}\rangle \\
\mathbf{x}_{t+1} &= (1-\gamma_{t+1})\,\mathbf{x}_t + \gamma_{t+1}\mathbf{v}_t.
\end{aligned} \tag{10.13}$$

We resort to three different gradient approximation schemes for approximating $g(\mathbf{x}_t, \mathbf{y}_t)$. In particular, in addition to the the KWSA scheme and the random directions scheme, as outlined in (10.4) and (10.5), we employ an improvised random directions gradient estimator (I-RDSA) by sampling m directions at each time followed by averaging, i.e., $\{\mathbf{z}_{i,t}\}_{i=1}^m$ for which we have,

$$\begin{aligned}&\mathbf{g}_m(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_{i,t})\\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{F(\mathbf{x}_t + c_t\mathbf{z}_{i,t}; \mathbf{y}) - F(\mathbf{x}_t; \mathbf{y})}{c_t}\mathbf{z}_{i,t}\right).\end{aligned} \tag{10.14}$$

It is to be noted that the above gradient approximation scheme uses more exactly one data point while utilizing m directional derivatives. In order to quantify the benefits of using such a scheme, we present the statistics concerning the gradient approximation of RDSA and I-RDSA. We have from Duchi et al. (2015) for RDSA,

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}_t\sim\mu, \mathbf{y}_t\sim\mathcal{P}}\left[\mathbf{g}(\mathbf{x}; \mathbf{y}_t, \mathbf{z}_t)\right] = \nabla f(\mathbf{x}) + c_t L\mathbf{v}(\mathbf{x}, c_t)\\ &\mathbb{E}_{\mathbf{z}_t\sim\mu, \mathbf{y}_t\sim\mathcal{P}}\left[\|\mathbf{g}(\mathbf{x}; \mathbf{y}_t, \mathbf{z}_t)\|^2\right] \leq 2s(d)\mathbb{E}\left[\|\nabla F(\mathbf{x}; \mathbf{y}_t)\|^2\right]\\ &+ \frac{c_t^2}{2}L^2M(\mu),\end{aligned} \tag{10.15}$$

Using (10.15), similar statistics for the improvised RDSA gradient estimator can be evaluated as follows:

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}_t\sim\mu, \mathbf{y}_t\sim\mathcal{P}}\left[\mathbf{g}_m(\mathbf{x}; \mathbf{y}_t, \mathbf{z}_t)\right] = \nabla f(\mathbf{x}) + \frac{c_t}{m}L\mathbf{v}(\mathbf{x}, c_t)\\ &\mathbb{E}_{\mathbf{z}_t\sim\mu, \mathbf{y}_t\sim\mathcal{P}}\left[\|\mathbf{g}_m(\mathbf{x}; \mathbf{y}_t, \mathbf{z}_t)\|^2\right] \leq \left(\frac{1+m}{2m}\right)c_t^2L^2M(\mu)\\ &+ 2\left(1 + \frac{s(d)}{m}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}; \mathbf{y}_t)\|^2\right],\end{aligned} \tag{10.16}$$

where $\|\mathbf{v}(\mathbf{x}, c_t)\| \leq \frac{1}{2}\mathbb{E}\left[\|\mathbf{z}\|^3\right]$. As we will see later the I-RDSA scheme improves the dimension dependence of the primal gap, but it comes at the cost of m calls to the SZO. We are now ready to state the zeroth order stochastic Frank-Wolfe algorithm which is presented in algorithm 2. Before the main results, we first study the evolution of the gradient estimates in (10.12) and the associated mean square error. The following Lemma studies the error of the process $\{\mathbf{d}_t\}$ as defined in (10.12).

Lemma 10.4.2. *Let Assumptions 10.4.1-10.4.5 hold. Given the recursion in (10.12), we have that* $\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2$ *satisfies*

1) for the RDSA gradient approximation scheme

$$\begin{aligned}&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq 2\rho_t^2\sigma^2 + 4\rho_t^2L_1^2 + 8\rho_t^2s(d)L_1^2\\ &+ 2\rho_t^2c_t^2L^2M(\mu) + \frac{2L^2R^2\gamma_t^2}{\rho_t} + \frac{\rho_t}{2}c_t^2L^2M(\mu)\\ &+ \left(1 - \frac{\rho_t}{2}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right],\end{aligned} \tag{10.17}$$

2) for the I-RDSA gradient approximation scheme

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq 2\rho_t^2\left(\sigma^2 + 2L_1^2\right)$$

Algorithm 2 Stochastic Gradient Free Frank Wolfe

Require: Input, Loss Function $F(x)$, Convex Set $\mathcal{C}$, number of directions m, sequences $\gamma_t = \frac{2}{t+8}$,

$(\rho_t, c_t)_{RDSA} = \left(\frac{4}{d^{1/3}(t+8)^{2/3}}, \frac{2}{d^{3/2}(t+8)^{1/3}}\right)$

$(\rho_t, c_t)_{I-RDSA} = \left(\frac{4}{(1+\frac{d}{m})^{1/3}(t+8)^{2/3}}, \frac{2\sqrt{m}}{d^{3/2}(t+8)^{1/3}}\right)$

$(\rho_t, c_t)_{KWSA} = \left(\frac{4}{(t+8)^{2/3}}, \frac{2}{d^{1/2}(t+8)^{1/3}}\right)$.

Output: $\mathbf{x}_T$ or $\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}_t$.

1: Initialize $\mathbf{x}_0 \in \mathcal{C}$
2: **for** $t = 0, 2, \ldots, T-1$ **do**
3: Compute
KWSA:
$\mathbf{g}(\mathbf{x}_t; \mathbf{y}) = \sum_{i=1}^{d} \frac{F(\mathbf{x}_t + c_t\mathbf{e}_i; \mathbf{y}) - F(\mathbf{x}_t; \mathbf{y})}{c_t}\mathbf{e}_i$
RDSA: Sample $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I}_d)$,
$\mathbf{g}(\mathbf{x}_t; \mathbf{y}, \mathbf{z}_t) = \frac{F(\mathbf{x}_t + c_t\mathbf{z}_t; \mathbf{y}) - F(\mathbf{x}_t; \mathbf{y})}{c_t}\mathbf{z}_t$
I-RDSA: Sample $\{\mathbf{z}_{i,t}\}_{i=1}^{m} \sim \mathcal{N}(0, \mathbf{I}_d)$,
$\mathbf{g}(\mathbf{x}_t; \mathbf{y}, \mathbf{z}_t) = \frac{1}{m}\sum_{i=1}^{m} \frac{F(\mathbf{x}_t + c_t\mathbf{z}_{i,t}; \mathbf{y}) - F(\mathbf{x}_t; \mathbf{y})}{c_t}\mathbf{z}_{i,t}$
4: Compute $\mathbf{d}_t = (1-\rho_t)\,\mathbf{d}_{t-1} + \rho_t \mathbf{g}\,(\mathbf{x}_t, \mathbf{y}_t)$
5: Compute $\mathbf{v}_t = \text{argmin}_{\mathbf{s}\in\mathcal{C}}\langle \mathbf{s}, \mathbf{d}_t)\rangle$,
6: Compute $\mathbf{x}_{t+1} = (1-\gamma_t)\,\mathbf{x}_t + \gamma_t \mathbf{v}_t$.
7: **end for**

$$
\begin{aligned}
&+ \frac{\rho_t}{2m^2} c_t^2 L^2 M(\mu) + 8\rho_t^2 \left(1 + \frac{s(d)}{m}\right) L_1^2 \\
&+ \left(\frac{1+m}{2m}\right) \rho_t^2 c_t^2 L^2 M(\mu) + \frac{2L^2 R^2 \gamma_t^2}{\rho_t} \\
&+ \left(1 - \frac{\rho_t}{2}\right) \mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \qquad (10.18)
\end{aligned}
$$

3) for the KWSA gradient approximation scheme

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq 2\rho_t^2\sigma^2 + 2\rho_t c_t^2 d L^2 \\
&+ \frac{2L^2R^2\gamma_t^2}{\rho_t} + \left(1 - \frac{\rho_t}{2}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right]. \qquad (10.19)
\end{aligned}
$$

We use the following Lemma so as to study the dynamics of the primal gap.

Lemma 10.4.3. *Consider the zeroth order Frank Wolfe Algorithm in 1. Let Assumptions 10.4.1-10.4.5 hold. Then, the primal gap $F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)$ satisfies*

$$
\begin{aligned}
&F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq (1-\gamma_{t+1})(F(\mathbf{x}_t) - F(\mathbf{x}^*)) \\
&+ \gamma_{t+1} R \|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\| + \frac{LR^2\gamma_{t+1}^2}{2}. \qquad (10.20)
\end{aligned}
$$

With the above recursions in place, we can now characterize the finite time rates of the mean square errors for the different error approximation schemes. In particular, using Lemma 10.4.2, we first state the main result concerning the setting, where the objective is convex.

10.5 Main Results

In this section, we state the main results, while the proofs are relegated to Appendix I. We first state the main results concerning the primal gap of the proposed algorithm.

Primal Gap: We state the main results involving the different gradient approximation schemes for the primal gap, which provide a characterization of $\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right]$.

Theorem 10.5.1. *Let Assumptions 10.4.1-10.4.5 hold. Let the sequence γ_t be given by $\gamma_t = \frac{2}{t+8}$.*

1) Then, we have the following primal sub-optimality gap for the algorithm in 2, with the RDSA gradient approximation scheme:

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] = O\left(\frac{d^{1/3}}{(t+9)^{1/3}}\right). \tag{10.21}$$

2) In case of the I-RDSA the gradient approximation scheme, the primal sub-optimality gap is given by,

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] = O\left(\frac{(d/m)^{1/3}}{(t+9)^{1/3}}\right). \tag{10.22}$$

3) Finally, for the KWSA gradient approximation scheme, the primal sub-optimality gap is given by,

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] = O\left(\frac{1}{(t+9)^{1/3}}\right). \tag{10.23}$$

Theorem 10.5.1 quantifies the dimension dependence of the primal gap to be $d^{1/3}$. At the same time the dependence on iterations, i.e., $O(T^{-1/3})$ matches that of the stochastic Frank-Wolfe which has access to first order information as in Mokhtari et al. (2018). The improvement of the rates for I-RDSA and KWSA are at the cost of extra directional derivatives at each iteration. The number of queries to the SZO so as to obtain a primal gap of ϵ, i.e., $\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \epsilon$ is given by $O\left(\frac{d}{\epsilon^3}\right)$, where the dimension dependence is consistent with zeroth order schemes and cannot be improved on as illustrated in Duchi et al. (2015)

Dual Gap: We state the main results involving the different gradient approximation schemes for the dual gap, which provide a characterization of $\mathcal{G}(\mathbf{x}) = \max_{\mathbf{v}\in\mathcal{C}}\langle\nabla F(\mathbf{x}), \mathbf{x} - \mathbf{v}\rangle$.

Theorem 10.5.2. *Let Assumptions 10.4.1-10.4.5 hold. Let the sequence γ_t be given by $\gamma_t = \frac{2}{t+8}$.*

1) Then, we have the following dual gap for the algorithm in 2, with the RDSA gradient approximation scheme:

$$\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}(t)\right)\right] \leq \frac{7(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{2T} + \frac{LR^2\ln(T+7)}{T} + \frac{Q^{'} + R\sqrt{2Q}}{2T}(T+7)^{2/3}, \tag{10.24}$$

where $Q = 32d^{-1/3}\sigma^2 + 64d^{-1/3}L_1^2 + 128d^{2/3}L_1^2 + 2L^2R^2d^{2/3} + 416d^{2/3}L^2$ and $Q^{'} = \max\{2(f(\mathbf{x}_0) - f(\mathbf{x}^*)), 2R\sqrt{2Q} + LR^2/2\}$.

2) In case of the I-RDSA the gradient approximation scheme, the dual gap is given by,

$$\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}(t)\right)\right] \leq \frac{7(F(\mathbf{x}_0)-F(\mathbf{x}^*))}{2T} + \frac{LR^2\ln(T+7)}{T} + \frac{Q_{ir}^{'}+R\sqrt{2Q_{ir}}}{2T}(T+7)^{2/3}, \tag{10.25}$$

where $Q_{ir} = 32\left(1+d/m\right)^{-1/3}\sigma^2 + 128\left(1+d/m\right)^{2/3}L_1^2 + 64\left(1+d/m\right)^{-1/3}L_1^2 + 2L^2R^2\left(1+d/m\right)^{2/3} + 416\left(1+d/m\right)^{2/3}L^2$ *and* $Q_{ir}^{'} = \max\{2(f(\mathbf{x}_0)-f(\mathbf{x}^*)), 2R\sqrt{2Q_{ir}}+LR^2/2$.

3) Finally, for the KWSA gradient approximation scheme, the dual gap is given by,

$$\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}(t)\right)\right] \leq \frac{7(F(\mathbf{x}_0)-F(\mathbf{x}^*))}{2T} + \frac{LR^2\ln(T+7)}{T} + \frac{Q_{kw}^{'}+R\sqrt{2Q_{kw}}}{2T}(T+7)^{2/3}, \tag{10.26}$$

where $Q_{kw} = \max\left\{4\|\nabla f(\mathbf{x}_0)-\mathbf{d}_0\|^2,\ 32\sigma^2+32L^2+2L^2R^2\right\}$ *and* $Q_{kw}^{'} = \max\{2(f(\mathbf{x}_0)-f(\mathbf{x}^*)), 2R\sqrt{Q_{kw}}+LR^2/2\}$.

Theorem 10.5.2 quantifies the dimension dependence of the Frank-Wolfe duality gap to be $d^{1/3}$. At the same time the dependence on iterations, i.e., $O(T^{-1/3})$ matches that of the primal gap and hence follows that the number of queries to the SZO so as to obtain a Frank-Wolfe duality gap of ϵ, i.e., $\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}(t)\right)\right] \leq \epsilon$ is given by $O\left(\frac{d}{\epsilon^3}\right)$. In particular, theorem 10.5.2 asserts that the initial conditions are forgotten as $O(1/T)$.

10.5.1 Zeroth-Order Frank-Wolfe Non-Convex

We employ the following algorithm for the non-convex stochastic Frank-Wolfe:

Algorithm 3 Stochastic Gradient Free Frank-Wolfe

Require: Input, Loss Function $F(x)$, Convex Set $\mathcal{C}$, number of directions m. Sequences $\gamma = \frac{1}{T^{3/4}}$,

$$(\rho_t, c_t) = \left(\frac{4}{\left(1+\frac{d}{m}\right)^{1/3}(t+8)^{2/3}}, \frac{2\sqrt{m}}{(d^{3/2}(t+8)^{1/3})}\right)$$

Output: $\mathbf{x}_T$.

1: Initialize $\mathbf{x}_0 \in \mathcal{C}$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Compute
Sample $\{\mathbf{z}_{i,t}\}_{i=1}^m \sim \mathcal{N}(0, \mathbf{I}_d)$, $\mathbf{g}(\mathbf{x}_t;\mathbf{y},\mathbf{z}_t) = \frac{1}{m}\sum_{i=1}^m \frac{F(\mathbf{x}_t+c_t\mathbf{z}_{i,t};\mathbf{y})-F(\mathbf{x}_t;\mathbf{y})}{c_t}\mathbf{z}_{i,t}$
4: Compute $\mathbf{d}_t = (1-\rho_t)\,\mathbf{d}_{t-1} + \rho_t g\left(\mathbf{x}_t, \mathbf{y}\right)$
5: Compute $\mathbf{v}_t = \text{argmin}_{\mathbf{s}\in\mathcal{C}}\langle \mathbf{s}, \mathbf{d}_t)\rangle$,
6: Compute $\mathbf{x}_{t+1} = (1-\gamma)\,\mathbf{x}_t + \gamma\mathbf{v}_t$.
7: **end for**

We use the following assumption concerning the smoothness of the non-convex loss function.

Assumption 10.5.1. The gradients ∇f are L-Lipschitz continuous over the set $\mathcal{C}$, i.e., for all $x, y \in \mathcal{C}$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|.$$

Theorem 10.5.3. *Let Assumptions 10.4.3-10.5.1 hold. Then, we have the following dual gap for iterations* $t = 0, 1, \cdots, T-1$ *for the algorithm as described in* (10.13)

$$\mathbb{E}\left[\min_{t=0,\cdots,T-1} \mathcal{G}\left(\mathbf{x}(t)\right)\right] \leq \frac{Q'}{T^{1/4}} = O\left(\frac{(d/m)^{1/3}}{T^{1/4}}\right), \tag{10.27}$$

where $Q' = \max\{9^{1/3}(f(\mathbf{x}_0) - f(\mathbf{x}^*)), Q_{nc}R(d/m)^{1/3}\}$.

Theorem 10.5.3 quantifies the dimension dependence of the Frank-Wolfe duality gap for non-convex functions to be $d^{1/3}$. At the same time the dependence on iterations, i.e., $O(T^{-1/4})$ matches that of the rate of SFW in Reddi et al. (2016) and hence follows that the number of queries to the SZO so as to obtain a Frank-Wolfe duality gap of ϵ, i.e., $\mathbb{E}\left[\min_{t=0,\cdots,T-1} \mathcal{G}\left(\mathbf{x}(t)\right)\right] \leq \epsilon$ is given by $O\left(\frac{d^{4/3}}{\epsilon^4}\right)$.

10.6 Experiments

We now present empirical results for zeroth order Frank-Wolfe optimization with an aim to highlight three aspects of our method: (i) it is accurate even in stochastic case (Section 10.6.1) (ii) it scales to relatively high dimensions (Section 10.6.2) (iii) it reaches stationary point in non-convex setting (Section 10.6.3).

Methods and Evaluation We look at the optimality gap $|f(x_{\text{optimizer}}) - f(x^*)|$ as the evaluation metric, where $x_{\text{optimizer}}$ denotes the solution obtained from the employed optimizer and x^* corresponds to true solution. Most existing zero order optimization techniques like Nelder-Mead simplex (Nelder and Mead, 1965) or bound optimization by quadratic approximation (BOBYQA; Powell 2009) can only handle bound constraints, but not arbitrary convex constraints as our method can. Thus, for all experiments, we could compare proposed zeroth order stochastic Frank-Wolf (0-FW) only with COBYLA, a constrained optimizer by linear approximation, which is popular in engineering fields (Powell, 1994). For experiments where SFO is available, we additionally compare with stochastic proximal gradient descent (PGD) and first order stochastic Frank-Wolfe method (1-FW).

10.6.1 Stochastic Lasso Regression

To study performance of various stochastic optimization, we solve a simple lasso regression on the dataset covtype ($n = 581012$, $d = 54$) from libsvm website[2]. We use the variant with feature values in $[0, 1]$ and solve the following problem:

$$\min_{\|\mathbf{w}\|_1 \leq 1} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|_2^2$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ represents the feature vectors and $\mathbf{y} \in \mathbb{R}^n$ are the corresponding targets.
For the 0-FW, we used I-RDSA with $m = 6$. This problem represents a stochastic setting and from Figure 10.1a we note that the performance of 0-FW matches that of 1-FW in terms of the number of oracle calls to their respective oracles in spite of the dimension involved being $d = 54$.

10.6.2 High Dimensional Cox Regression

To demonstrate efficacy of zeroth order Frank-Wolfe optimization in a moderately high dimensional case, we look at gene expression data. In particular, we perform patient survival analysis by solving Cox regression

[2]Available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

(also known as proportional hazards regression) to relate different gene expression profiles with survival time (Sohn et al., 2009). We use the Kidney renal clear cell carcinoma dataset[3], which contains gene expression data for 606 patients (534 with tumor and 72 without tumor) along with survival time information. We preprocess the dataset by eliminating the rarely expressed genes, i.e. we only keep genes expressed in 50% of the patients. This leads to a feature vector x_i of size 9376 for each patient i. Also, for each patient i, we have the censoring indicator variable y_i that takes the value 0 if patient is alive or 1 if death is observed with t_i denoting the time of death. In this setup, we can obtain a sparse solution to cox regression by solving the following problem (Park and Hastie, 2007; Sohn et al., 2009):

$$\min_{\|\mathbf{w}\|_1 \leq 10} \frac{1}{n} \sum_{i=1}^{n} y_i \left\{ -\mathbf{x}_i^\top \mathbf{x} + \log \left(\sum_{j \in \mathcal{R}_i} \exp(\mathbf{x}_j^\top \mathbf{w}) \right) \right\}$$

where $\mathcal{R}_i$ is the set of subjects at risk at time t_i, i.e. $\mathcal{R}_i = \{j : t_j \geq t_i\}$.
This problem represents a high-dimensional setting with $d = 9000$. For this setup, we take $m = 900$ for the I-RDSA scheme of our proposed algorithm. Due to the unavoidable dimension dependence of zeroth order schemes, Figure 10.1b shows the gap between 1-FW and 0-FW to be around $2\times$ and thereby reinforcing the result in Theorem 10.5.1 (2)

10.6.3 Black-Box Optimization

Finally, we show efficacy of zeroth order Frank-Wolfe optimization in a non-convex setting for a black-box optimization. Many engineering problems can be posed as optimizing forward models from physics, which are often complicated, do not posses analytical expression, and cannot be differentiated. We take the example of analyzing electron back-scatter diffraction (EBSD) patterns in order to determine crystal orientation of the sample material. Such analysis is useful in determining strength, malleability, ductility, etc. of the material along various directions. Brute-force search has been the primary optimization technique in use (Ram et al., 2017). For this problem, we use the forward model of EBSD provided by EMSoft[4]. There are $d = 6$ parameters to optimize over the L_∞-ball of radius 1.
This problem represents a non-convex black box optimization setting for which we used $m = 1$ for the I-RDSA, i.e. RDSA. Figure 10.1c shows that our proposed algorithm converges to a first order stationary point there by showing the effectiveness of our proposed algorithm for black-box optimization.

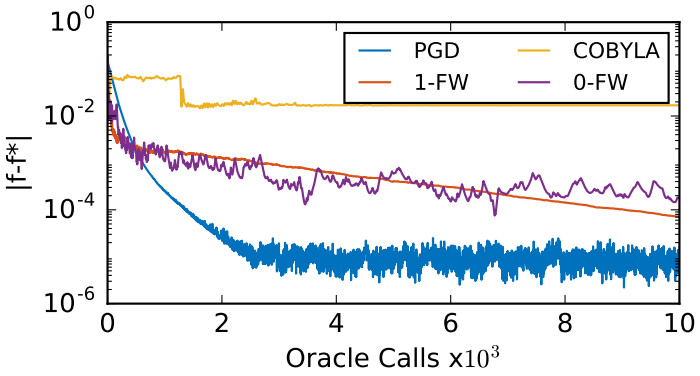
10.7 Contributions

- **Dimension dependence of $d^{1/3}$**: We proposed a zeroth order Frank-Wolfe method for which we showed that in the deterministic case it reduces to the inexact version of Frank-Wolfe method. In particular, when per epoch only one directional derivative is sampled, we showed that the primal gap and Frank-Wolfe gap have a dependence of $d^{1/3}$ which is the best known dimension dependence among all zeroth order schemes.

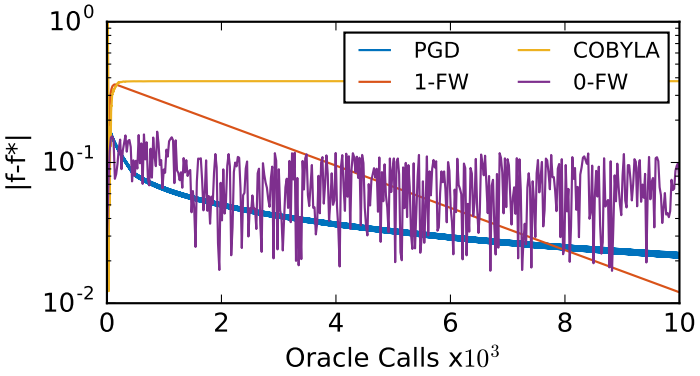
[3] Available at http://gdac.broadinstitute.org
[4] Software is available at https://github.com/EMsoft-org/EMsoft

10.8 Conclusion and Future Directions

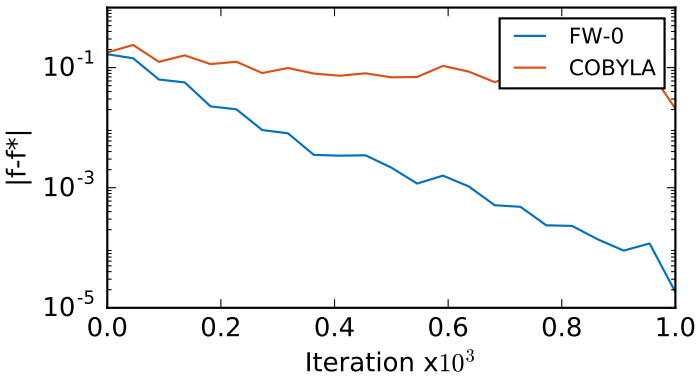
In this chapter, we proposed a stochastic zeroth order Frank-Wolfe algorithm. The proposed algorithm does not depend on hard to estimate quantities like Lipschitz constants and thus is easy to deploy in practice. For the proposed algorithm, we quantified the rates of convergence of the proposed algorithm in terms of the primal gap and the Frank-Wolfe duality gap, which we showed to match its first order counterpart in terms of iterations. In particular, we showed that the dimension dependence, when one directional derivative is sampled at each iteration to be $\mathrm{O}(d^{1/3})$. We demonstrated the efficacy of our proposed algorithm through experiments on multiple datasets. Natural future directions include extending the proposed algorithm to non-smooth functions and incorporating variance reduction techniques to get better rates.



(a) Covtype Dataset



(b) Cox Regression Dataset



(c) Black Box Optimization Dataset

Figure 10.1: Comparison of proposed zeroth order Frank-Wolfe (0-FW) with first order Frank-Wolfe (1-FW), proximal gradient descent (PGD), and another zero order constrained optimaztion by linear approximation (COBYLA) on various problems.

Chapter 11

Conclusions

The thesis develops methodology and algorithms to study distributed inference and optimization problems in resource constrained networked setups. Typical examples that fall under the scope of this study include cyberphysical systems and IoTs, which are typically deployed outside of a data center and hence have no central coordinator. Also, the aforementioned systems are deployed in random environments and are constrained in terms of resources, like communication bandwidth, computational power and sensing power in lieu with finite battery power. The networked entities need to collaborate with each other through local information exchange in terms of functions of data and not the data itself so as to be able to address the task at hand. The main thrust of this thesis centers around the fact that the energy needed to communicate is a few orders higher than the resources needed to sense data or perform local computations. In light of the aforementioned fact, one aspect of our work focuses on development of communication efficient distributed inference schemes. We explore different aspects of communication efficiency in terms of communicating in an increasingly sparse manner and reducing dimensions of messages. We study the performance of these distributed schemes keeping in mind the trade-off between optimality in terms of convergence rate and the communication cost. A major thrust of this thesis is to develop optimization schemes for loss functions which are analytically intractable and involve expensive gradient computations. We develop and explore communication efficient schemes for gradient free optimization and establish the communication optimality trade-off for them. The technical tools used in the study include mixed time scale stochastic approximation, large deviation theory and optimization. Our methods are generic and of independent interest to the general theory of these classical disciplines. We recapitulate the main contributions of this thesis.

Chapter 2 Distributed Sequential Detection: This chapter studies the problem of distributed sequential detection in the context of multi-agent networks. The proposed sequential detection procedure $\mathcal{CI}SPRT$ was inspired from Wald's SPRT. We derive thresholds for the proposed distributed sequential detection procedure which guarantee the algorithm to terminate while adhering to the pre-specified error tolerances. In particular, we established the dependence of the thresholds in terms of the network connectivity. In addition to the thresholds of the procedures, we characterized the tails of the stopping time distribution of $\mathcal{CI}SPRT$ and showed its proximity with respect to its centralized counterpart. The thresholds and stopping time distribution of the algorithm $\mathcal{CI}SPRT$ facilitate the comparison of the expected stopping time of the proposed algorithm to that of its counterpart and quantifying the dependence in terms of the network connectivity. In this chapter, we also provide a computationally tractable version of the stopping time distribution of the classical SPRT procedure.

Chapter 3 Distributed Composite Hypothesis Testing: This chapter studies the problem of recursive distributed composite hypothesis testing, where one of the hypotheses admits infinite parameterization. In such setups, it is customary to have an inherent maximization scheme to estimate the underlying parameter parameterizing the alternate hypothesis. We proposed algorithms $\mathcal{CIGLRT}-\mathcal{L}$ and $\mathcal{CILGRT}-\mathcal{NL}$ catering to linear and non-linear observation models respectively, where the aforementioned maximization and the decision statistic update are performed parallely and in an online manner ,i.e., as and when a new sample is collected. We characterize the convergence of the maximization to the true underlying parameter and algorithm parameters which ensure the decay of the probabilities of error concerning the hypothesis testing procedure at hand. In particular, for $\mathcal{CIGLRT}-\mathcal{L}$, we establish the exponential decay of the probabilities of error by studying concentration bounds for quadratic forms of Wishart matrices. Furthermore, we also considered a noisy communication model imposed to the setup in $\mathcal{CIGLRT}-\mathcal{L}$, for which we derived algorithm parameters as a function of every agent's local connectivity so as to ensure asymptotically decaying probabilities of errors.

Chapter 4 Communication Efficient Distributed Detection: This chapter studies convergence in probability of products of random, independent, but not identically distributed stochastic and symmetric matrices $\mathbf{W}_t$, where the topologies that underline the matrices have time-varying distributions. In particular, the chapter is motivated by the convergence properties of product of stochastic matrices that it is typically encountered while analyzing the probabilities of errors in distributed simple hypothesis testing. Technically speaking, we characterize the following quantity:

$$\mathcal{R} = -\lim_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(\|\mathbf{W}_t \cdot \ldots \cdot \mathbf{W}_1 - \mathbf{J}\| \geq \epsilon\right), \tag{11.1}$$

We show that the limit in (11.1) exists, and moreover we compute exactly the limit $\mathcal{R}$. Specifically, we show that $\mathcal{R}$ is given by the minimal vertex cut of the baseline graph, where the nodes' associated cut costs are defined by the nodes' limiting activation probabilities.We demonstrate the significance of the studied non-i.i.d. matrix model and the derived rate $\mathcal{R}$ in the context of *consensus+innovations* distributed detection. More precisely, we consider a distributed detector with a randomized and time-varying sparsified communication protocol, where neighborhood communications are probabilistically sparsified in a time varying fashion with the goal of reducing the detector's communication cost.By utilizing result (4.1), we first show theoretically that the detector with time-varying and sparsified protocol can be designed to achieve asymptotic optimality at all signal-to-noise ratio (SNR) regimes; this is achieved when the activation probabilities corresponding to each node converge to unity, possibly at a very slow rate, e.g., as $1-\Omega(1/\log(t))$.

Chapter 5 Communication Efficient Linear Parameter Estimation: $\mathcal{CREDO}$ This chapter studies the problem of communication efficient distributed linear parameter estimation, where we improve the communication cost of the distributed scheme without compromising on the optimality in terms of the convergence rate. We propose a scheme $\mathcal{CREDO}$, where each node at time t communicates only with a certain probability that decays sub linearly to zero in t. That is, communications are increasingly sparse, so that communication cost scales as $O(t^{\delta})$, where the growing rate δ is a tunable parameter strictly less than one that can go down to 0.5. We show that, despite significantly lower communication cost, the proposed method achieves the best possible $O(1/t)$ rate of MSE decay in time t (t also equals to per-worker number of data samples). Importantly, this result translates into significant improvements in the rate at which MSE decays with communication cost $\mathcal{C}_t$ –namely from $O(1/\mathcal{C}_t)$ with existing methods to $O(1/\mathcal{C}_t^{2-\zeta})$ with the proposed

method, where $\zeta > 0$ arbitrarily small. $\mathcal{CREDO}$ is build around the restriction that at each sampling epoch, the communication graph is restricted to being undirected. With further relaxation which allows for the communication graph to be directed while still being undirected in expectation, we show that the communication cost can be further reduced to $O(t^{\delta})$ where δ can be arbitrarily small. Hence, for the directed $\mathcal{CREDO}$, we show that the MSE decays with communication cost as $O(\mathcal{C}_t^{-1/\zeta})$, where $\zeta > 0$ is arbitrarily small. From a technical standpoint, the number of time scales involved in the proposed algorithm is 3 which further generalizes the *consensus+innovations* framework.

Chapter 6 Distributed Weighted Non-linear Least Squares: $\mathcal{CIWNLS}$ The chapter focuses on distributed nonlinear least squares estimation in distributed information settings. This chapter proposes a distributed recursive algorithm, namely, the $\mathcal{CIWNLS}$ (*Consensus+ innovations* Weighted Nonlinear Least Squares), which is of the *consensus + innovations* form Kar et al. (2012). We specifically focus on a setting in which the agents make i.i.d observations sequentially over time, only possess local model information, and update their parameter estimates by simultaneous assimilation of the information obtained from their neighboring agents (*consensus*) and current locally sensed information (*innovation*). We show that our recursive distributed estimator generates parameter estimate sequences that are strongly consistent at each agent. Furthermore, we also show that the proposed distributed estimation algorithm $\mathcal{CIWNLS}$ yields order-optimal pathwise convergence rate under certain smoothness conditions on the sensing model and characterize the asymptotic covariance of the estimator. Technically speaking, we quantify the inherent trade-off between the optimality of the estimation scheme in terms of the asymptotic covariance and the sharing of model information among the agents. Following as in lines of Chapter 5, we also develop a communication efficient version of $\mathcal{CIWNLS}$, namely $\mathcal{CREDO}-\mathcal{NL}$, for which we focus on a single time scale *consensus+innovations* algorithm. For $\mathcal{CREDO}-\mathcal{NL}$, we show that the MSE decays as $O(1/\mathcal{C}_t^{2-\zeta})$ in terms of communication cost $\mathcal{C}_t$, where $\zeta > 0$ is arbitrarily small.

Chapter 7 Communication Efficient Distributed Estimation: Random Fields Estimation The chapter focuses on a heterogeneous setup which involves distributed linear parameter estimation, where estimating the entire parameter at each agent is prohibitive due to excessive communication and memory overhead. Instead of estimating the entire parameter, agents choose to estimate only a few entries of the entire parameter, referred to as their interest set. We propose a scheme, namely $\mathcal{CIRFE}$, where each entity reconstructs only a subset of the components of the state modeled by a vector parameter, and thereby also reducing the dimension of messages being communicated among the agents. The proposed scheme allows heterogeneity in terms of agents' objectives, while still allowing for inter-agent collaboration. Through $\mathcal{CIRFE}$, we address communication efficiency for the class of distributed inference algorithm of the *consensus+innovations* form by reducing the dimension of vectors exchanged among the agents. In particular, we extended the idea of consensus to a heterogeneous version which exhibits consensus to subspaces which is common to a few agents. Under mild conditions of the connectivity of the network, we establish consistency of the estimate sequence at each agent with respect to the components of the parameters in its interest set.

Chapter 8 Communication Efficient Distributed Optimization: First Order The chapter focuses on first order distributed optimization schemes over random networks. We showed that, by carefully designing the consensus and the gradient weights (potentials), the considered distributed stochastic gradient algorithm achieves the order-optimal $O(1/k)$ rate of decay of the mean squared distance from the solution

(mean squared error – MSE). This is achieved for twice continuously differentiable strongly convex local costs, assuming also that the noisy gradients are unbiased estimates of the true gradients and that the noise in gradients has bounded second moment. We developed novel methods for first order distributed stochastic optimization, based on a probabilistic inter-agent communication protocol that increasingly sparsifies agent communications over time. For the first order distributed stochastic optimization, we propose a novel method that is shown to achieve the $O(1/(C_{\text{comm}})^{4/3-\zeta})$ MSE communication rate. At the same time, the proposed method retains the order-optimal $O(1/(C_{\text{comp}}))$ MSE rate in terms of the computational cost, the best achievable rate in the corresponding centralized setting.

Chapter 9 Communication Efficient Distributed Optimization: Zeroth Order The chapter focuses on optimization setups involving loss functions which do not admit to analytical forms and hence the associated gradient computations are analytically intractable. We first analyze a distributed zeroth order optimization scheme for strongly convex functions utilizing Kiefer Wolfowitz stochastic approximation. Furthermore, we develop novel methods for zeroth order distributed stochastic optimization, based on a probabilistic inter-agent communication protocol that increasingly sparsifies agent communications over time. We proposed a communication efficient distributed zeroth order scheme akin to the RDSA scheme albeit using three function evaluation at each epoch which in spirit draws from the twicing trick in kernel density estimation. In addition to the twicing trick, we have established for the proposed zeroth order method explicit mean square error (MSE) convergence rates with respect to (appropriately defined) computational cost C_{comp} and communication cost C_{comm}. Specifically, the proposed zeroth order method achieves the $O(1/(C_{\text{comm}})^{8/9-\zeta})$ MSE communication rate, which significantly improves over the rates of existing methods, while maintaining the order-optimal $O(1/(C_{\text{comp}})^{2/3})$ MSE computational rate.

Chapter 10 Zeroth Order Frank Wolfe The chapter focuses on stochastic constrained optimization involving stochastic zeroth order oracles. We develop a zeroth order projection free algorithm in lines of the celebrated Frank Wolfe algorithm so as to address the problem at hand. In particular, we establish the equivalence of the zeroth order Frank Wolfe scheme with that of the classical Frank Wolfe method in a deterministic optimization setting. n this chapter, we proposed a stochastic zeroth order Frank-Wolfe algorithm. The proposed algorithm does not depend on hard to estimate quantities like Lipschitz constants and thus is easy to deploy in practice. For the proposed algorithm, we quantified the rates of convergence of the proposed algorithm in terms of the primal gap and the Frank-Wolfe duality gap, which we showed to match its first order counterpart in terms of iterations. In particular, we showed that the dimension dependence, when one directional derivative is sampled at each iteration to be $\mathrm{O}(d^{1/3})$. We demonstrated the efficacy of our proposed algorithm through experiments on multiple datasets.

Bibliography

Delve datasets. http://www.cs.toronto.edu/~delve/data/datasets.html. 82

Libsvm regression datasets. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html. 82, 164

S. A. Alghunaim and A. H. Sayed. Distributed coupled multi-agent stochastic optimization. *arXiv preprint arXiv:1712.08817*, 2017. 108

T. W. Anderson. The non-central wishart distribution and certain problems of multivariate statistics. *The Annals of Mathematical Statistics*, pages 409–431, 1946. 219

T. C. Aysal and K. E. Barner. Constrained decentralized estimation over noisy channels for sensor networks. *IEEE Transactions on Signal Processing*, 56(4):1398–1410, 2008. 88

D. Bajović, J. Xavier, and B. Sinopoli. Products of stochastic matrices: large deviation rate for Markov chain temporal dependencies. In *Allerton'12, 50th Allerton Conference on Communication, Control, and Computing*, Monticello, Il, October 2012. 55

D. Bajović, J. Xavier, J. M. F. Moura, and B. Sinopoli. Consensus and products of random stochastic matrices: Exact rate for convergence in probability. *IEEE Transactions on Signal Processing*, 61(10): 2557–2571, May 2013. 54, 55, 57, 58

D. Bajovic, D. Jakovetic, J. Xavier, B. Sinopoli, and J. M. F. Moura. Distributed detection via gaussian running consensus: Large deviations asymptotic analysis. *IEEE Transactions on Signal Processing*, 59(9): 4381–4396, 2011. 13, 14, 32, 54, 55, 60, 61, 88

D. Bajović, J. M. F. Moura, J. Xavier, and B. Sinopoli. Distributed inference over directed networks: Performance limits and optimal design. *arXiv preprint arXiv:1504.07526*, 2015. 66

P. Baldi, L. Mazliak, and P. Priouret. *Martingales and Markov chains: solved exercises and elements of theory*. CRC Press, 2002. 203

D. Bertsekas, J. Tsitsiklis, and M. Athans. Convergence theories of distributed iterative processes: A survey. *Technical Report for Information and Decision Systems, Massachusetts Inst. of Technology, Cambridge, MA*, 1984. 87, 88

P. Billingsley. *Convergence of Probability Measures*. John Wiley and Sons, Inc., 1999. 236

S. D. Blostein and H. S. Richardson. A sequential detection approach to target tracking. *Aerospace and Electronic Systems, IEEE Transactions on*, 30(1):197–212, 1994. 6, 13

R. S. Blum, S. A. Kassam, and H. V. Poor. Distributed detection with multiple sensors i. advanced topics. *Proceedings of the IEEE*, 85(1):64–79, 1997. 13, 31

V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, Cambridge, UK, 2008. 67, 72, 84, 113

J.-C. Bourin and E.-Y. Lee. Decomposition and partial trace of positive matrices with hermitian blocks. *International Journal of Mathematics*, 24(01), 2013. 232

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning, Michael Jordan, Editor in Chief*, 3(1):1–122, 2011. 142

P. Braca, S. Marano, and V. Matta. Enforcing consensus while monitoring the environment in wireless sensor networks. *IEEE Transactions on Signal Processing*, 56(7):3375–3380, 2008. 13, 32, 66, 88

P. Braca, S. Marano, V. Matta, and P. Willett. Asymptotic optimality of running consensus in testing binary hypotheses. *IEEE Transactions on Signal Processing*, 58(2):814–825, 2010. 13, 32, 55, 88

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012. 167

S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. 167

F. Bullo, J. Cortes, and S. Martinez. *Distributed control of robotic networks: A mathematical approach to motion coordination algorithms*. Princeton University Press, 209. 123

F. Cattivelli and A. Sayed. Distributed detection over adaptive networks using diffusion adaptation. *IEEE Transactions on Signal Processing*, 59(5):1917–1932, May 2011a. 55

F. Cattivelli and A. H. Sayed. Diffusion LMS-based distributed detection over adaptive networks. In *Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, 2009*, pages 171–175. IEEE, 2009a. 13, 32

F. S. Cattivelli and A. H. Sayed. Distributed detection over adaptive networks based on diffusion estimation schemes. In *IEEE 10th Workshop on Signal Processing Advances in Wireless Communications, 2009. SPAWC'09*, pages 61–65. IEEE, 2009b. 13, 32

F. S. Cattivelli and A. H. Sayed. Diffusion LMS strategies for distributed estimation. *IEEE Transactions on Signal Processing*, 58(3):1035–1048, 2010. 66, 88

F. S. Cattivelli and A. H. Sayed. Distributed detection over adaptive networks using diffusion adaptation. *IEEE Transactions on Signal Processing*, 59(5):1917–1932, 2011b. 13, 32, 88

J.-F. Chamberland and V. V. Veeravalli. Decentralized detection in sensor networks. *IEEE Transactions on Signal Processing*, 51(2):407–416, 2003. 13

C. Chang and K. Dunn. A recursive generalized likelihood ratio test algorithm for detecting sudden changes in linear, discrete systems. In *Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*, pages 731–736. IEEE, 1979. 30

S. Chaudhari, V. Koivunen, and H. V. Poor. Autocorrelation-based decentralized sequential detection of ofdm signals in cognitive radios. *IEEE Transactions on Signal Processing*, 57(7):2690–2700, 2009. 6, 13

J. Chen and A. Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012. 88

J. Chen, C. Richard, and A. H. Sayed. Multitask diffusion adaptation over networks. *IEEE Transactions on Signal Processing*, 62(16):4129–4144, 2014. 66

P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017. 167

H. Chernoff. *Sequential analysis and optimal design*, volume 8. Siam, 1972. 16

K. W. Choi, W. S. Jeon, and D. G. Jeong. Sequential detection of cyclostationary signal for cognitive radio systems. *IEEE Transactions on Wireless Communications*, 8(9):4480–4485, 2009. 6, 13

K. Choromanski, M. Rowland, V. Sindhwani, R. E. Turner, and A. Weller. Structured evolution with compact architectures for scalable policy optimization. *arXiv preprint arXiv:1804.02395*, 2018. 167

F. R. Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997. 199

A. Daneshmand, F. Facchinei, V. Kungurtsev, and G. Scutari. Hybrid random/deterministic parallel algorithms for convex and nonconvex big data optimization. *IEEE Transactions on Signal Processing*, 63(15): 3914–3929, 2015. 123

D. A. Darling and A. Siegert. The first passage problem for a continuous markov process. *The Annals of Mathematical Statistics*, pages 624–639, 1953. 22

A. K. Das and M. Mesbahi. Distributed linear parameter estimation in sensor networks based on Laplacian dynamics consensus algorithm. In *3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks, 2006. SECON'06*, volume 2, pages 440–449. IEEE, 2006. 88, 105, 121

M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69:118–121, 1974. 55

S. N. Deming, L. R. Parker Jr, and M. Bonner Denton. A review of simplex optimization in analytical chemistry. *CRC Critical Reviews in Analytical Chemistry*, 7(3):187–202, 1978. 167

A. G. Dimakis, S. Kar, J. M. Moura, M. G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010. 13, 25, 40, 88

L. E. Dubins and D. A. Freedman. A sharper form of the Borel-Cantelli lemma and the strong law. *The Annals of Mathematical Statistics*, pages 800–807, 1965. 208

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5): 2788–2806, 2015. 143, 148, 159, 168, 170, 173, 175

V. Fabian. Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, 37(1):191–200, Feb 1967. 207

V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, August 1968. 209, 234, 252, 260, 272

V. Fabian. On asymptotically efficient recursive estimation. *The Annals of Statistics*, 6(4):854–866, Jul. 1978. 91

J. Font-Segura and X. Wang. Glrt-based spectrum sensing for cognitive radio with prior information. *IEEE Transactions on Communications*, 58(7):2137–2146, 2010. 6, 30

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956. 168

D. Garber and E. Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 541–549. JMLR. org, 2015. 168

S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM J. Control Optim.*, 29(5):999–1018, September 1991. 72, 113

I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 167

G. A. Gray, T. G. Kolda, K. Sale, and M. M. Young. Optimizing an empirical scoring function for transmembrane protein structure determination. *INFORMS Journal on Computing*, 16(4):406–418, 2004. 167

D. Hajinezhad, M. Hong, and A. Garcia. Zeroth order nonconvex multi-agent optimization over networks. *arXiv preprint arXiv:1710.09997*, 2017. 143

R. Has'minskij. Sequential estimation and recursive asymptotically optimal procedures of estimation and observation control. In *Proc. Prague Symp. Asymptotic Statist.*, volume 1, pages 157–178, Charles Univ., Prague, 1974. 91

E. Hazan and S. Kale. Projection-free online learning. In *International Conference on Machine Learning*, pages 1843–1850, 2012. 167, 168

E. Hazan and H. Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016. 167, 168

C. Heinze, B. McWilliams, and N. Meinshausen. Dual-loco: Distributing statistical estimation using random projections. In *Artificial Intelligence and Statistics*, pages 875–883, 2016. 1, 65, 67, 68, 69

P. Hieber and M. Scherer. A note on first-passage times of continuously time-changed brownian motion. *Statistics & Probability Letters*, 82(1):165–172, 2012. 22

M. D. Ilic' and J. Zaborszky. *Dynamics and Control of Large Electric Power Systems*. Wiley, 2000. 91, 92, 100

J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer-Verlag Berlin, 1987. 14

A. Jadbabaie, J. Lin, and A. S. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001, Jun. 2003. 13, 31, 88

M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013. 167, 168, 170

D. Jakovetic, J. Xavier, and J. M. F. Moura. Cooperative convex optimization in networked systems: Augmented Lagrangian algorithms with directed gossip communication. *IEEE Transactions on Signal Processing*, 59(8):3889–3902, August 2011. doi: 10.1109/TSP.2011.2146776. 66

D. Jakovetic, J. Xavier, and J. M. F. Moura. Fast distributed gradient methods. *IEEE Trans. Autom. Contr.*, 59(5):1131–1146, May 2014a. 142

D. Jakovetic, J. Xavier, and J. M. F. Moura. Convergence rates of distributed Nesterov-like gradient methods on random networks. *IEEE Transactions on Signal Processing*, 62(4):868–882, February 2014b. 123, 124, 143

D. Jakovetic, J. M. F. Moura, and J. Xavier. Distributed detection over noisy networks: Large deviations analysis. *IEEE Transactions on Signal Processing*, 60(8):4306–4320, 2012. 13, 14, 32, 88

D. Jakovetic, D. Bajovic, N. Krejic, and N. K. Jerinkic. Distributed gradient methods with variable number of working nodes. *IEEE Trans. Signal Processing*, 64(15):4080–4095, 2016. 68, 69, 102, 124, 144

D. Jakovetic, D. Bajovic, A. K. Sahu, and S. Kar. Convergence rates for distributed stochastic optimization over random networks. In *57th IEEE Conference on Decision and Control (CDC)*, Miami, 2018. Available at https://www.dropbox.com/s/zylonzrhypy29zj/MainCDC2018.pdf. 143

A. Jayaprakasam and V. Sharma. Sequential detection based cooperative spectrum sensing algorithms in cognitive radio. In *2009 First UK-India International Workshop on Cognitive Wireless Systems (UKI-WCWS)*, pages 1–6. IEEE, 2009. 6, 13

R. I. Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, 1969. 86, 89, 90, 102, 103

S. Kar and J. M. F. Moura. Distributed linear parameter estimation in sensor networks: convergence properties. In *42nd Asilomar Conference on Signals, Systems and Computers*, pages 1347–1351, Pacific Grove, CA, Oct. 26-29 2008a. 88

S. Kar. *Large scale networked dynamical systems: Distributed inference*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2010. URL http://gradworks.umi.com/34/21/3421734.html. 105, 115, 228

S. Kar and J. M. F. Moura. Asymptotically efficient distributed estimation with exponential family statistics. *IEEE Transactions on Information Theory*, 60(8):4811–4831, 2014. 88, 93, 96, 97, 102, 103, 207, 221, 224, 254, 261

S. Kar and J. M. Moura. Consensus based detection in sensor networks: Topology optimization under practical constraints. *Proc. 1st Intl. Wrkshp. Inform. Theory Sensor Networks*, 2007. 13, 31

S. Kar and J. M. Moura. Distributed linear parameter estimation in sensor networks: Convergence properties. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 1347–1351. IEEE, 2008b. 13

S. Kar and J. M. Moura. Sensor networks with random links: Topology design for distributed consensus. *IEEE Transactions on Signal Processing*, 56(7):3315–3326, 2008c. 25

S. Kar and J. M. Moura. Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise. *IEEE Transactions on Signal Processing*, 57(1):355–369, 2009. 25

S. Kar and J. M. Moura. Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):674–690, 2011. 66, 68, 69, 71, 93, 94, 105, 107, 114, 116, 118, 119, 120, 121, 123, 128, 258

S. Kar, S. Aldosari, and J. M. Moura. Topology for distributed inference on graphs. *IEEE Transactions on Signal Processing*, 56(6):2609, 2008. 13, 25

S. Kar, R. Tandon, H. V. Poor, and S. Cui. Distributed detection in noisy sensor networks. In *IEEE International Symposium on Information Theory Proceedings (ISIT), 2011*, pages 2856–2860. IEEE, 2011. 13, 14, 32

S. Kar, J. M. Moura, and K. Ramanan. Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication. *IEEE Transactions on Information Theory*, 58(6): 3575–3605, 2012. 13, 74, 86, 88, 91, 92, 93, 100, 121, 183

S. Kar, J. M. F. Moura, and H. V. Poor. Distributed linear parameter estimation: Asymptotically efficient adaptive strategies. *SIAM Journal on Control and Optimization*, 51(3):2200–2229, 2013a. 67, 68, 69, 72, 74, 84, 93, 94, 107, 110, 113, 208, 244

S. Kar, J. M. Moura, and H. V. Poor. On a consistent procedure for distributed recursive nonlinear least-squares estimation. In *Proceedings of the Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 891–894. IEEE, 2013b. 88, 131

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952. 160, 166, 168, 170

H. Kushner and G. Yin. Asymptotic properties of distributed and communicating stochastic approximation algorithms. *SIAM J. Control Optim.*, 25(5):1266–1290, Sept. 1987. 87

S. Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016. 168

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pages 496–504, 2015. 168

A. Lalitha, A. Sarwate, and T. Javidi. Social learning and distributed hypothesis testing. In *IEEE International Symposium on Information Theory (ISIT), 2014*, pages 551–555. IEEE, 2014. 32, 71

A. Lalitha, T. Javidi, and A. Sarwate. Social learning and distributed hypothesis testing. *arXiv preprint arXiv:1410.4307*, 2015. 32

G. Lan, S. Lee, and Y. Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *arXiv preprint arXiv:1701.03961*, 2017. 102, 124, 144

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml. 82

K. Liu and Y. Mei. Improved performance properties of the cisprt algorithm for distributed sequential detection, 2017. 24

Q. Liu and A. T. Ihler. Distributed estimation, information loss and exponential families. In *Advances in Neural Information Processing Systems*, pages 1098–1106, 2014. 67, 69

S. Liu, J. Chen, P.-Y. Chen, and A. Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In *International Conference on Artificial Intelligence and Statistics*, pages 288–297, 2018. 168

I. Lobel and A. E. Ozdaglar. Distributed subgradient methods for convex optimization over random networks. *IEEE Trans. Automat. Contr.*, 56(6):1291–1306, Jan. 2011. 123, 124, 143

I. Lobel, A. Ozdaglar, and D. Feijer. Distributed multi-agent optimization with state-dependent communication. *Mathematical Programming*, 129(2):255–284, 2011. 123, 124, 143

C. G. Lopes and A. H. Sayed. Diffusion least-mean squares over adaptive networks: Formulation and performance analysis. *IEEE Transactions on Signal Processing*, 56(7):3122–3136, July 2008. 66, 88, 105, 121

G. Lorden. On excess over the boundary. *The Annals of Mathematical Statistics*, pages 520–527, 1970. 205

Z.-Q. Luo. Universal decentralized estimation in a bandwidth constrained sensor network. *IEEE Transactions on Information Theory*, 51(6):2210–2219, 2005. 88

C. Ma and M. Takáč. Partitioning data on features or samples in communication-efficient distributed optimization? *arXiv preprint arXiv:1510.06688*, 2015. 67, 69

C. Ma, V. Smith, M. Jaggi, M. Jordan, P. Richtarik, and M. Takac. Adding vs. averaging in distributed primal-dual optimization. In *International Conference on Machine Learning*, pages 1973–1982, 2015. 1, 65, 67, 68, 69

A. L. Marsden, M. Wang, J. Dennis, and P. Moin. Trailing-edge noise reduction using derivative-free optimization and large-eddy simulation. *Journal of Fluid Mechanics*, 572:13–36, 2007. 167

A. L. Marsden, J. A. Feinstein, and C. A. Taylor. A computational framework for derivative-free optimization of cardiovascular geometries. *Computer methods in applied mechanics and engineering*, 197(21-24):1890–1905, 2008. 167

V. Matta, P. Braca, S. Marano, and A. H. Sayed. Diffusion-based adaptive distributed detection: Steady-state performance in the slow adaptation regime. *IEEE Transactions on Information Theory*, 62(8): 4710–4732, Aug 2016. ISSN 0018-9448. doi: 10.1109/TIT.2016.2580665. 88

A. Mokhtari and A. Ribeiro. Dsa: Decentralized double stochastic averaging gradient algorithm. *The Journal of Machine Learning Research*, 17(1):2165–2199, 2016. 124, 144

A. Mokhtari, H. Hassani, and A. Karbasi. Conditional gradient method for stochastic submodular maximization: Closing the gap. In *International Conference on Artificial Intelligence and Statistics*, pages 1886–1895, 2018. 167, 168, 169, 175

J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Püschel. Distributed optimization with local domains: Applications in mpc and network flows. *IEEE Transactions on Automatic Control*, 60(7):2004–2009, 2015. 108

A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48 61, Jan. 2009. 66, 102, 103

A. Nedic and A. Olshevsky. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Transactions on Automatic Control*, 61(12):3936–3947, Dec. 2016. 124, 143

A. Nedić, A. Olshevsky, and C. A. Uribe. Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs. *arXiv preprint arXiv:1410.1977*, 2014. 32, 55, 66

A. Nedić, A. Olshevsky, and C. A. Uribe. Network independent rates in distributed learning. In *American Control Conference (ACC), 2016*, pages 1072–1077. IEEE, 2016. 55

J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4): 308–313, 1965. 177

Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011. 158, 159, 160, 168, 170

M. B. Nevelson and R. Z. Khasminskiĭ. *Stochastic approximation and recursive estimation*, volume 47. Amer Mathematical Society, 1973. 113

R. Olfati-Saber and R. M. Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 49(9):15201533, Sept. 2004. 55

R. Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, January 2007. 13, 31, 88

R. Olfati-Saber, E. Franco, E. Frazzoli, and J. S. Shamma. Belief consensus and distributed hypothesis testing in sensor networks. In *Networked Embedded Sensing and Control*, pages 169–182. Springer, 2006. 31

D. B. Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, pages 1075–1090, 1956. 26

M. Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007. 178

J. Pfanzagl. Asymptotic optimum estimation and test procedures. In *Proceedings of the Prague Symposium on Asymptotic Statistics*, volume 1, Sept. 3 - 6 1973. 91

G. Poole and T. Boullion. A survey on m-matrices. *SIAM review*, 16(4):419–427, 1974. 47

M. J. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis*, pages 51–67. Springer, 1994. 177

M. J. Powell. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, pages 26–46, 2009. 177

F. Ram, S. Wright, S. Singh, and M. De Graef. Error analysis of the crystal orientations obtained by the dictionary approach to ebsd indexing. *Ultramicroscopy*, 181:17–26, 2017. 178

S. S. Ram, A. Nedic, and V. V. Veeravalli. Incremental stochastic subgradient algorithms for convex optimization. *SIAM Journal on Optimization*, 20(2):691–717, June 2009. 66, 102, 103

S. Ram, A. Nedić, and V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010a. 66, 102, 103

S. Ram, V. Veeravalli, and A. Nedic. Distributed and recursive parameter estimation in parametrized linear state-space models. *to appear in IEEE Transactions on Automatic Control*, 55(2):488– 492, February 2010b. 66, 102, 103, 105, 121

B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pages 693–701, 2011. 1, 65

S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Stochastic frank-wolfe methods for nonconvex optimization. In *Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on*, pages 1244–1251. IEEE, 2016. 167, 168, 169, 177

A. Ruszczyński. A merit function approach to the subgradient method with averaging. *Optimisation Methods and Software*, 23(1):161–172, 2008. 169

A. K. Sahu, D. Jakovetic, and S. Kar. CREDO: A communication-efficient distributed estimation algorithm. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 516–520, June 2018a. 102

A. K. Sahu and S. Kar. Distributed sequential detection for Gaussian shift-in-mean hypothesis testing. *IEEE Transactions on Signal Processing*, 64(1):89–103, 2016. 66, 105, 121

A. K. Sahu and S. Kar. Recursive distributed detection for composite hypothesis testing: Nonlinear observation models in additive gaussian noise. *IEEE Transactions on Information Theory*, 63(8):4797–4828, 2017. 60, 71, 88

A. K. Sahu, S. Kar, J. M. Moura, and H. V. Poor. Distributed constrained recursive nonlinear least-squares estimation: Algorithms and asymptotics. *IEEE Transactions on Signal and Information Processing over Networks*, 2(4):426–441, 2016. 102, 103, 107, 114

A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar. Distributed zeroth order optimization over random networks: A Kiefer-Wolfowitz stochastic approximation approach. In *57th IEEE Conference on Decision and Control (CDC)*, Miami, 2018b. Available at https://www.dropbox.com/s/kfc2hgbfcx5yhr8/MainCDC2018KWSA.pdf. 143, 168

A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar. Non-asymptotic rates for communication efficient distributed zeroth order strongly convex optimization. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018c. Available at https://www.dropbox.com/s/53rfp208rmysym3/globalsip2018.pdf. 143

A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar. Communication efficient distributed weighted non-linear least squares estimation. *arXiv preprint arXiv:1801.04050*, 2018d. 124, 144

A. K. Sahu, D. Jakovetic, and S. Kar. Communication optimality trade-offs for distributed estimation. *arXiv preprint arXiv:1801.04050*, 2018e. 54, 55, 102, 124, 142, 144

D. Sakrison. Efficient recursive estimation; application to estimating the parameters of a covariance function. *International Journal of Engineering Science*, 3(4):461–483, 1965. 91

L. L. Scharf. *Statistical signal processing*, volume 98. Addison-Wesley Reading, MA, 1991. 33

I. D. Schizas, A. Ribeiro, and G. B. Giannakis. Consensus in Ad Hoc WSNs with noisy links - part I: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 56(1):350–364, January 2008a. 105, 121

I. Schizas, G. Mateos, and G. Giannakis. Stability analysis of the consensus-based distributed LMS algorithm. In *Proceedings of the 33rd International Conference on Acoustics, Speech, and Signal Processing*, pages 3289–3292, Las Vegas, Nevada, USA, April 1-4 2008b. 105, 121

I. D. Schizas, A. Ribeiro, and G. B. Giannakis. Consensus in ad hoc WSNs with noisy links part I: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 56(1):350–364, 2008c. 88

S. Shahrampour, A. Rakhlin, and A. Jadbabaie. Distributed detection: Finite-time analysis and impact of network topology. *arXiv preprint arXiv:1409.8606*, 2014. 55

W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM J. Optim.*, 25(2):944?–966, 2015. 142

D. Siegmund and E. Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, pages 255–271, 1995. 30

I. Sohn, J. Kim, S.-H. Jung, and C. Park. Gradient lasso for cox proportional hazards model. *Bioinformatics*, 25(14):1775–1781, 2009. 178

J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992. 160

S. Stankovic, M. Stankovic, and D. Stipanovic. Decentralized parameter estimation by consensus based stochastic approximation. In *46th IEEE Conference on Decision and Control*, pages 1535–1540, New Orleans, LA, USA, 12-14 Dec. 2007. 105, 121

C. Stone. Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics*, 3(2): 267–284, Mar. 1975. 91

A. Tahbaz-Salehi and A. Jadbabaie. Consensus over ergodic stationary graph processes. *IEEE Transactions on Automatic Control*, 55(1):225–230, Jan. 2010. 55

A. Tajer, G. H. Jajamovich, X. Wang, and G. V. Moustakides. Optimal joint target detection and parameter estimation by mimo radar. *IEEE Journal of Selected Topics in Signal Processing*, 4(1):127–145, 2010. 6, 30

B. Touri and A. Nedic. Product of random stochastic matrices. *IEEE Transactions on Automatic Control*, 59(2):437–448, 2014. 55

Z. J. Towfic, J. Chen, and A. H. Sayed. Excess-risk of distributed stochastic learners. *IEEE Transactions on Information Theory*, 62(10):5753–5785, 2016. xi, xii, 101, 102, 124, 143, 161

K. Tsianos and M. Rabbat. Distributed strongly convex optimization. *50th Annual Allerton Conference onCommunication, Control, and Computing*, Oct. 2012. 124, 143

K. Tsianos, S. Lawlor, and M. G. Rabbat. Communication/computation tradeoffs in consensus-based distributed optimization. In *Advances in neural information processing systems*, pages 1943–1951, 2012. 68, 69, 102, 124, 144

K. I. Tsianos, S. F. Lawlor, J. Y. Yu, and M. G. Rabbat. Networked optimization with adaptive communication. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 579–582. IEEE, 2013. 68, 69, 102, 124, 144

J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, September 1986. 87

J. Tsitsiklis. *Problems in decentralized decision making and computation.* PHD, Massachusetts Institute of Technology, Cambridge, MA, 1984. 87

J. N. Tsitsiklis et al. Decentralized detection. *Advances in Statistical Signal Processing*, 2(2):297–344, 1993. 13, 31

N. D. Vanli, M. O. Sayin, and S. S. Kozat. Stochastic subgradient algorithms for strongly convex optimization over distributed networks. *IEEE Transactions on network science and engineering*, 4(4):248–260, Oct.-Dec. 2017. 124, 143

V. Vapnik. *Statistical learning theory. 1998*, volume 3. Wiley, New York, 1998. 141

V. V. Veeravalli, T. Basar, and H. V. Poor. Decentralized sequential detection with a fusion center performing the sequential test. *IEEE Transactions on Information Theory*, 39(2):433–442, 1993. 13

A. Wald. *Sequential analysis.* Courier Corporation, 1973. 205

A. Wald, J. Wolfowitz, et al. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948. 15

A. Wald et al. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 1945. 6, 13, 16, 25

Y. Wang, S. Du, S. Balakrishnan, and A. Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 1356–1365, 2018. 168

Z. Wang, Z. Yu, Q. Ling, D. Berberidis, and G. B. Giannakis. Decentralized rls with data-adaptive censoring for regressions over large-scale networks. *arXiv preprint arXiv:1612.08263*, 2016. 102, 124, 144

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938. 32, 52

A. S. Willsky and H. L. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, 21(1):108–112, 1976. 30

C.-F. Wu. Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, pages 501–513, 1981. 86

C. Xi and U. A. Khan. DEXTRA: A fast algorithm for optimization over directed graphs. *IEEE Transactions on Automatic Control*, 2017. to appear, DOI: 10.1109/TAC.2017.2672698. 142

L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1): 65–78, 2004. 24, 25

Y. Yang, G. Scutari, D. P. Palomar, and M. Pesavento. A parallel decomposition method for nonconvex stochastic multi-agent optimization problems. *IEEE Transactions on Signal Processing*, 64(11):2949–2964, 2016. 169

D. Yuan, Y. Hong, D. W. C. Ho, and G. Jiang. Optimal distributed stochastic mirror descent for strongly convex optimization. *Automatica*, 90:196–203, April 2018. 124, 143

K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM J. Optim.*, 26 (3):1835?–1854, 2016. 142

K. Yuan, B. Ying, X. Zhao, and A. H. Sayed. Exact diffusion for distributed optimization and learning — Part I: Algorithm development. 2017. arxiv preprint, arXiv:1702.05122. 142

S. Zarrin and T. J. Lim. Composite hypothesis testing for cooperative spectrum sensing in cognitive radio. In *IEEE International Conference on Communications, 2009. ICC'09.*, pages 1–5. IEEE, 2009. 6, 30

O. Zeitouni, J. Ziv, and N. Merhav. When is the generalized likelihood ratio test optimal? *IEEE Transactions on Information Theory*, 38(5):1597–1602, 1992. 6, 30

J. Zhang and R. S. Blum. Asymptotically optimal truncated multivariate gaussian hypothesis testing with application to consensus algorithms. *IEEE Transactions on Signal Processing*, 62(2):431–442, 2014. 13

J. Zhang, K. You, and T. Basar. Distributed discrete-time optimization in multi-agent networks using only sign of relative state. *arXiv preprint arXiv:1709.08360*, 2017. 121

Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013a. 1, 65

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013b. 1, 65, 67, 69

Q. Zhou, S. Kar, L. Huie, H. V. Poor, and S. Cui. Robust distributed least-squares estimation in sensor networks with node failures. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pages 1–6. IEEE, 2011. 13

Q. Zhou, S. Kar, L. Huie, and S. Cui. Distributed estimation in sensor networks with imperfect model information: an adaptive learning-based approach. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012*, pages 3109–3112. IEEE, 2012. 13

Q. Zou, S. Zheng, and A. H. Sayed. Cooperative sensing via sequential detection. *IEEE Transactions on Signal Processing*, 58(12):6266–6283, 2010. 6, 30, 32

Part IV

Appendix

Appendix A

Spectral Graph Theory: Preliminaries

Spectral Graph Theory For an undirected graph $G = (V, E)$, V denotes the set of agents or vertices with cardinality $|V| = N$, and E the set of edges with $|E| = M$. The unordered pair $(i, j) \in E$ if there exists an edge between agents i and j. We only consider simple graphs, i.e., graphs devoid of self loops and multiple edges. A path between agents i and j of length m is a sequence $(i = p_0, p_1, \cdots, p_m = j)$ of vertices, such that $(p_t, p_{t+1}) \in E$, $0 \leq t \leq m-1$. A graph is connected if there exists a path between all the possible agent pairings. The neighborhood of an agent n is given by $\Omega_n = \{j \in V | (n, j) \in E\}$. The degree of agent n is given by $d_n = |\Omega_n|$. The structure of the graph may be equivalently represented by the symmetric $N \times N$ adjacency matrix $\mathbf{A} = [A_{ij}]$, where $A_{ij} = 1$ if $(i, j) \in E$, and 0 otherwise. The degree matrix is represented by the diagonal matrix $\mathbf{D} = diag(d_1 \cdots d_N)$. The graph Laplacian matrix is represented by

$$\mathbf{L} = \mathbf{D} - \mathbf{A}. \tag{A.1}$$

The Laplacian is a positive semidefinite matrix, hence its eigenvalues can be sorted and represented in the following manner

$$0 = \lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \cdots \leq \lambda_N(\mathbf{L}). \tag{A.2}$$

Furthermore, a graph is connected if and only if $\lambda_2(\mathbf{L}) > 0$ (see Chung (1997) for instance).

Appendix B

Proofs of Theorems in Chapter 2

Proof of Lemma 2.5.5. Let us define the event A_s^i as $\{\gamma_{d,i}^l \leq S_{d,i}(s) \leq \gamma_{d,i}^h\}$. Now, note that

$$\mathbb{P}_1(T_{d,i} > t) = \mathbb{P}_1(\cap_{s=1}^t A_s^i), \tag{B.1}$$

and

$$\mathbb{P}_1(\cap_{s=1}^t A_s^i) \leq \mathbb{P}_1(A_t^i). \tag{B.2}$$

By Proposition 2.4.2, under H_1, for any t, the quantity $S_{d,i}(t)$ is Gaussian with mean mt and variance upper bounded by $\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}$. Hence we have, for all $i = 1, 2, \ldots, N$.

$$\mathbb{P}_1(T_{d,i} > t) \leq \mathbb{Q}\Big(\frac{-\gamma_{d,i}^h + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\Big). \tag{B.3}$$

□

Proof of Corollary 2.5.6. For simplicity of notation, let $a = \frac{Nm}{4}$ and $b = \frac{\sigma^2\pi^2}{2N(\gamma_c^h - \gamma_c^l)^2}$. From (2.41), we have,

$$\begin{aligned}
&\frac{1}{t}\log(\mathbb{P}_1(T_c > t)) \geq \frac{1}{t}\log\left(\exp\left(\frac{N\mu\gamma_c^l}{\sigma^2}\right)K_t^\infty\left(\gamma_c^h\right) - \exp\left(\frac{N\mu\gamma_c^h}{\sigma^2}\right)K_t^\infty\left(\gamma_c^l\right)\right) \\
&= \frac{1}{t}\log\left(\exp\left(-\left(a+b\right)t\right)\right) \\
&+ \frac{1}{t}\log\left(b\sum_{s=1}^{\infty}\frac{s(-1)^{s+1}}{a+s^2b}\exp\left(-b(s^2-1)t\right)\right. \\
&\left.\times\left(\exp\left(\frac{N\mu\gamma_c^l}{\sigma^2}\right)\sin\left(\frac{s\pi\gamma_c^h}{\gamma_c^h-\gamma_c^l}\right) - \exp\left(\frac{N\mu\gamma_c^h}{\sigma^2}\right)\sin\left(\frac{s\pi\gamma_c^l}{\gamma_c^h-\gamma_c^l}\right)\right)\right).
\end{aligned} \tag{B.4}$$

For all $t, S \geq 1$, let

$$\begin{aligned}
U(t,S) = \frac{1}{t}\log\left(b\sum_{s=1}^{S}\frac{s(-1)^{s+1}}{a+s^2b}\exp\left(-b(s^2-1)t\right)\right. \\
\left.\times\left(\exp\left(\frac{N\mu\gamma_c^l}{\sigma^2}\right)\sin\left(\frac{s\pi\gamma_c^h}{\gamma_c^h-\gamma_c^l}\right) - \exp\left(\frac{N\mu\gamma_c^h}{\sigma^2}\right)\sin\left(\frac{s\pi\gamma_c^l}{\gamma_c^h-\gamma_c^l}\right)\right)\right)
\end{aligned} \tag{B.5}$$

and let $g = \exp\left(\frac{N\mu\gamma_c^h}{\sigma^2}\right) + \exp\left(\frac{N\mu\gamma_c^l}{\sigma^2}\right)$.
Note that for all $t \geq 1$, the limit

$$\lim_{S\to\infty} U(t,S) \tag{B.6}$$

exists and is finite (by Theorem 2.5.4), and similarly for all $S \geq 1$,

$$\begin{aligned}\lim_{t\to\infty} U(t,S) &= \lim_{t\to\infty} \frac{1}{t}\log\left(b\sum_{s=1}^{S}\frac{s(-1)^{s+1}}{a+s^2b}\exp\left(-b(s^2-1)t\right)\right.\\ &\left.\times\sin\left(\frac{s\pi\gamma_c^h}{\gamma_c^h-\gamma_c^l}\right)\left(\exp\left(\frac{N\mu\gamma_c^l}{\sigma^2}\right)+(-1)^{s+1}\exp\left(\frac{N\mu\gamma_c^h}{\sigma^2}\right)\right)\right)\\ &= \lim_{t\to\infty}\frac{1}{t}\log\left(\frac{bg}{a+b}\sin\left(\frac{\pi\gamma_c^h}{\gamma_c^h-\gamma_c^l}\right)\right)\\ &= 0,\end{aligned} \tag{B.7}$$

where we use the fact that only the largest exponent in a finite summation of exponential terms contributes to its log-normalised limit as $t \to \infty$ and

$$\sin\left(\frac{s\pi\gamma_c^h}{\gamma_c^h-\gamma_c^l}\right) = (-1)^s\sin\left(\frac{s\pi\gamma_c^l}{\gamma_c^h-\gamma_c^l}\right). \tag{B.8}$$

Finally, using the fact that there exists a constant $c_5 > 0$ (independent of t and S) such that for all $t, S \geq 1$,

$$U(t,S) \leq \frac{1}{t}\log\left(bg\sum_{s=1}^{S}\frac{s}{a+s^2b}\exp\left(-b(s^2-1)t\right)\right) \leq c_5, \tag{B.9}$$

we may conclude that the convergence in (B.6)-(B.7) are uniform in S and t respectively. This in turn implies that the order of the limits may be interchanged and we have that

$$\lim_{t\to\infty}\lim_{S\to\infty} U(t,S) = \lim_{S\to\infty}\lim_{t\to\infty} U(t,S) = 0. \tag{B.10}$$

Hence, we have from (B.4) and (B.10),

$$\begin{aligned}&\liminf_{t\to\infty}\frac{1}{t}\log(\mathbb{P}_1(T_c > t)) \geq -(a+b)\\ &+\lim_{t\to\infty}\lim_{S\to\infty}\frac{1}{t}\log\left(b\sum_{s=1}^{S}\frac{s(-1)^{s+1}}{a+s^2b}\exp\left(-b(s^2-1)t\right)\right.\\ &\left.\times\sin\left(\frac{s\pi\gamma_c^h}{\gamma_c^h-\gamma_c^l}\right)\left(\exp\left(\frac{N\mu\gamma_c^l}{\sigma^2}\right)-(-1)^s\exp\left(\frac{N\mu\gamma_c^h}{\sigma^2}\right)\right)\right)\\ &= -(a+b)+\lim_{S\to\infty}\lim_{t\to\infty}\frac{1}{t}\log\left(b\sum_{s=1}^{S}\frac{s(-1)^{s+1}}{a+s^2b}\exp\left(-b(s^2-1)t\right)\right.\\ &\left.\times\sin\left(\frac{s\pi\gamma_c^h}{\gamma_c^h-\gamma_c^l}\right)\left(\exp\left(\frac{N\mu\gamma_c^l}{\sigma^2}\right)+(-1)^{s+1}\exp\left(\frac{N\mu\gamma_c^h}{\sigma^2}\right)\right)\right)\\ &= -(a+b)+\lim_{S\to\infty}\lim_{t\to\infty}\frac{1}{t}\log\left(\frac{bg}{a+b}\sin\left(\frac{\pi\gamma_c^h}{\gamma_c^h-\gamma_c^l}\right)\right)\\ &= -(a+b) = -\left(\frac{Nm}{4}+\frac{\sigma^2\pi^2}{2N(\gamma_c^h-\gamma_c^l)^2}\right).\end{aligned} \tag{B.11}$$

□

Proof of Theorem 2.5.7. We use the following upper bound for $\mathbb{Q}$ function in the proof below

$$\mathbb{Q}(x) \leq \frac{1}{x\sqrt{2\pi}} e^{-x^2/2} \tag{B.12}$$

From (2.43),(B.12) and (B.9), we have,

$$\begin{aligned}
&\limsup_{t\to\infty} \frac{1}{t} \log(\mathbb{P}_1(T_{d,i} > t)) \\
&\leq \limsup_{t\to\infty} \frac{1}{t} \log\left(\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\frac{2mt}{N} + 2m\frac{r^2(1-r^{2t})}{(1-r^2)}}}{(-\gamma^h_{d,i} + mt)} e^{\frac{-N(-\gamma^h_{d,i}+mt)^2}{4mt+4mN\frac{r^2(1-r^{2t})}{(1-r^2)}}} \right) \\
&\leq \limsup_{t\to\infty} \frac{1}{t} \left(\log\left(\frac{\sqrt{\frac{2mt}{N} + 2m\frac{r^2(1-r^{2t})}{(1-r^2)}}}{\sqrt{2\pi}(mt - \gamma^h_{d,i})} \right) - \frac{N(\gamma^h_{d,i})^2}{4mt + 4m\frac{r^2(1-r^{2t})}{(1-r^2)}} \right. \\
&\left. - \frac{Nmt}{4 + 4\frac{r^2(1-r^{2t})}{(t(1-r^2))}} + \frac{Nm\gamma^h_{d,i}t}{2mt + 2mN\frac{r^2(1-r^{2t})}{(1-r^2)}} \right) \\
&\Rightarrow \limsup_{t\to\infty} \frac{1}{t} \log(\mathbb{P}_1(T_{d,i} > t)) \leq -\frac{Nm}{4}.
\end{aligned} \tag{B.13}$$

□

The proof of Theorem 2.5.8 requires an intermediate result that estimates the divergence between the agent statistics over time.

Lemma B.0.1. *Let the Assumptions 2.3.1, 2.4.1 and 2.4.2 hold. Then, there exists a constant c_1, depending on the network topology and the Gaussian model statistics only, such that*

$$\mathbb{E}_1 \left[\sup_{t\geq 0} \|S_{d,i}(t) - S_{d,j}(t)\| \right] \leq c_1 \tag{B.14}$$

for all agent pairs (i, j).

Proof. Denoting by $\mathbf{S}_d(t) = t\mathbf{P}_d(t)$ the vector of the agent test statistics $S_{d,i}(t)$'s, we have,

$$\mathbf{S}_d(t+1) = W\left(\mathbf{S}_d(t) + \eta(t+1)\right). \tag{B.15}$$

Let $\overline{S}_d(t)$ denote the average of the $S_{d,i}(t)$'s, i.e.,

$$\overline{S}_d(t) = (1/N)\,.\,(S_{d,1}(t) + \cdots + S_{d,N}(t)), \tag{B.16}$$

Noting that $J\mathbf{S}_d(t) = \overline{S}_d(t)\mathbf{1}$ and $WJ = JW = J$, we have from (B.15)

$$\mathbf{v}_{t+1} = (W - J)\,\mathbf{v}_t + \mathbf{u}_{t+1}, \tag{B.17}$$

where $\mathbf{v}_t$ and $\mathbf{u}_t$, for all $t \geq 0$, are given by

$$\mathbf{v}_t = \mathbf{S}_d(t) - \overline{S}_d(t)\mathbf{1} \tag{B.18}$$

and

$$\mathbf{u}_{t+1} = (W - J)\,\eta(t+1). \tag{B.19}$$

It is important to note that the sequence $\{\mathbf{u}_t\}$ is i.i.d. Gaussian and, in particular, there exists a constant c_2 such that $\mathbb{E}_1[\|\mathbf{u}_t\|^2] \leq c_2$ for all t.
Now, by (B.17) we obtain

$$\|\mathbf{v}_{t+1}\| \leq r\|\mathbf{v}_t\| + \|\mathbf{u}_{t+1}\|, \tag{B.20}$$

where recall $r = \|\mathbf{W} - \mathbf{J}\| < 1$. Since the sequence $\{\mathbf{u}_t\}$ is i.i.d. and $\mathcal{L}_2$-bounded, an application of the Robbins-Siegmund's lemma (see Baldi et al. (2002)) yields

$$\mathbb{E}_1\left[\sup_{t\geq 0}\|\mathbf{v}_t\|\right] \leq c_3 < \infty, \tag{B.21}$$

where c_3 is a constant that may be chosen as a function of r, c_2 and $\mathbb{E}_1[\|\mathbf{v}_0\|]$. Now, noting that, for any pair (i,j),

$$\mathbb{E}_1\left[\sup_{t\geq 0}\|S_{d,i}(t) - S_{d,j}(t)\|\right] \leq \mathbb{E}_1\left[\sup_{t\geq 0}\|S_{d,i}(t) - \overline{S}_d(t)\|\right] + \mathbb{E}_1\left[\sup_{t\geq 0}\|S_{d,j}(t) - \overline{S}_d(t)\|\right] \leq 2c_3, \tag{B.22}$$

the desired assertion follows. □

Proof of Theorem 2.5.8. We prove the upper bound in Theorem 2.5.8 first. Since $\mathbb{P}_1(T_{d,i} < \infty) = 1$, for the upper bound we have,

$$\begin{aligned}
\mathbb{E}_1[T_{d,i}] &= \sum_{t=0}^{\infty}\mathbb{P}_1(T_{d,i} > t)\\
&\overset{(a)}{\leq} \sum_{0}^{\infty}\mathbb{Q}\Big(\frac{-\gamma_{d,i}^h + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\Big)\\
&= \underbrace{\sum_{0}^{\lfloor\frac{\gamma_{d,i}^h}{m}\rfloor}\mathbb{Q}\Big(\frac{-\gamma_{d,i}^h + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\Big)}_{(1)} + \underbrace{\sum_{\lfloor\frac{\gamma_{d,i}^h}{m}\rfloor+1}^{\lfloor\frac{3\gamma_{d,i}^h}{2m}\rfloor}\mathbb{Q}\Big(\frac{-\gamma_{d,i}^h + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\Big)}_{(2)}\\
&+ \underbrace{\sum_{\lfloor\frac{3\gamma_{d,i}^h}{2m}\rfloor+1}^{\lfloor\frac{2\gamma_{d,i}^h}{m}\rfloor}\mathbb{Q}\Big(\frac{-\gamma_{d,i}^h + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\Big)}_{(3)} + \underbrace{\sum_{\lfloor\frac{2\gamma_{d,i}^h}{m}\rfloor+1}^{\infty}\mathbb{Q}\Big(\frac{-\gamma_{d,i}^h + mt}{\sqrt{\frac{2mt}{N} + \frac{2mr^2(1-r^{2t})}{1-r^2}}}\Big)}_{(4)}\\
&\overset{(b)}{\leq} \frac{\gamma_{d,i}^h}{m} + \frac{\gamma_{d,i}^h}{4m} + \frac{1}{2}e^{\frac{N\gamma_{d,i}^h}{2(k+1)}}\sum_{\lfloor\frac{3\gamma_{d,i}^h}{2m}\rfloor+1}^{\lfloor\frac{2\gamma_{d,i}^h}{m}\rfloor} e^{\frac{-(N\gamma_{d,i}^h)^2 - Nm^2t^2}{4m(k+1)t}} + \frac{1}{2(1-e^{\frac{-Nm}{4(k+1)}})}
\end{aligned}$$

$$\leq \frac{5\gamma_{d,i}^h}{4m} + \frac{1}{2(1-e^{\frac{-Nm}{4(k+1)}})} + \frac{1}{2}e^{\frac{3N\gamma_{d,i}^h}{8(k+1)}} \sum_{\lfloor\frac{3\gamma_{d,i}^h}{2m}\rfloor+1}^{\lfloor\frac{2\gamma_{d,i}^h}{m}\rfloor} e^{\frac{-Nmt}{4(k+1)}}$$

$$\leq \frac{5\gamma_{d,i}^h}{4m} + \frac{1}{1-e^{\frac{-Nm}{4(k+1)}}}, \tag{B.23}$$

where (a) is due to the upper bound derived in Lemma 2.5.5 and (b) is due to the following : 1) $\forall t \in [0, \lfloor\frac{\gamma_{d,i}^h}{m}\rfloor]$ in (1), $-\gamma_{d,i}^h + mt$ is negative and hence every term in the summation can be upper bounded by 1; 2) $\forall t \in [\lfloor\frac{\gamma_{d,i}^h}{m}\rfloor + 1, \lfloor\frac{3\gamma_{d,i}^h}{2m}\rfloor]$ in (2), $-\gamma_{d,i}^h + mt$ is positive and hence every term in the summation can be upper bounded by $\frac{1}{2}$; and 3) for the terms (3) and (4), the inequality $\mathbb{Q}(x) \leq \frac{1}{2}e^{-x^2/2}$ is used and the sums are upper bounded by summing the resulting geometric series.

In order to obtain the lower bound, we first note that conditioned on hypothesis H_1, at the stopping time $T_{d,i}$, an agent exceeds the threshold $\gamma_{d,i}^h$ with probability at least $1-\epsilon$ and is lower than the threshold $\gamma_{d,i}^l$ with probability at most ϵ. Moreover, with $\alpha = \beta = \epsilon$, $\gamma_{d,i}^h = -\gamma_{d,i}^l$.

Now, denote by E_i^h the event $E_i^h = \{S_{d,i}(T_{d,i}) \geq \gamma_{d,i}^h\}$ and by E_i^l the event $E_i^l = \{S_{d,i}(T_{d,i}) \leq \gamma_{d,i}^l\}$. Since $\mathbb{P}_1(T_{d,i} < \infty) = 1$, we have that

$$\mathbb{E}_1\left[S_{d,i}(t)\right] = \mathbb{E}_1\left[S_{d,i}(t).\mathbb{I}_{E_i^h}\right] + \mathbb{E}_1\left[S_{d,i}(t).\mathbb{I}_{E_i^l}\right], \tag{B.24}$$

where $\mathbb{I}_{\{.\}}$ denotes the indicator function. We now lower bound the quantities on the R.H.S. of (B.24). Note that $\gamma_{d,i}^h \geq 0$ and $S_{d,i}(t) \geq \gamma_{d,i}^h$ on E_i^h. Hence

$$\mathbb{E}_1\left[S_{d,i}(t).\mathbb{I}_{E_i^h}\right] \geq \gamma_{d,i}^h \mathbb{P}_1\left(E_i^h\right) \geq (1-\epsilon)\gamma_{d,i}^h. \tag{B.25}$$

Now recall the construction in the proof of Lemma B.0.1 and note that by (B.15) we have

$$S_{d,i}(t) = S_{d,i}(t-1) - \sum_{j\in\Omega_i} w_{ij}\left(S_{d,i}(t-1) - S_{d,j}(t-1)\right) + \eta_i(t). \tag{B.26}$$

Hence, we have that

$$S_{d,i}(T_{d,i}).\mathbb{I}_{E_i^l} \tag{B.27}$$

$$\geq S_{d,i}(T_{d,i}-1).\mathbb{I}_{E_i^l} - \sum_{j\in\Omega_i} w_{ij}\|S_{d,i}(T_{d,i}-1) - S_{d,j}(T_{d,i}-1)\| - \|\eta_i(T_{d,i})\| \tag{B.28}$$

$$\geq S_{d,i}(T_{d,i}-1).\mathbb{I}_{E_i^l} - \sum_{j\in\Omega_i} w_{ij}\sup_{t\geq 0}\|S_{d,i}(t) - S_{d,j}(t)\| - \|\eta_i(T_{d,i})\|. \tag{B.29}$$

Now, observe that on the event E_i^l, $S_{d,i}(T_{d,i}-1) > \gamma_{d,i}^l$ a.s. Since $\gamma_{d,i}^l < 0$ and $\mathbb{P}_1(E_i^l) \leq \epsilon$ (by hypothesis), we have that

$$\gamma_{d,i}^l\epsilon \leq \gamma_{d,i}^l\mathbb{P}_1(E_i^l) \tag{B.30}$$

$$= \mathbb{E}_1\left[\gamma_{d,i}^l.\mathbb{I}_{E_i^l}\right] \leq \mathbb{E}_1\left[S_{d,i}(T_{d,i}-1).\mathbb{I}_{E_i^l}\right]. \tag{B.31}$$

Note that, by Lemma B.0.1, we have

$$\mathbb{E}_1\left[\sum_{j\in\Omega_i} w_{ij}\sup_{t\geq 0}\|S_{d,i}(t)-S_{d,j}(t)\|\right] \tag{B.32}$$

$$\leq \sum_{j\in\Omega_i} w_{ij}\mathbb{E}_1\left[\sup_{t\geq 0}\|S_{d,i}(t)-S_{d,j}(t)\|\right] \tag{B.33}$$

$$\leq |\Omega_i|c_1. \tag{B.34}$$

Finally, by arguments similar to Wald (1973); Lorden (1970) for characterizing expected overshoots in stopped random sums (see, in particular, Theorem 1 in Lorden (1970)) it follows that there exists a constant c_4 (depending on the Gaussian model statistics and the network topology only) such that

$$\mathbb{E}_1\left[\|\eta_i(T_{d,i})\|\right] \leq c_4. \tag{B.35}$$

In particular, note that, the constant c_4 in (B.35) may be chosen to be independent of the thresholds and, hence, the error tolerance parameter ϵ. Substituting (B.30)-(B.35) in (B.27) we obtain

$$\mathbb{E}_1\left[S_{d,i}(T_{d,i}).\mathbb{I}_{E_i^l}\right] \geq \gamma_{d,i}^l\epsilon - |\Omega_i|c_1 - c_4. \tag{B.36}$$

This together with (B.24)-(B.25) yield

$$\mathbb{E}_1\left[S_{d,i}(T_{d,i})\right] \geq (1-\epsilon)\,\gamma_{d,i}^h + \gamma_{d,i}^l\epsilon - |\Omega_i|c_1 - c_4 \tag{B.37}$$

$$= (1-2\epsilon)\,\gamma_{d,i}^h - c, \tag{B.38}$$

where the last equality follows by noting that $\gamma_{d,i}^h = -\gamma_{d,i}^l$ and taking the constant c to $c = |\Omega_i|c_1 + c_4$. We note that the event $\{T_{d,i} = t\}$ is independent of $\eta_i, i > t$. We also have from Theorem 2.5.1 that $\mathbb{P}_1(T_{d,i} < \infty) = 1$. Hence, we have,

$$\begin{aligned}
\mathbb{E}_1[S_{d,i}(T_{d,i})] &= \mathbb{E}_1[\sum_{j=1}^{T_{d,i}} \mathbf{e_i}^\top \mathbf{W}^{t+1-j}\eta(j)] \\
&= \mathbb{E}_1\left[\sum_{j=1}^{\infty}\mathbb{I}_{\{T_{d,i}\geq j\}}\mathbf{e_i}^\top\mathbf{W}^{T_{d,i}+1-j}\eta(j)\right] \\
&= \sum_{j=1}^{\infty}\mathbb{E}_1\left[\mathbb{I}_{\{T_{d,i}\geq j\}}\mathbf{e_i}^\top\mathbf{W}^{T_{d,i}+1-j}\right]\mathbb{E}_1\left[\eta(j)\right] \\
&= m\sum_{j=1}^{\infty}\mathbb{E}_1\left[\mathbb{I}_{\{T_{d,i}\geq j\}}\mathbf{e_i}^\top\mathbf{W}^{T_{d,i}+1-j}\right]\mathbf{1} \\
&= m\sum_{j=1}^{\infty}\mathbb{E}_1\left[\mathbb{I}_{\{T_{d,i}\geq j\}}\mathbf{e_i}^\top\mathbf{W}^{T_{d,i}+1-j}\mathbf{1}\right] \\
&= m\mathbb{E}_1\left[T_{d,i}\right].
\end{aligned} \tag{B.39}$$

Combining (B.39) and (B.37) we have,

$$\frac{(1-2\epsilon)\gamma_{d,i}^{h}}{m} - \frac{c}{m} \leq \mathbb{E}_1[T_{d,i}] \tag{B.40}$$

and the desired assertion follows.

□

Proof of Theorem 2.5.9. From (2.46), we first note that

$$\frac{\mathbb{E}_1[T_{d,i}]}{\mathbb{E}_1[T_c]} \geq 1, \forall i = 1, 2, \ldots, N. \tag{B.41}$$

From the upper bound for the stopping time distribution derived for the $\mathcal{CISPRT}$ in (B.3), we have the following upper bound for $\mathbb{E}_1[T_{d,i}]$

$$\mathbb{E}_1[T_{d,i}] \leq \frac{5\gamma_{d,i}^{h}}{4m} + \frac{1}{1-e^{\frac{-Nm}{4(k+1)}}}. \tag{B.42}$$

We choose the threshold $\gamma_{d,i}^{h}$ to be

$$\gamma_{d,i}^{h} = \gamma_{d}^{h,0} = \frac{8(k+1)}{7N}(\log(\frac{2}{\epsilon}) - \log(1-e^{\frac{-Nm}{4(k+1)}})). \tag{B.43}$$

Using (B.42) and (B.43), we have

$$\limsup_{\epsilon\to 0} \frac{\mathbb{E}_1[T_{d,i}]}{\mathbb{E}_1[T_c]} \leq \lim_{\epsilon\to 0} \frac{\frac{10}{7}(k+1)\log(\frac{2}{\epsilon}) + O(1)}{(1-2\epsilon)\log(\frac{1-\epsilon}{\epsilon})}. \tag{B.44}$$

Noting that,

$$\limsup_{\epsilon\to 0} \frac{O(1)}{(1-2\epsilon)\log(\frac{1-\epsilon}{\epsilon})} = 0, \tag{B.45}$$

we obtain

$$\limsup_{\epsilon\to 0} \frac{\mathbb{E}_1[T_{d,i}]}{\mathbb{E}_1[T_c]} \leq \frac{10(k+1)}{7}. \tag{B.46}$$

Combining (B.46) and (B.41), the result follows. □

Appendix C

Proofs of Theorems in Chapter 3

C.1 Proof of Main Results : $\mathcal{CIGLRT} - \mathcal{NL}$

C.1.1 Proof of Theorem 3.7.1

Proof. The proof of Theorem 3.7.1 is accomplished in steps, the key ingredients being Lemma C.1.1 and Lemma C.1.2 which concern the boundedness of the processes $\{\theta_n(t)\}$, $n = 1, \cdots, N$ and subsequently the consistency of the agent estimate sequences respectively. To this end, we follow the basic idea developed in Kar and Moura (2014), but with subtle modifications to take into account the state-dependent nature of the innovation gains. We state Lemma C.1.1 and Lemma C.1.2 here, with the proofs relegated to Appendix C.3.

Lemma C.1.1. *Let the hypothesis of Theorem 3.7.1 hold. Then, for each n and $\forall \theta^*$ the process $\{\theta_n(t)\}$*

$$\mathbb{P}_{\theta^*}\left(\sup_{t\geq 0}\|\theta_n(t)\| < \infty\right) = 1. \tag{C.1}$$

Lemma C.1.2. *Let the hypotheses of Theorem 3.7.1 hold. Then, for each n and $\forall \theta^*$, we have,*

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty}\theta_n(t) = \theta^*\right) = 1. \tag{C.2}$$

In the sequel, we analyze the rate of convergence of the parameter estimate sequence to the true parameter. We will use the following approximation result (Lemma C.1.3) and the generalized convergence criterion (Lemma C.1.4) for the proof of Theorem 3.7.1.

Lemma C.1.3 (Lemma 4.3 in Fabian (1967))**.** *Let $\{b_t\}$ be a scalar sequence satisfying*

$$b_{t+1} \leq \left(1 - \frac{c}{t+1}\right) b_t + d_t(t+1)^{-\tau}, \tag{C.3}$$

where $c > \tau, \tau > 0$, and the sequence $\{d_t\}$ is summable. Then, we have,

$$\limsup_{t\to\infty}\ (t+1)^{\tau} b_t < \infty. \tag{C.4}$$

Lemma C.1.4 (Lemma 10 in Dubins and Freedman (1965)). *Let $\{J(t)\}$ be an $\mathbb{R}$-valued $\{\mathcal{F}_{t+1}\}$-adapted process such that $\mathbb{E}[J(t)|\mathcal{F}_t] = 0$ a.s. for each $t \geq 1$. Then the sum $\sum_{t\geq 0} J(t)$ exists and is finite a.s. on the set where $\sum_{t\geq 0} \mathbb{E}\left[J^2(t)|\mathcal{F}_t\right]$ is finite.*

We now return to the proof of Theorem 4.1.

Proof of Theorem 3.7.1. We follow closely the corresponding development in Lemma 5.9 of Kar et al. (2013a). Define $\bar{\tau} \in [0, 1/2)$ such that,

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty} (t+1)^{\bar{\tau}} \left\|\mathbf{x}(t)\right\| = 0\right) = 1, \tag{C.5}$$

where $\mathbf{x}(t) = \theta(t) - \mathbf{1}_N \otimes \theta^*$. Note that such a $\bar{\tau}$ exists by Lemma C.1.2 (in particular, by taking $\bar{\tau} = 0$). We now analyze and finally show that there exists a τ such that $\bar{\tau} < \tau < 1/2$ for which the claim holds. Now, choose a $\hat{\tau} \in (\bar{\tau}, 1/2)$ and let $\mu = (\hat{\tau} + \bar{\tau})/2$. By standard algebraic manipulations, it can be readily seen that the recursion for $\{\mathbf{x}(t)\}$ satisfies

$$\begin{aligned}
&\left\|\mathbf{x}(t+1)\right\|^2 = \left\|\mathbf{x}(t)\right\|^2 - 2\beta_t \mathbf{x}^\top(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)\mathbf{x}(t)\\
&- 2\alpha_t \mathbf{x}^\top(t)\mathbf{G}\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right) - \mathbf{h}\left(\theta^*\right)\right)\\
&+ \beta_t^2 \mathbf{x}^\top(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)^2\mathbf{x}(t)\\
&+ 2\alpha_t\beta_t \mathbf{x}^\top(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)\mathbf{G}\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right) - \mathbf{h}\left(\theta^*\right)\right)\\
&+ \alpha_t^2 \left(\mathbf{y}(t) - \mathbf{h}\left(\theta^*\right)\right)^\top \mathbf{\Sigma}^{-1}\mathbf{G}^\top\left(\theta(t)\right)\mathbf{G}\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\left(\mathbf{y}(t) - \mathbf{h}\left(\theta^*\right)\right)\\
&+ \alpha_t^2 \left(\mathbf{h}\left(\theta(t)\right) - \mathbf{h}\left(\theta^*\right)\right)^\top \mathbf{\Sigma}^{-1}\mathbf{G}^\top\left(\theta(t)\right)\\
&\times \mathbf{G}\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right) - \mathbf{h}\left(\theta^*\right)\right)\\
&+ 2\alpha_t \mathbf{x}^\top(t)\mathbf{G}\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\left(\mathbf{y}(t) - \mathbf{h}(\theta^*)\right).
\end{aligned} \tag{C.6}$$

Let $\mathbf{J}(t) = \mathbf{G}\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\left(\mathbf{y}(t) - \mathbf{h}\left(\theta^*\right)\right)$. From Assumption 2.4.1, we have that $\left\|\nabla \mathbf{h}_n\left(\theta_n(t)\right)\right\|$ is uniformly bounded from above by k_n for all n. Hence, we have that $\left\|\mathbf{G}\left(\theta(t)\right)\right\| \leq \max_{n=1,\cdots,N} k_n$. Now, we consider the term $\alpha_t^2 \left\|\mathbf{J}(t)\right\|^2$. Since, the noise process under consideration is a temporally independent Gaussian sequence and $2\mu < 1$, we have,

$$\sum_{t\geq 0} (t+1)^{2\mu}\alpha_t^2 \left\|\mathbf{J}(t)\right\|^2 < \infty \text{ a.s.} \tag{C.7}$$

Let $\mathbf{W}(t) = \alpha_t \mathbf{x}^\top(t)\mathbf{G}\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\left(\mathbf{y}(t) - \mathbf{h}(\theta^*)\right)$. It follows that $\mathbb{E}_{\theta^*}\left[\mathbf{W}(t)|\mathcal{F}_t\right] = 0$.
We also have that $\mathbb{E}_{\theta^*}\left[\mathbf{W}^2(t)|\mathcal{F}_t\right] \leq \alpha_t^2 \left\|\mathbf{x}(t)\right\|^2 \left\|\mathbf{J}(t)\right\|^2$. Noting, that the noise under consideration is temporally independent with finite second moment, we have,

$$\mathbb{E}_{\theta^*}\left[\mathbf{W}^2(t)|\mathcal{F}_t\right] = o\left((t+1)^{-2-2\bar{\tau}}\right) \tag{C.8}$$

and hence,

$$\mathbb{E}_{\theta^*}\left[(t+1)^{4\mu}\mathbf{W}^2(t)|\mathcal{F}_t\right] = o\left((t+1)^{-2+2\hat{\tau}}\right). \tag{C.9}$$

Hence, by Lemma C.1.4, we conclude that $\sum_{t\geq 0}(t+1)^{2\mu}\mathbf{W}(t)$ exists and is finite, as $2\hat{\tau} < 1$ and hence the

left hand side (L.H.S) in (C.9) is summable. Using all the inequalities derived in (C.156)-(C.158), we have,

$$\begin{aligned}\|\mathbf{x}(t+1)\|^2 &\leq \left(1 - c_1\alpha_t + c_5\left(\alpha_t\beta_t + \alpha_t^2\right)\right)\|\mathbf{x}(t)\|^2 \\ &- c_6(\beta_t - \beta_t^2)\|\mathbf{x}_{C\perp}(t)\|^2 + \alpha_t^2\|\mathbf{J}(t)\|^2 + 2\mathbf{W}(t). \end{aligned} \tag{C.10}$$

Finally, noting that $c_1\alpha_t$ dominates $c_5\left(\alpha_t\beta_t + \alpha_t^2\right)$ and β_t dominates β_t^2, we obtain

$$\|\mathbf{x}(t+1)\|^2 \leq (1 - c_1\alpha_t)\|\mathbf{x}(t)\|^2 + \alpha_t^2\|\mathbf{J}(t)\|^2 + 2\mathbf{W}(t). \tag{C.11}$$

Now, using the analysis in (C.7)-(C.9), we have, from (C.11)

$$\|\mathbf{x}(t+1)\|^2 \leq (1 - c_1\alpha_t)\|\mathbf{x}(t)\|^2 + d_t(t+1)^{-2\mu}, \tag{C.12}$$

where

$$d_t(t+1)^{-2\mu} = \alpha_t^2\|\mathbf{J}(t)\|^2 + 2\mathbf{W}(t). \tag{C.13}$$

Finally, noting that $c_1\alpha_t(t+1) \geq 1 > 2\mu$, an immediate application of Lemma C.1.3 gives

$$\limsup_{t\to\infty}(t+1)^{2\mu}\|\mathbf{x}(t)\|^2 < \infty \text{ a.s.} \tag{C.14}$$

So, we have that, there exists a τ with $\bar{\tau} < \tau < \mu$ for which $(t+1)^{\tau}\|\mathbf{x}(t)\| \to 0$ as $t \to \infty$. Thus for every $\bar{\tau}$ for which (3.32) holds, there exists $\tau \in (\bar{\tau}, 1/2)$ for which the result in (3.32) continues to hold. We thus conclude that the result holds for all $\tau \in [0, 1/2)$.

□

C.1.2 Proof of Theorem 3.9.1

Proof. The proof of Theorem 3.9.1 needs the following Lemma from Fabian (1968) (stated in a form suitable to our needs) concerning the asymptotic normality of non-Markov stochastic recursions and an intermediate result which concerns with the asymptotic normality of the averaged decision statistic.

Lemma C.1.5 (Theorem 2.2 in Fabian (1968)). *Let $\{\mathbf{z}_t\}$ be an $\mathbb{R}^k$-valued $\{\mathcal{F}_t\}$-adapted process that satisfies*

$$\mathbf{z}_{t+1} = \left(\mathbf{I}_k - \frac{1}{t+1}\mathbf{\Gamma}_t\right)\mathbf{z}_t + (t+1)^{-1}\mathbf{\Phi}_t\mathbf{V}_t + (t+1)^{-3/2}\mathbf{T}_t, \tag{C.15}$$

where the stochastic processes $\{\mathbf{V}_t\}, \{\mathbf{T}_t\} \in \mathbb{R}^k$ while $\{\mathbf{\Gamma}_t\}, \{\mathbf{\Phi}_t\} \in \mathbb{R}^{k\times k}$. Moreover, for each t, $\mathbf{V}_{t-1}$ and $\mathbf{T}_t$ are $\mathcal{F}_t$ -adapted, whereas the processes $\{\mathbf{\Gamma}_t\}$, $\{\mathbf{\Phi}_t\}$ are $\{\mathcal{F}_t\}$ adapted.
Also, assume that

$$\mathbf{\Gamma}_t \to \mathbf{I}_k, \ \mathbf{\Phi}_t \to \mathbf{\Phi} \text{ and } \mathbf{T}_t \to 0 \text{ a.s. as } t \to \infty. \tag{C.16}$$

Furthermore, let the sequence $\{\mathbf{V}_t\}$ satisfy $\mathbb{E}[\mathbf{V}_t|\mathcal{F}_t] = 0$ for each t and suppose there exists a positive constant C and a matrix Σ such that $C > \left\|\mathbb{E}\left[\mathbf{V}_t\mathbf{V}_t^{\top}|\mathcal{F}_t\right] - \Sigma\right\| \to 0$ a.s. as $t \to \infty$ and with $\sigma_{t,r}^2 = \int_{\|\mathbf{V}_t\|^2 \geq r(t+1)}\|\mathbf{V}_t\|^2 d\mathbb{P}$, let $\lim_{t\to\infty}\frac{1}{t+1}\sum_{s=0}^{t}\sigma_{s,r}^2 = 0$ for every $r > 0$.

Then, we have,

$$(t+1)^{1/2}\mathbf{z}_t \xRightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{\Phi}\Sigma\mathbf{\Phi}^\top\right). \tag{C.17}$$

We state the lemma concerning the asymptotic normality of the averaged decision statistic here, while the proof is relegated to Appendix C.3.

Lemma C.1.6. *Let the hypotheses of Theorem 3.9.1 hold. Consider the averaged decision statistic sequence, $\{z_{avg}(t)\}$, defined as $z_{avg}(t) = \frac{1}{N}\sum_{n=1}^{N} z_n(t)$. Then, we have, under $\mathbb{P}_{\theta^*}$ for all $\|\theta^*\| > 0$,*

$$\begin{aligned}&\sqrt{t+1}\left(z_{avg}(t) - \frac{\mathbf{h}^\top(\theta_N^*)\,\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{2N}\right)\\ &\xRightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\mathbf{h}^\top(\theta_N^*)\,\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{N^2}\right), \forall n.\end{aligned} \tag{C.18}$$

We now use a lemma which establishes that the sequences $\{z_{\text{avg}}(t)\}$ and $\{z_n(t)\}$ are indistinguishable in the $\sqrt{t}$ time scale. We state the lemma here, while the proof is relegated to Appendix C.3.

Lemma C.1.7. *Given the averaged decision statistic sequence, $\{z_{avg}(t)\}$, for each $\delta_0 \in [0,1)$ we have*

$$\mathbb{P}_{\theta^*}(\lim_{t\to\infty}(t+1)^{\delta_0}(\mathbf{z}(t) - \mathbf{1}_N \otimes z_{avg}(t) = \mathbf{0}) = 1. \tag{C.19}$$

We now return to the proof of Theorem 3.9.1.

Proof of Theorem 3.9.1. Note that as δ_0 in Lemma C.1.7 can be chosen to be greater than $\frac{1}{2}$, we have for all n,

$$\begin{aligned}&\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty}\left\|\sqrt{t+1}\left(z_n(t) - \frac{\mathbf{h}^\top(\theta_N^*)\,\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{2N}\right)\right.\right.\\ &\left.\left.-\sqrt{t+1}\left(z_{\text{avg}}(t) - \frac{\mathbf{h}^\top(\theta_N^*)\,\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{2N}\right)\right\| = 0\right)\\ &= \mathbb{P}_{\theta^*}\left(\lim_{t\to\infty}\left\|\sqrt{t+1}\,(z_n(t) - z_{\text{avg}}(t))\right\| = 0\right)\\ &= \mathbb{P}_{\theta^*}\left(\lim_{t\to\infty}\left\|(t+1)^{0.5-\delta_0}(t+1)^{\delta_0}\,(z_n(t) - z_{\text{avg}}(t))\right\| = 0\right) = 1,\end{aligned} \tag{C.20}$$

where the last step follows from Lemma C.1.7 and the fact that $\delta_0 > 1/2$. Thus, the difference of the sequences $\left\{\sqrt{t+1}\left(z_n(t) - \frac{\mathbf{h}^\top(\theta_N^*)\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{2N}\right)\right\}$ and $\left\{\sqrt{t+1}\left(z_{\text{avg}}(t) - \frac{\mathbf{h}^\top(\theta_N^*)\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{2N}\right)\right\}$ converges a.s. to zero and hence we have,

$$\begin{aligned}&\sqrt{t+1}\left(z_n(t) - \frac{\mathbf{h}^\top(\theta_N^*)\,\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{2N}\right)\\ &\xRightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\mathbf{h}^\top(\theta_N^*)\,\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{N^2}\right).\end{aligned} \tag{C.21}$$

□

C.1.3 Proof of Theorem 3.7.2

Proof. From (3.14), we have,

$$\begin{aligned}
&\mathbb{P}_{M,\theta^*}(t) = \mathbb{P}_{1,\theta^*}\left(z_n(t) < \eta\right) \\
&= \mathbb{P}_{1,\theta^*}\left(z_n(t) - \frac{\mathbf{h}^\top\left(\theta_N^*\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N}\right. \\
&\left.< \eta - \frac{\mathbf{h}^\top\left(\theta_N^*\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N}\right) \\
&= \mathbb{P}_{1,\theta^*}\left(\sqrt{t+1}\left(z_n(t) - \frac{\mathbf{h}^\top\left(\theta_N^*\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N}\right)\right. \\
&\left.< \sqrt{t+1}\left(\eta - \frac{\mathbf{h}^\top\left(\theta_N^*\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N}\right)\right).
\end{aligned} \tag{C.22}$$

Now, invoking Theorem 3.9.1, where we have established the asymptotic normality for the decision statistic sequence $\{z_n(t)\}$, we have,

$$\begin{aligned}
&\lim_{t\to\infty}\mathbb{P}_{1,\theta^*}\left(\sqrt{t+1}\left(z_n(t) - \frac{\mathbf{h}^\top\left(\theta_N^*\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N}\right)\right. \\
&\left.< \sqrt{t+1}\left(\eta - \frac{\mathbf{h}^\top\left(\theta_N^*\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N}\right)\right) \\
&= \mathbb{P}_{1,\theta^*}(z < -\infty) = 0,
\end{aligned} \tag{C.23}$$

where z is a normal random variable with $z \sim \mathcal{N}\left(0, \frac{\mathbf{h}^\top(\theta_N^*)\boldsymbol{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{N^2}\right)$. In the derivation of (C.23) we have used the Portmanteau characterization for weak convergence and the fact that

$$\eta < \frac{\mathbf{h}^\top\left(\theta_N^*\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N}. \tag{C.24}$$

Hence, we have, from (C.22) and (C.23)

$$\lim_{t\to\infty}\mathbb{P}_{M,\theta^*}(t) = 0 \tag{C.25}$$

as long as (C.24) holds.

For the null hypothesis $\mathcal{H}_0$, from (3.14) and with $0 < \lambda < 1$, we have,

$$\begin{aligned}
&\mathbb{P}_{FA}(t) = \mathbb{P}_0\left(z_n(t) > \eta\right) \\
&= \mathbb{P}_0\left(\frac{1}{t}\sum_{s=0}^{t-1}\mathbf{e}_n^\top\mathbf{W}^{t-1-s}\mathbf{h}^\top\left(\theta(s)\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}(s) - \frac{\mathbf{h}\left(\theta(s)\right)}{2}\right) > \eta\right) \\
&= \mathbb{P}_0\left(\frac{1}{t}\sum_{s=0}^{t-1}\sum_{j=1}^{N}\phi_{n,j}(s,t-1)\left(\mathbf{h}_j^\top\left(\theta_j(s)\right)\boldsymbol{\Sigma}_j^{-1}\gamma_j(s)\right.\right. \\
&\left.\left. - \frac{\mathbf{h}_j^\top\left(\theta_j(s)\right)\boldsymbol{\Sigma}_j^{-1}\mathbf{h}_j\left(\theta_j(s)\right)}{2}\right) > \eta\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}_0 \left(\frac{1}{t} \sum_{s=0}^{t-1} \sum_{j=1}^{N} \phi_{n,j}(s, t-1) \left(\frac{\gamma_j^\top(s) \mathbf{\Sigma}_j^{-1} \gamma_j(s)}{2} \right.\right. \\
&\left.\left. - \frac{\left(\gamma_j(s) - \mathbf{h}_j\left(\theta_j(s)\right)\right)^\top \mathbf{\Sigma}_j^{-1} \left(\gamma_j(s) - \mathbf{h}_j\left(\theta_j(s)\right)\right)}{2} \right) > \eta \right) \\
&\leq \mathbb{P}_0 \left(\frac{1}{t} \sum_{s=0}^{t-1} \sum_{j=1}^{N} \phi_{n,j}(s, t-1) \frac{\gamma_j^\top(s) \mathbf{\Sigma}_j^{-1} \gamma_j(s)}{2} > \eta \right) \\
&\overset{(a)}{\leq} \mathbb{P}_0 \left(\frac{1}{t} \sum_{s=0}^{t-1} \sum_{j=1}^{N} \left(\frac{1}{N} + \sqrt{N} r^{t-1-s} \right) \frac{\gamma_j^\top(s) \mathbf{\Sigma}_j^{-1} \gamma_j(s)}{2} > \eta \right) \\
&\leq \exp\left(- \frac{t\eta\lambda}{\frac{1}{N} + \sqrt{N}} \right) \\
&\times \prod_{j=1}^{N} \prod_{s=0}^{t-1} \mathbb{E}_0 \left[\exp\left(\lambda \frac{\frac{1}{N} + \sqrt{N} r^{t-1-s}}{\frac{2}{N} + 2\sqrt{N}} \gamma_j(s)^\top \mathbf{\Sigma}_j^{-1} \gamma_j(s) \right) \right] \\
&\overset{(b)}{=} \exp\left(- \frac{t\eta\lambda}{\frac{1}{N} + \sqrt{N}} \right) \\
&\times \exp\left(- \sum_{s=0}^{t-1} \left(\frac{\sum_{n=1}^{N} M_n}{2} \right) \log\left(1 - \frac{\lambda\left(\frac{1}{N} + \sqrt{N} r^{t-1-s} \right)}{\frac{1}{N} + \sqrt{N}} \right) \right) \\
&\leq \exp\left(- \frac{t\eta\lambda}{\frac{1}{N} + \sqrt{N}} \right) \exp\left(- \left(\frac{\sum_{n=1}^{N} M_n}{2} \right) \log\left(1 - \lambda\right) \right) \\
&\times \exp\left(- (t-1) \left(\frac{\sum_{n=1}^{N} M_n}{2} \right) \log\left(1 - \frac{\lambda\left(\frac{1}{N} + \sqrt{N} r \right)}{\frac{1}{N} + \sqrt{N}} \right) \right),
\end{aligned} \tag{C.26}
$$

where $\phi_{n,j}\left(s, t-1\right)$ denotes the (n,j)-th element of $\mathbf{W}^{t-1-s}$, (a) follows due to $\|\phi_{n,j}\left(s, t-1\right) - \frac{1}{N}\| \leq \sqrt{N} r^{t-1-s}$ and (b) follows due to the fact that the random variable $\gamma_j(s)^\top \mathbf{\Sigma}_j^{-1} \gamma_j(s)$ is a chi-squared random variable with M_j degrees of freedom and the associated moment generating function exists since $\lambda < 1$. Now, taking limits on both sides of the equation (C.26), we have,

$$
\begin{aligned}
&\frac{1}{t} \log\left(\mathbb{P}_0\left(z_n(t) > \eta \right) \right) \\
&\leq - \frac{\eta\lambda}{\frac{1}{N} + \sqrt{N}} - \left(\frac{\sum_{n=1}^{N} M_n}{2t} \right) \log\left(1 - \lambda\right) \\
&- \frac{t-1}{t} \left(\frac{\sum_{n=1}^{N} M_n}{2} \right) \log\left(1 - \frac{\lambda\left(\frac{1}{N} + \sqrt{N} r \right)}{\frac{1}{N} + \sqrt{N}} \right) \\
&\Rightarrow \limsup_{t \to \infty} \frac{1}{t} \log\left(\mathbb{P}_0\left(z_n(t) > \eta \right) \right) \\
&\leq - \frac{\eta\lambda}{\frac{1}{N} + \sqrt{N}} - \left(\frac{\sum_{n=1}^{N} M_n}{2} \right) \log\left(1 - \frac{\lambda\left(\frac{1}{N} + \sqrt{N} r \right)}{\frac{1}{N} + \sqrt{N}} \right) \\
&= -LE(\lambda).
\end{aligned} \tag{C.27}
$$

First we note that, as (C.27) holds for all $\lambda \in (0,1)$, we have that

$$\limsup_{t\to\infty} \frac{1}{t} \log\left(\mathbb{P}_0\left(z_n(t) > \eta\right)\right) \leq -LE(1-\epsilon), \tag{C.28}$$

where $\epsilon \in (0,1)$. Moreover, as $LE(\lambda)$ is a continuous function of λ in the interval $\lambda \in (0,1]$, we can force ϵ to zero and thereby conclude that

$$\limsup_{t\to\infty} \frac{1}{t} \log\left(\mathbb{P}_0\left(z_n(t) > \eta\right)\right) \leq -LE(1). \tag{C.29}$$

Now consider λ^* which is given by

$$\lambda^* = \frac{\frac{1}{N} + \sqrt{N}}{\frac{1}{N} + \sqrt{N}r} - \frac{\left(\frac{1}{N} + \sqrt{N}\right)\sum_{n=1}^{N} M_n}{2\eta}. \tag{C.30}$$

It is to be noted that λ^* is positive when

$$\eta > \frac{\left(\frac{1}{N} + \sqrt{N}r\right)\sum_{n=1}^{N} M_n}{2}. \tag{C.31}$$

Furthermore, $LE(\lambda)$ is maximixed at $\lambda = \lambda^*$ when $\lambda^* \in (0,1)$. Hence, in the case when $\lambda^* \in (0,1)$, we have

$$\limsup_{t\to\infty} \frac{1}{t} \log\left(\mathbb{P}_0\left(z_n(t) > \eta\right)\right) \leq -LE(\lambda^*). \tag{C.32}$$

It is to be noted that $LE(\lambda)$ is an increasing function of λ in the interval $(0,\lambda^*)$ and hence in the case when $\lambda^* > 1$, we have that $LE(\lambda)$ is non-negative and increasing in the interval $(0,1)$ and we have the exponent as $LE(1)$ from (C.29). Finally, combining (C.29) and (C.32), we have,

$$\limsup_{t\to\infty} \frac{1}{t} \log\left(\mathbb{P}_0\left(z_n(t) > \eta\right)\right) \leq -LE\left(\min\{\lambda^*, 1\}\right). \tag{C.33}$$

Finally, the above arguments and the threshold choices obtained in (C.24) and (C.31) establish that as long as the true θ^* satisfies the following condition

$$\frac{\mathbf{h}^\top\left(\theta_N^*\right)\mathbf{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N} > \frac{\left(\frac{1}{N} + \sqrt{N}r\right)\sum_{n=1}^{N} M_n}{2}, \tag{C.34}$$

any η satisfying

$$\frac{\left(\frac{1}{N} + \sqrt{N}r\right)\sum_{n=1}^{N} M_n}{2} < \eta < \frac{\mathbf{h}^\top\left(\theta_N^*\right)\mathbf{\Sigma}^{-1}\mathbf{h}\left(\theta_N^*\right)}{2N}. \tag{C.35}$$

would guarantee $\mathbb{P}_{M,\theta^*}(t), \mathbb{P}_{FA}(t) \to 0$ as $t \to \infty$. Hence, the assertion is proved. □

C.2 Proof of Main Results : $\mathcal{CIGLRT} - \mathcal{L}$

C.2.1 Proof of Theorem 3.7.3

Proof. The following result which characterizes $\left\|\mathbf{I}_{NM}-\beta_t\left(\mathbf{L}\otimes\mathbf{I}_M\right)-\alpha_t\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right\|$, will be crucial for the subsequent analysis. We state the result here, while the proof is relegated to Appendix C.4.

Lemma C.2.1. *Let the Assumptions 3.6.1-3.6.2 hold. Consider the parameter estimate update of the $\mathcal{CIGLRT}-\mathcal{L}$ algorithm in (3.20). Then, we have,*

$$\left\|\mathbf{I}_{NM}-\beta_t\left(\mathbf{L}\otimes\mathbf{I}_M\right)-\alpha_t\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right\|\leq 1-c_1\alpha_t,\quad\forall t\geq t_1, \tag{C.36}$$

where

$$\begin{aligned}c_1&=\min_{\|x\|=1}x^\top\left(\mathbf{L}\otimes\mathbf{I}_M+\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right)x\\&=\lambda_{min}\left(\mathbf{L}\otimes\mathbf{I}_M+\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right),\end{aligned} \tag{C.37}$$

$$t_1=\max\{t_2,t_3\}, \tag{C.38}$$

and t_2,t_3 are positive constants (integers) chosen such that $\forall t\geq t_2$,

$$\beta_t\lambda_N\left(\mathbf{L}\right)+\alpha_t\lambda_{max}\left(\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right)\leq 1, \tag{C.39}$$

and $\forall t\geq t_3$,

$$\alpha_t\lambda_{min}\left(\mathbf{L}\otimes\mathbf{I}_M+\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right)<1 \tag{C.40}$$

respectively.

Under the null hypothesis, we have, for all $\lambda\in(0,1)$,

$$\begin{aligned}z_n(kt)&=\mathbf{e}_n^\top\mathbf{W}^{k-1}\mathbf{G}_\theta(k(t-1))\mathbf{\Sigma}^{-1}\\&\times\left(\mathbf{s}(k(t-1))-\frac{\mathbf{G}_H^\top\theta(k(t-1))}{2}\right).\end{aligned} \tag{C.41}$$

From (C.41), we have,

$$\begin{aligned}&\mathbb{P}_0\left(z_n(kt)>\eta\right)\leq e^{-\frac{(k(t-1)+1)\eta\lambda}{\frac{1}{N}+\sqrt{N}r^{k-1}}}\mathbb{E}_0\left[e^{\frac{(k(t-1)+1)}{\frac{1}{N}+\sqrt{N}r^{k-1}}\lambda z_n(kt)}\right]\\&\overset{(a)}{=}e^{-\frac{(k(t-1)+1)\eta\lambda}{\frac{1}{N}+\sqrt{N}r^{k-1}}}\times\\&\mathbb{E}_0\left[\exp\left(\frac{\lambda}{\frac{1}{N}+\sqrt{N}r^{k-1}}\sum_{i=0}^{k(t-1)}\sum_{j=1}^{N}\phi_{n,j}(k-1)\left(\frac{\gamma_j^\top(i)\mathbf{\Sigma}_j^{-1}\gamma_j(i)}{2}\right.\right.\right.\\&\left.\left.\left.-\frac{\left(\gamma_j(i)-\mathbf{H}_j\theta_j(t-1)\right)^\top\mathbf{\Sigma}_j^{-1}\left(\gamma_j(i)-\mathbf{H}_j\theta_j(t-1)\right)}{2}\right)\right)\right]\\&\overset{(b)}{\leq}e^{-\frac{(k(t-1)+1)\eta\lambda}{\frac{1}{N}+\sqrt{N}r^{k-1}}}\mathbb{E}_0\left[\exp\left(\frac{\lambda}{\frac{1}{N}+\sqrt{N}r^{k-1}}\sum_{j=1}^{N}\phi_{n,j}(k-1)\right.\right.\end{aligned}$$

$$\times \sum_{i=0}^{k(t-1)} \frac{\gamma_j^\top(i)\mathbf{\Sigma}_j^{-1}\gamma_j(i)}{2}\Bigg)\Bigg]$$
$$\overset{(c)}{\leq} e^{-\frac{(k(t-1)+1)\eta\lambda}{\frac{1}{N}+\sqrt{N}r^{k-1}}}\mathbb{E}_0\left[\exp\left(\lambda\sum_{j=1}^{N}\sum_{i=0}^{k(t-1)}\frac{\gamma_j^\top(i)\mathbf{\Sigma}_j^{-1}\gamma_j(i)}{2}\right)\right]$$
$$\overset{(d)}{=} e^{-\frac{(k(t-1)+1)\eta\lambda}{\frac{1}{N}+\sqrt{N}r^{k-1}}}\prod_{j=1}^{N}\prod_{i=0}^{k(t-1)}\mathbb{E}_0\left[\exp\left(\frac{\lambda\gamma_j^\top(i)\mathbf{\Sigma}_j^{-1}\gamma_j(i)}{2}\right)\right]$$
$$\overset{(e)}{=} \exp\left(-\frac{\lambda\eta(k(t-1)+1)}{\frac{1}{N}+\sqrt{N}r^{k-1}} - \frac{(k(t-1)+1)\sum_{n=1}^{N}M_n}{2}\log(1-\lambda)\right), \tag{C.42}$$

where $\phi_{n,j}(k-1)$ denotes the (n,j)-th entry of $\mathbf{W}^{k-1}$ and r denotes $\|\mathbf{W}-\mathbf{J}\|$. It is to be noted that (a) follows due to the fact that under the null hypothesis the observations made at the agents are of the form $\mathbf{y}_n(t)=\gamma_n(t)$, (b) follows due to the fact that the inverse covariances are positive definite and hence the quadratic forms are positive, (c) follows due to $|\phi_{n,j}(k-1)-\frac{1}{N}|\leq\sqrt{N}r^{k-1}$, (d) follows due to the independence of the noise processes over time and space and (e) follows due to the fact that for each i,j the random variable $\gamma_j(i)^\top\Sigma_j^{-1}\gamma_j(i)$ corresponds to a standard chi-squared random variable with M_j degrees of freedom and the associated moment generating functions[1] exists since $\lambda<1$.
Taking limits on both sides, we have,

$$\limsup_{t\to\infty}\frac{1}{kt}\log\left(\mathbb{P}_0\left(z_n(kt)>\eta\right)\right)$$
$$\leq -\frac{\lambda\eta}{\frac{1}{N}+\sqrt{N}r^{k-1}} - \frac{\sum_{n=1}^{N}M_n}{2}\log(1-\lambda), \tag{C.43}$$

which holds for all λ with $0<\lambda<1$. Now, supposing that

$$\eta > \frac{\left(\frac{1}{N}+\sqrt{N}r^{k-1}\right)\sum_{n=1}^{N}M_n}{2}, \tag{C.44}$$

it can be shown that the right-hand side (RHS) of (C.43) is minimized at $\lambda^*=1-\frac{\left(\frac{1}{N}+\sqrt{N}r^{k-1}\right)\sum_{n=1}^{N}M_n}{2\eta}$. It is to be noted that with the condition in (C.44) in force, $\lambda^*\in(0,1)$. Hence, by substituting $\lambda=\lambda^*$ in (C.43) we have,

$$\limsup_{t\to\infty}\frac{1}{kt}\log\left(\mathbb{P}_0\left(z_n(kt)>\eta\right)\right)\leq -\frac{\eta}{\frac{1}{N}+\sqrt{N}r^{k-1}}$$
$$-\frac{\sum_{n=1}^{N}M_n}{2}\left(1+\log\frac{2\eta}{\left(\frac{1}{N}+\sqrt{N}r^{k-1}\right)\sum_{n=1}^{N}M_n}\right). \tag{C.45}$$

We specifically focused on the sub-sequence $\{z_n(kt)\}$ for the derivation of large deviations[2] exponent in this proof. It can be readily seen that other time-shifted sub-sequences (with constant time-shifts upto k units) also inherit a similar large deviations upper bound as by construction, (see (3.24) for example), the decision

[1] The moment generating function $\mathbb{E}\left[\exp(\rho z)\right]$ of a chi-squared random variable z with M_n degrees of freedom exists and is given by $(1-2\rho)^{-\frac{M_n}{2}}$ for all $\rho<1/2$.

[2] By large deviations exponent, we mean the exponent associated with our large deviations upper bound.

statistic $z_n(kt)$ stays constant on the time interval $[kt, kt+k-1]$. Hence, the large deviations upper bound can be extended as a large deviations upper bound for the sequence $\{z_n(t)\}$.

For notational simplicity we denote $\mathbf{1}_N \otimes \theta^*$ as θ^*_N. Before analyzing the probability of miss $\mathbb{P}_{1,\theta^*}(z_n(kt) < \eta)$ and its error exponent, we first analyze the term $\left\|\mathbf{G}_H^\top(\theta(t) - \theta^*_N)\right\|^2$. We have,

$$\left\|\mathbf{G}_H^\top(\theta(t) - \theta^*_N)\right\| \leq \|\mathbf{G}_H\| \|\theta(t) - \theta^*_N\|. \tag{C.46}$$

From (3.20), we have that,

$$\begin{aligned}\theta(t+1) - \theta^*_N = &\underbrace{\left(\mathbf{I}_{NM} - \beta_t(\mathbf{L} \otimes \mathbf{I}_M) - \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right)}_{\mathbf{A}(t)} \\ &\times (\theta(t) - \theta^*_N) + \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \gamma(t).\end{aligned} \tag{C.47}$$

Let

$$\gamma_G(t) = \mathbf{G}_H \mathbf{\Sigma}^{-1} \gamma(t). \tag{C.48}$$

Then, we have,

$$\begin{aligned}&\|\theta(t) - \theta^*_N\|^2 = (\theta(t) - \theta^*_N)^\top (\theta(t) - \theta^*_N) \\ &= \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} \alpha_i \alpha_j \gamma_G(i)^\top \alpha_i \alpha_j \prod_{u=0}^{t-2-i} \mathbf{A}(t-1-u) \prod_{v=j+1}^{t-1} \mathbf{A}(v) \gamma_G(j) \\ &= \gamma_{G,t}^\top \mathbf{P}_t \gamma_{G,t} = \mathrm{tr}\left(\mathbf{P}_t \gamma_{G,t} \gamma_{G,t}^\top\right),\end{aligned} \tag{C.49}$$

where

$$\gamma_{G,t} = [\gamma_G^\top(0)\ \gamma_G^\top(1)\ \cdots\ \gamma_G^\top(t-1)]^\top \tag{C.50}$$

and $\mathbf{P}_t$ is a block matrix of dimension $NMt \times NMt$, whose (i,j)-th block $i,j = 0, \cdots, t-1$ is given as follows:

$$[\mathbf{P}_t]_{ij} = \alpha_i \alpha_j \prod_{u=0}^{t-2-i} \mathbf{A}(t-1-u) \prod_{v=j+1}^{t-1} \mathbf{A}(v). \tag{C.51}$$

First, note that the $\mathbf{A}(i)$'s commute and are symmetric and hence the individual blocks $[\mathbf{P}_t]_{ij}$-s and $\mathbf{P}_t$ is symmetric. We also note that, $\mathbf{P}_t$ is positive semi-definite, as using an expansion similar to (C.49) it can be shown that any quadratic form of $\mathbf{P}_t$ is non-negative.

Before, characterizing the large deviation exponents, we state the following lemma, the proof of which is provided in Appendix C.4.

Lemma C.2.2. *Let Assumptions 3.6.1-3.6.3 and 3.6.5 hold. Given, the block matrix $\mathbf{P}_t$ as defined in (C.51), we have the following upper bound,*

$$t\|\mathbf{P}_t\| \leq c_3 \frac{(t_1+1)^{2c_1\alpha_0}}{t^{2c_1\alpha_0 - 1}} + \frac{\alpha_0^2}{t} + \frac{\alpha_0^2}{2c_1\alpha_0 - 1}, \quad \forall t \geq t_1, \tag{C.52}$$

where t_1 *is as defined in* (C.38)-(C.40) *and* $c_3 = \sum_{v=0}^{t_1-1} \alpha_v^2 \prod_{u=v+1}^{t_1-1} \|\mathbf{A}(u)\|$.

For $\mathcal{H}_1$, we have,

$$\begin{aligned} z_n(kt) &= \frac{1}{(k(t-1)+1)} \sum_{j=1}^{N} \phi_{n,j}(k-1) \\ &\times \sum_{i=0}^{k(t-1)} \theta_j^\top (k(t-1)) \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \gamma_j(i) \\ &- \frac{\left(\mathbf{H}_j \left(\theta_j(k(t-1)) - \theta^*\right)\right)^\top \mathbf{\Sigma}_j^{-1} \left(\mathbf{H}_j \left(\theta_j(k(t-1)) - \theta^*\right)\right)}{2} \\ &+ \frac{(\theta^*)^\top \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \mathbf{H}_j \theta^*}{2}. \end{aligned} \tag{C.53}$$

For notational simplicity, we denote,

$$\eta_2 = \frac{-2N\eta + (\theta^*)^\top \mathbf{G}\theta^* \left(1 - N\sqrt{N} r^{k-1}\right)}{4 \left\|\mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right\| \left(1 + N\sqrt{N} r^{k-1}\right)}. \tag{C.54}$$

Moreover, supposing that

$$\eta < \frac{(\theta^*)^\top \mathbf{G}\theta^* \left(1 - N\sqrt{N} r^{k-1}\right)}{2N}, \tag{C.55}$$

we have $\eta_2 > 0$, and the probability of miss can be characterized as follows:

$$\begin{aligned} &\mathbb{P}_{1,\theta^*}\left(z_n(kt) < \eta\right) \\ &= \mathbb{P}_{1,\theta^*}\Bigg(\frac{1}{(k(t-1)+1)} \sum_{j=1}^{N} \phi_{n,j}(k-1) \\ &\times \sum_{i=0}^{k(t-1)} \theta_j^\top (k(t-1)) \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \gamma_j(i) \\ &- \frac{\left(\mathbf{H}_j \left(\theta_j(k(t-1)) - \theta^*\right)\right)^\top \mathbf{\Sigma}_j^{-1} \left(\mathbf{H}_j \left(\theta_j(k(t-1)) - \theta^*\right)\right)}{2} \\ &+ \frac{(\theta^*)^\top \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \mathbf{H}_j \theta^*}{2} < \eta \Bigg) \\ &\overset{(a)}{\leq} \mathbb{P}_{1,\theta^*}\Bigg(\frac{1}{(k(t-1)+1)} \sum_{j=1}^{N} \phi_{n,j}(k-1) \\ &\times \sum_{i=0}^{k(t-1)} \theta_j^\top (k(t-1)) \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \gamma_j(i) \\ &- \frac{\left(\mathbf{H}_j \left(\theta_j(k(t-1)) - \theta^*\right)\right)^\top \mathbf{\Sigma}_j^{-1} \left(\mathbf{H}_j \left(\theta_j(k(t-1)) - \theta^*\right)\right)}{2} \\ &< \eta - \frac{(\theta^*)^\top \mathbf{G}\theta^* \left(\frac{1}{N} - \sqrt{N} r^{k-1}\right)}{2} \Bigg) \end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} \mathbb{P}_{1,\theta^*}\left(\sum_{j=1}^{N}\phi_{n,j}(k-1)\right.\\
&\times\frac{\left(\mathbf{H}_j\left(\theta_j(k(t-1))-\theta^*\right)\right)^\top\mathbf{\Sigma}_j^{-1}\left(\mathbf{H}_j\left(\theta_j(k(t-1))-\theta^*\right)\right)}{2}\\
&\left.>-\frac{\eta}{2}+\frac{(\theta^*)^\top\mathbf{G}\theta^*\left(\frac{1}{N}-\sqrt{N}r^{k-1}\right)}{4}\right)\\
&+\mathbb{P}_{1,\theta^*}\left(\frac{1}{(k(t-1)+1)}\sum_{j=1}^{N}\phi_{n,j}(k-1)\right.\\
&\times\sum_{i=0}^{k(t-1)}\theta_j^\top(k(t-1))\mathbf{H}_j^\top\mathbf{\Sigma}_j^{-1}\gamma_j(i)\\
&\left.<\frac{\eta}{2}-\frac{(\theta^*)^\top\mathbf{G}\theta^*\left(\frac{1}{N}-\sqrt{N}r^{k-1}\right)}{4}\right)\\
&\overset{(c)}{\leq} t1+t2+t3,
\end{aligned}
\tag{C.56}
$$

where (a) follows from $|\phi_{n,j}(k-1)-\frac{1}{N}|\leq\sqrt{N}r^{k-1}$, (b) follows from the union bound, (c) follows from the union bound and the inequality $|\phi_{n,j}(k-1)-\frac{1}{N}|\leq\sqrt{N}r^{k-1}$ and, $(t1)$, $(t2)$ and $(t3)$ are as defined in (C.136). Note that, Assumption 3.6.4 ensures that $\frac{1}{N}-\sqrt{N}r^{k-1}$ is positive.

$$
\begin{aligned}
t1&=\mathbb{P}_{1,\theta^*}\left(\left\|\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right\|\frac{\|\theta(k(t-1))-\theta_N^*\|^2}{2}>\frac{-2N\eta+(\theta^*)^\top\mathbf{G}\theta^*\left(1-N\sqrt{N}r^{k-1}\right)}{4\left(1+N\sqrt{N}r^{k-1}\right)}\right)\\
t2&=\mathbb{P}_{1,\theta^*}\left(\frac{1}{(k(t-1)+1)}\sum_{j=1}^{N}\phi_{n,j}(k-1)\sum_{i=0}^{k(t-1)}\left(\theta_j(k(t-1))-\theta^*\right)^\top\mathbf{H}_j^\top\mathbf{\Sigma}_j^{-1}\gamma_j(i)<\frac{\eta}{4}-\frac{(\theta^*)^\top\mathbf{G}\theta^*\left(\frac{1}{N}-\sqrt{N}r^{k-1}\right)}{8}\right)\\
t3&=\mathbb{P}_{1,\theta^*}\left(\frac{1}{(k(t-1)+1)}\sum_{j=1}^{N}\phi_{n,j}(k-1)\sum_{i=0}^{k(t-1)}(\theta^*)^\top\mathbf{H}_j^\top\mathbf{\Sigma}_j^{-1}\gamma_j(i)<\frac{\eta}{4}-\frac{(\theta^*)^\top\mathbf{G}\theta^*\left(\frac{1}{N}-\sqrt{N}r^{k-1}\right)}{8}\right),
\end{aligned}
\tag{C.136}
$$

First, we analyze the term (t1) in (C.56). We first note that, if λ is chosen to be $\lambda\leq c_4$, where

$$
c_4=\frac{1}{\left\|\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right\|\left(c_3\frac{(t_1+1)^{2c_1\alpha_0}}{kt_1^{2c_1\alpha_0-1}}+\frac{\alpha_0^2}{kt_1}+\frac{\alpha_0^2}{2c_1\alpha_0-1}\right)},
\tag{C.137}
$$

we have that $kt\lambda\left\|\mathbf{P}_{kt}\left(\mathbf{I}_{kt}\otimes\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right)\right\|<1$. Hence, we finally have that $\forall t\geq t_1$, with t_1 as defined in (C.38)

$$
\begin{aligned}
&\det\left(\mathbf{I}_{NMkt}-kt\lambda\mathbf{P}_{kt}\left(\mathbf{I}_{kt}\otimes\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right)\right)\\
&\geq\left(1-kt\lambda\left\|\mathbf{P}_{kt}\right\|\left\|\mathbf{G}_H\mathbf{\Sigma}^{-1}\mathbf{G}_H^\top\right\|\right)^{NMkt},
\end{aligned}
\tag{C.138}
$$

which ensures the existence of the moment generating function of the Wishart distribution under consider-

ation (to be specified shortly). We have,

$$\begin{aligned}
&\mathbb{P}_{1,\theta^*}\left(\frac{\|\theta(k(t-1))-\theta_N^*\|^2}{2}>\eta_2\right)\\
&\leq e^{-\lambda\eta_2 kt}\mathbb{E}_{1,\theta^*}\left[\exp\left(kt\lambda\,\|\theta(k(t-1))-\theta_N^*\|^2\right)\right]\\
&\overset{(a)}{=} e^{-\lambda\eta_2 kt}\mathbb{E}_{1,\theta^*}\left[\exp\left(kt\lambda\,\mathrm{tr}\left(\frac{\mathbf{P}_{kt}}{2}\gamma_{G,kt}\gamma_{G,kt}^\top\right)\right)\right]\\
&\overset{(b)}{=} e^{-\lambda\eta_2 kt}\times\left(\det\left(\mathbf{I}_{NMkt}-kt\lambda\mathbf{P}_{kt}\left(\mathbf{I}_{kt}\otimes\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)\right)\right)^{-1/2},
\end{aligned}\tag{C.139}$$

where in (a), we use the definition of $\mathbf{P}_{kt}$ and $\gamma_{G,kt}$ as defined in (C.51) and (C.50) respectively and in (b) we use the moment generating function of the Wishart distribution (see, for example, Anderson (1946)) as $\gamma_{G,kt}\gamma_{G,kt}^\top$ follows a Wishart distribution. Moreover, from (C.237), we have that,

$$\limsup_{t\to\infty} kt\,\|\mathbf{P}_{kt}\|\leq\frac{\alpha_0^2}{2c_1\alpha_0-1}.\tag{C.140}$$

Now, on using (C.140) and (C.138) in (C.139), we have,

$$\begin{aligned}
&\mathbb{P}_{1,\theta^*}\left(\|\theta(k(t-1))-\theta_N^*\|^2>\eta_2\right)\\
&\leq e^{-\lambda\eta_2 kt}\times\left(\det\left(\mathbf{I}_{NMkt}-kt\lambda\mathbf{P}_{kt}\left(\mathbf{I}_{kt}\otimes\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)\right)\right)^{-1/2}\\
&\leq e^{-\lambda\eta_2 kt}\times\left(1-kt\lambda\,\|\mathbf{P}_{kt}\|\,\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|\right)^{-NMkt/2}\\
&\Rightarrow\frac{1}{kt}\log\left(\mathbb{P}_{1,\theta^*}\left(\|\theta(k(t-1))-\theta_N^*\|^2>\eta_2\right)\right)\\
&\leq-\lambda\eta_2-\frac{NM}{2}\log\left(1-kt\lambda\,\|\mathbf{P}_{kt}\|\,\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|\right)\\
&\Rightarrow\limsup_{t\to\infty}\frac{1}{kt}\log\left(\mathbb{P}_{1,\theta^*}\left(\|\theta(k(t-1))-\theta_N^*\|^2>\eta_2\right)\right)\\
&\leq-\lambda\eta_2-\frac{NM}{2}\log\left(1-\frac{\lambda\alpha_0^2\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|}{2c_1\alpha_0-1}\right).
\end{aligned}\tag{C.141}$$

Let $LD(\lambda)=\lambda\eta_2+\frac{NM}{2}\log\left(1-\frac{\lambda\alpha_0^2\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\|}{2c_1\alpha_0-1}\right)$. We first note that $LD(0)=0$. In order to ensure that the term $(t1)$ decays exponentially, the function $LD(.)$ needs to be increasing in an interval of the form, $[0,c_5]$, where $0<c_4\leq c_5$, with c_4 as defined in (C.137) which is formalized as follows:

$$\lambda<\frac{2c_1\alpha_0-1}{\alpha_0^2\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|}-\frac{NM}{2\eta_2}=c_4^*,\tag{C.142}$$

with η_2 as defined in (C.54). In order to have a positive large deviations upper bound, the RHS of (C.142) needs to be positive and hence, we require,

$$\begin{aligned}
&\frac{2c_1\alpha_0-1}{\alpha_0^2\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right\|}-\frac{NM}{2\eta_2}>0\\
&\Rightarrow\eta<\frac{(\theta^*)^\top\mathbf{G}\theta^*\left(1-N\sqrt{N}r^{k-1}\right)}{2N}
\end{aligned}$$

$$- \frac{M\alpha_0^2 \left\|\mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right\|^2 \left(1 + N\sqrt{N} r^{k-1}\right)}{2c_1\alpha_0 - 1}. \tag{C.143}$$

We note that the condition derived in (C.143) is tighter than (C.55). Now, combining the threshold condition derived above in (C.143) and the one derived in (C.44), we have the following condition on the parameter θ^*

$$\begin{aligned} &\frac{(\theta^*)^\top \mathbf{G}\theta^* \left(1 - N\sqrt{N} r^{k-1}\right)}{2N} \\ &> \frac{M\alpha_0^2 \left\|\mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right\|^2 \left(1 + N\sqrt{N} r^{k-1}\right)}{2c_1\alpha_0 - 1} \\ &+ \frac{\left(\frac{1}{N} + \sqrt{N} r^{k-1}\right) \sum_{n=1}^N M_n}{2} \end{aligned} \tag{C.144}$$

which ensures the exponential decay of the term $(t1)$. Now, when we analyze (t2) and (t3) in (C.56), we note that (t2) involves an additional time-decaying term, i.e., $\theta_j(k(t-1)) - \theta^*$ which contributes to the large deviations upper bound as well. Hence, the exponent which will dominate among (t2) and (t3), would be the exponent of their sum. Using the condition derived in (C.55) and the union bound on (t3), we have,

$$\begin{aligned} &\mathbb{P}_{1,\theta^*} \Bigg(\frac{1}{(k(t-1)+1)} \sum_{j=1}^N \phi_{n,j}(k-1) \sum_{i=0}^{k(t-1)} (\theta^*)^\top \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \gamma_j(i) \\ &< \frac{\eta}{4} - \frac{(\theta^*)^\top \mathbf{G}\theta^* \left(\frac{1}{N} - \sqrt{N} r^{k-1}\right)}{8} \Bigg) \\ &\leq \mathbb{Q} \left(\frac{-\frac{\eta\sqrt{k(t-1)+1}}{4} + \frac{\sqrt{k(t-1)+1}(\theta^*)^\top \mathbf{G}\theta^* \left(\frac{1}{N} - \sqrt{N} r^{k-1}\right)}{8}}{\sqrt{\sum_{j=1}^N (\theta^*)^\top \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \mathbf{H}_j \theta^* \left(\frac{1}{N} + \sqrt{N} r^{k-1}\right)}} \right) \\ &\Rightarrow \limsup_{t\to\infty} \frac{1}{kt} \log \Bigg(\mathbb{P}_{1,\theta^*} \Bigg(\frac{1}{(k(t-1)+1)} \sum_{j=1}^N \phi_{n,j}(k-1) \\ &\times \sum_{i=0}^{k(t-1)} (\theta^*)^\top \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \gamma_j(i) \\ &< \frac{\eta}{4} - \frac{(\theta^*)^\top \mathbf{G}\theta^* \left(\frac{1}{N} - \sqrt{N} r^{k-1}\right)}{8} \Bigg) \Bigg) \\ &\leq - \left(\frac{\left(-\frac{\eta}{4} + \frac{(\theta^*)^\top \mathbf{G}\theta^* \left(\frac{1}{N} - \sqrt{N} r^{k-1}\right)}{8}\right)^2}{2\sum_{j=1}^N (\theta^*)^\top \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \mathbf{H}_j \theta^* \left(\frac{1}{N} + \sqrt{N} r^{k-1}\right)^2} \right). \end{aligned} \tag{C.145}$$

Combining (C.145) and (C.141), we have,

$$\limsup_{t\to\infty} \frac{1}{kt} \log \left(\mathbb{P}_{1,\theta^*} \left(z_n(kt) < \eta \right) \right)$$

$$\leq \max \left\{ -\frac{\left(-\frac{\eta}{4} + \frac{(\theta^*)^\top \mathbf{G}\theta^* \left(\frac{1}{N} - \sqrt{N} r^{k-1} \right)}{8} \right)^2}{2 \sum_{j=1}^{N} (\theta^*)^\top \mathbf{H}_j^\top \mathbf{\Sigma}_j^{-1} \mathbf{H}_j \theta^* \left(\frac{1}{N} + \sqrt{N} r^{k-1} \right)^2} \right.$$
$$\left. , -LD \left(\min \{c_4, c_4^*\} \right) \right\} = LD_1 (\eta) , \tag{C.146}$$

We specifically focused on the sub-sequence $\{z_n(kt)\}$ for the derivation of large deviations[3] exponent in this proof. It can be readily seen that other time-shifted sub-sequences (with constant time-shifts upto k units) also inherit a similar large deviations upper bound as by construction, (see (3.24) for example), the decision statistic $z_n(kt)$ stays constant on the time interval $[kt, kt + k - 1]$. Hence, the large deviations upper bound can be extended as a large deviations upper bound for the sequence $\{z_n(t)\}$. □

C.3 Proofs of Lemmas in Section C.1

Proof of Lemma C.1.1. The proof follows similarly as the proof of Lemma IV.1 in Kar and Moura (2014) with appropriate modifications to take into account the state-dependent nature of the innovation gains. Define the process $\{\mathbf{x}(t)\}$ as $\mathbf{x}(t) = \theta(t) - \mathbf{1}_N \otimes \theta^*$ where θ^* denotes the true but unknown parameter. The process $\{\mathbf{x}(t)\}$ satisfies the following recursion:

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{x}(t) - \beta_t (\mathbf{L} \otimes \mathbf{I}_M) x(t) \\ &+ \alpha_t \mathbf{G}(\theta(t)) \mathbf{\Sigma}^{-1} (\mathbf{y}(t) - \mathbf{h}(\theta(t))) , \end{aligned} \tag{C.147}$$

which implies that,

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{x}(t) - \beta_t (\mathbf{L} \otimes \mathbf{I}_M) x(t) \\ &+ \alpha_t \mathbf{G}(\theta(t)) \mathbf{\Sigma}^{-1} (\mathbf{y}(t) - \mathbf{h}(\theta_N^*)) \\ &- \alpha_t \mathbf{G}(\theta(t)) \mathbf{\Sigma}^{-1} (\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*)) . \end{aligned} \tag{C.148}$$

It follows from basic properties of the Laplacian $\mathbf{L}$, that

$$(\mathbf{L} \otimes \mathbf{I}_M)(\mathbf{1}_N \otimes \theta^*) = (\mathbf{L}\mathbf{1}_N) \otimes (\mathbf{I}_M \theta^*) = \mathbf{0}. \tag{C.149}$$

Taking norms of both sides of (C.147), we have,

$$\begin{aligned} &\|\mathbf{x}(t+1)\|^2 = \|\mathbf{x}(t)\|^2 - 2\beta_t \mathbf{x}^\top(t) (\mathbf{L} \otimes \mathbf{I}_M) \mathbf{x}(t) \\ &- 2\alpha_t \mathbf{x}^\top(t) \mathbf{G}(\theta(t)) \mathbf{\Sigma}^{-1} (\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*)) \\ &+ \beta_t^2 \mathbf{x}^\top(t) (\mathbf{L} \otimes \mathbf{I}_M)^2 \mathbf{x}(t) \\ &+ 2\alpha_t \beta_t \mathbf{x}^\top(t) (\mathbf{L} \otimes \mathbf{I}_M) \mathbf{G}(\theta(t)) \mathbf{\Sigma}^{-1} (\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*)) \\ &- 2\alpha_t \beta_t \mathbf{x}^\top(t) (\mathbf{L} \otimes \mathbf{I}_M) \mathbf{G}(\theta(t)) \mathbf{\Sigma}^{-1} (\mathbf{y}(t) - \mathbf{h}(\theta_N^*)) \\ &+ \alpha_t^2 (\mathbf{y}(t) - \mathbf{h}(\theta_N^*))^\top \mathbf{\Sigma}^{-1} \mathbf{G}^\top(\theta(t)) \\ &\times \mathbf{G}(\theta(t)) \mathbf{\Sigma}^{-1} (\mathbf{y}(t) - \mathbf{h}(\theta_N^*)) \end{aligned}$$

[3] By large deviations exponent, we mean the exponent associated with out large deviations upper bound.

$$
\begin{aligned}
&+\alpha_t^2\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right)^\top\boldsymbol{\Sigma}^{-1}\mathbf{G}^\top\left(\theta(t)\right)\\
&\times\mathbf{G}\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right)\\
&+2\alpha_t\mathbf{x}^\top(t)\mathbf{G}\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}(t)-\mathbf{h}(\theta_N^*)\right)\\
&-2\alpha_t^2\left(\mathbf{y}(t)-\mathbf{h}\left(\theta_N^*\right)\right)^\top\boldsymbol{\Sigma}^{-1}\mathbf{G}^\top\left(\theta(t)\right)\\
&\times\mathbf{G}\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right).
\end{aligned}
\tag{C.150}
$$

Consider the orthogonal decomposition,

$$
\mathbf{x}=\mathbf{x}_c+\mathbf{x}_{c\perp},
\tag{C.151}
$$

where $\mathbf{x}_c$ denotes the projection of $\mathbf{x}$ to the consensus subspace $\mathcal{C}$ with

$$
\mathcal{C}=\{\mathbf{x}\in\mathbb{R}^{MN}\mid\mathbf{x}=1_N\otimes a,\textit{for some }\ a\in\mathbb{R}^M\}.
\tag{C.152}
$$

From, (3.1), we have that,

$$
\mathbb{E}_{\theta^*}\left[\mathbf{y}(t)-\mathbf{h}\left(\theta_N^*\right)\right]=\mathbf{0}.
\tag{C.153}
$$

Consider the process

$$
V_2(t)=\left\|\mathbf{x}(t)\right\|^2.
\tag{C.154}
$$

Using conditional independence properties we have,

$$
\begin{aligned}
&\mathbb{E}_{\theta^*}[V_2(t+1)|\mathcal{F}_t]=V_2(t)+\beta_t^2\mathbf{x}^\top(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)^2\mathbf{x}(t)\\
&+\alpha_t^2\mathbb{E}_{\theta^*}\Big[\left(\mathbf{y}(t)-\mathbf{h}\left(\theta_N^*\right)\right)^\top\boldsymbol{\Sigma}^{-1}\mathbf{G}^\top\left(\theta(t)\right)\\
&\times\mathbf{G}\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}(t)-\mathbf{h}\left(\theta_N^*\right)\right)\Big]-2\beta_t\mathbf{x}^\top(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)\mathbf{x}(t)\\
&-2\alpha_t\mathbf{x}^\top(t)\mathbf{G}\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right)\\
&+2\alpha_t\beta_t\mathbf{x}^\top(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)\mathbf{G}\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right)\\
&+\alpha_t^2\left\|\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right)^\top\mathbf{G}^\top\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\right\|^2.
\end{aligned}
\tag{C.155}
$$

We use the following inequalities $\forall t\geq t_1$,

$$
\begin{aligned}
&\mathbf{x}^\top(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)^2\mathbf{x}(t)\overset{(q1)}{\leq}\lambda_N^2(\mathbf{L})||\mathbf{x}_{C\perp}(t)||^2;\\
&\mathbf{x}^\top(t)\mathbf{G}\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right)\geq c_1||\mathbf{x}(t)||^2\overset{(q2)}{\geq}0;\\
&\mathbf{x}^\top(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)\mathbf{x}(t)\overset{(q3)}{\geq}\lambda_2(\mathbf{L})\left\|\mathbf{x}_{C\perp}(t)\right\|^2;\\
&\mathbf{x}^\top(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)\mathbf{G}\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right)\overset{(q4)}{\leq}c_2\left\|\mathbf{x}(t)\right\|^2,
\end{aligned}
\tag{C.156}
$$

for c_1 as defined in Assumption 2.4.2, and a positive constant c_2, where $(q2)$ follows from Assumption 2.4.2 and $(q4)$ follows from Assumption 2.4.1 by which we have that $\|\nabla\mathbf{h}_n\left(\theta_n(t)\right)\|$ is uniformly bounded from above by k_n for all n, and hence, we have that $\|\mathbf{G}\left(\theta(t)\right)\|\leq\max_{n=1,\cdots,N}k_n$. We also have

$$\mathbb{E}_{\theta^*}\Big[\left(\mathbf{y}(t)-\mathbf{h}\left(\theta_N^*\right)\right)^\top \mathbf{\Sigma}^{-1}\mathbf{G}^\top(\theta(t)) \\ \times\mathbf{G}\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\left(\mathbf{y}(t)-\mathbf{h}\left(\theta_N^*\right)\right)\Big] \leq c_4, \tag{C.157}$$

for some constant $c_4 > 0$. In (C.157), we use the fact that the noise process under consideration is Gaussian and hence has finite moments. We also use the fact that $\|\mathbf{G}\left(\theta(t)\right)\| \leq \max_{n=1,\cdots,N} k_n$, which in turn follows from Assumption 2.4.1.
We further have that,

$$\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right)^\top \mathbf{\Sigma}^{-1}\mathbf{G}^\top\left(\theta(t)\right) \\ \mathbf{G}\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\left(\mathbf{h}\left(\theta(t)\right)-\mathbf{h}\left(\theta_N^*\right)\right) \leq c_3\left\|\mathbf{x}(t)\right\|^2, \tag{C.158}$$

where $c_3 > 0$ is a constant. It is to be noted that (C.158) follows from the Lipschitz continuity in Assumption 2.4.1 and the fact that $\|\mathbf{G}\left(\theta(t)\right)\| \leq \max_{n=1,\cdots,N} k_n$.
Using (C.155)-(C.158), we have,

$$\mathbb{E}_{\theta^*}[V_2(t+1)|\mathcal{F}_t] \leq \left(1+c_5\left(\alpha_t\beta_t+\alpha_t^2\right)\right)V_2(t) \\ -c_6(\beta_t-\beta_t^2)||\mathbf{x}_{C\perp}(t)||^2+c_4\alpha_t^2, \tag{C.159}$$

for some positive constants c_5 and c_6. As β_t^2 goes to zero faster than β_t, $\exists t_2$ such that $\forall t \geq t_2$, $\beta_t \geq \beta_t^2$. Hence $\exists t_2$ and $\exists \tau_1, \tau_2 > 1$ such that for all $t \geq t_2$

$$c_5\left(\alpha_t\beta_t+\alpha_t^2\right) \leq \frac{c_7}{(t+1)^{\tau_1}} = \gamma_t,\ c_4\alpha_t^2 \leq \frac{c_8}{(t+1)^{\tau_2}} = \hat{\gamma}_t \tag{C.160}$$

where $c_7, c_8 > 0$ are constants.
By the above construction we obtain, $\forall t \geq t_2$,

$$\mathbb{E}_{\theta^*}[V_2(t+1)|\mathcal{F}_t] \leq (1+\gamma_t)V_2(t)+\hat{\gamma}_t, \tag{C.161}$$

where the positive weight sequences $\{\gamma_t\}$ and $\{\hat{\gamma}_t\}$ are summable, i.e.,

$$\sum_{t\geq 0}\gamma_t < \infty,\ \sum_{t\geq 0}\hat{\gamma}_t < \infty. \tag{C.162}$$

By (C.162), the product $\prod_{s=t}^{\infty}(1+\gamma_s)$ exists for all t. Now let $\{W(t)\}$ be such that

$$W(t) = \left(\prod_{s=t}^{\infty}(1+\gamma_s)\right)V_2(t)+\sum_{s=t}^{\infty}\hat{\gamma}_s, \quad \forall t \geq t_2. \tag{C.163}$$

By (C.163), it can be shown that $\{W(t)\}$ satisfies,

$$\mathbb{E}_{\theta^*}[W(t+1)|\mathcal{F}_t] \leq W(t). \tag{C.164}$$

Hence, $\{W(t)\}$ is a non-negative super martingale and converges a.s. to a bounded random variable W^* as $t \to \infty$. It then follows from (C.163) that $V_2(t) \to W^*$ as $t \to \infty$. Thus, we conclude that the sequences

$\{\theta_n(t)\}$ are bounded for all n.

□

Proof of Lemma C.1.2. The proof follows exactly the development in theorem IV.1 of Kar and Moura (2014).
Let $\mathbf{x}(t)$ denote the residual $\theta(t) - \mathbf{1}_N \otimes \theta^*$.
For $\epsilon \in (0,1)$, define the set Γ_ϵ

$$\Gamma_\epsilon = \left\{ \theta \in \mathbb{R}^{NM} : \epsilon \leq \|\theta - \mathbf{1}_N \otimes \theta^*\| \leq \frac{1}{\epsilon} \right\}. \tag{C.165}$$

Let ρ_ϵ denote the $\{\mathcal{F}_t\}$ stopping time

$$\rho_\epsilon = \inf\{t \geq 0 : \theta(t) \notin \Gamma_\epsilon\}, \tag{C.166}$$

where Γ_ϵ is defined in (C.165). Let $\{V^\epsilon(t)\}$ denote the stopped process

$$V^\epsilon(t) = V_2(\max\{t, \rho_\epsilon\}), \forall t, \tag{C.167}$$

with $V_2(t)$ as defined in (C.154).
Then, we have,

$$V^\epsilon(t+1) = V_2(t+1)\mathbb{I}\left(\rho_\epsilon > t\right) + V_2(\rho_\epsilon)\mathbb{I}\left(\rho_\epsilon \leq t\right), \tag{C.168}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Due to the fact that $\mathbb{I}\left(\rho_\epsilon > t\right)$ and $V_2(\rho_\epsilon)\mathbb{I}\left(\rho_\epsilon \leq t\right)$ are adapted to $\mathcal{F}_t$ for all t, we have,

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[V^\epsilon(t+1)|\mathcal{F}_t\right] &= \mathbb{E}_{\theta^*}\left[V_2(t+1)\right]\mathbb{I}\left(\rho_\epsilon > t\right) \\ &+ V_2(\rho_\epsilon)\mathbb{I}\left(\rho_\epsilon \leq t\right),\end{aligned} \tag{C.169}$$

for all t.
First, noting the inequality derived in (C.156) in $(q2)$ and rewriting it as,

$$-\mathbf{x}(t)^T \mathbf{G}\left(\theta(t)\right) \mathbf{\Sigma}^{-1} \left(\mathbf{h}\left(\theta(t)\right) - \mathbf{h}\left(\theta_N^*\right)\right) \leq -c_1 ||\mathbf{x}(t)||^2, \tag{C.170}$$

we have with a slight rearrangement of terms from the expansion in (C.155),

$$\begin{aligned}\mathbb{E}_{\theta^*}[V_2(t+1)|\mathcal{F}_t] &= V_2(t) + \beta_t^2 \mathbf{x}^T(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right)^2 \mathbf{x}(t) \\ &+ \alpha_t^2 \mathbb{E}_{\theta^*}\Big[\left(\mathbf{y}(t) - \mathbf{h}\left(\theta_N^*\right)\right)^\top \mathbf{\Sigma}^{-1} \mathbf{G}^\top\left(\theta(t)\right) \\ &\times \mathbf{G}\left(\theta(t)\right) \mathbf{\Sigma}^{-1} \left(\mathbf{y}(t) - \mathbf{h}\left(\theta_N^*\right)\right)\Big] - 2\beta_t \mathbf{x}^\top(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{x}(t) \\ &- 2\alpha_t \mathbf{x}^\top(t) \mathbf{G}\left(\theta(t)\right) \mathbf{\Sigma}^{-1} \left(\mathbf{h}\left(\theta(t)\right) - \mathbf{h}\left(\theta_N^*\right)\right) \\ &+ 2\alpha_t \beta_t \mathbf{x}^\top(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{G}\left(\theta(t)\right) \mathbf{\Sigma}^{-1} \left(\mathbf{h}\left(\theta(t)\right) - \mathbf{h}\left(\theta_N^*\right)\right) \\ &+ \alpha_t^2 \left\| \left(\mathbf{h}\left(\theta(t)\right) - \mathbf{h}^\top\left(\theta_N^*\right)\right) \mathbf{G}^\top\left(\theta(t)\right) \mathbf{\Sigma}^{-1} \right\|^2.\end{aligned} \tag{C.171}$$

Now, using (C.170) in (C.171) and the inequalities derived in (C.156)-(C.158), we have,

$$\begin{aligned}\mathbb{E}_{\theta^*}[V_2(t+1)|\mathcal{F}_t] &\leq \left(1 - c_1\alpha_t + c_5\left(\alpha_t\beta_t + \alpha_t^2\right)\right) V_2(t) \\ &- c_6(\beta_t - \beta_t^2)||\mathbf{x}_{C\perp}(t)||^2 + c_4\alpha_t^2, \end{aligned} \tag{C.172}$$

where c_5, c_6, c_4 are appropriately chosen constants.
Now, by choosing a large enough t_ϵ, such that for all $t \geq t_\epsilon$, we can assert that,

$$\begin{aligned}&\beta_t - \beta_t^2 \geq 0, \\ &c_1\alpha_t - c_5\left(\alpha_t\beta_t + \alpha_t^2\right) \geq c_7\alpha_t. \end{aligned} \tag{C.173}$$

Thus, we have for $t \geq t_\epsilon$,

$$\mathbb{E}_{\theta^*}[V_2(t+1)|\mathcal{F}_t] \leq (1 - c_1\alpha_t)\, V_2(t) + c_4\alpha_t^2. \tag{C.174}$$

Furthermore, by the definition of Γ_ϵ, we have,

$$\|\mathbf{x}(t)\|^2 \geq \epsilon^2 \text{ on } \{\mathbf{x}(t) \in \Gamma_\epsilon\}, \tag{C.175}$$

and hence by the definition of $V_2(t)$, we have that there exists a constant $c_7(\epsilon) > 0$ such that

$$V_2(t) \geq c_7(\epsilon) \text{ on } \{\mathbf{x}(t) \in \Gamma_\epsilon\}. \tag{C.176}$$

Using the above relation in (C.174), we then have for all $t \geq t_\epsilon$,

$$\begin{aligned}&\mathbb{E}_{\theta^*}[V_2(t+1)|\mathcal{F}_t]\mathbb{I}(\rho_\epsilon > t) \\ &\leq \left[V_2(t) - c_8(\epsilon)\alpha_t + c_4\alpha_t^2\right]\mathbb{I}(\rho_\epsilon > t), \end{aligned} \tag{C.177}$$

where $c_8(\epsilon) > 0$ is an appropriately chosen constant. Finally, the observation that $\alpha_t > \alpha_t^2$ establishes that

$$\mathbb{E}_{\theta^*}[V_2(t+1)|\mathcal{F}_t]\mathbb{I}(\rho_\epsilon > t) \leq [V_2(t) - c_9(\epsilon)\alpha_t]\,\mathbb{I}(\rho_\epsilon > t), \tag{C.178}$$

where $c_9(\epsilon) > 0$ is an appropriately chosen constant Finally, from (C.169), we have that

$$\begin{aligned}&\mathbb{E}_{\theta^*}[V^\epsilon(t+1)|\mathcal{F}_t] \leq V_2(t)\mathbb{I}(\rho_\epsilon > t) + V_2)(\rho_\epsilon)\mathbb{I}(\rho_\epsilon \leq t) \\ &- c_9(\epsilon)\alpha_t\mathbb{I}(\rho_\epsilon > t) \\ &= V^\epsilon(t) - c_9(\epsilon)\alpha_t\mathbb{I}(\rho_\epsilon > t). \end{aligned} \tag{C.179}$$

It is to be noted that $\{V^\epsilon(t)\}_{t\geq t_\epsilon}$ satisfies $\mathbb{E}_{\theta^*}[V^\epsilon(t+1)|\mathcal{F}_t] \leq V^\epsilon(t)$ for all $t \geq t_\epsilon$, which being a non-negative supermartingale, there exists an a.s. finite V^ϵ such that $V^\epsilon(t+1) \to V^\epsilon$ a.s. as $t \to \infty$. To this end, define the process $\{V_1^\epsilon(t)\}$ given by

$$V_1^\epsilon(t) = V^\epsilon(t) + c_9(\epsilon)\sum_{s=0}^{t-1}\alpha_s\mathbb{I}(\rho_\epsilon > s), \tag{C.180}$$

and by (C.179) we have that

$$\mathbb{E}_{\theta^*}[V_1^\epsilon(t+1)|\mathcal{F}_t] \leq V^\epsilon(t) - c_9(\epsilon)\alpha_t \mathbb{I}(\rho_\epsilon > t) + c_9(\epsilon)\sum_{s=0}^{t-1}\alpha_s \mathbb{I}(\rho_\epsilon > s) = V_1^\epsilon(t), \tag{C.181}$$

for all $t \geq t_\epsilon$. Hence, we have that $\{V_1^\epsilon(t)\}_{t\geq t_\epsilon}$ is a non-negative supermartingale and there exists a finite random variable V_1^ϵ such that $V_1^\epsilon(t) \to V_1^\epsilon$ a.s. as $t \to \infty$. From the definition in (C.180), we have that the following limit exists:

$$\lim_{t\to\infty} c_9(\epsilon)\sum_{s=0}^{t-1}\alpha_s \mathbb{I}(\rho_\epsilon > s) = V_1^\epsilon - V^\epsilon < \infty \text{ a.s.} \tag{C.182}$$

We also have that as $t \to \infty$, $\sum_{s=0}^{t-1}\alpha_s \to \infty$, the limit condition in (C.182) is satisfied only if $\rho_\epsilon < \infty$ a.s. Let's define the sequence $\{\mathbf{x}(\rho_{1/p})\}$, by choosing $\epsilon = 1/p$, for each positive integer $p > 1$. By definition, we have,

$$\left\|\mathbf{x}(\rho_{1/p})\right\| \in [1, 1/p) \cup (p, \infty) \text{ a.s.} \tag{C.183}$$

We also have from Lemma C.1.1 that

$$\mathbb{P}_{\theta^*}\left(\left\|\mathbf{x}(\rho_{1/p})\right\| > p \text{ i.o.}\right) = 0, \tag{C.184}$$

where i.o. denotes infinitely often as $p \to \infty$. Hence, by (C.183) we have that there exists a finite integer valued random variable p^* such that $\left\|\mathbf{x}(\rho_{1/p})\right\| < 1/p^*$, $\forall p \geq p^*$, which in turn implies that $\left\|\mathbf{x}(\rho_{1/p})\right\| \to 0$ as $p \to \infty$. Finally, we have that

$$\mathbb{P}_{\theta^*}\left(\liminf_{t\to\infty}\left\|\mathbf{x}(\rho_{1/p})\right\| = 0\right) = 1. \tag{C.185}$$

With the above development in place we have from (C.154) that $\liminf_{t\to\infty} V_2(t) = 0$ a.s. Noting that the limit of $\{V_2(t)\}$ exists, we have that $V_2(t) \to 0$ as $t \to \infty$ a.s. and again from (C.154), we have that $\mathbf{x}(t) \to 0$ as $t \to \infty$ a.s. □

Proof of Lemma C.1.6. Define, the process $\{\hat{z}_{\text{avg}}(t)\}$ as follows :

$$\hat{z}_{\text{avg}}(t) = z_{\text{avg}}(t) - \frac{\mathbf{h}^\top(\theta_N^*)\,\mathbf{\Sigma}^{-1}\mathbf{h}(\theta_N^*)}{2N}. \tag{C.186}$$

The recursion for $\{\hat{z}_{\text{avg}}(t)\}$ can then be represented as

$$\begin{aligned}\hat{z}_{\text{avg}}(t+1) &= \left(1 - \frac{1}{t+1}\right)\hat{z}_{\text{avg}}(t) \\ &+ \frac{1}{N(t+1)}\sum_{n=1}^{N}\mathbf{h}_n^\top(\theta_n(t))\,\mathbf{\Sigma}_n^{-1}\left(\mathbf{y}_n(t) - \mathbf{h}_n(\theta^*)\right) \\ &- \frac{1}{2N(t+1)}\sum_{n=1}^{N}\left(\mathbf{h}_n(\theta_n(t)) - \mathbf{h}_n(\theta^*)\right)^\top\mathbf{\Sigma}_n^{-1}\left(\mathbf{h}_n(\theta_n(t)) - \mathbf{h}_n(\theta^*)\right)\end{aligned}$$

$$= \left(1 - \frac{1}{t+1}\right) \hat{z}_{\text{avg}}(t) + \frac{1}{N(t+1)} \mathbf{h}^\top(\theta(t)) \boldsymbol{\Sigma}^{-1} (\mathbf{y}(t) - \mathbf{h}(\theta_N^*))$$
$$- \frac{1}{2N(t+1)} (\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*))^\top \boldsymbol{\Sigma}^{-1} (\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*)). \tag{C.187}$$

In order to apply Lemma C.1.5 to the process $\{\hat{z}_{\text{avg}}(t)\}$, define

$$\begin{aligned}
\boldsymbol{\Gamma}_t &= \mathbf{I}, \\
\boldsymbol{\Phi}_t &= \frac{1}{N} \mathbf{h}^\top(\theta(t)) \boldsymbol{\Sigma}^{-1}, \\
\mathbf{V}_t &= \mathbf{y}(t) - \mathbf{h}(\theta_N^*), \\
\mathbf{T}_t &= \sqrt{t+1} (\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*))^\top \boldsymbol{\Sigma}^{-1} (\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*)).
\end{aligned} \tag{C.188}$$

From Assumption 2.4.1, we have that,

$$\|\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*)\| \leq k_{\max} \|\theta(t) - \theta^*\|, \tag{C.189}$$

where $k_{\max} = \max_{n=1,\cdots,N} k_n$, with the k_n's defined in Assumption 2.4.1. Moreover, from theorem 3.7.1 we have that, with $\tau = 1/4$,

$$\lim_{t\to\infty} \sqrt{t+1} \|\theta(t) - \theta_N^*\|^2 = 0 \text{ a.s.} \tag{C.190}$$

The above implies that

$$\begin{aligned}
&\lim_{t\to\infty} \sqrt{t+1} (\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*))^\top \boldsymbol{\Sigma}^{-1} (\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*)) \\
&\leq \lim_{t\to\infty} \sqrt{t+1} \|\mathbf{h}(\theta(t)) - \mathbf{h}(\theta_N^*)\|^2 \left\|\boldsymbol{\Sigma}^{-1}\right\| = 0.
\end{aligned} \tag{C.191}$$

From Theorem 3.7.1, we have $\boldsymbol{\Phi}_t = \frac{1}{N} \mathbf{h}^\top(\theta(t)) \boldsymbol{\Sigma}^{-1} \to \frac{1}{N} \mathbf{h}^\top(\theta_N^*) \boldsymbol{\Sigma}^{-1}$ a.s. as $t \to \infty$.
Clearly, $\mathbb{E}_{\theta^*}[\mathbf{V}_t | \mathcal{F}_t] = 0$ and $\mathbb{E}_{\theta^*}\left[\mathbf{V}_t \mathbf{V}_t^\top | \mathcal{F}_t\right] = \boldsymbol{\Sigma}$. Due to the i.i.d nature of the noise process, the required uniform integrability condition for the process $\{\mathbf{V}_t\}$ is also verified. Hence, $\{z_{\text{avg}}(t)\}$ falls under the purview of Lemma C.1.5 and the assertion follows. □

Proof of Lemma C.1.7. Define the process $\{\mathbf{p}(t)\}$ as follows:

$$\mathbf{p}(t) = \mathbf{z}(t) - \mathbf{1}_N \otimes z_{\text{avg}}(t). \tag{C.192}$$

Then $\{\mathbf{p}(t)\}$ evolves as

$$\begin{aligned}
\mathbf{p}(t+1) &= \frac{t}{t+1}(\mathbf{W} - \mathbf{J})\mathbf{p}(t) \\
&+ \frac{1}{t+1} \left(\mathbf{h}^*(\theta(t)) - \frac{\mathbf{1}\mathbf{1}^\top}{N} \mathbf{h}^\top(\theta(t))\right) J(\mathbf{y}(t)) \\
&- \frac{1}{2(t+1)} \left(\mathbf{h}^*(\theta(t)) \boldsymbol{\Sigma}^{-1} \mathbf{h}(\theta(t)) \right. \\
&\left. - \frac{\mathbf{1}_N}{N} \otimes \left(\mathbf{h}^\top(\theta(t)) \boldsymbol{\Sigma}^{-1} \mathbf{h}(\theta(t))\right)\right),
\end{aligned} \tag{C.193}$$

where $J(\mathbf{y}(t)) = \mathbf{\Sigma}^{-1}\mathbf{y}(t)$. The following lemmas are instrumental for the subsequent analysis. Lemma C.3.1 is concerned with a stochastic approximation type result which will be used later in the proof, whereas, Lemma C.3.2 establishes the a.s. boundedness of $J(\mathbf{y}(t))$.

Lemma C.3.1 (Kar (2010))**.** *Consider the scalar time-varying linear system*

$$u(t+1) = (1 - r_1(t))u(t) + r_2(t), \tag{C.194}$$

where $\{r_1(t)\}$ is a sequence, such that, $0 \leq r_1(t) \leq 1$ and is given by

$$r_1(t) = \frac{a_1}{(t+1)^{\delta_1}} \tag{C.195}$$

with $a_1 > 0, 0 \leq \delta_1 \leq 1$, whereas the sequence $\{r_2(t)\}$ is given by

$$r_2(t) = \frac{a_2}{(t+1)^{\delta_2}} \tag{C.196}$$

with $a_2 > 0, \delta_2 \geq 0$. Then, if $u(0) \geq 0$ and $\delta_1 < \delta_2$, we have

$$\lim_{t\to\infty} (t+1)^{\delta_0} u(t) = 0, \tag{C.197}$$

for all $0 \leq \delta_0 < \delta_2 - \delta_1$.

Proof. A proof of this Lemma can be found in Kar (2010) in the proof of Lemma 3.3.3 in Chapter 3. □

Lemma C.3.2. *Define $J(\mathbf{y}(t))$ as follows:*

$$J(\mathbf{y}(t)) = \mathbf{\Sigma}^{-1}\mathbf{y}(t) \tag{C.198}$$

Then we have

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty} \frac{1}{(t+1)^{\delta}} ||J(\mathbf{y}(t))|| = 0\right) = 1. \tag{C.199}$$

Proof. Consider any $\epsilon_1 > 0$. By Chebyshev's inequality, we have,

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\frac{1}{(t+1)^{\delta}} \|J(\mathbf{y}(t))\| > \epsilon_1\right) \\
&\leq \frac{1}{\epsilon_1^{1+\frac{1}{\delta}}(t+1)^{1+\delta}} \mathbb{E}_{\theta^*}\left[\|J(\mathbf{y}(t))\|^{1+\frac{1}{\delta}}\right] \\
&= \frac{\mathcal{K}(\theta^*)}{(t+1)^{1+\frac{1}{\delta}}}
\end{aligned} \tag{C.200}$$

where $\mathbb{E}_{\theta^*}[\|J(\mathbf{y}(t))\|^{1+\frac{1}{\delta}}] = \mathcal{K}(\theta^*) < \infty$ because the noise in consideration is Gaussian and has finite moments. Moreover, since $\delta > 0$, the sequence $(t+1)^{1+\frac{1}{\delta}}$ is square summable and we obtain

$$\sum_{t>0} \mathbb{P}_{\theta^*}\left(\frac{1}{(t+1)^{\delta}} \|J(\mathbf{y}(t))\| > \epsilon_1\right) < \infty. \tag{C.201}$$

Hence, we have from the Borel-Cantelli Lemma, for arbitrary $\epsilon_1 > 0$,

$$\mathbb{P}_{\theta^*}\left(\frac{1}{(t+1)^{\delta}}\left\|J(\mathbf{y}(t))\right\| > \epsilon_1 \text{ i.o.}\right) = 0, \tag{C.202}$$

where i.o. stands for infinitely often and the claim follows from standard arguments. □

We also have from Lemma C.1.1 that

$$\mathbb{P}\left(\sup_{t\geq 0}\left\|\left(\mathbf{h}^*\left(\theta(t)\right) - \frac{\mathbf{1}\mathbf{1}^\top}{N}\mathbf{h}^\top\left(\theta(t)\right)\right)\right\| < \infty\right) = 1, \tag{C.203}$$

and combining this with lemma C.3.2, we have,

$$\mathbb{P}\left(\sup_{t\geq 0}\left\|\left(\mathbf{h}^*\left(\theta(t)\right) - \frac{\mathbf{1}\mathbf{1}^\top}{N}\mathbf{h}^\top\left(\theta(t)\right)\right)J(\mathbf{y}(t))\right\| < \infty\right) = 1. \tag{C.204}$$

To prove uniform bounds, we use truncation arguments. For a scalar d, let its truncation $(d)^{A_0}$ be defined at level A_0 by

$$(d)^{A_0} = \begin{cases} \frac{d}{|d|}\min(|d|, A_0), & \text{if } d \neq 0 \\ 0, & \text{if } d = 0, \end{cases} \tag{C.205}$$

while for a vector, the truncation operator is applied component-wise. To this end, we consider sequences $\{\mathbf{p}_{A_0}(t)\}$, which is in turn given by,

$$\begin{aligned}\mathbf{p}_{A_0}(t+1) = {} & \frac{t}{t+1}\left(\mathbf{W}-\mathbf{J}\right)\mathbf{p}_{A_0}(t) + \frac{1}{t+1}\left(J_1\left(\mathbf{y}(t)\right)\right)^{A_0(t+1)^{\delta}} \\ & - \frac{1}{2(t+1)}\Big(\left(\mathbf{h}^*\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\mathbf{h}\left(\theta(t)\right)\right. \\ & \left.-\frac{\mathbf{1}_N}{N}\otimes\left(\mathbf{h}^\top\left(\theta(t)\right)\mathbf{\Sigma}^{-1}\mathbf{h}\left(\theta(t)\right)\right)\Big)\Big)^{A_0},\end{aligned} \tag{C.206}$$

where $J_1\left(\mathbf{y}(t)\right) = \left(\mathbf{h}^*\left(\theta(t)\right) - \frac{\mathbf{1}\mathbf{1}^\top}{N}\mathbf{h}^\top\left(\theta(t)\right)\right)J(\mathbf{y}(t))$, $A_0 > 0$ and $\delta > 0$.
In order to prove the assertion,

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty}(t+1)^{\delta_0}\mathbf{p}(t) = \mathbf{0}\right) = 1, \tag{C.207}$$

it is sufficient to prove that for every $A_0 > 0$,

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty}(t+1)^{\delta_0}\mathbf{p}_{A_0}(t) = \mathbf{0}\right) = 1, \tag{C.208}$$

which is due to the following standard arguments. The pathwise boundedness of the different terms in the recursion for $\mathbf{p}(t)$ as defined in (C.206) implies that, for every $\epsilon > 0$, there exists A_ϵ such that

$$\mathbb{P}_{\theta^*}\left(\sup\left\|J_1(\mathbf{y}(t))\right\| < A_\epsilon(t+1)^{\delta_0}\right) > 1-\epsilon, \tag{C.209}$$

and

$$\mathbb{P}_{\theta^*}\Big(\sup\Big\|\big(\mathbf{h}^*\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta(t)\right) - \frac{\mathbf{1}_N}{N}\otimes\left(\mathbf{h}^\top\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta(t)\right)\right)\big)\Big\| < A_\epsilon\Big) > 1-\epsilon. \tag{C.210}$$

In particular, (C.209) follows from the pathwise boundedness of $\{\theta(t)\}$ proved in Lemma C.1.1, whereas, (C.210) follows from the a.s. convergence in Lemma C.3.2. The processes $\{\mathbf{p}(t)\}$ and $\{\mathbf{p}_{A_\epsilon}(t)\}$ agree on the set where both of the above mentioned events occur. Hence, it follows that,

$$\mathbb{P}_{\theta^*}\left(\sup\|\mathbf{p}(t)-\mathbf{p}_{A_\epsilon}(t)\|=0\right) > 1-2\epsilon. \tag{C.211}$$

Invoking the claim in (C.208), we have,

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty}(t+1)^{\delta_0}\mathbf{p}(t)=\mathbf{0}\right) > 1-2\epsilon. \tag{C.212}$$

The assertion then can be proved by taking ϵ to 0.
In order to establish the claim in (C.208), for every $A_0>0$, consider the scalar process $\{\hat{p}_{A_0}(t)\}_{t\geq 0}$ defined as

$$\hat{p}_{A_0}(t+1) = \|\mathbf{I}_N-\delta\mathbf{L}-\mathbf{J}\|\,\hat{p}_{A_0}(t) + \frac{NA_0}{2(t+1)} + \frac{NA_0(t+1)^{\delta_0}}{t+1}, \tag{C.213}$$

where $\hat{p}_{A_0}(0)$ is initialized as $\hat{p}_{A_0}(0)=\|\mathbf{p}_{A_0}(0)\|$ and δ is as defined in (3.12). From (C.206), we have,

$$\begin{aligned}\|\mathbf{p}_{A_0}(t+1)\| &\leq \frac{t}{t+1}\|(\mathbf{W}-\mathbf{J})\|\,\|\mathbf{p}_{A_0}(t)\| \\ &+ \frac{1}{t+1}\left\|\left(J_1\left(\mathbf{y}(t)\right)\right)^{A_0(t+1)^\delta}\right\| \\ &+ \frac{1}{2(t+1)}\left\|\left(\left(\mathbf{h}^*\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta(t)\right)\right.\right.\right. \\ &\left.\left.\left. - \frac{\mathbf{1}_N}{N}\otimes\left(\mathbf{h}^\top\left(\theta(t)\right)\boldsymbol{\Sigma}^{-1}\mathbf{h}\left(\theta(t)\right)\right)\right)^{A_0}\right\| \\ &\leq \|\mathbf{I}_N-\delta\mathbf{L}-\mathbf{J}\|\,\|\mathbf{p}_{A_0}(t)\| + \frac{NA_0}{2(t+1)} + \frac{NA_0(t+1)^{\delta_0}}{t+1}.\end{aligned} \tag{C.214}$$

Given the initial condition for $\hat{p}_{A_0}(0)$, through an induction argument we have that

$$\|\mathbf{p}_{A_0}(t+1)\| \leq \hat{p}_{A_0}(t+1), \forall t. \tag{C.215}$$

Moreover, we also have that,

$$\|\mathbf{I}_N-\delta\mathbf{L}-\mathbf{J}\| = \frac{\lambda_N\left(\mathbf{L}\right)-\lambda_2\left(\mathbf{L}\right)}{\lambda_N\left(\mathbf{L}\right)+\lambda_2\left(\mathbf{L}\right)}. \tag{C.216}$$

Using (C.216) in (C.213), we have,

$$\hat{p}_{A_0}(t+1) \leq \left(1 - \frac{2\lambda_2(\mathbf{L})}{\lambda_N(\mathbf{L}) + \lambda_2(\mathbf{L})}\right) \hat{p}_{A_0}(t) + \frac{2NA_0}{(t+1)^{1-\delta_0}}, \tag{C.217}$$

where $\frac{2\lambda_2(\mathbf{L})}{\lambda_N(\mathbf{L})+\lambda_2(\mathbf{L})} < 1$ and hence the recursion in (C.217) comes under the purview of Lemma C.3.1. Hence, we have

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty} (t+1)^{\delta_0} \hat{p}_{A_0}(t) = 0\right) = 1. \tag{C.218}$$

Finally, the assertion follows from by invoking (C.215) and noting that, for arbitrary $A_0 > 0$,

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty} (t+1)^{\delta_0} \mathbf{p}_{A_0}(t) = 0\right) = 1. \tag{C.219}$$

□

C.4 Proofs of Lemmas in Section C.2

Proof of Lemma C.2.1. First, we note that both the matrices $\mathbf{L} \otimes \mathbf{I}_M$ and $\mathbf{G}_H \boldsymbol{\Sigma}^{-1} \mathbf{G}_H^\top$ are symmetric and positive semi-definite. Then the matrix $\mathbf{L} \otimes \mathbf{I}_M + \mathbf{G}_H \boldsymbol{\Sigma}^{-1} \mathbf{G}_H^\top$ is positive semi-definite as it is the sum of two positive semi-definite matrices. To prove that the matrix $\mathbf{L} \otimes \mathbf{I}_M + \mathbf{G}_H \boldsymbol{\Sigma}^{-1} \mathbf{G}_H^\top$ is positive definite, let's assume that it's not positive definite. Hence there exists $\mathbf{x} \in \mathbb{R}^{NM}$, where $\mathbf{x} \neq \mathbf{0}$ such that

$$\mathbf{x}^\top \left(\mathbf{L} \otimes \mathbf{I}_M + \mathbf{G}_H \boldsymbol{\Sigma}^{-1} \mathbf{G}_H^\top\right) \mathbf{x} = 0, \tag{C.220}$$

which further implies that

$$\mathbf{x}^\top \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{x} = 0 \quad \text{and} \quad \mathbf{x}^\top \left(\mathbf{G}_H \boldsymbol{\Sigma}^{-1} \mathbf{G}_H^\top\right) \mathbf{x} = 0. \tag{C.221}$$

Moreover, $\mathbf{x}$ can be written as $\mathbf{x} = \left[\mathbf{x}_1^\top, \cdots, \mathbf{x}_N^\top\right]^\top$, with $\mathbf{x}_n \in \mathbb{R}^M$ for all n. Now note that, by the properties of the graph Laplacian (C.221) holds if and only if (*iff*)

$$\mathbf{x}_n = \mathbf{g}, \ \forall n, \tag{C.222}$$

where $\mathbf{g} \in \mathbb{R}^M$ and $\mathbf{g} \neq \mathbf{0}$. Hence, from (C.221),we have,

$$\sum_{n=1}^{N} \mathbf{g}^\top \mathbf{H}_n^\top \boldsymbol{\Sigma}_n^{-1} \mathbf{H}_n \mathbf{g} = \mathbf{g}^\top \mathbf{G} \mathbf{g} = 0, \tag{C.223}$$

which is a contradiction from Assumption 3.6.1 as $\mathbf{G}$ is invertible. Hence, we have that $\mathbf{L} \otimes \mathbf{I}_M + \mathbf{G}_H \boldsymbol{\Sigma}^{-1} \mathbf{G}_H^\top$ is positive definite. Since $\beta_t/\alpha_t \to \infty$ as $t \to \infty$, there exists an integer t_4 (sufficiently large) such that $\forall t \geq t_4$ and for all $\mathbf{x}$ with $\|\mathbf{x}\| = 1$,

$$\begin{aligned}
&\mathbf{x}^\top \left(\beta_t \left(\mathbf{L} \otimes \mathbf{I}_M\right) + \alpha_t \mathbf{G}_H \boldsymbol{\Sigma}^{-1} \mathbf{G}_H^\top\right) \mathbf{x} \\
&= \alpha_t \mathbf{x}^\top \left(\frac{\beta_t}{\alpha_t} \left(\mathbf{L} \otimes \mathbf{I}_M\right) + \mathbf{G}_H \boldsymbol{\Sigma}^{-1} \mathbf{G}_H^\top\right) \mathbf{x}
\end{aligned}$$

$$\geq \alpha_t \mathbf{x}^\top \left((\mathbf{L} \otimes \mathbf{I}_M) + \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top \right) \mathbf{x} \geq c_1 \alpha_t, \tag{C.224}$$

where

$$c_1 = \lambda_{\min} \left((\mathbf{L} \otimes \mathbf{I}_M) + \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top \right). \tag{C.225}$$

We now choose a $t_3 > t_4$ such that $\forall t \geq t_3$, $c_1 \alpha_t < 1$.
In order to ensure that all the eigenvalues of $\left(\mathbf{I}_{NM} - \beta_t (\mathbf{L} \otimes \mathbf{I}_M) - \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top \right)$ are positive, we choose a t_2 such that $\forall t \geq t_2$,

$$\beta_t \lambda_N(\mathbf{L}) + \alpha_t \lambda_{\max}(\mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top) < 1. \tag{C.226}$$

It is to be noted that such choices of t_3 and t_2 are possible as $\beta_t, \alpha_t \to 0$ as $t \to \infty$. Moreover, the condition in (C.226) readily implies that $\lambda_{\max} \left(\beta_t (\mathbf{L} \otimes \mathbf{I}_M) + \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top \right) \leq \beta_t \lambda_N(\mathbf{L}) + \alpha_t \lambda_{\max}(\mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top) < 1$ for all $t \geq t_2$. Hence, from (C.224), we have $\forall t \geq t_1$, with $t_1 = \max\{t_2, t_3\}$, and for all $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$,

$$\mathbf{x}^\top \left(\mathbf{I}_{NM} - \beta_t (\mathbf{L} \otimes \mathbf{I}_M) - \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top \right) \mathbf{x} \leq 1 - c_1 \alpha_t, \tag{C.227}$$

which implies that

$$\left\| \mathbf{I}_{NM} - \beta_t (\mathbf{L} \otimes \mathbf{I}_M) - \alpha_t \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top \right\| \leq 1 - c_1 \alpha_t, \tag{C.228}$$

for all $t \geq t_1$. □

Proof of Lemma C.2.2. The following Lemma from Bourin and Lee (2013), will be used in the subsequent analysis.

Lemma C.4.1 (Bourin and Lee (2013))**.** *Given a positive-semidefinite matrix* $\mathbf{P}$ *$(Nt \times Nt)$, with each of its blocks $(N \times N)$ being symmetric, the following result holds for any invariant norm,*

$$\|\mathbf{P}\| \leq \left\| \sum_{i=1}^{t} [\mathbf{P}]_{ii} \right\|. \tag{C.229}$$

From Lemma C.4.1, we have that,

$$\|\mathbf{P}_t\| \leq \sum_{i=1}^{t} \|[\mathbf{P}_t]_{ii}\|. \tag{C.230}$$

From Lemma C.2.1. we have that, $\forall t \geq t_1$,

$$\|\mathbf{A}(u)\| \leq (1 - c_1 \alpha_u), \tag{C.231}$$

which implies

$$\|[\mathbf{P}_t]_{ii}\| \leq \alpha_i^2 \prod_{u=i}^{t-1} (1 - c_1 \alpha_u)^2, \tag{C.232}$$

for all $t \geq t_1$. Using (C.231), the RHS of (C.230) can be rewritten as

$$\sum_{i=1}^{t} \|[\mathbf{P}_t]_{ii}\| \leq c_3 \prod_{u=t_1}^{t-1} (1 - c_1\alpha_u)^2 + \sum_{v=t_1}^{t} \alpha_v^2 \prod_{u=v+1}^{t-1} (1 - c_1\alpha_u)^2, \tag{C.233}$$

where c_3 is given by

$$c_3 = \sum_{v=0}^{t_1-1} \alpha_v^2 \prod_{u=v+1}^{t_1-1} \|\mathbf{A}(u)\|. \tag{C.234}$$

Using the properties of Riemann integration and the inequality $(1-x) \leq e^{-x}$, for $x \in (0,1)$, we have,

$$\prod_{u=i}^{t-1} (1 - c_1\alpha_u)^2 \leq \left(\frac{i+1}{t}\right)^{2c_1\alpha_0}, \tag{C.235}$$

where, in the derivation, we also use the property that

$$\sum_{u=i+1}^{t} \frac{1}{u} > \ln\left(\frac{t}{i+1}\right). \tag{C.236}$$

On using (C.235), in (C.233) we have $\forall t \geq t_1$,

$$\begin{aligned}
&\sum_{i=1}^{t} \|[\mathbf{P}_t]_{ii}\| \leq c_3 \prod_{u=t_1}^{t-1} (1 - c_1\alpha_u)^2 \\
&+ \sum_{v=t_1}^{t} \alpha_v^2 \prod_{u=v+1}^{t-1} (1 - c_1\alpha_u)^2 \\
&\leq c_3 \left(\frac{t_1+1}{t}\right)^{2c_1\alpha_0} + \sum_{u=t_1+1}^{t-1} \alpha_u^2 \left(\frac{u+1}{t}\right)^{2c_1\alpha_0} \\
&= c_3 \left(\frac{t_1+1}{t}\right)^{2c_1\alpha_0} + \alpha_0^2 \sum_{u=t_1+1}^{t-1} \frac{1}{t^{2c_1\alpha_0}(u+1)^{2-2c_1\alpha_0}} \\
&= c_3 \left(\frac{t_1+1}{t}\right)^{2c_1\alpha_0} + \frac{\alpha_0^2}{t^2} \\
&+ \alpha_0^2 \sum_{u=t_1+1}^{t-2} \frac{1}{t^{2c_1\alpha_0}(u+1)^{2-2c_1\alpha_0}} \\
&\overset{(a)}{\leq} c_3 \left(\frac{t_1+1}{t}\right)^{2c_1\alpha_0} + \frac{\alpha_0^2}{t^2} \\
&+ \frac{\alpha_0^2}{t^{2c_1\alpha_0}} \int_{t_1}^{t-1} \frac{1}{(s+1)^{2-2c_1\alpha_0}} ds \\
&\leq c_3 \left(\frac{t_1+1}{t}\right)^{2c_1\alpha_0} + \frac{\alpha_0^2}{t^2} + \frac{\alpha_0^2}{t^{2c_1\alpha_0}} \left(\frac{t^{2c_1\alpha_0-1}}{2c_1\alpha_0 - 1}\right).
\end{aligned} \tag{C.237}$$

The above implies that, for all $t \geq t_1$,

$$t \sum_{i=1}^{t} \|[\mathbf{P}_t]_{ii}\|$$

$$\overset{(b)}{\leq} c_3 \frac{(t_1+1)^{2c_1\alpha_0}}{t^{2c_1\alpha_0-1}} + \frac{\alpha_0^2}{t} + \frac{\alpha_0^2}{2c_1\alpha_0-1}. \tag{C.238}$$

where in (a) and (b), we use the fact that $2c_1\alpha_0 - 1 > 1$ by Assumption 3.6.5. The proof follows by noting that the RHS of (C.238) is a non-increasing function of t. □

Proof of Theorem 3.9.1. The proof of Theorem 3.9.1 needs the following Lemma from Fabian (1968) concerning the asymptotic normality of non-Markov stochastic recursions.

Lemma C.4.2 (Theorem 2.2 in Fabian (1968))**.** *Let $\{\mathbf{z}_t\}$ be an $\mathbb{R}^k$-valued $\{\mathcal{F}_t\}$-adapted process that satisfies*

$$\mathbf{z}_{t+1} = \left(\mathbf{I}_k - \frac{1}{t+1}\Gamma_t\right)\mathbf{z}_t + (t+1)^{-1}\mathbf{\Phi}_t\mathbf{V}_t + (t+1)^{-3/2}\mathbf{T}_t, \tag{C.239}$$

where the stochastic processes $\{\mathbf{V}_t\}, \{\mathbf{T}_t\} \in \mathbb{R}^k$ while $\{\mathbf{\Gamma}_t\}, \{\mathbf{\Phi}_t\} \in \mathbb{R}^{k\times k}$. Moreover, for each t, $\mathbf{V}_{t-1}$ and $\mathbf{T}_t$ are $\mathcal{F}_t$ -adapted, whereas the processes $\{\mathbf{\Gamma}_t\}$, $\{\mathbf{\Phi}_t\}$ are $\{\mathcal{F}_t\}$ adapted.
Also, assume that

$$\mathbf{\Gamma}_t \to \mathbf{\Gamma}, \mathbf{\Phi}_t \to \mathbf{\Phi} \text{ and } \mathbf{T}_t \to 0 \text{ as } t \to \infty, \tag{C.240}$$

where $\mathbf{\Gamma}$ is symmetric and positive definite, and admits an eigen decomposition of the form $\mathbf{P}^\top\mathbf{\Gamma}\mathbf{P} = \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is diagonal and $\mathbf{P}$ is orthogonal. Furthermore, let the sequence $\{\mathbf{V}_t\}$ satisfy $\mathbb{E}[\mathbf{V}_t|\mathcal{F}_t] = 0$ for each t and there exists a positive constant C and a matrix Σ such that $C > \left\|\mathbb{E}\left[\mathbf{V}_t\mathbf{V}_t^\top|\mathcal{F}_t\right] - \Sigma\right\| \to 0$ as $t \to \infty$ and with $\sigma_{t,r}^2 = \int_{\|\mathbf{V}_t\|^2 \geq r(t+1)} \|\mathbf{V}_t\|^2 \, d\mathbb{P}$, let $\lim_{t\to\infty} \frac{1}{t+1}\sum_{s=0}^t \sigma_{s,r}^2 = 0$ for every $r > 0$.
Then, we have,

$$(t+1)^{1/2}\mathbf{z}_t \overset{\mathcal{D}}{\Longrightarrow} \mathcal{N}\left(\mathbf{0}, \mathbf{P}\mathbf{M}\mathbf{P}^\top\right), \tag{C.241}$$

where

$$[\mathbf{M}]_{ij} = \left[\mathbf{P}^\top\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top\mathbf{P}\right]_{ij}\left([\mathbf{\Lambda}]_{ii} + [\mathbf{\Lambda}]_{jj} - 1\right)^{-1}. \tag{C.242}$$

Define the process $\{\hat{\mathbf{z}}(t)\}$ as

$$\hat{\mathbf{z}}(t) = \mathbf{z}(t) - (b\mathbf{L} + \mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N \otimes \theta^*)\mathbf{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N \otimes \theta^*)}{2}. \tag{C.243}$$

It can be shown that the process $\{\hat{\mathbf{z}}(t)\}$ satisfies the recursion

$$\begin{aligned}\hat{\mathbf{z}}(t+1) = &\left(\mathbf{I}_N - \frac{1}{t+1}\mathbf{\Gamma}(t)\right)\hat{\mathbf{z}}(t) + \frac{1}{t+1}\mathbf{\Phi}(t)\mathbf{V}(t) \\ &+ (t+1)^{-3/2}\mathbf{T}(t),\end{aligned} \tag{C.244}$$

where the processes $\{\mathbf{\Gamma}(t)\}$, $\{\mathbf{\Phi}(t)\}$, $\{\mathbf{V}(t)\}$ and $\{\mathbf{T}_t\}$ are given by

$$\begin{aligned}&\mathbf{\Gamma}(t) = \mathbf{\Gamma} = b\mathbf{L} + \mathbf{I}, \mathbf{\Phi}(t) = \mathbf{\Phi} = \mathbf{I}, \\ &\mathbf{V}(t) = \mathbf{K}(t)\mathbf{\Sigma}^{-1}\gamma(t) + b\zeta(t) = \mathbf{h}^*(\theta(t))\mathbf{\Sigma}^{-1}\gamma(t) + b\zeta(t) \\ &\mathbf{T}(t) = (t+1)^{1/2}\mathbf{U}(t)\end{aligned}$$

$$= (t+1)^{1/2} \left(\mathbf{h}^*(\theta(t)) - \mathbf{h}^*(\mathbf{1}_N \otimes \theta^*)\right) \mathbf{\Sigma}^{-1} \left(\mathbf{h}(\mathbf{1}_N \otimes \theta^*) - \mathbf{h}(\theta(t))\right). \tag{C.245}$$

From, Theorem 3.7.1, we have the following convergences : $(t+1)^{1/2}\mathbf{U}(t) \to \mathbf{0}$ and $\mathbf{K}(t) \to \mathbf{K}^* = \mathbf{h}^*(\theta^*)$ a.s. as $t \to \infty$. By Egorov's Theorem the a.s. convergence can be taken to be uniform convergence on sets of arbitrarily large probability measure, and hence for every $\delta > 0$, there exist uniformly bounded processes $\left\{\mathbf{U}^\delta(t)\right\}$ and $\left\{\mathbf{K}^\delta(t)\right\}$ satisfying

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\sup_{s \geq t_\epsilon^\delta} \left\|\mathbf{K}^\delta(s) - \mathbf{K}^*\right\| > \epsilon\right) = 0 \\
&\mathbb{P}_{\theta^*}\left(\sup_{s \geq t_\epsilon^\delta} (t+1)^{1/2} \left\|\mathbf{U}^\delta(s)\right\| > \epsilon\right) = 0
\end{aligned} \tag{C.246}$$

for each $\epsilon > 0$ and some t_ϵ^δ chosen appropriately large, such that

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\sup_{t \geq 0} \max\left\{\left\|\mathbf{K}^\delta(t) - \mathbf{K}(t)\right\|, \left\|\mathbf{U}^\delta(t) - \mathbf{U}(t)\right\|\right\} = 0\right) \\
&> 1 - \delta.
\end{aligned} \tag{C.247}$$

With the above development in place, we define the process $\{\hat{\mathbf{z}}^\delta(t)\}$ for each $\delta > 0$ by,

$$\begin{aligned}
\hat{\mathbf{z}}^\delta(t+1) &= \left(\mathbf{I}_N - \frac{1}{t+1}\mathbf{\Gamma}\right)\hat{\mathbf{z}}^\delta(t) + \frac{1}{t+1}\mathbf{\Phi}\mathbf{V}^\delta(t) \\
&+ (t+1)^{-3/2}\mathbf{T}^\delta(t),
\end{aligned} \tag{C.248}$$

where $\hat{\mathbf{z}}^\delta(0) = \hat{\mathbf{z}}(t)$, $\mathbf{V}^\delta(t) = \mathbf{K}^\delta(t)\mathbf{\Sigma}^{-1}\gamma(t) + b\mathbf{\Psi}(t)$, $\mathbf{T}^\delta(t) = (t+1)^{1/2}\mathbf{U}^\delta(t)$ and

$$\mathbb{P}_{\theta^*}\left(\sup_{t \geq 0} \left\|\hat{\mathbf{z}}^\delta(t) - \hat{\mathbf{z}}(t)\right\| = 0\right) > 1 - \delta. \tag{C.249}$$

Since, $\mathbf{T}(t)$ and $\mathbf{K}(t)$ may not converge uniformly, Lemma C.1.5 might not be directly applicable. Hence, we first consider the process $\{\hat{\mathbf{z}}^\delta(t)\}$ for some $\delta > 0$. It is to noted that by construction, we have,

$$\begin{aligned}
&\lim_{t\to\infty} \mathbf{T}^\delta(t) = \mathbf{0}, \mathbb{E}_{\theta^*}\left[\mathbf{V}^\delta(t)|\mathcal{F}_t\right] = 0 \ \forall t, \\
&\lim_{t\to\infty} \mathbb{E}\left[\mathbf{V}^\delta(t)(\mathbf{V}^\delta(t))^\top|\mathcal{F}_t\right] \\
&\overset{(a)}{=} \mathbf{h}^*(\mathbf{1}_N \otimes \theta^*)\mathbf{\Sigma}^{-1}(\mathbf{h}^*(\mathbf{1}_N \otimes \theta^*))^\top + b^2\mathbf{\Sigma}_c \\
&= \mathbf{\Sigma}_1^*,
\end{aligned} \tag{C.250}$$

where (a) follows from the fact that the agent observation noises and channel noises are mutually uncorrelated. Moreover, the uniform boundedness of the process $\left\{\mathbf{K}_t^\delta\right\}$ ensures the existence of a constant C such that

$$\left\|\mathbb{E}\left[\mathbf{V}^\delta(t)(\mathbf{V}^\delta(t))^\top|\mathcal{F}_t\right] - \mathbf{\Sigma}_1^*\right\| < C, \ \forall t \geq 0. \tag{C.251}$$

Furthermore, the process $\{\mathbf{V}^\delta(t)\}$ satisfies the uniform integrability assumption of Lemma C.1.5. Hence, by

Lemma C.1.5, we have that,

$$(t+1)^{1/2}\left(\hat{\mathbf{z}}^{\delta}(t)\right) \xRightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{\Sigma}^{*}_{c,1}), \tag{C.252}$$

for each $\delta > 0$ where $\mathbf{\Sigma}^{*}_{c,1}$ is as defined in (3.51). To extend the asymptotic normality to the process $\{\hat{\mathbf{z}}(t)\}$, define any continuous bounded function $g : \mathbb{R}^N \to \mathbb{R}$. By, Portmanteau's theorem (Billingsley (1999)), we have that

$$\lim_{t\to\infty} \mathbb{E}_{\theta^*}\left[g\left((t+1)^{1/2}\hat{\mathbf{z}}^{\delta}(t)\right)\right] = \mathbb{E}_{\theta^*}\left[g\left(\mathbf{z}\right)\right], \tag{C.253}$$

for each δ, where $\mathbf{z}$ is a normal random variable with $\mathbf{z} \sim \mathcal{N}(0, \mathbf{\Sigma}^{*}_{c,1})$. We denote the sup-norm of $g(.)$ as $\|g\|_{\infty}$ which is necessarily finite and hence from (C.249), we have that,

$$\begin{aligned}&\left\|\mathbb{E}_{\theta^*}\left[g\left((t+1)^{1/2}\hat{\mathbf{z}}^{\delta}(t)\right)\right] - \mathbb{E}_{\theta^*}\left[g\left((t+1)^{1/2}\hat{\mathbf{z}}(t)\right)\right]\right\| \\ &\leq 2\delta\left\|g\right\|_{\infty}. \end{aligned} \tag{C.254}$$

We then have, from (C.253),

$$\limsup_{t\to\infty}\left\|\mathbb{E}_{\theta^*}\left[g\left((t+1)^{1/2}\hat{\mathbf{z}}(t)\right)\right] - \mathbb{E}_{\theta^*}\left[g\left(\mathbf{z}\right)\right]\right\| \leq 2\delta\left\|g\right\|_{\infty}. \tag{C.255}$$

Since the above development holds for each $\delta > 0$, we have,

$$\lim_{t\to\infty} \mathbb{E}_{\theta^*}\left[g\left((t+1)^{1/2}\hat{\mathbf{z}}(t)\right)\right] = \mathbb{E}_{\theta^*}\left[g\left(\mathbf{z}\right)\right]. \tag{C.256}$$

As the weak convergence derived above holds for all bounded continuous functions $g(.)$, we have the required weak convergence (convergence in distribution) of the sequence $\left\{(t+1)^{1/2}\hat{\mathbf{z}}(t)\right\}$.
The proof for the asymptotic normality under $\mathbb{P}_0$ follows exactly in a similar way.

□

Proof of Theorem 3.9.2. From (3.28), we have,

$$\begin{aligned}&\mathbb{P}_{M,\theta^*}(t) = \mathbb{P}_{1,\theta^*}\left(z_n(t) < \eta_n\right) \\ &= \mathbb{P}_{1,\theta^*}\left(z_n(t) - \left[(b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N\otimes\theta^*)\mathbf{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N\otimes\theta^*)}{2}\right]_n\right. \\ &\left. < \eta_n - \left[(b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N\otimes\theta^*)\mathbf{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N\otimes\theta^*)}{2}\right]_n\right) \\ &= \mathbb{P}_{1,\theta^*}\left(\sqrt{t+1}\left(z_n(t) - \left[(b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N\otimes\theta^*)\mathbf{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N\otimes\theta^*)}{2}\right]_n\right)\right. \\ &\left. < \sqrt{t+1}\left(\eta_n - \left[(b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N\otimes\theta^*)\mathbf{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N\otimes\theta^*)}{2}\right]_n\right)\right). \end{aligned} \tag{C.257}$$

To this end, invoking Theorem 3.9.1, where we established asymptotic normality for the decision statistic

sequence $\{z_n(t)\}$, we have,

$$\begin{aligned}
&\lim_{t\to\infty} \mathbb{P}_{1,\theta^*}\left(\sqrt{t+1}\left(z_n(t) - \left[(b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N\otimes\theta^*)\boldsymbol{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N\otimes\theta^*)}{2}\right]_n\right)\right.\\
&\left.< \sqrt{t+1}\left(\eta_n - \left[(b\mathbf{L}+\mathbf{I})^{-1}\frac{\mathbf{h}^*(\mathbf{1}_N\otimes\theta^*)\boldsymbol{\Sigma}^{-1}\mathbf{h}(\mathbf{1}_N\otimes\theta^*)}{2}\right]_n\right)\right)\\
&= \mathbb{P}_{1,\theta^*}(z < -\infty) = 0, \qquad \text{(C.258)}
\end{aligned}$$

where z is a normal random variable with $z \sim \mathcal{N}\left(0, \left[\boldsymbol{\Sigma}^*_{c,1}\right]_{nn}\right)$.
Under $\mathcal{H}_0$, we have,

$$\begin{aligned}
&\mathbb{P}_{FA}(t) = \mathbb{P}_0\left(z_n(t) > \eta_n\right)\\
&= \mathbb{P}_0\left(\frac{1}{t}\sum_{s=0}^{t-1}\mathbf{e}_n^\top\boldsymbol{\Phi}(s,t-1)\left(\mathbf{h}^*(\theta(s))\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}(s) - \frac{\mathbf{h}(\theta(s))}{2}\right)\right.\right.\\
&\left.\left. + \; b\zeta(s)\right) > \eta_n\right)\\
&\leq \underbrace{\mathbb{P}_0\left(\frac{1}{t}\sum_{s=0}^{t-1}\mathbf{e}_n^\top\boldsymbol{\Phi}(s,t-1)\left(\mathbf{h}^*(\theta(s))\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}(s) - \frac{\mathbf{h}(\theta(s))}{2}\right)\right) > \frac{\eta_n}{2}\right)}_{\text{(a1)}}\\
&+ \underbrace{\mathbb{P}_0\left(\frac{b}{t}\sum_{s=0}^{t-1}\mathbf{e}_n^\top\boldsymbol{\Phi}(s,t-1)\zeta(s) > \frac{\eta_n}{2}\right)}_{\text{(a2)}}, \qquad \text{(C.259)}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\Phi}(s,t-1) &= \prod_{q=s}^{t-1}\left(\mathbf{I} - \alpha_q\mathbf{L}\right), \; \textit{if } s \neq t-1,\\
\boldsymbol{\Phi}(s,t-1) &= \mathbf{I}, \; \textit{if } s = t-1. \qquad \text{(C.260)}
\end{aligned}$$

We first analyze the term $\boldsymbol{\Phi}(s,t-1) - \mathbf{J}$. We have,

$$\begin{aligned}
\|\boldsymbol{\Phi}(0,t-1) - \mathbf{J}\| &\leq \prod_{s=0}^{t-1}(1 - b\alpha_s\lambda_2(\mathbf{L}))\\
&\leq \exp\left(-b\lambda_2(\mathbf{L})\sum_{s=0}^{t-1}\alpha_s\right)\\
&\leq \left(\frac{1}{t}\right)^{b\lambda_2(\mathbf{L})}. \qquad \text{(C.261)}
\end{aligned}$$

Similarly, it may be shown that for each $0 < s \leq t-1$,

$$\|\boldsymbol{\Phi}(s,t-1) - \mathbf{J}\| \leq \left(\frac{s+1}{t}\right)^{b\lambda_2(\mathbf{L})} \leq 1, \qquad \text{(C.262)}$$

which implies that $\left\|\boldsymbol{\Phi}_{n,j}(s,t-1) - \frac{1}{N}\right\| \leq \sqrt{N}$, where $\boldsymbol{\Phi}_{n,j}(s,t-1)$ denotes (n,j)-th entry of $\boldsymbol{\Phi}(s,t-1)$.

Now, we analyze the term $(a1)$. From (C.25), we have,

$$\begin{aligned}
&\mathbb{P}_0\left(\frac{1}{t}\sum_{s=0}^{t-1}\mathbf{e}_n^\top\mathbf{\Phi}(s,t-1)\left(\mathbf{h}^*(\theta(s))\mathbf{\Sigma}^{-1}\left(\mathbf{y}(s)-\frac{\mathbf{h}(\theta(s))}{2}\right)\right)\right.\\
&\left.>\frac{\eta_n}{2}\right)\\
&=\mathbb{P}_0\left(\frac{1}{t}\sum_{s=0}^{t-1}\sum_{j=1}^{N}\mathbf{\Phi}_{n,j}(s,t-1)\left(\frac{\gamma_j^\top(s)\mathbf{\Sigma}_j^{-1}\gamma_j(s)}{2}\right.\right.\\
&\left.\left.-\frac{\left(\gamma_j(s)-\mathbf{h}_j\left(\theta_j(s)\right)\right)^\top\mathbf{\Sigma}_j^{-1}\left(\gamma_j(s)-\mathbf{h}_j\left(\theta_j(s)\right)\right)}{2}\right)>\frac{\eta_n}{2}\right)\\
&\leq\mathbb{P}_0\left(\frac{1}{t}\sum_{s=0}^{t-1}\sum_{j=1}^{N}\mathbf{\Phi}_{n,j}(s,t-1)\left(\frac{\gamma_j^\top(s)\mathbf{\Sigma}_j^{-1}\gamma_j(s)}{2}\right)>\frac{\eta_n}{2}\right)\\
&\leq\underbrace{\mathbb{P}_0\left(\frac{1}{t}\sum_{s=0}^{t-1}\sum_{j=1}^{N}\left\|\mathbf{\Phi}_{n,j}(s,t-1)-\frac{1}{N}\right\|\left(\frac{\gamma_j^\top(s)\mathbf{\Sigma}_j^{-1}\gamma_j(s)}{2}\right)>\frac{\eta_n}{4}\right)}_{(t1)}\\
&+\underbrace{\mathbb{P}_0\left(\frac{1}{Nt}\sum_{s=0}^{t-1}\sum_{j=1}^{N}\left(\frac{\gamma_j^\top(s)\mathbf{\Sigma}_j^{-1}\gamma_j(s)}{2}\right)>\frac{\eta_n}{4}\right)}_{(t2)}\\
&\leq\mathbb{P}_0\left(\frac{1}{t}\sum_{s=0}^{t-1}\sum_{j=1}^{N}\sqrt{N}\left(\frac{\gamma_j^\top(s)\mathbf{\Sigma}_j^{-1}\gamma_j(s)}{2}\right)>\frac{\eta_n}{4}\right)\\
&+\mathbb{P}_0\left(\frac{1}{Nt}\sum_{s=0}^{t-1}\sum_{j=1}^{N}\left(\frac{\gamma_j^\top(s)\mathbf{\Sigma}_j^{-1}\gamma_j(s)}{2}\right)>\frac{\eta_n}{4}\right),
\end{aligned}\tag{C.263}$$

where $\mathbf{\Phi}_{n,j}\left(s,t-1\right)$ denotes the (n,j)th element of $\mathbf{\Phi}(s,t-1)$. Hence, it can be concluded the decay exponent of $(t2)$ will dominate the decay exponent of $(a1)$. For $0<\lambda<1$, we have,

$$\begin{aligned}
&\mathbb{P}_0\left(\frac{1}{Nt}\sum_{s=0}^{t-1}\sum_{j=1}^{N}\left(\frac{\gamma_j^\top(s)\mathbf{\Sigma}_j^{-1}\gamma_j(s)}{2}\right)>\frac{\eta_n}{4}\right)\\
&\leq\exp\left(-\frac{Nt\eta_n\lambda}{4}\right)\prod_{j=1}^{N}\prod_{s=0}^{t-1}\mathbb{E}_0\left[\exp\left(\lambda\frac{\gamma_j(s)^\top\mathbf{\Sigma}_j^{-1}\gamma_j(s)}{2}\right)\right]\\
&=\exp\left(-\frac{Nt\eta_n\lambda}{4}-t\left(\frac{\sum_{n=1}^{N}M_n}{2}\right)\log\left(1-\lambda\right)\right)
\end{aligned}\tag{C.264}$$

which implies that,

$$\begin{aligned}
&\frac{1}{t}\log\left(\mathbb{P}_0\left(z_n(t)>\frac{\eta_n}{4}\right)\right)\\
&\leq-\frac{N\eta_n\lambda}{4}-\left(\frac{\sum_{n=1}^{N}M_n}{2}\right)\log\left(1-\lambda\right).
\end{aligned}\tag{C.265}$$

Hence, we finally have,

$$\begin{aligned}&\limsup_{t\to\infty}\frac{1}{t}\log\left(\mathbb{P}_0\left(z_n(t)>\frac{\eta_n}{4}\right)\right)\\&\leq-\frac{N\eta_n\lambda}{4}-\left(\frac{\sum_{n=1}^{N}M_n}{2}\right)\log\left(1-\lambda\right)=-LE(\lambda).\end{aligned}\tag{C.266}$$

Assuming η_n to be,

$$\eta_n>\frac{2\sum_{n=1}^{N}M_n}{N},\tag{C.267}$$

we have that, on maximizing $LE(\lambda)$

$$\lambda^*=1-\frac{2\sum_{n=1}^{N}M_n}{N\eta_n},\tag{C.268}$$

λ^* is a feasible maximizer.
Hence, using (C.268) and (C.267) in (C.266), we have,

$$\begin{aligned}&\limsup_{t\to\infty}\frac{1}{t}\log\left(\mathbb{P}_0\left(z_n(t)>\frac{\eta_n}{4}\right)\right)\\&\leq-\frac{N\eta_n}{4}-\frac{\sum_{n=1}^{N}M_n}{2}\left(\frac{1}{N}+\log\frac{N\eta_n}{2\sum_{n=1}^{N}M_n}\right).\end{aligned}\tag{C.269}$$

Now, we analyze the term $(a2)$ in (C.259) and for $\lambda>0$, we have,

$$\begin{aligned}&\mathbb{P}_0\left(\frac{b}{t}\sum_{s=0}^{t-1}\mathbf{e}_n^\top\mathbf{\Phi}(s,t-1)\zeta(s)>\frac{\eta_n}{2}\right)\\&\leq\exp\left(-\frac{t\lambda\eta_n}{2}\right)\mathbb{E}_0\left[\exp\left(b\lambda\sum_{s=0}^{t-1}\mathbf{e}_n^\top\mathbf{\Phi}(s,t-1)\zeta(s)\right)\right]\\&=\exp\left(-\frac{t\lambda\eta_n}{2}\right)\prod_{s=0}^{t-1}\mathbb{E}_0\left[\exp\left(b\lambda\mathbf{e}_n^\top\mathbf{\Phi}(s,t-1)\zeta(s)\right)\right]\\&=\exp\left(-\frac{t\lambda\eta_n}{2}\right)\prod_{s=0}^{t-1}\exp\left(\frac{\lambda^2b^2}{2}\mathbf{e}_n^\top\mathbf{\Phi}(s,t-1)\mathbf{\Sigma}_c\mathbf{\Phi}(s,t-1)\mathbf{e}_n\right)\\&\leq\exp\left(-\frac{t\lambda\eta_n}{2}\right)\prod_{s=0}^{t-1}\exp\left(\frac{\lambda^2b^2}{2}\left\|\mathbf{\Sigma}_c\right\|\right)\\&\overset{(a)}{=}\exp\left(-\frac{t\lambda\eta_n}{2}+\frac{t\lambda^2b^2\left\|\mathbf{\Sigma}_c\right\|}{2}\right).\end{aligned}\tag{C.270}$$

Taking limits on both sides, we have,

$$\limsup_{t\to\infty}\frac{1}{t}\log\left(\mathbb{P}_0\left(\frac{b}{t}\sum_{s=0}^{t-1}\mathbf{e}_n^\top\mathbf{\Phi}(s,t-1)\zeta(s)>\frac{\eta_n}{2}\right)\right)$$

$$\leq \exp\left(-\frac{\lambda\eta_n}{2}+\frac{\lambda^2 b^2\left\|\mathbf{\Sigma}_c\right\|}{2}\right). \tag{C.271}$$

In order to have a positive decay exponent for $(a2)$, we have,

$$\lambda < \frac{\eta_n}{b^2\left\|\mathbf{\Sigma}_c\right\|}. \tag{C.272}$$

To this end, minimizing the right hand side (RHS) of (C.271), we have,

$$\begin{aligned}&\limsup_{t\to\infty}\frac{1}{t}\log\left(\mathbb{P}_0\left(\frac{b}{t}\sum_{s=0}^{t-1}\mathbf{e}_n^\top\mathbf{\Phi}(s,t-1)\zeta(s)>\frac{\eta_n}{2}\right)\right)\\&\leq -\frac{\eta_n^2}{8b^2\left\|\mathbf{\Sigma}_c\right\|},\end{aligned} \tag{C.273}$$

where the minimizer is at $\lambda^* = \frac{\eta_n}{2b^2\|\mathbf{\Sigma}_c\|}$.
Finally, combining the exponents of $(a1)$ and $(a2)$ obtained in (C.269) and (C.273), we obtain the following large deviations upper bound characterization for the probability of false alarm for $\eta_n > \frac{2\sum_{n=1}^N M_n}{N}$,

$$\begin{aligned}&\limsup_{t\to\infty}\frac{1}{t}\log\left(\mathbb{P}_0\left(z_n(t)>\eta_n\right)\right)\\&\leq \max\left\{-\frac{\eta_n^2}{8b^2\left\|\mathbf{\Sigma}_c\right\|}, -LE(\lambda^*)\right\}.\end{aligned} \tag{C.274}$$

□

Appendix D

Proofs of Theorems in Chapter 5

Proof sketch of Theorem 5.4.3
The proof of almost sure convergence of the estimate sequence to $\boldsymbol{\theta}$ involves establishing the boundedness of the estimate sequence. With the boundedness of the estimate sequence in place, we show the convergence of the estimate sequence to its averaged estimate sequence $\{\mathbf{x}_{\text{avg}}(t)\}$, where $\mathbf{x}_{\text{avg}}(t) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n(t)$ at a rate faster $t^{1/2}$ and finally show that the averaged estimate sequence converges to $\boldsymbol{\theta}$ with a rate $\{(t+1)^{\tau}\}$ $\tau \in [0, 1/2)$. The final result follows by noting that, the averaged estimate sequence and the estimate sequence are indistinguishable in the $\{(t+1)^{\tau}\}$ time scale, where $\tau \in [0, 1/2)$.
Proof sketch of Theorem 5.4.5
The proof of the asymptotic normality of the estimate sequence proceeds in the following procedure. The first step involves establishing the asymptotic normality of the averaged estimate sequence $\mathbf{x}_{\text{avg}}(t)$. Moreover, an intermediate result ensures that the averaged estimate sequence and the estimate sequence are indistinguishable in the $\{(t+1)^{\frac{1}{2}}\}$ time scale. With the above development in place, it follows that the asymptotic normality of the averaged estimate sequence $\mathbf{x}_{\text{avg}}(t)$ can be extended to that of the estimate sequence $\{\mathbf{x}_n(t)\}$.

Lemma D.0.1. *For each n, the process $\{\mathbf{x}_n(t)\}$ satisfies*

$$\mathbb{P}_{\theta}\left(\sup_{t\geq 0}\|\mathbf{x}(t)\| < \infty\right) = 1. \tag{D.1}$$

Proof. We first note that,

$$\mathbf{L}(t) = \beta_t\overline{\mathbf{L}} + \widetilde{\mathbf{L}}(t), \tag{D.2}$$

where $\mathbb{E}\left[\widetilde{\mathbf{L}}(t)\right] = \mathbf{0}$ and $\mathbb{E}\left[\widetilde{\mathbf{L}}^2_{i,j}(t)\right] = \frac{c_4}{(t+1)^{\tau_1+\epsilon}} - \frac{c_3^2}{(t+1)^{2\tau_1}}$.
Define, $\mathbf{z}(t) = \mathbf{x}(t) - \mathbf{1}_N \otimes \boldsymbol{\theta}^*$ and $V(t) = \|\mathbf{z}(t)\|^2$. By conditional independence, we have that,

$$\begin{aligned}
&\mathbb{E}\left[V(t+1)|\mathcal{F}_t\right] = V(t)\\
&+ \mathbf{z}^{\top}(t)\left(\mathbf{I}_{NM} - \beta_t\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right) - \alpha_t\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^{\top}\right)^2\mathbf{z}(t)\\
&+ \mathbf{z}^{\top}(t)\mathbb{E}_{\boldsymbol{\theta}^*}\left[\left(\widetilde{\mathbf{L}}(t)\otimes\mathbf{I}_M\right)^2\right]\mathbf{z}(t)
\end{aligned}$$

$$+\alpha^2(t)\mathbb{E}_{\boldsymbol{\theta}^*}\left[\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}(t)-\mathbf{G}_H^\top\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right\|^2\right]$$
$$-2\mathbf{z}^\top(t)\left(\beta_t\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)+\alpha_t\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)\mathbf{z}(t), \tag{D.3}$$

where the filtration $\{\mathcal{F}_t\}$ may be taken to be the natural filtration generated by the random observations, the random Laplacians i.e.,

$$\mathcal{F}_t=\sigma\left(\left\{\{\mathbf{y}_n(s)\}_{n=1}^N,\{\mathbf{L}(s)\}\right\}_{s=0}^{t-1}\right), \tag{D.4}$$

which is the σ-algebra induced by the observation processes. For $t\geq t_1$, it can be shown that,

$$\mathbf{z}^\top(t)\left(\mathbf{I}_{NM}-\beta_t\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)-\alpha_t\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)^2\mathbf{z}(t)$$
$$\leq\left(1-c_4\alpha_t\right)^2\left\|\mathbf{z}(t)\right\|^2. \tag{D.5}$$

We use the following inequalities so as to analyze the recursion in (D.3).

$$\mathbf{z}^\top(t)\mathbb{E}_{\boldsymbol{\theta}^*}\left[\left(\widetilde{\mathbf{L}}(t)\otimes\mathbf{I}_M\right)^2\right]\mathbf{z}(t)\leq\frac{c_5\left\|\mathbf{z}_{\mathcal{C}^\perp}\right\|^2}{(t+1)^{\tau_1+\epsilon}}$$
$$\mathbb{E}_{\boldsymbol{\theta}^*}\left[\left\|\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}(t)-\mathbf{G}_H^\top\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right\|^2\right]\leq c_6$$
$$\mathbf{z}^\top(t)\left(\beta_t\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)+\alpha_t\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^\top\right)\mathbf{z}(t)$$
$$\geq\beta_t\lambda_2\left(\overline{\mathbf{L}}\right)\left\|\mathbf{z}_{\mathcal{C}^\perp}\right\|^2+c_7\alpha_t\left\|\mathbf{z}(t)\right\|^2. \tag{D.6}$$

Using the inequalities derived in (D.6), we have,

$$\mathbb{E}\left[V(t+1)|\mathcal{F}_t\right]\leq\left(1+c_8\alpha^2(t)\right)V(t)$$
$$-c_9\left(\beta_t-\frac{c_5}{(t+1)^{\tau_1+\epsilon}}\right)\left\|\mathbf{z}_{\mathcal{C}^\perp}\right\|^2+c_6\alpha^2(t). \tag{D.7}$$

As $\frac{c_5}{(t+1)^{\tau_1+\epsilon}}$ goes to zero faster than β_t, $\exists t_2$ such that $\forall t\geq t_2$, $\beta_t\geq\frac{c_5}{(t+1)^{\tau_1+\epsilon}}$. By the above construction we obtain $\forall t\geq t_2$,

$$\mathbb{E}_{\theta^*}[V(t+1)|\mathcal{F}_t]\leq(1+\alpha^2(t))V(t)+\widehat{\alpha}_t^2, \tag{D.8}$$

where $\widehat{\alpha}(t)=\sqrt{c_6}\alpha_t$. The product $\prod_{s=t}^{\infty}(1+\alpha_s^2)$ exists for all t. Now let $\{W(t)\}$ be such that

$$W(t)=\left(\prod_{s=t}^{\infty}(1+\alpha_s^2)\right)V_2(t)+\sum_{s=t}^{\infty}\widehat{\alpha}_s^2,\ \forall t\geq t_2. \tag{D.9}$$

By (D.9), it can be shown that $\{W(t)\}$ satisfies,

$$\mathbb{E}_{\theta^*}[W(t+1)|\mathcal{F}_t]\leq W(t). \tag{D.10}$$

Hence, $\{W(t)\}$ is a non-negative super martingale and converges a.s. to a bounded random variable W^* as $t\to\infty$. It then follows from (D.9) that $V(t)\to W^*$ as $t\to\infty$. Thus, we conclude that the sequences $\{\mathbf{x}_n(t)\}$ are bounded for all n. □

We now prove the almost sure convergence of the estimate sequence to the true parameter. In the sequel, we establish the order optimal convergence of the estimate sequence in the regime of $0 < \tau_1 < \frac{1}{2} - \frac{1}{2+\epsilon_1}$.

Lemma D.0.2. *Let the hypothesis of Theorem 5.4.3 hold. Then, we have,*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty} \mathbf{x}_n(t) = \boldsymbol{\theta}\right) = 1. \tag{D.11}$$

Proof of Lemma D.0.2. Following as in the proof of Lemma D.0.1, for t large enough

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)|\mathcal{F}_t] &\leq \left(1 - 2c_4\alpha_t + c_7\alpha_t^2\right) V(t) + c_6\alpha_t^2 \\ &\leq V(t) + c_6\alpha_t^2,\end{aligned} \tag{D.12}$$

as for t large enough, $-2c_4\alpha_t + c_7\alpha_t^2 < 0$. Now, consider the $\{\mathcal{F}_t\}$-adapted process $\{V_1(t)\}$ defined as follows

$$\begin{aligned}V_1(t) &= V(t) + c_6\sum_{s=t}^{\infty}\alpha_s^2 \\ &= V(t) + c_8\sum_{s=t}^{\infty}(t+1)^{-2},\end{aligned} \tag{D.13}$$

for appropriately chosen positive constant c_8.Since, $\{(t+1)^{-2}\}$ is summable, the process $\{V_1(t)\}$ is bounded from above. Moreover, it also follows that $\{V_1(t)\}_{t\geq t_1}$ is a supermartingale and hence converges a.s. to a finite random variable. By definition from (D.13), we also have that $\{V(t)\}$ converges to a non-negative finite random variable V^*. Finally, from (D.12), we have that,

$$\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)] \leq (1 - c_7\alpha_t)\,\mathbb{E}_{\boldsymbol{\theta}}[V(t)] + c_9(t+1)^{-2}, \tag{D.14}$$

for $t \geq t_1$. The sequence $\{V(t)\}$ then falls under the purview of Lemma C.3.1, and we have $\mathbb{E}_{\boldsymbol{\theta}}[V(t)] \to 0$ as $t \to \infty$. Finally, by Fatou's Lemma, where we use the non-negativity of the sequence $\{V(t)\}$, we conclude that

$$0 \leq \mathbb{E}_{\boldsymbol{\theta}}[V^*] \leq \liminf_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}}[V(t)] = 0, \tag{D.15}$$

which thus implies that $V^* = 0$ a.s. Hence, $\|\mathbf{z}(t)\| \to 0$ as $t \to \infty$ and the desired assertion follows. □

Consider the averaged estimate sequence, $\{\mathbf{x}_{\text{avg}}(t)\}\}$, which follows the following update:

$$\begin{aligned}\mathbf{x}_{\text{avg}}(t+1) &= \left(\mathbf{I}_M - \frac{\alpha_t}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\mathbf{H}_n\right)\mathbf{x}_{\text{avg}}(t) \\ &+ \frac{\alpha_t}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\left(\mathbf{x}_n(t) - \mathbf{x}_{\text{avg}}(t)\right) \\ &+ \frac{\alpha_t}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\gamma_n(t).\end{aligned} \tag{D.16}$$

The following Lemmas will be used to quantify the rate of convergence of distributed vector or matrix valued recursions to their network-averaged behavior.

Lemma D.0.3. *Let $\{z_t\}$ be an $\mathbb{R}^+$ valued $\mathcal{F}_t$-adapted process that satisfies*

$$z_{t+1} \leq (1 - r_1(t))\, z_t + r_2(t) U_t (1 + J_t),$$

where $\{r_1(t)\}$ is an $\mathcal{F}_{t+1}$-adapted process, such that for all t, $r_1(t)$ satisfies $0 \leq r_1(t) \leq 1$ and

$$a_1 \leq \mathbb{E}\left[r_1(t) | \mathcal{F}_t\right] \leq \frac{1}{(t+1)^{\delta_1}}$$

with $a_1 > 0$ and $0 \leq \delta_1 < 1$. The sequence $\{r_2(t)\}$ is deterministic and $\mathbb{R}^+$ valued and satisfies $r_2(t) \leq \frac{a_2}{(t+1)^{\delta_2}}$ with $a_2 > 0$ and $\delta_2 > 0$. Further, let $\{U_t\}$ and $\{J_t\}$ be $\mathbb{R}^+$ valued $\mathcal{F}_t$ and $\mathcal{F}_{t+1}$ adapted processes, respectively, with $\sup_{t \geq 0} \|U_t\| < \infty$ a.s. The process $\{J_t\}$ is i.i.d. with J_t independent of $\mathcal{F}_t$ for each t and satisfies the moment condition $\mathbb{E}\left[\|J_t\|^{2+\epsilon_1}\right] < \kappa < \infty$ for some $\epsilon_1 > 0$ and a constant $\kappa > 0$. Then, for every δ_0 such that $0 \leq \delta_0 < \delta_2 - \delta_1 - \frac{1}{2+\epsilon_1}$, we have $(t+1)^{\delta_0} z_t \to 0$ a.s. as $t \to \infty$.

Lemma D.0.4 (Lemma 4.1 in Kar et al. (2013a))**.** *Consider the scalar time-varying linear system*

$$u(t+1) \leq (1 - r_1(t)) u(t) + r_2(t), \tag{D.17}$$

where $\{r_1(t)\}$ is a sequence, such that

$$\frac{a_1}{(t+1)^{\delta_1}} \leq r_1(t) \leq 1 \tag{D.18}$$

with $a_1 > 0, 0 \leq \delta_1 \leq 1$, whereas the sequence $\{r_2(t)\}$ is given by

$$r_2(t) \leq \frac{a_2}{(t+1)^{\delta_2}} \tag{D.19}$$

with $a_2 > 0, \delta_2 \geq 0$. Then, if $u(0) \geq 0$ and $\delta_1 < \delta_2$, we have

$$\lim_{t \to \infty} (t+1)^{\delta_0} u(t) = 0, \tag{D.20}$$

for all $0 \leq \delta_0 < \delta_2 - \delta_1$. Also, if $\delta_1 = \delta_2$, then the sequence $\{u(t)\}$ stays bounded, i.e. $\sup_{t \geq 0} \|u(t)\| < \infty$.

Lemma D.0.5. *Let the Assumptions 5.3.1-5.4.3 hold. Consider the averaged estimate sequence as in (D.16). Then, we have,*

$$\mathbb{P}\left(\lim_{t \to \infty} (t+1)^{\frac{1}{2}+\delta} \left(\mathbf{x}(t) - \mathbf{1}_N \otimes \mathbf{x}_{avg}(t)\right) = 0\right) = 1 \tag{D.21}$$

Proof. Let $\mathcal{L}_t$ denote the set of possible Laplacian matrices (necessarily finite) at time t. Note, that the finiteness property of the cardinality of the set $\mathcal{L}_t$ holds for all t. Since the set of Laplacians is finite, we have,

$$\underline{p} = \inf_{\mathbf{L} \in \mathcal{L}_t} p_{\mathbf{L}} > 0, \tag{D.22}$$

with $p_L = \mathbb{P}\left(\mathbf{L}(t) = \mathbf{L}\right)$ for each $\mathbf{L} \in \mathcal{L}_t$ such that $\sum_{\mathbf{L} \in \mathcal{L}_t} p_{\mathbf{L}} = 1$. The connectedness of the network in an

average sense, i.e., $\lambda_2\left(\overline{\mathbf{L}}(t)\right) > 0$ implies that for every $\mathbf{z} \in \mathcal{C}^{\perp}$, where,

$$\mathcal{C} = \left\{\mathbf{x} | \mathbf{x} = \mathbf{1}_N \otimes \mathbf{a}, \mathbf{a} \in \mathbb{R}^M\right\}, \tag{D.23}$$

we have,

$$\sum_{\mathbf{L}\in\mathcal{L}_t} \mathbf{z}^{\top}\mathbf{L}\mathbf{z} \geq \sum_{\mathbf{L}\in\mathcal{L}_t} \mathbf{z}^{\top} p_{\mathbf{L}}\mathbf{L}\mathbf{z} = \mathbf{z}^{\top}\overline{\mathbf{L}}(t)\mathbf{z} \geq \lambda_2\left(\overline{\mathbf{L}}(t)\right)\|\mathbf{z}\|^2. \tag{D.24}$$

Owing to the finite cardinality of $\mathcal{L}_t$ and (D.24), we also have that for each $\mathbf{z} \in \mathcal{C}^{\perp}$,$\exists \mathbf{L}_{\mathbf{z}} \in \mathcal{L}_t$ such that,

$$\mathbf{z}^{\top}\mathbf{L}_{\mathbf{z}}\mathbf{z} \geq \frac{\lambda_2\left(\overline{\mathbf{L}}(t)\right)}{|\mathcal{L}_t|}\|\mathbf{z}\|^2 \tag{D.25}$$

Moreover, since $\mathcal{L}_t$ is finite, the mapping $L_{\mathbf{z}} : \mathcal{C}^{\perp} \mapsto \mathcal{L}_t$ can be realized as a measurable function. It is also to be noted that, $\mathbf{L}(t) = \rho_t^2\widehat{\mathbf{L}}$, where $\widehat{\mathbf{L}}$ is a Laplacian such that $[\widehat{\mathbf{L}}]_{ij} \in \mathbb{Z}$. For each, $\mathbf{L} \in \mathcal{L}_t$, the eigen values of $\mathbf{I}_{NM} - \rho_t^2\left(\widehat{\mathbf{L}} \otimes \mathbf{I}_M\right)$ are given by M repetitions of1 and $1 - \rho_t^2\lambda_n\left(\widehat{\mathbf{L}}\right)$, where $2 \leq n \leq N$. Thus, for $t \geq t_0$, $\left\|\mathbf{I}_{NM} - \rho_t^2\left(\widehat{\mathbf{L}} \otimes \mathbf{I}_M\right)\right\| \leq 1$ and $\left\|\left(\mathbf{I}_{NM} - \rho_t^2\left(\widehat{\mathbf{L}} \otimes \mathbf{I}_M\right)\right)\mathbf{z}\right\| \leq \|\mathbf{z}\|$. Hence, we can define a jointly measurable function $r_{\mathbf{L},\mathbf{z}}$ given by,

$$r_{\mathbf{L},\mathbf{z}} = \begin{cases} 1 & \text{if } t < t_0 \text{ or } \mathbf{z} = \mathbf{0} \\ 1 - \frac{\|(\mathbf{I}_{NM} - \rho_t^2(\widehat{\mathbf{L}}\otimes\mathbf{I}_M))\mathbf{z}\|}{\|\mathbf{z}\|} & \text{otherwise,} \end{cases} \tag{D.26}$$

which satisfies $0 \leq r_{\mathbf{L},\mathbf{z}} \leq 1$ for each $(\mathbf{L}, \mathbf{z})$. Define $\{r_t\}$ to be a $\mathcal{F}_{t+1}$ process given by, $r_t = r_{\mathbf{L}(t),\mathbf{z}_t}$ for each t and $\left\|\left(\mathbf{I}_{NM} - \rho_t^2\left(\widehat{\mathbf{L}} \otimes \mathbf{I}_M\right)\right)\mathbf{z}_t\right\| = (1 - r_t)\|\mathbf{z}_t\|$ a.s. for each t. Then, we have,

$$\begin{aligned}
&\left\|\left(\mathbf{I}_{NM} - \rho_t^2\left(\widehat{\mathbf{L}}_{\mathbf{z}_t} \otimes \mathbf{I}_M\right)\right)\mathbf{z}_t\right\|^2 \\
&= \mathbf{z}_t^{\top}\left(\mathbf{I}_{NM} - 2\rho_t^2\left(\widehat{\mathbf{L}} \otimes \mathbf{I}_M\right)\right)\mathbf{z}_t \\
&+ \mathbf{z}_t^{\top}\rho_t^4\left(\widehat{\mathbf{L}}_{\mathbf{z}_t} \otimes \mathbf{I}_M\right)^2\mathbf{z}_t \\
&\leq \left(1 - 2\beta_t\frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{|\mathcal{L}_t|}\right)\|\mathbf{z}_t\|^2 + c_1\rho_t^4\|\mathbf{z}_t\|^2 \\
&\leq \left(1 - \beta_t\frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{|\mathcal{L}_t|}\right)\|\mathbf{z}_t\|^2
\end{aligned} \tag{D.27}$$

where we have used the boundedness of the Laplacian matrix and the fact that $\overline{\mathbf{L}}_t = \beta_t\overline{\mathbf{L}}$. With the above development in place, choosing an appropriate t_1 (making t_0 larger if necessary), for all $t \geq t_1$, we have,

$$\left\|\left(\mathbf{I}_{NM} - \rho_t^2\left(\widehat{\mathbf{L}}_{\mathbf{z}_t} \otimes \mathbf{I}_M\right)\right)\mathbf{z}_t\right\| \leq \left(1 - \beta_t\frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{4|\mathcal{L}_t|}\right)\|\mathbf{z}_t\|^2. \tag{D.28}$$

Then, from (D.28), we have,

$$\mathbb{E}\left[\left\|\left(\mathbf{I}_{NM} - \rho_t^2\left(\widehat{\mathbf{L}}_{\mathbf{z}_t} \otimes \mathbf{I}_M\right)\right)\mathbf{z}_t\right\| \middle| \mathcal{F}_t\right]$$

$$
\begin{aligned}
&= \sum_{\mathbf{L} \in \mathcal{L}_t} p_{\mathbf{L}} \left(1 - r_{\mathbf{L}, \mathbf{z}_t}\right) \|\mathbf{z}_t\| \\
&\leq \left(1 - \left(\underline{p} \beta_t \frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{4|\mathcal{L}_t|} + \sum_{\mathbf{L} \neq \mathbf{L}_{\mathbf{z}_t}}\right)\right) \|\mathbf{z}_t\| . \qquad \text{(D.29)}
\end{aligned}
$$

Since, $\sum_{\mathbf{L} \neq \mathbf{L}_{\mathbf{z}_t}} p_{\mathbf{L}} r_{\mathbf{L}, \mathbf{z}_t} \geq 0$, we have for all $t \geq t_1$,

$$
\begin{aligned}
&\left(1 - \mathbb{E}\left[r_t | \mathcal{F}_t\right]\right) \|\mathbf{z}_t\| \\
&= \mathbb{E}\left[\left\| \left(\mathbf{I}_{NM} - \rho_t^2 \left(\widehat{\mathbf{L}}_{\mathbf{z}_t} \otimes \mathbf{I}_M\right)\right) \mathbf{z}_t \right\| \Big| \mathcal{F}_t\right] \\
&\leq \left(1 - \underline{p} \beta_t \frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{4|\mathcal{L}_t|}\right) \|\mathbf{z}_t\| . \qquad \text{(D.30)}
\end{aligned}
$$

As $r_t = 1$ on the set $\{\mathbf{z}_t = 0\}$, we have that,

$$
\mathbb{E}\left[r_t | \mathcal{F}_t\right] \geq \underline{p} \beta_t \frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{4|\mathcal{L}_t|}. \qquad \text{(D.31)}
$$

Thus, we have established that,

$$
\left\|\left(\mathbf{I}_{NM} - \left(\mathbf{L}(t) \otimes \mathbf{I}_M\right)\right) \mathbf{z}_t\right\| \leq \left(1 - r_t\right) \|\mathbf{z}_t\| , \qquad \text{(D.32)}
$$

where $\{r_t\}$ is a $\mathbb{R}^+$ valued $\mathcal{F}_{t+1}$ process satisfying (D.31). With the above development in place, consider the residual process $\{\widetilde{\mathbf{x}}(t)\}$ given by $\widetilde{\mathbf{x}}(t) = \mathbf{x}(t) - \mathbf{x}_{\text{avg}}(t)$. Thus, we have that the process $\{\widetilde{\mathbf{x}}(t)\}$ satisfies the recursion,

$$
\widetilde{\mathbf{x}}(t+1) = \left(\mathbf{I}_{NM} - \mathbf{L}(t) \otimes \mathbf{I}_M\right) \widetilde{\mathbf{x}}(t) + \alpha_t \widetilde{\mathbf{z}}(t), \qquad \text{(D.33)}
$$

where the process $\{\widetilde{\mathbf{z}}(t)\}$ is given by

$$
\begin{aligned}
\widetilde{\mathbf{z}}(t) &= \left(\mathbf{I}_{NM} - \frac{1}{N} \mathbf{1}_N \otimes \left(\mathbf{1}_N \otimes \mathbf{I}_M\right)^\top\right) \\
&\times \mathbf{G}_H \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}(t) - \mathbf{G}_H^\top \mathbf{x}(t)\right) . \qquad \text{(D.34)}
\end{aligned}
$$

From (D.34), we also have,

$$
\widetilde{\mathbf{z}}(t) = \overline{\mathbf{J}}_t + \overline{\mathbf{U}}_t, \qquad \text{(D.35)}
$$

where,

$$
\begin{aligned}
\overline{\mathbf{J}}_t &= \left(\mathbf{I}_{NM} - \frac{1}{N} \mathbf{1}_N \otimes \left(\mathbf{1}_N \otimes \mathbf{I}_M\right)^\top\right) \\
&\times \mathbf{G}_H \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}(t) - \mathbf{G}_H^\top \boldsymbol{\theta}\right) \\
\overline{\mathbf{U}}_t &= \left(\mathbf{I}_{NM} - \frac{1}{N} \mathbf{1}_N \otimes \left(\mathbf{1}_N \otimes \mathbf{I}_M\right)^\top\right) \\
&\times \mathbf{G}_H \boldsymbol{\Sigma}^{-1} \left(\mathbf{G}_H^\top \boldsymbol{\theta} - \mathbf{G}_H^\top \mathbf{x}(t)\right) . \qquad \text{(D.36)}
\end{aligned}
$$

By Lemma D.0.1, we also have that, the process $\{\mathbf{x}(t)\}$ is bounded. Hence, there exists an $\mathcal{F}_t$-adapted process $\{\widetilde{U}_t\}$ such that $\left\|\overline{\mathbf{U}}_t\right\| \leq \widetilde{U}_t$ and $\sup_{t\geq 0}\widetilde{U}_t < \infty$ a.s.. Furthermore, denote the process U_t as follows,

$$U_t = \max\left\{\widetilde{U}_t, \left\|\mathbf{I}_{NM} - \frac{1}{N}\mathbf{1}_N \otimes (\mathbf{1}_N \otimes \mathbf{I}_M)^\top\right\|\right\}. \tag{D.37}$$

With the above development in place, we conclude,

$$\left\|\overline{\mathbf{U}}_t\right\| + \left\|\overline{\mathbf{J}}_t\right\| \leq U_t\left(1 + J_t\right), \tag{D.38}$$

where $J_t = \mathbf{y}(t) - \mathbf{G}_H^\top \boldsymbol{\theta}$. Then, from (D.32)-(D.33) and noting that $\widetilde{\mathbf{x}}(t) \in \mathcal{C}^\perp$, we have,

$$\|\widetilde{\mathbf{x}}(t+1)\| \leq (1 - r_t)\|\widetilde{\mathbf{x}}(t)\| + \alpha_t U_t(1 + J_t), \tag{D.39}$$

which then falls under the purview of Lemma D.0.3 and hence we have the assertion,

$$\mathbb{P}\left(\lim_{t\to\infty}(t+1)^{\delta_0}\left(\mathbf{x}(t) - \mathbf{1}_N \otimes \mathbf{x}_{avg}(t)\right) = 0\right) = 1, \tag{D.40}$$

where $0 < \delta_0 < 1 - \tau_1$ and hence δ_0 can be chosen to be $1/2 + \delta$, where $\delta > 0$ and we finally have,

$$\mathbb{P}\left(\lim_{t\to\infty}(t+1)^{\frac{1}{2}+\delta}\left(\mathbf{x}(t) - \mathbf{1}_N \otimes \mathbf{x}_{avg}(t)\right) = 0\right) = 1. \tag{D.41}$$

□

Lemma D.0.6. *Let the Assumptions 5.3.1-5.4.3 hold. Consider the averaged estimate sequence as in (D.16). Then, we have,*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}\mathbf{x}_{avg}(t) = \boldsymbol{\theta}\right) = 1. \tag{D.42}$$

Proof. Define, the residual sequence, $\{\mathbf{z}_t\}$, where $\mathbf{z}(t) = \mathbf{x}_{\text{avg}}(t) - \boldsymbol{\theta}$, which can be then shown to satisfy the recursion

$$\mathbf{z}_{t+1} = \left(\mathbf{I}_M - \alpha_t\Gamma\right)\mathbf{z}_t + \alpha_t\mathbf{U}_t + \alpha_t\mathbf{J}_t, \tag{D.43}$$

where

$$\begin{aligned}
\Gamma &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\mathbf{H}_n \\
\mathbf{U}_t &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\left(\mathbf{x}_n(t) - \mathbf{x}_{\text{avg}}(t)\right) \\
\mathbf{J}_t &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\gamma_n(t).
\end{aligned} \tag{D.44}$$

From, Lemma D.0.5, we have that,

$$\mathbb{P}\left(\lim_{t\to\infty}(t+1)^{\delta_0}\left(\mathbf{x}(t) - \mathbf{1}_N \otimes \mathbf{x}_{avg}(t)\right) = 0\right) = 1, \tag{D.45}$$

where $0 < \delta_0 < 1 - \tau_1$. Fix, a δ_0 and then by convergence of $(t+1)^{\delta_0}\mathbf{U}_t \to 0$ a.s. as $t \to \infty$ and Egorov's theorem, the a.s. convergence may be assumed to be uniform on sets of arbitrarily large probability measure and hence for every $\delta > 0$, there exists uniformly bounded process $\{\mathbf{U}_t^\delta\}$ satisfying,

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\sup_{s \geq t_\epsilon^\delta}(s+1)^{\delta_0}\left\|\mathbf{U}_t^\delta\right\| > \epsilon\right) = 0, \tag{D.46}$$

for each $\epsilon > 0$ and some t_ϵ^δ chosen appropriately large enough such that

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\sup_{t \geq 0}\left\|\mathbf{U}_t^\delta - \mathbf{U}_t\right\| = 0\right) > 1 - \delta. \tag{D.47}$$

With the above development in the place, for each $\delta > 0$, define the $\mathcal{F}_t$-adapted process $\{\mathbf{z}_t^\delta\}$ which satisfies the recursion

$$\mathbf{z}_{t+1}^\delta = (\mathbf{I}_M - \alpha_t\Gamma)\,\mathbf{z}_t^\delta + \alpha_t\mathbf{U}_t^\delta + \alpha_t\mathbf{J}_t, \mathbf{z}_0^\delta = \mathbf{z}_0, \tag{D.48}$$

and

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\sup_{t \geq 0}\left\|\mathbf{z}_t^\delta - \mathbf{z}_t\right\| = 0\right) > 1 - \delta. \tag{D.49}$$

It is to be noted that, in order to show that $\mathbf{z}_t \to 0$ as $t \to \infty$, it suffices to show that $\mathbf{z}_t^\delta \to 0$ for each $\delta > 0$. We now focus on the process $\{\mathbf{z}_t^\delta\}$ for a fixed but arbitrary $\delta > 0$. Let $\{V_t^\delta\}$ denote the $\mathcal{F}_t$-adapted process such that $V_t^\delta = \left\|\mathbf{z}_t^\delta\right\|^2$. Then, we have,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[V_{t+1}^\delta\right] &\leq \left\|\mathbf{I}_M - \alpha_t\boldsymbol{\Gamma}\right\|^2 V_t^\delta + 2\alpha_t\left(\mathbf{U}_t^\delta\right)^\top\left(\mathbf{I}_M - \alpha_t\boldsymbol{\Gamma}\right)\mathbf{z}_t^\delta \\ &+ \alpha_t^2\left\|\mathbf{U}_t^\delta\right\|^2 + \alpha_t^2\,\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\mathbf{J}_t\right\|^2\middle|\mathcal{F}_t\right].\end{aligned} \tag{D.50}$$

For large enough t, we have,

$$\begin{aligned}&\left\|2\alpha_t\left(\mathbf{U}_t^\delta\right)^\top\left(\mathbf{I}_M - \alpha_t\boldsymbol{\Gamma}\right)\mathbf{z}_t^\delta\right\| \leq 2\alpha_t\left\|\mathbf{U}_t^\delta\right\|\left\|\mathbf{z}_t^\delta\right\| \\ &\leq 2\alpha_t\left\|\mathbf{U}_t^\delta\right\|\left\|\mathbf{z}_t^\delta\right\|^2 + 2\alpha_t\left\|\mathbf{U}_t^\delta\right\|.\end{aligned} \tag{D.51}$$

We note that $\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\mathbf{J}_t\right\|^2\middle|\mathcal{F}_t\right]$ is bounded and making t_ϵ^δ larger if necessary in order to ensure $\left\|\mathbf{U}_t^\delta\right\| \leq \epsilon(t+1)^{-\delta_0}$, it follows that $\exists c_1, c_2$ such that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[V_{t+1}^\delta\right] &\leq \left(1 - c_1\alpha_t + c_2\alpha_t(t+1)^{-\delta_0}\right)V_t^\delta \\ &+ c_2\left(\alpha_t(t+1)^{-\delta_0} + \alpha_t^2(t+1)^{-2\delta_0} + \alpha_t^2\right) \\ &\leq (1 - c_3\alpha_t)\,V_t^\delta + c_4\alpha_t(t+1)^{-\delta_0} \leq V_t^\delta + c_4\alpha_t(t+1)^{-\delta_0}\end{aligned} \tag{D.52}$$

which is ensured by making $c_4 > c_2$ and $c_3 < c_1$ respectively. As the process $\{\alpha_t(t+1)^{-\delta_0}\}$ is summable, the process $\{\overline{V}_t^\delta\}$ given by,

$$\overline{V}_t^\delta = V_t^\delta + c_4\sum_{s=t}^{\infty}\alpha_s(s+1)^{-\delta_0}, \tag{D.53}$$

is bounded from above. Thus, we have that $\{\overline{V}_t^{\delta}\}_{t\geq t_\epsilon^\delta}$ is a supermartingale and hence converges to a finite random variable. From (D.53), we have that the process $\{V_t^\delta\}$ converges to a finite random variable V^δ. We also have from (D.52), for $t \geq t_\epsilon^\delta$

$$\mathbb{E}_{\boldsymbol{\theta}}\left[V_{t+1}^\delta\right] \leq (1 - c_3\alpha_t)\,\mathbb{E}_{\boldsymbol{\theta}}\left[V_t^\delta\right] + c_4\alpha_t(t+1)^{-\delta_0}.. \tag{D.54}$$

Since $\delta_0 > 0$, the recursion in (D.54) falls under the purview of Lemma D.0.4 and thus we have, $\mathbb{E}_{\boldsymbol{\theta}}\left[V_t^\delta\right] \to 0$ as $t \to \infty$. The sequence $\{V_t^\delta\}$ is non-negative, so by Fatou's Lemma, we have,

$$0 \leq \mathbb{E}_{\boldsymbol{\theta}}\left[V^\delta\right] \leq \liminf_{t\to\infty} \mathbb{E}_{\boldsymbol{\theta}}\left[V_t^\delta\right] = 0. \tag{D.55}$$

Hence $V^\delta = 0$ a.s. and thus $\left\|\mathbf{z}_t^\delta\right\| \to 0$ as $t \to \infty$ and the assertion follows. □

We will use the approximation result (Lemma C.1.3) and the generalized convergence criterion (Lemma C.1.4) for the proof of Theorem 5.4.3. Lemma D.0.6 establishes the almost sure convergence of the averaged estimate sequence $\{\mathbf{x}_{\text{avg}}(t)\}$ to the true underlying parameter. We now establish the order optimal convergence of the estimate sequence in terms of t.

Proof of Theorem 5.4.3. We first analyze the rate of convergence of the process $\{\mathbf{z}_t^\delta\}$ as developed in Lemma D.0.6 and note that the rate of convergence of the process $\{\mathbf{z}_t^\delta\}$ suffices for the rate of convergence of the process $\{\mathbf{z}_t\}$. For each $\delta > 0$, recall the process $\{\mathbf{z}_t^\delta\}$ as in (C.199)-(D.48). Let $\overline{\tau} \in [0, 1/2)$ be such that,

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}(t+1)^{\overline{\tau}}\left\|\mathbf{z}_t^\delta\right\| = 0\right) = 1. \tag{D.56}$$

It is to be noted that such a $\overline{\tau}$ always exists from Lemma D.0.6. We now focus on showing that there exists τ such that $\overline{\tau} < \tau < 1/2$ for which the assertion holds. Define $\widetilde{\tau} \in (\tau, 1/2)$ and $\mu = \frac{1}{2}(\overline{\tau} + \widetilde{\tau})$. Then, for each $\delta > 0$,

$$\begin{aligned}
\left\|\mathbf{z}_{t+1}^\delta\right\|^2 &\leq \left\|\mathbf{I}_M - \alpha_t\boldsymbol{\Gamma}\right\|^2 \left\|\mathbf{z}_t^\delta\right\|^2 + \alpha_t^2\left\|\mathbf{U}_t^\delta\right\|^2 \\
&+ \alpha_t^2\left\|\mathbf{J}_t\right\|^2 \\
&+ 2\alpha_t\left(\mathbf{z}_t^\delta\right)^\top\left(\mathbf{I}_M - \alpha_t\boldsymbol{\Gamma}\right)\mathbf{J}_t \\
&+ 2\alpha_t\left\|\mathbf{U}_t^\delta\right\|\left(\left\|\mathbf{I}_M - \alpha_t\boldsymbol{\Gamma}\right\|\left\|\mathbf{z}_t^\delta\right\| + \alpha_t\left\|\mathbf{J}_t\right\|\right).
\end{aligned} \tag{D.57}$$

We have that, $1 > \tau_1 + \frac{1}{2+\epsilon_1} + \frac{1}{2}$, hence the process $\{\mathbf{U}_t^\delta\}$ may be chosen such that, $\left\|\mathbf{U}_t^\delta\right\| = o\left((t+1)^{-1/2}\right)$. Moreover, as $\left\|\mathbf{z}_t^\delta\right\| = o\left((t+1)^{-\overline{\tau}}\right)$, we have,

$$2\alpha_t\left\|\mathbf{U}_t^\delta\right\|\left\|\mathbf{I}_M - \alpha_t\boldsymbol{\Gamma}\right\|\left\|\mathbf{z}_t^\delta\right\| = o\left((t+1)^{-3/2-\overline{\tau}}\right). \tag{D.58}$$

From Assumption 5.3.1, we have that,

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}(t+1)^{-1/2-\epsilon}\left\|\mathbf{J}_t^\delta\right\|\right) = 1, \text{ for each } \epsilon > 0, \tag{D.59}$$

and hence we conclude that

$$2\alpha_t^2\left\|\mathbf{U}_t^\delta\right\|\left\|\mathbf{J}_t\right\| = o\left((t+1)^{-3/2-\overline{\tau}}\right). \tag{D.60}$$

Since, $2\mu = \overline{\tau} + \widetilde{\tau}$ and $\widetilde{\tau} < 1/2$, we have the following conclusions

$$\begin{aligned}
&\sum_{t\geq 0}(t+1)^{2\mu}\alpha_t \left\|\mathbf{U}_t^{\delta}\right\| \left\|\mathbf{I}_M - \alpha_t\boldsymbol{\Gamma}\right\| \left\|\mathbf{z}_t^{\delta}\right\| < \infty \\
&\sum_{t\geq 0}(t+1)^{2\mu}\alpha_t^2 \left\|\mathbf{U}_t^{\delta}\right\| \left\|\mathbf{J}_t\right\| < \infty \\
&\sum_{t\geq 0}(t+1)^{2\mu}\alpha_t^2 \left\|\mathbf{U}_t^{\delta}\right\|^2 < \infty \\
&\sum_{t\geq 0}(t+1)^{2\mu}\alpha_t^2 \left\|\mathbf{J}_t\right\|^2 < \infty.
\end{aligned} \tag{D.61}$$

With the above development in place, let $\{W_t^{\delta}\}$ denote the $\mathcal{F}_{t+1}$-adapted sequence given by

$$W_t^{\delta} = \alpha_t \left(\mathbf{z}_t^{\delta}\right)^{\top} \left(\mathbf{I} - \alpha_t\boldsymbol{\Gamma}\right) \mathbf{J}_t, \tag{D.62}$$

where $\mathbb{E}_{\boldsymbol{\theta}}\left[W_t^{\delta}\middle|\mathcal{F}_t\right] = 0$ and for t chosen sufficiently large, we have that,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}}\left[\left(W_t^{\delta}\right)^2\middle|\mathcal{F}_t\right] = o\left((t+1)^{-2-2\overline{\tau}}\right) \\
&\Rightarrow \mathbb{E}_{\boldsymbol{\theta}}\left[(t+1)^{4\mu}\left(W_t^{\delta}\right)^2\middle|\mathcal{F}_t\right] = o\left((t+1)^{-2-2\overline{\tau}+4\mu}\right) \\
&= o\left((t+1)^{-2+2\widetilde{\tau}}\right).
\end{aligned} \tag{D.63}$$

Since, $2\widetilde{\tau} < 1$, the sequence $\mathbb{E}_{\boldsymbol{\theta}}\left[(t+1)^{4\mu}\left(W_t^{\delta}\right)^2\middle|\mathcal{F}_t\right]$ is summable and by Lemma C.1.4, $\sum_{t\geq 0}(t+1)^{2\mu}W_t^{\delta}$ exists. It may be shown that as $\alpha_t \to 0$ as $t \to \infty$,

$$\left\|\mathbf{I} - \alpha_t\boldsymbol{\Gamma}\right\|^2 \leq 1 - c_1\alpha_t, \tag{D.64}$$

where $c_1 = \lambda_{min}\left(\boldsymbol{\Gamma}\right)$. Then, from (D.57), we have,

$$\left\|\mathbf{z}_{t+1}^{\delta}\right\|^2 \leq (1 - c_1\alpha_t)\left\|\mathbf{z}_t^{\delta}\right\|^2 + d_t(t+1)^{-2\mu}, \tag{D.65}$$

where the term $d_t(t+1)^{-2\mu}$ represents all the residual terms in (D.57). The fact that $\lim_{t\to\infty}\sum_{s=0}^{t} d_s$ exists and is finite in conjunction with $c_1\alpha_t(t+1) \geq 1 \geq 2\mu$ (from Assumption 5.4.3) brings (D.65) under the purview of Lemma C.1.3 and yields

$$\limsup_{t\to\infty}(t+1)^{2\mu}\left\|\mathbf{z}_t^{\delta}\right\|^2 < \infty \ a.s., \tag{D.66}$$

which leads to the conclusion that there exists τ with $\overline{\tau} < \tau < \mu$, such that $(t+1)^{\tau}\left\|\mathbf{z}_t^{\delta}\right\| \to 0$ as $t \to \infty$. The fact that the above development holds for all $\delta > 0$, we conclude that $(t+1)^{\tau}\left\|\mathbf{z}_t\right\| \to 0$ as $t \to \infty$. Hence, for every $\overline{\tau}$ for which

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}(t+1)^{\overline{\tau}}\|\mathbf{x}_{\text{avg}}(t) - \boldsymbol{\theta}\| = 0\right) = 1 \tag{D.67}$$

holds, then there exists $\tau \in (\overline{\tau}, 1/2)$ for which the convergence continues to hold. Finally, an application of

induction yields the result

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}(t+1)^{\tau}\|\mathbf{x}_{\text{avg}}(t)-\boldsymbol{\theta}\|=0\right)=1,\forall\tau\in[0,1/2) \tag{D.68}$$

The above result in conjunction with Lemma D.0.5 and the usage of triangle inequality yields $\forall\tau\in[0,1/2)$

$$\begin{aligned}&(t+1)^{\tau}\left\|\mathbf{x}_n(t)-\boldsymbol{\theta}\right\|\leq(t+1)^{\tau}\left\|\mathbf{x}_{\text{avg}}(t)-\boldsymbol{\theta}\right\|\\&+(t+1)^{\tau}\left\|\mathbf{x}_n(t)-\mathbf{x}_{\text{avg}}(t)\right\|\\&\Rightarrow\lim_{t\to\infty}(t+1)^{\tau}\left\|\mathbf{x}_n(t)-\boldsymbol{\theta}\right\|=0\ a.s.\end{aligned} \tag{D.69}$$

□

Proof of Theorem 5.4.4. Proceeding as in proof of Lemma D.0.2, we have, for t large enough

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)|\mathcal{F}_t]\leq\left(1-2c_4\alpha_t+c_7\alpha_t^2\right)V(t)+c_6\alpha_t^2\\&\leq V(t)+c_6\alpha_t^2,\end{aligned} \tag{D.70}$$

as for t large enough, $-c_4\alpha_t+c_7\alpha_t^2<0$. Before proceeding further, we note that, from (D.5),

$$\begin{aligned}&\mathbf{x}^{\top}\left(\beta_t\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)+\alpha_t\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^{\top}\right)\mathbf{x}\\&=\alpha_t\mathbf{x}^{\top}\left(\frac{\beta_t}{\alpha_t}\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)+\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^{\top}\right)\mathbf{x}\\&\geq\alpha_t\mathbf{x}^{\top}\left(\left(\mathbf{L}\otimes\mathbf{I}_M\right)+\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^{\top}\right)\mathbf{x}\geq c_4\alpha_t,\end{aligned} \tag{D.71}$$

where

$$c_4=\lambda_{\min}\left(\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)+\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^{\top}\right). \tag{D.72}$$

Thus, we have that

$$\left\|\mathbf{I}_{NM}-\beta_t\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)-\alpha_t\mathbf{G}_H\boldsymbol{\Sigma}^{-1}\mathbf{G}_H^{\top}\right\|\leq1-c_4\alpha_t, \tag{D.73}$$

for all $t\geq t_1$, where t_1 is chosen to be appropriately large. Now, consider the $\{\mathcal{F}_t\}$-adapted process $\{V_1(t)\}$ defined as follows

$$\begin{aligned}V_1(t)&=V(t)+c_6\sum_{s=t}^{\infty}\alpha_s^2\\&=V(t)+c_8\sum_{s=t}^{\infty}(t+1)^{-2},\end{aligned} \tag{D.74}$$

for appropriately chosen positive constant c_8.Since, $\{(t+1)^{-2}\}$ is summable, the process $\{V_1(t)\}$ is bounded from above. Moreover, it also follows that $\{V_1(t)\}_{t\geq t_1}$ is a supermartingale and hence converges a.s. to a finite random variable. By definition from (D.13), we also have that $\{V(t)\}$ converges to a non-negative finite random variable V^*. Finally, from (D.70), we have that,

$$\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)]\leq(1-c_4\alpha_t)\,\mathbb{E}_{\boldsymbol{\theta}}[V(t)]+c_8(t+1)^{-2}$$

$$\Rightarrow E_{\boldsymbol{\theta}}[V(t+1)] \leq (1 - c_4\alpha_t)\,\mathbb{E}_{\boldsymbol{\theta}}[V(t)] + c_{10}\alpha_t(t+1)^{-1} \tag{D.75}$$

for $t \geq t_1$. The summability of $\{\alpha_t\}$ in conjunction with assumption 5.4.3 ensures that the sequence $\{V(t)\}$ then falls under the purview of Lemma C.1.3, and we have

$$\begin{aligned}&\limsup_{t\to\infty}(t+1)\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)] < \infty\\ &\Rightarrow \mathbb{E}_{\boldsymbol{\theta}}[V(t)] = O\left(\frac{1}{t}\right).\end{aligned} \tag{D.76}$$

Furthermore, from (D.74), we also have that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}[V_1(t)] \leq \mathbb{E}_{\boldsymbol{\theta}}[V(t)] + \frac{c_6\pi^2}{6}\\ &\Rightarrow \mathbb{E}_{\boldsymbol{\theta}}[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2] = O\left(\frac{1}{t}\right).\end{aligned} \tag{D.77}$$

It is to be noted that the communication cost $\mathcal{C}_t$ for the proposed $\mathcal{CREDO}$ algorithm, is given by $\mathcal{C}_t = \Theta\left(t^{1+\frac{\epsilon-\tau_1}{2}}\right)$ and thus the assertion follows in conjunction with (D.77). □

D.0.1 Asymptotic Normality and Covariance

The proof of Theorem 5.4.5 needs Lemma C.1.5 from Fabian (1968) concerning the asymptotic normality of the stochastic recursions. In order to establish asymptotic normality and characterize the estimator in terms of asymptotic covariance, the following Lemma plays a crucial role.

Proof of Theorem 5.4.5. We invoke the definition of the process $\{\mathbf{z}_t\}$ as defined in (C.199)-(D.44). We rewrite the recursion for $\{\mathbf{z}_t\}$ as follows:

$$\mathbf{z}_{t+1} = (\mathbf{I}_M - \alpha_t\Gamma_t)\,\mathbf{z}_t + (t+1)^{-3/2}\mathbf{T}_t + (t+1)^{-1}\boldsymbol{\Phi}_t\mathbf{V}_t, \tag{D.78}$$

where

$$\begin{aligned}&\boldsymbol{\Gamma}_t = \boldsymbol{\Gamma} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\mathbf{H}_n\\ &\mathbf{T}_t = a(t+1)^{1/2}\mathbf{U}_t\\ &= \frac{a}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}(t+1)^{1/2}\left(\mathbf{x}_n(t) - \mathbf{x}_{\text{avg}}(t)\right) \to 0,\ t\to\infty\\ &\boldsymbol{\Phi}_t = a\mathbf{I}\\ &\mathbf{V}_t = \mathbf{J}_t = \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\gamma_n(t),\ \mathbb{E}\left[\mathbf{V}_t|\mathcal{F}_t\right] = 0,\\ &\mathbb{E}\left[\mathbf{V}_t\mathbf{V}_t^\top|\mathcal{F}_t\right] = \frac{1}{N^2}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\mathbf{H}_n,\end{aligned} \tag{D.79}$$

and the convergence of $\mathbf{T}_t$ follows from Lemma D.0.5. Due to the i.i.d nature of the noise process, we have the uniform integrability condition for the process $\{\mathbf{V}_t\}$. Hence, $\{\mathbf{x}_{\text{avg}}(t)\}$ falls under the purview of Lemma

C.1.5 and we thus conclude that

$$(t+1)^{1/2}\left(\mathbf{x}_{\text{avg}}(t)-\boldsymbol{\theta}\right) \overset{\mathcal{D}}{\Longrightarrow} \mathcal{N}(0, \mathbf{P}\mathbf{M}\mathbf{P}^{\top}), \tag{D.80}$$

where

$$\begin{aligned}
&a\mathbf{P}^{\top}\boldsymbol{\Gamma}\mathbf{P} = a\boldsymbol{\Lambda},\\
&[\mathbf{M}]_{ij} = \left[a^2\mathbf{P}^{\top}\boldsymbol{\Phi}\left(\frac{1}{N^2}\sum_{n=1}^{N}\mathbf{H}_n^{\top}\boldsymbol{\Sigma}_n^{-1}\mathbf{H}_n\right)\boldsymbol{\Phi}^{\top}\mathbf{P}\right]_{ij}\\
&\times\left(a\left[\boldsymbol{\Lambda}\right]_{ii}+a\left[\boldsymbol{\Lambda}\right]_{jj}-1\right)^{-1}\\
&=\frac{a^2}{N}\left[\boldsymbol{\Lambda}\right]_{ij}\left(a\left[\boldsymbol{\Lambda}\right]_{ii}+a\left[\boldsymbol{\Lambda}\right]_{jj}-1\right)^{-1},
\end{aligned} \tag{D.81}$$

which also implies that $\mathbf{M}$ is a diagonal matrix with its i-th diagonal element given by $\frac{a^2\boldsymbol{\Lambda}_{ii}}{2aN\boldsymbol{\Lambda}_{ii}-N}$. Note that, Assumption 5.4.3 ensures that $\frac{a^2\boldsymbol{\Lambda}_{ii}}{2aN\boldsymbol{\Lambda}_{ii}-N} > 0$, $\forall i$. We already have that $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\top} = \boldsymbol{\Gamma}$. Hence, the matrix with eigenvalues as $\frac{a^2\boldsymbol{\Lambda}_{ii}}{2aN\boldsymbol{\Lambda}_{ii}-N}$ is given by

$$\mathbf{P}\mathbf{M}\mathbf{P}^{\top} = \frac{a\mathbf{I}}{2N} + \frac{\left(\boldsymbol{\Gamma}-\frac{\mathbf{I}}{2a}\right)^{-1}}{4N}. \tag{D.82}$$

Now from Lemma D.0.5, we have that the processes $\{\mathbf{x}_n(t)\}$ and $\{\mathbf{x}_{\text{avg}}(t)\}$ are indistinguishable in the $(t+1)^{1/2}$ time scale, which is formalized as follows:

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}\left\|\sqrt{t+1}\left(\mathbf{x}_n(t)-\boldsymbol{\theta}\right)-\sqrt{t+1}\left(\mathbf{x}_{\text{avg}}(t)-\boldsymbol{\theta}\right)\right\|=0\right)\\
&=\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}\left\|\sqrt{t+1}\left(\mathbf{x}_n(t)-\mathbf{x}_{\text{avg}}(t)\right)\right\|=0\right)=1.
\end{aligned} \tag{D.83}$$

Thus, the difference of the sequences $\left\{\sqrt{t+1}\left(\mathbf{x}_n(t)-\boldsymbol{\theta}\right)\right\}$ and $\left\{\sqrt{t+1}\left(\mathbf{x}_{\text{avg}}(t)-\boldsymbol{\theta}\right)\right\}$ converges a.s. to zero as $t\to\infty$ and hence we have,

$$\sqrt{t+1}\left(\mathbf{x}_n(t)-\boldsymbol{\theta}\right) \overset{\mathcal{D}}{\Longrightarrow} \mathcal{N}\left(0, \frac{a\mathbf{I}}{2N}+\frac{\left(\boldsymbol{\Gamma}-\frac{\mathbf{I}}{2a}\right)^{-1}}{4N}\right). \tag{D.84}$$

□

Appendix E

Proofs of Theorems in Chapter 6

In this section, we provide the proofs of Theorems 6.5.1 and 6.5.2.

E.0.1 Proof of Theorem 6.5.1

Proof. The proof of Theorem 6.5.1 is accomplished in three steps. First, we establish the boundedness of the estimate sequence followed by proving the strong consistency of the estimate sequence $\{\mathbf{x}_n(t)\}$ and then in the sequel we establish the rate of convergence of the estimate sequence to the true underlying parameter. We follow the basic idea developed in Kar and Moura (2014).

Lemma E.0.1. *Let the hypothesis of Theorem 6.5.1 hold. Then, for each n, the process $\{\mathbf{x}_n(t)\}$ satisfies*

$$\mathbb{P}_\theta\left(\sup_{t\geq 0}\|\mathbf{x}(t)\|<\infty\right)=1. \tag{E.1}$$

Proof. The proof is built around a similar framework as the proof of Lemma IV.1 in Kar and Moura (2014) with appropriate modifications to take into account the state-dependent nature of the innovation gain and the projection operator used in (6.16). Define the process $\{\mathbf{z}(t)\}$ as follows $\mathbf{z}(t)=\mathbf{x}(t)-\mathbf{1}_N\otimes\boldsymbol{\theta}$ where $\boldsymbol{\theta}$ denotes the true (but unknown) parameter. Note the following recursive relationship:

$$\begin{aligned}\widehat{\mathbf{x}}(t+1)-\mathbf{1}_N\otimes\boldsymbol{\theta}&=\mathbf{z}(t)-\beta_t(\mathbf{L}\otimes\mathbf{I}_M)z(t)\\&+\alpha_t\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{y}(t)-\mathbf{f}(\mathbf{x}(t))\right),\end{aligned} \tag{E.2}$$

which further implies that,

$$\begin{aligned}\widehat{\mathbf{x}}(t+1)-\mathbf{1}_N\otimes\boldsymbol{\theta}&=\mathbf{z}(t)-\beta_t(\mathbf{L}\otimes\mathbf{I}_M)z(t)\\&+\alpha_t\mathbf{G}\left(\mathbf{x}(t)\right)\left(\mathbf{y}(t)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\\&-\alpha_t\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right).\end{aligned} \tag{E.3}$$

In the above, we have used a basic property of the Laplacian $\mathbf{L}$,

$$\left(\mathbf{L}\otimes\mathbf{I}_M\right)\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)=\mathbf{0}, \tag{E.4}$$

Since the projection is onto a convex set it is non-expansive. It follows that the inequality

$$\|\mathbf{x}_n(t+1) - \boldsymbol{\theta}\| \leq \|\widehat{\mathbf{x}}_n(t+1) - \boldsymbol{\theta}\| \tag{E.5}$$

holds for all n and t. Taking norms on both sides of (E.2) and using (E.5), we have,

$$\begin{aligned}
&\|\mathbf{z}(t+1)\|^2 \leq \|\mathbf{z}(t)\|^2 - 2\beta_t \mathbf{z}^\top(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{z}(t) \\
&- 2\alpha_t \mathbf{z}^\top(t) \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \\
&+ \beta_t^2 \mathbf{z}^\top(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right)^2 \mathbf{z}(t) \\
&+ 2\alpha_t \beta_t \mathbf{z}^\top(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \theta\right)\right) \\
&+ \alpha_t^2 \left\|\mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)\right\|^2 \\
&+ \alpha_t^2 \left\|\mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)\right\|^2 \\
&+ 2\alpha_t \mathbf{z}^\top(t) \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}(\mathbf{1}_N \otimes \boldsymbol{\theta})\right) \\
&+ 2\alpha_t^2 \left(\mathbf{y}(t) - \mathbf{f}(\mathbf{1}_N \otimes \boldsymbol{\theta})\right)^\top \mathbf{R}^{-1} \mathbf{G}^\top\left(\mathbf{x}(t)\right) \\
&\times \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right) - \mathbf{f}\left(\mathbf{x}(t)\right)\right).
\end{aligned} \tag{E.6}$$

Consider the orthogonal decomposition

$$\mathbf{z} = \mathbf{z}_c + \mathbf{z}_{c\perp}, \tag{E.7}$$

where $\mathbf{z}_c$ denotes the projection of $\mathbf{z}$ to the consensus subspace $\mathcal{C} = \{\mathbf{z} \in \mathbb{R}^{MN} | \mathbf{z} = 1_N \otimes a, \text{for some } a \in \mathbb{R}^M\}$. From (6.1), we have that,

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{y}(t) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right] = \mathbf{0}. \tag{E.8}$$

Consider the process

$$V_2(t) = \|\mathbf{z}(t)\|^2. \tag{E.9}$$

Using conditional independence properties, we have,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}}[V_2(t+1)|\mathcal{F}_t] \leq V_2(t) + \beta_t^2 \mathbf{z}^\top(t) \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right)^2 \mathbf{z}(t) \\
&+ \alpha_t^2 \mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)\right\|^2\right] \\
&- 2\beta_t \mathbf{z}^\top(t) \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right) \mathbf{z}(t) \\
&- 2\alpha_t \mathbf{z}^\top(t) \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \\
&+ 2\alpha_t \beta_t \mathbf{z}^\top(t) \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right) \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \\
&+ \alpha_t^2 \left\|\left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)^\top \mathbf{G}^\top\left(\mathbf{x}(t)\right) \mathbf{R}^{-1}\right\|^2.
\end{aligned} \tag{E.10}$$

We use the following inequalities $\forall t \geq t_1$,

$$\mathbf{z}^\top(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right)^2 \mathbf{z}(t) \overset{(q1)}{\leq} \lambda_N^2(\mathbf{L}) ||\mathbf{z}_{C\perp}(t)||^2;$$

$$\begin{aligned}
&\mathbf{z}^{\top}(t)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\geq c_1||\mathbf{z}(t)||^2\overset{(q2)}{\geq}0;\\
&\mathbf{z}^{\top}(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)\mathbf{z}(t)\overset{(q3)}{\geq}\lambda_2(\overline{\mathbf{L}})\left\|\mathbf{z}_{C\perp}(t)\right\|^2;\\
&\mathbf{z}^{\top}(t)\left(\mathbf{L}\otimes\mathbf{I}_M\right)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\\
&\overset{(q4)}{\leq}c_2\left\|\mathbf{z}(t)\right\|^2,
\end{aligned}\tag{E.11}$$

for c_1 as defined in Assumption 6.4.3 and a positive constant c_2. Inequalities $(q1)$ and $(q4)$ follow from the properties of the Laplacian. Inequality $(q2)$ follows from Assumption 6.4.3 and $(q4)$ follows from Assumption 6.4.2 since we have that $\|\nabla\mathbf{f}_n\left(\mathbf{x}_n(t)\right)\|$ is uniformly bounded from above by k_n for all n and hence, we have that $\|\mathbf{G}\left(\mathbf{x}(t)\right)\|\leq\max_{n=1,\cdots,N}k_n$. We also have

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{y}(t)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\right\|^2\right]\leq c_4,\tag{E.12}$$

for some constant $c_4>0$. In (E.12), we use the fact that the noise process under consideration has finite covariance. We also use the fact that $\|\mathbf{G}\left(\mathbf{x}(t)\right)\|\leq\max_{n=1,\cdots,N}k_n$, which in turn follows from Assumption 6.4.1. We further have that,

$$\left\|\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\right\|^2\leq c_3\left\|\mathbf{z}(t)\right\|^2,\tag{E.13}$$

where $c_3>0$ is a constant. It is to be noted that (E.13) follows from the Lipschitz continuity in Assumption 6.4.1 and the result that $\|\mathbf{G}\left(\mathbf{x}(t)\right)\|\leq\max_{n=1,\cdots,N}k_n$. Using (E.10)-(E.13), we have,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}}[V_2(t+1)|\mathcal{F}_t]\leq\left(1+c_5\left(\alpha_t\beta_t+\alpha_t^2\right)\right)V_2(t)\\
&-c_6(\beta_t-\beta_t^2)||\mathbf{x}_{C\perp}(t)||^2+c_4\alpha_t^2,
\end{aligned}\tag{E.14}$$

for some positive constants c_5 and c_6. As β_t^2 goes to zero faster than β_t, $\exists t_2$ such that $\forall t\geq t_2$, $\beta_t\geq\beta_t^2$. Hence $\exists t_2$ and $\exists\tau_1,\tau_2>1$ such that for all $t\geq t_2$

$$c_5\left(\alpha_t\beta_t+\alpha_t^2\right)\leq\frac{c_7}{(t+1)^{\tau_1}}=\gamma_t\quad\text{and}\quad c_4\alpha_t^2\leq\frac{c_8}{(t+1)^{\tau_2}}=\hat{\gamma}_t,\tag{E.15}$$

where $c_7,c_8>0$ are constants. By the above construction we obtain $\forall t\geq t_2$,

$$\mathbb{E}_{\theta^*}[V_2(t+1)|\mathcal{F}_t]\leq(1+\gamma_t)V_2(t)+\hat{\gamma}_t,\tag{E.16}$$

where the positive weight sequences $\{\gamma_t\}$ and $\{\hat{\gamma}_t\}$ are summable i.e.,

$$\sum_{t\geq0}\gamma_t<\infty,\sum_{t\geq0}\hat{\gamma}_t<\infty.\tag{E.17}$$

By (E.17), the product $\prod_{s=t}^{\infty}(1+\gamma_s)$ exists for all t. Now let $\{W(t)\}$ be such that

$$W(t)=\left(\prod_{s=t}^{\infty}(1+\gamma_s)\right)V_2(t)+\sum_{s=t}^{\infty}\hat{\gamma}_s,\quad\forall t\geq t_2.\tag{E.18}$$

By (E.18), it can be shown that $\{W(t)\}$ satisfies,

$$\mathbb{E}_{\theta^*}[W(t+1)|\mathcal{F}_t] \leq W(t). \tag{E.19}$$

Hence, $\{W(t)\}$ is a non-negative super martingale and converges a.s. to a bounded random variable W^* as $t \to \infty$. It then follows from (E.18) that $W(t) \to W^*$ as $t \to \infty$. Thus, we conclude that the sequences $\{\theta_n(t)\}$ are bounded for all n. □

Due to inherent stochasticity associated with the noisy observations, there need not be *uniform* boundedness of the estimate sequences. Hence, while Lemma E.0.1 establishes the pathwise boundedness of the parameter estimate sequence, it does not guarantee uniform boundedness over almost all sample paths.

Lemma E.0.2. *Let the hypotheses of Theorem 6.5.1 hold. Then, we have,*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty} \mathbf{x}_n(t) = \boldsymbol{\theta}\right) = 1, \ \forall n. \tag{E.20}$$

Proof. Denote the processes $\{\mathbf{z}_n(t)\}$ and $\{\widehat{\mathbf{z}}_n(t)\}$ as

$$\mathbf{z}_n(t) = \mathbf{x}_n(t) - \boldsymbol{\theta} \ \text{ and } \ \widehat{\mathbf{z}}_n(t) = \widehat{\mathbf{x}}_n(t) - \boldsymbol{\theta} \tag{E.21}$$

respectively. Let $\mathbf{z}(t) = \left[\mathbf{z}_1^\top(t) \cdots \mathbf{z}_n^\top(t)\right]^\top$ and $\widehat{\mathbf{z}}(t) = \left[\widehat{\mathbf{z}}_1^\top(t) \cdots \widehat{\mathbf{z}}_n^\top(t)\right]^\top$. From, (6.17), we have,

$$\begin{aligned}\widehat{\mathbf{z}}(t+1) &= \mathbf{z}(t) - \beta_t \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{z}(t) \\ &+ \alpha_t \mathbf{G}(\mathbf{x}(t)) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{x}(t)\right)\right),\end{aligned} \tag{E.22}$$

where we have used the fact that $\left(\mathbf{L} \otimes \mathbf{I}_M\right)\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right) = \mathbf{0}$. Define the $\{\mathcal{F}_t\}$-adapted process $\{V(t)\}$ by

$$V(t) = \left\|\mathbf{z}(t)\right\|^2. \tag{E.23}$$

Now, using (E.2) and the fact that $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{y}(t) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right] = \mathbf{0}$, we have,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)|\mathcal{F}_t] \leq V(t) + \beta_t^2 \mathbf{z}^\top(t) \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right)^2 \mathbf{z}(t) \\ &+ \alpha_t^2 \mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)\right\|^2\right] \\ &- 2\beta_t \mathbf{z}^\top(t) \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right) \mathbf{z}(t) \\ &- 2\alpha_t \mathbf{z}(t)^\top \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \\ &+ 2\alpha_t \beta_t \mathbf{z}(t)^\top \left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right) \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \\ &+ \alpha_t^2 \left\|\left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)^\top \mathbf{G}\left(\mathbf{x}(t)\right)^\top \mathbf{R}^{-1}\right\|^2.\end{aligned} \tag{E.24}$$

Following the steps as in the proof of Lemma C.1.1 and using (E.11)-(E.13), we have,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)|\mathcal{F}_t] &\leq \left(1 - c_7\alpha_t\right) V(t) + c_7\alpha_t^2 \\ &\leq V(t) + c_7\alpha_t^2,\end{aligned} \tag{E.25}$$

for an appropriately chosen positive constant c_7. Now, consider the $\{\mathcal{F}_t\}$-adapted process $\{V_1(t)\}$ defined

as follows

$$V_1(t) = V(t) - c_7 \sum_{s=t}^{\infty} \alpha_s^2$$
$$= V(t) - bc_7 \sum_{s=t}^{\infty} (t+1)^{-2}. \tag{E.26}$$

Since, $\{(t+1)^{-2}\}$ is summable, the process $\{V_1(t)\}$ is bounded from below. Moreover, it also follows that $\{V_1(t)\}_{t \geq t_1}$ is a supermartingale and hence converges a.s. to a finite random variable. By definition from (E.26), we also have that $\{V(t)\}$ converges to a non-negative finite random variable V^*. Finally, from (E.25), we have that,

$$\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)] \leq (1 - c_1\alpha_t)\, \mathbb{E}_{\boldsymbol{\theta}}[V(t)] + bc_7(t+1)^{-2}, \tag{E.27}$$

for $t \geq t_1$. The sequence $\{V(t)\}$ then falls under the purview of Lemmas 4 and 5 of Kar and Moura (2011), and we have $\mathbb{E}_{\boldsymbol{\theta}}[V(t)] \to 0$ as $t \to \infty$. Finally, by Fatou's Lemma, where we use the non-negativity of the sequence $\{V(t)\}$, we conclude that

$$0 \leq \mathbb{E}_{\boldsymbol{\theta}}[V^*] \leq \liminf_{t \to \infty} \mathbb{E}_{\boldsymbol{\theta}}[V(t)] = 0, \tag{E.28}$$

which thus implies that $V^* = 0$ a.s. Hence, $\|\mathbf{z}(t)\| \to 0$ as $t \to \infty$ and the desired assertion follows. □

We will use the approximation result from Lemma C.1.3 and the generalized convergence criterion from Lemma C.1.4 for the proof of Theorem 6.5.1. We now return to the proof of Theorem 6.5.1. Define $\bar{\tau} \in [0, 1/2)$ such that,

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t \to \infty} (t+1)^{\bar{\tau}} \|\mathbf{z}(t)\| = 0\right) = 1, \tag{E.29}$$

where $\{\mathbf{z}(t)\}$ is as defined in (E.21) and note that such a $\bar{\tau}$ always exists by Lemma E.0.2 (in particular $\bar{\tau} = 0$). We now analyze and show that there exists a τ such that $\bar{\tau} < \tau < 1/2$ for which the claim in (E.29) holds. Now, choose a $\widehat{\tau} \in (\tau, 1/2)$ and let $\mu = (\widehat{\tau} + \bar{\tau})/2$. The recursion for $\{\mathbf{z}(t)\}$ can be written as follows:

$$\begin{aligned}
&\|\mathbf{z}(t+1)\|^2 \leq \|\mathbf{z}(t)\|^2 - 2\beta_t \mathbf{z}^\top(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{z}(t) \\
&- 2\alpha_t \mathbf{z}^\top(t) \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \\
&+ \beta_t^2 \mathbf{z}^\top(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right)^2 \mathbf{z}(t) \\
&+ 2\alpha_t \beta_t \mathbf{z}^\top(t) \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \theta\right)\right) \\
&+ \alpha_t^2 \left\|\mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)\right\|^2 \\
&+ \alpha_t^2 \left\|\mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)\right\|^2 \\
&+ 2\alpha_t \mathbf{z}^\top(t) \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}(\mathbf{1}_N \otimes \boldsymbol{\theta})\right) \\
&+ 2\alpha_t^2 \left(\mathbf{y}(t) - \mathbf{f}(\mathbf{1}_N \otimes \boldsymbol{\theta})\right)^\top \mathbf{R}^{-1} \mathbf{G}^\top\left(\mathbf{x}(t)\right) \\
&\times \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right) - \mathbf{f}\left(\mathbf{x}(t)\right)\right).
\end{aligned} \tag{E.30}$$

Let $\mathbf{J}(t) = \mathbf{G}\left(\mathbf{x}(t)\right) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)$. Now, we consider the term $\alpha_t^2 \|\mathbf{J}(t)\|^2$. Since, the noise process

under consideration has finite second moment and $2\mu < 1$, we have,

$$\sum_{t\geq 0}(t+1)^{2\mu}\alpha_t^2 \left\|\mathbf{J}(t)\right\|^2 < \infty. \tag{E.31}$$

Let $\mathbf{W}(t) = \alpha_t \mathbf{z}(t)^\top \mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{y}(t) - \mathbf{f}(\mathbf{1}_N \otimes \boldsymbol{\theta})\right)$. It follows that $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{W}(t)|\mathcal{F}_t\right] = 0$. We also have that $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{W}^2(t)|\mathcal{F}_t\right] \leq \alpha_t^2 \left\|\mathbf{z}(t)\right\|^2 \left\|\mathbf{J}(t)\right\|^2$. Noting that the noise under consideration has finite second order moment, we have,

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{W}^2(t)|\mathcal{F}_t\right] = o\left((t+1)^{-2-2\bar{\tau}}\right), \tag{E.32}$$

and, hence,

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(t+1)^{4\mu}\mathbf{W}^2(t)|\mathcal{F}_t\right] = o\left((t+1)^{-2+2\widehat{\tau}}\right). \tag{E.33}$$

Hence, by Lemma C.1.4, we conclude that $\sum_{t\geq 0}(t+1)^{2\mu}\mathbf{W}(t)$ exists and is finite, as $2\widehat{\tau} < 1$. Similarly, it can be shown that, for $\mathbf{W}_1(t) = \alpha_t\beta_t \mathbf{z}(t)^\top \mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}(\mathbf{1}_N \otimes \boldsymbol{\theta}) - \mathbf{y}(t)\right)$, the sum $\sum_{t\geq 0}(t+1)^{2\mu}\mathbf{W}_1(t)$ exists and is finite. Finally, consider $\mathbf{W}_2(t) = \alpha_t^2 \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)^\top \mathbf{R}^{-1}\mathbf{G}\left(\mathbf{x}(t)\right)^\top \times \mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right) - \mathbf{f}\left(\mathbf{x}(t)\right)\right)$. It follows that $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{W}_2(t)|\mathcal{F}_t\right] = 0$. We also have that $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{W}_2^2(t)|\mathcal{F}_t\right] \leq \alpha_t^4 \left\|\mathbf{z}(t)\right\|^2 \left\|\mathbf{J}(t)\right\|^2$. Following as in (E.32) and (E.33), we have that $\sum_{t\geq 0}(t+1)^{2\mu}\mathbf{W}_2(t)$ exists and is finite. Using all the inequalities derived in (E.11)-(E.13), we have,

$$\begin{aligned}\left\|\mathbf{z}(t+1)\right\|^2 &\leq \left(1 - c_1\alpha_t + c_2\alpha_t\beta_t + c_3\alpha_t^2\right)\left\|\mathbf{z}(t)\right\|^2 + \alpha_t^2\left\|\mathbf{J}(t)\right\|^2 \\ &- c_6(\beta_t - \beta_t^2)\left\|\mathbf{z}_{C\perp}(t)\right\|^2 + 2\mathbf{W}(t) + 2\mathbf{W}_1(t) + 2\mathbf{W}_2(t).\end{aligned} \tag{E.34}$$

Finally, noting that $c_1\alpha_t$ dominates $c_2\alpha_t\beta_t$ and $c_3\alpha_t^2$, β_t dominates β_t^2, we have eventually

$$\begin{aligned}\left\|\mathbf{z}(t+1)\right\|^2 &\leq \left(1 - c_1\alpha_t\right)\left\|\mathbf{z}(t)\right\|^2 \\ &+ \alpha_t^2\left\|\mathbf{J}(t)\right\|^2 + 2\mathbf{W}(t) + 2\mathbf{W}_1(t) + 2\mathbf{W}_2(t).\end{aligned} \tag{E.35}$$

To this end, using the analysis in (E.31)-(E.33), we have, from (E.35)

$$\left\|\mathbf{z}(t+1)\right\|^2 \leq \left(1 - c_1\alpha_t\right)\left\|\mathbf{z}(t)\right\|^2 + d_t(t+1)^{-2\mu}, \tag{E.36}$$

where

$$d_t(t+1)^{-2\mu} = \alpha_t^2\left\|\mathbf{J}(t)\right\|^2 + 2\mathbf{W}(t) + +2\mathbf{W}_1(t) + 2\mathbf{W}_2(t). \tag{E.37}$$

Finally, noting that $c_1\alpha_t(t+1) = 1 > 2\mu$, an immediate application of Lemma C.1.3 gives

$$\limsup_{t\to\infty}(t+1)^{2\mu}\left\|\mathbf{z}(t)\right\|^2 < \infty \ a.s. \tag{E.38}$$

So, we have that there exists a τ with $\bar{\tau} < \tau < \mu$ for which $(t+1)^\tau \left\|\mathbf{z}(t)\right\| \to 0$ as $t \to \infty$. Thus, for every $\bar{\tau}$ for which (5.12) holds, there exists $\tau \in (\bar{\tau}, 1/2)$ for which the result in (5.12) still continues to hold. By a simple application of induction, we conclude that the result holds for all $\tau \in [0, 1/2)$. □

E.0.2 Proof of Theorem 6.5.2

Proof. The proof of Theorem 6.5.2 uses Lemma C.1.5 from Fabian (1968) which concerns with the asymptotic normality of the stochastic recursions. From Theorem 6.5.1 and the fact that $\boldsymbol{\theta}$ lies in the interior of the parameter set Θ, we have that there exists an $\epsilon > 0$, such that $B_\epsilon(\boldsymbol{\theta}) \in \Theta$, where $B_\epsilon(\boldsymbol{\theta})$ denotes the open ball centered at $\boldsymbol{\theta}$ with radius ϵ. In particular, fix an $\epsilon > 0$ for which $B_\epsilon(\boldsymbol{\theta}) \in \Theta$. Then, we have that there exists a random time $t_\epsilon(\omega)$, which is almost surely finite, i.e., $\mathbb{P}\left(t_\epsilon(\omega) < \infty\right) = 1$, such that $\|\mathbf{x}_\omega(t) - \boldsymbol{\theta}\| < \epsilon$ for all $t \geq t_\epsilon(\omega)$, where ω denotes the sample path. In the above, we introduce the ω-argument to emphasize that the time $t_\epsilon(\omega)$ is random and sample-path dependent and our analysis is pathwise. With the above development in place, we note that (6.14) and (6.15) can be rewritten as follows:

$$\begin{aligned}\mathbf{x}(t+1) &= \mathbf{x}(t) - \beta_t \left(\mathbf{L} \otimes \mathbf{I}_M\right) \mathbf{x}(t) \\ &+ \alpha_t \mathbf{G}(\mathbf{x}(t)) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{x}(t)\right)\right) + \mathbf{e}_{\mathcal{P}}(t), \ \forall t \geq 0, \end{aligned} \tag{E.39}$$

where $\mathbf{e}_{\mathcal{P}}(t)$ is the projection error which is given by,

$$\mathbf{e}_{\mathcal{P}}(t) = \mathbf{x}(t+1) - \widehat{\mathbf{x}}(t+1), \tag{E.40}$$

and in particular $\mathbf{e}_{\mathcal{P}}(t) = 0$ for all $t \geq t_\epsilon$.
Define the process $\{\mathbf{x}_{\text{avg}}(t)\}$ as

$$\mathbf{x}_{\text{avg}}(t) = \left(\frac{\mathbf{1}_N^\top}{N} \otimes \mathbf{I}_M\right) \mathbf{x}(t). \tag{E.41}$$

It is readily seen that the process $\{\mathbf{x}_{\text{avg}}(t)\}$ satisfies the recursion

$$\begin{aligned}\mathbf{x}_{\text{avg}}(t+1) &= \mathbf{x}_{\text{avg}}(t) + \left(\frac{\mathbf{1}_N^\top}{N} \otimes \mathbf{I}_M\right) \mathbf{e}_{\mathcal{P}}(t) \\ &+ \alpha_t \left(\frac{\mathbf{1}_N^\top}{N} \otimes \mathbf{I}_M\right) \mathbf{G}(\mathbf{x}(t)) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{x}(t)\right)\right) \\ &= \mathbf{x}_{\text{avg}}(t) + \left(\frac{\mathbf{1}_N^\top}{N} \otimes \mathbf{I}_M\right) \mathbf{e}_{\mathcal{P}}(t) \\ &+ \frac{\alpha_t}{N} \sum_{n=1}^{N} \nabla \mathbf{f}_n\left(\mathbf{x}_n(t)\right) \mathbf{R}_n^{-1} \left(\mathbf{y}_n(t) - \mathbf{f}_n\left(\mathbf{x}_n(t)\right)\right). \end{aligned} \tag{E.42}$$

Now noting that, for all $t \geq 0$, $\mathbf{x}_n(t) \in \Theta$ for each n and as Θ is a convex set, we have that $\mathbf{x}_{\text{avg}}(t) \in \Theta$ for all $t \geq 0$. Then, we have by the mean-value theorem for each agent n

$$\mathbf{f}_n\left(\mathbf{x}_{\text{avg}}(t)\right) = \mathbf{f}_n\left(\boldsymbol{\theta}\right) + \nabla^\top \mathbf{f}_n\left(c\boldsymbol{\theta} + (1-c)\mathbf{x}_{\text{avg}}(t)\right)\left(\mathbf{x}_{\text{avg}}(t) - \boldsymbol{\theta}\right), \tag{E.43}$$

where $0 < c < 1$. It is to be noted that $\nabla^\top \mathbf{f}_n\left(c\boldsymbol{\theta} + (1-c)\mathbf{x}_{\text{avg}}(t)\right) \to \nabla^\top \mathbf{f}_n\left(\boldsymbol{\theta}\right)$ as $\mathbf{x}_{\text{avg}}(t) \to \boldsymbol{\theta}$ in the limit $t \to \infty$. Using (E.43) in (E.42), we have for all $t \geq 0$,

$$\begin{aligned}\mathbf{x}_{\text{avg}}(t+1) - \boldsymbol{\theta} = \Bigg(\mathbf{I} - \frac{\alpha_t}{N} \Bigg(\sum_{n=1}^{N} \nabla \mathbf{f}_n\left(\mathbf{x}_n(t)\right) \mathbf{R}_n^{-1} \\ \times \ \nabla^\top \mathbf{f}_n\left(c\boldsymbol{\theta} + (1-c)\mathbf{x}_{\text{avg}}(t)\right)\Bigg)\Bigg)\left(\mathbf{x}_{\text{avg}}(t) - \boldsymbol{\theta}\right)\end{aligned}$$

$$+\frac{\alpha_t}{N}\sum_{n=1}^{N}\nabla\mathbf{f}_n\left(\mathbf{x}_n(t)\right)\mathbf{R}_n^{-1}\zeta_n(t)+\left(\frac{\mathbf{1}_N^\top}{N}\otimes\mathbf{I}_M\right)\mathbf{e}_{\mathcal{P}}(t)$$
$$+\frac{\alpha_t}{N}\sum_{n=1}^{N}\nabla\mathbf{f}_n\left(\mathbf{x}_n(t)\right)\mathbf{R}_n^{-1}\left(\mathbf{f}_n\left(\mathbf{x}_{\text{avg}}(t)\right)-\mathbf{f}_n\left(\mathbf{x}_n(t)\right)\right). \tag{E.44}$$

The following Lemma will be crucial for the subsequent part of the proof.

Lemma E.0.3. *For every τ_0 such that $0\leq\tau_0<1-\tau_1-1/(2+\epsilon_1)$, we have,*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}(t+1)^{\tau_0}\left(\mathbf{x}_n(t)-\mathbf{x}_{avg}(t)\right)=0\right)=1. \tag{E.45}$$

Proof. The proof follows exactly like the proof of Lemma IV.2 in Kar and Moura (2014). Note the additional term that comes up in $\mathbf{x}_n(t)-\mathbf{x}_{\text{avg}}(t)$ in the current context due to the projection error is given by $\mathbf{e}_{\mathcal{P},n}(t)-\left(\frac{\mathbf{1}_N^\top}{N}\otimes\mathbf{I}_M\right)\mathbf{e}_{\mathcal{P}}(t)$; nonetheless, this term satisfies the property that

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}(t+1)^{\tau_0}\left(\mathbf{e}_{\mathcal{P},n}(t)-\left(\frac{\mathbf{1}_N^\top}{N}\otimes\mathbf{I}_M\right)\mathbf{e}_{\mathcal{P}}(t)\right)=0\right)=1$$

as $\mathbf{e}_{\mathcal{P}}(t)=0$ for all $t\geq t_\epsilon$. Hence, the techniques employed in the proof of Lemma IV.2 also apply here. Lemma IV.2 in Kar and Moura (2014) is concerned with the asymptotic agreement of the estimates across any pair of agents, but as an intermediate result the agreement of the estimate at an agent and the averaged estimate is established. □

As $\tau_1+1/(2+\epsilon_1)<1/2$, from Lemma E.0.3 we have that there exists an $\epsilon_2>0$ (sufficiently small) such that

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}(t+1)^{\frac{1}{2}+\epsilon_2}\left(\mathbf{x}_n(t)-\mathbf{x}_{\text{avg}}(t)\right)=0\right)=1. \tag{E.46}$$

We consider the process $\{\mathbf{x}_{\text{avg}}(t)\}$ for the application of Lemma C.1.5. Hence, comparing term by term of (E.44) with (C.15), we have,

$$\begin{aligned}
&\boldsymbol{\Gamma}_t=\frac{a}{N}\sum_{n=1}^{N}\nabla\mathbf{f}_n\left(\mathbf{x}_n(t)\right)\mathbf{R}_n^{-1}\nabla^\top\mathbf{f}_n\left(c\boldsymbol{\theta}+(1-c)\mathbf{x}_{\text{avg}}(t)\right)\\
&\to\frac{a}{N}\sum_{n=1}^{N}\nabla\mathbf{f}_n\left(\boldsymbol{\theta}\right)\mathbf{R}_n^{-1}\nabla\mathbf{f}_n^\top\left(\boldsymbol{\theta}\right)=a\boldsymbol{\Gamma}_{\boldsymbol{\theta}},\\
&\boldsymbol{\Phi}_t=a\left(\frac{\mathbf{1}_N^\top}{N}\otimes\mathbf{I}_M\right)\mathbf{G}(\mathbf{x}(t))\mathbf{R}^{-1}\to a\left(\frac{\mathbf{1}_N^\top}{N}\otimes\mathbf{I}_M\right)\mathbf{G}(\mathbf{1}\otimes\boldsymbol{\theta})\mathbf{R}^{-1}\\
&=a\boldsymbol{\Phi},\\
&\mathbf{V}_t=\zeta(t),\mathbb{E}\left[\mathbf{V}_t|\mathcal{F}_t\right]=0,\mathbb{E}\left[\mathbf{V}_t\mathbf{V}_t^\top|\mathcal{F}_t\right]=\mathbf{R},\\
&\mathbf{T}_t=a(t+1)^{1/2}\left(\frac{\mathbf{1}_N^\top}{N}\otimes\mathbf{I}_M\right)\mathbf{G}(\mathbf{x}(t))\mathbf{R}^{-1}\times\\
&\left(\mathbf{f}\left(\mathbf{1}\otimes\mathbf{x}_{\text{avg}}(t)\right)-\mathbf{f}\left(\mathbf{x}(t)\right)\right)+(t+1)^{3/2}\left(\frac{\mathbf{1}_N^\top}{N}\otimes\mathbf{I}_M\right)\mathbf{e}_{\mathcal{P}}(t)\to 0,
\end{aligned} \tag{E.47}$$

where the convergence of $\mathbf{T}_t$ follows from Lemma E.0.3. Due to the i.i.d nature of the noise process, we have the uniform integrability condition for the process $\{\mathbf{V}_t\}$. Hence, $\{\mathbf{x}_{\text{avg}}(t)\}$ falls under the purview of

Lemma C.1.5, and we thus conclude that

$$(t+1)^{1/2}\left(\mathbf{x}_{\text{avg}}(t)-\boldsymbol{\theta}\right) \stackrel{\mathcal{D}}{\Longrightarrow} \mathcal{N}(0, \mathbf{PMP}^{\top}), \tag{E.48}$$

where

$$\begin{aligned} &a\mathbf{P}^{\top}\boldsymbol{\Gamma}_{\boldsymbol{\theta}}\mathbf{P} = a\boldsymbol{\Lambda}_{\boldsymbol{\theta}}, \\ &[\mathbf{M}]_{ij} = \left[a^2\mathbf{P}^{\top}\boldsymbol{\Phi}\mathbf{R}\boldsymbol{\Phi}^{\top}\mathbf{P}\right]_{ij}\left(a\left[\boldsymbol{\Lambda}\right]_{\boldsymbol{\theta},ii} + a\left[\boldsymbol{\Lambda}\right]_{\boldsymbol{\theta},jj} - 1\right)^{-1} \\ &= \frac{a^2}{N}\left[\boldsymbol{\Lambda}\right]_{ij}\left(a\left[\boldsymbol{\Lambda}\right]_{\boldsymbol{\theta},ii} + a\left[\boldsymbol{\Lambda}\right]_{\boldsymbol{\theta},jj} - 1\right)^{-1}, \end{aligned} \tag{E.49}$$

which also implies that $\mathbf{M}$ is a diagonal matrix with its i-th diagonal element given by $\frac{a^2\boldsymbol{\Lambda}_{\boldsymbol{\theta},ii}}{2aN\boldsymbol{\Lambda}_{\boldsymbol{\theta},ii}-N}$. We already have that $\mathbf{P}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\mathbf{P}^{\top} = \boldsymbol{\Gamma}_{\boldsymbol{\theta}}$. Hence, the matrix with eigenvalues as $\frac{a^2\boldsymbol{\Lambda}_{\boldsymbol{\theta},ii}}{2aN\boldsymbol{\Lambda}_{\boldsymbol{\theta},ii}-N}$ is given by

$$\mathbf{PMP}^{\top} = \frac{a\mathbf{I}}{2N} + \frac{\left(\boldsymbol{\Gamma}_{\boldsymbol{\theta}} - \frac{\mathbf{I}}{2a}\right)^{-1}}{4N}. \tag{E.50}$$

Now from (E.46), which is a consequence of Lemma E.0.3, we have that the processes $\{\mathbf{x}_n(t)\}$ and $\{\mathbf{x}_{\text{avg}}(t)\}$ are indistinguishable in the $t^{1/2}$ time scale, which is formalized as follows:

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}\left\|\sqrt{t+1}\left(\mathbf{x}_n(t)-\boldsymbol{\theta}\right) - \sqrt{t+1}\left(\mathbf{x}_{\text{avg}}(t)-\boldsymbol{\theta}\right)\right\| = 0\right) \\ &= \mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty}\left\|\sqrt{t+1}\left(\mathbf{x}_n(t)-\mathbf{x}_{\text{avg}}(t)\right)\right\| = 0\right) = 1. \end{aligned} \tag{E.51}$$

Thus, the difference of the sequences $\left\{\sqrt{t+1}\left(\mathbf{x}_n(t)-\boldsymbol{\theta}\right)\right\}$ and $\left\{\sqrt{t+1}\left(\mathbf{x}_{\text{avg}}(t)-\boldsymbol{\theta}\right)\right\}$ converges a.s. to zero as $t\to\infty$, and hence we have,

$$\sqrt{t+1}\left(\mathbf{x}_n(t)-\boldsymbol{\theta}\right) \stackrel{\mathcal{D}}{\Longrightarrow} \mathcal{N}\left(0, \frac{a\mathbf{I}}{2N} + \frac{\left(\boldsymbol{\Gamma}_{\boldsymbol{\theta}} - \frac{\mathbf{I}}{2a}\right)^{-1}}{4N}\right). \tag{E.52}$$

□

Proof of Theorem 6.7.1.

Lemma E.0.4. *For each n, the process $\{\mathbf{x}_n(t)\}$ satisfies*

$$\mathbb{P}_{\theta}\left(\sup_{t\geq 0}\|\mathbf{x}(t)\| < \infty\right) = 1. \tag{E.53}$$

Proof. Consider (6.15). Since the projection is onto a convex set it is non-expansive. It follows that the inequality

$$\|\mathbf{x}_n(t+1)-\boldsymbol{\theta}\| \leq \|\widehat{\mathbf{x}}_n(t+1)-\boldsymbol{\theta}\| \tag{E.54}$$

holds for all n and t. We first note that,

$$\mathbf{L}(t) = \beta_t\overline{\mathbf{L}} + \widetilde{\mathbf{L}}(t), \tag{E.55}$$

where $\mathbb{E}\left[\widetilde{\mathbf{L}}(t)\right] = \mathbf{0}$ and $\mathbb{E}\left[\widetilde{\mathbf{L}}^2_{i,j}(t)\right] = \frac{\rho_0\beta_0}{(t+1)^{1+\epsilon}} - \frac{\beta_0^2}{(t+1)^2}$, for $\{i,j\} \in E, i \neq j$.
Define, $\mathbf{z}(t) = \mathbf{x}(t) - \mathbf{1}_N \otimes \boldsymbol{\theta}$ and $V(t) = \|\mathbf{z}(t)\|^2$. Note that $\mathbf{z}(t)$ corresponds to the estimation error vector at time t; its squared norm $V(t)$ will first serve us as a Lyapunov function to establish the almost sure boundedness of $\mathbf{x}(t)$ as in Lemma E.0.4. Let $\{\mathcal{F}_t\}$ be the natural filtration generated by the random observations and the random Laplacians i.e.,

$$\mathcal{F}_t = \sigma\left(\left\{\{\mathbf{y}_n(s)\}_{n=1}^N, \{\mathbf{L}(s)\}\right\}_{s=0}^{t-1}\right). \tag{E.56}$$

By algebraic manipulations, conditional independence, we have that,

$$\begin{aligned}
&\mathbb{E}\left[V(t+1)|\mathcal{F}_t\right] \leq V(t) + \beta_t^2 \mathbf{z}^\top(t)\left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right)^2 \mathbf{z}(t) \\
&+ \alpha_t^2 \mathbb{E}\left[\left\|\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{y}(t) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)\right\|^2\right] \\
&- 2\beta_t \mathbf{z}^\top(t)\left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right)\mathbf{z}(t) \\
&- 2\alpha_t \mathbf{z}^\top(t)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \\
&+ 2\alpha_t\beta_t \mathbf{z}^\top(t)\left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \\
&+ \alpha_t^2 \left\|\left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right)^\top \mathbf{G}^\top\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\right\|^2 + \mathbf{z}^\top(t)\mathbb{E}\left[\left(\widetilde{\mathbf{L}}(t) \otimes \mathbf{I}_M\right)^2\right]\mathbf{z}(t).
\end{aligned} \tag{E.57}$$

Consider the orthogonal decomposition

$$\mathbf{z} = \mathbf{z}_C + \mathbf{z}_{C\perp}, \tag{E.58}$$

where $\mathbf{z}_C$ denotes the projection of $\mathbf{z}$ to the consensus subspace $\mathcal{C} = \{\mathbf{z} \in \mathbb{R}^{MN} | \mathbf{z} = \mathbf{1}_N \otimes a, \text{for some } a \in \mathbb{R}^M\}$. The following inequalities hold for all $t \geq t_1$, where t_1 is a sufficiently large positive integer:

$$\begin{aligned}
&\mathbf{z}^\top(t)\mathbb{E}\left[\left(\widetilde{\mathbf{L}}(t) \otimes \mathbf{I}_M\right)^2\right]\mathbf{z}(t) \overset{(q0)}{\leq} \frac{c_5\left\|\mathbf{z}_{\mathcal{C}^\perp}\right\|^2}{(t+1)^{1+\epsilon}} \\
&\mathbf{z}^\top(t)\left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right)^2\mathbf{z}(t) \overset{(q1)}{\leq} \lambda_N^2(\overline{\mathbf{L}})||\mathbf{z}_{C\perp}(t)||^2; \\
&\mathbf{z}^\top(t)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \geq c_1||\mathbf{z}(t)||^2 \overset{(q2)}{\geq} 0; \\
&\mathbf{z}^\top(t)\left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right)\mathbf{z}(t) \overset{(q3)}{\geq} \lambda_2(\overline{\mathbf{L}})\left\|\mathbf{z}_{C\perp}(t)\right\|^2; \\
&\mathbf{z}^\top(t)\left(\overline{\mathbf{L}} \otimes \mathbf{I}_M\right)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right) - \mathbf{f}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}\right)\right) \\
&\overset{(q4)}{\leq} c_2\left\|\mathbf{z}(t)\right\|^2.
\end{aligned} \tag{E.59}$$

Here, we recall that $\lambda_N(\overline{\mathbf{L}})$ is the largest eigenvalue of matrix $\overline{\mathbf{L}}$. Further, c_1 is defined in Assumption 6.4.3, and c_2, c_5 are appropriately chosen positive constants. Here, $\mathbf{z}_{C\perp}(t) = \mathbf{z}(t) - \mathbf{z}_C(t)$, where $\mathbf{z}_C(t)$ is the projection of $\mathbf{z}(t)$ on the consensus subspace $\mathcal{C}$. Inequality $(q0)$ holds because, as noted above, there holds that $\mathbb{E}\left[\widetilde{\mathbf{L}}^2_{i,j}(t)\right] \leq \frac{\rho_0\beta_0}{(t+1)^{1+\epsilon}}$, for $\{i,j\} \in E, i \neq j$. Specifically, constant c_5 can be taken to equal $2\,N^3\,\rho_0\beta_0$. Next, inequalities $(q1)$ and $(q3)$ follow from the properties of the Laplacian. Inequality $(q2)$ follows from Assumption 6.4.3; and $(q4)$ follows from Assumption 6.4.2 since we have that $\|\nabla \mathbf{f}_n\left(\mathbf{x}_n(t)\right)\|$ is uniformly bounded from above by k_n for all n and hence, we have that $\|\mathbf{G}\left(\mathbf{x}(t)\right)\| \leq \max_{n=1,\cdots,N} k_n$. That is, c_2 can

be taken as $(\max_{n=1,\cdots,N} k_n)^2$ $(\max_{n=1,\cdots,N} \|\mathbf{R}_n^{-1}\|)$ $\|\overline{\mathbf{L}}\|$. We also have

$$\mathbb{E}\left[\left\|\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{y}(t)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\right\|^2\right]\leq c_4, \tag{E.60}$$

for some constant $c_4 > 0$. In (E.60), we use the fact that the noise process under consideration has finite covariance. We also use the fact that, almost surely, $\|\mathbf{G}\left(\mathbf{x}(t)\right)\| \leq \max_{n=1,\cdots,N} k_n$, which in turn follows from Assumption 6.4.2. In particular, c_4 may be taken as $(\max_{n=1,\cdots,N} k_n)^2$ $(\max_{n=1,\cdots,N} \|\mathbf{R}_n^{-1}\|)^2$ $(\max_{n=1,\cdots,N} \|\mathbf{R}_n\|)^2$. We further have that,

$$\left\|\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\right\|^2\leq c_3\left\|\mathbf{z}(t)\right\|^2, \tag{E.61}$$

where $c_3 > 0$ is a constant. It is to be noted that (E.61) follows from the Lipschitz continuity in Assumption 6.4.2 and the result that $\|\mathbf{G}\left(\mathbf{x}(t)\right)\| \leq \max_{n=1,\cdots,N} k_n$. That is, c_3 may be taken as $(\max_{n=1,\cdots,N} k_n)^4$ $(\max_{n=1,\cdots,N} \|\mathbf{R}_n^{-1}\|)^2$. Applying the above bounds, we obtain, after some algebraic manipulations,

$$\begin{aligned}\mathbb{E}\left[V(t+1)|\mathcal{F}_t\right] &\leq (1+c_8\alpha_t^2)V(t)\\ &-c_9\left(\beta_t-\frac{c_5}{(t+1)^{\tau_1+\epsilon}}\right)\left\|\mathbf{z}_{\mathcal{C}^\perp}\right\|^2+c_6\alpha_t^2,\end{aligned} \tag{E.62}$$

where c_6, c_8, c_9 are appropriately chosen positive constants. In particular, c_6 may be taken as $c_6 = c_4$; c_8 may be taken as $\beta_0^2\left(\lambda_N(\overline{\mathbf{L}})\right)^2/\alpha_0^2+2\beta_0\sqrt{c_3}+c_3$; and c_9 may be taken as $2\,\lambda_2(\overline{\mathbf{L}})$. As $\frac{c_5}{(t+1)^{1+\epsilon}}$ goes to zero faster than β_t, $\exists t_2$ such that $\forall t \geq t_2$, $\beta_t \geq \frac{c_5}{(t+1)^{1+\epsilon}}$. By the above construction we obtain $\forall t \geq t_2$,

$$\mathbb{E}[V(t+1)|\mathcal{F}_t]\leq(1+\alpha_t^2)V(t)+\widehat{\alpha}_t^2, \tag{E.63}$$

where $\widehat{\alpha}(t)=\sqrt{c_6}\alpha_t$. The product $\prod_{s=t}^{\infty}(1+\alpha_s^2)$ exists for all t. Now let $\{W(t)\}$ be such that

$$W(t)=\left(\prod_{s=t}^{\infty}(1+\alpha_s^2)\right)V(t)+\sum_{s=t}^{\infty}\widehat{\alpha}_s^2,\ \forall t\geq t_2. \tag{E.64}$$

By (E.64), it can be shown that $\{W(t)\}$ satisfies,

$$\mathbb{E}[W(t+1)|\mathcal{F}_t]\leq W(t). \tag{E.65}$$

Hence, $\{W(t)\}$ is a non-negative supermartingale and converges a.s. to a bounded random variable W^* as $t \to \infty$. It then follows from (E.64) that $V(t) \to W^*$ as $t \to \infty$. Thus, we conclude that the desired claim holds. □

We now use Lemma C.3.1 for establishing the convergence of the estimate sequence. Following similar steps as in the proof of Lemma D.0.1, for t large enough

$$\begin{aligned}\mathbb{E}[V(t+1)|\mathcal{F}_t] &\leq \left(1-2c_1\alpha_t+c_7\alpha_t^2\right)V(t)+c_6\alpha_t^2\\ &\leq V(t)+c_6\alpha_t^2,\end{aligned} \tag{E.66}$$

as for t large enough, $-2c_1\alpha_t + c_7\alpha_t^2 < 0$. Here, c_7 is appropriately chosen positive constant that may be

taken as $\beta_0^2 \left(\lambda_N(\overline{\mathbf{L}})\right)^2/\alpha_0^2 + 2\beta_0\sqrt{c_3} + c_3$. Now, consider the $\{\mathcal{F}_t\}$-adapted process $\{V_1(t)\}$ defined as follows

$$\begin{aligned} V_1(t) &= V(t) + c_6 \sum_{s=t}^{\infty} \alpha_s^2 \\ &= V(t) + c_6 \alpha_0^2 \sum_{s=t}^{\infty} (t+1)^{-2}. \end{aligned} \tag{E.67}$$

Since $\{(t+1)^{-2}\}$ is summable, the process $\{V_1(t)\}$ is bounded from above. Moreover, it also follows that $\{V_1(t)\}_{t \geq t_1}$ is a supermartingale and hence converges a.s. to a finite random variable. By definition from (E.67), we also have that $\{V(t)\}$ converges to a non-negative finite random variable V^*. Finally, from (E.66), we have that,

$$\mathbb{E}[V(t+1)] \leq (1 - c_1\alpha_t)\, \mathbb{E}[V(t)] + c_6 \alpha_0^2\, (t+1)^{-2}, \tag{E.68}$$

for t large enough. The sequence $\{V(t)\}$ then falls under the purview of Lemma C.3.1, and we have $\mathbb{E}[V(t)] \to 0$ as $t \to \infty$. Finally, by Fatou's Lemma, where we use the non-negativity of the sequence $\{V(t)\}$, we conclude that

$$0 \leq \mathbb{E}[V^*] \leq \liminf_{t \to \infty} \mathbb{E}[V(t)] = 0, \tag{E.69}$$

which thus implies that $V^* = 0$ a.s. Hence, $\|\mathbf{z}(t)\| \to 0$ a.s. as $t \to \infty$, and the desired assertion follows. □

Proof of Theorem 6.7.2. We can now see that the sequence $\{V(t)\}$ then falls under the purview of Lemma C.1.3, and we have

$$\begin{aligned} &\limsup_{t \to \infty} (t+1)\mathbb{E}[V(t+1)] < \infty \\ &\Rightarrow \mathbb{E}[V(t)] = O\left(\frac{1}{t}\right). \end{aligned} \tag{E.70}$$

Inequality (58) now clearly implies that, for each agent n, there holds:

$$\mathbb{E}[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2] = O\left(\frac{1}{t}\right). \tag{E.71}$$

The communication cost $\mathcal{C}_t$ for the proposed $\mathcal{CREDO} - \mathcal{NL}$ algorithm is given by $\mathcal{C}_t = \Theta\left(t^{\frac{\epsilon+1}{2}}\right)$, and thus the assertion follows in conjunction with (E.71). □

Appendix F

Proofs of Theorems in Chapter 7

Proof of Theorem 7.5.1. Define the sequence, $\{\widehat{\mathbf{x}}(t)\}$, as $\widehat{\mathbf{x}}(t) = \widetilde{\mathbf{x}}(t) - \mathcal{P}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)$. Then, we have,

$$\begin{aligned}\widehat{\mathbf{x}}(t+1) &= \widehat{\mathbf{x}}(t) - \left(\beta_t \overline{\mathbf{L}}_{\mathcal{P}} + \alpha_t \mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\mathbf{G}_H^\top\mathcal{P}\right)\widehat{\mathbf{x}}(t) \\ &- \beta_t \widetilde{\mathbf{L}}_{\mathcal{P}}(t)\widehat{\mathbf{x}}(t) + \alpha_t \mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\left(\mathbf{y}(t) - \mathbf{G}_H^\top\mathcal{P}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)\right). \end{aligned} \tag{F.1}$$

It is clear that $\{\widehat{\mathbf{x}}(t)\}$ is Markov with respect to its natural filtration $\{\mathcal{F}_t^{\widehat{\mathbf{x}}}\}$. Now, define the function $V : \mathbb{R}^{N^2} \longmapsto \mathbb{R}_+$ as, $V(\mathbf{y}) = \|\mathbf{y}\|^2$, for all $\mathbf{y}$. We note that

$$\mathbb{E}_{\theta^*}\left[V(\widehat{\mathbf{x}}(t+1)) \mid \mathcal{F}_t^{\widetilde{\mathbf{x}}}\right] = \mathbb{E}_{\theta^*}\left[V(\widehat{\mathbf{x}}(t+1)) \mid \widehat{\mathbf{x}}(t)\right] \tag{F.2}$$

By basic algebraic manipulations, we have,

$$\begin{aligned}&\mathbb{E}_{\theta^*}\left[V(\widehat{\mathbf{x}}(t+1)) \mid \widehat{\mathbf{x}}(t)\right] \\ &\leq \widehat{\mathbf{x}}(t)^\top\left(\mathbf{I} - \beta_t\overline{\mathbf{L}}_{\mathcal{P}} - \alpha_t\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\mathbf{G}_H^\top\mathcal{P}\right)^2\widehat{\mathbf{x}}(t) \\ &+ \beta_t^2\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{L}}_{\mathcal{P}}(t)\widehat{\mathbf{x}}(t)\right\|^2\right] \\ &+ \alpha_t^2\mathbb{E}_{\theta^*}\left[\left\|\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\left(\mathbf{y}(t) - \mathbf{G}_H^\top\mathcal{P}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)\right)\right\|^2\right]. \end{aligned} \tag{F.3}$$

We note that $\beta_t\overline{\mathbf{L}}_{\mathcal{P}} + \alpha_t\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\mathbf{G}_H^\top\mathcal{P}$ is uniformly elliptic on the subspace $\mathcal{S}_{\mathcal{P}}$, and it is precisely the subspace where $\{\widehat{\mathbf{x}}(t)\}$ resides. We thus prove the result by showing convergence to zero of the sequence $\{\widehat{\mathbf{x}}(t)\}$ through the subspace $\mathcal{S}_{\mathcal{P}}$. To this end, using the fact, that, for $\mathbf{y} \in \mathcal{S}_{\mathcal{P}}$,

$$\mathbf{y}^\top\left(\frac{\beta_0}{\alpha_0}\mathbf{L}_{\mathcal{P}} + \mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\mathbf{G}_H^\top\mathcal{P}\right)\mathbf{y} \geq c_1\|\mathbf{y}\|^2, \text{ a.s.} \tag{F.4}$$

By choosing, t_1 sufficiently large, we have for $\widehat{\mathbf{x}}(t)^\top \in \mathcal{S}_{\mathcal{P}}$ for all $t \geq t_1$,

$$\begin{aligned}&\widehat{\mathbf{x}}(t)^\top\left(\beta_t^2\overline{\mathbf{L}}_{\mathcal{P}}^2 + \beta_t^2\mathbb{E}_{\theta^*}\left\|\widetilde{\mathbf{L}}_{\mathcal{P}}(t)\right\|^2 - \beta_t\overline{\mathbf{L}}_{\mathcal{P}}\right)\widehat{\mathbf{x}}(t) \\ &\leq \left(c_1'\beta_t^2 - c_3'\beta_t\right)\|\widehat{\mathbf{x}}(t)\|^2 \leq 0, \end{aligned} \tag{F.5}$$

where equality exists if $\widehat{\mathbf{x}}(t) = \mathcal{P}\left(\mathbf{1}_N \otimes \mathbf{a}\right)$, where $\mathbf{a} \in \mathbb{R}^N$. Thus, we obtain the following inequality:

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[V(\widehat{\mathbf{x}}(t+1)) \mid \widehat{\mathbf{x}}(t) = \mathbf{y}\right] - V(\mathbf{y}) \leq c_{11}\alpha_t^2\left(1 + \|\mathbf{y}\|^2\right) \\ - \alpha_t c_{10}\|\mathbf{y}\|^2 \end{aligned} \tag{F.6}$$

for all $\mathbf{y} \in \mathcal{S}_{\mathcal{P}}$. Now, define the function $W : \mathbb{T}_+ \times \mathbb{R}^{N^2} \longmapsto \mathbb{R}_+$:

$$W(t, \mathbf{y}) = (1 + V(\mathbf{y}))\prod_{j=t}^{\infty}(1 + c_{11}\alpha_j^2). \tag{F.7}$$

From (F.6) it can be shown that, for $\mathbf{y} \in \mathcal{S}_{\mathcal{P}}$,

$$\begin{aligned}&\mathbb{E}_{\theta^*}\left[W(t+1, \widehat{\mathbf{x}}(t+1)) \mid \widehat{\mathbf{x}}(t) = \mathbf{y}\right] - W(t, \mathbf{y}) \\ &\leq -\alpha_t c_{10}\|\mathbf{y}\|^2\left(\prod_{j=t+1}^{\infty}(1 + c_{11}\alpha_j^2)\right) \\ &\leq -\alpha_t c_{10}\|\mathbf{y}\|^2 \end{aligned} \tag{F.8}$$

Now consider $\varepsilon > 0$, and let V_ε denote the set

$$V_\varepsilon = \{\mathbf{y} \in \mathbb{R}^{N^2} \mid \|\mathbf{y}\| \geq \varepsilon\} \cap \mathcal{S}_{\mathcal{P}} \tag{F.9}$$

Also, define τ_ε to be the exit time of the process $\{\widehat{\mathbf{x}}(t)\}$ from V_ε, i.e.,

$$\tau_\varepsilon = \inf\{i \in \mathbb{T}_+ \mid \widehat{\mathbf{x}}(t) \notin V_\varepsilon\} \tag{F.10}$$

We now show that $\tau_\varepsilon < \infty$ a.s. For mathematical simplicity, assume $\widehat{\mathbf{x}}(0) \in V_\varepsilon$. Consider the function

$$\widetilde{W}(t, \mathbf{y}) = W(t, \mathbf{y}) + c_{10}\varepsilon^2\sum_{j=0}^{t-1}\alpha_j \tag{F.11}$$

By (F.8) it follows that, for $\mathbf{y} \in V_\varepsilon$,

$$\mathbb{E}_{\theta^*}\left[W(t+1, \widehat{\mathbf{x}}(t+1)) \mid \widehat{\mathbf{x}}(t) = \mathbf{y}\right] - W(t, \mathbf{y}) \leq -\alpha_t c_{10}\varepsilon^2 \tag{F.12}$$

and hence, it can be shown that, for $\mathbf{y} \in V_\varepsilon$,

$$\mathbb{E}_{\theta^*}\left[\widetilde{W}(t+1, \widehat{\mathbf{x}}(t+1)) \mid \widehat{\mathbf{x}}(t) = \mathbf{y}\right] - \widetilde{W}(t, \mathbf{y}) \leq 0 \tag{F.13}$$

Hence, we have that the stopped process $\{\widetilde{W}(\max\{t, \tau_\varepsilon\}, \widehat{\mathbf{x}}(\max\{t, \tau_\varepsilon\}))\}$ is a super martingale. Being nonnegative it converges a.s. as $t \to \infty$. By (F.12), we then conclude that the following term converges,

$$\lim_{t\to\infty} c_{10}\varepsilon^2\sum_{j=0}^{(t\wedge\tau_\varepsilon)-1}\alpha_j \text{ converges a.s.} \tag{F.14}$$

Since, $\sum_{t\in\mathbb{T}_+}\alpha_t = \infty$, the above is possible, only if, $\tau_\varepsilon < \infty$ a.s.

We thus note, that the process $\{\widehat{\mathbf{x}}(t)\}$ leaves the set V_ε almost surely in finite time. Since, the process is constrained to lie in $\mathcal{S}_\mathcal{P}$ at all times, the finite time exit from V_ε suggests,

$$\mathbb{P}_{\theta^*}\left(\inf\{t \in \mathbb{T}_+ \mid \|\widehat{\mathbf{x}}(t)\| < \varepsilon\} < \infty\right) = 1 \tag{F.15}$$

Since $\varepsilon > 0$ is arbitrary, a subsequence almost surely converges to zero, and we have

$$\mathbb{P}_{\theta^*}\left(\liminf_{t\to\infty} \|\widehat{\mathbf{x}}(t)\| = 0\right) = 1 \tag{F.16}$$

Now going back to (F.6) and noting that $\{\widehat{\mathbf{x}}(t)\}$ takes values in $\mathcal{S}_\mathcal{P}$, we conclude that the process $\{V(\widehat{\mathbf{x}}(t))\}$ is a nonnegative supermartingale. Hence,

$$\mathbb{P}_{\theta^*}\left(\lim_{t\to\infty} V(\widehat{\mathbf{x}}(t)) \text{ exists}\right) = 1 \tag{F.17}$$

Also, by (F.16)

$$\mathbb{P}_{\theta^*}\left(\liminf_{t\to\infty} V(\widehat{\mathbf{x}}(t)) = 0\right) = 1 \tag{F.18}$$

and we conclude that

$$\mathbb{P}_{\theta^*}\left(\lim_{i\to\infty} \|\widehat{\mathbf{x}}(t)\| = 0\right) = 1 \tag{F.19}$$

□

Proof of Theorem 7.5.2. From (F.2)-(F.4) in the proof of Theorem 7.5.1 we have, for $t \geq t_1$ (t_1 chosen appropriately large) and using the property that $\widehat{\mathbf{x}}(t)$ resides in $\mathcal{S}_\mathcal{P}$

$$\begin{aligned}
&\mathbb{E}_{\theta^*}\left[V(\widehat{\mathbf{x}}(t+1)) \mid \widehat{\mathbf{x}}(t)\right] \leq (1 - c_1\alpha_t)\|\widehat{\mathbf{x}}(t)\|^2 + \alpha_t^2 c_2 \\
&\Rightarrow \mathbb{E}_{\theta^*}\left[\|\widehat{\mathbf{x}}(t+1)\|^2\right] \leq (1 - c_1\alpha_t)\|\widehat{\mathbf{x}}(t)\|^2 + \alpha_t^2 c_2 \\
&\Rightarrow \mathbb{E}\left[\|\widetilde{\mathbf{x}}(t) - \mathcal{P}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)\|^2\right] = O\left(\frac{1}{t}\right).
\end{aligned}$$

for appropriately chosen constants c_1 and c_2, where the conclusion in the last line follows from Lemma C.3.1. □

Proof of Theorem 7.5.3. Let the number of agents interested in the i-th entry of $\boldsymbol{\theta}^*$ be Q_i. To get the vector of estimates of the i-th entry of $\boldsymbol{\theta}^*$, left multiply the selector matrix $\mathcal{S}_i \in \mathbb{R}^{Q_i \times N^2}$ and noting that $S_i\mathbf{L}_\mathcal{P}(t)\widetilde{\mathbf{x}}(t) = \mathbf{L}_{\mathcal{P},i}(t)\widetilde{\mathbf{x}}(i,t)$, where $\mathbf{L}_{\mathcal{P},i}(t) \in \mathbb{R}^{Q_i \times Q_i}$ is the subgraph induced by the interest sets for the i-th entry of $\boldsymbol{\theta}^*$, which is connected as a result of a sufficient condition which enforced Assumption 7.4.3 and $\widetilde{\mathbf{x}}(i,t) \in \mathbb{R}^{Q_i}$ is the vector of estimates for the i-th entry of $\boldsymbol{\theta}^*$.
A vector $\mathbf{z} \in \mathbb{R}^{N^2}$ may be decomposed as $\mathbf{z} = \mathbf{z}_\mathcal{C} + \mathbf{z}_{\mathcal{C}^\perp}$ with $\mathbf{z}_\mathcal{C}$ denoting its projection on the consensus or agreement subspace $\mathcal{C}$, $\mathcal{C} = \left\{\mathbf{z} \in \mathbb{R}^{N^2} | \mathbf{z} = \mathbf{1}_N \otimes \mathbf{a} \textit{ for some } \mathbf{a} \in \mathbb{R}^N\right\}$. We first prove the following Lemma regarding the mean connectedness of the subgraphs $\mathbf{L}_{\mathcal{P},i}(t)$.

Lemma F.0.1. *Let $\{\mathbf{z}_t\}$ be an $\mathbb{R}^{N^2}$ valued $\mathcal{F}_t$-adapted process such that $\mathbf{z}_t \in \mathcal{C}^\perp$ for all t. Also, let $\{\mathbf{L}_t\}$ be an i.i.d. sequence of Laplacian matrices as in assumption 7.4.2 that satisfies*

$$\lambda_2\left(\overline{\mathbf{L}}\right) = \lambda_2\left(\mathbb{E}\left[\mathbf{L}_t\right]\right) > 0, \tag{F.20}$$

where $\mathbf{L}_t$ is $\mathcal{F}_{t+1}$-adapted and independent of $\mathcal{F}_t$ for all t.

$$\left\|\left(\mathbf{I}_{N^2} - \left(\mathbf{L}(t) \otimes \mathbf{I}_N\right)\right)\mathbf{z}_t\right\| \leq (1 - r_t)\left\|\mathbf{z}_t\right\|, \tag{F.21}$$

where $\{r_t\}$ is a $\mathbb{R}^+$ valued $\mathcal{F}_{t+1}$ process satisfying

$$\mathbb{E}\left[r_t|\mathcal{F}_t\right] \geq \underline{p}\beta_t \frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{4|\mathcal{L}|}, \tag{F.22}$$

where $\mathcal{L}$ denotes the set of all possible Laplacians.

The following Lemma in addition to Lemma C.3.1 will be used to quantify the rate of convergence of distributed vector or matrix valued recursions to their network-averaged behavior.

Lemma F.0.2. *Let $\{z_t\}$ be an $\mathbb{R}^+$ valued $\mathcal{F}_t$-adapted process that satisfies*

$$z_{t+1} \leq (1 - r_1(t))\, z_t + r_2(t) U_t (1 + J_t),$$

where $\{r_1(t)\}$ is an $\mathcal{F}_{t+1}$-adapted process, such that for all t, $r_1(t)$ satisfies $0 \leq r_1(t) \leq 1$ and

$$a_1 \leq \mathbb{E}\left[r_1(t)|\mathcal{F}_t\right] \leq \frac{1}{(t+1)^{\delta_1}}$$

with $a_1 > 0$ and $0 \leq \delta_1 < 1$. The sequence $\{r_2(t)\}$ is deterministic and $\mathbb{R}^+$ valued and satisfies $r_2(t) \leq \frac{a_2}{(t+1)^{\delta_2}}$ with $a_2 > 0$ and $\delta_2 > 0$. Further, let $\{U_t\}$ and $\{J_t\}$ be $\mathbb{R}^+$ valued $\mathcal{F}_t$ and $\mathcal{F}_{t+1}$ adapted processes, respectively, with $\sup_{t\geq 0}\|U_t\| < \infty$ a.s. The process $\{J_t\}$ is i.i.d. with J_t independent of $\mathcal{F}_t$ for each t and satisfies the moment condition $\mathbb{E}\left[\|J_t\|^{2+\epsilon_1}\right] < \kappa < \infty$ for some $\epsilon_1 > 0$ and a constant $\kappa > 0$. Then, for every δ_0 such that $0 \leq \delta_0 < \delta_2 - \delta_1 - \frac{1}{2+\epsilon_1}$, we have $(t+1)^{\delta_0} z_t \to 0$ a.s. as $t \to \infty$.

Proof of Lemma F.0.1. Let $\mathcal{L}$ denote the set of possible Laplacian matrices which is necessarily finite. Since the set of Laplacians is finite, we have,

$$\underline{p} = \inf_{\mathbf{L}\in\mathcal{L}} p_{\mathbf{L}} > 0, \tag{F.23}$$

with $p_L = \mathbb{P}\left(\mathbf{L}(t) = \mathbf{L}\right)$ for each $\mathbf{L} \in \mathcal{L}$ such that $\sum_{\mathbf{L}\in\mathcal{L}} p_{\mathbf{L}} = 1$. We also have that $\lambda_2\left(\overline{\mathbf{L}}\right) > 0$ implies that for every $\mathbf{z} \in \mathcal{C}^\perp$, where,

$$\mathcal{C} = \left\{\mathbf{x} | \mathbf{x} = \mathbf{1}_N \otimes \mathbf{a}, \mathbf{a} \in \mathbb{R}^N\right\}, \tag{F.24}$$

we have,

$$\sum_{\mathbf{L}\in\mathcal{L}} \mathbf{z}^\top \mathbf{L}\mathbf{z} \geq \sum_{\mathbf{L}\in\mathcal{L}} \mathbf{z}^\top p_{\mathbf{L}} \mathbf{L}\mathbf{z} = \mathbf{z}^\top \overline{\mathbf{L}}\mathbf{z} \geq \lambda_2\left(\overline{\mathbf{L}}\right)\|\mathbf{z}\|^2. \tag{F.25}$$

Owing to the finite cardinality of $\mathcal{L}$ and (F.25), we also have that for each $\mathbf{z} \in \mathcal{C}^{\perp}$,$\exists \mathbf{L}_{\mathbf{z}} \in \mathcal{L}$ such that,

$$\mathbf{z}^{\top}\mathbf{L}_{\mathbf{z}}\mathbf{z} \geq \frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{|\mathcal{L}_t|}\left\|\mathbf{z}\right\|^2 \tag{F.26}$$

Moreover, since $\mathcal{L}$ is finite, the mapping $L_{\mathbf{z}} : \mathcal{C}^{\perp} \mapsto \mathcal{L}$ can be realized as a measurable function. For each, $\mathbf{L} \in \mathcal{L}$, the eigen values of $\mathbf{I}_{N^2} - \beta_t\left(\mathbf{L} \otimes \mathbf{I}_N\right)$ are given by N repetitions of 1 and $1 - \beta_t\lambda_n\left(\mathbf{L}\right)$, where $2 \leq n \leq N$. Thus, for $t \geq t_0$, $\left\|\mathbf{I}_{N^2} - \beta_t\left(\mathbf{L} \otimes \mathbf{I}_N\right)\right\| \leq 1$ and $\left\|\left(\mathbf{I}_{N^2} - \beta_t\left(\mathbf{L} \otimes \mathbf{I}_N\right)\right)\mathbf{z}\right\| \leq \left\|\mathbf{z}\right\|$. Hence, we can define a jointly measurable function $r_{\mathbf{L},\mathbf{z}}$ given by,

$$r_{\mathbf{L},\mathbf{z}} = \begin{cases} 1 & \text{if } t < t_0 \text{ or } \mathbf{z} = \mathbf{0} \\ 1 - \frac{\left\|\left(\mathbf{I}_{NM} - \beta_t(\mathbf{L}\otimes\mathbf{I}_M)\right)\mathbf{z}\right\|}{\left\|\mathbf{z}\right\|} & \text{otherwise,} \end{cases} \tag{F.27}$$

which satisfies $0 \leq r_{\mathbf{L},\mathbf{z}} \leq 1$ for each $(\mathbf{L}, \mathbf{z})$. Define $\{r_t\}$ to be a $\mathcal{F}_{t+1}$ process given by, $r_t = r_{\mathbf{L},\mathbf{z}_t}$ for each t and $\left\|\left(\mathbf{I}_{N^2} - \beta_t\left(\mathbf{L} \otimes \mathbf{I}_N\right)\right)\mathbf{z}_t\right\| = \left(1 - r_t\right)\left\|\mathbf{z}_t\right\|$ a.s. for each t. Then, we have,

$$\begin{aligned} &\left\|\left(\mathbf{I}_{N^2} - \beta_t\left(\mathbf{L}_{\mathbf{z}_t} \otimes \mathbf{I}_N\right)\right)\mathbf{z}_t\right\|^2 \\ &= \mathbf{z}_t^{\top}\left(\mathbf{I}_{N^2} - 2\beta_t\left(\mathbf{L}_{\mathbf{z}_t} \otimes \mathbf{I}_N\right)\right)\mathbf{z}_t \\ &+ \mathbf{z}_t^{\top}\beta_t^2\left(\mathbf{L}_{\mathbf{z}_t} \otimes \mathbf{I}_N\right)^2\mathbf{z}_t \\ &\leq \left(1 - 2\beta_t\frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{|\mathcal{L}|}\right)\left\|\mathbf{z}_t\right\|^2 + c_1\beta_t^2\left\|\mathbf{z}_t\right\|^2 \\ &\leq \left(1 - \beta_t\frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{|\mathcal{L}|}\right)\left\|\mathbf{z}_t\right\|^2 \end{aligned} \tag{F.28}$$

where we have used the boundedness of the Laplacian matrix. With the above development in place, choosing an appropriate t_1 (making t_0 larger if necessary), for all $t \geq t_1$, we have,

$$\left\|\left(\mathbf{I}_{N^2} - \beta_t\left(\mathbf{L}_{\mathbf{z}_t} \otimes \mathbf{I}_N\right)\right)\mathbf{z}_t\right\| \leq \left(1 - \beta_t\frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{4|\mathcal{L}|}\right)\left\|\mathbf{z}_t\right\|^2. \tag{F.29}$$

Then, from (F.29), we have,

$$\begin{aligned} &\mathbb{E}\left[\left\|\left(\mathbf{I}_{N^2} - \beta_t\left(\mathbf{L}_{\mathbf{z}_t} \otimes \mathbf{I}_N\right)\right)\mathbf{z}_t\right\|\middle|\mathcal{F}_t\right] \\ &= \sum_{\mathbf{L}\in\mathcal{L}} p_{\mathbf{L}}\left(1 - r_{\mathbf{L},\mathbf{z}_t}\right)\left\|\mathbf{z}_t\right\| \\ &\leq \left(1 - \left(\underline{p}\beta_t\frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{4|\mathcal{L}|} + \sum_{\mathbf{L}\neq\mathbf{L}_{\mathbf{z}_t}}\right)\right)\left\|\mathbf{z}_t\right\|. \end{aligned} \tag{F.30}$$

Since, $\sum_{\mathbf{L}\neq\mathbf{L}_{\mathbf{z}_t}} p_{\mathbf{L}} r_{\mathbf{L},\mathbf{z}_t} \geq 0$, we have for all $t \geq t_1$,

$$\begin{aligned} &\left(1 - \mathbb{E}\left[r_t|\mathcal{F}_t\right]\right)\left\|\mathbf{z}_t\right\| \\ &= \mathbb{E}\left[\left\|\left(\mathbf{I}_{N^2} - \beta_t\left(\mathbf{L}_{\mathbf{z}_t} \otimes \mathbf{I}_N\right)\right)\mathbf{z}_t\right\|\middle|\mathcal{F}_t\right] \\ &\leq \left(1 - \underline{p}\beta_t\frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{4|\mathcal{L}|}\right)\left\|\mathbf{z}_t\right\|. \end{aligned} \tag{F.31}$$

As $r_t = 1$ on the set $\{\mathbf{z}_t = 0\}$, we have that,

$$\mathbb{E}\left[r_t|\mathcal{F}_t\right] \geq \underline{p}\beta_t \frac{\lambda_2\left(\overline{\mathbf{L}}\right)}{4|\mathcal{L}|}. \tag{F.32}$$

Thus, we have established that,

$$\left\|\left(\mathbf{I}_{N^2} - \left(\mathbf{L}(t) \otimes \mathbf{I}_N\right)\right)\mathbf{z}_t\right\| \leq (1 - r_t)\left\|\mathbf{z}_t\right\|, \tag{F.33}$$

where $\{r_t\}$ is a $\mathbb{R}^+$ valued $\mathcal{F}_{t+1}$ process satisfying (F.32). □

With the above development in place, consider the residual process $\{\mathbf{x}^\dagger(t)\}$ given by $\mathbf{x}^\dagger(i,t) = \widetilde{\mathbf{x}}(i,t) - \mathbf{1}_{Q_i} \otimes \widetilde{\mathbf{x}}_{\text{avg,i}}(t)$, where i denotes the i-th entry of $\boldsymbol{\theta}^*$ and $\mathbf{x}^\dagger(t) = \left[\mathbf{x}^\dagger(1,t), \cdots, \mathbf{x}^\dagger(N,t)\right]^\top$. Thus, we have that the process $\{\mathbf{x}^\dagger(i,t)\}$ satisfies the recursion,

$$\mathbf{x}^\dagger(i,t+1) = \left(\mathbf{I}_{Q_i} - \mathbf{L}_{\mathcal{P},i}(t)\right)\mathbf{x}^\dagger(i,t) + \alpha_t \widetilde{\mathbf{z}}(i,t), \tag{F.34}$$

where the process $\{\widetilde{\mathbf{z}}(i,t)\}$ is given by

$$\widetilde{\mathbf{z}}(i,t) = \left(\mathbf{I}_{Q_i} - \frac{1}{Q_i}\mathbf{1}_{Q_i}\mathbf{1}_{Q_i}^\top\right) \times \mathcal{S}_i\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\left(\mathbf{y}(t) - \mathbf{G}_H^\top\mathcal{P}\widetilde{\mathbf{x}}(t)\right). \tag{F.35}$$

From (F.35), we also have,

$$\widetilde{\mathbf{z}}(i,t) = \overline{\mathbf{J}}_{i,t} + \overline{\mathbf{U}}_{i,t}, \tag{F.36}$$

where,

$$\begin{aligned}
\overline{\mathbf{J}}_{i,t} &= \left(\mathbf{I}_{Q_i} - \frac{1}{Q_i}\mathbf{1}_{Q_i}\mathbf{1}_{Q_i}^\top\right) \\
&\times \mathcal{S}_i\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\left(\mathbf{y}(t) - \mathbf{G}_H^\top\mathcal{P}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)\right) \\
\overline{\mathbf{U}}_t &= \left(\mathbf{I}_{Q_i} - \frac{1}{Q_i}\mathbf{1}_{Q_i}\mathbf{1}_{Q_i}^\top\right) \\
&\times \mathcal{S}_i\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\left(\mathbf{G}_H^\top\mathcal{P}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^*\right) - \mathbf{G}_H^\top\mathcal{P}\widetilde{\mathbf{x}}(t)\right).
\end{aligned} \tag{F.37}$$

By Theorem 7.5.1, we also have that, the process $\{\widetilde{\mathbf{x}}(i,t)\}$ is bounded. Hence, there exists an $\mathcal{F}_t$-adapted process $\{\widetilde{U}_{i,t}\}$ such that $\left\|\overline{\mathbf{U}}_{i,t}\right\| \leq \widetilde{U}_{i,t}$ and $\sup_{t\geq 0}\widetilde{U}_{i,t} < \infty$ a.s.. Furthermore, denote the process $U_{i,t}$ as follows,

$$U_{i,t} = \max\left\{\widetilde{U}_{i,t}, \left\|\mathbf{I}_{Q_i} - \frac{1}{Q_i}\mathbf{1}_{Q_i}\mathbf{1}_{Q_i}^\top\right\|\right\}. \tag{F.38}$$

With the above development in place, we conclude,

$$\left\|\overline{\mathbf{U}}_{i,t}\right\| + \left\|\overline{\mathbf{J}}_{i,t}\right\| \leq U_{i,t}\left(1 + J_{i,t}\right), \tag{F.39}$$

where $J_{i,t} = \left\|\mathbf{y}(t) - \mathbf{G}_H^\top\mathcal{P}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)\right\|$ and $\mathbb{E}_{\boldsymbol{\theta}}\left[J_{i,t}^{2+\epsilon}\right] < \infty$. Then, from (F.21)-(F.34) we have,

$$\left\|\mathbf{x}^\dagger(i,t+1)\right\| \leq (1 - r_t)\left\|\mathbf{x}^\dagger(i,t)\right\| + \alpha_t U_{i,t}(1 + J_{i,t}), \tag{F.40}$$

which then falls under the purview of Lemma D.0.3 and hence we have the assertion,

$$\mathbb{P}\left(\lim_{t\to\infty}(t+1)^{\delta_0}\left(\widetilde{\mathbf{x}}(i,t)-\mathbf{1}_{Q_i}\otimes\widetilde{\mathbf{x}}_{\text{avg,i}}(t)\right)=0\right)=1, \tag{F.41}$$

where $0<\delta_0<1-\tau_1$ and hence δ_0 can be chosen to be $1/2+\delta$, where $\delta>0$ and we finally have,

$$\mathbb{P}\left(\lim_{t\to\infty}(t+1)^{\frac{1}{2}+\delta}\left(\widetilde{\mathbf{x}}(t)-\mathbf{1}_N\otimes\widetilde{\mathbf{x}}_{\text{avg}}(t)\right)=0\right)=1, \tag{F.42}$$

as the above analysis can be repeated each entry i of the parameter of interest $\boldsymbol{\theta}^*$.
The proof of Theorem 7.5.3 needs Lemma C.1.5 from Fabian (1968) concerning the asymptotic normality of the stochastic recursions. Multiplying the selection matrix, we have,

$$\begin{aligned}
&\widetilde{\mathbf{x}}(i,t+1)=\widetilde{\mathbf{x}}(i,t)-\mathbf{L}_{\mathcal{P},i}(t)\widetilde{\mathbf{x}}(i,t)+\alpha_t\mathcal{S}_i\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\\
&\times\left(\mathbf{y}(t)-\mathbf{G}_H^\top\mathcal{P}\widetilde{\mathbf{x}}(t)\right)\\
&\Rightarrow\frac{\mathbf{1}_{Q_i}^\top}{Q_i}\widetilde{\mathbf{x}}(i,t+1)=\frac{\mathbf{1}_{Q_i}^\top}{Q_i}\widetilde{\mathbf{x}}(i,t)-\frac{\mathbf{1}_{Q_i}^\top}{Q_i}\mathbf{L}_{\mathcal{P},i}(t)\widetilde{\mathbf{x}}(i,t)\\
&+\alpha_t\frac{\mathbf{1}_{Q_i}^\top}{Q_i}\mathcal{S}_i\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\left(\mathbf{y}(t)-\mathbf{G}_H^\top\mathcal{P}\widetilde{\mathbf{x}}(t)\right)\\
&\Rightarrow\widetilde{\mathbf{x}}_{\text{avg},i}(t+1)=\widetilde{\mathbf{x}}_{\text{avg},i}(t)+\alpha_t\frac{\mathbf{1}_{Q_i}^\top}{Q_i}\mathcal{S}_i\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\\
&\times\left(\mathbf{y}(t)-\mathbf{G}_H^\top\mathcal{P}\widetilde{\mathbf{x}}(t)\right),
\end{aligned} \tag{F.43}$$

where $\{\widetilde{\mathbf{x}}_{\text{avg},i}(t)\}$ is the averaged estimate sequence for the i-th entry of the parameter $\boldsymbol{\theta}^*$. Stacking, all such averages together we have,

$$\begin{aligned}
&\widetilde{\mathbf{x}}_{\text{avg}}(t+1)=\widetilde{\mathbf{x}}_{\text{avg}}(t)+\alpha_t\mathcal{S}_{\text{avg}}\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\left(\mathbf{y}(t)-\mathbf{G}_H^\top\mathcal{P}\widetilde{\mathbf{x}}(t)\right)\\
&\Rightarrow\widetilde{\mathbf{x}}_{\text{avg}}(t+1)-\boldsymbol{\theta}^*=\left(\mathbf{I}-\alpha_t\mathbf{Q}\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}^{-1}\mathbf{H}_n\mathcal{P}_{\mathcal{I}_n}\right)\\
&\times\left(\widetilde{\mathbf{x}}_{\text{avg}}(t)-\boldsymbol{\theta}^*\right)\\
&+\alpha_t\mathcal{S}_{\text{avg}}\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\gamma(t)\\
&+\alpha_t\mathbf{Q}\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}_n^{-1}\mathbf{H}_n\left(\widetilde{\mathbf{x}}_n(t)-\mathcal{P}_{\mathcal{I}_n}\widetilde{\mathbf{x}}_{\text{avg}}(t)\right),
\end{aligned} \tag{F.44}$$

where $\mathcal{S}_{\text{avg}}=\left[\frac{\mathbf{1}_{Q_1}^\top}{Q_1}\mathcal{S}_1,\frac{\mathbf{1}_{Q_2}^\top}{Q_2}\mathcal{S}_2,\cdots,\frac{\mathbf{1}_{Q_N}^\top}{Q_N}\mathcal{S}_N\right]$ and $\mathbf{Q}=diag\left[\frac{1}{Q_1},\frac{1}{Q_2},\cdots,\frac{1}{Q_N}\right]$. In the above derivation, we make use of the fact that $\mathcal{S}_{\text{avg}}\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\mathbf{G}_H^\top\mathcal{P}\mathbf{1}_N\otimes(\widetilde{\mathbf{x}}_{\text{avg}}(t)-\boldsymbol{\theta}^*)=\mathbf{Q}\sum_{n=1}^{N}\mathcal{P}_n\mathbf{H}_n^\top\mathbf{R}^{-1}\mathbf{H}_n\left(\widetilde{\mathbf{x}}_{\text{avg}}(t)-\boldsymbol{\theta}^*\right)$, which in turn follows from the fact that,

$$\begin{aligned}
&\mathcal{S}_{\text{avg}}=\mathbf{Q}\left[\mathcal{P}_{\mathcal{I}_1}\ \mathcal{P}_{\mathcal{I}_2}\cdots\mathcal{P}_{\mathcal{I}_N}\right]=\left[\mathbf{Q}\mathcal{P}_{\mathcal{I}_1}\ \mathbf{Q}\mathcal{P}_{\mathcal{I}_2}\cdots\mathbf{Q}\mathcal{P}_{\mathcal{I}_N}\right]\\
&\Rightarrow\mathcal{S}_{\text{avg}}\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\mathbf{G}_H^\top\mathcal{P}\mathbf{1}_N\otimes(\widetilde{\mathbf{x}}_{\text{avg}}(t)-\boldsymbol{\theta}^*)\\
&=\left[\mathbf{Q}\mathcal{P}_{\mathcal{I}_1}\ \mathbf{Q}\mathcal{P}_{\mathcal{I}_2}\cdots\mathbf{Q}\mathcal{P}_{\mathcal{I}_N}\right]\\
&\times\left[\mathcal{P}_{\mathcal{I}_1}\mathbf{H}_1^\top\mathbf{R}_1^{-1}\mathbf{H}_1\left(\widetilde{\mathbf{x}}_{\text{avg}}(t)-\boldsymbol{\theta}^*\right)\cdots\right.
\end{aligned}$$

$$\mathcal{P}_{\mathcal{I}_N}\mathbf{H}_N^\top\mathbf{R}_N^{-1}\mathbf{H}_N\left(\widetilde{\mathbf{x}}_{\text{avg}}(t)-\boldsymbol{\theta}^*\right)\Big]^\top$$
$$=\mathbf{Q}\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}^{-1}\mathbf{H}_n\left(\widetilde{\mathbf{x}}_{\text{avg}}(t)-\boldsymbol{\theta}^*\right). \tag{F.45}$$

Define, the residual sequence, $\{\mathbf{z}_t\}$, where $\mathbf{z}(t)=\widetilde{\mathbf{x}}_{\text{avg}}(t)-\boldsymbol{\theta}^*$, which can be then shown to satisfy the recursion

$$\mathbf{z}_{t+1}=\left(\mathbf{I}_N-\alpha_t\Gamma\right)\mathbf{z}_t+\alpha_t\mathbf{U}_t+\alpha_t\mathbf{J}_t, \tag{F.46}$$

where

$$\Gamma=\mathbf{Q}\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}^{-1}\mathbf{H}_n\mathcal{P}_{\mathcal{I}_n}$$
$$\mathbf{U}_t=\mathbf{Q}\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}_n^{-1}\mathbf{H}_n\left(\widetilde{\mathbf{x}}_n(t)-\mathcal{P}_{\mathcal{I}_n}\widetilde{\mathbf{x}}_{\text{avg}}(t)\right)$$
$$\mathbf{J}_t=\mathcal{S}_{\text{avg}}\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\gamma(t). \tag{F.47}$$

We rewrite the recursion for $\{\mathbf{z}_t\}$ as follows:

$$\mathbf{z}_{t+1}=\left(\mathbf{I}_N-\alpha_t\Gamma_t\right)\mathbf{z}_t+(t+1)^{-3/2}\mathbf{T}_t+(t+1)^{-1}\boldsymbol{\Phi}_t\mathbf{V}_t, \tag{F.48}$$

where

$$\boldsymbol{\Gamma}_t=\boldsymbol{\Gamma}=\mathbf{Q}\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}^{-1}\mathbf{H}_n\mathcal{P}_{\mathcal{I}_n},\boldsymbol{\Phi}_t=a\mathbf{I}$$
$$\mathbf{T}_t=a(t+1)^{1/2}\mathbf{U}_t$$
$$=a\mathbf{Q}\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}_n^{-1}\mathbf{H}_n(t+1)^{0.5}\left(\widetilde{\mathbf{x}}_n(t)-\mathcal{P}_{\mathcal{I}_n}\widetilde{\mathbf{x}}_{\text{avg}}(t)\right)\xrightarrow{t\to\infty}0$$
$$\mathbf{V}_t=\mathbf{J}_t=\mathcal{S}_{\text{avg}}\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\gamma(t),\ \mathbb{E}\left[\mathbf{V}_t|\mathcal{F}_t\right]=0,$$
$$\mathbb{E}\left[\mathbf{V}_t\mathbf{V}_t^\top|\mathcal{F}_t\right]=\mathcal{S}_{\text{avg}}\mathcal{P}\mathbf{G}_H\mathbf{R}^{-1}\mathbf{G}_H^\top\mathcal{P}\mathcal{S}_{\text{avg}}$$
$$=\mathbf{Q}\left(\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}^{-1}\mathbf{H}_n\mathcal{P}_{\mathcal{I}_n}\right)\mathbf{Q} \tag{F.49}$$

Due to the i.i.d nature of the noise process, we have the uniform integrability condition for the process $\{\mathbf{V}_t\}$. Hence, $\{\mathbf{x}_{\text{avg}}(t)\}$ falls under the purview of Lemma C.1.5 and we thus conclude that

$$(t+1)^{1/2}\left(\widetilde{\mathbf{x}}_{\text{avg}}(t)-\boldsymbol{\theta}\right)\overset{\mathcal{D}}{\Longrightarrow}\mathcal{N}(0,\mathbf{PMP}^\top), \tag{F.50}$$

in which,

$$[\mathbf{M}]_{ij}=\left[\mathbf{PQ}\left(\sum_{n=1}^{N}\mathcal{P}_{\mathcal{I}_n}\mathbf{H}_n^\top\mathbf{R}_n^{-1}\mathbf{H}_n\mathcal{P}_{\mathcal{I}_n}\right)\mathbf{QP}\right]_{ij}$$

$$\times \left([\mathbf{\Lambda}]_{ii} + [\mathbf{\Lambda}]_{jj} - 1 \right)^{-1}, \tag{F.51}$$

where $\mathbf{P}$ and $\mathbf{\Lambda}$ are orthonormal and diagonal matrices such that $\mathbf{P}^\top \mathbf{Q} \left(\sum_{n=1}^{N} \mathcal{P}_{\mathcal{I}_n} \mathbf{H}_n^\top \mathbf{R}_n^{-1} \mathbf{H}_n \mathcal{P}_{\mathcal{I}_n} \right) \mathbf{Q}\mathbf{P} = \mathbf{\Lambda}$. Now from (F.42), we have that the processes $\{\widetilde{\mathbf{x}}_n(t)\}$ and $\{\widetilde{\mathbf{x}}_{\text{avg}}(t)\}$ are indistinguishable in the $(t+1)^{1/2}$ time scale, which is formalized as follows:

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\theta}} \left(\lim_{t \to \infty} \left\| \sqrt{t+1} \, (\widetilde{\mathbf{x}}(t) - \boldsymbol{\theta}) - \sqrt{t+1} \, (\widetilde{\mathbf{x}}_{\text{avg}}(t) - \boldsymbol{\theta}) \right\| = 0 \right) \\ &= \mathbb{P}_{\boldsymbol{\theta}} \left(\lim_{t \to \infty} \left\| \sqrt{t+1} \, (\widetilde{\mathbf{x}}(t) - \widetilde{\mathbf{x}}_{\text{avg}}(t)) \right\| = 0 \right) = 1. \end{aligned} \tag{F.52}$$

Thus, the difference of the sequences $\left\{ \sqrt{t+1} \, (\widetilde{\mathbf{x}}_n(t) - \boldsymbol{\theta}) \right\}$ and $\left\{ \sqrt{t+1} \, (\widetilde{\mathbf{x}}_{\text{avg}}(t) - \boldsymbol{\theta}) \right\}$ converges a.s. to zero as $t \to \infty$ and hence we have,

$$\sqrt{t+1} \, (\widetilde{\mathbf{x}}_n(t) - \boldsymbol{\theta}) \overset{\mathcal{D}}{\Longrightarrow} \mathcal{N}(0, \mathbf{P}\mathbf{M}\mathbf{P}^\top). \tag{F.53}$$

□

Appendix G

Proofs of Theorems in Chapter 8

G.1 Proof of the main result: First order optimization

Lemma G.1.1. *Consider algorithm (8.2), and let the hypotheses of Theorem 8.6.1 hold. Then, we have that for all $k = 0, 1, ...,$ there holds:*

$$\mathbb{E}\left[\|\mathbf{x}(k) - \mathbf{x}^o\|^2\right] \leq q_{k_0}(N, \alpha_0) + \frac{\pi^2}{6}\alpha_0^2\left(2c_u N\|\mathbf{x}^o\|^2 + N\sigma_u^2\right) + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} \doteq q_\infty(N, \alpha_0),$$

where $\mathbb{E}\left[\|\mathbf{x}(k_2) - \mathbf{x}^o\|^2\right] \leq q_{k_2}(N, \alpha_0)$, $k_2 = \max\{k_0, k_1\}$, $k_0 = \inf\{k|\mu^2\alpha_k^2 < 1\}$ *and* $k_1 = \inf\left\{k|\frac{\mu}{2} > 2c_u\alpha_k\right\}$.

Proof. Proceeding as in the proof of Lemma H.1.1, with $c_k = 1$ and $\mathbf{b}(\mathbf{x}(k)) = 0$, we have that, $\forall k \geq \max\{k_0, k_1\}$,

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{\zeta}(k+1)\|^2\right] \leq \prod_{l=k_0}^{k}\left(1 - \frac{\mu\alpha_l}{2}\right)\mathbb{E}\left[\|\boldsymbol{\zeta}(k_0)\|^2\right] \\
&+ \frac{\pi^2}{6}\alpha_0^2\left(2c_u N\|\mathbf{x}^o\|^2 + N\sigma_u^2\right) \\
&+ 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} \\
&\mathbb{E}\left[\|\boldsymbol{\zeta}(k+1)\|^2\right] \leq q_{k_2}(N, \alpha_0) + \frac{\pi^2}{6}\alpha_0^2\left(2c_u N\|\mathbf{x}^o\|^2 + N\sigma_u^2\right) \\
&+ 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} \\
&\doteq q_\infty(N, \alpha_0), \qquad \text{(G.1)}
\end{aligned}$$

where $k_0 = \inf\{k|\mu^2\alpha_k^2 < 1\}$ and

$$k_1 = \inf\left\{k|\frac{\mu}{2} > 2c_u\alpha_k\right\}.$$

and $k_2 = \max\{k_0, k_1\}$. It is to be noted that k_1 is necessarily finite as $\alpha_k \to 0$ as $k \to \infty$. Hence, we have that $\mathbb{E}\left[\|\mathbf{x}(k+1) - \mathbf{x}^o\|^2\right]$ is finite and bounded from above, where $\mathbb{E}\left[\|\mathbf{x}(k_2) - \mathbf{x}^o\|^2\right] \leq q_{k_2}(N, \alpha_0)$. From the boundedness of $\mathbb{E}\left[\|\mathbf{x}(k) - \mathbf{x}^o\|^2\right]$, we have also established the boundedness of $\mathbb{E}\left[\|\nabla F(\mathbf{x}(k))\|^2\right]$ and $\mathbb{E}\left[\|\mathbf{x}(k)\|^2\right]$.

With the above development in place, we can bound the variance of the noise process $\{\mathbf{v}(k)\}$ as follows:

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{u}(k)\|^2|\mathcal{F}_k\right] \leq 2c_u q_\infty(N,\alpha_0)\\&+2N\underbrace{\left(\sigma_u^2+\|\mathbf{x}^*\|^2\right)}_{\sigma_1^2}.\end{aligned} \tag{G.2}$$

The proof of Lemma G.1.1 is now complete. □

Recall the (hypothetically available) global average of nodes' estimates $\overline{\mathbf{x}}(k)=\frac{1}{N}\sum_{i=1}^N\mathbf{x}_i(k)$, and denote by $\widetilde{\mathbf{x}}_i(k)=\mathbf{x}_i(k)-\overline{\mathbf{x}}(k)$ the quantity that measures how far apart is node i's solution estimate from the global average. Introduce also vector $\widetilde{\mathbf{x}}(k)=(\,\widetilde{\mathbf{x}}_1(k),...,\widetilde{\mathbf{x}}_N(k)\,)^\top$, and note that it can be represented as $\widetilde{\mathbf{x}}(k)=(\mathbf{I}-\mathbf{J})\,\mathbf{x}(k)$, where we recall $\mathbf{J}=\frac{1}{N}\mathbf{1}\mathbf{1}^\top$. We have the following Lemma.

Lemma G.1.2. *Let the hypotheses of Theorem 8.6.1 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] &\leq Q_k+\frac{2\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)}\\&=O\left(\frac{1}{k}\right),\end{aligned}$$

where Q_k is a term which decays faster than $(k+1)^{-1}$.

Lemma G.1.2 is important as it allows to sufficiently tightly bound the bias in the gradient estimates according to which the global average $\overline{\mathbf{x}}(k)$ evolves.

Proof. Proceeding as in the proof of Lemma H.1.3 in (H.19)-(H.22), we have,

$$\begin{aligned}&\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2|\mathcal{F}_k\right] \leq (1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2\\&+\left(1+\frac{1}{\theta_k}\right)\alpha_k^2\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{F}_k\right],\end{aligned} \tag{G.3}$$

where

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{F}_k\right] \leq 2\|\nabla F(\mathbf{x}(k))\|^2+2\mathbb{E}\left[\|\mathbf{v}(k)\|^2|\mathcal{F}_k\right]\\&\leq\underbrace{2\|\nabla F(\mathbf{x}(k))\|^2+4c_u q_\infty(N,\alpha_0)+4N\sigma_1^2}_{\Delta_{1,\infty}}\\&\Rightarrow\mathbb{E}\left[\|\mathbf{w}(k)\|^2\right]<\infty.\end{aligned} \tag{G.4}$$

With the above development in place, we then have,

$$\begin{aligned}&\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] \leq (1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2\\&+\left(1+\frac{1}{\theta_k}\right)\alpha_k^2\Delta_{1,\infty}.\end{aligned} \tag{G.5}$$

In particular, we choose $\theta(k) = \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)$. From (H.25), we have,

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] &\leq \left(1 - \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right] \\ &+ \left(1 + \frac{2}{\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)}\right)\alpha_k^2\Delta_{1,\infty} \\ &= \left(1 - \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right] \\ &+ \frac{2\alpha_k^2}{\lambda_2\left(\overline{\mathbf{L}}\right)\beta_k}\Delta_{1,\infty} + \alpha_k^2\Delta_{1,\infty}. \end{aligned} \tag{G.6}$$

Applying lemma H.1.4 to $q_k = \mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right]$, $d_k = \Delta_k$, $b_k = \alpha_k^2$, and $s_k = \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)$, we obtain for $m(k) = \lfloor\frac{k-1}{2}\rfloor$:

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] &\leq \underbrace{\exp\left(-\sum_{l=0}^{k} s(l)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(0)\|^2\right]}_{t_1} \\ &\underbrace{+ \Delta_{1,\infty}\exp\left(-\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k} s(m)\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}\left(\frac{2\alpha_l^2}{\lambda_2\left(\overline{\mathbf{L}}\right)\beta_l} + \alpha_l^2\right)}_{t_2} \\ &+ \underbrace{\frac{2\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)}}_{t_3} + \underbrace{\frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2\left(\overline{\mathbf{L}}\right)\beta_0(k+1)^{3/2}}}_{t_4}. \end{aligned} \tag{G.7}$$

In the proof of Lemma H.1.4, the splitting in the interval $[0, k]$ was done at $\lfloor\frac{k-1}{2}\rfloor$ for ease of book keeping. The division can be done at an arbitrary point. It is to be noted that the sequence $\{s(k)\}$ is not summable and hence terms t_1 and t_2 decay faster than $(k+1)$. Also, note that term t_4 decays faster than t_3. For notational ease, henceforth we refer to $t_1 + t_2 + t_4 = Q_k$, while keeping in mind that Q_k decays faster than $(k+1)$. Hence, we have the disagreement given by,

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] = O\left(\frac{1}{k}\right).$$

□

Lemma G.1.3. *Consider algorithm (8.2) and let the hypotheses of Theorem 8.6.1 hold. Then, there holds:*

$$\mathbb{E}[\,\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\,] = O(1/k).$$

and

$$\mathbb{E}[\,\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\,] = O\left(\frac{1}{\mathcal{C}_k^{4/3-\zeta}}\right),$$

where $\zeta > 0$ can be arbitrarily small, for all $i = 1, \cdots, N$.

Proof. Denote $\overline{\mathbf{x}}(k) = \frac{1}{N}\sum_{n=1}\mathbf{x}_i(k)$. Then, we have,

$$\overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k)-\alpha_k\left[\frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right)+\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i(k)}_{\overline{\mathbf{u}}(k)}\right] \tag{G.8}$$

which implies:

$$\begin{aligned}\overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k)-\frac{\alpha_k}{N}\Bigg[\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right)\\ -\nabla f_i\left(\overline{\mathbf{x}}(k)\right)+\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\Bigg]-\alpha_k\overline{\mathbf{u}}(k).\end{aligned}$$

where

$$\begin{aligned}\mathbf{e}(k)&=N\overline{\mathbf{u}}(k)\\&+\underbrace{\sum_{i=1}^{N}\left(\nabla f_i\left(\mathbf{x}_i(k)\right)-\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right)}_{\boldsymbol{\epsilon}(k)}.\end{aligned} \tag{G.9}$$

Proceeding as in (H.34)-(H.40), with $c_k=1$ and $\mathbf{b}(\mathbf{x}_i(k))=0$, $\forall i=1,\cdots,N$, we have on choosing $\theta_k=\frac{\mu\alpha_0}{k+1}$, where $\alpha_0>\frac{1}{\mu}$,

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{m}(k)\|^2\right]\leq\left(1-\frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]\\&+\frac{8NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)^2}+\frac{2NL^2Q_k}{\mu(k+1)}\\&\Rightarrow\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right]\leq\left(1-\frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]\\&+\frac{8NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)^2}+\frac{2NL^2Q_k}{\mu(k+1)}+2\alpha_k^2\left(c_uq_\infty(N,\alpha_0)+N\sigma_1^2\right)\\&\Rightarrow\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right]\leq\left(1-\frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]\\&+\frac{8NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)^2}+2\alpha_k^2\left(c_uq_\infty(N,\alpha_0)+N\sigma_1^2\right)+P_k,\end{aligned} \tag{G.10}$$

where P_k decays faster as compared to the other terms. Proceeding as in (H.30), we have

$$\begin{aligned}&\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right]\\&\leq\underbrace{\exp\left(-\mu\sum_{l=0}^{k}\alpha_l\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]}_{t_6}\\&\underbrace{+\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{8L^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(l+1)^2}}_{t_7}\\&\underbrace{+\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}P_l}_{t_{10}}\end{aligned}$$

$$+\underbrace{\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{2\alpha_0^2\left(c_u q_\infty(N,\alpha_0)+N\sigma_1^2\right)}{(l+1)^2}}_{t_{11}}$$
$$+\underbrace{\frac{16NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)}}_{t_{12}}$$
$$+\underbrace{\frac{N(k+1)P_k}{\mu\alpha_0}}_{t_{14}}+\underbrace{\frac{4N\alpha_0\left(c_u q_\infty(N,\alpha_0)+N\sigma_1^2\right)}{\mu(k+1)}}_{t_{15}}. \tag{G.11}$$

It is to be noted that the term t_6 decays as $1/k$. The terms t_7, t_{10} and t_{11} decay faster than its counterparts in the terms t_{12} and t_{15} respectively. We note that Q_l also decays faster. Hence, the rate of decay of $\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\right\|^2\right]$ is determined by the terms t_{12} and t_{15}. Thus, we have that, $\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\right\|^2\right]=O\left(\frac{1}{k}\right)$. For notational ease, we refer to $t_6+t_7+t_{10}+t_{11}+t_{14}=M_k$ from now on. Finally, we note that,

$$\begin{aligned}
&\|\mathbf{x}_i(k)-\mathbf{x}^*\|\leq\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|+\left\|\underbrace{\mathbf{x}_i(k)-\overline{\mathbf{x}}(k)}_{\widetilde{\mathbf{x}}_i(k)}\right\|\\
&\Rightarrow\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\leq 2\|\widetilde{\mathbf{x}}_i(k)\|^2+2\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\\
&\Rightarrow\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]\leq 2M_k+\frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)}\\
&+2Q_k+\frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)}\\
&\Rightarrow\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]=O\left(\frac{1}{k}\right),\quad\forall i.
\end{aligned}\tag{G.12}$$

The communication cost is given by,

$$\mathbb{E}\left[\sum_{t=1}^{k}\zeta_t\right]=O\left(k^{\frac{3}{4}+\frac{\epsilon}{2}}\right).$$

Thus, we achieve the communication rate to be,

$$\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^\star\|^2\right]=O\left(\frac{1}{\mathcal{C}_k^{\frac{4}{3}-\zeta}}\right). \tag{G.13}$$

□

Appendix H

Proofs of Theorems in Chapter 9

H.1 Proof of the main result: Zeroth order optimization

The proof of the main result proceeds through three main steps. The first step involves establishing the boundedness of the iterate sequence, while the second step involves establishing the convergence rate of the optimizer sequence at each agent to the network averaged optimizer sequence. The convergence of the network averaged optimizer is then analyzed as the final step and in combination with the second step results in the establishment of bounds on MSE of the optimizer sequence at each agent.

Lemma H.1.1. *Let the hypotheses of Theorem 9.7.1 hold. Then, we have,*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}(k)-\mathbf{x}^o\|^2\right] \leq q_{k_2}(N,d,\alpha_0,c_0)+4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}\\
&+\frac{\sqrt{N}s_1(P)M\alpha_0 c_0^2}{8\delta}+\frac{Ns_1^2(P)M^2\alpha_0^2c_0^4}{16(1+4\delta)}\\
&+\frac{d\alpha_0^2\left(2c_vN\|\mathbf{x}^o\|^2+N\sigma_v^2\right)}{c_0^2(1-2\delta)}+\frac{\alpha_0^2c_0^2\sqrt{N}s_1(P)L\|\nabla F(\mathbf{x}^o)\|}{1+2\delta}\\
&+\frac{N\alpha_0^2c_0^4s_2(P)}{1+4\delta}+\frac{4\alpha_0^2c_0^2Ns_1(P)}{1+2\delta}\|\nabla F(\mathbf{x}^o)\|^2\\
&\doteq q_\infty(N,d,\alpha_0,c_0),
\end{aligned}$$

where $\mathbb{E}\left[\|\mathbf{x}(k_2)-\mathbf{x}^o\|^2\right] \leq q_{k_2}(N,d,\alpha_0,c_0)$, $k_2=\max\{k_0,k_1\}$, $k_0=\inf\{k|\mu^2\alpha_k^2<1\}$ *and* $k_1=\inf\left\{k|\frac{\mu}{2}>\frac{\sqrt{N}}{4}s_1(P)Mc_k^2+\frac{2dc_v\alpha_k}{c_k^2}+4\right.$

Proof.

$$\begin{aligned}
&\mathbf{x}(k+1)=\mathbf{W}_k\mathbf{x}(k)\\
&-\frac{\alpha_k}{c_k}\left(c_k\nabla F(\mathbf{x}(k))+c_k^2\mathbf{b}(\mathbf{x}(k))+c_k\mathbf{h}(\mathbf{x}(k))\right).
\end{aligned}\tag{H.1}$$

Denote $\mathbf{x}^o=\mathbf{1}_N\otimes x^*$. Then, we have,

$$\begin{aligned}
&\mathbf{x}(k+1)-\mathbf{x}^o=\mathbf{W}_k(\mathbf{x}(k)-\mathbf{x}^o)\\
&-\alpha_k\left(\nabla F(\mathbf{x}(k))-\nabla F(\mathbf{x}^o)\right)\\
&-\alpha_k\mathbf{h}(\mathbf{x}(k))-\alpha_k\nabla F(\mathbf{x}^o)-\alpha_kc_k\mathbf{b}(\mathbf{x}(k)).
\end{aligned}\tag{H.2}$$

Moreover, note that, $\mathbb{E}\left[\mathbf{h}(\mathbf{x}(k)) \mid \mathcal{F}_k\right] = 0$. By Leibnitz rule, we have,

$$\begin{aligned}
&\nabla F(\mathbf{x}(k)) - \nabla F(\mathbf{x}^o) \\
&= \left[\int_{s=0}^{1} \nabla^2 F\left(\mathbf{x}^o + s(\mathbf{x}(k) - \mathbf{x}^o)\right) ds\right] (\mathbf{x}(k) - \mathbf{x}^o) \\
&= \mathbf{H}_k \left(\mathbf{x}(k) - \mathbf{x}^o\right). \qquad \text{(H.3)}
\end{aligned}$$

By Lipschitz continuity of the gradients and strong convexity of $f(\cdot)$, we have that $L\mathbf{I} \succcurlyeq \mathbf{H}_k \succcurlyeq \mu\mathbf{I}$. Denote by $\boldsymbol{\zeta}(k) = \mathbf{x}(k) - \mathbf{x}^o$ and by $\boldsymbol{\xi}(k) = \left(\mathbf{W}_k - \alpha_k \mathbf{H}_k\right)\left(\mathbf{x}(k) - \mathbf{x}^o\right) - \alpha_k \nabla F(\mathbf{x}^o)$. Then, there holds:

$$\begin{aligned}
&\mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2 \,|\, \mathcal{F}_k\,] \leq \mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2 | \mathcal{F}_k\right] \\
&- 2\alpha_k c_k\, \mathbb{E}\left[\boldsymbol{\xi}(k)^\top | \,\mathcal{F}_k\right] \mathbb{E}[\,\mathbf{h}(\mathbf{x}(k)) \,|\, \mathcal{F}_k\,] + \alpha_k^2 c_k^2\, \mathbb{E}[\,\|\mathbf{h}(\mathbf{x}(k))\|^2 \,|\, \mathcal{F}_k\,] \\
&+ \alpha_k^2 c_k^2 \mathbf{b}^\top(\mathbf{x}(k))\mathbf{b}(\mathbf{x}(k)) - 2\alpha_k c_k \mathbf{b}^\top(\mathbf{x}(k)) \mathbb{E}\left[\boldsymbol{\xi}(k) | \mathcal{F}_k\right] \\
&+ \mathbf{b}\left(\mathbf{x}(k)\right)^\top \mathbb{E}\left[\mathbf{h}(\mathbf{x}(k)) | \mathcal{F}_k\right]. \qquad \text{(H.4)}
\end{aligned}$$

We use the following inequalities:

$$\begin{aligned}
&c_k \mathbf{b}(\mathbf{x}_i(k)) \\
&= \frac{c_k}{2} \mathbb{E}\left[\langle \mathbf{z}_{i,k}, \nabla^2 f_i \left(\mathbf{x}_i(k) + \frac{(1-\theta_1)}{2} c_k \mathbf{z}_{i,k}\right) \mathbf{z}_{i,k} \rangle \mathbf{z}_{i,k} | \mathcal{F}_k\right] \\
&- \frac{c_k}{2} \mathbb{E}\left[\langle \mathbf{z}_{i,k}, \nabla^2 f_i \left(\mathbf{x}_i(k) + (1-\theta_2)\, c_k \mathbf{z}_{i,k}\right) \mathbf{z}_{i,k} \rangle \mathbf{z}_{i,k} | \mathcal{F}_k\right] \\
&\Rightarrow c_k \left\|\mathbf{b}(\mathbf{x}_i(k))\right\| \leq \frac{c_k^2}{4} M s_1(P). \qquad \text{(H.5)}
\end{aligned}$$

$$\begin{aligned}
&- \mathbf{b}^\top(\mathbf{x}(k)) \mathbb{E}\left[\boldsymbol{\xi}(k) | \mathcal{F}_k\right] \\
&= -2\mathbf{b}^\top(\mathbf{x}(k)) \left(\mathbf{I} - \beta_k \overline{\mathbf{L}} - \alpha_k \mathbf{H}_k\right) \left(\mathbf{x}(k) - \mathbf{x}^o\right) \\
&+ 2\alpha_k \mathbf{b}^\top(\mathbf{x}(k)) \nabla F(\mathbf{x}^o) \\
&\leq 2 \left\|\mathbf{b}(\mathbf{x}(k))\right\| \left\|\mathbf{I} - \beta_k \overline{\mathbf{L}} - \alpha_k \mathbf{H}_k\right\| \left\|\mathbf{x}(k) - \mathbf{x}^o\right\| \\
&+ 2\alpha_k \left\|\mathbf{b}(\mathbf{x}(k))\right\| \left\|\nabla F(\mathbf{x}^o)\right\| \\
&\leq \frac{\sqrt{N}}{4} s_1(P) M c_k \left(1 - \mu\alpha_k\right) \left(1 + \left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2\right) \\
&+ \alpha_k c_k \frac{\sqrt{N}}{2} s_1(P) M \left\|\nabla F(\mathbf{x}^o)\right\| \\
&\leq \frac{\sqrt{N}}{4} s_1(P) M c_k + \frac{\sqrt{N}}{4} s_1(P) M c_k \left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2 \\
&+ \alpha_k c_k \frac{\sqrt{N}}{2} s_1(P) M \left\|\nabla F(\mathbf{x}^o)\right\|, \qquad \text{(H.6)}
\end{aligned}$$

$$\mathbf{b}^\top(\mathbf{x}(k))\mathbf{b}(\mathbf{x}(k)) \leq \frac{N}{16} s_1^2(P) M^2 c_k^2, \qquad \text{(H.7)}$$

$$\begin{aligned}
&\mathbb{E}[\,\|\mathbf{h}(\mathbf{x}(k))\|^2 \,|\, \mathcal{F}_k\,] = \mathbb{E}\left[\|\mathbf{v}_{\mathbf{z}}(k; \mathbf{x}(k))\|^2 | \mathcal{F}_k\right] \\
&+ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}(k)) - \mathbb{E}\left[\widehat{\mathbf{g}}(\mathbf{x}(k)) \mid \mathcal{F}_k\right]\|^2 \mid \mathcal{F}_k\right], \qquad \text{(H.8)}
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}(k)) - \mathbb{E}\left[\widehat{\mathbf{g}}(\mathbf{x}(k)) \mid \mathcal{F}_k\right]\|^2 \mid \mathcal{F}_k\right] \\
&\leq \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}(k))\|^2 \mid \mathcal{F}_k\right] \\
&\leq 4Ns_1(P)L^2\left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2 + 4Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2 + 2Nc_k^2 s_2(P),
\end{aligned} \tag{H.9}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{v}_{\mathbf{z}}(k;\mathbf{x}(k))\|^2 |\mathcal{F}_k\right] \leq dc_v\left\|\mathbf{x}(k)\right\|^2 + dN\sigma_v^2 \\
&\leq 2dc_v\left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2 + \left(2dc_v\left\|\mathbf{x}^o\right\|^2 + N\sigma_v^2\right).
\end{aligned} \tag{H.10}$$

Then from (H.4), we have,

$$\begin{aligned}
&\mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2 \,|\, \mathcal{F}_k\,] \leq \mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2 | \mathcal{F}_k\right] \\
&+ \frac{\sqrt{N}}{4}s_1(P)M\alpha_k c_k^2\|\boldsymbol{\zeta}(k)\|^2 + 2\frac{d\alpha_k^2}{c_k^2}c_v\|\boldsymbol{\zeta}(k)\|^2 \\
&+ \frac{d\alpha_k^2}{c_k^2}\left(2c_v\left\|\mathbf{x}^o\right\|^2 + N\sigma_v^2\right) + \frac{\sqrt{N}}{4}s_1(P)M\alpha_k c_k^2 \\
&+ \frac{N}{16}s_1^2(P)M^2\alpha_k^2 c_k^4 + \alpha_k^2 c_k^2\frac{\sqrt{N}}{2}s_1(P)M\left\|\nabla F(\mathbf{x}^o)\right\| \\
&+ 4\alpha_k^2 c_k^2 Ns_1(P)L^2\|\boldsymbol{\zeta}(k)\|^2 + 4\alpha_k^2 c_k^2 Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2 \\
&+ 2N\alpha_k^2 c_k^4 s_2(P).
\end{aligned} \tag{H.11}$$

We next bound $\mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2 | \,\mathcal{F}_k\,\right]$. Note that $\|\mathbf{W}_k - \alpha_k\,\boldsymbol{H}_k\| \leq 1 - \mu\,\alpha_k$. Therefore, we have:

$$\|\boldsymbol{\xi}(k)\| \leq (1 - \mu\,\alpha_k)\,\|\boldsymbol{\zeta}(k)\| + \alpha_k\,\|\nabla F(\mathbf{x}^o)\|. \tag{H.12}$$

We now use the following inequality:

$$(a+b)^2 \leq (1+\theta)\,a^2 + \left(1 + \frac{1}{\theta}\right)b^2, \tag{H.13}$$

for any $a, b \in \mathbb{R}$ and $\theta > 0$. We set $\theta = \mu\alpha_k$. Using the inequality (H.13) in (H.12) and we have $\forall k \geq k_0$, where $k_0 = \inf\{k|\mu^2\alpha_k^2 < 1\}$:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2 |\mathcal{F}_k\right] \leq (1+\mu\alpha_k)\,(1-\alpha_k\mu)^2\,\|\boldsymbol{\zeta}(k)\|^2 \\
&+ \left(1 + \frac{1}{\mu\alpha_k}\right)\alpha_k^2\|\nabla F(\mathbf{x}^o)\|^2 \\
&\leq (1-\alpha_k\mu)\,\|\boldsymbol{\zeta}(k)\|^2 + 2\frac{\alpha_k}{\mu}\|\nabla F(\mathbf{x}^o)\|^2.
\end{aligned} \tag{H.14}$$

Using (H.14) in (H.11), we have for all $k \geq k_0$

$$\begin{aligned}
&\mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2 \,|\, \mathcal{F}_k\,] \\
&\leq \Big(1 - \alpha_k\mu + \frac{\sqrt{N}}{4}s_1(P)M\alpha_k c_k^2 + 2\frac{d\alpha_k^2}{c_k^2}c_v \\
&+4\alpha_k^2 c_k^2 Ns_1(P)L^2\Big) \times \|\boldsymbol{\zeta}(k)\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\frac{d\alpha_k^2}{c_k^2}\left(2c_v\left\|\mathbf{x}^o\right\|^2+N\sigma_v^2\right)+\frac{\sqrt{N}}{4}s_1(P)L\alpha_k c_k^2\\
&+\frac{N}{16}s_1^2(P)M^2\alpha_k^2c_k^4+2\frac{\alpha_k}{\mu}\|\nabla F(\mathbf{x}^o)\|^2+2N\alpha_k^2c_k^4s_2(P)\\
&+\alpha_k^2c_k^2\frac{\sqrt{N}}{2}s_1(P)M\left\|\nabla F(\mathbf{x}^o)\right\|+4\alpha_k^2c_k^2Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2. \qquad \text{(H.15)}
\end{aligned}
$$

Define k_1 as follows:

$$
k_1=\inf\left\{k|\frac{\mu}{2}>\frac{\sqrt{N}}{4}s_1(P)Mc_k^2+\frac{2dc_v\alpha_k}{c_k^2}+4\alpha_kc_k^2Ns_1(P)L^2\right\}.
$$

It is to be noted that k_1 is necessarily finite as $c_k\to 0$ and $\alpha_k c_k^{-2}\to 0$ as $k\to\infty$. We proceed by using the following auxiliary lemma.

Lemma H.1.2. *Let $a_k\in(0,1)$, $u\leq 0$ and $d_k\geq 0$, for all $k\geq 1$. If $q_{k_0}\geq 0$ and for all $k\geq k_0$ there holds $q_{k+1}\leq(1-a_k)q_k+a_ku+d_k$, then, for all $k\geq k_0$,*

$$
q_{k+1}\leq q_{k_0}+u+\sum_{l=l_0}^{k}d_l. \qquad \text{(H.16)}
$$

Proof. Introduce $p(k,l)=(1-a_k)\cdots(1-a_l)$, for $l\leq k$ and also $p(k,k+1)=1$. It is easy to see that, for every $k\geq k_0$, $q_{k+1}\leq p(k,k_0)q_{k_0}+u\sum_{l=k_0}^{k}p(k,l+1)a_l+\sum_{l=k_0}^{k}p(k,l+1)d_l$. Note now that $p(k,l+1)a_l=p(k,l+1)-p(k,l)$, and hence $\sum_{l=k_0}^{k}p(k,l+1)a_l=1-p(k,k_0)\leq 1$. Using the latter together with the fact that $p(k,l+1)\leq 1$ proves the claim of the lemma. □

Applying Lemma H.1.2 to $q_k=\mathbb{E}\left[\|\boldsymbol{\zeta}(k)\|^2\right]$, $a_k=\frac{\mu\alpha_k}{2}$, $u=4\frac{\|\nabla F(x^o)\|^2}{\mu^2}$, and d_k defined as the remaining term in (H.15) we have, $\forall k\geq\max\{k_0,k_1\}\doteq k_2$,

$$
\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{\zeta}(k+1)\|^2\right]\leq q_{k_2}(N,d,\alpha_0,c_0)+4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}\\
&+\frac{\sqrt{N}s_1(P)M\alpha_0c_0^2}{8\delta}+\frac{Ns_1^2(P)M^2\alpha_0^2c_0^4}{16(1+4\delta)}\\
&+\frac{d\alpha_0^2\left(2c_vN\left\|\mathbf{x}^o\right\|^2+N\sigma_v^2\right)}{c_0^2(1-2\delta)}+\frac{\alpha_0^2c_0^2\sqrt{N}s_1(P)L\left\|\nabla F(\mathbf{x}^o)\right\|}{1+2\delta}\\
&+\frac{2N\alpha_0^2c_0^4s_2(P)}{1+4\delta}+\frac{4\alpha_0^2c_0^2Ns_1(P)}{1+2\delta}\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2\\
&\doteq q_\infty(N,d,\alpha_0,c_0), \qquad \text{(H.17)}
\end{aligned}
$$

□

From (H.17), we have that $\mathbb{E}\left[\left\|\mathbf{x}(k+1)-\mathbf{x}^o\right\|^2\right]$ is finite and bounded from above, where $\mathbb{E}\left[\left\|\mathbf{x}(k_2)-\mathbf{x}^o\right\|^2\right]\leq q_{k_2}(N,d,\alpha_0,c_0)$. From the boundedness of $\mathbb{E}\left[\left\|\mathbf{x}(k)-\mathbf{x}^o\right\|^2\right]$, we have also established the boundedness of $\mathbb{E}\left[\left\|\nabla F(\mathbf{x}(k))\right\|^2\right]$ and $\mathbb{E}\left[\left\|\mathbf{x}(k)\right\|^2\right]$.

With the above development in place, we can bound the variance of the noise process $\{\mathbf{v}_\mathbf{z}(k;\mathbf{x}(k))\}$ as

follows:

$$\mathbb{E}\left[\|\mathbf{v}_{\mathbf{z}}(k;\mathbf{x}(k))\|^2 |\mathcal{F}_k\right] \leq 2dc_v q_\infty(N,d,\alpha_0,c_0) + 2Nd\underbrace{\left(\sigma_v^2 + \|\mathbf{x}^*\|^2\right)}_{\sigma_1^2}. \tag{H.18}$$

We also have the following bound:

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}(k)) - \mathbb{E}\left[\widehat{\mathbf{g}}(\mathbf{x}(k)) \mid \mathcal{F}_k\right]\|^2 \mid \mathcal{F}_k\right] \leq 4Ns_1(P)L^2 q_\infty(N,d,\alpha_0,c_0) + 4Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2 + 2Nc_k^2 s_2(P).$$

We now study the disagreement of the optimizer sequence $\{\mathbf{x}_i(k)\}$ at a node i with respect to the (hypothetically available) network averaged optimizer sequence, i.e., $\overline{\mathbf{x}}(k) = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i(k)$. Define the disagreement at the i-th node as $\widetilde{\mathbf{x}}_i(k) = \mathbf{x}_i(k) - \overline{\mathbf{x}}(k)$. The vectorized version of the disagreements $\widetilde{\mathbf{x}}_i(k)$, $i = 1, \cdots, N$, can then be written as $\widetilde{\mathbf{x}}(k) = (\mathbf{I} - \mathbf{J})\,\mathbf{x}(k)$, where $\mathbf{J} = \frac{1}{N}\left(\mathbf{1}_N \otimes \mathbf{I}_d\right)\left(\mathbf{1}_N \otimes \mathbf{I}_d\right)^\top = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top \otimes \mathbf{I}_d$. We have the following Lemma:

Lemma H.1.3. *Let the hypotheses of Theorem 9.7.1 hold. Then, we have*

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] \leq Q_k + \frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2 c_0^2 (k+1)^{2-2\tau-2\delta}} = O\left(\frac{1}{k^{2-2\delta-2\tau}}\right),$$

where Q_k is a term which decays faster than $(k+1)^{-2+2\tau+2\delta}$.

Lemma H.1.3 plays a crucial role in providing a tight bound for the bias in the gradient estimates according to which the global average $\overline{\mathbf{x}}(k)$ evolves.

Proof. The process $\{\widetilde{\mathbf{x}}(k)\}$ follows the recursion:

$$\widetilde{\mathbf{x}}(k+1) = \widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k) - \frac{\alpha_k}{c_k}(\mathbf{I} - \mathbf{J})\underbrace{\left(c_k\nabla F(\mathbf{x}(k)) + c_k\mathbf{h}(\mathbf{x}(k)) + c_k^2\mathbf{b}\left(\mathbf{x}(k)\right)\right)}_{\mathbf{w}(k)}, \tag{H.19}$$

where $\widetilde{\mathbf{W}}_k = \mathbf{W}_k - \mathbf{J}$. Then, we have,

$$\|\widetilde{\mathbf{x}}(k+1)\| \leq \left\|\widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\right\| + \frac{\alpha_k}{c_k}\|\mathbf{w}(k)\|. \tag{H.20}$$

Using (H.13) in (H.19), we have,

$$\|\widetilde{\mathbf{x}}(k+1)\|^2 \leq (1+\theta_k)\left\|\widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\right\|^2 + \left(1 + \frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\|\widetilde{\mathbf{w}}(k)\|^2. \tag{H.21}$$

We, now bound the term $\mathbb{E}\left[\left\|\widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\right\|^2|\mathcal{F}_k\right]$.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\mathbf{W}}(k)\widetilde{\mathbf{x}}(k)\right\|^2|\mathcal{F}_k\right]=\widetilde{\mathbf{x}}^\top(k)\mathbb{E}\left[\widetilde{\mathbf{W}}^2(k)-\mathbf{J}|\mathcal{F}_k\right]\widetilde{\mathbf{x}}(k)\\
&=\widetilde{\mathbf{x}}^\top(k)\left(\mathbf{I}-2\beta_k\overline{\mathbf{L}}+\beta_k^2\overline{\mathbf{L}}^2+\widetilde{\mathbf{L}}(k)^2-\mathbf{J}\right)\widetilde{\mathbf{x}}(k)\\
&\leq\left(1-2\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)+\beta_k^2\lambda_N^2\left(\overline{\mathbf{L}}\right)\right.\\
&\left.+\frac{4N^2\beta_0\rho_0^2}{(k+1)^{\tau+\epsilon}}-4\beta_k^2N^2\right)\|\widetilde{\mathbf{x}}(k)\|^2\\
&\leq\left(1-2\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)+\frac{4N^2\beta_0\rho_0^2}{(k+1)^{\tau+\epsilon}}\right)\|\widetilde{\mathbf{x}}(k)\|^2\\
&\leq\left(1-\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2,
\end{aligned}\tag{H.22}$$

where the last inequality follows from assumption 9.6.3. Then, we have,

$$\begin{aligned}
&\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2|\mathcal{F}_k\right]\leq(1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2\\
&+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{F}_k\right],
\end{aligned}\tag{H.23}$$

where

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{F}_k\right]\leq3c_k^2\|\nabla F(\mathbf{x}(k))\|^2+3c_k^2\mathbb{E}\left[\|\mathbf{h}(\mathbf{x}(k))\|^2|\mathcal{F}_k\right]\\
&+3c_k^2\|\mathbf{b}\left(\mathbf{x}(k)\right)\|^2\\
&\leq3c_k^2\|\nabla F(\mathbf{x}(k))\|^2+\frac{3}{16}c_k^4Ns_1^2(P)M^2\\
&+6dc_vq_\infty(N,d,\alpha_0,c_0)+6dN\sigma_1^2+6Nc_k^4s_2(P)\\
&+12c_k^2Ns_1(P)L^2q_\infty(N,d,\alpha_0,c_0)+12c_k^2Ns_1(P)\|\nabla F\left(\mathbf{x}^o\right)\|^2\\
&\Rightarrow\mathbb{E}\left[\|\mathbf{w}(k)\|^2\right]\leq3\left(2dc_v+c_k^2L^2(1+4Ns_1(P))\right)\\
&\times q_\infty(N,d,\alpha_0,c_0)\\
&+\frac{3}{16}c_k^4Ns_1^2(P)M^2+6Nc_k^4s_2(P)\\
&+6dN\sigma_1^2+12c_k^2Ns_1(P)\|\nabla F\left(\mathbf{x}^o\right)\|^2\\
&=\Delta_{1,\infty}+c_k^2\Delta_{2,\infty}\doteq\Delta_k\\
&\Rightarrow\mathbb{E}\left[\|\mathbf{w}(k)\|^2\right]<\infty,
\end{aligned}\tag{H.24}$$

where $\Delta_{1,\infty}=6dc_vq_\infty(N,d,\alpha_0,c_0)+6dN\sigma_1^2$ and $c_k^2\Delta_{2,\infty}=\frac{3}{16}c_k^4Ns_1^2(P)M^2+3c_k^2L^2(1+4Ns_1(P))q_\infty(N,d,\alpha_0,c_0)+12c_k^2Ns_1(P)\|\nabla F\left(\mathbf{x}^o\right)\|^2+6Nc_k^4s_2(P)$. With the above development in place, we then have,

$$\begin{aligned}
&\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right]\leq(1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2\\
&+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\Delta_k.
\end{aligned}\tag{H.25}$$

In particular, we choose $\theta(k)=\frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)$. From (H.25), we have,

$$\begin{aligned}
&\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right]\leq\left(1-\frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right]\\
&+\left(1+\frac{2}{\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)}\right)\frac{\alpha_k^2}{c_k^2}\Delta_k
\end{aligned}$$

$$= \left(1 - \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right] + \frac{2\alpha_k^2}{\lambda_2\left(\overline{\mathbf{L}}\right)c_k^2\beta_k}\Delta_k + \frac{\alpha_k^2}{c_k^2}\Delta_k. \tag{H.26}$$

For ease of analysis, define $s(k) = \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)$. We proceed by using the following technical lemma.

Lemma H.1.4. *If for all $k \geq k_0$ there holds*

$$q_{k+1} \leq (1 - s_k)q_k + \left(1 + \frac{1}{s_k}\right)b_k d_k, \tag{H.27}$$

where $q_{k_0} \geq 0$, $s_k \in (0,1)$, d_k, $b_k \geq 0$ are monotonously decreasing, then, for any $k \geq m(k) \geq k_0$

$$\begin{aligned} q_{k+1} &\leq e^{-\sum_{l=k_0}^{k} s_l} q_{k_0} + d_{k_0} e^{-\sum_{l=m(k)}^{k} s_l} \sum_{l=k_0}^{m(k)-1} \left(1 + \frac{1}{s_l}\right) b_l \\ &+ d_{m(k)} b_{m(k)} \frac{s_k + 1}{s_k^2}. \end{aligned} \tag{H.28}$$

Proof. Similarly as before, define $p(k,l) = (1 - s_k)\cdots(1 - s_l)$ for $k_0 \leq l \leq k$, and let also $p(k, k+1) = 1$. Recall that $p(k, l+1)s_l$ can be expressed as $p(k, l+1)s_l = p(k, l+1) - p(k, l)$. Then, we have:

$$\begin{aligned} q_{k+1} &\leq p(k,k_0)q_{k_0} + \sum_{l=k_0}^{k} p(k,l)\left(1 + \frac{1}{s_l} b_l d_l\right) \\ &\leq p(k,k_0)q_{k_0} + d_{k_0} p(k, m(k)) \sum_{l=k_0}^{m(k)} \left(1 + \frac{1}{s_l}\right) b_l \\ &+ b_{m(k)} d_{m(k)} \frac{s_k + 1}{s_k^2} \sum_{m(k)}^{k} \left(p(k, l+1) - p(k,l)\right), \end{aligned} \tag{H.29}$$

where we break the sum in (H.29) at $l = m(k)$, and use the fact that $p\left(k, m(k) - 1\right) \geq p(k,l)$ for every $l \leq m(k) - 1$, together with the fact that $1/s_l \leq 1/s_k$, for every $l \leq k$. Finally, noting that, for every $l \leq k$, $p(k,l) \leq e^{-\sum_{m=1}^{k} s_l}$, and also recalling that $\sum_{m(k)}^{k}\left(p(k,l+1) - p(k,l)\right) \leq 1$, proves the claim of the lemma. □

Applying the preceding lemma to $q_k = \mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right]$, $d_k = \Delta_k$, $b_k = \frac{\alpha_k^2}{c_k^2}$, and $s_k = \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)$ we have,

$$\begin{aligned} &\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] \\ &\leq \underbrace{\exp\left(-\sum_{l=0}^{k} s(l)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(0)\|^2\right]}_{t_1} \\ &+ \underbrace{\Delta_0 \exp\left(-\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k} s(m)\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor - 1}\left(\frac{2\alpha_l^2}{\lambda_2\left(\overline{\mathbf{L}}\right)c_l^2\beta_l} + \frac{\alpha_l^2}{c_l^2}\right)}_{t_2} \\ &+ \underbrace{\frac{4\Delta_{\lfloor\frac{k-1}{2}\rfloor}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2 c_0^2 (k+1)^{2-2\tau-2\delta}}}_{t_3} + \underbrace{\frac{2\Delta_{\lfloor\frac{k-1}{2}\rfloor}\alpha_0^2}{\lambda_2\left(\overline{\mathbf{L}}\right)\beta_0 c_0^2 (k+1)^{2-\tau-2\delta}}}_{t_4}. \end{aligned} \tag{H.30}$$

In the above proof, the splitting in the interval $[0,k]$ was done at $\lfloor\frac{k-1}{2}\rfloor$ for ease of book keeping. The division can be done at an arbitrary point. It is to be noted that the sequence $\{s(k)\}$ is not summable and hence terms t_1 and t_2 decay faster than $(k+1)^{2-2\tau-2\delta}$. Also, note that term t_4 decays faster than t_3.

Furthermore, t_3 can be written as

$$\frac{4\Delta_{\lfloor\frac{k-1}{2}\rfloor}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}} = \underbrace{\frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}}}_{t_{31}}$$
$$+\underbrace{\frac{4c_{\lfloor\frac{k-1}{2}\rfloor}^2\Delta_{2,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}}}_{t_{32}},$$

from which we have that t_{32} decays faster than t_{31}. For notational ease, henceforth we refer to $t_1+t_2+t_{32}+t_4=Q_k$, while keeping in mind that Q_k decays faster than $(k+1)^{2-2\tau-2\delta}$. Hence, we have the disagreement given by,

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] = O\left(\frac{1}{k^{2-2\delta-2\tau}}\right).$$

□

We now proceed to the proof of Theorem 9.7.1. Denote $\overline{\mathbf{x}}(k)=\frac{1}{N}\sum_{n=1}\mathbf{x}_i(k)$. Then, we have,

$$\begin{aligned}
&\overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k)\\
&-\frac{\alpha_k}{c_k}\Bigg[\frac{c_k}{N}\sum_{i=1}^N\nabla f_i\left(\mathbf{x}_i(k)\right)+\underbrace{\frac{c_k^2}{N}\sum_{i=1}^N\mathbf{b}_i\left(\mathbf{x}_i(k)\right)}_{\overline{\mathbf{b}}\left(\mathbf{x}(k)\right)}\\
&+\underbrace{\frac{c_k}{N}\sum_{i=1}^N\mathbf{h}_i(\mathbf{x}_i(k))}_{\overline{\mathbf{h}}(\mathbf{x}(k))}\Bigg]\\
&\Rightarrow\overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k)-\frac{\alpha_k}{c_k}\left(\overline{\mathbf{h}}(\mathbf{x}(k))+\overline{\mathbf{b}}\left(\mathbf{x}(k)\right)\right)\\
&-\frac{\alpha_k}{Nc_k}\left[c_k\sum_{i=1}^N\nabla f_i\left(\mathbf{x}_i(k)\right)-\nabla f_i\left(\overline{\mathbf{x}}(k)\right)+\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right].
\end{aligned}\tag{H.31}$$

Recall that $f(\cdot)=\sum_{i=1}^N f_i(\cdot)$. Then, we have,

$$\begin{aligned}
&\overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k)-\frac{\alpha_k}{c_k}\left(\overline{\mathbf{h}}(\mathbf{x}(k))+\overline{\mathbf{b}}\left(\mathbf{x}(k)\right)\right)\\
&-\frac{\alpha_k}{N}\nabla f\left(\overline{\mathbf{x}}(k)\right)-\frac{\alpha_k}{N}\left[\sum_{i=1}^N\nabla f_i\left(\mathbf{x}_i(k)\right)-\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right]\\
&\Rightarrow\overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k)-\frac{\alpha_k}{Nc_k}\left[c_k\nabla f\left(\overline{\mathbf{x}}(k)\right)+\mathbf{e}(k)\right],
\end{aligned}\tag{H.32}$$

where

$$\begin{aligned}
&\mathbf{e}(k)=N\overline{\mathbf{h}}(\mathbf{x}(k))\\
&\underbrace{+N\overline{\mathbf{b}}\left(\mathbf{x}(k)\right)+c_k\sum_{i=1}^N\left(\nabla f_i\left(\mathbf{x}_i(k)\right)-\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right)}_{\boldsymbol{\epsilon}(k)}.
\end{aligned}\tag{H.33}$$

Note that, $c_k\left\|\nabla f_i\left(\mathbf{x}_i(k)\right)-\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right\|\leq c_kL\left\|\mathbf{x}_i(k)-\overline{\mathbf{x}}(k)\right\|=c_kL\left\|\widetilde{\mathbf{x}}_i(k)\right\|$. We also have that, $\left\|\overline{\mathbf{b}}\left(\mathbf{x}(k)\right)\right\|\leq$

$\frac{M}{4}s_1(P)c_k^3$. Thus, we can conclude that, $\forall k \geq k_3$

$$\begin{aligned}
&\boldsymbol{\epsilon}(k) = c_k \sum_{i=1}^{N} \left(\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right) + N\overline{\mathbf{b}}\left(\mathbf{x}(k)\right) \\
&\Rightarrow \left\|\boldsymbol{\epsilon}(k)\right\|^2 \leq 2NL^2c_k^2\left\|\widetilde{\mathbf{x}}(k)\right\|^2 + \frac{N}{8}M^2d^2(P)c_k^6 \\
&\Rightarrow \mathbb{E}\left[\left\|\boldsymbol{\epsilon}(k)\right\|^2\right] \leq \frac{8NL^2\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2(k+1)^{2-2\tau}} + \frac{NM^2d^2(P)c_0^6}{8(k+1)^{6\delta}} \\
&+ \frac{2NL^2Q_kc_0^2}{(k+1)^{2\delta}}.
\end{aligned} \tag{H.34}$$

With the above development in place, we rewrite (H.32) as follows:

$$\begin{aligned}
&\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{N}\nabla f\left(\overline{\mathbf{x}}(k)\right) - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{h}}(\mathbf{x}(k)) \\
&\Rightarrow \overline{\mathbf{x}}(k+1) - \mathbf{x}^* = \overline{\mathbf{x}}(k) - \mathbf{x}^* - \frac{\alpha_k}{N}\left[\nabla f\left(\overline{\mathbf{x}}(k)\right) - \underbrace{\nabla f\left(\mathbf{x}^*\right)}_{=0}\right] \\
&- \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{h}}(\mathbf{x}(k)).
\end{aligned} \tag{H.35}$$

By Leibnitz rule, we have,

$$\begin{aligned}
&\nabla f\left(\overline{\mathbf{x}}(k)\right) - \nabla f\left(\mathbf{x}^*\right) \\
&= \underbrace{\left[\int_{s=0}^{1} \nabla^2 f\left(\mathbf{x}^* + s\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right)\right) ds\right]}_{\overline{\mathbf{H}}_k}\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right),
\end{aligned} \tag{H.36}$$

where it is to be noted that $NL \succcurlyeq \overline{\mathbf{H}}_k \succcurlyeq N\mu$. Using (H.36) in (H.35), we have,

$$\begin{aligned}
&\left(\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right) = \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right]\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right) \\
&- \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{h}}(\mathbf{x}(k)).
\end{aligned} \tag{H.37}$$

Denote by $\mathbf{m}(k) = \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right]\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right) - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k)$ and note that $\mathbf{m}(k)$ is conditionally independent from $\overline{\mathbf{h}}(\mathbf{x}(k))$ given the history $\mathcal{F}_k$. Then (H.37) can be rewritten as:

$$\begin{aligned}
&\left(\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right) = \mathbf{m}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{h}}(\mathbf{x}(k)) \\
&\Rightarrow \left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right\|^2 \leq \left\|\mathbf{m}(k)\right\|^2 - 2\frac{\alpha_k}{c_k}\mathbf{m}(k)^\top\overline{\mathbf{h}}(\mathbf{x}(k)) \\
&+ \frac{\alpha_k^2}{c_k^2}\left\|\overline{\mathbf{h}}(\mathbf{x}(k))\right\|^2.
\end{aligned} \tag{H.38}$$

Using the properties of conditional expectation and noting that $\mathbb{E}\left[\mathbf{h}(\mathbf{x}(k))|\mathcal{F}_k\right] = \mathbf{0}$, we have,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right\|^2|\mathcal{F}_k\right] \leq \left\|\mathbf{m}(k)\right\|^2 + \frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\left\|\overline{\mathbf{h}}(\mathbf{x}(k))\right\|^2|\mathcal{F}_k\right] \\
&\Rightarrow \mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right\|^2\right] \leq \mathbb{E}\left[\left\|\mathbf{m}(k)\right\|^2\right] + 2N\alpha_k^2c_k^2s_2(P) \\
&+ \frac{2\alpha_k^2\left(dc_vq_\infty(N,d,\alpha_0,c_0) + dN\sigma_1^2\right)}{c_k^2} \\
&+ 4\alpha_k^2Ns_1(P)L^2q_\infty(N,d,\alpha_0,c_0) + 4\alpha_k^2Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2.
\end{aligned} \tag{H.39}$$

For notational simplicity, we denote $\alpha_k^2\sigma_h^2 = 2N\alpha_k^2c_k^2s_2(P) + 4\alpha_k^2Ns_1(P)L^2q_\infty(N,d,\alpha_0,c_0) + 4\alpha_k^2Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2$.

Using (H.13), we have for $\mathbf{m}(k)$,

$$\begin{aligned}
\|\mathbf{m}(k)\|^2 &\leq (1+\theta_k)\left\|\mathbf{I}-\frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right\|^2\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2 \\
&+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2c_k^2}\|\boldsymbol{\epsilon}(k)\|^2 \\
&\leq (1+\theta_k)\left(1-\frac{\mu\alpha_0}{k+1}\right)^2\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2 \\
&+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2c_k^2}\|\boldsymbol{\epsilon}(k)\|^2. \qquad \text{(H.40)}
\end{aligned}$$

On choosing $\theta_k=\frac{\mu\alpha_0}{k+1}$, where $\alpha_0>\frac{1}{\mu}$, we have,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{m}(k)\|^2\right]\leq\left(1-\frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right] \\
&+\frac{16L^2\Delta_{1,\infty}N\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}+\frac{4M^2Nd^2(P)c_0^4\alpha_0}{\mu(k+1)^{1+4\delta}}+\frac{4L^2NQ_k}{\mu(k+1)} \\
&\Rightarrow\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right]\leq\left(1-\frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right] \\
&+\frac{16NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}+\frac{4NM^2d^2(P)c_0^4\alpha_0}{\mu(k+1)^{1+4\delta}}+\frac{4NL^2Q_k}{\mu(k+1)} \\
&+\frac{2\alpha_k^2\left(dc_vq_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{c_k^2}+\alpha_k^2\sigma_h^2 \\
&\Rightarrow\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right]\leq\left(1-\frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right] \\
&+\frac{16NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}} \\
&+\frac{4M^2Nd^2(P)c_0^4\alpha_0}{\mu(k+1)^{1+4\delta}}+\frac{2\alpha_0^2\left(dc_vq_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{c_0^2(k+1)^{2-2\delta}}+P_k, \qquad \text{(H.41)}
\end{aligned}$$

where $P_k=\frac{4NL^2Q_k}{\mu(k+1)}+\frac{\alpha_0^2\sigma_h^2}{(k+1)^2}$ decays faster as compared to the other terms. Proceeding as in (H.30), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right] \\
&\leq\underbrace{\exp\left(-\mu\sum_{l=0}^{k}\alpha_l\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]}_{t_6} \\
&\underbrace{+\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{16NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}}_{t_7} \\
&\underbrace{+\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{4M^2Nd^2(P)c_0^4\alpha_0}{\mu(k+1)^{1+4\delta}}}_{t_8} \\
&\underbrace{+\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}P_l+\frac{2\alpha_0^2dc_vq_\infty(N,d,\alpha_0,c_0)}{c_0^2(l+1)^{2-2\delta}}}_{t_{10}}
\end{aligned}$$

$$+\underbrace{\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{2\alpha_0^2dN\sigma_1^2}{c_0^2(l+1)^{2-2\delta}}}_{t_{11}}$$
$$+\underbrace{\frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{2-2\tau-2\delta}}}_{t_{12}}$$
$$+\underbrace{\frac{8NM^2d^2(P)c_0^4}{\mu^2(k+1)^{4\delta}}}_{t_{13}}+\underbrace{\frac{N(k+1)P_k}{\mu\alpha_0}}_{t_{14}}$$
$$+\underbrace{\frac{4N\alpha_0\left(dc_vq_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{1-2\delta}}}_{t_{15}}. \tag{H.42}$$

It is to be noted that the term t_6 decays as $1/k$. The terms t_7, t_8, t_{10}, t_{11} and t_{14} decay faster than its counterparts in the terms t_{12}, t_{13} and t_{15} respectively. We note that Q_l also decays faster. Hence, the rate of decay of $\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\right\|^2\right]$ is determined by the terms t_{12}, t_{13} and t_{15}. Thus, we have that, $\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\right\|^2\right]=O\left(k^{-\delta_1}\right)$, where $\delta_1=\min\{1-2\delta,2-2\tau-2\delta,4\delta\}$. For notational ease, we refer to $t_6+t_7+t_8+t_{10}+t_{11}+t_{14}=M_k$ from now on. Finally, we note that,

$$\begin{aligned}
&\|\mathbf{x}_i(k)-\mathbf{x}^*\|\leq\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|+\left\|\underbrace{\mathbf{x}_i(k)-\overline{\mathbf{x}}(k)}_{\widetilde{\mathbf{x}}_i(k)}\right\|\\
&\Rightarrow\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\leq2\|\widetilde{\mathbf{x}}_i(k)\|^2+2\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\\
&\Rightarrow\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]\leq2M_k+\frac{64NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{2-2\tau-2\delta}}\\
&+\frac{16NM^2d^2(P)c_0^4}{\mu^2(k+1)^{4\delta}}+2Q_k+\frac{8\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}}\\
&+\frac{4N\alpha_0\left(dc_vq_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{1-2\delta}}\\
&\Rightarrow\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]=O\left(\frac{1}{k^{\delta_1}}\right),\quad\forall i,
\end{aligned}\tag{H.43}$$

where $\delta_1=\min\{1-2\delta,2-2\tau-2\delta,4\delta\}$. By, optimizing over τ and δ, we obtain that for $\tau=1/2$ and $\delta=1/6$,

$$\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]=O\left(\frac{1}{k^{\frac{2}{3}}}\right),\quad\forall i.$$

The communication cost is given by,

$$\mathbb{E}\left[\sum_{t=1}^{k}\zeta_t\right]=O\left(k^{\frac{3}{4}+\frac{\epsilon}{2}}\right).$$

Thus, we achieve the communication rate to be,

$$\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^\star\|^2\right]=O\left(\frac{1}{\mathcal{C}_k^{8/9-\zeta}}\right), \tag{H.44}$$

where ζ can be arbitrarily small.

Appendix I

Proofs of Theorems in Chapter 10

Lemma I.0.1. *Consider the proposed zeroth order Frank Wolfe Algorithm. Let Assumptions 10.4.1-10.4.5 hold. Then, the sub-optimality $F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)$ satisfies*

$$\begin{aligned} &F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq (1-\gamma_{t+1})(F(\mathbf{x}_t) - F(\mathbf{x}^*)) \\ &+ \gamma_{t+1} R\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\| + \frac{LR^2\gamma_{t+1}^2}{2}. \end{aligned} \tag{I.1}$$

Proof. The L-smoothness of the function f yields the following upper bound on $f(\mathbf{x}_{t+1})$:

$$\begin{aligned} &f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^T(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \mathbf{d}_t)^T(\mathbf{v}_t - \mathbf{x}_t) + \gamma_{t+1}\mathbf{d}_t^T(\mathbf{v}_t - \mathbf{x}_t) \\ &+ \frac{L\gamma_{t+1}^2}{2}\|\mathbf{v}_t - \mathbf{x}_t\|^2 \end{aligned} \tag{I.2}$$

Since $\langle \mathbf{x}^*, \mathbf{d}_t\rangle \geq \min_{v\in\mathcal{C}}\{\langle \mathbf{v}, \mathbf{d}_t\rangle\} = \langle \mathbf{v}_t, \mathbf{d}_t\rangle$, we have,

$$\begin{aligned} &f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \mathbf{d}_t)^T(\mathbf{v}_t - \mathbf{x}_t) \\ &+ \gamma_{t+1}\mathbf{d}_t^T(\mathbf{x}^* - \mathbf{x}_t) + \frac{L\gamma_{t+1}^2}{2}\|\mathbf{v}_t - \mathbf{x}_t\|^2 \\ &\leq f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \mathbf{d}_t)^T(\mathbf{v}_t - \mathbf{x}^*) \\ &+ \gamma_{t+1}\nabla f(\mathbf{x}_t)^T(\mathbf{x}^* - \mathbf{x}_t) + \frac{LR\gamma_{t+1}^2}{2}\|\mathbf{v}_t - \mathbf{x}_t\|^2. \end{aligned} \tag{I.3}$$

Using Cauchy-Schwarz inequality, we have,

$$\begin{aligned} &f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \gamma_{t+1}\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|\|\mathbf{v}_t - \mathbf{x}^*\| \\ &- \gamma_{t+1}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{L\gamma_{t+1}^2}{2}\|\mathbf{v}_t - \mathbf{x}^*\|^2 \\ &\leq f(\mathbf{x}_t) + \gamma_{t+1}R\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\| - \gamma_{t+1}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\ &+ \frac{LR^2\gamma_{t+1}^2}{2}, \end{aligned} \tag{I.4}$$

and subtracting $f(\mathbf{x}^*)$ from both sides of (I.4), we have,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq (1-\gamma_{t+1})(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \gamma_{t+1}R\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\| + \frac{LR^2\gamma_{t+1}^2}{2}. \tag{I.5}$$

□

Proof of Theorem 10.4.1. We have, from Lemma I.0.1,

$$\begin{aligned} &F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq (1-\gamma_{t+1})(F(\mathbf{x}_t) - F(\mathbf{x}^*)) \\ &+ \gamma_{t+1}R\|\nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t)\| + \frac{LR^2\gamma_{t+1}^2}{2} \\ &\Rightarrow F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq (1-\gamma_{t+1})(F(\mathbf{x}_t) - F(\mathbf{x}^*)) \\ &+ \frac{c_{t+1}d}{2}\gamma_{t+1}R^2 + \frac{LR^2\gamma_{t+1}^2}{2}. \end{aligned} \tag{I.6}$$

From, (I.6), we have,

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq (1-\gamma_{t+1})(F(\mathbf{x}_t) - F(\mathbf{x}^*)) + LR^2\gamma_{t+1}^2. \tag{I.7}$$

We use Lemma I.1.1 to derive the primal gap which then yields,

$$F(\mathbf{x}_t) - F(\mathbf{x}^*) = \frac{Q_{ns}}{t+2}, \tag{I.8}$$

where $Q_{ns} = \max\{2(F(\mathbf{x}_0) - F(\mathbf{x}^*)), 4LR^2\}$. □

I.1 Proofs of Zeroth Order Stochastic Frank Wolfe: RDSA

Proof of Lemma 10.4.2 (1). Use the definition $\mathbf{d}_t := (1-\rho_t)\mathbf{d}_{t-1} + \rho_t g(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_t)$ to write the difference $\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2$ as

$$\begin{aligned} &\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2 = \|\nabla f(\mathbf{x}_t) - (1-\rho_t)\mathbf{d}_{t-1} \\ &- \rho_t g(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_t)\|^2. \end{aligned} \tag{I.9}$$

Add and subtract the term $(1-\rho_t)\nabla f(\mathbf{x}_{t-1})$ to the right hand side of (I.9), regroup the terms and expand the squared term to obtain

$$\begin{aligned} &\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2 \\ &= \|\nabla f(\mathbf{x}_t) - (1-\rho_t)\nabla f(\mathbf{x}_{t-1}) + (1-\rho_t)\nabla f(\mathbf{x}_{t-1}) \\ &- (1-\rho_t)\mathbf{d}_{t-1} - \rho_t g(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_t)\|^2 \\ &= \rho_t^2\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_t)\|^2 + (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\ &+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\ &+ 2\rho_t(1-\rho_t)(\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_t))^T(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})) \end{aligned}$$

$$+ 2\rho_t(1-\rho_t)(\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t))^T(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1})$$
$$+ 2(1-\rho_t)^2(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}))^T(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}). \tag{I.10}$$

Compute the expectation $\mathbb{E}\left[(.) \mid \mathcal{F}_t\right]$ for both sides of (I.10), where $\mathcal{F}_t$ is the σ-algebra given by $\{\{\mathbf{y}_s\}_{s=0}^{t-1}, \{\mathbf{z}_s\}_{s=0}^{t-1}\}$to obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2 \mid \mathcal{F}_t\right] \\
&= \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right] \\
&+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\
&+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\
&+ 2(1-\rho_t)^2(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}))^T(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}) \\
&+ 2\rho_t(1-\rho_t)\mathbb{E}\left[(\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t))^T(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})) \mid \mathcal{F}_t\right] \\
&+ 2\rho_t(1-\rho_t)\mathbb{E}\left[(\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t))^T(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}) \mid \mathcal{F}_t\right] \\
&\leq \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right] \\
&+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\
&+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\
&+ (1-\rho_t)^2\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + \frac{(1-\rho_t)^2}{\beta_t}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\
&+ 2\rho_t(1-\rho_t)(c_tL\mathbf{v}\left(\mathbf{x},c_t\right))^\top(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})) \\
&+ 2\rho_t(1-\rho_t)(c_tL\mathbf{v}\left(\mathbf{x},c_t\right))^\top(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}) \\
&\leq \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right] \\
&+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\
&+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\
&+ (1-\rho_t)^2\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + \frac{(1-\rho_t)^2}{\beta_t}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\
&+ 2\rho_t(1-\rho_t)c_t^2\left\|L\mathbf{v}\left(\mathbf{x},c_t\right)\right\|^2 + \rho_t(1-\rho_t)\left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\right\|^2 \\
&+ \rho_t(1-\rho_t)\left\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\right\|^2 \\
&\Rightarrow \mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\
&\leq \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla F(\mathbf{x}_t,\mathbf{y}_t) + \nabla F(\mathbf{x}_t,\mathbf{y}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2\right] \\
&+ (1-\rho_t)^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ (1-\rho_t)^2\|\mathbb{E}\left[\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ (1-\rho_t)^2\beta_t\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{(1-\rho_t)^2}{\beta_t}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \frac{\rho_t}{4}(1-\rho_t)c_t^2L^2M(\mu) + \rho_t(1-\rho_t)\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\right\|^2\right] \\
&+ \rho_t(1-\rho_t)\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\right\|^2\right] \\
&\leq 2\rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla F(\mathbf{x}_t,\mathbf{y}_t)\|^2\right] \\
&+ 2\rho_t^2\mathbb{E}\left[\|\nabla F(\mathbf{x}_t,\mathbf{y}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+\left(1-\rho_t+\frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right]\\
&+\left(1-\rho_t+(1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right]\\
&+\frac{\rho_t}{2}(1-\rho_t)c_t^2L^2M(\mu)\\
&\leq 2\rho_t^2\sigma^2+4\rho_t^2\mathbb{E}\left[\|\nabla F(\mathbf{x}_t,\mathbf{y}_t)\|^2\right]+4\rho_t^2\mathbb{E}\left[\|g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2\right]\\
&+\left(1-\rho_t+\frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right]\\
&+\left(1-\rho_t+(1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right]\\
&+\frac{\rho_t}{2}(1-\rho_t)c_t^2L^2M(\mu)\\
&\leq 2\rho_t^2\sigma^2+4\rho_t^2L_1^2+8\rho_t^2s(d)L_1^2+2\rho_t^2c_t^2L^2M(\mu)\\
&+\left(1-\rho_t+\frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right]\\
&+\left(1-\rho_t+(1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right]\\
&+\frac{\rho_t}{2}c_t^2L^2M(\mu),
\end{aligned}
\tag{I.11}
$$

where we used the gradient approximation bounds as stated in (10.15) and used Young's inequality to substitute the inner products and in particular substituted $2\langle\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1}),\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\rangle$ by the upper bound $\beta_t\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2+(1/\beta_t)\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2$ where $\beta_t>0$ is a free parameter.
By assumption 10.4.4, the norm $\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|$ is bounded above by $L\|\mathbf{x}_t-\mathbf{x}_{t-1}\|$. In addition, the condition in Assumption 10.4.1 implies that $L\|\mathbf{x}_t-\mathbf{x}_{t-1}\|=L\gamma_t\|\mathbf{v}_t-\mathbf{x}_t\|\leq\gamma_tLR$. Therefore, we can replace $\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|$ by its upper bound γ_tLR and since we assume that $\rho_t\leq 1$ we can replace all the terms $(1-\rho_t)^2$. Furthermore, using $\beta_t:=\rho_t/2$ we have,

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]\\
&\leq 2\rho_t^2\sigma^2+4\rho_t^2L_1^2+8\rho_t^2s(d)L_1^2+2\rho_t^2c_t^2L^2M(\mu)\\
&+\gamma_t^2(1-\rho_t)\left(1+\frac{2}{\rho_t}\right)L^2R^2+\frac{\rho_t}{2}c_t^2L^2M(\mu)\\
&+(1-\rho_t)\left(1+\frac{\rho_t}{2}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right].
\end{aligned}
\tag{I.12}
$$

Now using the inequalities $(1-\rho_t)(1+(2/\rho_t))\leq(2/\rho_t)$ and $(1-\rho_t)(1+(\rho_t/2))\leq(1-\rho/2)$ we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]\leq 2\rho_t^2\sigma^2+4\rho_t^2L_1^2\\
&+8\rho_t^2s(d)L_1^2+2\rho_t^2c_t^2L^2M(\mu)\\
&+\frac{2L^2R^2\gamma_t^2}{\rho_t}+\frac{\rho_t}{2}c_t^2L^2M(\mu)\\
&+\left(1-\frac{\rho_t}{2}\right))\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right].
\end{aligned}
\tag{I.13}
$$

□

Then, we have, from Lemma I.0.1

$$
\mathbb{E}\left[f(\mathbf{x}_{t+1})-f(\mathbf{x}^*)\right]\leq(1-\gamma_{t+1})\mathbb{E}\left[(f(\mathbf{x}_t)-f(\mathbf{x}^*))\right]
$$

$$+\gamma_{t+1}R\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|\right]+\frac{LR^2\gamma_{t+1}^2}{2}, \tag{I.14}$$

and then by using Jensen's inequality, we obtain,

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1})-f(\mathbf{x}^*)\right] &\leq (1-\gamma_{t+1})\mathbb{E}\left[(f(\mathbf{x}_t)-f(\mathbf{x}^*))\right]\\ &+\gamma_{t+1}R\sqrt{\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]}+\frac{LR^2\gamma_{t+1}^2}{2}.\end{aligned} \tag{I.15}$$

We state a Lemma next which will be crucial for the rest of the paper.

Lemma I.1.1. *Let $z(k)$ be a non-negative (deterministic) sequence satisfying:*

$$z(k+1)\leq(1-r_1(k))\,z_1(k)+r_2(k),$$

where $\{r_1(k)\}$ and $\{r_2(k)\}$ are deterministic sequences with

$$\frac{a_1}{(k+1)^{\delta_1}}\leq r_1(k)\leq 1 \text{ and } r_2(k)\leq\frac{a_2}{(k+1)^{2\delta_1}},$$

with $a_1>0$, $a_2>0$, $1>\delta_1>1/2$ and $k_0\geq 1$. Then,

$$z(k+1)\leq\exp\left(-\frac{a_1\delta_1(k+k_0)^{1-\delta_1}}{4(1-\delta_1)}\right)\left(z(0)+\frac{a_2}{k_0^{\delta_1}(2\delta_1-1)}\right)+\frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}}.$$

Proof of Lemma I.1.1. We have,

$$\begin{aligned}z(k+1)&\leq\prod_{l=0}^{k}\left(1-\frac{a_1}{(l+k_0)^{\delta_1}}\right)z(0)\sum_{l=0}^{\lfloor\frac{k}{2}\rfloor-1}\prod_{m=l+1}^{k}\left(1-\frac{a_1}{(m+k_0)^{\delta_1}}\right)\frac{a_2}{(k+k_0)^{2\delta_1}}\\
&+\sum_{l=\lfloor\frac{k}{2}\rfloor}^{k}\prod_{m=l+1}^{k}\left(1-\frac{a_1}{(m+k_0)^{\delta_1}}\right)\frac{a_2}{(k+k_0)^{2\delta_1}}\\
&\leq\exp\left(\sum_{l=0}^{k}\left(1-\frac{a_1}{(l+k_0)^{\delta_1}}\right)\right)z(0)+\prod_{m=l+1}^{k}\left(1-\frac{a_1}{(m+k_0)^{\delta_1}}\right)\sum_{l=0}^{\lfloor\frac{k}{2}\rfloor-1}\frac{a_2}{(k+k_0)^{2\delta_1}}\\
&+\frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}}\sum_{l=\lfloor\frac{k}{2}\rfloor}^{k}\prod_{m=l+1}^{k}\left(1-\frac{a_1}{(m+k_0)^{\delta_1}}\right)\frac{a_1}{(k+k_0)^{\delta_1}}\\
&\leq\exp\left(-\sum_{l=0}^{k}\frac{a_1}{(l+k_0)^{\delta_1}}\right)z(0)+\frac{a_2}{a_1k_0^{\delta_1}}\exp\left(-\sum_{m=\lfloor\frac{k}{2}\rfloor}^{k}\frac{a_1}{(m+k_0)^{\delta_1}}\right)\sum_{l=0}^{\lfloor\frac{k}{2}\rfloor-1}\frac{a_1}{(k+k_0)^{2\delta_1}}\\
&+\frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}}\sum_{l=\lfloor\frac{k}{2}\rfloor}^{k}\left(\prod_{m=l+1}^{k}\left(1-\frac{a_1}{(m+k_0)^{\delta_1}}\right)-\prod_{m=l}^{k}\left(1-\frac{a_1}{(m+k_0)^{\delta_1}}\right)\right)\\
&\leq\exp\left(-\sum_{l=0}^{k}\frac{a_1}{(l+k_0)^{\delta_1}}\right)z(0)+\frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}}+\frac{a_2}{a_1k_0^{\delta_1}}\exp\left(-\sum_{m=\lfloor\frac{k}{2}\rfloor}^{k}\frac{a_1}{(m+k_0)^{\delta_1}}\right)\sum_{l=0}^{\lfloor\frac{k}{2}\rfloor-1}\frac{a_1}{(k+k_0)^{2\delta_1}}\end{aligned}$$

$$\leq \exp\left(-\sum_{l=0}^{k}\frac{a_1}{(l+k_0)^{\delta_1}}\right)z(0)+\frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}}+\frac{a_2}{k_0^{\delta_1}}\exp\left(-\frac{a_1\delta_1}{4(1-\delta_1)}(k+k_0)^{1-\delta_1}\right)\frac{1}{2\delta_1-1}, \tag{I.16}$$

where we used the inequality that,

$$\sum_{m=\lfloor\frac{k}{2}\rfloor}^{k}\frac{1}{(m+k_0)^{\delta_1}}\geq\frac{1}{2(1-\delta_1)}(k+k_0)^{1-\delta_1}-\frac{1}{2(1-\delta_1)}\left(\frac{k}{2}+k_0\right)^{1-\delta_1}$$
$$\geq\frac{1}{2^{1+\delta_1}(1-\delta_1)}(k+k_0)^{1-\delta_1}\left(2^{1-\delta_1}-1-\frac{(1-\delta_1)k_0}{k+k_0}\right)\geq\frac{\delta_1}{4(1-\delta_1)}(k+k_0)^{1-\delta_1}$$

Following up with (I.16), we have,

$$z(k+1)\leq\exp\left(-\sum_{l=0}^{k}-\frac{a_1}{(l+k_0)^{\delta_1}}\right)z(0)+\frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}}+\frac{a_2}{k_0^{\delta_1}}\exp\left(-\frac{a_1\delta_1}{4(1-\delta_1)}(k+k_0)^{1-\delta_1}\right)\frac{1}{2\delta_1-1}$$
$$\leq\exp\left(-\frac{a_1\delta_1(k+k_0)^{1-\delta_1}}{4(1-\delta_1)}\right)\left(z(0)+\frac{a_2}{k_0^{\delta_1}(2\delta_1-1)}\right)+\frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}}. \tag{I.17}$$

For $\delta=2/3$, we have,

$$z(k+1)\leq\exp\left(-\frac{a_1(k+k_0)^{1/3}}{2}\right)\left(z(0)+\frac{3a_2}{k_0^{2/3}}\right)+\frac{a_2 2^{2/3}}{a_1\left(k+k_0\right)^{2/3}}.$$

□

Proof of Theorem 10.5.1 (1). Now using the result in Lemma I.1.1 we can characterize the convergence of the sequence of expected errors $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]$ to zero. To be more precise, using the result in Lemma 10.4.2 and setting $\gamma_t=2/(t+8)$, $\rho_t=4/d^{1/3}(t+8)^{2/3}$ and $c_t=2/\sqrt{M(\mu)}(t+8)^{1/3}$ for any $\epsilon>0$ to obtain

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]$$
$$\leq\left(1-\frac{2}{d^{1/3}(t+8)^{2/3}}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right]$$
$$+\frac{32d^{-1/3}\sigma^2+64d^{-1/3}L_1^2+128d^{2/3}L_1^2+2L^2R^2d^{2/3}+416d^{2/3}L^2}{(t+8)^{4/3}}. \tag{I.18}$$

According to the result in Lemma I.1.1, the inequality in (I.18) implies that

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]\leq\overline{Q}+\frac{Q}{(t+8)^{2/3}}\leq\frac{2Q}{(t+8)^{2/3}}, \tag{I.19}$$

where $Q=32d^{-1/3}\sigma^2+64d^{-1/3}L_1^2+128d^{2/3}L_1^2+2L^2R^2d^{2/3}+416d^{2/3}L^2$, where $\overline{Q}$ is a function of $\mathbb{E}\left[\|\nabla f(\mathbf{x}_0)-\mathbf{d}_0\|^2\right]$ and decays exponentially. Now we proceed by replacing the term $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]$ in (I.15) by its upper bound in (I.19) and γ_{t+1} by $2/(t+9)$ to write

$$\mathbb{E}\left[f(\mathbf{x}_{t+1})-f(\mathbf{x}^*)\right]\leq\left(1-\frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t)-f(\mathbf{x}^*))\right]$$
$$+\frac{R\sqrt{Q}}{(t+9)^{4/3}}+\frac{2LR^2}{(t+9)^2}. \tag{I.20}$$

Note that we can write $(t+9)^2 = (t+9)^{4/3}(t+9)^{2/3} \geq (t+9)^{4/3}9^{2/3} \geq 4(t+9)^{4/3}$.Therefore,

$$\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] \leq \left(1 - \frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right] + \frac{2R\sqrt{Q} + LD^2/2}{(t+9)^{4/3}}. \tag{I.21}$$

We use induction to prove for $t \geq 0$,

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{Q'}{(t+9)^{1/3}},$$

where $Q' = \max\{9^{1/3}(f(\mathbf{x}_0) - f(\mathbf{x}^*)), 2R\sqrt{2Q} + LR^2/2\}$. For $t = 0$, we have that $\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{Q'}{9^{1/3}}$, which is turn follows from the definition of Q'. Assume for the induction hypothesis holds for $t = k$. Then, for $t = k+1$, we have,

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right] &\leq \left(1 - \frac{2}{k+9}\right)\mathbb{E}\left[(f(\mathbf{x}_k) - f(\mathbf{x}^*))\right] \\ &+ \frac{2R\sqrt{2Q} + LD^2/2}{(k+9)^{4/3}} \\ &\leq \left(1 - \frac{2}{k+9}\right)\frac{Q'}{(t+9)^{1/3}} + \frac{Q'}{(t+9)^{4/3}} \leq \frac{Q'}{(t+10)^{1/3}}.\end{aligned}$$

Thus, for $t \geq 0$ from Lemma I.1.1 we have that,

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{Q'}{(t+9)^{1/3}} = O\left(\frac{d^{1/3}}{(t+9)^{1/3}}\right). \tag{I.22}$$

where $Q' = \max\{2(f(\mathbf{x}_0) - f(\mathbf{x}^*)), 2R\sqrt{2Q} + LR^2/2\}$. □

Proof of Theorem 10.5.2(1). Then, we have,

$$\begin{aligned}&F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \gamma_t\langle \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \mathbf{x}_t\rangle \\ &+ \gamma_t\langle \nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{LR^2\gamma_t^2}{2} \\ &\Rightarrow F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \gamma_t\langle \mathbf{g}(\mathbf{x}_t), \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle - \mathbf{x}_t\rangle \\ &+ \gamma_t\langle \nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{LR^2\gamma_t^2}{2} \\ &\Rightarrow F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \gamma_t\langle \nabla F(\mathbf{x}_t), \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle - \mathbf{x}_t\rangle \\ &+ \gamma_t\langle \nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle\rangle + \frac{LR^2\gamma_t^2}{2} \\ &\Rightarrow F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) - \gamma_t\mathcal{G}\left(\mathbf{x}_t\right) \\ &+ \gamma_t\langle \nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle\rangle + \frac{LR^2\gamma_t^2}{2} \\ &\Rightarrow \gamma_t\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[F(\mathbf{x}_t) - F(\mathbf{x}_{t+1})\right] + \gamma_t R\frac{\sqrt{2Q}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t^2}{2} + \\ &\Rightarrow \mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[\frac{t+7}{2}F(\mathbf{x}_t) - \frac{t+8}{2}F(\mathbf{x}_{t+1}) + \frac{1}{2}F(\mathbf{x}_t)\right] + R\frac{\sqrt{2Q}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}\end{aligned}$$

$$
\begin{aligned}
&\Rightarrow \sum_{t=0}^{T-1} \mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0) - \frac{T+7}{2}F(\mathbf{x}_T) + \sum_{t=0}^{T-1}\left(\frac{1}{2}F(\mathbf{x}_t)\right] + R\frac{\sqrt{2Q}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}\right) \\
&\Rightarrow \sum_{t=0}^{T-1} \mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*)\right] + \sum_{t=0}^{T-1}\left(\frac{1}{2}\left(F(\mathbf{x}_t) - F(\mathbf{x}^*)\right) + R\frac{\sqrt{2Q}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}\right) \\
&\Rightarrow \sum_{t=0}^{T-1} \mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*) + \sum_{t=0}^{T-1}\left(\frac{Q^{'} + R\sqrt{2Q}}{2(t+8)^{1/3}} + \frac{LR^2}{(t+8)}\right) \\
&\Rightarrow T\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*) + LR^2\ln(T+7) + \frac{Q^{'} + R\sqrt{2Q}}{2}(T+7)^{2/3} \\
&\Rightarrow \mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \frac{7(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{2T} + \frac{LR^2\ln(T+7)}{T} + \frac{Q^{'} + R\sqrt{2Q}}{2T}(T+7)^{2/3}. \qquad \text{(I.23)}
\end{aligned}
$$

□

I.2 Proofs for Improvised RDSA

Proof of Lemma 10.4.2(2). Following as in the proof of RDSA, we have,

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2 \mid \mathcal{F}_t\right] \\
&\leq \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right] \\
&+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\
&+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\
&+ (1-\rho_t)^2\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\
&+ \frac{(1-\rho_t)^2}{\beta_t}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\
&+ 2\rho_t(1-\rho_t)\frac{c_t^2}{m^2}\left\|L\mathbf{v}\left(\mathbf{x},c_t\right)\right\|^2 + \rho_t(1-\rho_t)\left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\right\|^2 \\
&+ \rho_t(1-\rho_t)\left\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\right\|^2 \\
&\Rightarrow \mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq 2\rho_t^2\sigma^2 + 4\rho_t^2\mathbb{E}\left[\|\nabla F(\mathbf{x}_t,\mathbf{y}_t)\|^2\right] \\
&+ 4\rho_t^2\mathbb{E}\left[\|g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2\right] \\
&+ \left(1 - \rho_t + \frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1 - \rho_t + (1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{\rho_t}{2}(1-\rho_t)c_t^2L^2M(\mu) \\
&\leq 2\rho_t^2\sigma^2 + 4\rho_t^2L_1^2 + 8\rho_t^2\left(1 + \frac{s(d)}{m}\right)L_1^2 + \left(\frac{1+m}{2m}\right)\rho_t^2c_t^2L^2M(\mu) \\
&+ \left(1 - \rho_t + \frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1 - \rho_t + (1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{\rho_t}{2m^2}c_t^2L^2M(\mu), \qquad \text{(I.24)}
\end{aligned}
$$

where we used the gradient approximation bounds as stated in (10.15) and used Young's inequality to substitute the inner products and in particular substituted $2\langle\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\rangle$ by the

upper bound $\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + (1/\beta_t)\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2$ where $\beta_t > 0$ is a free parameter. According to Assumption 2.4.2, the norm $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|$ is bounded above by $L\|\mathbf{x}_t - \mathbf{x}_{t-1}\|$. In addition, the condition in Assumption 2.3.1 implies that $L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| = L\gamma_t\|\mathbf{v}_t - \mathbf{x}_t\| \leq \gamma_t LR$. Therefore, we can replace $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|$ by its upper bound $\gamma_t LR$ and since we assume that $\rho_t \leq 1$ we can replace all the terms $(1-\rho_t)^2$. Furthermore, using $\beta_t := \rho_t/2$ we have,

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\
&\leq 2\rho_t^2\sigma^2 + 4\rho_t^2 L_1^2 + 8\rho_t^2\left(1 + \frac{s(d)}{m}\right)L_1^2 + \frac{\rho_t}{2m^2}c_t^2 L^2 M(\mu) \\
&+ \gamma_t^2(1-\rho_t)\left(1 + \frac{2}{\rho_t}\right)L^2R^2 + \left(\frac{1+m}{2m}\right)\rho_t^2 c_t^2 L^2 M(\mu) \\
&+ (1-\rho_t)\left(1 + \frac{\rho_t}{2}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right].
\end{aligned} \tag{I.25}$$

Now using the inequalities $(1-\rho_t)(1+(2/\rho_t)) \leq (2/\rho_t)$ and $(1-\rho_t)(1+(\rho_t/2)) \leq (1-\rho/2)$ we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq 2\rho_t^2\sigma^2 + 4\rho_t^2 L_1^2 + 8\rho_t^2\left(1 + \frac{s(d)}{m}\right)L_1^2 \\
&+ \left(\frac{1+m}{2m}\right)\rho_t^2 c_t^2 L^2 M(\mu) + \frac{2L^2R^2\gamma_t^2}{\rho_t} + \frac{\rho_t}{2m^2}c_t^2 L^2 M(\mu) \\
&+ \left(1 - \frac{\rho_t}{2}\right))\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right].
\end{aligned} \tag{I.26}$$

□

Proof of Theorem 10.5.1(2). Now using the result in Lemma I.1.1 we can characterize the convergence of the sequence of expected errors $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ to zero. To be more precise, using the result in Lemma 10.4.2 and setting $\gamma_t = 2/(t+8)$, $\rho_t = 4/\left(1+\frac{d}{m}\right)^{1/3}(t+8)^{2/3}$ and $c_t = 2\sqrt{m}/\sqrt{M(\mu)}(t+8)^{1/3}$, we have,

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\
&\leq \left(1 - \frac{2}{\left(1+\frac{d}{m}\right)^{1/3}(t+8)^{2/3}}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{32\left(1+\frac{d}{m}\right)^{-1/3}\sigma^2 + 64L_1^2\left(1+\frac{d}{m}\right)^{-1/3} + 128\left(1+\frac{d}{m}\right)^{2/3}L_1^2}{(t+8)^{4/3}} \\
&+ \frac{2L^2R^2\left(1+\frac{d}{m}\right)^{2/3} + 416\left(1+\frac{d}{m}\right)^{2/3}L^2}{(t+8)^{4/3}}.
\end{aligned} \tag{I.27}$$

According to the result in Lemma I.1.1, the inequality in (I.18) implies that

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq \overline{Q}_{ir} + \frac{Q_{ir}}{(t+8)^{2/3}} \leq \frac{Q_{ir}}{(t+8)^{2/3}}, \tag{I.28}$$

where $Q_{ir} = 32\left(1+\frac{d}{m}\right)^{-1/3}\sigma^2 + 128\left(1+\frac{d}{m}\right)^{2/3}L_1^2 + 64\left(1+\frac{d}{m}\right)^{-1/3}L_1^2 + 2L^2R^2\left(1+\frac{d}{m}\right)^{2/3} + 416\left(1+\frac{d}{m}\right)^{2/3}L^2$ and $\overline{Q}_{ir}$ is a function of $\mathbb{E}\left[\|\nabla f(\mathbf{x}_0) - \mathbf{d}_0\|^2\right]$ and decays exponentially. Now we proceed by replacing the term $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ in (I.15) by its upper bound in (I.28) and γ_{t+1} by $2/(t+9)$ to write

$$\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] \leq \left(1 - \frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right]$$

$$+\frac{R\sqrt{2Q_{ir}}}{(t+9)^{4/3}}+\frac{2LR^2}{(t+9)^2}. \tag{I.29}$$

Note that we can write $(t+9)^2=(t+9)^{4/3}(t+9)^{2/3}\geq(t+9)^{4/3}9^{2/3}\geq 4(t+9)^{4/3}$.Therefore,

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1})-f(\mathbf{x}^*)\right]&\leq\left(1-\frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t)-f(\mathbf{x}^*))\right]\\&+\frac{2R\sqrt{Q}+LD^2/2}{(t+9)^{4/3}}.\end{aligned} \tag{I.30}$$

Following the induction steps as in (I.22), we have,

$$\mathbb{E}\left[f(\mathbf{x}_t)-f(\mathbf{x}^*)\right]\leq\frac{Q_{ir}^{'}}{(t+8)^{1/3}}=O\left(\frac{(d/m)^{1/3}}{(t+9)^{1/3}}\right). \tag{I.31}$$

where $Q_{ir}'=\max\{2(f(\mathbf{x}_0)-f(\mathbf{x}^*)),2R\sqrt{2Q_{ir}}+LR^2/2\}$. □

Proof of Theorem 10.5.2(2). Following as in (I.23), we have,

$$\begin{aligned}&\gamma_t\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\mathbb{E}\left[F(\mathbf{x}_t)-F(\mathbf{x}_{t+1})\right]+\gamma_tR\frac{\sqrt{2Q_{ir}}}{(t+8)^{1/3}}+\frac{LR^2\gamma_t^2}{2}+\\&\Rightarrow\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\mathbb{E}\left[\frac{t+7}{2}F(\mathbf{x}_t)-\frac{t+8}{2}F(\mathbf{x}_{t+1})+\frac{1}{2}F(\mathbf{x}_t)\right]+R\frac{\sqrt{2Q_{ir}}}{(t+8)^{1/3}}+\frac{LR^2\gamma_t}{2}\\&\Rightarrow\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0)-\frac{T+7}{2}F(\mathbf{x}_T)+\sum_{t=0}^{T-1}\left(\frac{1}{2}F(\mathbf{x}_t)\right]+R\frac{\sqrt{2Q_{ir}}}{(t+8)^{1/3}}+\frac{LR^2\gamma_t}{2}\right)\\&\Rightarrow\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0)-\frac{7}{2}F(\mathbf{x}^*)\right]+\sum_{t=0}^{T-1}\left(\frac{1}{2}\left(F(\mathbf{x}_t)-F(\mathbf{x}^*)\right)+R\frac{\sqrt{2Q_{ir}}}{(t+8)^{1/3}}+\frac{LR^2\gamma_t}{2}\right)\\&\Rightarrow\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\frac{7}{2}F(\mathbf{x}_0)-\frac{7}{2}F(\mathbf{x}^*)+\sum_{t=0}^{T-1}\left(\frac{Q_{ir}^{'}+R\sqrt{2Q_{ir}}}{2(t+8)^{1/3}}+\frac{LR^2}{(t+8)}\right)\\&\Rightarrow T\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\frac{7}{2}F(\mathbf{x}_0)-\frac{7}{2}F(\mathbf{x}^*)+LR^2\ln(T+7)+\frac{Q_{ir}^{'}+R\sqrt{2Q_{ir}}}{2}(T+7)^{2/3}\\&\Rightarrow\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\frac{7(F(\mathbf{x}_0)-F(\mathbf{x}^*))}{2T}+\frac{LR^2\ln(T+7)}{T}+\frac{Q_{ir}^{'}+R\sqrt{2Q_{ir}}}{2T}(T+7)^{2/3}\end{aligned} \tag{I.32}$$

□

I.3 Proofs for KWSA

Proof of Lemma 10.4.2(3). Following as in the proof of Lemma 10.4.2, we have,

$$\begin{aligned}&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]\\&\leq(1-\rho_t)^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right]\\&+(1-\rho_t)^2\|\mathbb{E}\left[\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right]\\&+(1-\rho_t)^2\beta_t\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right]\\&+\frac{(1-\rho_t)^2}{\beta_t}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\rho_t}{2}(1-\rho_t)c_t^2 L^2 d \\
&+ \rho_t(1-\rho_t)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \rho_t(1-\rho_t)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&\leq 2\rho_t^2\sigma^2 + 2\rho_t^2 c_t^2 dL^2 \\
&+ \left(1 - \rho_t + \frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1 - \rho_t + (1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{\rho_t}{2}(1-\rho_t)c_t^2 L^2 d \\
&\leq 2\rho_t^2\sigma^2 + 2\rho_t c_t^2 dL^2 \\
&+ \left(1 - \rho_t + \frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1 - \rho_t + (1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right], \qquad \text{(I.33)}
\end{aligned}
$$

where we used the gradient approximation bounds as stated in (10.15) and used Young's inequality to substitute the inner products and in particular substituted $2\langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\rangle$ by the upper bound $\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + (1/\beta_t)\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2$ where $\beta_t > 0$ is a free parameter. According to Assumption 2.4.2, the norm $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|$ is bounded above by $L\|\mathbf{x}_t - \mathbf{x}_{t-1}\|$. In addition, the condition in Assumption 2.3.1 implies that $L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| = L\gamma_t\|\mathbf{v}_t - \mathbf{x}_t\| \leq \gamma_t LR$. Therefore, we can replace $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|$ by its upper bound $\gamma_t LR$ and since we assume that $\rho_t \leq 1$ we can replace all the terms $(1-\rho_t)^2$. Furthermore, using $\beta_t := \rho_t/2$ we have,

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\
&\leq 2\rho_t^2\sigma^2 + 2\rho_t c_t^2 dL^2 + \gamma_t^2(1-\rho_t)\left(1 + \frac{2}{\rho_t}\right)L^2R^2 \\
&+ (1-\rho_t)\left(1 + \frac{\rho_t}{2}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right]. \qquad \text{(I.34)}
\end{aligned}
$$

Now using the inequalities $(1-\rho_t)(1 + (2/\rho_t)) \leq (2/\rho_t)$ and $(1-\rho_t)(1 + (\rho_t/2)) \leq (1 - \rho/2)$ we obtain

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] &\leq 2\rho_t^2\sigma^2 + 2\rho_t c_t^2 dL^2 + \frac{2L^2R^2\gamma_t^2}{\rho_t} \\
&+ \left(1 - \frac{\rho_t}{2}\right))\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right]. \qquad \text{(I.35)}
\end{aligned}
$$

□

Proof of Theorem 10.5.1(3). Now using the result in Lemma I.1.1 we can characterize the convergence of the sequence of expected errors $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ to zero. To be more precise, using the result in Lemma 10.4.2 and setting $\gamma_t = 2/(t+8)$, $\rho_t = 4/(t+8)^{2/3}$ and $c_t = 2/\sqrt{d}(t+8)^{1/3}$ for any $\epsilon > 0$ to obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq \\
&\left(1 - \frac{2}{(t+8)^{2/3}}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{32\sigma^2 + 32L^2 + 2L^2R^2}{(t+8)^{4/3}}. \qquad \text{(I.36)}
\end{aligned}
$$

According to the result in Lemma I.1.1, the inequality in (I.18) implies that

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq \frac{Q_{kw}}{(t+8)^{2/3}}, \tag{I.37}$$

where

$$Q = \max\left\{4\|\nabla f(\mathbf{x}_0) - \mathbf{d}_0\|^2, 32\sigma^2 + 32L^2 + 2L^2R^2\right\}$$

Now we proceed by replacing the term $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ in (I.15) by its upper bound in (I.28) and γ_{t+1} by $2/(t+9)$ to write

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] &\leq \left(1 - \frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right] \\ &+ \frac{R\sqrt{Q_{kw}}}{(t+9)^{4/3}} + \frac{2LR^2}{(t+9)^2}.\end{aligned} \tag{I.38}$$

Note that we can write $(t+9)^2 = (t+9)^{4/3}(t+9)^{2/3} \geq (t+9)^{4/3}9^{2/3} \geq 4(t+9)^{4/3}$.Therefore,

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] &\leq \left(1 - \frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right] \\ &+ \frac{2R\sqrt{Q_{kw}} + LD^2/2}{(t+9)^{4/3}}.\end{aligned} \tag{I.39}$$

Thus, for $t \geq 0$ by induction we have,

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{Q'}{(t+9)^{1/3}} = O\left(\frac{d^0}{(t+9)^{1/3}}\right). \tag{I.40}$$

where $Q' = \max\{2(f(\mathbf{x}_0) - f(\mathbf{x}^*)), 2R\sqrt{Q_{kw}} + LR^2/2\}$. □

Proof of Theorem 10.5.2(3). Following as in (I.23), we have,

$$\begin{aligned}&\gamma_t\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[F(\mathbf{x}_t) - F(\mathbf{x}_{t+1})\right] + \gamma_t R\frac{\sqrt{2Q_{kw}}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t^2}{2} + \\ &\Rightarrow \mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[\frac{t+7}{2}F(\mathbf{x}_t) - \frac{t+8}{2}F(\mathbf{x}_{t+1}) + \frac{1}{2}F(\mathbf{x}_t)\right] + R\frac{\sqrt{2Q_{kw}}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2} \\ &\Rightarrow \sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0) - \frac{T+7}{2}F(\mathbf{x}_T) + \sum_{t=0}^{T-1}\left(\frac{1}{2}F(\mathbf{x}_t)\right] + R\frac{\sqrt{2Q_{kw}}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}\right) \\ &\Rightarrow \sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*)\right] + \sum_{t=0}^{T-1}\left(\frac{1}{2}\left(F(\mathbf{x}_t) - F(\mathbf{x}^*)\right) + R\frac{\sqrt{2Q_{kw}}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}\right) \\ &\Rightarrow \sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*) + \sum_{t=0}^{T-1}\left(\frac{Q'_{kw} + R\sqrt{2Q_{kw}}}{2(t+8)^{1/3}} + \frac{LR^2}{(t+8)}\right) \\ &\Rightarrow T\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*) + LR^2\ln(T+7) + \frac{Q'_{kw} + R\sqrt{2Q_{kw}}}{2}(T+7)^{2/3} \\ &\Rightarrow \mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \frac{7(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{2T} + \frac{LR^2\ln(T+7)}{T} + \frac{Q'_{kw} + R\sqrt{2Q_{kw}}}{2T}(T+7)^{2/3}\end{aligned} \tag{I.41}$$

□

I.4 Proofs for Non Convex Stochastic Frank Wolfe

Proof of Theorem 10.5.3. We reuse the following characterization derived earlier:

Lemma I.4.1. *Let Assumptions 10.4.3-10.5.1 hold. Given the recursion in (2.18), we have that* $\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2$ *satisfies*

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq 2\rho_t^2\sigma^2 + 4\rho_t^2 L_1^2 \\
&+ 8\rho_t^2\left(1 + \frac{s(d)}{m}\right)L_1^2 + \left(\frac{1+m}{2m}\right)\rho_t^2 c_t^2 L^2 M(\mu) \\
&+ \frac{2L^2R^2\gamma^2}{\rho_t} + \frac{\rho_t}{2m^2}c_t^2 L^2 M(\mu) \\
&+ \left(1 - \frac{\rho_t}{2}\right))\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right].
\end{aligned} \tag{I.42}$$

Now using the result in Lemma I.1.1 we can characterize the convergence of the sequence of expected errors $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ to zero. To be more precise, using the result in Lemma 10.4.2 and setting $\gamma = T^{-3/4}$, $\rho_t = 4/\left(1 + \frac{d}{m}\right)^{1/3}(t+8)^{1/2}$ and $c_t = 2\sqrt{m}/\sqrt{M(\mu)}(t+8)^{1/4}$ to obtain for all $t = 0, \cdots, T-1$,

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\
&\leq \left(1 - \frac{2}{\left(1 + \frac{d}{m}\right)^{1/3}(t+8)^{1/2}}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{32\sigma^2 + 64L_1^2 + 128\left(1 + \frac{d}{m}\right)^{1/3}L_1^2}{(t+8)} \\
&+ \frac{8L^2R^2\left(1 + \frac{d}{m}\right)^{1/3} + 416L^2}{(t+8)}.
\end{aligned} \tag{I.43}$$

Using Lemma I.1.1, we then have,

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] = O\left(\frac{(d/m)^{2/3}}{(t+9)^{1/2}}\right), \forall\ t = 0, \cdots, T-1 \tag{I.44}$$

Finally, we have,

$$\begin{aligned}
&F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \gamma_t\langle \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle \\
&+ \gamma\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{LR^2\gamma^2}{2} \\
&\leq F(\mathbf{x}_t) + \gamma\langle \mathbf{d}_t, \operatorname{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle - \mathbf{x}_t\rangle \\
&+ \gamma\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{LR^2\gamma^2}{2} \\
&\leq F(\mathbf{x}_t) + \gamma\langle \nabla F(\mathbf{x}_t), \operatorname{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle - \mathbf{x}_t\rangle \\
&+ \gamma\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \operatorname{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle\rangle + \frac{LR^2\gamma^2}{2} \\
&\leq F(\mathbf{x}_t) - \gamma\mathcal{G}(\mathbf{x}_t) + \frac{LR^2\gamma^2}{2}
\end{aligned}$$

$$
\begin{aligned}
&+\gamma\langle\nabla F(\mathbf{x}_t)-\mathbf{d}_t,\mathbf{v}_t-\underset{\mathbf{v}\in\mathcal{C}}{\operatorname{argmin}}\langle\mathbf{v},\nabla F(\mathbf{x}_t)\rangle\rangle\\
&\Rightarrow\gamma\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\mathbb{E}\left[F(\mathbf{x}_t)\right]-\mathbb{E}\left[F(\mathbf{x}_{t+1})\right]\\
&+\gamma R\mathbb{E}\left[\|\nabla F(\mathbf{x}_t)-\mathbf{d}_t\|\right]+\frac{LR^2\gamma^2}{2}\\
&\leq\mathbb{E}\left[F(\mathbf{x}_t)\right]-\mathbb{E}\left[F(\mathbf{x}_{t+1})\right]+\gamma_t R\sqrt{\mathbb{E}\left[\|\nabla F(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]}+\frac{LR^2\gamma^2}{2}\\
&\leq\mathbb{E}\left[F(\mathbf{x}_t)\right]-\mathbb{E}\left[F(\mathbf{x}_{t+1})\right]+Q_{nc}\gamma\rho_t^{1/2}R(d/m)^{1/3}+\frac{LR^2\gamma^2}{2}\\
&\Rightarrow\mathbb{E}\left[\mathcal{G}_{min}\right]T\gamma\leq\mathbb{E}\left[F(\mathbf{x}_0)\right]-\mathbb{E}\left[F(\mathbf{x}_{t+1})\right]\\
&+Q_{nc}\gamma R(d/m)^{1/3}\sum_{t=0}^{T-1}\rho_t^{1/2}+\frac{LR^2T\gamma^2}{2}\\
&\Rightarrow\mathbb{E}\left[\mathcal{G}_{min}\right]\leq\frac{\mathbb{E}\left[F(\mathbf{x}_0)\right]-\mathbb{E}\left[F(\mathbf{x}^*)\right]}{T\gamma}\\
&+\gamma Q_{nc}R(d/m)^{1/3}\frac{\sum_{t=0}^{T-1}\rho_t^{1/2}}{T\gamma}+\frac{LR^2T\gamma^2}{2T\gamma}\\
&\Rightarrow\mathbb{E}\left[\mathcal{G}_{min}\right]\leq\frac{\mathbb{E}\left[F(\mathbf{x}_0)\right]-\mathbb{E}\left[F(\mathbf{x}^*)\right]}{T^{1/4}}\\
&+\frac{Q_{nc}Rd^{1/3}}{T^{1/4}m^{1/3}}+\frac{LR^2}{2T},
\end{aligned}
\tag{I.45}
$$

where $\mathcal{G}_{min}=\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)$. □