EFFICIENT ALGORITHMS FOR DISTRIBUTED LEARNING, OPTIMIZATION AND BELIEF SYSTEMS OVER NETWORKS

BY

CÉSAR A. URIBE MENESES

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Tamer Başar, Chair
Professor Angelia Nedić, Director of Research
Assistant Professor Alex Olshevsky, Co-Director of Research
Professor Rayadurgam Srikant
Associate Professor Maxim Raginsky

ABSTRACT

A distributed system is composed of independent agents, machines, processing units, etc., where interactions between them are usually constrained by a network structure. In contrast to centralized approaches where all information and computation resources are available at a single location, agents on a distributed system can only use locally available information. The particular flexibilities induced by a distributed structure make it suitable for large-scale problems involving large quantities of data. Specifically, the increasing amount of data generated by inherently distributed systems such as social media, sensor networks, and cloud-based databases has brought considerable attention to distributed data processing techniques on several fronts of applied and theoretical machine learning, robotics, resource allocation, among many others. As a result, much effort has been put into the design of efficient distributed algorithms that take into account the communication constraints and make coordinated decisions in a fully distributed manner.

In this dissertation, we focus on the principled design and analysis of distributed algorithms for optimization, learning and belief systems over networks. Particularly, we are interested in the non-asymptotic analysis of various distributed algorithms and the explicit influence of the topology of the network they ought to be solved over.

Initially, we analyze a recently proposed model for opinion dynamics in belief systems with logic constraints. Opinion dynamics are a natural model for a distributed system and serve as an introductory topic for the further study of learning and optimization over networks. We assume there is an underlying structure of social relations, represented by a social network, and people in this social group interact by exchanging opinions about a number of truth statements. We analyze, from a graph-theoretic point of view, this belief system when a set of logic constraints relate the opinions on the several topics being discussed. We provide novel graph-theoretic conditions for convergence, explicit estimates of the convergence rate and the limiting value of the opinions for all agents in the network in terms of the topology of the social structure of the agents and the topology induced by the set of logic constraints. We derive explicit dependencies for a number of well-known graph topologies.

We then shift our attention to the distributed learning problem of cooperative inference

where a group of agents interact over a network and seek to estimate a joint parameter that best explains a set of network-wide observations using the local information only. Again, we assume there is an underlying network that defines the communication constraints between the agents and derive explicit, non-asymptotic, and geometric convergence rates for the concentration of beliefs on the optimal parameter. For the case of having a finite number of hypotheses, we propose distributed learning algorithms for time-varying undirected graphs, time-varying directed graphs and a new acceleration scheme for fixed undirected graphs. For each of the network structures, we present explicit dependencies for the worst case network topology. Furthermore, we extend these belief concentration results to hypotheses sets being a compact subset of the real numbers, for a simplified static undirected network assumption. Moreover, we present a generic distributed parameter estimation algorithm for observational models belonging to the exponential family of distributions. We further extend the distributed mean estimation from Gaussian observations to time-varying directed networks.

The graph-theoretical analysis of belief systems with logic constraints and the distributed learning for cooperative inference are specific instances of convex optimization problems where the objective function is decomposable as the sum of convex functions. Particularly, these problems assume each of the summands is held by a node on a graph and agents are oblivious to the network topology. As a final object of interest, we study the optimality of first-order distributed optimization algorithms for general convex optimization problems. We focus on understanding the fundamental limits induced by the distributed networked structure of the problem and how it compares with the hypothetical case of having centralized computations available. We show that for large classes of convex optimization problems, we can design optimal algorithms that can be executed over a network in a distributed manner while matching lower complexity bounds of their centralized counterparts with an additional iteration cost that depends on the network structure. We design optimal distributed algorithms for various convexity and smoothness properties that can be executed over arbitrary fixed, connected and undirected graphs. Furthermore, we explore the application of these distributed algorithms to the problem of distributed computation of Wasserstein barycenters of finite distributions.

Finally, we discuss some future directions of research for the design and analysis of distributed algorithms, both from theoretical and applied perspectives.

To Amalia, Ariadna, Guillermo, Dylan, Donovan and Luna.

ACKNOWLEDGMENTS

Every thesis acknowledgment starts with something like: "This wouldn't have been possible without Prof..." This one will not be any different, not to keep the cliché alive, but because nothing could be closer to reality. I cannot imagine having better advisers than Prof. Angelia Nedić and Prof. Alex Olshevsky. They provided me with the unique opportunity to have the intellectual freedom to explore several research problems. Moreover, they provided me constant and uninterrupted guidance and were extremely generous with their patience. Their advice was invaluable at every scale, from a missing comma in a draft and a loop-hole in a proof to the big picture on relevant open problems. Most importantly I am grateful for their trust. This is the first necessary condition to enjoy and complete any doctoral program.

Of course, the second necessary condition for a successful Ph.D. is to have constant support from the loved ones. *Mamá, Papá, Angelica, Mary, muchas gracias por confiar en mí y darme todo el apoyo que necesité. Ustedes son la esperanza que mantiene mi motivación siempre alta. Nunca los saqué de mi mente.* Sandra, there are no words to express the infinite value you have brought to my life. Your company is calm in the hardest times.

The third necessary condition is to have a great committee for the defense. My gratitude goes to Prof. Tamer Başar for allowing me to take part in his group meetings, being attentive to the development of my dissertation and kindly agreeing to be the chair of my defense. No other course influenced more the contents of this dissertation than Statistical Learning Theory by Prof. Maxim Raginsky. His unlimited knowledge of the literature both in English and Russian was always readily available. Finally, I am thankful to Prof. Srikant; he always has the precise personal and professional advice that can only come from experience. His kind invitation to be his TA and attend his group meetings gave me a number of new experiences that only strengthened my professional profile.

The fourth necessary condition is to have unbelievably talented colleagues, and CSL is the place to find them. Being surrounded by extremely smart people is humbling and puts things in perspective. Some people say one is the average of the five people one spends the most time with. CSL always brought this average higher. Many thanks, Philip, Thinh, Amir, Jaeho, James, Prof. Rasoul Etesami, Prof. Behrouz Touri, Dr. Soomin Lee, Dr. Wei Shi, and all the

ones I am missing. I also thank Dr. Hoi-To Wai for his friendship and welcoming attitude during my time in Arizona. Finally, I am most thankful to Prof. Alexander Gasnikov. His illuminating discussions and unparalleled understanding of mathematical programming were fundamental for the realization of the later parts of this dissertation.

Being away from home is never easy. However, no matter how cold Chambana got, the Colombian crew made me always feel like I was in the tropics. During these years, I met great Colombian and Latin-American scientists. I was always amazed at how much talent our region is able to produce in spite of all the adverse circumstances. Thanks a lot Ian & Angela, Jorge & Monica, Wladimir & Monika, Santiago & Ana, Agustin & Fernanda, Camila & Pedro, Ximena & Mauricio, Ian & Lina, Catalina & Fabian, Eliana, Daniel, Felipe, Santiago & Mariam, Rafa & Catalina, Liz, JJ, Rocio, and all others I'm unconsciously forgetting to keep this dissertation finite.

Life in Chambana would have been much sadder without volleyball. Here is where I would always be thankful for the Rec Volleyball Club. Being able to play every Wednesday allowed me to meet great players, but foremost, great people who made my winter days warm. Especially, I want to do a shout-out to the No-Name team. Thank you Pavel, Dima, Evelyn, Eric and Drake for being the most amazing undefeated team in the whole of Urbana-Champaign. It was an honor to play with you guys.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$x \in \mathbb{R}^d$	A column vector in $\mathbb{R}^d$
$\mathbf{1}_d$	A column vector of dimension d with all entries equal to 1
$\mathcal{G}(V, E)$	A static undirected graph with node set V and edge set E
$a.s.$	Almost surely, with probability 1
$\|x\|$ or $\|x\|_2$	Euclidean norm of vector x
$i.i.d.$	Independent and identically distributed
$\langle x, y \rangle$	Inner product of two vectors x and y
$\cdot_k$ or $\cdot_t$	Subindices k and t are usually reserved to iteration or time indices
$\cdot^i$ or $\cdot^j$	Superindices i and j are usually reserved to indicated values at the nodes
$\|S\|$	The cardinality of a set S
S^c	The complement of a set S
$A_{k_f:k_i}$	The cumulative product $A_{k_f:k_i} = A_{k_f} \cdot A_{k_i+1} A_{k_i}$ for all $k_f \geq k_i \geq 0$
d^i or $deg(i)$	The degree of a node i
E^*	The dual space of the finite-dimensional vector space E
$[A]_{ij}$ or a_{ij}	The entry of the matrix A at the i-th row and the j-th column
$\mathbb{E}[\cdot]$	The expectation of an event
$S_1(n)$	The finite probability simplex such that $S_1(n) = \{p \in \mathbb{R}^n_+ \mid p^T \mathbf{1} = 1\}$
$h^2(P, Q)$	The Hellinger distance between distributions P and Q
$[x]_i$ or x_i	The i-th entry of a vector x
I_d	The identity matrix of size d

$\ker(W)$	The kernel of matrix W
$A \otimes B$	The Kronecker product between objects A and B
$D_{KL}(P\|Q)$	The Kullback-Leibler divergence between distributions P and Q
$\lambda_{\max}(W)$	The largest eigenvalue of a matrix W
$d_{\max}$ and $d_{\min}$	The maximum and minimum degree of a graph
$\sigma_{\max}(W)$	The maximum eigenvalue of the matrix $W^T W$, i.e., $\lambda_{\max}(W^T W)$
$\|A\|_{E\to H}$	The norm of a linear operator $A : E \to H$ such that $\|A\|_{E\to H} = \max_{x\in E, u\in H^*}\{\langle u, Ax\rangle \mid \|x\|_E = 1, \|u\|_{H^*} = 1\}$
$\perp$	The orthogonality symbol between two vectors
$\|x\|_p$	The p-norm of vector x
$d(\mathcal{G})$	The period of a graph $\mathcal{G}$
$\mathbb{P}[\cdot]$	The probability of an event
$X \sim P$	The random variable X is distributed according to P
$\lambda_2(W)$	The second largest eigenvalue of a matrix W
$\mathbb{R}^d$	The set of d-dimensional vectors with components in $\mathbb{R}$
N_k^i	The set of neighbors of agent i at time k
$\mathbb{R}_+$	The set of nonnegative real numbers
$\mathbb{R}$	The set of real numbers
$\lambda_{\min}^+(W)$	The smallest positive eigenvalue of a matrix W
$\sigma_{\min}^+(W)$	The smallest positive eigenvalue of the matrix $W^T W$, i.e., $\lambda_{\min}^+(W^T W)$
$\text{span}(W)$	The span of matrix W
x' or x^T	The transpose of a vector x

CHAPTER 1

INTRODUCTION

Large numbers of interconnected components add to the complexity of engineering systems. Developing models and tools for the analysis of such distributed systems is necessary, not only from the engineering point of view but for effective decision-making and policy design. For example, the control of autonomous vehicles for exploration, rescue, and surveillance depends on the coordination abilities of fleets of robots; each robot should make decisions based on local information and limited communications. Power networks (e.g., the electric grid) need several generating and consuming stations to coordinate offer and demand to improve efficiency. In traffic control, the goal is to avoid jams distributively and to increase traffic flow based on limited infrastructure (e.g., roads). Economic systems need modeling, estimation, and control of markets at the micro and macroeconomic scales. Market dynamics depend on several agents influencing the system, each of which might have conflicting goals. In telecommunication networks, several stations need to communicate over non-perfect channels to optimize information transmission. The control of industrial processes requires communication and coordination between different parts of the process in hazardous environments. The modeling and control of ecological systems requires the analysis of several actors interacting with each other, subject to changing environments.

The increasing amount of data generated by recent applications of distributed systems such as social media, sensor networks, and cloud-based databases has brought considerable attention to distributed data processing, in particular the design of distributed algorithms that take into account the communication constraints and make coordinated decisions in a distributed manner [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. In a distributed system, interactions between agents are usually constrained by the network structure and agents can only use locally available information. This contrasts with centralized approaches where all information and computation resources are available at a single location [12, 13, 14, 15].

Traditional approaches for the design of distributed inference algorithms, for inherently distributed systems, assume a fusion center exists. The fusion center gathers all the information and makes centralized decisions [12, 13, 14, 15]. Nonetheless, communication constraints, limited memory and lack of physical accessibility to certain measurements hin-

der this task. Therefore, it is necessary to develop algorithmic protocols that take into account such constraints and use only locally available information.

The adoption of distributed optimization algorithms on several fronts of applied and theoretical machine learning, robotics, and resource allocation has increased the attention on such methods in recent years [16, 17, 18, 19, 20]. The particular flexibilities induced by the distributed setup make them suitable for large-scale learning problems involving large quantities of data [21, 22, 23, 24, 25]. Although many results on these themes have appeared in recent years, the study of distributed decision-making and computation traces back to classic papers from the 70s and 80s [5, 6, 7, 8, 9, 10, 11].

In [26] the authors defined society as "wise" if the influence of the most influential agents vanishes with the size of the network. This assumes there exists some balancedness in the network in terms of the agents' *centrality*. Knowledge about the topology of the network can be used to design algorithms that take the agents' connectivity into account, but this introduces additional information requirements and limits the ad-hoc nature of a distributed solution. Specifically, in evolving networks the connectivity of the agents' changes with time and thus so does their influence, introducing variability in the group confidence.

The object of study of this dissertation is twofold: on the one hand we have a convex optimization problem $\sum_{i=1}^{m} f_i(z)$ assumed to be the sum of a finite number n of convex functions, and on the other hand we have a network, modeled as a graph $\mathcal{G}(V, E)$, with $|V| = n$ nodes and a set of edges E between them that represent their ability to share information, see Fig. 1.1. The main property of this object is the assumption that each node i in the network has access to a single $f_i(z)$ only. Nonetheless, one seeks to solve the network-wide optimization problem by local interactions constrained by the network edges.

Now that we have defined the main focus of this theses we can describe the specific problems we are interested in. We study four specific problems:

1. Graph-theoretic analysis of belief systems with logic constraints.

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \|x_i - x_j\|^2 \qquad \forall i \in V.$$

2. Distributed learning with finite hypotheses sets.

$$\min_{\theta \in \Theta} \sum_{i=1}^{n} D_{KL}(P^i \| P^i_\theta) \qquad \Theta \text{ is finite.}$$

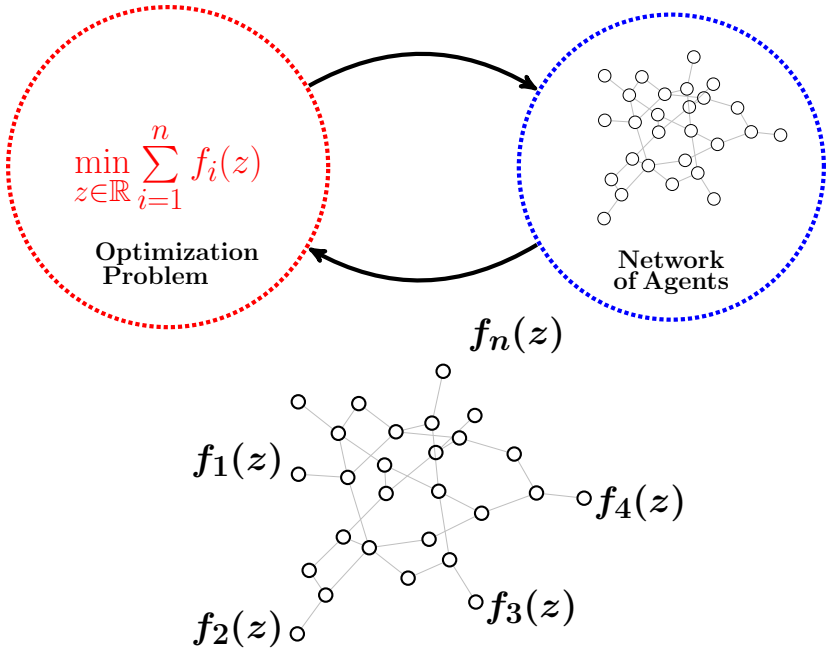


Figure 1.1: The object of study of this dissertation.

3. Distributed learning on compact hypotheses sets.

$$\min_{\theta\in\Theta\subset\mathbb{R}^d} \sum_{i=1}^{n} D_{KL}(P^i \| P^i_\theta) \qquad \Theta \text{ is compact.}$$

4. Optimal algorithms for distributed optimization.

$$\min_{z\in\mathbb{R}} \sum_{i=1}^{m} f_i(z),$$

such that

(a) Each f_i is strongly convex and smooth.

(b) Each f_i is strongly convex.

(c) Each f_i is smooth and convex.

(d) Each f_i is convex.

1.1 Motivation and Past Work

In this section, we introduce each of the problems studied in this dissertation and motivate the open problems for each case. We provide some classic references and guide the reader towards a more comprehensive literature review.

1.1.1 Opinion Dynamics in Belief Systems with Logic Constraints

The analysis and modeling of opinion dynamics spans several decades of interdisciplinary research [27, 28, 29, 30, 31, 32, 33, 34, 35]. Belief systems are modeled as a process where agents continuously update their beliefs by repeated interactions where opinions are exchanged over some social structure (e.g., social network) [10, 36]. New opinions are formed by aggregating operations weighted by the relative importance assigned by an individual to others. This simple characterization has provided tools for analyzing the long-term behaviors using systems theory. Nevertheless, the characterization has been shown insufficient to explain the existence of shared beliefs in a population [37].

Opinion formation cannot be described solely as an ideological deduction from a set of principles about the social world. Repeated social interactions and logic constraints on truth statements are consequential for the construction of belief systems as well. Recently proposed generalizations of opinion dynamic models integrate functional interdependencies among issues that coherently bound ideas and attitudes [38]. Mainly, *logic constraints* in belief systems provide a successful model for the evolution of opinions in both large-scale populations and small groups [37]. *Logic constraints* build upon the natural idea that believing a specific statement is true may depend on the belief that other statements are true as well. Nonetheless, existing algebraic tools can be too complicated to use when facing large-scale and complex networks [38]. Understanding the role of the networks involved in the structural features of a belief system is of critical importance and can have direct implications for better decision-making and policy design [39, 40, 41, 42, 37].

We seek to provide graph-theoretic answers for a model of opinion dynamics of a belief system with logic constraints. Particularly, we are interested in showing how the belief system properties depend on the social network where agents interact and the set of logic constraints that relate beliefs on different truth statements. Moreover, we search for explicit dependencies for a variety of commonly used large-scale network models.

1.1.2 Distributed (Non-Bayesian) Learning over Networks

Numerous engineered and natural systems can be modeled as a group of agents interacting (e.g., people, robots, sensors). Distributed non-Bayesian learning studies groups of agents that try to "learn" a distribution (from a parametrized family) that best explains some observed data [1, 43, 44, 45, 46, 25, 47]. Specifically, agents seek to learn this parameter in a distributed manner where each agent accesses local information without the involvement of any centralized coordination.

One traditional problem in decision-making is that of parameter estimation. Given a set of noisy observations coming from a joint distribution, one would like to estimate a parameter or distribution that minimizes a certain loss function. For example, maximum a posteriori (MAP) or minimum least squared error (MLSE) estimators fit a parameter to some model of the observations. Both MAP and MLSE estimators require some form of Bayesian posterior computation based on models that explain the observations for a given parameter. Computation of such a posteriori distributions depends on having exact models about the likelihood of the corresponding observations. This is one of the main difficulties of using Bayesian approaches in a distributed setting. A fully Bayesian approach is not possible because full knowledge of the network structure, or of other agents' likelihood models, may not be available [48, 49, 29].

In [29], the authors describe results on learning in social networks based on computing posterior distributions using Bayes' rule. That is, given some assumed prior knowledge and new observations, an agent computes a posterior based on likelihood models, see [50]. Nevertheless, a fully Bayesian approach might not be possible because full knowledge of the network structure, or other agents' likelihood models, need not be available [48, 49]. Other authors showed that non-Bayesian methods can be used in learning task as well [51, 1, 52, 43]. In this case, agents are assumed to be boundedly rational (i.e., fail to aggregate information in a fully Bayesian manner [26]). They repeatedly communicate with others and use naive approaches to aggregate information.

Several groundbreaking papers have described distributed methods to achieve global behaviors by repeatedly aggregating local information without complete knowledge of the network [1, 2, 3, 4]. For example, in distributed hypothesis testing using belief propagation, convergence and its dependence on the communication structure were shown [3]. Later, extensions to finite capacity channels, packet losses, delayed communications, and tracking were developed [53, 54]. In [2], the authors proved convergence in probability, the asymptotic normality of the distributed estimation and provided conditions under which the distributed estimation is as good as a centralized one. Later in [1], the almost sure convergence of a

non-Bayesian rule based on the arithmetic mean was shown for fixed topology graphs. Extensions to information heterogeneity and asymptotic convergence rates have been derived as well [52]. Following [1], other methods to aggregate Bayes estimates in a network have been explored. In [55], geometric means are used for fixed topologies as well. However, the consensus and learning steps are separated. The work in [56] extends the results of [1] to time-varying undirected graphs. In [43], local exponential rates of convergence for undirected gossip-like graphs are studied. The authors in [46, 45, 57, 56] proposed a non-Bayesian learning algorithm where a local Bayes' update is followed by a consensus step. In [46], convergence result for fixed graphs is provided, and large deviation convergence rates are given, proving the existence of a random time after which the beliefs will concentrate exponentially fast. In [45], similar probabilistic bounds for the rate of convergence are derived for fixed graphs, and comparisons with the centralized version of the learning rule are provided. Other variations of the non-Bayesian approach have been proposed for continuum set of hypotheses [58], weakly connected graphs [59], bisection search algorithms [60], transmission node failures [61, 62, 63] and time-varying graphs [64, 65, 66]. See [67, 68] for an extended literature review.

1.1.3 Optimal Convergence Rates in Distributed Optimization over Networks

Early algorithms for distributed optimization, such as distributed subgradient methods, were shown successful for solving optimization problems in a distributed manner over networks [69, 70, 71, 72]. Nevertheless, these algorithms are particularly slow compared with their centralized counterparts. Recently, distributed methods that achieve linear convergence rates for minimizing a sum of strongly convex and smooth (network) objective functions have been proposed. One can identify three main approaches to the study of distributed algorithms. In [73], a new method was proposed where it was shown that $O((n^2 + \sqrt{L/\mu}n)\log\varepsilon^{-1})$ iterations are required to find an ε solution to the optimization problem when the function is μ-strongly convex and L-smooth, and m is the number of nodes in the network. In [74], a new analysis technique for the convergence rate of distributed optimization algorithms via a semidefinite programming characterization was proposed. This approach provides an innovative procedure to numerically certify worst-case rates of a plethora of distributed algorithms, which can be useful to fine-tune parameters in existing algorithms based on feasibility conditions of a semidefinite program. In [75], a unifying approach was proposed, that recovers rate results from several existing algorithms such as those in [76, 77]. This

newly proposed general method is able to recover existing rates and achieves an ε precision in $O(\sqrt{L/(\mu\lambda_2)}\log\varepsilon^{-1})$ iterations, where λ_2 is the second largest eigenvalue of the interaction matrix. These results require some minimal information about the topology of the network and provide explicit statements about the dependency of the convergence rate on the problem parameters. Specifically, polynomial scalability is shown with the network parameter for particular choices of small enough step-sizes, and even uncoordinated step-sizes are allowed [78]. One particular advantage of this approach is that it can handle time-varying and directed graphs. Nevertheless, optimal dependencies on the problem parameters and tight convergence rate bounds are far less understood. A third approach was recently introduced in [79], where the first optimal algorithm for distributed optimization problems was proposed. This new method achieves an ε precision in $O(\sqrt{L/\mu}(1+\tau/\sqrt{\gamma})\log\varepsilon^{-1})$ iterations for μ-strongly convex and L-smooth problems, where τ is the diameter of the network and γ is the normalized eigengap of the interaction matrix. Even though extra information about the topology of the network is required, the work in [79] provides a coherent understanding of the optimal convergence rates and its dependencies on the communication network.

One particular area of interest is the large-scale optimal transport problems. Optimal transport distances (also known as *earth mover's distances* or *Wasserstein distances*) design an optimal plan to move "mass" from one probability distribution to another. This problem can be traced back to the early work of Monge [80] and Kantorovich [81] and has been of constant interest for allowing natural formulations to the problems of comparing, interpolating, and measuring distances of functions [82]. On the other hand, computational optimal transport has gained popularity for its applications in learning theory [83], computer vision [84], computer graphics [85], statistical inference [86], information fusion [87], and its relative complexity advantages with respect to classical methods [88]. Particularly, *large-scale* optimal transport has been of recent interest for the latest applications where large quantities of data are available and efficient algorithms are required [89, 90, 91]. Comprehensive accounts of the optimal transport problem and its computational aspects can be found in [92, 93, 94, 82].

1.2 Dissertation Structure and Contributions

As indicated earlier, this dissertation is devoted to the study of the relation between optimization problems in the form of a sum of convex functions and distributed networks. Moreover, we are particularly interested in the design of distributed algorithms that can be executed over a network where each node only requires local information and yet global

performance goals are achieved. For each of the studied problems and algorithms, we focused on non-asymptotic performance analysis by looking into their efficiency and scalability concerning the structural properties of the problem and the topology of the network where the problem needs to be solved. Next, we provide a summary of the main contributions of this dissertation.

In Chapter 2, we study how the structural properties of the social network of agents and the set of logic constraints influence the dynamics of a belief system from a graph-theoretic point of view. We describe this influence for the convergence of beliefs, the expected convergence time and the stationary value of the belief system. Informally, we answer the following three questions with graph-theoretic conditions that are easily accessible for a number of commonly used topologies in large-scale complex networks: When does a belief system converge? How long does it take converge? What does it converge to?

In Chapter 3, we consider the problem of distributed learning, where a network of agents collectively aims to agree on a hypothesis that best explains a set of distributed observations of conditionally independent random processes. We focus on the case where the number of hypotheses is finite and propose a distributed algorithm and establish consistency, as well as a nonasymptotic, explicit, and geometric convergence rate for the concentration of the beliefs around the set of optimal hypotheses. Additionally, if the agents interact over static networks, we provide an improved learning protocol with better scalability with respect to the number of nodes in the network. Also, we propose a novel belief update algorithm for distributed learning over time-varying directed graphs. Our main results state that, after a transient time, all agents will concentrate their beliefs at a network independent rate.

In Chapter 4, we revisit the problem of distributed (non-Bayesian) learning. In contrast with Chapter 3, we focus on the problem of having compact hypothesis sets. We explore a variational interpretation of the Bayesian posterior and its relation to the stochastic mirror descent algorithm to propose a new distributed learning algorithm. We show that, under appropriate assumptions, the beliefs generated by the proposed algorithm concentrate around the true parameter exponentially fast. We provide explicit non-asymptotic bounds for the convergence rate. Moreover, we develop explicit and computationally efficient algorithms for observation models in the exponential families. The algorithm is expressed as explicit updates on the parameters of the conjugate distribution of the observational model (i.e., means and precision for Gaussian beliefs). As an application example, we present a distributed algorithm for the problem of parameter estimation with Gaussian noise for the general case of time-varying directed graphs. We show a convergence rate of $O(1/k)$ with the constant term depending on the number of agents and the topology of the network.

In Chapter 5, we study the optimal convergence rates for distributed convex optimization

problems over networks, where the objective is to minimize the sum $\sum_{i=1}^{n} f_i(z)$ of local functions of the nodes in the network. We provide optimal complexity bounds for four different cases: the case when each function f_i is strongly convex and smooth, the cases when it is either strongly convex or smooth, and the case when it is convex but neither strongly convex nor smooth. Our approach is based on the dual of an appropriately formulated primal problem, which includes the underlying static graph that models the communication restrictions. Our results show distributed algorithms that achieve the same optimal rates as their centralized counterparts (up to constant and logarithmic factors), with an additional cost related to the spectral gap of the interaction matrix that captures the local communications of the nodes in the network. As an application example, we propose a new class-optimal algorithm for the distributed computation of Wasserstein barycenters over networks. Assuming that each node in a graph has a probability distribution, we prove that every node is able to reach the barycenter of all distributions held in the network by using local interactions compliant with the topology of the graph. We show the minimum number of communication rounds required for the proposed method to achieve arbitrary relative precision both in the optimality of the solution and the consensus among all agents for undirected fixed networks.

1.3 Mathematical Preliminaries

1.3.1 Networks and Graph Theory

We model the communication structure that defines the ability of the group of agents to exchange information between them as a graph. Particularly, throughout this dissertation *we will assume the number of agents n remains fixed*, and the interactions between them are enabled by the edges of a graph $\mathcal{G}(V, E)$, where $V = \{1, 2, \cdots, n\}$ and $E \in V \times V$ is a set of directed edges such that an ordered pair $(j, i) \in E$ if an agent j can communicate or share information to agent i. In the general case, we will denote this as a *directed* graph. A path $\mathbf{P}$ of $\mathcal{G}$ is a finite sequence $\{p_i\}_{i=0}^{l}$ such that $(p_i, p_{i+1}) \in E$ for $0 \leq i \leq l-1$. Moreover, define $n(\mathbf{P})$ as the number of edges in the path $\mathbf{P}$. A cycle $\mathbf{C}$ of a graph $\mathcal{G}$ is a path $\mathbf{P}$ such that $p_0 = p_l$, i.e., the start and end nodes of the path are the same. We denote the *period* of a directed graph as $d(\mathcal{G})$, and define it as the greatest common divisor of the length of all cycles in the graph $\mathcal{G}$. If all edges in the network are bidirectional, we will refer to the graph as *undirected*. Figure 1.2 shows some examples of common undirected graph topologies.

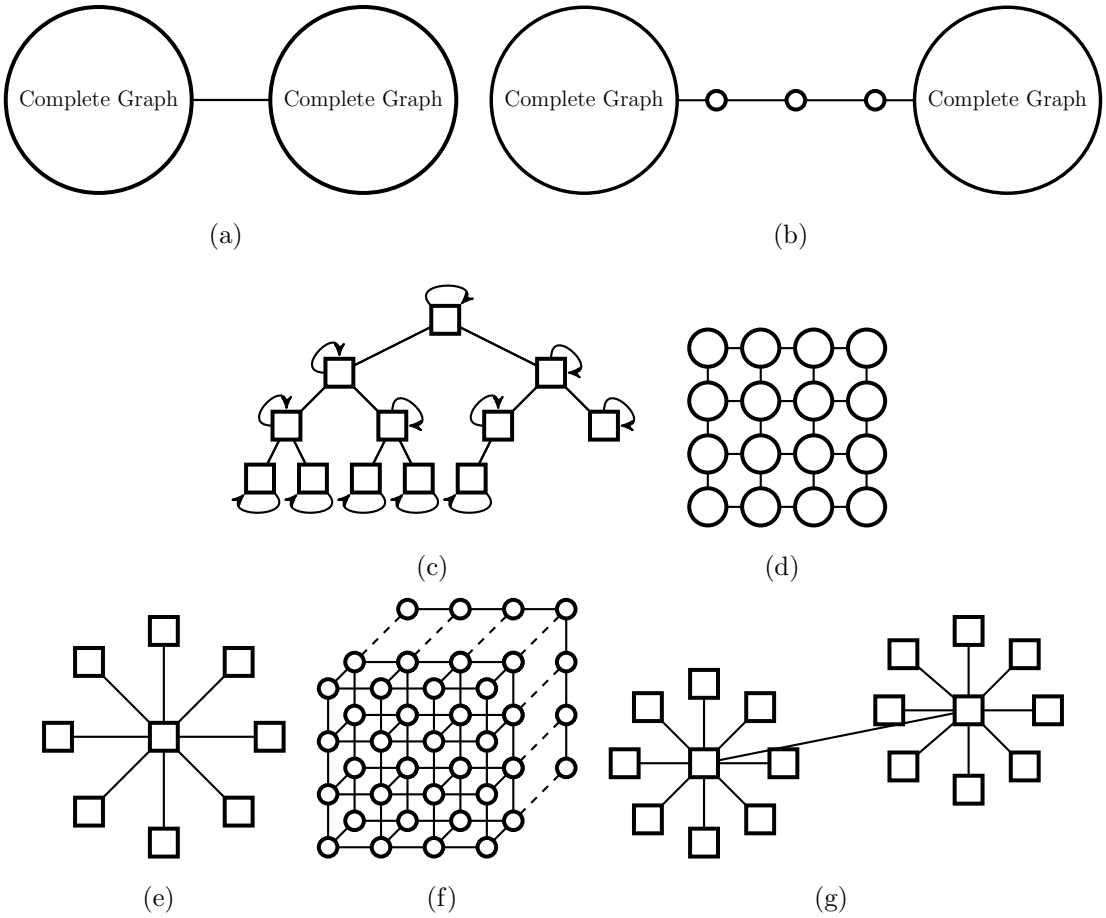


Figure 1.2: Examples of common graphs. (a) Dumbbell graph, two complete graphs connected by an edge. (b) Bolas graph, two complete graphs connected by a path. (c) Complete binary tree. (d) 2-d grid or lattice. (e) Star graph. (f) 3-d grid. (g) Two star graph connected on their centers.

Even though we assume the set of nodes in the graph remains constant, we might allow for the edges to change with time. In this scenario, we will refer to the graph as a *time-varying* graph and we will define a particular graph at an instant k as $\mathcal{G}_k(V, E_k)$. Moreover, we denote the graph sequence a $\{\mathcal{G}_k\}$.

Next, we provide three useful definitions regarding the connectivity of a graph, or a sequence of graphs, for the cases when the edges are directed, undirected, or changing with time.

Definition 1. *An undirected and static graph is called connected, if there is a path between any pair of nodes or vertices.*

Definition 2. *A directed and static graph is called:*

- *Weakly connected if by replacing all directed edges by undirected ones creates a connected graph.*
- *Connected if it contains a directed path, for any two pair of nodes $i, j \in V$, from i to j or from j to i.*
- *Strongly connected if it contains a directed path, for any two pair of nodes $i, j \in V$, from i to j and from j to i.*

Definition 3. *A sequence of directed and time-varying graph is called B-strongly connected if there is an integer $B \geq 1$ such that the graph $\left\{V, \bigcup_{i=kB}^{(k+1)B-1} E_i\right\}$ is strongly connected for all $k \geq 0$.*

We define the Laplacian matrix $L \in \mathbb{R}^{n\times n}$ of the static directed graph $\mathcal{G}$ as a squared matrix whose elements are defined as

$$[L]_{ij} = \begin{cases} -1, & \text{if } (j,i) \in E, \\ \deg(i), & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

where $\deg(i)$.

In addition, we will define weighted adjacency matrix $A \in \mathbb{R}^{n\times n}$ associated with a graph $\mathcal{G}$ as a squared matrix such that $[A]_{ij} \neq 0$ if $(j,i) \in E$ and $[A]_{ij} = 0$ if $(j,i) \notin E$. That is, we assume each of the edges in the graph gets assigned a weight. Particularly, we will use positive matrices A where every element is nonnegative. We will say a matrix A is *row stochastic* or simply *stochastic* if $[A]_{ij} \geq 0$ and $\sum_{j=1}^{n}[A]_{ij} = 1$ for all $i \in V$. Moreover, we will say a matrix A is *column stochastic* if $[A]_{ij} \geq 0$ and $\sum_{i=1}^{n}[A]_{ij} = 1$ for all $j \in V$. Finally, a matrix A is doubly stochastic if it is row stochastic and column stochastic.

There are several ways to construct a set of stochastic weight matrices. If the graph is undirected one can construct row stochastic or doubly stochastic weight matrices from undirected local interactions. For example, one can construct doubly stochastic weight matrices by considering a lazy Metropolis (stochastic) matrix of the form $\bar{A}_k = \frac{1}{2}I_n + \frac{1}{2}\hat{A}_k$, where I_n is the identity matrix and $\hat{A}_k$ is a stochastic matrix whose off-diagonal entries satisfy

$$[\hat{A}_k]_{ij} = \begin{cases} \frac{1}{\max\left\{d_k^i+1, d_k^j+1\right\}} & \text{if } (i,j) \in E_k, \\ 0 & \text{if } (i,j) \notin E_k, \end{cases}$$

where d_k^i is the degree (the number of neighbors) of node i at time k. Note that the lazy Metropolis weights require undirected communications since each weight $[\hat{A}_k]_{ij}$ depends on the degree of both agent i and agent j.

Next, we present a series of assumptions for different cases of the network connectivity and directedness. We will use different assumptions for time-varying directed graphs, time-varying undirected graphs and fixed graphs.

Assumption 1. *The graph sequence $\{\mathcal{G}_k\}$ and the matrix sequence $\{A_k\}$ are such that:*

(a) A_k is doubly-stochastic with $[A_k]_{ij} > 0$ if $(i,j) \in E_k$.

(b) If $(i,j) \notin E_k$ for some $i \neq j$ then $A_{ij} = 0$.

(c) A_k has positive diagonal entries, $[A_k]_{ii} > 0$ for all $i = 1, \ldots, n$.

(d) If $[A_k]_{ij} > 0$, then $[A_k]_{ij} \geq \eta$ for some positive constant η.

(e) $\{\mathcal{G}_k\}$ is B-strongly connected.

Assumption 1(a) and Assumption 1(b) characterize the communication between agents. If two agents can exchange information at a certain time instant k, the underlying communication graph will have an edge between the corresponding nodes. This also implies a positive weighting of the information shared. The graph sequence $\{\mathcal{G}_k\}$ and the matrix sequence $\{A_k\}$ define a corresponding inhomogeneous Markov chain with transition probabilities A_k. Assumption 1(c) guarantees the aperiodicity of this Markov chain. Additionally, Assumptions 1(d) and 1(e) guarantee that this Markov chain is ergodic by ensuring there is sufficient connectivity and that the entries of A_k do not vanish. Assumption 1 is common in distributed optimization and consensus literature [69, 72]. It guarantees convergence of the associated Markov chain and defines bounds on relevant eigenvalues in terms of the number of agents.

Assumption 2. *The graph $\mathcal{G}$ and matrix A are such that:*

(a) A is doubly-stochastic with $[A]_{ij} = a_{ij} > 0$ for $i \neq j$ if and only if $(i,j) \in E$.

(b) A has positive diagonal entries, $a_{ii} > 0$ for all $i \in V$.

(c) The graph $\mathcal{G}$ is connected.

Analogous to Assumption 1, we use the following assumption when the interaction between the agents happens over static graphs.

Assumption 3. *The graph sequence $\{\mathcal{G}_k\}$ is static (i.e. $\mathcal{G}_k = \mathcal{G}$ for all k) and undirected and the weight matrix $\bar{A}$ is a lazy Metropolis matrix, defined by*

$$\bar{A} = \frac{1}{2} I_n + \frac{1}{2} \hat{A},$$

where $\hat{A}$ is the Metropolis matrix, which is the unique stochastic matrix whose off-diagonal entries satisfy

$$\hat{A}_{ij} = \begin{cases} \frac{1}{\max\{deg(i)+1, deg(j)+1\}} & \text{if } (i,j) \in E, \\ 0 & \text{if } (i,j) \notin E. \end{cases}$$

1.3.2 Lemmas for Left Product of Weighted Adjacency Matrices

One of the main theoretical tools we are going to exploit in the analysis of distributed algorithms over networks is the left product of stochastic matrices. Next, we present a number of auxiliary lemmas that will allow us to analyze the convergence and convergence rate of distributed algorithms. For a more comprehensive account of this results see [95].

First, we recall few results from [72] about the convergence of a product of doubly stochastic matrices.

Lemma 1. *[72, 69] Under Assumption 1 on a matrix sequence $\{A_k\}$, we have*

$$\left| [A_{k:t}]_{ij} - \frac{1}{n} \right| \leq \sqrt{2} \lambda^{k-t} \quad \forall\ k \geq t \geq 0,$$

where $\lambda \in (0,1)$ is given by:

$$\lambda = \left(1 - \frac{\eta}{4n^2}\right)^{\frac{1}{B}}.$$

If each A_k is the lazy Metropolis matrix associated with $\mathcal{G}_k$ and $B = 1$, then

$$\lambda = 1 - \frac{1}{\mathcal{O}(n^2)}.$$

.

Proof. The proof may be found in [72], except the bounds on λ for the lazy Metropolis chains which may be found in [96]. □

Lemma 2. *[Corollary 2.a in [72]] Let the graph sequence $\{\mathcal{G}_k\}$, with $\mathcal{G}_k = (E_k, V)$ be uniformly strongly connected. Then, there is a sequence $\{\boldsymbol{\phi}_k\}$ of stochastic vectors such that*

$$|\,[A_{k:t}]_{ij} - \phi_k^i| \leq C\lambda^{k-t} \qquad \textit{for all} \;\; k \geq t \geq 0.$$

The constants C, δ and λ satisfy the following relations:
(1) For general B-strongly-connected graph sequences $\{\mathcal{G}_k\}$,

$$C = 4, \quad \lambda = \left(1 - \frac{1}{n^{nB}}\right)^{\frac{1}{B}}, \quad \delta \geq \frac{1}{n^{nB}}.$$

(2) If every graph G_k is regular with $B = 1$,

$$C = \sqrt{2}, \quad \lambda = \left(1 - \frac{1}{4n^3}\right)^{\frac{1}{B}}, \quad \delta = 1,$$

and $\{A_k\}$ is a sequence of matrices where A_k is a stochastic matrix such that

$$[A_k]_{ij} = \begin{cases} \frac{1}{d_k^j} & \textit{if } (j,i) \in E_k, \\ 0 & \textit{otherwise.} \end{cases}$$

.

Lemma 3. *[Corollary 2.b in [72]] Let the graph sequence $\{\mathcal{G}_k\}$ satisfy the B-strong connectivity assumption. Define*

$$\delta \triangleq \inf_{k \geq 0} \left(\min_{1 \leq i \leq n} [A_{k:0}\mathbf{1}_n]_i \right). \tag{1.1}$$

Then, $\delta \geq 1/n^{nB}$, and if all $\mathcal{G}_k$ with $B = 1$ are regular, then $\delta = 1$. Furthermore, the sequence $\boldsymbol{\phi_k}$ from Lemma 2 satisfies $\phi_k^j \geq \delta/n$ for all $k \geq 0, j = 1, \ldots, n$.

The next lemma is an extension of Lemma 2 in [45] to the case of time-varying graphs. It provides a technical result that will help us later in the computation of the non-asymptotic convergence rate for the distributed learning algorithms.

Lemma 4. *Let Assumption 1 hold for a matrix sequence $\{A_k\}$. Then for all i,*

$$\sum_{t=1}^{k} \sum_{j=1}^{n} \left| [A_{k:t}]_{ij} - \frac{1}{n} \right| \leq \frac{4 \log n}{1 - \lambda},$$

where $\lambda = 1 - \eta/4n^2$, and if every A_k is a lazy Metropolis matrix then $\lambda = 1 - 1/\mathcal{O}(n^2)$.

Proof. In [45], the authors assume the weight matrix is static and diagonalizable, then they use the following inequality from [97]:

$$\|\boldsymbol{e}_j' A^k - \pi'\|_1 \leq n\lambda_2(A)^k,$$

where $\boldsymbol{e}_j$ is a vector with its j-th entry equal to one and zero otherwise, π is the stationary distribution of the Markov chain with transition matrix A and $\lambda_2(A)$ is the second largest eigenvalue of the matrix A.

For time-varying graphs, one can use the inequality in Lemma 1 instead. The remainder of the proof remains the same as in [45]. □

Finally, we will state an enabling theorem presented in [96], which presents a distributed consensus protocol that achieves a consensus with linear growth in the number of agents.

Theorem 5. *[96] Suppose each node i in a fixed undirected connected graph updates its variable x_k^i at each time instant $k \geq 2$ as follows:*

$$y_{k+1}^i = x_k^i + \frac{1}{2}\sum_{j\in N_i} \frac{x_k^j - x_k^i}{\max\{d^i+1, d^j+1\}}, \tag{1.2a}$$

$$x_{k+1}^i = y_{k+1}^i + \left(1 - \frac{2}{9U+1}\right)\left(y_{k+1}^i - y_k^i\right), \tag{1.2b}$$

where N_i is the set of neighbors of agent i and d^i is its corresponding degree. Then, if $U \geq n$ we have that

$$\|\boldsymbol{y}_k - \bar{x}\mathbf{1}\|_2^2 \leq 2\left(1 - \frac{1}{9U}\right)^{k-1} \|\boldsymbol{y}_1 - \bar{x}\mathbf{1}\|_2^2 \quad \forall k \geq 1, \tag{1.3}$$

where $[\boldsymbol{y}_k]_i = y_k^i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_1^i$, and the process is initialized with $y_1^i = x_1^i$.

1.3.3 Random Walks, Mixing and Markov chains

Consider a finite graph $\mathcal{G} = (V, E)$ composed of V nodes with a set of edges E and a compliant associated row-stochastic matrix A. A random walk on the graph $\mathcal{G}$ is the event of a token moving from one node to another according to some probability distribution. These dynamics are captured by a Markov chain $X = (X_k)_0^\infty$ such that $\mathbb{P}\{X_{k+1} = y | X_k = x\} = P(x, y)$. This Markov chain is called *ergodic* if it is irreducible and aperiodic. For an ergodic Markov chain, there exists a unique stationary distribution π, which describes the probability that

a random walk visits a particular node in the graph as the time goes to infinity, that is $\mathbb{P}\{X_k = j\} \to \pi_j$ as $k \to \infty$. The stationary distribution is invariant to the transition matrix, that is $\pi' P = \pi'$. It follows immediately that its convergence reduces to analyzing powers of P (Theorem 4.9 in Levin et al.[98]).

Now, define the distance to stationarity as

$$d(k) = \max_{x \in \Omega} \|P^k(x, \cdot) - \pi\|_{TV}.$$

Moreover, define the mixing time of the Markov chain as

$$t_{\text{mix}}(\epsilon) = \min_k \{k : d(k) \leq \epsilon\},$$

and we say the Markov chain is rapid mixing if $t_{\text{mix}}(\epsilon) = \text{poly}(\log n, \log \frac{1}{\epsilon})$. Finally, it holds that

$$\frac{\lambda_2}{2(1-\lambda_2)} \log\left(\frac{1}{2\epsilon}\right) \leq t_{\text{mix}}(\epsilon) \leq \frac{\log n + \log(1/\epsilon)}{1-\lambda_2}, \tag{1.4}$$

where λ_2 is the second largest left-eigenvalue of the transition matrix P [99].

Table 1.1 shows estimates for the dependency of the mixing time of a random walk on a graph for several common well-studied topologies and the number of nodes in the network.

1.3.4 The Coupling Method

Consider two independent Markov chains $X = (X_k)_0^\infty$ and $Y = (Y_k)_0^\infty$, with the same transition matrix P. Then, define the *coupling time* K as the smallest k such that $X_k = Y_k$, that is, $K = \min_{k \geq 0}\{X_k = Y_k\}$. Note that K is a random variable and it depends on P as well as the initial distributions of the processes X_k and Y_k. Finally, define the quantity L_P as the maximum expected coupling time of a Markov chain with transition matrix P over all possible initial distributions of the processes X_k and Y_k, then

$$L_P = \max_{u,v} \mathbb{E}[K] \qquad \text{where} \qquad X_0 = u \text{ and } Y_0 = v.$$

In words, this L_P is the maximum expected time it takes for two random walks, with the same transition matrix and arbitrary initial states, to intersect. If we assume X starts from a distribution π, and Y from some other arbitrary stochastic vector v and we *couple* the

Table 1.1: Maximum expected convergence time for a random walk on networks with n nodes.

Network Topology	**Mixing Time**
Complete	$O(n \log 1/\epsilon)$
Cycle	$O(n^2 \log 1/\epsilon)$
Path [100]	$O(n^2 \log 1/\epsilon)$
Dumbbell Graph [101]	$O(n^2 \log 1/\epsilon)$
Complete Binary Tree [102, 103, 98]-Section 5.3.4	$O(n \log 1/\epsilon)$
k-d Cube with Loops [104]	$O((1 - 1/k) \log 1/\epsilon)$
k-d Hypercube $\{0,1\}^k$ [98]-Section 5.3.3	$O(k \log k \log 1/\epsilon)$
Lovasz Graph $\mathcal{C}_n^k$ [104]	$O((1 - 1/(kn^2)) \log 1/\epsilon)$
2-d Grid [105, 106]	$O(n \log n \log 1/\epsilon)$
Star Graph [107]	$O(n \log 1/\epsilon)$
3-d Grid [105, 106]	$O(n^{2/3} \log n \log 1/\epsilon)$
Two Joined Star Graphs	$O(n \log 1/\epsilon)$
k-d Grid [105, 106]	$O(k^2 n^{2/k} \log n \log 1/\epsilon)$
2-d Torus [108]	$O(n^2 \log 1/\epsilon)$
3-d Torus [108]	$O(n^2 \log 1/\epsilon)$
k-d Torus [108]	$O(n^2 k \log k) \log 1/\epsilon)$
Lollipop [109]	$O(n^3 \log 1/\epsilon)$
Barbell [109]	$O(n^3 \log 1/\epsilon)$
Eulerian: d-degree and expansion [110]	$O(\lvert E\rvert^2 \log 1/\epsilon)$
Lazy Eulerian with degree d-degree [111]	$O(n\lvert E\rvert \log 1/\epsilon)$
Eulerian: d-degree, max-degree weights and expansion [110]	$O(n^2 d \log 1/\epsilon)$
Lamplighter on k-Hypercube [108]	$O(k2^k \log 1/\epsilon)$
Lamplighter on (k, n)-Torus [108]	$O(kn^k \log 1/\epsilon)$
Bolas Graph [112]	$O(n^3 \log 1/\epsilon)$
Geometric Random Graph: $\mathcal{G}^d(n, r)$ [113]	$O(r^{-2} \log n \log 1/\epsilon)$
Geometric Random Graph: $\mathcal{G}^2(n, \Omega(\text{polylog}(n)))$ [114]	$O(\text{polylog}(n) \log 1/\epsilon)$
Erdős-Rényi: $\mathcal{G}(n, c/n)$, $c > 1$ [115, 116]	$O(\log^2 n \log 1/\epsilon)$
Erdős-Rényi: $\mathcal{G}(n, (1+\delta)/n)$, $\delta^3 n \to \infty$ [117, 118]	$O((1/\delta^3) \log^2(\delta^3 n) \log 1/\epsilon)$
Erdős-Rényi: $\mathcal{G}(n, 1/n)$ [119]	$O(n \log 1/\epsilon)$
Newman-Watts (small-world) Graph [120]	$O(\log^2 n \log 1/\epsilon)$
Expander Graph [121]	$O(n^2 \log 1/\epsilon)$
Exponential Random Graph: High temperature [122]	$O(n^2 \log n \log 1/\epsilon)$
Exponential Random Graph: Low temperature [122]	$O(\exp(n) \log 1/\epsilon)$
Any Connected Undirected Graph [96]	$O(n^2 \log 1/\epsilon)$
Any Connected Graph	$O(\lvert E\rvert \text{diam}(\mathcal{G}) \log 1/\epsilon)$

processes Y and X by defining a new process W such that

$$W_k = \begin{cases} Y_k, & \text{if } k < K, \\ X_k, & \text{if } k \geq K, \end{cases}$$

then

$$\|v'P^k - \pi\|_1 \leq \max_v \mathbb{P}_v \{K > k\}$$

and by the Markov inequality

$$\|v'P^k - \pi\|_1 \leq \frac{\max_v \mathbb{E}[K]}{k}.$$

Thus, after $T = O(L_P \log 1/\epsilon)$ steps, $\|v^T P^T - \pi\|_1 \leq \epsilon$, for any v.

1.3.5 Some Basic Notions on Convex Analysis

In this subsection, we will present a sequence of basic definitions from convex analysis. For a comprehensive account of definitions and results of convex analysis see [123].

Definition 4 (Definition 1.2.1 in [123])**.** *A subset C of $\mathbb{R}^d$ is called convex if*

$$\alpha x + (1-\alpha)y \in C, \qquad \forall x, y \in C, \forall \alpha \in [0,1].$$

Definition 5 (Definition 1.2.2 in [123])**.** *Let C be a convex subset of $\mathbb{R}^d$. A function $f : C \to \mathbb{R}$ is called convex if*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \qquad \forall x, y \in C, \forall \alpha \in [0,1].$$

Definition 6. *Let $f : \mathbb{R}^d \to R$ be a convex function and X be a bounded set in $\mathbb{R}^d$. We say f is Lipschitz continuous over X with constant L, or simply L-Lipschitz over X, if*

$$\|f(x) - f(y)\| \leq L\|x - y\|, \qquad \forall x, y \in X.$$

Definition 7. *We will refer to a function $f(\cdot)$ as μ-strongly convex with $\mu > 0$, if for any x, y it holds that*

$$f(y) \geq f(x) + \left\langle \tilde{\nabla} f(x), y - x \right\rangle + \frac{\mu}{2}\|x - y\|_2^2,$$

where $\tilde{\nabla} f(x)$ is any subgradient of $f(\cdot)$ at x.

Definition 8. *We will refer to a function $f(\cdot)$ as having L-Lipschitz continuous gradients*

(or L-smooth), if it is differentiable and for any x and y it holds that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

1.3.6 Additional Definitions

McDiarmid's Inequality

In the proof of some of the non-asymptotic converge rates bounds we will use McDiarmid's inequality [124], which provides bounds for the concentration of functions of random variables. This inequality allows us to show bounds on the probability that the beliefs exceed a given value ϵ. For completeness, next, we state the McDiarmid's inequality.

Theorem 6. *(McDiarmid's inequality [124]) Let $X_1, \ldots, X_k$ be a sequence of independent random variables with $X_t \in \mathcal{X}$ for $1 \leq t \leq k$. Further, let $g : \mathcal{X}^k \to \mathbb{R}$ be a function of bounded differences, i.e., for all $1 \leq t \leq k$,*

$$\sup_{x_t \in \mathcal{X}} g\left(\ldots, x_t, \ldots\right) - \inf_{x_t \in \mathcal{X}} g\left(\ldots, x_t, \ldots\right) \leq c_t,$$

then for any $\epsilon > 0$ and all $k \geq 1$,

$$\mathbb{P}\left(g(\{X_t\}_{t=1}^k) - \mathbb{E}[g(\{X_t\}_{t=1}^k)] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^k c_t^2}\right).$$

Distances between Probability Distributions

Next, we provide three definitions of the most common "distance" functions between probability distributions.

Definition 9. *The squared Hellinger distance between two probability distributions P and Q is given by*

$$h^2\left(P, Q\right) = \frac{1}{2}\int\left(\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}}\right)^2 d\lambda,$$

where P and Q are dominated by λ. Moreover, the Hellinger distance satisfies the property that $0 \leq h(P, Q) \leq 1$.

Definition 10. *If P and Q are probability measures over a set X, and P is absolutely continuous with respect to Q, then the Kullback-Leibler divergence from Q to P is defined as*

$$D_{KL}(P\|Q) = \int_X \log \frac{dP}{dQ} dP,$$

where dP/dQ is the Radon-Nikodym derivative of P with respect to Q.

Definition 11. *The total variation distance between two probability measures P and Q on a sigma-algebra $\mathcal{F}$ of subsets of the sample space Ω is defined as*

$$\|P - Q\|_{TV} = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|.$$

The Kronecker Product

In the next definition, we recall some basic properties of the Kronecker product and the corresponding Kronecker product of two graphs.

Definition 12. *[125] Let A be a $m \times n$ matrix, and C be a $p \times q$ matrix, the **Kronecker product** $A \otimes C$ is the $mp \times nq$ matrix defined as:*

$$A \otimes C = \begin{bmatrix} a_{11}C & \dots & a_{1n}C \\ \vdots & \ddots & \vdots \\ a_{m1}C & \dots & a_{mn}C \end{bmatrix}$$

or explicitly

$$A \otimes C = \begin{bmatrix} a_{11}\begin{bmatrix} c_{11} & \dots & c_{1q} \\ \vdots & \ddots & \vdots \\ c_{p1} & \dots & c_{pq} \end{bmatrix} & \dots & a_{1n}\begin{bmatrix} c_{11} & \dots & c_{1q} \\ \vdots & \ddots & \vdots \\ c_{p1} & \dots & c_{pq} \end{bmatrix} \\ \vdots & \ddots & \vdots \\ a_{m1}\begin{bmatrix} c_{11} & \dots & c_{1q} \\ \vdots & \ddots & \vdots \\ c_{p1} & \dots & c_{pq} \end{bmatrix} & \dots & a_{mn}\begin{bmatrix} c_{11} & \dots & c_{1q} \\ \vdots & \ddots & \vdots \\ c_{p1} & \dots & c_{pq} \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}c_{11} & \dots & a_{11}c_{1q} & \dots & a_{1n}c_{11} & \dots & a_{1n}c_{1q} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ a_{11}c_{p1} & \dots & a_{11}c_{pq} & \dots & a_{1n}c_{p1} & \dots & a_{1n}c_{pq} \\ \vdots & & \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{m1}c_{11} & \dots & a_{m1}c_{1q} & \dots & a_{mn}c_{11} & \dots & a_{mn}c_{1q} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ a_{m1}c_{p1} & \dots & a_{m1}c_{pq} & \dots & a_{mn}c_{p1} & \dots & a_{mn}c_{pq} \end{bmatrix}.$$

Moreover, the following properties hold:

1. *Bilinearity and associativity: for matrices A, B and C, and a scalar k, it holds:*

$$\begin{aligned} A \otimes (B + C) &= A \times B + A \otimes C \\ (A + B) \times C &= A \otimes C + B \otimes C \\ (kA) \otimes C &= A \otimes (kB) = k(A \otimes B) \\ (A \otimes B) \otimes C &= A \otimes (B \otimes C). \end{aligned}$$

2. *Non-Commutative: In general $A \times B \neq B \otimes A$. However, there exists commutation matrices P and Q such that:*

$$A \otimes B = P(B \otimes A)Q,$$

   and if A and B are square matrices then $P = Q'$.

3. *Mixed-product property: for matrices A, B, C and D:*

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD).$$

Additionally, we will define the Kronecker product of graphs as follows. The Kronecker (also known as categorical, direct, cardinal, relational, tensor, weak direct or conjunction) product $\mathcal{G} = \mathcal{G}_1 \otimes \mathcal{G}_2$ of two graphs $\mathcal{G}_1 = (V_1, E_1)$ and $\mathcal{G}_2 = (V_1, E_1)$ is a graph $\mathcal{G} = (V, E)$ where $V = V_1 \times V_2$ and $|V| = |V_1||V_2|$; and $(u, u') \to (v, v') \in E$ if and only if $u \to v \in E_1$ and $u' \to v' \in E_2$. Moreover, the adjacency matrix of the graph $\mathcal{G}$ is the Kronecker product of the adjacency matrices of $\mathcal{G}_1$ and $\mathcal{G}_2$.

CHAPTER 2

GRAPH-THEORETIC ANALYSIS OF BELIEF SYSTEMS UNDER LOGIC CONSTRAINTS

In this chapter, we study how the structural properties of the social network of agents and the set of logic constraints influence the dynamics of a belief system from a *graph-theoretic* point of view. We describe this influence for the convergence of beliefs, the expected convergence time and the stationary value of the belief system. Informally, we answer the following three questions with graph-theoretic conditions that are easily accessible for a number of commonly used topologies in large-scale complex networks:

1. When does a belief system converge?
2. How long does it take for a belief system to converge?
3. Where does a belief system converge?

2.1 Problem Formulation

Friedkin et al.[37, 38] describe a belief system with logic constraints as a group of n agents that periodically exchange and update their opinions about a set of m different truth statements with logical dependencies among them. After each social interaction, the agents use shared opinions as well as underlying logical dependencies among the opinions to update their beliefs. The agents exchange their opinions by interacting over a social network captured by a graph $\mathcal{G} = (V, E)$, where V is the set of agents, and E is a set of edges. A directed edge towards an agent indicates that it receives the opinion of another agent, i.e., the directed flow of information. Analogously, the logical dependencies among the truth statements are modeled by a graph $\mathcal{T} = (W, D)$, where an edge between two statements exists if the belief in one statement affects belief in the other.

The generalized dynamics of a belief system are defined as follows. First, every agent aggregates its opinions on every truth statement according to the imposed logic constraints (i.e., modifying the opinions to take into account the dependencies on the other truth statements). Second, the agents share their opinions over a social network, where the opinions

are aggregated again to take into account those coming from the neighboring agents (i.e., social interactions). Finally, a new opinion is formed as a combination of the most recent aggregation and the initial opinion, which models adversity to deviate from the initial beliefs or stubbornness. The opinion of an agent on a specific statement being true or false is modeled by a scalar value between zero and one. A value of zero indicates that the given agent strongly believes a specific statement is false, whereas a value of one indicates that the agent believes the statement is true. Similarly, a value of 0.5 indicates the maximal uncertainty about a statement. The aggregation steps consist of weighted (convex) combinations of the available values, where the weights represent the relative influence. This model is described in the following equations (2.1) for an arbitrary agent $i \in V$ and an arbitrary statement $u \in W$:

$$\hat{x}_k^i(u) = \sum_{v=1}^{m} C_{uv} x_k^i(v) \qquad \text{(Aggregation by logic constraints)} \tag{2.1a}$$

$$\bar{x}_k^i(u) = \sum_{j=1}^{n} A_{ij} \hat{x}_k^j(u) \qquad \text{(Aggregation by social network)} \tag{2.1b}$$

$$x_{k+1}^i(u) = \lambda^i \bar{x}_k^i(u) + (1 - \lambda^i) x_0^i(u) \qquad \text{(Influence of initial beliefs)} \tag{2.1c}$$

where $0 \leq x_k^i(u) \leq 1$ represents the opinion of an agent i at time k on a certain statement u, while $\hat{x}_k^i(u)$ and $\bar{x}_k^i(u)$ are the intermediate aggregation steps. Specifically, the intermediate aggregated opinion $\hat{x}_k^i(u)$ of agent i on statement u is formed by using the opinions of the same agent about the other statements v. The parameters $0 \leq C_{uv} \leq 1$ are compliant with the graph $\mathcal{T}$ that models the logic constraints in the sense that C_{uv} is nonzero if the statement u depends on statement v, and otherwise $C_{uv} = 0$. These parameters represent the strength of the logic constraints, i.e., the influence that an opinion on a statement has on the opinion on other statements.

Subsequently, the intermediate aggregated opinion $\bar{x}_k^i(u)$ of agent i on statement u is formed by combining all the intermediate opinions $\bar{x}_k^i(u)$ of neighboring agents j. In this update, the parameters $0 \leq A_{ij} \leq 1$ represent the weights that an agent i assigns to the information coming from its neighbor j, for example A_{13} is how agent 1 weights the opinions shared by agent 3. These parameters are compliant with the network $\mathcal{G}$ in the sense that if there is an incoming edge to agent i from agent j in the graph, then the corresponding weight A_{ij} is nonzero.

The last update in Eq. (2.1) indicates that, at time $k+1$, the new opinion $x_{k+1}^i(u)$ of agent i on statement u is obtained as a weighted combination of its intermediate aggregated opinion $\bar{x}_k^i(u)$ at time k and its initial opinion $x_0^i(u)$ on statement u. The parameter $0 \leq \lambda^i \leq 1$ that

agent i uses models its stubbornness. If $\lambda^i < 1$ we say an agent is *stubborn*, where $\lambda^i = 0$ indicates that the agent i is *maximally closed* to the influence of others. If $\lambda^i = 1$, agent i is said to be *maximally open* to the influence of others, and *oblivious* if additionally it is not influenced by stubborn agents.

We can group the parameters $\{A_{ij}\}$ into an n-by-n matrix A, known as the *social influence structure*, and the parameters $\{C_{uv}\}$ into an m-by-m matrix C, known as the *multi-issues dependent structure* [38]. These matrices are nonnegative. Furthermore, the weights A_{ij} assigned by an agent i to its neighbors j sum up to one, i.e., the sum of the entries in each row of the matrix A is 1; likewise, the sum of the entries in each row of the matrix C is 1. Thus, the matrices A and C are row-stochastic

Figure 2.1(c) shows the belief system generated by the network of agents in Fig. 2.1(a) and the set of logic constraints in Fig. 2.1(b). This new graph depicted in Fig. 2.1(c) is much larger than the network of agents or the network of statements taken separately; effectively, it has $2nm$ nodes. The belief of each agent on each truth statement is a separate node; also, the initial beliefs are separate nodes.

The model of this larger graph of the belief system can be compactly restated as

$$x_{k+1} = Px_k, \tag{2.2}$$

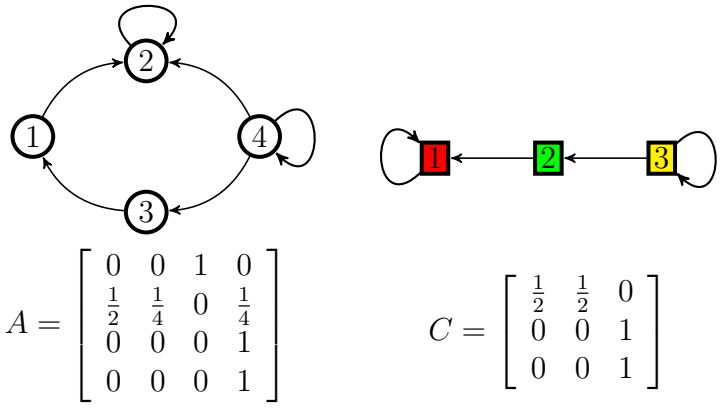
where $x_k \in [0,1]^{2nm}$ is a state that stacks the current beliefs of all agents on all topics along side with the initial beliefs, i.e.,

$$x_k = \Bigg[\underbrace{x_k^1(1), \ldots, x_k^1(m)}_{\text{Beliefs of Agent 1}}, \underbrace{x_k^2(1), \ldots, x_k^2(m)}_{\text{Beliefs of Agent 2}}, \ldots, \underbrace{x_k^n(1), \ldots, x_k^n(m)}_{\text{Beliefs of Agent } n},$$
$$\underbrace{x_0^1(1), \ldots, x_0^1(m)}_{\text{Initial Beliefs of Agent 1}}, \underbrace{x_0^2(1), \ldots, x_0^2(m)}_{\text{Initial Beliefs of Agent 2}}, \ldots, \underbrace{x_0^n(1), \ldots, x_0^n(m)}_{\text{Initial Beliefs of Agent } n} \Bigg]'$$

and

$$P = \left[\begin{array}{c|c} (\Lambda A) \otimes C & (\boldsymbol{I}_n - \Lambda) \otimes \boldsymbol{I}_m \\ \hline \boldsymbol{0}_{nm} & \boldsymbol{I}_{nm} \end{array} \right],$$

where $\boldsymbol{0}_{nm}$ is a zero matrix of size $n \times m$, $\boldsymbol{I}_{nm}$ is an identity matrix of size $n \times m$, $\otimes$ indicates the Kronecker product, Λ is a diagonal matrix with the i-th diagonal entry being λ^i, and x' denotes the transpose of a vector or matrix x. This allows for the definition of the belief system graph $\mathcal{P}$, which is compliant with the matrix P, where an edge from ℓ to r exists if $P_{r\ell} > 0$. Equation (2.3) shows an example of a matrix P for the belief system in Fig. 2.1(c)

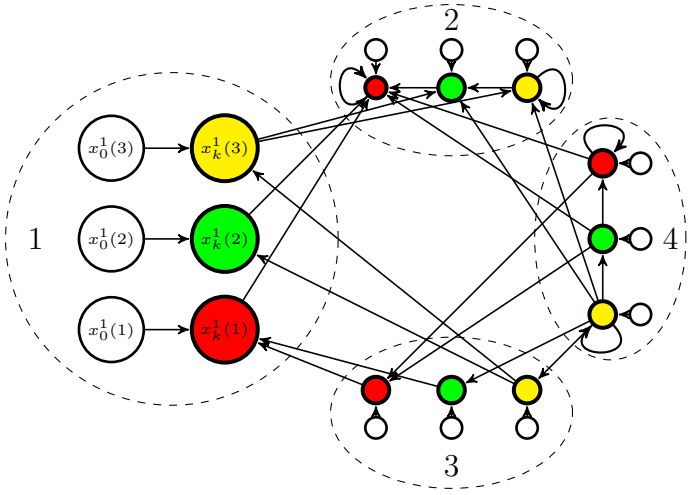


$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

(a)

$$C = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

(b)



(c)

Figure 2.1: A belief system with 4 agents and 3 truth statements. (a) Agents are represented as nodes/circles, numbered from 1 to 4, and the network of influences among them is shown as edges between nodes. The truth statements or topics are color-coded, e.g., the truth statement 1 is represented as a red square. Agent 2 is influenced by its own opinion and agents 4 and 1, agent 1 follows the opinion of agent 3 which in turn follows the opinion of agent 4, agent 4 follows its own opinion only. A possible matrix A for this social network is shown below the graph. This indicates that agent 2 assigns a higher weight of $\frac{1}{2}$ to the opinion of agent 1 than the weight it assigns to the opinion of communicated by agent 4. (b) The truth statement 1 is influenced by the belief that statement 2 is true, statement 2 directly follows the belief in statement 3. A possible matrix C for this set of logic constraints is shown below the graph. The belief that the truth statement 1 is true is influenced (with a weight of $\frac{1}{2}$) by the opinion that the truth statement 2 is true. (c) The beliefs system, see equation 2.2, composed by the agent's interaction graph and the logic constraints.

assuming that $\lambda^i = 0.5$ for all agents.

$$P = \left[\begin{array}{c|c} \left[\begin{array}{cccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{8} & \frac{1}{8} & 0 & \frac{1}{16} & \frac{1}{16} & 0 & 0 & 0 & 0 & \frac{1}{16} & \frac{1}{16} & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} \\ 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \end{array}\right] & \frac{1}{2} \cdot \boldsymbol{I}_{12} \\ \hline \mathbf{0}_{12} & \boldsymbol{I}_{12} \end{array}\right]. \tag{2.3}$$

Figure 2.2 shows an example where a network of 5 agents forms a cycle graph, given in Fig. 2.2(a), a set of 4 logic constraints forms a directed path, given in Fig. 2.2(b), and $\lambda^i = 1$ for all i. The belief system graph is shown in Fig. 2.2(c). Figure 2.2(d) shows dynamics of the belief vector as the number of social interactions increases. The opinion on all 4 topics converges to a single value for all agents. Figure 2.2(e) shows the dynamics of the belief vector when no logic constraints are considered. In this case, the agents reach some agreement on the final value, but this consensual value is different for each of the statements. See Fig. 2.3 for an additional example of the influence of the logic constraints on the resulting belief system and Fig. 2.4 for a variation of the example discussed in Fig. 2.2 when the network of agents is a complete graph.

2.2 Convergence, Convergence Time and Convergence Value

In this section, we provide graph-theoretic answers to the questions of convergence, convergence time and convergence value of a belief system with logic constraints. Particularly, we are interested in how the topology of the graph of agent interactions and the graph of topic relations, as well as the number of agents and the number of truth statements, affect the dynamics of a belief system.

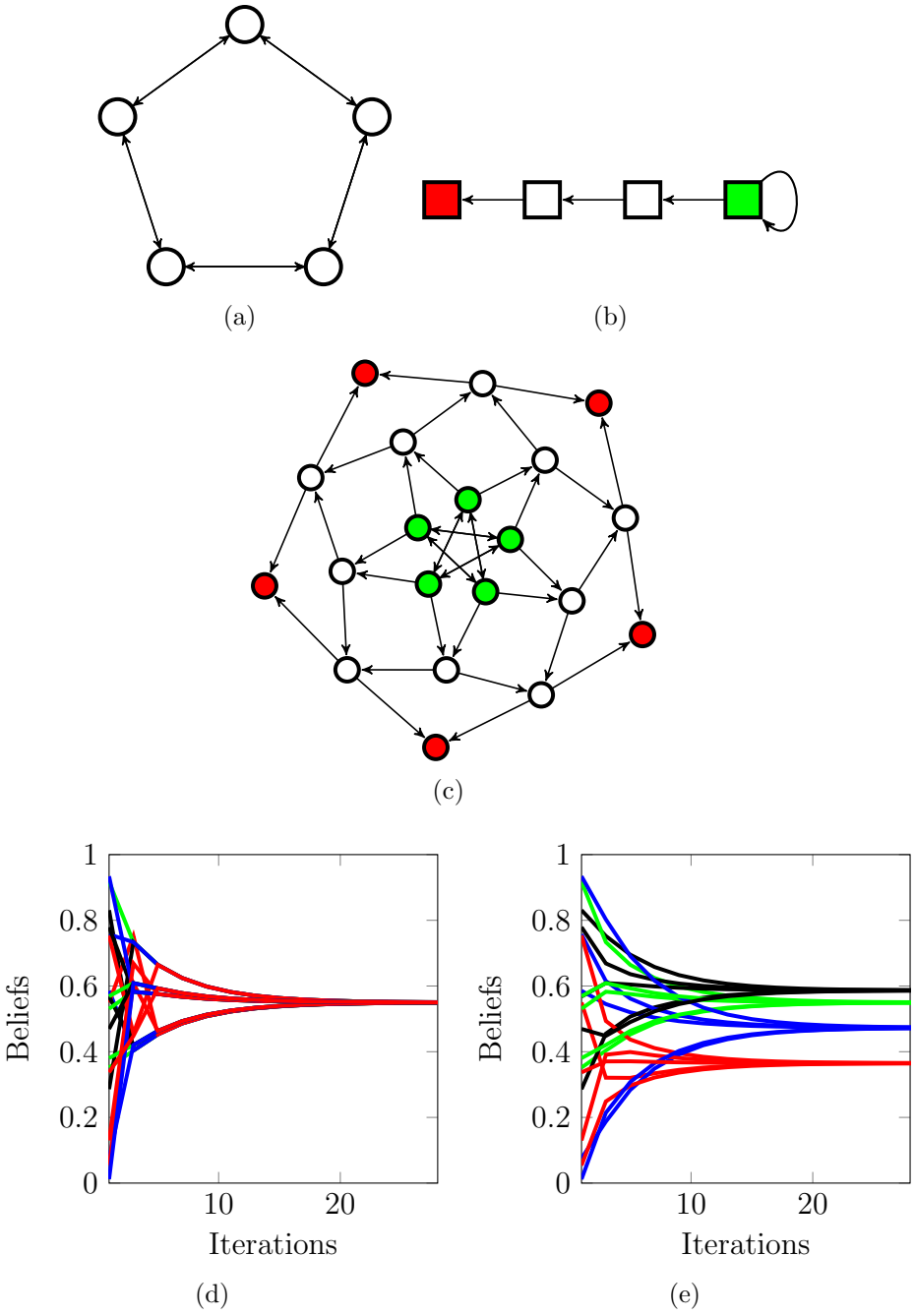


Figure 2.2: A belief system with agents on a cycle graph and logic constraints on a path graph. (a) A network of 5 oblivious agents forming a cycle graph. (b) A set of 4 truth statements with logic constraints forming a path graph. (c) The belief system graph $\mathcal{P}$. (d) The belief dynamics with logic constraints. (e) The belief dynamics with no logic constraints. The beliefs of all agents have been color coded per truth statement. The agents reach an agreement on each of the truth statements.

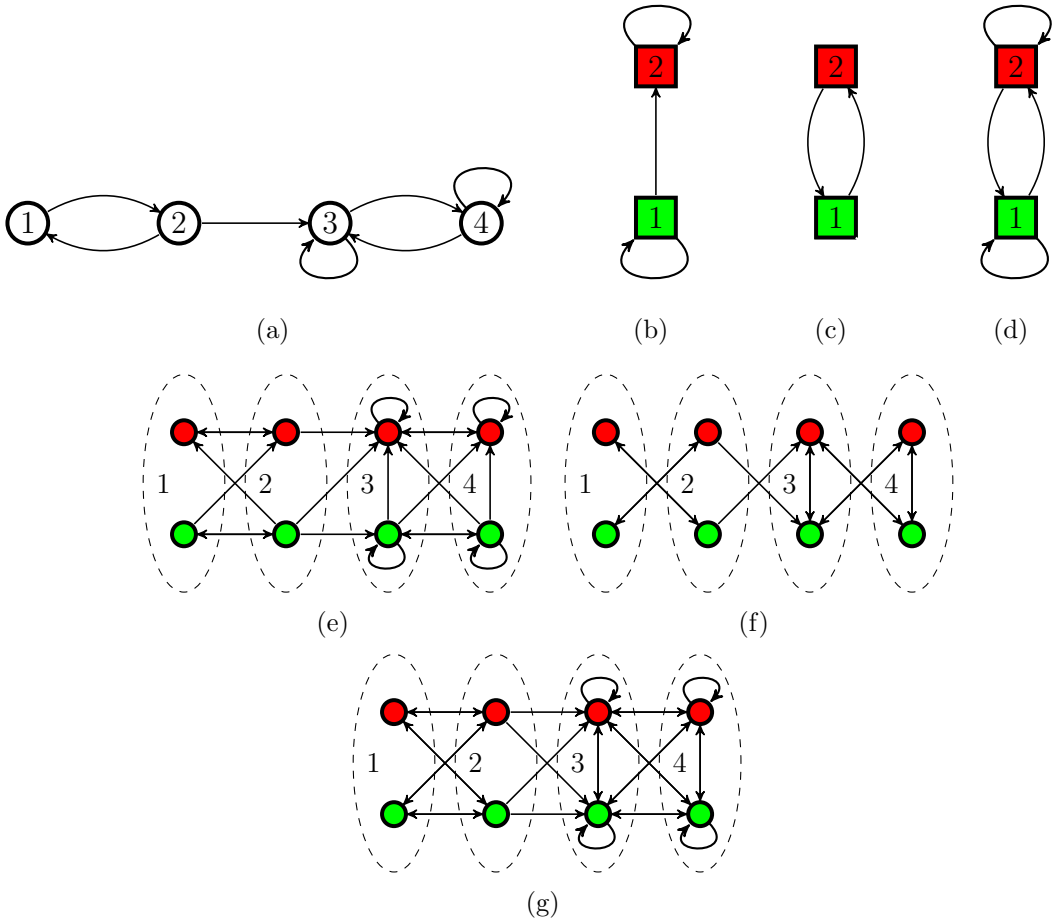


Figure 2.3: The influence of the logic constraints in the resulting aggregated belief system. (a) The network of agents, where agent 1 follows the opinion of agent 2, agent 2 is influenced by agent 1 and 3, agent 3 is influenced by its own opinion, and the opinion of agent 4 and agent 4 is influenced by agent 3 as well as its own. (b) The opinion on statement 1 is influenced by the belief on statement 2. (c) The opinion on statements 2 and 1 follow each other. (d) The opinion on statements 2 and 1 influence each other (e-g) The belief systems with the network of agents in (a) and logic constraints in (b-d).

2.2.1 Does it converge?

The convergence of the belief system can be stated as a question of the existence of a limit of the beliefs in time, as the social interactions continue with time. That is, whether or not there exists a vector of opinions x_∞ such that

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} P^k x_0 = x_\infty$$

for any initial value x_0.

Friedkin et al. [37, 38] showed that a belief system with logic constraints will converge

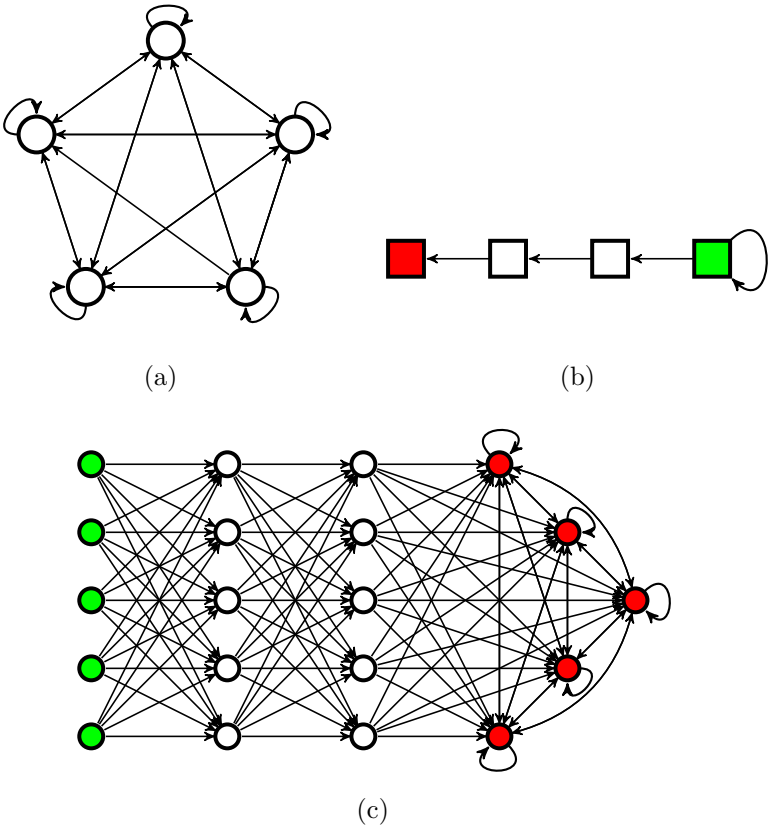


Figure 2.4: Two examples of graph product between a complete graph/cycle graph with 5 nodes and a path graph of 4 logical belief constraints. (a) A complete graph with 5 agents. (b) A path graph with 5 nodes. (c) A cycle graph with 5 agents. (d) The resulting belief system graph from the network of agents in (a) and the network of logic constraints in (b).

to equilibrium if and only if either $\lim_{k\to\infty}(\Lambda A)^k = 0$, or $\lim_{k\to\infty}(\Lambda A)^k \neq 0$ and $\lim_{k\to\infty} C^k$ exists. Moreover, if we represent the matrices A and Λ with a block structure as

$$A = \begin{bmatrix} A^{11} & A^{12} \\ 0 & A^{22} \end{bmatrix} \qquad \Lambda = \begin{bmatrix} \Lambda^{11} & 0 \\ 0 & I \end{bmatrix}$$

where A^{22} is the subgraph of oblivious agents, then the belief system is convergent if and only if $\lim_{k\to\infty} C^k$ and $\lim_{k\to\infty}(A^{22})^k$ exists. We next consider how these conditions may be interpreted in terms of the the topology of the network of agents and the set of logic constraints.

The belief system in Eq. (2.2) converges to equilibrium if and only if every closed strongly connected component of the graph $\mathcal{P}$ is aperiodic [30, 126]. Recall that a strongly connected component is closed if it has no incoming links from other agents; otherwise it is called open, see Fig. 2.5. In general, the set of strongly connected components can be computed

efficiently for large-complex networks[127].

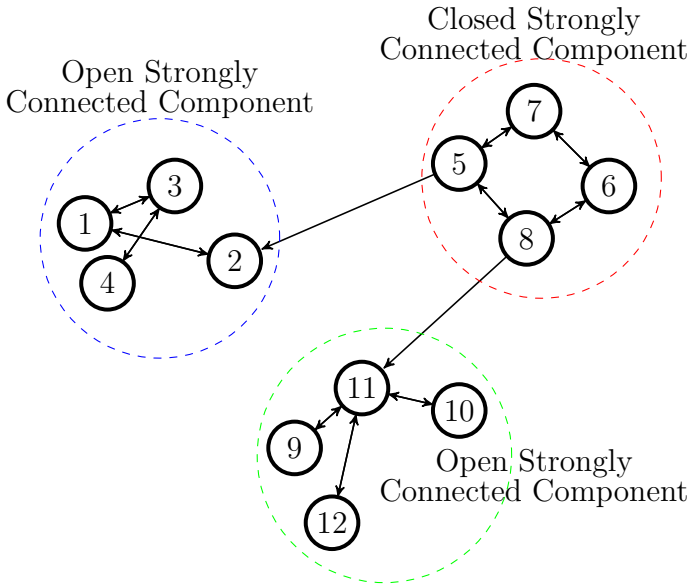


Figure 2.5: Open and closed strongly connected components of a graph. A graph with 12 nodes and 3 strongly connected components. The strongly connected component composed of nodes 5, 6, 7 and 8 is closed since it has no incoming edges from other components.

We can now recall an auxiliary result that will help us later in the convergence analysis of the belief system. Namely, the next theorem relates the connectedness of the Kronecker product of directed graphs with the connectedness of the factor graphs.

Theorem 7 (Theorem 1 in [128])**.** *Let $\mathcal{G}$ and $\mathcal{H}$ be strongly connected directed graphs. Let $d_1 = d(\mathcal{G})$, $d_2 = d(\mathcal{H})$, $d_3 = gcd(d_1, d_2)$ and $D = lcm(d1, d2)$. Then, the number of components in $\mathcal{G} \otimes \mathcal{H}$ is d_3. Moreover, for any component $\mathcal{B}$ of $\mathcal{G} \otimes \mathcal{H}$, $d(\mathcal{B}) = D$.*

Theorem 7 allows us to state our main result in this section. Namely, we relate the connectivity properties of the strongly connected components of the product of two graphs with the connectivity properties of the strongly connected components of the each of the factor graphs.

Lemma 8. *Every strongly connected component of the Kronecker product graph $\mathcal{G}_1 \otimes \mathcal{G}_2$ is the result of the Kronecker product of a strongly connected component of $\mathcal{G}_1$ and a strongly connected component of $\mathcal{G}_2$.*

Proof. Let A_1 and A_2 denote the adjacency matrices for the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. We can construct a condensation og the graph $\mathcal{G}$ by contracting every strongly connected component to a single vertex, resulting in a directed acyclic graph. Thus, a topological

ordering is possible (see Cormen et al. [129] Section 22.4) and there always exists two permutation matrices P_1 and P_2 such that we can rearrange the matrices A_1 and A_2 into a block upper triangular form where each of the blocks is a strongly connected component, that is

$$P_1' A_1 P_1 = \begin{bmatrix} A_1^1 & * & * & * \\ 0 & A_1^2 & * & * \\ 0 & 0 & \ddots & * \\ 0 & 0 & \dots & A_1^{n_1} \end{bmatrix} \quad \text{and} \quad P_2' A_2 P_2 = \begin{bmatrix} A_2^1 & * & * & * \\ 0 & A_2^2 & * & * \\ 0 & 0 & \ddots & * \\ 0 & 0 & \dots & A_2^{n_2} \end{bmatrix}.$$

Moreover, define $P = P_1 \otimes P_2$ and by properties of the Kronecker product, cf. Definition 12, it follows that

$$(P_1' A_1 P_1) \otimes (P_2' A_2 P_2) = P'(A_1 \otimes A_2)P,$$

where P is also a permutation matrix and

$$P'(A_1 \otimes A_2)P = \begin{bmatrix} A_1^1 \otimes A_2 & * & * \\ 0 & \ddots & * \\ 0 & \cdots & A_1^{n_1} \otimes A_2 \end{bmatrix}.$$

Finally, by property 2 in Definition 12 there exists a permutation matrix Q such that

$$Q'(P'(A_1 \otimes A_2)P)Q = \begin{bmatrix} A_2 \otimes A_1^1 & * & * \\ 0 & \ddots & * \\ 0 & \cdots & A_2 \otimes A_1^{n_1} \end{bmatrix},$$

$$= \begin{bmatrix} A_2^1 \otimes A_1^1 & * & * & * & * & * & * \\ 0 & \ddots & * & * & * & * & * \\ 0 & \cdots & A_2^{n_2} \otimes A_1^1 & * & * & * & * \\ 0 & \cdots & 0 & \ddots & * & * & \\ 0 & \cdots & \cdots & 0 & A_2^1 \otimes A_1^{n_1} & * & * \\ 0 & \cdots & \cdots & \cdots & 0 & \ddots & * \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & A_2^{n_2} \otimes A_1^{n_1} \end{bmatrix}.$$

Therefore, every block in the block diagonal form of the product of two adjacency matrices is the product of two strongly connected components, one from each graph. □

The matrix P has two diagonal blocks, one corresponding to the initial beliefs and one in-

volving the product $\Lambda A \otimes C$. The initial belief nodes are aperiodic closed strongly connected components, each consisting of a single node. Therefore, the diagonal block in P corresponding to the initial beliefs induces an aperiodic graph. Moreover, strongly connected components with stubborn agents do not affect the convergence of the belief system. Thus, one can focus on the closed strongly connected components of the graph induced by $A^{22} \otimes C$.

The product $A^{22} \otimes C$ can be written in its block upper triangular form, where each of the blocks in the diagonal is the product of one strongly connected component from the graph induced by A^{22} and one from $\mathcal{T}$ (see Supplementary Lemma 8).Theorem 7 shows that the period of a product graph is the lowest common multiple of the periods of the two factor graphs. If the factor graphs are not coprime the resulting product graph is a disconnected set of components. Nevertheless, each of the resulting components will have the same period as defined above. Therefore, in order for a product graph to be aperiodic we require the factors to be aperiodic as well. An immediate conclusion drawn from this fact is that the process (2.2) converges to an equilibrium if an only if every closed strongly connected component of the graph $\mathcal{T}$ is aperiodic and every closed strongly connected component of the graph $\mathcal{G}$ composed by oblivious agents only is aperiodic. This is a graph-theoretic interpretation of the algebraic criteria derived in [37, 38].

In Fig. 2.1, the network of agents has a single closed strongly connected component which consists of the node 4. The network of truth statements also has a single closed strongly connected component, consisting of the node 3. Thus, the belief system will converge to a set of final beliefs. In Fig. 2.2, the belief system has one closed strongly connected component shown in green with the topology of a cycle graph. This strongly connected component corresponds to the product of the cycle graph and the green node of the logic constraints. The cycle graph is aperiodic if and only if the number of nodes is odd. Thus, if the cycle network of agents has an even number of nodes, the belief system will not converge.

2.2.2 How long does a belief system take to converge?

We seek to characterize the time required by the process in Eq. (2.2) to be arbitrarily close to its limiting value in terms of properties of the graphs $\mathcal{G}$ and $\mathcal{T}$, such as the number of agents and truth statements, and the topology of the graphs. We will provide an estimate on the number of iterations required for the beliefs to be at a distance of at most ϵ from their final value (assuming they converge). We will express this estimate in terms of the total

variation distance, denoted by $\|\cdot\|_{TV}$. For this we define the convergence time as follows:

$$T(\epsilon) = \min_{k\geq 0}\left\{\frac{\|x_k - x_\infty\|_{TV}}{\|x_0 - x_\infty\|_{TV}} \leq \epsilon\right\},$$

where x_k evolves according to Eq. (2.2). Informally, the value $T(\epsilon)$ shows the minimum number of social interactions required for the belief system to be arbitrarily close to their final value as a function of the initial disagreement.

The dynamics of the belief system in Eq. (2.2) are closely related to the dynamics of a Markov chain with a transition matrix P. Convergence to a stationary distribution of a random walk with the transition probability P on the graph $\mathcal{P}$ implies convergence of the dynamics in Eq. (2.2). Therefore, bounds on the convergence time based on the mixing properties of this Markov chain provide rates of convergence for the belief system. Notably, the convergence time is proportional to the maximum time required for the random walk to get *absorbed* into a closed strongly connected component plus the time needed for such component to *mix* sufficiently. Figure 2.6 illustrates this by considering two random walks X and Y with the same transition matrix P.

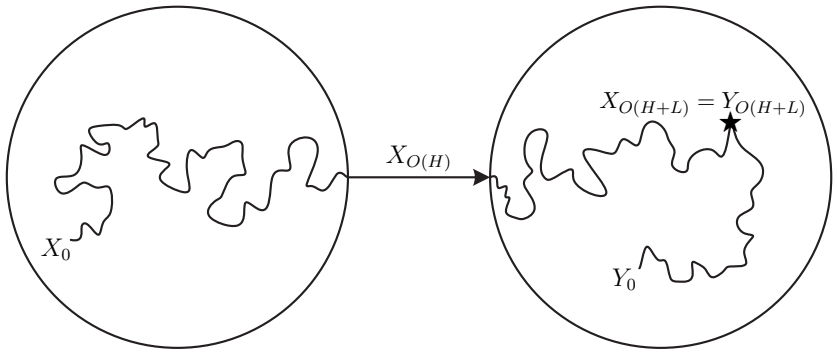


Figure 2.6: Hitting and absorbing time of a random walk. A random walk starts at X_0 in a transient state and evolves according to some transition matrix P; after $O(H)$ time steps (the absorbing time), it gets absorbed into a closed connected component. Then, after $O(L)$ time steps (the mixing time) it crosses paths with another random walk Y_k starting at π the stationary distribution of P. Then after $O((L+H)\log(1/\epsilon))$ time steps, the random walk X_0 is arbitrarily close to its limit value.

If we denote by L the maximum expected mixing time among all closed strongly connected components, and by H the maximum expected time to get absorbed into a closed component, then if the graph $\mathcal{P}$ has no bipartite closed strongly connected components, the belief system will be ϵ close to its limiting distribution after $O((L+H)\log(1/\epsilon))$ steps. Therefore, not

only do we have an estimate of the convergence time of the belief system in terms of the topology of the graph $\mathcal{P}$, but we also know this convergence happens exponentially fast. We formalize this statement in Theorem 9.

Theorem 9. *Assume that there exists at least one closed strongly connected component in the graph $\mathcal{P}$, and that all closed strongly connected components are aperiodic. Let L be the maximum expected coupling time of a random walk in a closed strongly connected component of $\mathcal{P}$. Moreover, let H be maximum expected time for a random walk, starting at an arbitrary node, to get absorbed into a closed strongly connected component. Then, for $k \geq 4(L+H)\log(1/\epsilon)$, it holds for the belief system described in equation (2.2) that $\|x_k - x_\infty\|_{TV} \leq \epsilon$.*

Proof. We use the coupling method to bound the convergence time of the belief system [130]. The main conceptual idea of the proof is relation between the convergence time of the belief system described in equation (2.2) and the ergodic properties of a random walk over on the graph $\mathcal{P}$. Particularly, consider a random walk on the state space $\{1, \ldots, 2nm\}$ which, at time k jumps to a random neighbor of its current state. The relation between a random walk on a graph and the convergence properties of systems of the form of the belief system in (2.2) has been previously explored in [131].

Initially, we show that all opinions x_k^i, such that i lies in a closed strongly connected component, will converge to some stationary point. Thus, in what follows we will find the required time to reach some ϵ-consensus via coupling arguments, which in turn will provide the required time for a belief system to be ϵ close to its stationary distribution.

Let i be a node belonging to a closed strongly connected component S and let P_S be the matrix obtained by looking at the minor of P corresponding to entries in S. If S is closed then P_S is row-stochastic, and Perron-Frobenius theory tells us there exists some vector π_S such that

$$\pi_S' P_S = \pi_S'.$$

Now, define two independent random walks $X = (X_k)_0^\infty$ and $Y = (Y_n)_0^\infty$ with the same transition matrix P_S. X starts from a distribution π_S, and Y from some other arbitrary stochastic vector v. Moreover, *couple* the processes Y and X by defining a new process W such that

$$W_k = \begin{cases} Y_k, & \text{if } k < K, \\ X_k, & \text{if } k \geq K, \end{cases}$$

where $K = \min\{k \geq 0 : Y_k = X_k\}$ is called the *coupling time*. Each random walk moves according to P_S, so if we correlate them by moving them together after they intersect, we have not changed the fact that, individually, they move according to P_S. With this construction of the coupling (Theorem 5.2 in Levin et. al.[98]), we have that

$$\|v' P_S^k - \pi_S\|_{TV} \leq \max_v \mathbb{P}\{K > k\},$$

and by the Markov inequality

$$\|v' P_S^k - \pi_S\|_{TV} \leq \frac{\max_v \mathbb{E}[K]}{k}. \tag{2.4}$$

Therefore, to be at a distance of at most $1/4$ we require $k = 4 \max_v \mathbb{E}[K]$. We say the mixing time of the random walk is $4L$ where we have that $L = \max_v \mathbb{E}[K]$ is the maximum expected time it takes for the random walks X and Y in the source S to intersect. Then, it follows from Eq. 4.36 in Levin et. al.[98] that in order to be ϵ close to the stationary distribution we require at least $k \geq 4L \log(1/\epsilon)$ steps, for any v. Therefore, we have shown that x_k^i for i in a closed strongly connected component S converges to $\pi_S' x_0^S$ at a geometric rate. Here x_0^S stacks those x_0^i that belong to S.

Now, consider the case where i belongs to an open strongly connected component. Let M be the set of states in such connected component. Stacking up x_k^i over i in M into the vector x_k^M, observe that

$$x_{k+1}^M = Z x_k^M + R y_k, \tag{2.5}$$

where Z is strongly connected and substochastic, meaning some rows add up to less than 1. The entries of y_k come from nodes in other strongly connected components and the matrix R represents how they influence the nodes in M.

Initially, assume that y_k converges and call its limit y_∞. Now, consider a random walk that moves around M according to Z; the moment it steps out of M into another strongly connected component we say it is absorbed by it since it can not return to M.

Let q_k^i be the probability the walk is at state i in M at time k. Then

$$q_{k+1}' = q_k{}' Z,$$

and let H_i be the expected time to get absorbed into any other strongly connected compo-

nent, the set of nodes in M is connected to, starting from node i and let

$$H^1 = \max_{i \in M} H_i.$$

If the absorbing strongly connected component is closed, then $H = H^1$. On the other hand, the absorbing strongly connected component will have some other absorbing time H^2, i.e., the time to get absorbed into another strongly connected component. Thus, the total absorbing time H is the sum of the absorbing times of the strongly connected components on the longest path on the condensation of the graph $\mathcal{G}$ from an open strongly connected component to a closed strongly connected component. The condensation of the graph $\mathcal{G}$ is a directed acyclic graph and such path always exist.

By Markov inequality, regardless of where the random walk starts, the probability that it takes more than $4H$ iterations to get absorbed is at most $1/4$. Thus, for all $k \geq 4H \log(1/\epsilon)$ steps we have that $\|q_k\|_1 < \epsilon$.

Now, let z_∞ be the vector that satisfies

$$z_\infty = Z z_\infty + R y_\infty, \tag{2.6}$$

which we know exists since every eigenvalue of Z must be strictly less than 1 (since $Z^k \to 0$). If we define

$$\Delta_k = x_k^M - z_\infty,$$

then subtracting the updates of x_M and z_∞,

$$\Delta_{k+1} = Z\Delta_k + R(y_k - y_\infty). \tag{2.7}$$

It follows that Δ_k goes to zero since we have assumed that $y_k \to y_\infty$, and $Z^k \to 0$.

In conclusion, this argument shows that for all $k \geq 4(L + H) \log(1/\epsilon)$ steps every node is within ϵ of its limiting value. □

The next lemma states the relation of the coupling and absorbing time for random walks on product graphs. Specifically, it shows a maximum-type behavior where the coupling and absorbing time of the product system is the maximum of coupling and aborning of the factors.

Lemma 10. *Consider two aperiodic strongly connected directed graphs $\mathcal{G}_1$ and $\mathcal{G}_2$. The expected coupling time of two random walks on the graph $\mathcal{G}_1 \otimes \mathcal{G}_2$ is $L = \max\{L_1, L_2\}$,*

where L_1 and L_2 are the expected coupling times for random walks on the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively. Similarly, a random walk on an open strongly connected component of a graph $\mathcal{G}_1 \otimes \mathcal{G}_2$ has an expected absorbing time (into another strongly connected component) of $H = \max\{H_1, H_2\}$, where H_1 and H_2 are the expected absorbing times for random walks on the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively.

Proof. Given that both graphs are aperiodic and strongly connected, their product is also aperiodic and strongly connected, and there exists a limiting distribution π for a random walk moving on the Kronecker product graph $\mathcal{G}_1 \otimes \mathcal{G}_2$.

Consider a random walk $X = (X_k)_0^\infty$, on the graph $\mathcal{G}_1 \otimes \mathcal{G}_2$, with transition matrix $A_1 \otimes A_2$ starting with some arbitrary distribution v, where A_1 is the transition probability on a random walk on the graph $\mathcal{G}_1$ and A_2 is the transition probability on a random walk on the graph $\mathcal{G}_2$. Moreover, from the definition of the Kronecker product of graphs, we have that the state space of $\mathcal{G}_1 \otimes \mathcal{G}_2$ is the Cartesian product $V = V_1 \times V_2$, composed by the ordered pairs (i, j) for $i \in V_1$ and $j \in V_2$. Thus, the probability that the random walk X jumps from the node (i, j) to the node $(\bar{i}, \bar{j})$ is $[A_1]_{i,\bar{i}}[A_2]_{j,\bar{j}}$.

Following the coupling method, define another random walk $Y = (Y_k)_0^\infty$ with the same transition matrix $A_1 \otimes A_2$ but starting at the stationary distribution π. Now, construct an new random walk as follows:

$$W_k = \begin{cases} Y_k, & \text{if } k < K, \\ X_k, & \text{if } k \geq K, \end{cases}$$

where $K = \min\{k \geq 0 : Y_k = X_k\}$. Clearly, if the state of the random walk X at time k is $X_k = (i_k, j_k)$ and the state of the random walk Y at time k is $Y_k = (\bar{i}_k, \bar{j}_k)$, then the condition $Y_k = X_k$ implies that $i_k = \bar{i}_k$ and $j_k = \bar{j}_k$. Thus, the coupling time K can alternatively be expressed in terms of the two separate conditions $i_k = \bar{i}_k$ and $j_k = \bar{j}_k$, which in turn represents the coupling conditions for two separate random walks on each individual coordinate where each coordinate represents one of the factor graphs. Therefore, we write the coupling time between the random walks X and Y as $K = \min\{k \geq 0 : Y_k = X_k\} = \min\{k \geq 0 : i_k = \bar{i}_k, j_k = \bar{j}_k\}$ which is equivalent to

$$\begin{aligned} K &= \min\{k \geq 0 : Y_k = X_k\} \\ &= \min\{k \geq 0 : i_k = \bar{i}_k, j_k = \bar{j}_k\} \\ &= \max\{\min\{k \geq 0 : i_k = \bar{i}_k\}, \min\{k \geq 0 : j_k = \bar{j}_k\}\} \\ &= \max\{K_1, K_2\}, \end{aligned}$$

where K_1 and K_2 are the coupling times for the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively. Thus,

$$\begin{aligned}\mathbb{P}\{K > k\} &= \mathbb{P}\{\max\{K_1, K_2\} > k\} \\ &= 1 - \mathbb{P}\{\max\{K_1, K_2\} \leq k\} \\ &= 1 - \mathbb{P}\{K_1 \leq k\}\,\mathbb{P}\{K_2 \leq k\} \\ &\leq \mathbb{P}\{K_1 \leq k\} + \mathbb{P}\{K_2 \leq k\}.\end{aligned}$$

Note that given that the initial state of the random walk X is v, the random walks on each of its coordinates have some well defined initial state, $v_1(i) = \sum_{j \in V_2} v((i,j))$ and $v_2(j) = \sum_{i \in V_2} v((i,j))$, where $v_1(i)$ is the probability of staring in node $i \in V_1$, $v_2(j)$ is the probability of starting in node $j \in V_2$, and $v((i,j))$ is the probability of the random walk X to start in the node (i,j).

It follows from Theorem 5.2 in Levin et. al.[98] that

$$\begin{aligned}\|v'(A_1 \otimes A_2)_S - \pi\|_{TV} &\leq \max_v \mathbb{P}\{K > k\} \\ &\leq \max_{v_1} \mathbb{P}\{K_1 > k\} + \max_{v_2} \mathbb{P}\{K_2 > k\} \\ &\leq \max_{v_1} \frac{\mathbb{E}[K_1]}{k} + \max_{v_2} \frac{\mathbb{E}[K_2]}{k} \\ &= \max_{v_1} \frac{L_1}{k} + \max_{v_2} \frac{L_2}{k}.\end{aligned}$$

Thus, in order to be at a distance at most $1/4$ from the stationary distribution we require $k \geq 8\max\{L_1, L_2\}$. Moreover, in order to be ϵ close to the stationary distribution we require at least $k \geq 8\max\{L_1, L_2\}\log(1/\epsilon)$ steps in the random walk for any initial state v. Finally, the coupling time of X is $L = \max\{L_1, L_2\}$.

A similar argument follows for the absorbing time of a random walk on a transient component defined by a product graph requires both coordinates be absorbed individually, thus $H = \max\{H_1, H_2\}$. □

We have shown that each of the strongly connected components of the graph $\mathcal{P}$ is the product of two such components, one from the graph $\mathcal{G}$ and the other from the graph $\mathcal{T}$. Moreover, the expected mixing and absorbing times for a random walk on a product graph are the maximum of the expected mixing and absorbing times of the individual factor graphs (see Lemma 10). Thus, we have an explicit characterization of the convergence time in terms of the components of the network of agents and the network of logic constraints. For example, in Fig. 2.2, the expected absorbing time is of the order of the number of nodes in the path,

that is m, while the expected mixing time of cycle [133, 134, 98] graphs is of the order of the number of the nodes squared, which is n^2 in this example. Thus, the expected convergence time for the belief system is $O(\max(n^2, m)\log(1/\epsilon))$. Figure 2.7 depicts simulation results for this bound that demonstrate its validity. In particular, Fig. 2.7(a) shows how the convergence time changes when the number of nodes in the cycle graph increases, while Fig. 2.7(b) shows how the convergence time changes when the number of truth statements in the directed path graph increases. Moreover, Fig. 2.7(c) shows that the convergence to the final beliefs is exponentially fast.

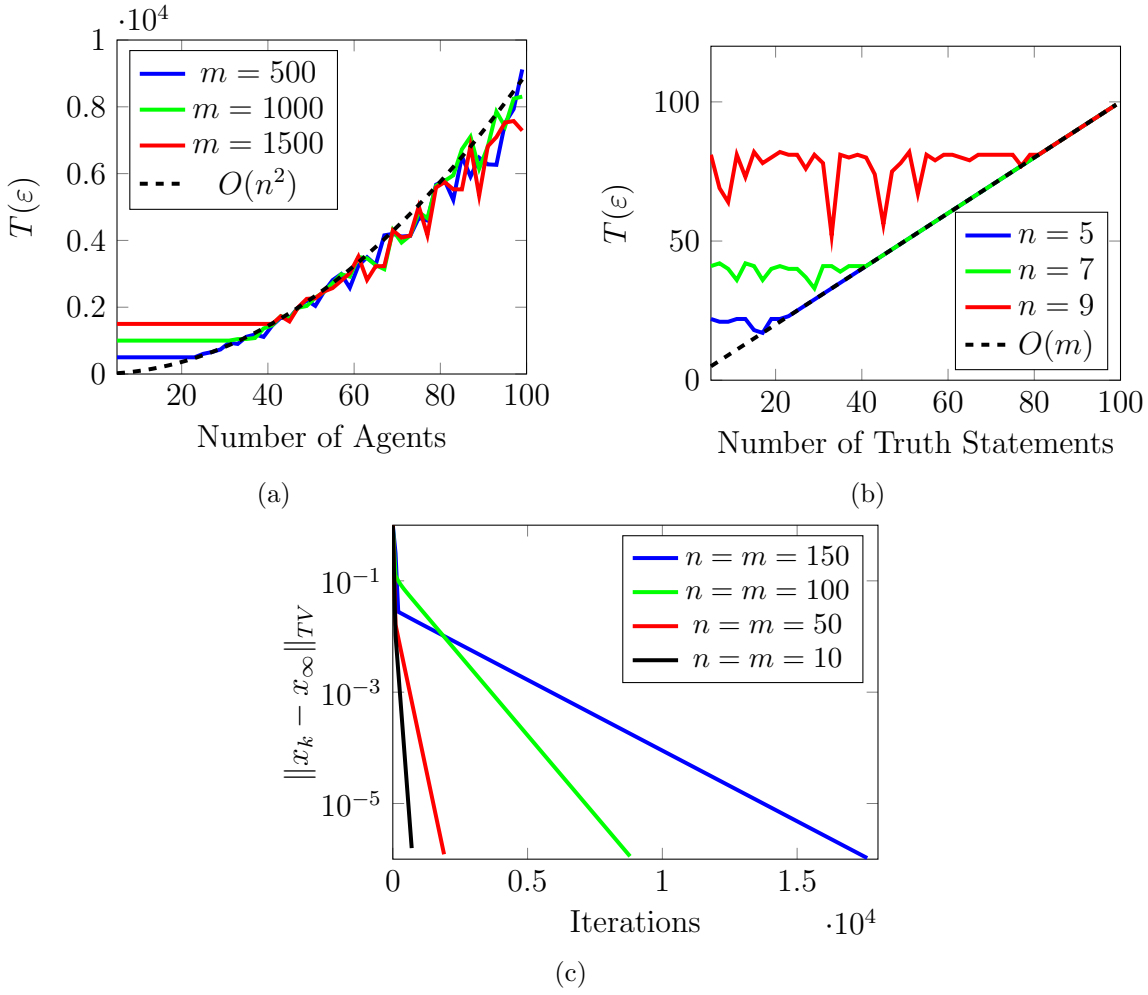


Figure 2.7: Convergence time for a belief system with an undirected cycle as a social network and a directed path as a network for the logic constraints. (a) Varying the number of the agents in the social graph. (b) Varying the number of the truth statements for a directed path. (c) The exponential convergence rate of the belief system.

Table 2.1 presents the estimates for the expected convergence time for belief systems composed of well-known classic graphs. We use the existing results about the mixing time

for these graphs to provide an estimate of the convergence time of the resulting belief system when all agents are oblivious. Table 1.1 shows a detailed list of references on each of the studied graphs. Particularly, our method allows the direct estimation of the dynamics of a belief system when large-scale complex networks are involved. For example, we provide convergence time bounds for the case where networks follow random graph models, namely: the geometric random graphs, the Erdős-Rényi model, and the Newman-Watts small-world networks. These graphs are usually considered for their ability to represent the behavior of complex networks encountered in a variety of fields [135, 136, 137, 138] (see Fig. 2.8).

Figure 2.9 shows experimental results for the convergence time of a belief system for a subset of the graphs given in Table 2.1. For every pair of graphs, we show how the convergence time increases as the number of agents or the number of truth statements change. One can particularly observe that the maximum type behavior on the convergence time as predicted by the theoretical bounds. See Fig. 2.10 and Fig. 2.11 for additional numerical results on other combinations of graphs from Table 2.1, and Fig. 2.12 and Fig. 2.13 for their linear convergence rates, respectively.

2.2.3 What does it converge to?

So far we have discussed the conditions for convergence of belief system dynamics and the corresponding convergence time. Convergence implies the existence of a vector x_∞ where the set of beliefs settles as the number of interactions increases. Particularly, Proskurnikov and Tempo [126] characterize the limiting distribution as a solution of

$$X_\infty = \Lambda A X_\infty C' + (I - \Lambda) X_0,$$

which can be intractable to compute when the matrices A and C are large. We are interested in a characterization of this limit vector that admits a rapid computation of its value.

Lemma 8 shows that one can always group the nodes in the graph $\mathcal{P}$ into open and closed strongly connected components. In order to guarantee convergence we assume that every closed strongly connected component is aperiodic. For example, assume there is a closed strongly connected component S and let P_S be the minor of the matrix P obtained by taking into account only the nodes in the set S. Then, P_S corresponds to the transition matrix of an irreducible and aperiodic Markov chain with a stationary distribution π_S, where $\pi_S' P_S = \pi_S'$. The vector π_S is effectively the left-eigenvector of the matrix P_S corresponding to the eigenvalue 1. Let x_k^S be the vector obtained from the state vector x_k by taking only

Table 2.1: Maximum expected convergence time for the belief system with Logic constraints for different networks of agents with n nodes and networks of truth statements with m nodes. The approximated maximum expected convergence time identified as $\approx$ should be understood in terms of the order $O(\cdot)$, that is, an estimate up to constant terms. Additionally, all the estimates provided should be multiplied by the accuracy term $\log(1/\epsilon)$.

Network of Agents	**Logic Constraints**	**Maximum Expected Convergence Time $\approx$**
Complete	Directed Path	m
Cycle	Directed Path	$\max(n^2, m)$
Cycle	Path	$\max(n^2, m^2)$
Dumbbell Graph	Complete Binary Tree	$\max(n^2, m)$
k-d Cube with Loops	Complete Binary Tree	$\max((1-1/k), m$
k-d Hypercube $\{0,1\}^k$	Complete Binary Tree	$\max(k \log k, m)$
Lovasz Graph $\mathcal{C}_n^k$	Dumbbell	$\max(1-1/(kn^2), m^2)$
2-d Grid	Star	$\max(n \log n, m)$
3-d Grid	Two Joined Star	$\max(n^{2/3} \log n, m)$
k-d Grid	Star	$\max(k^2 n^{2/k} \log n, m)$
2-d Torus	2-d Grid	$\max(n^2, m \log m)$
3-d Torus	Star	$\max(n^2, m)$
k-d Torus	k-d Grid	$\max(n^2 k \log k, k^2 m^{2/k} \log m)$
Lollipop	Star	$\max(n^3, m)$
Barbell	Star	$\max(n^3, m)$
Eulerian: d-degree and expansion	Dumbbell	$\max(\|E\|^2, m^2)$
Eulerian: d-degree, max-degree weights	Dumbbell	$\max(n^2 d, m^2)$
Lazy Eulerian with degree d-degree	Dumbbell	$\max(n\|E\|, m^2)$
Lamplighter on k-Hypercube	Bolas	$\max(k2^k, m^3)$
Lamplighter on (k,n)-Torus	Bolas	$\max(kn^k, m^3)$
Geometric Random: $\mathcal{G}^d(n,r)$	Bolas	$\max(r^{-2} \log n, m^3)$
Geometric Random: $r = \Omega(\text{polylog}(n))$	Bolas	$\max(\text{polylog}(n), m^3)$
Erdős-Rényi: $\mathcal{G}(n, c/n)$, $c > 1$	Dumbbell	$\max(\log^2 n, m^2)$
Erdős-Rényi: $\mathcal{G}(n, c/n)$, $c > 1$	Newman-Watts	$\max(\log^2 n, \log^2 m)$
Erdős-Rényi: $\mathcal{G}(n, (1+\delta)/n)$, $\delta^3 n \to \infty$	Dumbbell	$\max((1/\delta^3) \log^2(\delta^3 n), m^2))$
Erdős-Rényi: $\mathcal{G}(n, 1/n)$	Dumbbell	$\max(n, m^2)$
Newman-Watts : $\mathcal{G}(n, k, c/n)$, $c > 0$	Path	$\max(\log^2 n, m^2)$
Expander	Path	m^2
Exponential Random: High temperature	Path	$\max(n^2 \log n, m^2)$
Exponential Random: Low temperature	Path	$\max(\exp(n), m^2)$
Any Connected Undirected Graph with Metropolis Weights	Expander	n^2
Any Connected Undirected Graph	Expander	$\|E\|\text{diam}(\mathcal{G})$

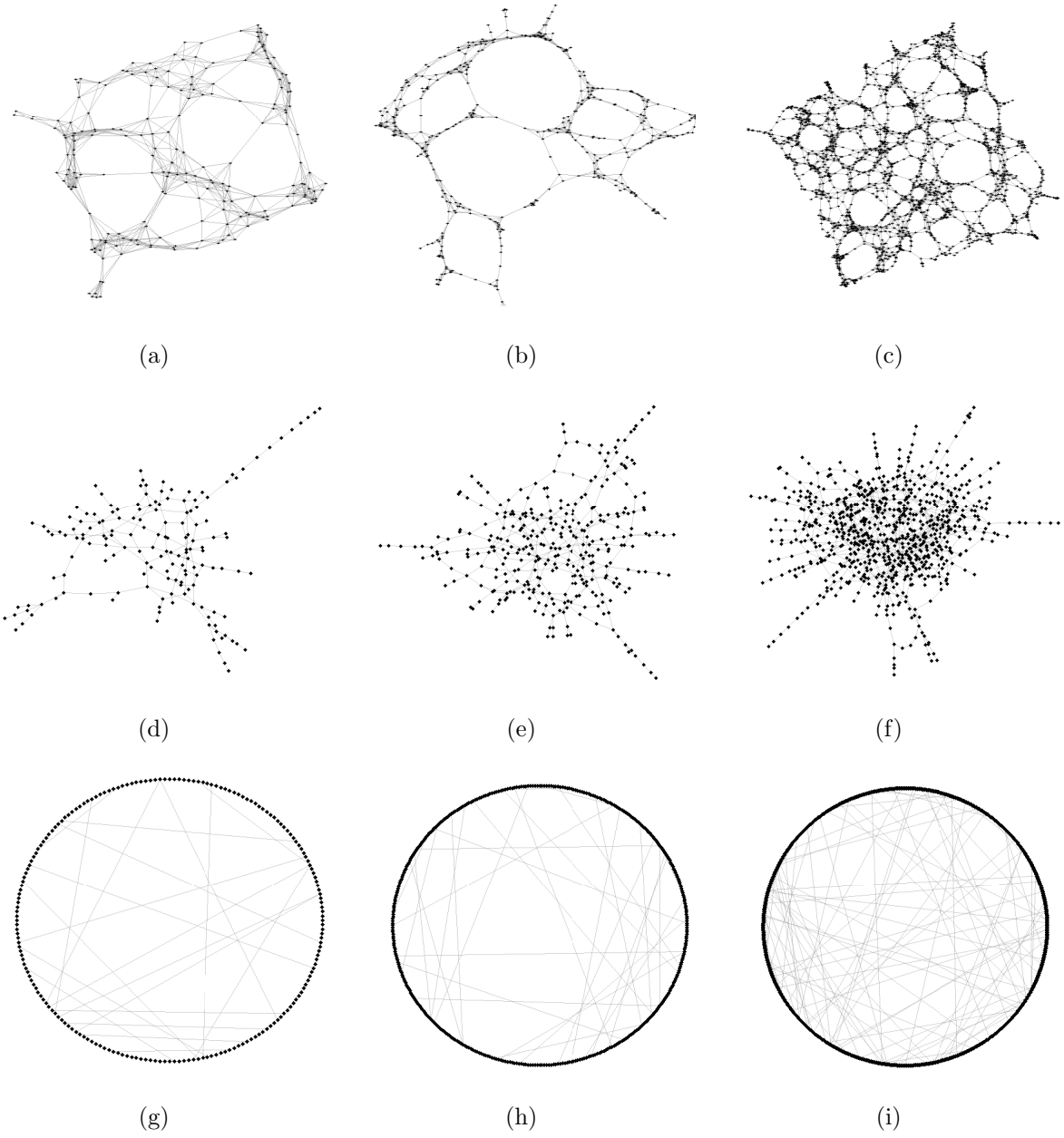


Figure 2.8: (a-c) Geometric random graphs with 200, 400 and 2000 nodes respectively. A geometric random graph is the result of randomly placing n nodes in a metric space and adding an edge between two nodes if and only if their distance is smaller than a certain radius r [139]. (d-f) Erdős-Rényi random graphs with 200, 400 and 1000 nodes respectively. An $\mathcal{G}_{n,p}$ Erdős-Rényi graph is the result of adding edges independently with probability p to a set of n nodes [140]. (g-i) Small-World Random Graphs with 200, 400 and 1000 nodes respectively. The Newman-Watts graph $H_{n,k,p}$ is the random graph obtained from a (n, k)-ring graph by independently adding edges with probability p [141].

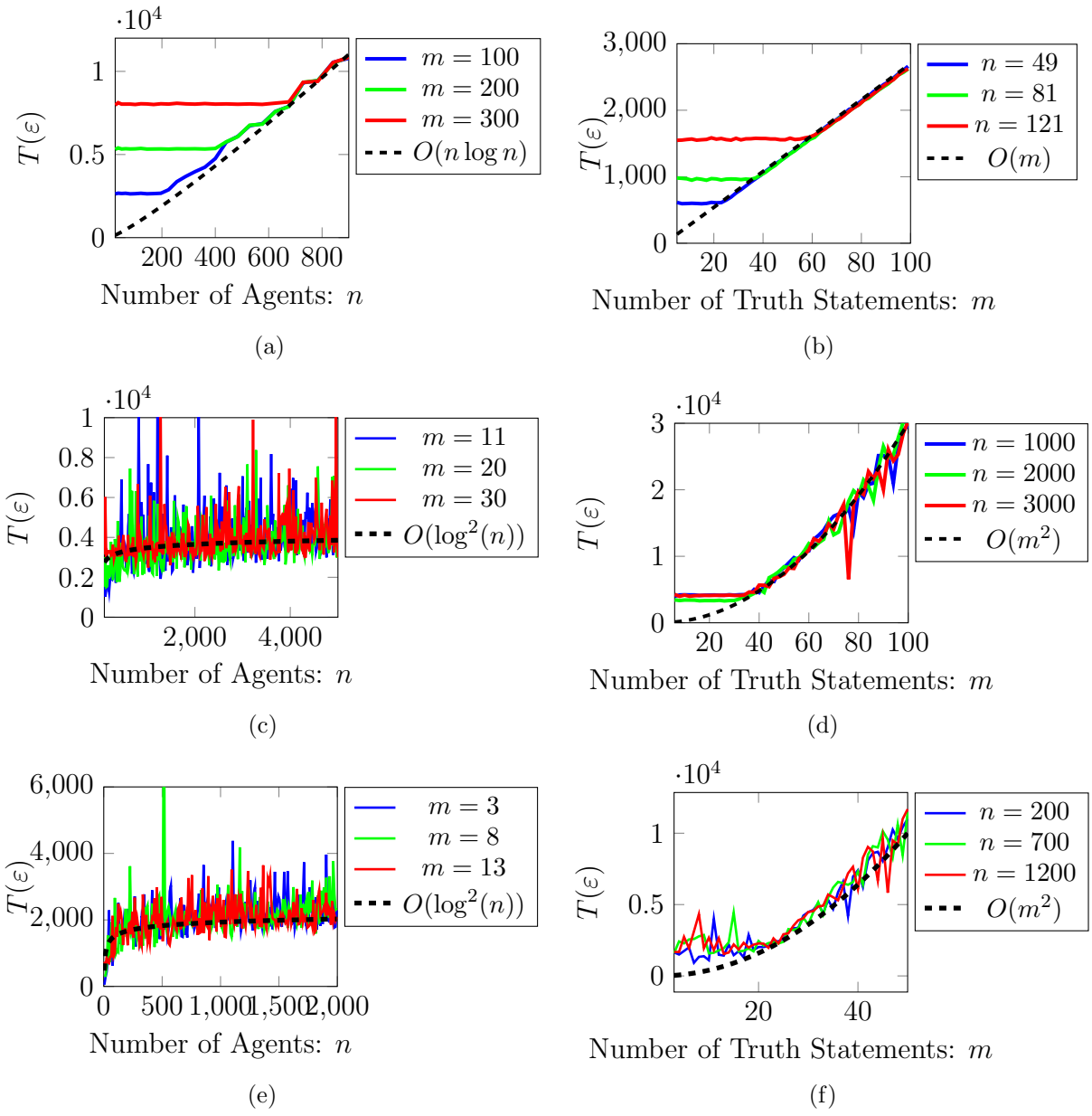


Figure 2.9: Convergence time of various belief systems. (a) Varying the number of agents on a 2d-grid with fixed the number of truth statements on a star graph. (b) Varying the number of truth statements on a star graph with fixed number of agents on a 2d-grid. (c) Varying the number of agents on a Erdős-Rényi graph with fixed the number of truth statements on a dumbbell graph. (d) Varying the number of truth statements on a dumbbell graph with fixed number of agents on a Erdős-Rényi graph. (e) Varying the number of agents on a Newman-Watts small-world graph with fixed the number of truth statements on a path graph. (f) Varying the number of truth statements on a path graph with fixed number of agents on a Newman-Watts small-world graph.

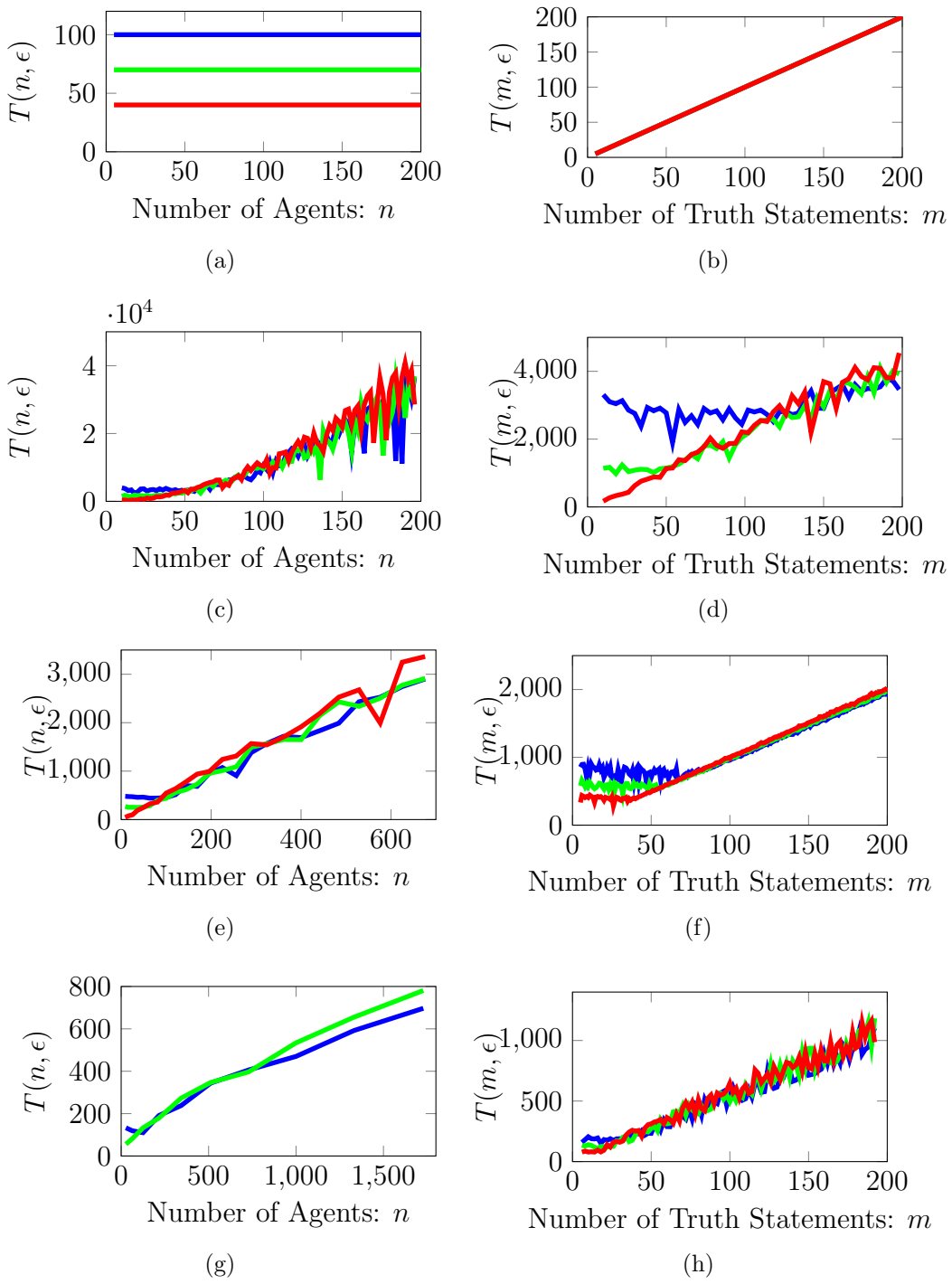


Figure 2.10: Convergence time for different examples of networks of agents and network of truth statements in a belief system. Varying the number of agents for a: (a) complete graph, (c) undirected cycle, (c) dumbbell graph, (e) 2-d grid and (g) 3-d grid. Varying the number of truth statements for a: (b) directed path, (d) complete binary tree, (f) star graph and (h) two joined star graphs.

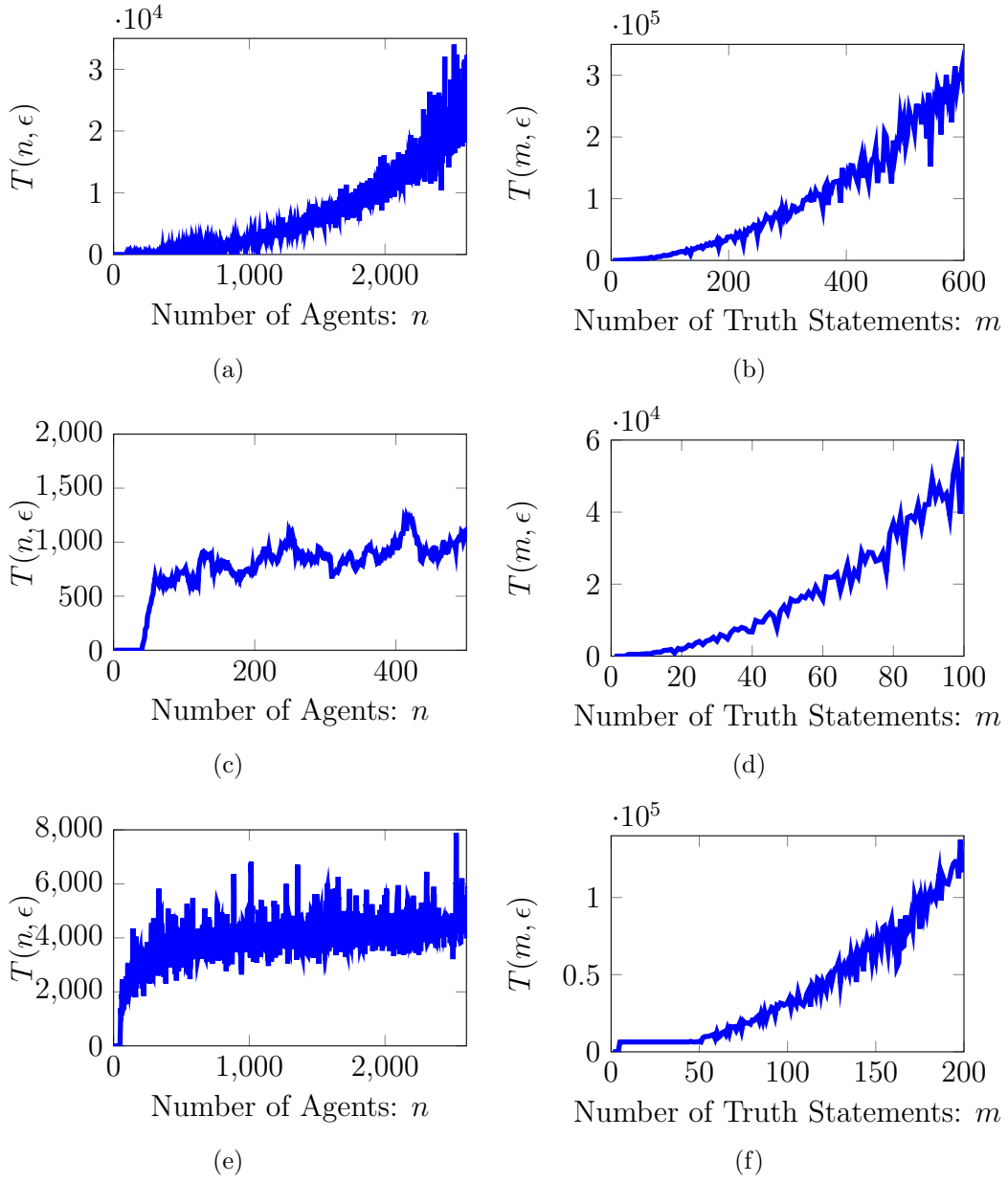


Figure 2.11: Convergence time dependency for random graphs. (a) Varying the number of agents in a geometric random graph with a fixed number of truth statements in a Bolas graph. (b) Varying the number of truth statements in a Bolas graph with a fixed number of agents in a geometric random graph. (c) Varying the number of agents in a Erdős-Rényi random graph with a fixed number of truth statements in a dumbbell graph. (d) Varying the number of truth statements in a dumbbell graph with a fixed number of agents in an Erdős-Rényi random graph. (e) Varying the number of agents in an Newman-Watts random graph with a fixed number of truth statements in an undirected path graph. (f) Varying the number of truth statements in an undirected path graph with a fixed number of agents in a Newman-Watts random graph random graph.

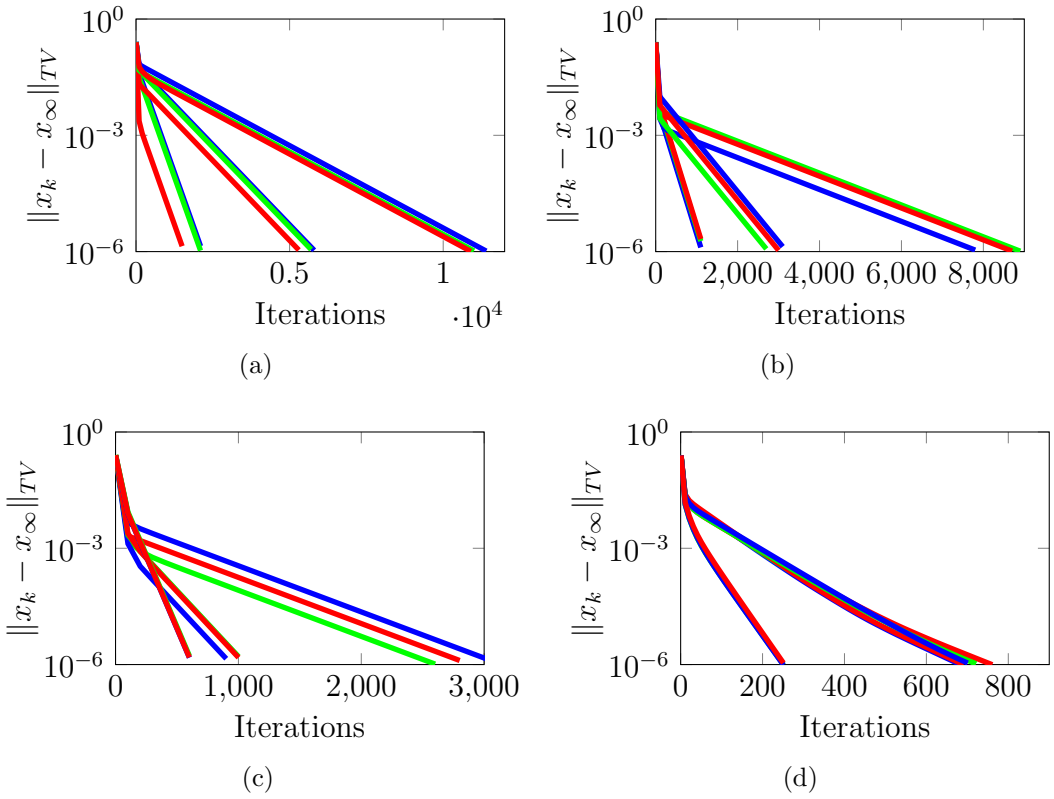


Figure 2.12: Exponentially fast convergence of the belief system. Distance to the final value of a belief system with: (a) a directed cycle network of agents and a directed path of truth statements, (b) a dumbbell network of agents and a complete binary tree of truth statements, (c) a 2-d grid of agents and a star network of truth statements, (d) a 3-d grid of agents and a two-joined star network of truth statements.

the components of x_k corresponding to the nodes in the set S. Then,

$$\lim_{k\to\infty} x_k^S = \pi_S' x_0^S \mathbf{1}_{|S|},$$

where $|S|$ is the cardinality of the set S, and 1_p is the vector of size p with all entries equal to 1 [126, 98]. Additionally, recall that every strongly connected component of $\mathcal{P}$ is the product of two strongly connected components, one from the network of agents and one from the logic constraint network. Thus, $P_S = A_S \otimes C_S$ for some matrices A_S and C_S (sub-matrices of A and C, respectively), which implies that $\pi_S = \pi_S^A \otimes \pi_S^C$, i.e., the vectors π_S^A and π_S^C are the corresponding left eigenvalues of the factor components of P_S associated with the eigenvalue 1. Therefore, the final beliefs of those nodes in the closed strongly connected component S are a weighted average of their initial beliefs, and the weights (sometimes referred to as the social power) are determined by the product of the left-eigenvectors of the factors A_S and

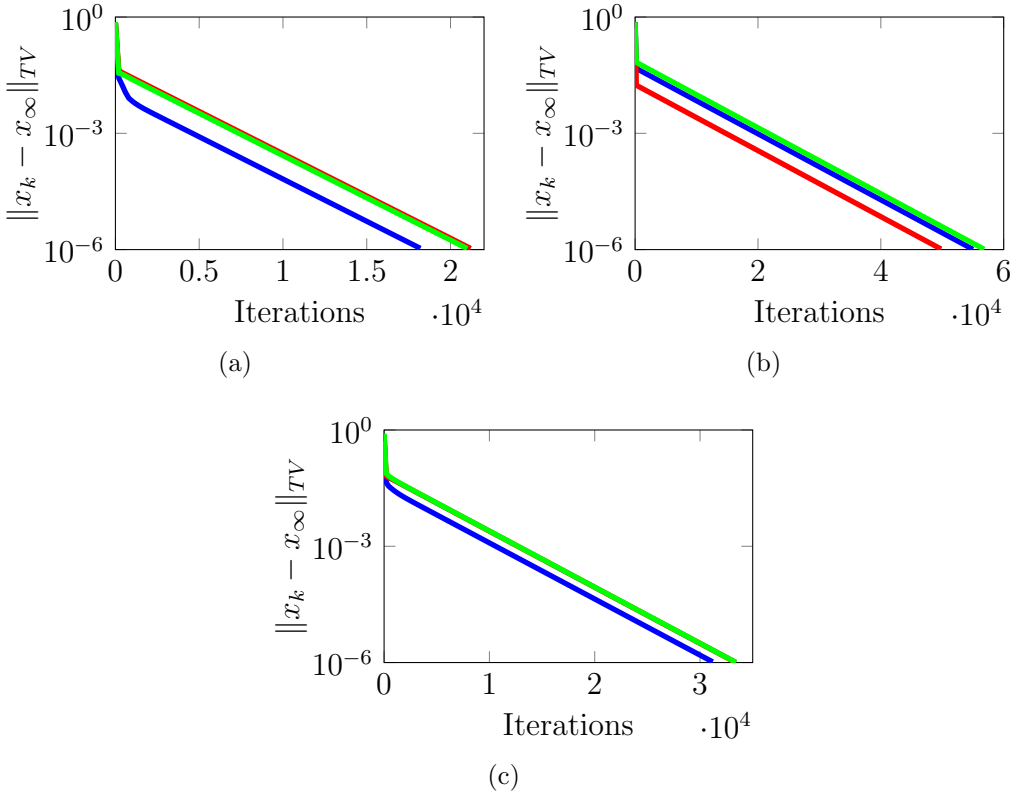


Figure 2.13: Geometric convergence of the belief system with random networks of agents. (a) Distance to the stationary distribution for a network of 200 agents modeled as a geometric random graph and a network of 150 truth statements modeled as a Bolas graph. (b) Distance to the stationary distribution for a network of 500 agents modeled as a Erdős-Rényi random graph and a network of 100 truth statements modeled as a dumbbell graph. (c) Distance to the stationary distribution for a network of 500 agents modeled as a small-world random graph and a network of 100 truth statements modeled as a undirected path graph.

C_S. Particularly, the value π_S indicates the limit distribution of a random walk in S, that is, it tells the probability that a random walk visits a particular node in S after a long time.

On the other hand, now consider an open strongly connected component M with incoming edges from nodes grouped into a set S^M, in this case, the belief x_k^i, for $i \in M$, will converge to

$$\lim_{k\to\infty} x_k^i = \sum_{j\in S^M} p_{ij} x_\infty^j$$

where p_{ij} is the probability of absorption of a random walk starting at node i into a node $j \in S^M$ with limiting value x_∞^j. Therefore, the limiting value of nodes in an open strongly

connected components is a convex combination of the limiting values of the nodes it is connected to.

In order to compute the limiting values of the belief system consider a random walk starting in an open strongly connected component M of a graph $\mathcal{G}$. Moreover, assume the open strongly connected component M has incoming edges from nodes in other strongly connected components, and group those nodes in a set defined as $S^M = \{j \mid (j,i) \in E, i \in M\}$.

The dynamics of the nodes in M are described in equation (2.5). Similarly as in Theorem 9, we can assume that y_k converges to some value y_∞. Therefore, we can analyze the dynamics in the strongly connected component M as follows: Initially define the following two systems

$$\begin{aligned}\bar{x}^M_{k+1} &= Z\bar{x}^M_k + Ry_\infty \\ x^M_{k+1} &= Zx^M_k + Ry_k,\end{aligned}$$

where Z is the set of weights assigned to nodes inside the component M and R is the set of weights assigned to each of the incoming edges from other components.

It follows that

$$\begin{aligned}\lim_{k\to\infty}(\bar{x}^M_{k+1} - x^M_{k+1}) &= Z\lim_{k\to\infty}(\bar{x}^M_k - x^M_k) + R\lim_{k\to\infty}(y_\infty - y_k) \\ &= Z\lim_{k\to\infty}(\bar{x}^M_k - x^M_k).\end{aligned}$$

Moreover, given that Z is substochastic, the magnitude of its eigenvalues are strictly less than 1 and $1 - Z$ is invertible. Thus, we can conclude that $\lim_{k\to\infty}\bar{x}^M_{k+1} = \lim_{k\to\infty} x^M_{k+1}$.

Stacking the vector $\bar{x}^M_k$ and y_∞ into a single vector we obtain the following recursion:

$$\begin{bmatrix} \bar{x}^M_{k+1} \\ y_\infty \end{bmatrix} = P^M \begin{bmatrix} \bar{x}^M_k \\ y_\infty \end{bmatrix}, \tag{2.8}$$

where

$$P_M = \begin{bmatrix} Z & R \\ 0 & I \end{bmatrix}.$$

Thus, in order to find the limit value of the set of beliefs in the component M we can focus in the analysis of the powers of the matrix P^M.

We have that

$$\lim_{k\to\infty} P^k_M = \begin{bmatrix} 0 & NR \\ 0 & I \end{bmatrix},$$

where $N = I + Z + Z^2 + \cdots = (1 - Z)^{-1}$. The matrix NR is the absorbing probability matrix, where $p_{ij} \triangleq [NR]_{ij}$ is the probability of being absorbed by into the node $j \in S^M$ starting from node $i \in M$. Moreover, it follows that for any node $i \in M$

$$\lim_{k\to\infty} x_k^i = \sum_{j\in S^M} p_{ij} x_\infty^j,$$

where x_∞^j is the limiting value of the components in the closed strongly connected component $j \in S^M$.

Therefore, the limiting value of nodes in an open strongly connected component is a convex combination of the limiting values of the closed strongly connected components it is connected to.

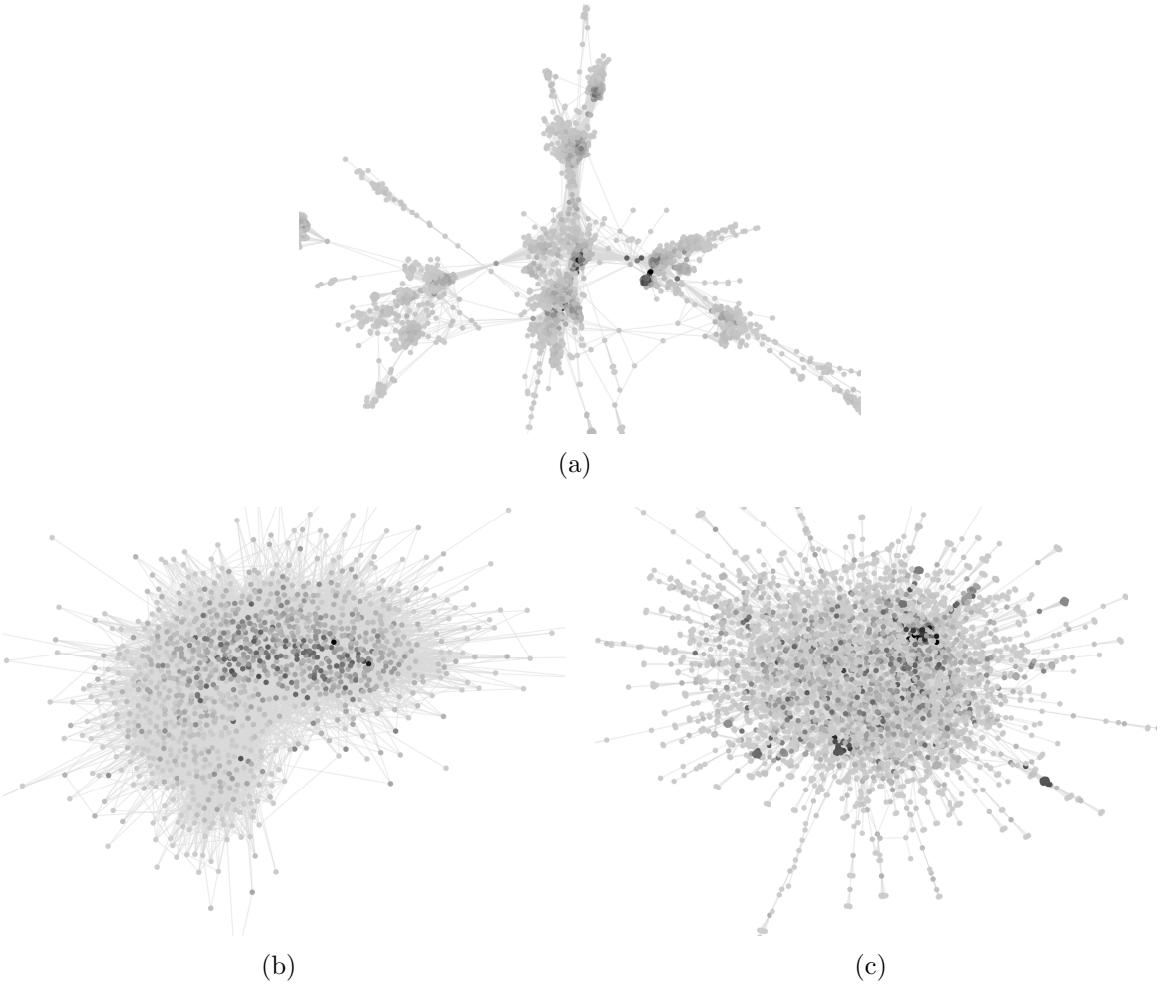
2.3 Numerical Analysis

Next, we provide a numerical analysis of three large-scale networks from the Stanford Network Analysis Project (SNAP)[142], see Fig. 2.14. Table 2.2 shows the description of the networks used. In the three cases, we select the largest strongly connected component of the graph and use it as a representative of the network structure and the mixing properties of the graph. Furthermore, we assume that the agents use equal weights for all their (in)neighbors.

Table 2.2: Datasets of large-scale networks. Description, the number of nodes, the number of edges, simulated mixing time and an upper bound on the mixing time of the three datasets used in the numerical analysis. The upper bound on the mixing time is computed from the second largest eigenvalue bound in Eq. (1.4)

Graph	**Nodes**	**Edges**	**Type**	**Upper Bound on Mixing Time**	**Description**
`wiki-Vote` [143]	1300	103663	Directed	145	Wikipedia who-votes-on-whom network
`ca-GrQc` [144]	4158	13428	Undirected	12308	Collaboration network of arXiv General Relativity
`ego-Facebook`[145]	3927	88234	Undirected	53546	Social circles from Facebook

Random graph generating models, such at the Erdős-Rényi graphs, the Newman-Watts graph, and the geometric random graphs, have been proposed to model the dynamics and the properties of real large-scale complex networks, for example, rapid mixing or linear



(a)

(b) (c)

Figure 2.14: Large-scale complex networks from the Stanford Network Analysis Project (SNAP). (a) The `ego-Facebook`, nodes are anonymized users from Facebook and edges indicate friendship status between them. (b) The `wiki-Vote` graph, each node represents a Wikipedia administrator and an directed edge represents a vote used for promoting a user to admin status. (c) The `ca-GrQc` graph is a collaboration network from arXiv authors with papers submitted to the General Relativity and Quantum Cosmology category, edges indicate co-authorship of a manuscript. The gray scale in the node colors shows the relative social power according to the left-eigenvector corresponding to the eigenvalue 1.

convergence of the beliefs. The existing approaches for the computation of such properties in real-world social networks are mainly simulation-based or require extensive computations for the approximation of the spectral properties of the graphs [146, 147]. Our analysis based on the graph-theoretical properties of the networks provides a structural explanation of the exhibited behavior. Specifically, we can explain fast mixing or equivalently linear convergence of beliefs from the existence of highly influential cliques that drive the dynamics of the complete belief system. In particular, suppose we want to study whether a specific

graph has a rapid mixing and, for example, there is a subset of $\bar{V}$ of M nodes that affect the 20% of the final opinion. Then, it is enough to check if there is a finite number K such that after K time steps, a random walk in the graph has a probability of 1/5 to be in $\bar{V}$, resulting in the probability of being at any of these particularly influential nodes of $\frac{1}{5M}$. The next theorem describes how the existence of a clique of a well-connected subset of nodes can guarantee fast mixing of a random walk on a graph.

Theorem 11. *Consider a random walk on a connected undirected and static graph $\mathcal{G} = (V, E)$ with $|V| = n$ nodes, and assume there is a subset $\bar{V} \subset V$ with M nodes such that after K steps, the probability of being in $\bar{V}$ is at least $\frac{1}{5}$ and the probability of being on a specific node in $\bar{V}$ is at least $\frac{1}{5M}$. Then the mixing time of the corresponding Markov chain is of the order $O(MK \log(1/\epsilon))$.*

Proof. The proof follows immediately since any two random walks will intersect with probability $\frac{1}{M}$ every K steps. □

Figure 2.15 shows the cumulative influence of the nodes in each of the graphs, that is, the weight an ordered subset of the nodes has on the final value of the beliefs. In this case, since we are considering a single strongly connected component, the weights are determined by the left-eigenvalue of the weight matrix corresponding to the eigenvector 1. Table 2.3 shows the values for K and M of the graphs studied in this section.

Figure 2.16 shows the convergence time of a belief system when the network of agents is the three large-scale complex networks described in Table 2.2. Results show that the predicted maximum type behavior holds; that is, the convergence time of the belief system is upper bounded by the maximum mixing time of a random walk on the graph of agents and the graph of logic constraints. The convergence time remains constant and of the order of the convergence time of the network of agents, until the mixing time of the network formed by the logic constraints is larger. Then, the total convergence time increases based on the specific topology of the graph of logic constraints.

Table 2.3: Size of the highly influential cliques and the number of iterations required for them to drive the fast mixing of a random walk on the three examples of large scale graphs.

Graph	K	**% of Nodes**	M
`wiki-Vote` [143]	121	10%	25
`ca-GrQc` [144]	365	11%	1700
`ego-Facebook`[145]	365	11%	1829

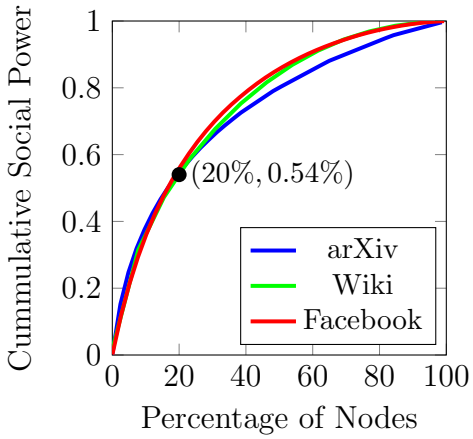


Figure 2.15: Cumulative social power of the agents. Each of the nodes in the graphs considered has some weight in the final value achieved by the belief system. In all three cases, the 20% most important nodes account for 50% of the final value.

Figure 2.17 shows the exponential convergence rate of the belief system. It shows a linear convergence rate of the total variation distance between the beliefs and its limiting value as the number of iteration increases.

2.4 Conclusions

Friedkin et al. [37] proposed a new model that integrates logic constraints into the evolution opinions of a group of agents in a belief system. Logic constraints among truth statements have a significant impact on the evolution of opinion dynamics. Such restrictions can be modeled as graphs that represent the positive or negative influence the beliefs on specific topics have on others. Starting from this context, we have here approached this model from its extended representation of a belief system, where opinions of all agents on all topics as well as their corresponding initial values are nodes in a larger graph. This larger graph is composed of the Kronecker product of the graphs corresponding to the network of agents and the network of logical constraints respectively.

In this chapter, we have provided graph-theoretic arguments for the characterization of the convergence properties of such opinion dynamic models based on extensive existing knowledge of convergence and mixing time of random walks on graphs using the theory of Markov chains. We have shown that convergence occurs if every strongly connected component of the network of logic constraints is non-bipartite and every strongly connected component of oblivious agents is non-bipartite as well. Moreover, to be arbitrarily close

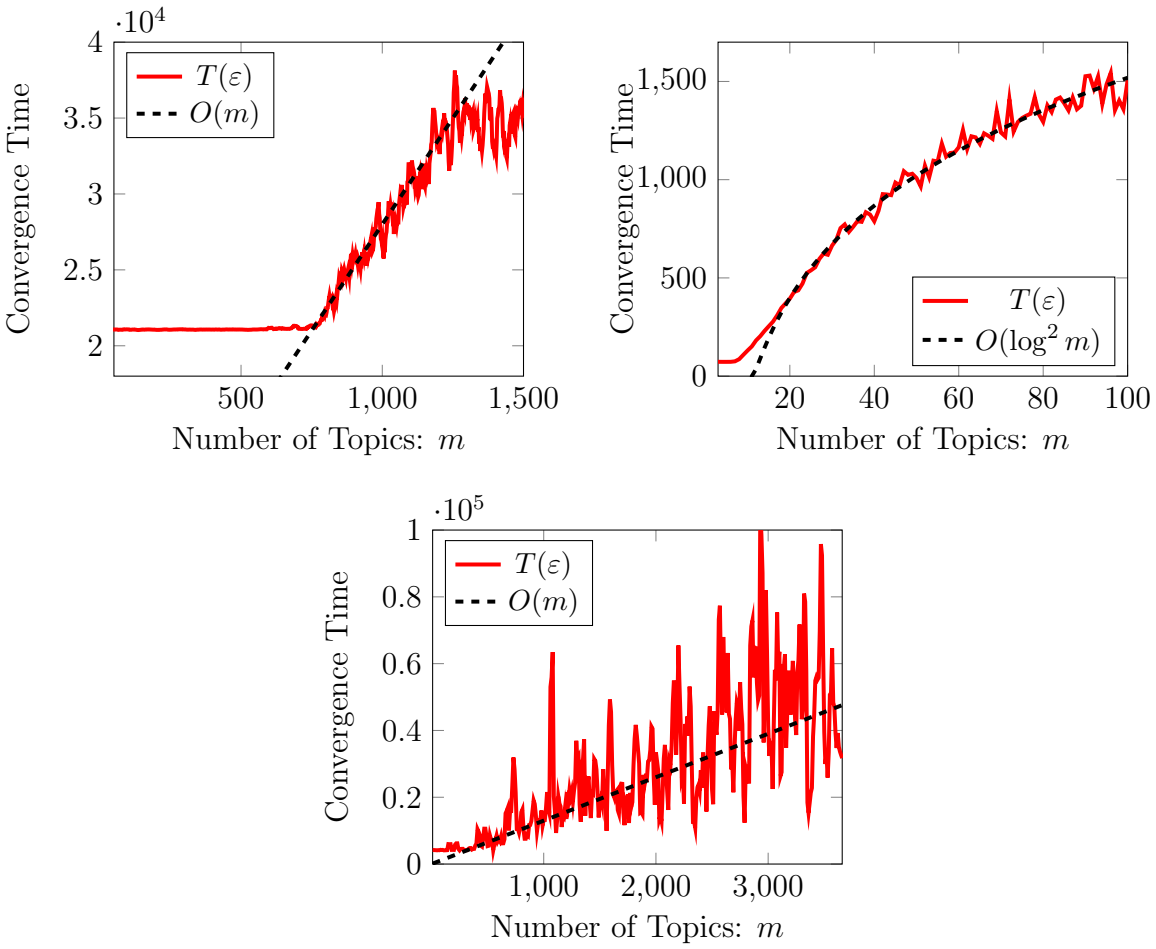


Figure 2.16: Convergence time of belief system with large-scale complex networks. (a) The social network is the `ego-Facebook` graph, and the logic constraints form a complete binary tree with an increasing number of topics. (b) The social network is the `wiki-Vote` graph and the logic constraints form Newman-Watts small-world graph with an increasing number of topics. (c) The social network is the `ca-GrQc` arXiv collaboration graph, and the logic constraints form an Erdős-Rényi graph with an increasing number of topics.

to their limiting value we require $O((L + H)\log(1/\epsilon))$ time steps. The parameter L is the maximum coupling time for a random walk among the closed strongly connected components of the product graph, and H is the maximum time required for a random walk, that starts on an open component, to get absorbed by a closed component. Our analysis applies to broad classes of networks of agents and logic constraints for which we have provided bounds regarding the number of nodes in the graphs. Finally, we show that the limiting opinion value is a convex combination of the nodes in the closed strongly connected components and this convergence happens exponentially fast.

Our framework offers analytical tools that deepen our abilities for modeling, control and

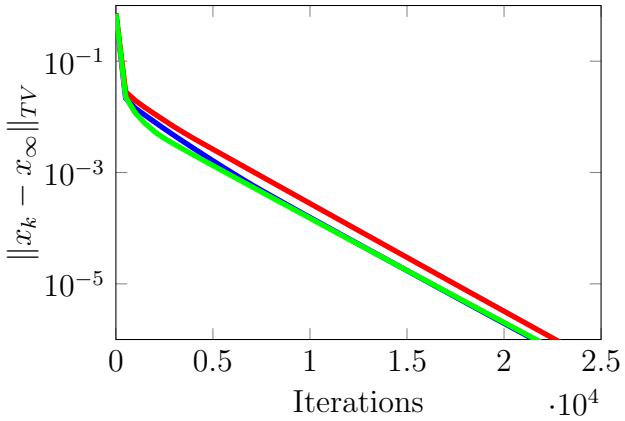


Figure 2.17: Total variation distance between the beliefs and its limiting value as the number of iteration increases. Results are shown for a particular subset of randomly selected agents.

synthesis of complex network systems, particularly man-made, and can inspire further research in domains where opinion formation and networks interact naturally, such as neuroscience and social sciences. Finally, extending this analysis to other opinion formation models that use different aggregating strategies may require further study of Markov processes and random walks.

CHAPTER 3

DISTRIBUTED (NON-BAYESIAN) LEARNING WITH FINITE HYPOTHESES SETS

In this chapter, we begin with a variational analysis of Bayesian posterior and derive an optimization problem for which the posterior is a step of the stochastic mirror descent method. We then use this interpretation to propose a distributed stochastic mirror descent method for distributed learning. We show that this distributed learning algorithm concentrates the beliefs of all agents around the true parameter at an exponential rate. We derive high probability non-asymptotic bounds for the convergence rate.

3.1 Problem Formulation

Initially, we introduce the learning problem from a centralized perspective, where all information is available at a single location. Later, we will generalize the setup to the distributed setting where only partial and distributed information is available.

3.1.1 The Bayesian Approach to Statistical Inference

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a sample space, $\mathcal{F}$ is a σ-algebra and $\mathbb{P}$ a probability measure. Assume that we observe a sequence of independent random variables $X_1, X_2, \ldots$, all taking values in some measurable space $(\mathcal{X}, \mathcal{A})$ and identically distributed with a common *unknown* distribution P on $\mathcal{X}$, i.e. $X_k \sim P$ for all k. In addition, we have a statistical model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ composed by a parametrized family of probability measures on the sample space $(\mathcal{X}, \mathcal{A})$, where the map $\Theta \to \mathscr{P}$ from parameter to distribution is injective. Moreover, all distributions in the model are dominated[1] by a σ-finite measure λ, with corresponding densities $p_\theta = dP_\theta/d\lambda$. Assume also that the model $\mathscr{P}$ is well-specified, thus there exists a θ^* such that $P_{\theta^*} = P$. The objective is to estimate θ^* based on the sequence of received observations $x_1, x_2, \ldots$. For example, the maximum likelihood

[1] A measure μ is dominated by (or absolutely continuous with respect to) a measure λ if $\lambda(B) = 0$ implies $\mu(B) = 0$ for every measurable set B.

estimator (MLE) can be defined as

$$\hat{\theta}(X) = \arg\sup_{\theta\in\theta} p_\theta(X) = \arg\sup_{P\in\mathcal{P}} p(X).$$

Following a Bayesian approach, the parameter is represented as a random variable ϑ on the set Θ that is equipped with a σ-algebra $\mathcal{T}$ and a prior probability measure μ_0 on the measurable space $(\Theta, \mathcal{T})$. Moreover, we assume the existence of a probability measure Π on the product space $(\mathcal{X} \times \Theta)$ with σ-algebra $(\mathcal{A} \times \mathcal{T})$. Therefore one can pair the elements of the parametric model with the conditional distributions $\Pi_{X|\vartheta}$. Furthermore, the densities $p_\theta(x)$ are measurable functions of θ for any $x \in \mathcal{X}$. We then define the belief μ_k as the posterior distribution given the sequence of observations up to time k, i.e.,

$$\mu_k(B) = \Pi(\vartheta \in B|X_1, \ldots, X_k) = \frac{\int_B \prod_{t=1}^k p_\theta(X_t) d\mu_0(\theta)}{\int_\Theta \prod_{t=1}^k p_\theta(X_t) d\mu_0(\theta)}, \tag{3.1}$$

for all $B \in \mathcal{T}$ (note that we used the independence of the observations at each time step).

Assuming that all observations, up to time k, are readily available at a centralized location, under appropriate conditions, the recursive Bayesian posterior in Eq. (3.1) will be consistent in the sense that the beliefs μ_k will concentrate around θ^*; see [148, 149] and [150] for a formal statement. Furthermore, several authors have studied the rate at which this concentration occurs, in both asymptotic and non-asymptotic regimes [151, 152, 153].

3.1.2 The Distributed Statistical Inference Problem

Now, consider the case where there is a network of n agents observing the process $X_1, X_2, \ldots$, where X_k is now a random vector belonging to the product space $\prod_{i=1}^n \mathcal{X}^i$ and $X_k = [X_k^1, X_k^2, \ldots, X_k^n]'$. Specifically, agent i observes the sequence $X_1^i, X_2^i, \ldots$, where X_k^i is now distributed according to an unknown distributions P^i, effectively making $X_k \sim \boldsymbol{P} = \prod_{i=1}^n P^i$. The statistical model is now distributed, where each agent i has a private family of distributions $\mathscr{P}^i = \{P_\theta^i : \theta \in \Theta\}$ it would like to fit to the observations. However, the goal is for *all* agents to agree on a *single* θ that best explains the complete set of observations instead of their local observations only. In other words, the agents collaboratively seek to find θ^* such that $\boldsymbol{P}_{\theta^*} = \prod_{i=1}^n P_{\theta^*}^i = \prod_{i=1}^n P^i = \boldsymbol{P}$.

Agents interact over a network defined by an undirected graph $\mathcal{G} = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of agents and E is a set of undirected edges, i.e., $(i, j) \in E$ if and only if agents i and j can communicate with each other. We study a simple interaction

model where, at each step, agents exchange their beliefs with their neighbors in the graph. Thus at every time step k, agent i will receive the sample x_k^i from X_k^i as well as the beliefs of its neighboring agents, i.e., it will receive μ_{k-1}^j for all j such that $(i,j) \in E$. Applying a fully Bayesian approach runs into some obstacles in this setting, as agents know neither the network topology nor the private family of distributions of other agents. Our goal is to design a learning procedure which is both distributed and consistent. That is, we are interested in a belief update algorithm that aggregates information in a non-Bayesian manner and guarantees that the beliefs of all agents will concentrate around θ^*.

3.2 Bayesian Posterior as Stochastic Mirror Descent

In this section, we observe that the posterior in Eq. (3.1) corresponds to an iteration of a first-order optimization algorithm, namely stochastic mirror descent [154, 155, 156, 157]. Closely related variational interpretations of Bayes' rule are well-known, and in particular have been given in [158, 159, 160]. The specific connection to stochastic mirror descent has not been noted, as far as we are aware. This connection will serve to motivate a distributed learning method which will be the main focus of this chapter.

3.2.1 Bayes' Rule as Stochastic Mirror Descent

Suppose we want to solve the following optimization problem:

$$\min_{\theta \in \Theta} F(\theta) \triangleq D_{KL}(P \| P_\theta), \tag{3.2}$$

where P is an unknown distribution and $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is a parametrized family of distributions. Here, $D_{KL}(P\|Q)$ is the Kullback-Leibler (KL) divergence[2] between distributions P and Q.

First note that we can rewrite the optimization problem in Eq. (3.2) as

$$\begin{aligned}
\min_{\theta \in \Theta} D_{KL}(P\|P_\theta) &= \min_{\pi \in \Delta_\Theta} \mathbb{E}_\pi D_{KL}(P\|P_\vartheta) \qquad \text{where } \vartheta \sim \pi \\
&= \min_{\pi \in \Delta_\Theta} \mathbb{E}_\pi \mathbb{E}_P \left[-\log \frac{dP_\vartheta(X)}{dP(X)} \right] \qquad \text{where } \vartheta \sim \pi, X \sim P,
\end{aligned}$$

[2] $D_{KL}(P\|Q)$ between distributions P and Q (with P dominated by Q) is defined to be $D_{KL}(P\|Q) = -\mathbb{E}_P \left[\log dQ/dP\right]$.

where Δ_Θ is the set of all possible distributions on the parameter space Θ. Since the distribution P does not depend on ϑ, it follows that

$$\begin{aligned}\arg\min_{\pi\in\Delta_\Theta} \mathbb{E}_\pi \mathbb{E}_P \left[-\log \frac{dP_\vartheta(X)}{dP(X)}\right] &= \arg\min_{\pi\in\Delta_\Theta} \mathbb{E}_\pi \mathbb{E}_P \left[-\log p_\vartheta(X)\right] \\ &= \arg\min_{\pi\in\Delta_\Theta} \mathbb{E}_P \mathbb{E}_\pi \left[-\log p_\vartheta(X)\right]. \end{aligned} \tag{3.3}$$

The equality in Eq. (3.3), where we exchange the order of the expectations, follows from the Fubini-Tonelli theorem. Clearly, if θ^* minimizes Eq. (3.2), then a distribution π^* which puts all the mass on θ^* (i.e. $\pi^*(\vartheta = \theta^*) = 1$) minimizes Eq. (3.3).

The difficulty in evaluating the objective function in Eq. (3.3) lies in the fact that the distribution P is unknown. A generic approach to solving such problems is using algorithms from stochastic approximation methods, where the objective is minimized by constructing a sequence of gradient-based iterates whereby the true gradient of the objective (which is not available) is replaced with a gradient sample that is available at a given time.

A particular method that is relevant for the solution of stochastic programs as in Eq. (3.3) is the *stochastic mirror descent* method [161, 155, 154, 162]. The stochastic mirror descent approach constructs a sequence of densities $\{d\mu_k\}$, as follows:

$$d\mu_{k+1} = \arg\min_{\pi\in\Delta_\Theta} \left\{ \langle -\log p_\theta(x_{k+1}), \pi\rangle + \frac{1}{\alpha_k} D_w(\pi, d\mu_k) \right\}, \tag{3.4}$$

where $\alpha_k > 0$ is the step-size, the inner product is defined as $\langle p, q\rangle = \int_\Theta p(\theta)q(\theta)d\sigma$, and $D_w(x, x_k)$ is a Bregman distance function associated with a distance-generating function w, i.e.,

$$D_w(x, z) = w(z) - w(x) - \delta w[z; x - z],$$

where $\delta w[z; x - z]$ is the Fréchet derivative of w at z in the direction of $x - z$. If we choose $w(x) = \int x \log x$ as the distance-generating function, then the corresponding Bregman distance is the Kullback-Leibler (KL) divergence D_{KL}. Additionally, by selecting $\alpha_k = 1$, the solution to the optimization problem in Eq. (3.4) can be computed explicitly, where for each $\theta \in \Theta$,

$$d\mu_{k+1}(\theta) \propto p_\theta(x_{k+1}) d\mu_k(\theta),$$

which is the particular definition for the posterior distribution according to Eq. (3.1) (a

formal proof of this assertion is a special case of Proposition 12 shown later in this chapter).

3.2.2 Entropic Distributed Stochastic Mirror Descent

Now, consider the distributed problem where the network of agents want to collectively solve the following optimization problem:

$$\min_{\theta\in\Theta} F(\theta) \triangleq D_{KL}\left(\boldsymbol{P}\|\boldsymbol{P}_\theta\right) = \sum_{i=1}^{n} D_{KL}(P^i\|P^i_\theta). \tag{3.5}$$

Recall that the distribution $\boldsymbol{P}$ is unknown (though, of course, agents gain information about it by observing samples from $X^i_1, X^i_2, \ldots$ and interacting with other agents) and that $\mathscr{P}^i$ containing all the distributions P^i_θ is a private family of distributions and is only available to agent i.

We propose the following algorithm as a distributed version of the stochastic mirror descent for the solution of problem Eq. (3.5):

$$d\mu^i_{k+1} = \arg\min_{\pi\in\Delta_\Theta} \left\{ \langle -\log p^i_\theta(x^i_{k+1}), \pi\rangle + \sum_{j=1}^{n} a_{ij} D_{KL}(\pi\|d\mu^j_k) \right\} \qquad \text{where } \theta\sim\pi, \tag{3.6}$$

with $a_{ij} > 0$ denoting the weight that agent i assigns to beliefs coming from its neighbor j. Specifically, $a_{ij} > 0$ if $(i,j) \in E$ or $j = i$, and $a_{ij} = 0$ if $(i,j) \notin E$. The optimization problem in Eq. (3.6) has a closed form solution. In particular, the posterior density at each $\theta \in \Theta$ is given by

$$d\mu^i_{k+1}(\theta) \propto p^i_\theta(x^i_{k+1}) \prod_{j=1}^{n} (d\mu^j_k(\theta))^{a_{ij}},$$

or equivalently, the belief on a measurable set B of an agent i at time $k+1$ is

$$\mu^i_{k+1}(B) \propto \int_B p^i_\theta(x^i_{k+1}) \prod_{j=1}^{n} (d\mu^j_k(\theta))^{a_{ij}}. \tag{3.7}$$

We state the correctness of this claim in the following proposition.

Proposition 12. *The probability measure μ^i_{k+1} over the set Θ defined by the update protocol Eq.* (3.7) *coincides, almost everywhere, with the update the distributed stochastic mirror descent algorithm applied to the optimization problem in Eq.* (3.5)*.*

Proof. We need to show that the density $d\mu_{k+1}^i$ associated with the probability measure μ_{k+1}^i defined by Eq. (3.7) minimizes the problem in Eq. (3.6). To do so, let $G(\pi)$ be the objective function for the problem in Eq. (3.6), i.e.,

$$G(\pi) = \langle -\log p_\theta^i(x_{k+1}^i), \pi \rangle + \sum_{j=1}^n a_{ij} D_{KL}(\pi \| d\mu_k^j).$$

Next, we add and subtract the KL divergence between π and the density $d\mu_{k+1}^i$ to obtain

$$\begin{aligned} G(\pi) &= \langle -\log p_\theta^i(x_{k+1}^i), \pi \rangle + \sum_{j=1}^n a_{ij} D_{KL}(\pi \| d\mu_k^j) \\ &\quad - D_{KL}\left(\pi \| d\mu_{k+1}^i\right) + D_{KL}\left(\pi \| d\mu_{k+1}^i\right) \\ &= \langle -\log p_\theta^i(x_{k+1}^i), \pi \rangle + D_{KL}\left(\pi \| d\mu_{k+1}^i\right) + \sum_{j=1}^n a_{ij} \mathbb{E}_\pi \log \frac{d\mu_{k+1}^i}{d\mu_k^j}. \end{aligned}$$

Now, from Eq. (3.7) it follows that

$$\begin{aligned} G(\pi) &= \langle -\log p_\theta^i(x_{k+1}^i), \pi \rangle + D_{KL}\left(\pi \| d\mu_{k+1}^i\right) + \\ &\qquad \sum_{j=1}^n a_{ij} \mathbb{E}_\pi \log \left(\frac{1}{d\mu_k^j} \frac{1}{Z_{k+1}^i} \prod_{l=1}^n \left(d\mu_k^l\right)^{a_{il}} p_\theta^i(x_{k+1}^i) \right) \\ &= \langle -\log p_\theta^i(x_{k+1}^i), \pi \rangle + D_{KL}\left(\pi \| d\mu_{k+1}^i\right) \\ &\qquad - \log Z_{k+1}^i + \langle \log p_\theta^i(x_{k+1}^i), \pi \rangle + \sum_{j=1}^n a_{ij} \mathbb{E}_\pi \log \left(\frac{1}{d\mu_k^j} \prod_{l=1}^n \left(d\mu_k^l\right)^{a_{il}} \right) \\ &= -\log Z_{k+1}^i + D_{KL}\left(\pi \| d\mu_{k+1}^i\right) - \sum_{j=1}^n a_{ij} \mathbb{E}_\pi \log d\mu_k^j + \sum_{l=1}^n a_{il} \mathbb{E}_\pi \log d\mu_k^l \\ &= -\log Z_{k+1}^i + D_{KL}\left(\pi \| d\mu_{k+1}^i\right), \end{aligned} \tag{3.8}$$

where $Z_{k+1}^i = \int_\theta p_\theta^i(x_{k+1}^i) \prod_{j=1}^n (d\mu_k^j(\theta))^{a_{ij}}$ is the corresponding normalizing constant.

The first term in Eq. (3.8) does not depend on the distribution π. Thus, we conclude that the solution to the problem in Eq. (3.6) is the density $\pi^* = d\mu_{k+1}^i$ as defined in Eq. (3.7) (almost everywhere).

□

We remark that the update in Eq. (3.7) can be viewed as a two-step process: first, every agent constructs an aggregate belief using a weighted geometric average of its own belief and the beliefs of its neighbors, and then each agent performs a Bayes' update using the

aggregated belief as a prior. We note that similar arguments in the context of distributed optimization have been proposed in [157, 163] for general Bregman distances. In the case when the number of hypotheses is finite, variations on this update rule were previously analyzed in [45, 25, 46].

Example 1 (Distributed Bernoulli Filter)**.** *Consider a group of* 4 *agents, connected over a network as shown in Fig. 3.1. A set of metropolis weights for this network is given by the following matrix:*

$$A = \begin{bmatrix} 2/3 & 1/6 & 0 & 1/6 \\ 1/6 & 2/3 & 1/6 & 0 \\ 0 & 1/6 & 2/3 & 1/6 \\ 1/6 & 0 & 1/6 & 2/3 \end{bmatrix}.$$

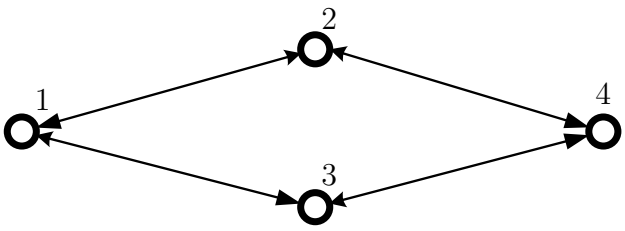


Figure 3.1: A network of 4 agents.

Furthermore, assume that each agent is observing a Bernoulli random variable such that $X_k^1 \sim Bern(0.2)$*,* $X_k^2 \sim Bern(0.4)$*,* $X_k^3 \sim Bern(0.6)$ *and* $X_k^4 \sim Bern(0.8)$*. In this case, the parameter space is* $\Theta = [0, 1]$*. Thus, the objective is to collectively find a parameter* θ^* *that best explains the joint observations in the sense of the problem in Eq.* (3.5)*, i.e.*

$$\begin{aligned} \min_{\theta \in [0,1]} F(\theta) &= \sum_{j=1}^{4} D_{KL}(Bern(\theta^j) \| Bern(\theta)) \\ &= \sum_{j=1}^{4} \left(\theta \log \frac{\theta}{\theta^j} + (1-\theta) \log \frac{1-\theta}{1-\theta^j} \right), \end{aligned}$$

where $\theta^1 = 0.2$*,* $\theta^2 = 0.4$*,* $\theta^3 = 0.6$ *and* $\theta^4 = 0.8$*. We can see that the optimal solution is* $\theta^* = 0.5$ *by determining it explicitly via the first-order optimality conditions or by exploiting the symmetry in the objective function.*

To summarize, we have given an interpretation of Bayes' rule as an instance of stochastic mirror descent. We have shown how this interpretation motivates a distributed update rule.

In the next section, we discuss explicit forms of this update rule for parametric models coming from exponential families.

Distribution P^i's are unknown. Therefore agents try to "learn" the solution to this optimization problem based on local observations and interactions, see Fig. 3.2.

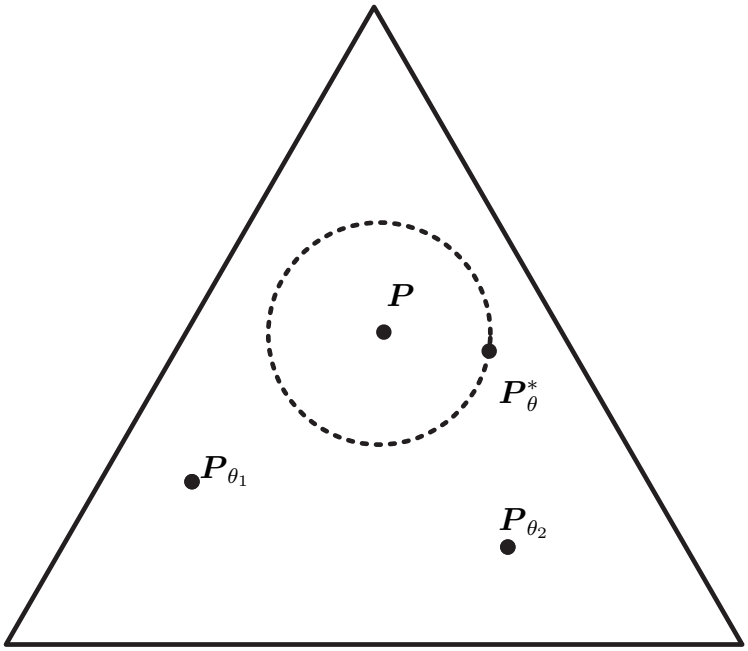


Figure 3.2: Geometric interpretation of the learning objective. The triangle represents the probability simplex; observations of the agents are generated according to a joint probability distribution $\boldsymbol{P}$.

Moreover, agents interact over a sequence of graphs $\{\mathcal{G}_k\}$ where $\mathcal{G} = \{V, E_k\}$, with $V = \{1, 2, \cdots, n\}$ being the set of agents where each agent is denoted as a node, and E_k being the set of edges where $(j, i) \in E_k$ if agent j can communicate with node i at time instant k. Specifically, agents communicate with each other by sharing their beliefs about the hypotheses set, denoted as μ^i_k, which is a probability distribution over Θ.

Next, we introduce three algorithms. The first is a version of Eq. (3.7) for finite random variables with beliefs dominated by the counting measure and agents interacting over *undirected time-varying graphs*. Then, for *fixed graphs* we develop an additional algorithm that scales better with respect to the number of agents. Finally, we propose an algorithm that works on *time-varying directed graphs*. For the first algorithm, we show consistency. Then for the three proposed algorithms, we show explicit, non-asymptotic, and geometric concentration rates of the beliefs on the correct hypotheses.

3.3 Distributed Learning over Time-Varying Undirected Graphs

For agents interacting over undirected time-varying graphs, we consider the following rule which is a specific version of Eq. (3.7) when $\Theta = \{\theta_1, \ldots, \theta_m\}$ is a finite set:

For each $\theta \in \Theta$,

$$\mu^i_{k+1}(\theta) = \frac{1}{Z^i_{k+1}} \prod_{j=1}^{n} \mu^j_k(\theta)^{[A_k]_{ij}} p^i_\theta(x^i_{k+1})^{\beta^i_k}, \tag{3.9}$$

where Z^i_{k+1} is a normalization factor to make the beliefs a probability distribution, i.e.,

$$Z^i_{k+1} = \sum_{r=1}^{m} \prod_{j=1}^{n} \mu^j_k(\theta_r)^{[A_k]_{ij}} p^i_{\theta_r}(x^i_{k+1})^{\beta^i_k},$$

where the A_k is a non-negative matrix of "weights", which is compliant with the connectivity structure of the underlying communication network. The network at each time instant k is modeled as a graph $\mathcal{G}_k$ composed by a node set $V = \{1, 2, \ldots, n\}$ and a set E_k of undirected links. The variable β^i_k is a stationary Bernoulli random process with mean q^i, which indicates if an agent obtained a new realization of X^i_{k+1}. Specifically, $\beta^i_k = 1$ indicates that agent i obtained a new observation, while $\beta^i_k = 0$ indicates that it did not.

3.3.1 Preliminaries

Next, we provide three important definitions that we use in the sequel to describe some learning-related quantities.

Definition 13. *The group confidence of a nonempty subset $W \subseteq V$ of agents is given by*

$$\mathsf{C}^W_{\boldsymbol{q}}(\theta) = -\sum_{i \in W} q^i D_{KL}\left(P^i \| P^i_\theta\right) \qquad \text{for all } \theta \in \Theta,$$

where q^i is the mean-value of the i.i.d. Bernoulli variable β^i_k characterizing the availability of measurements for agent i. If $W = V$, we simply write $\mathsf{C}_{\boldsymbol{q}}$.

The group confidence provides a way to quantify the quality of a hypothesis from the perspective of a subset of the agents. The quality of a hypothesis for individual agents is weighted by the mean of the i.i.d. Bernoulli process governing the availability of observations.

Definition 14. *Two distinct hypotheses θ_i and θ_j are said to be W-observationally equivalent if $\mathsf{C}^W_{\boldsymbol{q}}(\theta_i) = \mathsf{C}^W_{\boldsymbol{q}}(\theta_j)$.*

This definition extends the idea of observational equivalence introduced in [1]. Group observational equivalence provides a general definition where a group of agents cannot differentiate between two hypotheses even if their corresponding likelihood models are not the same.

Finally, we introduce the optimal set of hypotheses as the set with the maximum group confidence.

Definition 15. *The optimal hypothesis set is defined as* $\Theta^* = \arg\max_{\theta\in\Theta} \mathsf{C}_{\boldsymbol{q}}(\theta)$*, and the confidence of the optimal hypothesis set is denoted as* $\mathsf{C}^*_{\boldsymbol{q}}$*, i.e.,* $\mathsf{C}^*_{\boldsymbol{q}} = \mathsf{C}_{\boldsymbol{q}}(\theta^*)$ *for* $\theta^* \in \Theta^*$*.*

The optimal set is always non-empty, and we assume it is a strict subset of Θ to avoid the trivial case where all hypotheses are observationally equivalent. This holds if there is a unique true state, $\hat{\theta} \in \Theta$, such that each agent i sees distributions generated according to $P^i = P^i_{\hat{\theta}}$, and Θ contains other hypotheses besides $\hat{\theta}$.

Informally, we will refer to our assumptions above as describing a setup with *conflicting models*; by this, we mean that the hypothesis which best describes the observations of agent i (i.e., the hypothesis θ which minimizes $D_{\mathrm{KL}}(P^i \| P^i_\theta)$) may not be the hypothesis which best describes the observations of a different agent, and may in fact not belong to the optimal set Θ^*.

We will further require the following assumption on the agents' prior distributions and likelihood functions. The first of these is sometimes referred to as the zero probability property [8].

Assumption 4. *For all agents* $i = 1, \ldots, n$*,*

(a) The set $\hat{\Theta}^* = \cap_{i=1}^n \Theta^{*i}$ *is nonempty, where* $\Theta^{*i} \subseteq \Theta^*$ *is the subset of optimal hypotheses with positive initial beliefs for agent* i*, i.e.,* $\mu^i_0(\theta) > 0$ *for all* $\theta \in \Theta^{*i}$ *and* $\mu^i_0(\theta) = 0$ *for all* $\theta \in \Theta^* \setminus \Theta^{*i}$*.*

(b) The support of the true distribution of the observations is contained in the support of the likelihood models for all hypothesis, i.e., there exists an $\alpha > 0$ *such that if* $P^i(x) > 0$ *then* $P^i_\theta(x) > \alpha$ *for all* $\theta \in \Theta$*.*

Uniform prior beliefs satisfy the Assumption 4(a), which is a reasonable assumption if there is no initial information about the hypotheses quality. In Eq. (3.9), if $\mu^i_k(\theta) = 0$ for some hypothesis θ and for some agent i, at some instance k, then all beliefs of all agents will eventually become zero at that hypothesis. Assumption 4(a) removes the undesired effects of this property which could lead to the inability to learn. Also, Assumption 4(b) guarantees the sub-Gaussian behavior of the observed random variables. Specifically, the

derived convergence rates use results from the measure concentration of random variables. In the most common setting, the random variables must have a sub-Gaussian or sub-exponential behavior [164].

3.3.2 Consistency

The next theorem shows that the dynamics in Eq. (3.9) concentrate the beliefs on the optimal set Θ^*, which is precisely the set that best describes the observations.

Theorem 13. *Under Assumptions 1 and 4, the update rule of Eq. (3.9) has the following property:*

$$\lim_{k\to\infty} \mu_k^i(\theta) = 0 \quad a.s. \quad \textit{for all } \theta \notin \hat{\Theta}^*,\ i = 1, \ldots, n.$$

Next, we present a result regarding the weighted average of random variables with a finite variance.

Lemma 14. *Assume that the graph sequence $\{\mathcal{G}_k\}$ satisfies Assumption 1. Also, let Assumption 4 hold. Then, for $\theta_v \notin \Theta^*$ and $\theta_w \in \hat{\Theta}^*$,*

$$\lim_{k\to\infty} \left(\frac{1}{k} \sum_{t=1}^{k} A_{k:t} \mathcal{L}_t^{\theta_v,\theta_w} + \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \boldsymbol{H}(\theta_v, \theta_w) \right) = 0 \ a.s. \tag{3.10}$$

where $\mathcal{L}_t^{\theta_v,\theta_w}$ is a random vector with coordinates given by

$$[\mathcal{L}_t^{\theta_v,\theta_w}]_i = \beta_{t-1}^i \log \frac{p_{\theta_v}^i(X_t^i)}{p_{\theta_w}^i(X_t^i)} \qquad \forall i = 1, \ldots, n,$$

while the vector $\boldsymbol{H}(\theta_v, \theta_w)$ has coordinates given by

$$H^i(\theta_v, \theta_w) = q^i \left(D_{KL}(P^i \| P_{\theta_v}^i) - D_{KL}\left(P^i \| P_{\theta_w}^i\right) \right).$$

Proof. Adding and subtracting $\frac{1}{k}\sum_{t=1}^{k} \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \mathcal{L}_t^{\theta_v,\theta_w}$ to the expression under the limit in Eq. (3.10) yields

$$\begin{aligned} &\frac{1}{k} \sum_{t=1}^{k} A_{k:t} \mathcal{L}_t^{\theta_v,\theta_w} + \frac{1}{k} \sum_{t=1}^{k} \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \boldsymbol{H}(\theta_v, \theta_w) = \\ &\qquad \frac{1}{k} \sum_{t=1}^{k} \left(A_{k:t} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right) \mathcal{L}_t^{\theta_v,\theta_w} + \frac{1}{k} \sum_{t=1}^{k} \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \left(\mathcal{L}_t^{\theta_v,\theta_w} + \boldsymbol{H}(\theta_v, \theta_w) \right). \end{aligned} \tag{3.11}$$

By Lemma 1, $\lim_{k\to\infty} A_{k:t} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ for all $t \geq 0$. Moreover, by Assumption 4(b), we have that $\log\alpha \leq [\mathcal{L}_t^{\theta_v,\theta_w}]_i \leq \log\frac{1}{\alpha}$. Thus, the first term on the right-hand side of Eq. (3.11) goes to zero a.s. as we take the limit over $k \to \infty$.

Regarding the second term on the right side of Eq. (3.11), by the definition of the KL divergence, and the assumption of each β_t^i being independent, we have that

$$\begin{aligned}
\mathbb{E}\left[\beta_{t-1}^i \log\frac{p_{\theta_v}^i(x_t^i)}{p_{\theta_w}^i(x_t^i)}\right] &= q^i \sum_{x\in\mathcal{X}^i} p^i(x)\log\frac{p_{\theta_v}^i(x)}{p_{\theta_w}^i(x)} \\
&= q^i \sum_{x\in\mathcal{X}^i} p^i(x)\log\left(\frac{p_{\theta_v}^i(x)}{p_{\theta_w}^i(x)}\frac{p^i(x)}{p^i(x)}\right) \\
&= q^i\left(\sum_{x\in\mathcal{X}^i} p^i(x)\log\left(\frac{p^i(x)}{p_{\theta_w}^i(x)}\right) - \sum_{x\in\mathcal{X}^i} p^i(x)\log\left(\frac{p^i(x)}{p_{\theta_v}^i(x)}\right)\right) \\
&= q^i\left(D_{KL}\left(P^i\|P_{\theta_w}^i\right) - D_{KL}\left(P^i\|P_{\theta_v}^i\right)\right),
\end{aligned}$$

or equivalently

$$\mathbb{E}[\mathcal{L}_t^{\theta_v,\theta_w}] = -\boldsymbol{H}(\theta_v,\theta_w).$$

Kolmogorov's strong law of large numbers states that if $\{X_t\}$ is a sequence of independent random variables with variances such that $\sum_{k=1}^{\infty}\frac{\mathrm{Var}(X_k)}{k^2} < \infty$, then $\frac{1}{n}\sum_{k=1}^n X_k - \frac{1}{n}\sum_{k=1}^n \mathbb{E}[X_k] \to 0$ a.s. Let $X_t = \frac{1}{n}\mathbf{1}_n'\mathcal{L}_t^{\theta_v,\theta_w}$, then by Assumption 4(b), it can be seen that $\sup_{t\geq 0}\mathrm{Var}\,(X_t) < \infty$. The final result follows by Lemma 1 and Kolmogorov's strong law of large numbers. □

Lemma 14 provides the necessary results to complete the proof of Theorem 13.

Proof. (*Theorem 13*) Initially, define the following quantities: for all $i = 1,\ldots,n$ and $k \geq 0$,

$$\varphi_k^i(\theta_v,\theta_w) \triangleq \log\frac{\mu_k^i(\theta_v)}{\mu_k^i(\theta_w)}, \tag{3.12}$$

defined for any $\theta_v \notin \hat{\Theta}^*$ and $\theta_w \in \hat{\Theta}^*$. We also use these quantities later in the proof of Theorem 17.

Let agent i be arbitrary and consider the update rule of Eq. (3.9). We will show that $\mu_k^i(\theta_v) \to 0$ as $k \to \infty$ for all $i = 1,\ldots,n$. Note that if $\theta_v \in \Theta^* \setminus \hat{\Theta}^*$, then as a consequence of Assumption 4(a) we have that $\mu_k^i(\theta_v) = 0$ for all i and large enough k. Thus, we consider the case when $\theta_v \notin \Theta^*$ in the remainder of this proof.

Using the definition of $\varphi_k^i(\theta_v, \theta_w)$, it follows from Eq. (3.9) that

$$\begin{aligned}\varphi_{k+1}^i(\theta_v, \theta_w) &= \log \frac{\mu_{k+1}^i(\theta_v)}{\mu_{k+1}^i(\theta_w)} \\ &= \log \frac{\prod_{j=1}^n \mu_k^j(\theta_v)^{[A_k]_{ij}} p_{\theta_v}^i(X_{k+1}^i)^{\beta_k^i}}{\prod_{j=1}^n \mu_k^j(\theta_w)^{[A_k]_{ij}} p_{\theta_w}^i(X_{k+1}^i)^{\beta_k^i}} \\ &= \sum_{j=1}^n [A_k]_{ij} \varphi_k^j(\theta_v, \theta_w) + \beta_k^i \log \frac{p_{\theta_v}^i(X_{k+1}^i)}{p_{\theta_w}^i(X_{k+1}^i)}.\end{aligned}$$

Stacking up the values $\varphi_{k+1}^i(\theta_v, \theta_w)$ for $i = 1, \ldots, n$, into a single vector $\boldsymbol{\varphi}_{k+1}(\theta_v, \theta_w)$, we can compactly write the preceding relations, as follows:

$$\boldsymbol{\varphi}_{k+1}(\theta_v, \theta_w) = A_k \boldsymbol{\varphi}_k(\theta_v, \theta_w) + \mathcal{L}_{k+1}^{\theta_v, \theta_w}, \tag{3.13}$$

where $\mathcal{L}_{k+1}^{\theta_v, \theta_w}$ is defined in the statement of Lemma 14. Now, the relation in Eq. (3.13) implies that for all $k \geq 0$,

$$\boldsymbol{\varphi}_{k+1}(\theta_v, \theta_w) = A_{k:0} \boldsymbol{\varphi}_0(\theta_v, \theta_w) + \sum_{t=1}^k A_{k:t} \mathcal{L}_t^{\theta_v, \theta_w} + \mathcal{L}_{k+1}^{\theta_v, \theta_w}. \tag{3.14}$$

The, if we add and subtract $\sum_{t=1}^k \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \boldsymbol{H}(\theta_v, \theta_w)$ in Eq. (3.14), where $\boldsymbol{H}(\theta_v, \theta_w)$ is as in Lemma 14, it follows that

$$\begin{aligned}\boldsymbol{\varphi}_{k+1}(\theta_v, \theta_w) = A_{k:0} \boldsymbol{\varphi}_0(\theta_v, \theta_w) &- \frac{k}{n} \sum_{i=1}^n H^i(\theta_v, \theta_w) \mathbf{1}_n \\ &+ \sum_{t=1}^k \left(A_{k:t} \mathcal{L}_t^{\theta_v, \theta_w} + \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \boldsymbol{H}(\theta_v, \theta_w) \right) + \mathcal{L}_{k+1}^{\theta_v, \theta_w}.\end{aligned}$$

By the definition of group confidence (cf. Definition 13), we have

$$\sum_{i=1}^n H^i(\theta_v, \theta_w) = \mathsf{C}_{\boldsymbol{q}}(\theta_w) - \mathsf{C}_{\boldsymbol{q}}(\theta_v) = \mathsf{C}_{\boldsymbol{q}}^* - \mathsf{C}_{\boldsymbol{q}}(\theta_v), \tag{3.15}$$

where the last equality follows from $\theta_w \in \hat{\Theta}^*$ and the definition of the optimal value $\mathsf{C}_{\boldsymbol{q}}^*$ (Definition 15). Therefore,

$$\boldsymbol{\varphi}_{k+1}(\theta_v, \theta_w) = A_{k:0} \boldsymbol{\varphi}_0(\theta_v, \theta_w) - \frac{k}{n} \left(\mathsf{C}_{\boldsymbol{q}}^* - \mathsf{C}_{\boldsymbol{q}}(\theta_v) \right) \mathbf{1}_n$$

$$+ \sum_{t=1}^{k} \left(A_{k:t} \mathcal{L}_t^{\theta_v, \theta_w} + \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \boldsymbol{H}(\theta_v, \theta_w) \right) + \mathcal{L}_{k+1}^{\theta_v, \theta_w}.$$

By dividing both sides of the preceding equation by k and taking the limit as k goes to infinity, almost surely we have

$$\lim_{k\to\infty} \frac{1}{k} \boldsymbol{\varphi}_{k+1}(\theta_v, \theta_w) = \lim_{k\to\infty} \frac{1}{k} A_{k:0} \boldsymbol{\varphi}_0(\theta_v, \theta_w) + \lim_{k\to\infty} \frac{1}{k} \sum_{t=1}^{k} \left(A_{k:t} \mathcal{L}_t^{\theta_v, \theta_w} + \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \boldsymbol{H}(\theta_v, \theta_w) \right)$$
$$+ \lim_{k\to\infty} \frac{1}{k} \mathcal{L}_{k+1}^{\theta_v, \theta_w} - \frac{1}{n} \left(\mathsf{C}_{\boldsymbol{q}}^* - \mathsf{C}_{\boldsymbol{q}}(\theta_v) \right) \mathbf{1}_n. \tag{3.16}$$

The limit on the left-hand side of Eq. (3.16) is justified since all the limits on the right-hand side exist. Specifically, the first term on the right-hand side of Eq. (3.16) converges to zero deterministically. The second term converges to zero almost surely by Lemma 14, while the third term goes to zero since $\mathcal{L}_t^{\theta_v, \theta_w}$ is bounded almost surely (cf. Assumption 4(b)).

Consequently,

$$\lim_{k\to\infty} \frac{1}{k} \boldsymbol{\varphi}_{k+1}(\theta_v, \theta_w) = -\frac{1}{n} \left(\mathsf{C}_{\boldsymbol{q}}^* - \mathsf{C}_{\boldsymbol{q}}(\theta_v) \right) \mathbf{1}_n \quad \text{a.s.}$$

Since $\mathsf{C}_{\boldsymbol{q}}^*$ is the maximum value and $\theta_v \notin \Theta^*$, it follows that $\mathsf{C}_{\boldsymbol{q}}^* - \mathsf{C}_{\boldsymbol{q}}(\theta_v) > 0$, implying that $\boldsymbol{\varphi}_k(\theta_v, \theta_w) \to -\infty$ almost surely. Also, by $\mu_k^i(\theta_v) \leq \exp\left(\varphi_k^i(\theta_v, \theta_w)\right)$ for all i, we have $\mu_k^i(\theta_v) \to 0$ *a.s.* □

One specific instance of our setup is when there exists a unique hypothesis that matches the distribution of the observations of all agents. This case relates to the previously proposed approaches for distributed learning. Specifically, in [46, 43, 1], the authors assume that there is a "true state" of the world, i.e., there is a unique hypothesis such that the distance between such hypothesis and the true distribution of the data is zero for all agents. This case could be expressed, as a consequence of Theorem 13, as follows:

Corollary 15. *Under assumptions of Theorem 13, if there is a unique hypothesis θ^* with $C_q^* = 0$, then*

$$\lim_{k\to\infty} \mu_k^i(\theta^*) = 1 \quad a.s. \quad \forall i \in V.$$

Proof. By Theorem 13 for every $\theta \neq \theta^*$ we have that $\lim_{k\to\infty} \mu_k^i(\theta) = 0$ a.s. □

In general, one can consider several closed social *cliques* where the same hypothesis can represent different distributions for different groups. For example, in a social network, what one community might consider as a good hypothesis, need not be good for other communities. Each disconnected social *clique* could have a different optimal hypothesis, even if all observations come from the same distribution, see Fig. 3.3. If such social cliques interact, Theorem 13 provides the conditions for which all agents will agree on the hypothesis that is the closest to the best one considering the models of all agents in the network and not only those in a specific *clique*.

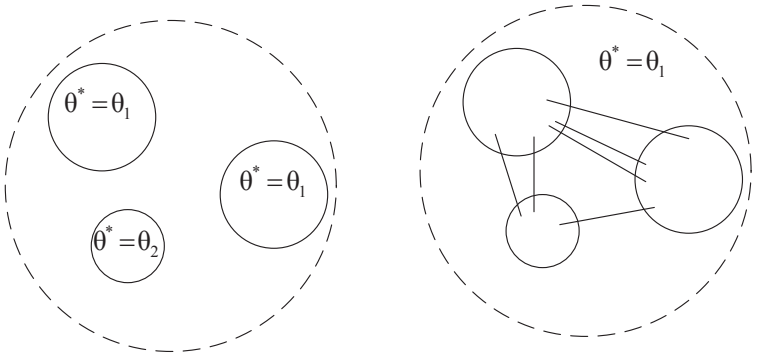


Figure 3.3: Conflicting social groups interacting. Initially, on the left, there are three isolated social cliques, each with a different optimal hypothesis. Once such groups interact (on the right), others might influence the local decision, and a clique changes its beliefs to the optimal with respect to the complete set of agents. In this case, one of the groups was convinced that θ_1 was a better solution than θ_2.

The previous statement is formally stated in the next corollary.

Corollary 16. *Let the agent set V be partitioned into $\hat{p}$ disjoint sets $V_j, j = 1, \ldots, \hat{p}$. Under assumptions of Theorem 13 where each agent updates its beliefs according to Eq.* (3.9)*, if there exists a hypothesis θ^* such that*

$$\sum_{j=1}^{\hat{p}} C_q^{V_j}(\theta^*) > \max_{\theta \neq \theta^*} \sum_{j=1}^{\hat{p}} C_q^{V_j}(\theta),$$

then $\lim_{k\to\infty} \mu_k^i(\theta^*) = 1$ a.s. for all i.

Proof. If the hypothesis θ^* exists, then the group confidence on θ^* is larger than the group confidence for any other hypothesis. Thus, $\hat{\Theta}^* = \{\theta^*\}$ and the result follows by Theorem 13. □

3.3.3 Convergence Rate for Time-Varying Undirected Graphs

This subsection presents our result regarding the non-asymptotic explicit convergence rate of the update rules in Eq. (3.9).

Theorem 17. *Let Assumptions 1 and 4 hold and let $\rho \in (0,1)$. The update rule of Eq. (3.9) has the following property: There is an integer $\boldsymbol{N}(\rho)$ such that, with probability $1-\rho$, for all $k \geq \boldsymbol{N}(\rho)$ and for all $\theta_v \notin \Theta^*$, we have*

$$\mu_k^i(\theta_v) \leq \exp\left(-\frac{k}{2}\gamma_2 + \gamma_1^i\right) \quad \textit{for all } i = 1, \ldots, n,$$

where

$$\begin{aligned}
\boldsymbol{N}(\rho) &\triangleq \left\lceil \frac{1}{\gamma_2^2} 8 \left(\log \alpha\right)^2 \log \frac{1}{\rho} \right\rceil, \\
\gamma_1^i &\triangleq \max_{\substack{\theta_w \in \hat{\Theta}^* \\ \theta_v \notin \Theta^*}} \left\{ \max_i \log \frac{\mu_0^i(\theta_v)}{\mu_0^i(\theta_w)} \right\} + \frac{12 \log n}{1-\lambda} \log \frac{1}{\alpha}, \\
\gamma_2 &\triangleq \frac{1}{n} \min_{\theta_v \notin \Theta^*} \left(\mathsf{C}_{\boldsymbol{q}}^* - \mathsf{C}_{\boldsymbol{q}}(\theta_v) \right),
\end{aligned}$$

with α from Assumption 4(b), η from Assumption 1(d) and λ given by:

$$\lambda = \left(1 - \frac{\eta}{4n^2}\right)^{\frac{1}{B}}.$$

If each A_k is the lazy Metropolis matrix associated with $\mathcal{G}_k$ and $B = 1$, then

$$\lambda = 1 - \frac{1}{\mathcal{O}(n^2)}.$$

In words, the belief of each agent on any hypothesis outside the optimal set decays at a network-independent rate which scales with the constant γ_2, which is the average Kullback-Leibler divergence to the next best hypothesis. However, there is a transient due to the γ_1^i term (since the bound of Theorem 17 is not below 1 until $k \geq 2\gamma_1^i/\gamma_2$), and the size of this transient depends on the network and the number of nodes through the constant λ.

Observe that the term γ_1^i represents the influence of the initial beliefs as well as the mixing properties of the graph. If all agents use uniform initial beliefs, i.e., $\mu_0^i \equiv 1/|\Theta|$, then the effect of the initial beliefs is zero and γ_1^i reduces to

$$\gamma_1^i = \frac{12 \log n}{1-\lambda} \log \frac{1}{\alpha},$$

where the constant λ may be thought of as the "time to ergodicity" of the inhomogeneous Markov chain associated with the matrix sequence A_k. On the other hand, if one can start with an informative prior where $\mu_0^i(\theta^*) > \mu_0^i(\theta)$, the influence of the initial beliefs will be a negative term, effectively reducing the transient time.

Before proving Theorem 17, we will provide an auxiliary result regarding bounds on the expectation of the random variables $\varphi_k^i(\theta_v, \theta_w)$ as defined in Eq. (3.12).

Lemma 18. *Consider $\varphi_k^i(\theta_v, \theta_w)$ as defined in Eq. (3.12), with $\theta_w \in \hat{\Theta}^*$. Then, for any $\theta_v \notin \Theta^*$ we have*

$$\mathbb{E}[\varphi_k^i(\theta_v, \theta_w)] \leq \gamma_1^i - k\gamma_2, \quad \textit{for all } i \textit{ and } k \geq 0,$$

with γ_1^i and γ_2 as defined in Theorem 17.

Proof. Taking the expected value in Eq. (3.14) we can see that for all $k \geq 0$,

$$\mathbb{E}[\varphi_{k+1}^i(\theta_v, \theta_w)] = \sum_{j=1}^{n} [A_{k:0}]_{ij} \varphi_0^j(\theta_v, \theta_w) - \sum_{t=1}^{k} \sum_{j=1}^{n} [A_{k:t}]_{ij} H^j(\theta_v, \theta_w) - H^i(\theta_v, \theta_w).$$

By adding and subtracting $\sum_{t=1}^{k+1} \sum_{j=1}^{n} \frac{1}{n} H^j(\theta_v, \theta_w)$, we obtain

$$\begin{aligned}\mathbb{E}[\varphi_{k+1}^i(\theta_v, \theta_w)] = \sum_{j=1}^{n} [A_{k:0}]_{ij} \varphi_0^j(\theta_v, \theta_w) - \sum_{t=1}^{k} \sum_{j=1}^{n} \left([A_{k:t}]_{ij} - \frac{1}{n} \right) H^j(\theta_v, \theta_w) \\ - \frac{k+1}{n} \sum_{j=1}^{n} H^j(\theta_v, \theta_w) - \left(H^i(\theta_v, \theta_w) - \frac{1}{n} \sum_{j=1}^{n} H^j(\theta_v, \theta_w) \right). \end{aligned} \tag{3.17}$$

For the first term in Eq. (3.17), since $A_{k:0}$ is stochastic matrix, we have that

$$\sum_{j=1}^{n} [A_{k:0}]_{ij} \varphi_0^j(\theta_v, \theta_w) \leq \max_i \log \frac{\mu_0^i(\theta_v)}{\mu_0^i(\theta_w)}.$$

The second term in Eq. (3.17) can be bounded using Lemma 4, thus

$$\sum_{t=1}^{k} \sum_{j=1}^{n} \left([A_{k:t}]_{ij} - \frac{1}{n} \right) H^j(\theta_v, \theta_w) \leq \frac{4 \log n}{1 - \lambda} \log \frac{1}{\alpha},$$

since $\log \alpha \leq H^j(\theta_v, \theta_w) \leq \log \frac{1}{\alpha}$.

The last term in Eq. (3.17) is bounded as

$$\frac{1}{n}\sum_{j=1}^{n}\left(H^i(\theta_v,\theta_w)-H^j(\theta_v,\theta_w)\right) \le 2\log\frac{1}{\alpha} \le \frac{8\log n}{1-\lambda}\log\frac{1}{\alpha},$$

where the last inequality follows from $2 \le 8\log n$ for $n \ge 2$ and $1-\lambda < 1$.

Finally we have that

$$\mathbb{E}[\varphi^i_{k+1}(\theta_v,\theta_w)] \le \max_i \log\frac{\mu^i_0(\theta_v)}{\mu^i_0(\theta_w)} + \frac{12\log n}{1-\lambda}\log\frac{1}{\alpha} - \frac{k+1}{n}\sum_{j=1}^{n} H^j(\theta_v,\theta_w),$$

from which the desired result follows by using the definitions of γ^i_1, γ_2, $H^j(\theta_v,\theta_w)$ and taking the appropriate maximum values over θ_v and θ_w on the right hand side of the preceding inequality. □

Now, we are ready to prove Theorem 17.

Proof. (*Theorem* 17) First, we will express the belief $\mu^i_{k+1}(\theta_v)$ in terms of the variable $\varphi^i_{k+1}(\theta_v,\theta_w)$. This will allow us to use McDiarmid's inequality to obtain the concentration bounds. By the dynamics of the beliefs in Eq. (3.9) and Assumption 4(a), since $\mu^i_k(\theta_w) \in (0,1]$ for $\theta_w \in \hat{\Theta}^*$, we have

$$\mu^i_k(\theta_v) \le \frac{\mu^i_k(\theta_v)}{\mu^i_k(\theta_w)} = \exp\left(\varphi^i_k(\theta_v,\theta_w)\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(\mu^i_k(\theta_v) \ge \exp\left(-\frac{k}{2}\gamma_2+\gamma^i_1\right)\right) &\le \mathbb{P}\left(\exp\left(\varphi^i_k(\theta_v,\theta_w)\right) \ge \exp\left(-\frac{k}{2}\gamma_2+\gamma^i_1\right)\right)\\ &= \mathbb{P}\left(\varphi^i_k(\theta_v,\theta_w) \ge -\frac{k}{2}\gamma_2+\gamma^i_1\right)\\ &\le \mathbb{P}\left(\varphi^i_k(\theta_v,\theta_w) - \mathbb{E}[\varphi^i_k(\theta_v,\theta_w)] \ge \frac{k}{2}\gamma_2\right),\end{aligned}$$

where the last inequality follows from Lemma 18.

We now view $\varphi^i_{k+1}(\theta_v,\theta_w)$ as a function of the random vectors $\boldsymbol{S}_1,\ldots,\boldsymbol{S}_k$ (see Eq. (3.14)), where $\boldsymbol{S}_t = (S^1_t,\ldots,S^n_t)$ for $t \ge 1$, and the random variable S^i_{k+1}. Next, we will establish that this function has bounded differences in order to apply McDiarmid's inequality.

For all t with $1 \leq t \leq k$ and j with $1 \leq j \leq n$, we have

$$\begin{aligned}
&\max_{s_t^j \in \mathcal{S}^j} \varphi_{k+1}^i(\theta_v, \theta_w) - \min_{s_t^j \in \mathcal{S}^j} \varphi_{k+1}^i(\theta_v, \theta_w) \\
&= \max_{s_t \in \mathcal{S}^j} [A_{k:t}]_{ij} \log \frac{p_{\theta_v}^j(x_k)}{p_{\theta_w}^j(x_k)} - \min_{s_t \in \mathcal{S}^j} [A_{k:t}]_{ij} \log \frac{p_{\theta_v}^j(x_k)}{p_{\theta_w}^j(x_k)} \\
&\leq [A_{k:t}]_{ij} \log \frac{1}{\alpha} + [A_{k:t}]_{ij} \log \frac{1}{\alpha} \\
&= 2[A_{k:t}]_{ij} \log \frac{1}{\alpha}.
\end{aligned}$$

Similarly, from Eq. (3.14) we can see that

$$\max_{s_{k+1}^j \in \mathcal{S}^j} \varphi_{k+1}^i(\theta_v, \theta_w) - \min_{s_{k+1}^j \in \mathcal{S}^j} \varphi_{k+1}^i(\theta_v, \theta_w) \leq 2 \log \frac{1}{\alpha}.$$

It follows that $\varphi_{k+1}^i(\theta_v, \theta_w)$ has bounded variations, with

$$\begin{aligned}
\sum_{t=1}^{k} \sum_{j=1}^{n} (2[A_{k:t}]_{ij} \log \frac{1}{\alpha})^2 + \left(2 \log \frac{1}{\alpha}\right)^2 &= 4 \left(\log \frac{1}{\alpha}\right)^2 \left(\sum_{t=1}^{k} \sum_{j=1}^{n} ([A_{k:t}]_{ij})^2 + 1\right) \\
&\leq 4 \left(\log \frac{1}{\alpha}\right)^2 (k+1),
\end{aligned}$$

where the last inequality follows from the fact that $A_{k:t}$ is row stochastic.

Thus,

$$\mathbb{P}\left(\varphi_k^i(\theta_v, \theta_w) - \mathbb{E}[\varphi_k^i(\theta_v, \theta_w)] \geq \frac{k}{2}\gamma_2\right) = \exp\left(-\frac{2\left(\frac{1}{2}k\gamma_2\right)^2}{4k\left(\log \frac{1}{\alpha}\right)^2}\right).$$

Therefore, for a given confidence level ρ, in order to have

$$\mathbb{P}\left(\mu_k^i(\theta_v) \geq \exp\left(-\frac{1}{2}k\gamma_2 + \gamma_1^i\right)\right) \leq \rho,$$

we require that

$$k \geq \frac{1}{\gamma_2^2} 8 \left(\log \alpha\right)^2 \log \frac{1}{\rho}.$$

□

3.4 Distributed Learning over Time-varying Directed Graphs

In this section we present our results for distributed learning over *directed* time-varying graphs. We propose a new algorithm in which every agent i updates its beliefs on a hypothesis θ given some observation x^i_{k+1} following the next protocol

$$y^i_{k+1} = \sum_{j \in N^i_k} \frac{y^j_k}{d^j_k} \tag{3.18a}$$

$$\mu^i_{k+1}(\theta) = \frac{1}{Z^i_{k+1}} \left(\prod_{j \in N^i_k} \mu^j_k(\theta)^{\frac{y^j_k}{d^j_k}} p^i_\theta\left(x^i_{k+1}\right) \right)^{\frac{1}{y^i_{k+1}}}, \tag{3.18b}$$

where at time k: N^i_k is the set of in-neighbors of node i, that is $N^i_k = \{j | (j,i) \in E_k\}$ (a node is assumed to be its own neighbor) and the value d^i_k its the out degree of node i. The term Z^i_{k+1} is the corresponding normalization factor.

The proposed update rule in Eqs. (3.18) is inspired by the push-sum protocol recently studied in [165, 166] and its application to distributed optimization in time-varying directed graphs [167, 72, 168, 169, 170]. At each time step, each node shares its beliefs on the hypothesis set Θ to its out neighbors. Additionally, it also shares a self-assigned weight y^j_k/d^j_k to be used by its neighbors. Then, each node i computes the geometric average of the beliefs of its in-neighbor set with weights corresponding to a normalized version of the self-assigned weights it received. Then a Bayesian update step is performed based on the local observations with a learning rate parameter of $1/y^i_{k+1}$.

In this subsection, we show a non-asymptotic convergence rate for the proposed update rule in Eq. (3.18).

Theorem 19. *Let Assumption 4 hold, and let the sequence $\{\mathcal{G}_k\}$ be B-strongly connected. Also, let $\rho \in (0,1)$ be a given error percentile (or confidence value). Then, the update rule in Eqs. (3.18), with $y^i_0 = 1$ and uniform initial beliefs, has the following property: There is an integer $\boldsymbol{N}(\rho)$ such that, with probability $1-\rho$, for all $k \geq \boldsymbol{N}(\rho)$ and for all $\theta_v \notin \Theta^*$ there holds*

$$\mu^i_k(\theta_v) \leq \exp\left(-\frac{k}{2}\gamma_2 + \frac{1}{\delta}\gamma^i_1\right) \quad \textit{for all } i = 1, \ldots, n,$$

where

$$\boldsymbol{N}(\rho) \triangleq \left\lceil \frac{1}{\delta^2 \gamma_2^2} 8 \left(\log(\alpha)\right)^2 \log\left(\frac{1}{\rho}\right) + 1 \right\rceil,$$

$$\gamma_1^i = \max_{\substack{\theta_w \in \hat{\Theta}^* \\ \theta_v \notin \Theta^*}} \left\{ \frac{2C}{1-\lambda} \| \boldsymbol{H}(\theta_v, \theta_w) \|_1 - [\boldsymbol{H}(\theta_v, \theta_w)]_i \right\},$$
$$\gamma_2 = \frac{1}{n} \min_{\theta_v \notin \Theta^*} (\mathsf{C}^* - \mathsf{C}(\theta_v)),$$

with $\mathsf{C}(\theta)$ *being the group confidence on hypothesis* θ *and* $\mathsf{C}^* = \mathsf{C}(\theta)$ *for all* $\theta \in \Theta^*$ *and the vector* $\boldsymbol{H}(\theta_v, \theta_w)$ *has coordinates given by*

$$[\boldsymbol{H}(\theta_v, \theta_w)]_i = D_{KL}(P^i \| P^i_{\theta_v}) - D_{KL}\left(P^i \| P^i_{\theta_w}\right).$$

The constants C, δ *and* λ *satisfy the following relations:*
(1) For general B*-strongly-connected graph sequences* $\{\mathcal{G}_k\}$*,*

$$C = 4, \quad \lambda = \left(1 - \frac{1}{n^{nB}}\right)^{\frac{1}{B}}, \quad \delta \geq \frac{1}{n^{nB}}.$$

(2) If every graph G_k *is regular with* $B = 1$*,*

$$C = \sqrt{2}, \quad \lambda = \left(1 - \frac{1}{4n^3}\right)^{\frac{1}{B}}, \quad \delta = 1.$$

This theorem shows that the network of agents will collectively solve the optimization problem in Eq. (3.5). After a transient time $\boldsymbol{N}(\rho)$, the belief in the hypothesis outside the optimal hypothesis set, that maximizes the group confidence, will decay exponentially fast. Moreover, this will happen at a rate that depends on explicitly characterized terms γ_1^i and γ_2. Additionally, after a transient time of $2\gamma_1^i/\delta\gamma_2$ for which the beliefs are bounded by 1 the exponential decay will occur at a rate that depends on γ_2 only, i.e., the average difference between the optimal confidence and the second best hypothesis. This exponential rate is network independent and holds for all the nodes in the network.

This result generalizes previously proposed algorithms [25] when the optimal set of the hypothesis is also optimal from the local perspective [64]. Moreover, in contrast with previous literature, the convergence rate induced by the parameter γ_2 does not depend on the parameter δ, that is, after a transient time, the convergence rate is as if the sequence of graphs were regular. Without this regularization behavior, the amount an agent contributes to the group confidence is determined by its location in the network, i.e., δ. Then in the case of time-varying graphs, the importance of the nodes might change as well, and since we allow for disjoint node optimal hypothesis, the concentration of the beliefs would oscillate with the confidence as a weighted sum of local confidences are changing with the topology

of the network.

Remark 1. *If the auxiliary sequence y_k^i is not used in the update rule, i.e.*

$$\mu_{k+1}^i(\theta) = \frac{1}{Z_{k+1}^i} \prod_{j \in N_k^i} \mu_k^j(\theta)^{\frac{1}{d_k^j}} p_\theta^i\left(x_{k+1}^i\right),$$

with the corresponding normalization term Z_{k+1}^i, we obtain a similar result as in Theorem 19 with the exponential rate

$$\mu_k^i(\theta_v) \leq \exp\left(-\delta \frac{k}{2}\gamma_2 + \gamma_1^i\right), \quad \text{for all } i = 1, \ldots, n,$$

with the same constants δ, C, γ_2, γ_1^i and $\boldsymbol{N}(\rho)$. However, after the same transient time $2\gamma_1^i/\delta\gamma_2$, the exponential decay occurs at a rate that depends on $\delta\gamma_2$, where δ might be very small (note that $\delta \geq 1/n^{nB}$).

In this section we analyze the dynamics of the proposed learning rule in Eqs. (3.18). First, define the following quantities that simplify the analysis procedure: For all $i = 1, \ldots, n$ and $k \geq 0$ let

$$\varphi_k^i(\theta_v, \theta_w) \triangleq \log \frac{\mu_k^i(\theta_v)}{\mu_k^i(\theta_w)}, \tag{3.19}$$

$$\hat{\varphi}_k^i(\theta_v, \theta_w) \triangleq y_k^i \varphi_k^i(\theta_v, \theta_w), \tag{3.20}$$

for any $\theta_v \notin \hat{\Theta}^*$ and $\theta_w \in \hat{\Theta}^*$. With this definition in place we can focus on analyzing the dynamics of $\hat{\varphi}_k^i(\theta_v, \theta_w)$.

Proposition 20. *The quantity $\hat{\varphi}_k^i(\theta_v, \theta_w)$ evolves as*

$$\hat{\varphi}_{k+1}^i(\theta_v, \theta_w) = \sum_{j=1}^n \frac{1}{d_k^j} \hat{\varphi}_k^j(\theta_v, \theta_w) + \log \frac{p_{\theta_v}^i(x_{k+1}^i)}{p_{\theta_w}^i(x_{k+1}^i)}. \tag{3.21}$$

Moreover, by staking all $\hat{\varphi}_k^i(\theta_v, \theta_w)$ into a single vector, $\hat{\boldsymbol{\varphi}}_{k+1}(\theta_v, \theta_w)$ can be compactly stated as

$$\hat{\boldsymbol{\varphi}}_{k+1}(\theta_v, \theta_w) = A_k \hat{\boldsymbol{\varphi}}_k(\theta_v, \theta_w) + \mathcal{L}_{k+1}^{\theta_v, \theta_w}, \tag{3.22}$$

where A_k is a stochastic matrix such that

$$[A_k]_{ij} = \begin{cases} \frac{1}{d_k^j} & \text{if } (j,i) \in E_k, \\ 0 & \text{otherwise,} \end{cases}$$

and $\left[\mathcal{L}_{k+1}^{\theta_v,\theta_w}\right]_i = \log \frac{p_{\theta_v}^i(x_{k+1}^i)}{p_{\theta_w}^i(x_{k+1}^i)}$

Proof. By the definitions provided in Eqs. (3.19) and (3.20) we have that

$$\begin{aligned}
\hat{\varphi}_{k+1}^i(\theta_v,\theta_w) &= y_{k+1}^i \varphi_{k+1}^i(\theta_v,\theta_w) \\
&= y_{k+1}^i \log \frac{\mu_{k+1}^i(\theta_v)}{\mu_{k+1}^i(\theta_w)} \\
&= y_{k+1}^i \log \frac{\left(\prod_{j=1}^n \mu_k^j(\theta_v)^{[A_k]_{ij} y_k^j} p_{\theta_v}^i(x_{k+1}^i)\right)^{\frac{1}{y_{k+1}^i}}}{\left(\prod_{j=1}^n \mu_k^j(\theta_w)^{[A_k]_{ij} y_k^j} p_{\theta_w}^i(x_{k+1}^i)\right)^{\frac{1}{y_{k+1}^i}}} \\
&= \log \frac{\left(\prod_{j=1}^n \mu_k^j(\theta_v)^{[A_k]_{ij} y_k^j} p_{\theta_v}^i(x_{k+1}^i)\right)}{\left(\prod_{j=1}^n \mu_k^j(\theta_w)^{[A_k]_{ij} y_k^j} p_{\theta_w}^i(x_{k+1}^i)\right)} \\
&= \sum_{j=1}^n [A_k]_{ij} y_k^j \log \frac{\mu_k^j(\theta_v)}{\mu_k^j(\theta_w)} + \log \frac{p_{\theta_v}^i(x_{k+1}^i)}{p_{\theta_w}^i(x_{k+1}^i)}.
\end{aligned}$$

The first three equalities follow from Eqs. (3.18), (3.19) and (3.20). Cancellation of the term y_{k+1}^i leads to the fourth equality. The rest of the proof follows from arithmetic properties of logarithms. □

We can now proceed to further analyze the sequence $\hat{\varphi}_{k+1}^i(\theta_v,\theta_w)$. First by adding and subtracting the term $\sum_{t=1}^k \boldsymbol{\phi}_k \mathbf{1}' \mathcal{L}_t^{\theta_v,\theta_w}$ from Eq. (3.22) we obtain

$$\begin{aligned}
\hat{\boldsymbol{\varphi}}_{k+1}(\theta_v,\theta_w) &= A_{k:0}\hat{\boldsymbol{\varphi}}_0(\theta_v,\theta_w) + \sum_{t=1}^k A_{k:t}\mathcal{L}_t^{\theta_v,\theta_w} + \mathcal{L}_{k+1}^{\theta_v,\theta_w} - \sum_{t=1}^k \boldsymbol{\phi}_k \mathbf{1}' \mathcal{L}_t^{\theta_v,\theta_w} + \sum_{t=1}^k \boldsymbol{\phi}_k \mathbf{1}' \mathcal{L}_t^{\theta_v,\theta_w} \\
&= A_{k:0}\hat{\boldsymbol{\varphi}}_0(\theta_v,\theta_w) + \sum_{t=1}^k D_{k:t}\mathcal{L}_t^{\theta_v,\theta_w} + \mathcal{L}_{k+1}^{\theta_v,\theta_w} + \sum_{t=1}^k \phi_k \mathbf{1}' \mathcal{L}_t^{\theta_v,\theta_w},
\end{aligned}$$

with $D_{k:t} = A_{k:t} - \boldsymbol{\phi}_k \mathbf{1}'$.

From now on we will ignore the first term in $\hat{\boldsymbol{\varphi}}_{k+1}(\theta_v, \theta_w)$, assuming all agents use a uniform distribution as their initial beliefs; thus $\hat{\varphi}_0^i(\theta_v, \theta_w) = 0$. This simplifies the notation and facilitates the exposition of the results. Moreover, it does not limit the generality of our method since this term can be upper bounded and it will depend at most linearly on the number of agents.

Now we have that

$$\varphi_{k+1}^i(\theta_v, \theta_w) = \frac{\sum_{t=1}^{k} \left[D_{k:t} \mathcal{L}_t^{\theta_v, \theta_w} \right]_i + \left[\mathcal{L}_{k+1}^{\theta_v, \theta_w} \right]_i + \sum_{t=1}^{k} \phi_k^i \mathbf{1}' \mathcal{L}_t^{\theta_v, \theta_w}}{y_{k+1}^i}.$$

Similarly the dynamics of $\boldsymbol{y}_k$ can be expressed as

$$\begin{aligned} \boldsymbol{y}_{k+1} &= A_{k:0} \boldsymbol{y}_0 \\ &= A_{k:0} \boldsymbol{y}_0 - \boldsymbol{\phi}_k \mathbf{1}' \boldsymbol{y}_0 + \boldsymbol{\phi}_k \mathbf{1}' \boldsymbol{y}_0 \\ &= D_{k:0} \mathbf{1} + \boldsymbol{\phi}_k n, \end{aligned}$$

which leads us to

$$\varphi_{k+1}^i(\theta_v, \theta_w) = \frac{\sum_{t=1}^{k} \left[D_{k:t} \mathcal{L}_t^{\theta_v, \theta_w} \right]_i + \left[\mathcal{L}_{k+1}^{\theta_v, \theta_w} \right]_i + \sum_{t=1}^{k} \phi_k^i \mathbf{1}' \mathcal{L}_t^{\theta_v, \theta_w}}{[D_{k:0} \mathbf{1}]_i + \phi_k^i n}. \tag{3.23}$$

The next lemma will provide a general tool for analyzing the non-asymptotic properties of a learning rule that can be expressed as a log-linear function of bounded variations and upper bounded expectation as it was recently used in [64, 25]. It can be interpreted as a specialized version of the McDiarmid concentration [124] for log-linear update rules.

Lemma 21. *Consider a learning update rule that can be expressed as a log-linear function, i.e.,*

$$\mu_{k+1}^i(\theta_v) \leq \exp\left(\varphi_{k+1}^i(\theta_v, \theta_w)\right).$$

If the term $\varphi_{k+1}^i(\theta)$ is of bounded variations with bounds $\{c_k^i\}$ at each time k and its expected value is upper bounded by an affine function as $\mathbb{E}\left[\varphi_{k+1}^i(\theta)\right] \leq \frac{1}{\delta}\gamma_1^i - k\gamma_2$, then

$$\mathbb{P}\left(\mu_{k+1}^i(\theta_v) \geq \exp\left(-\frac{k}{2}\gamma_2 + \frac{1}{\delta}\gamma_1^i\right)\right) \leq \exp\left(-\frac{\frac{1}{2}(k\gamma_2)^2}{\sum_{t=1}^{k+1}(c_t^i)^2}\right).$$

Proof. Following simple set properties of the probability measure on the desired set

$$\mu_{k+1}^{i}\left(\theta_v\right) \geq \exp\left(-\frac{k}{2}\gamma_2 + \frac{1}{\delta}\gamma_1^i\right),$$

we have that

$$\begin{aligned}
\mathbb{P}\left(\mu_{k+1}^{i}\left(\theta_v\right) \geq \exp\left(-\frac{k}{2}\gamma_2 + \frac{1}{\delta}\gamma_1^i\right)\right) &\leq \mathbb{P}\left(\exp\left(\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right)\right) \geq \exp\left(-\frac{k}{2}\gamma_2 + \frac{1}{\delta}\gamma_1^i\right)\right) \\
&= \mathbb{P}\left(\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right) \geq -\frac{k}{2}\gamma_2 + \frac{1}{\delta}\gamma_1^i\right) \\
&= \mathbb{P}\left(\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right) - \mathbb{E}\left[\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right)\right] \geq \right. \\
&\qquad \left. -\frac{k}{2}\gamma_2 + \frac{1}{\delta}\gamma_1^i - \mathbb{E}\left[\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right)\right]\right) \\
&= \mathbb{P}\left(\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right) - \mathbb{E}\left[\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right)\right] \geq \frac{k}{2}\gamma_2\right).
\end{aligned}$$

Finally, use McDiarmid's inequality to get the desired result. □

The next lemma will show the desired properties required in Lemma 21 to get the non-asymptotic results. First, we will show the bounds on the expected value and then the bounded variation property.

Lemma 22. *Consider* $\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right)$ *as defined in Eq.* (3.19)*, then*

$$\mathbb{E}\left[\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right)\right] \leq \frac{1}{\delta}\gamma_1^i - k\gamma_2$$

for all i *and* $k \geq 0$*, with* γ_1^i *and* γ_2 *as in Theorem 19.*

Proof. First by taking the expected value of Eq. (3.23) we have that for all $k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right)\right] \\
&= \frac{\sum_{t=1}^{k}\left[D_{k:t}\boldsymbol{H}\left(\theta_v, \theta_w\right)\right]_i + \left[\boldsymbol{H}\left(\theta_v, \theta_w\right)\right]_i + \sum_{t=1}^{k}\phi_k^i \mathbf{1}'\boldsymbol{H}\left(\theta_v, \theta_w\right)}{\left[D_{k:0}\mathbf{1}\right]_i + \phi_k^i n} \\
&= \frac{\sum_{t=1}^{k}\left[D_{k:t}\boldsymbol{H}\left(\theta_v, \theta_w\right)\right]_i + \left[\boldsymbol{H}\left(\theta_v, \theta_w\right)\right]_i + k\phi_k^i \mathbf{1}'\boldsymbol{H}\left(\theta_v, \theta_w\right)}{\left[D_{k:0}\mathbf{1}\right]_i + \phi_k^i n}.
\end{aligned}$$

The main idea is to analyze how the term $\mathbb{E}\left[\varphi_{k+1}^{i}\left(\theta_v, \theta_w\right)\right]$ differs from a dynamic term where all agents have the same importance in the network and thus the learning occurs at

a rate $\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)/n$.

The first step will be to add and subtract the term $k\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)/n$; therefore, we obtain

$$\mathbb{E}\left[\varphi_{k+1}^{i}(\theta_v,\theta_w)\right] = \frac{\sum_{t=1}^{k}\left[D_{k:t}\boldsymbol{H}(\theta_v,\theta_w)\right]_i + \left[\boldsymbol{H}(\theta_v,\theta_w)\right]_i + k\phi_k^i\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{\left[D_{k:0}\mathbf{1}\right]_i + \phi_k^i n} - k\frac{\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{n} + k\frac{\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{n}.$$

By working out the arithmetic we have

$$\begin{aligned}\mathbb{E}\left[\varphi_{k+1}^{i}(\theta_v,\theta_w)\right] &= \frac{n\left(\sum_{t=1}^{k}\left[D_{k:t}\boldsymbol{H}(\theta_v,\theta_w)\right]_i + \left[\boldsymbol{H}(\theta_v,\theta_w)\right]_i + k\phi_k^i\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)\right)}{n\left(\left[D_{k:0}\mathbf{1}\right]_i + \phi_k^i n\right)} \\ &\quad - \frac{\left(\left[D_{k:0}\mathbf{1}\right]_i + \phi_k^i n\right)k\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{n\left(\left[D_{k:0}\mathbf{1}\right]_i + \phi_k^i n\right)} + k\frac{\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{n} \\ &= \frac{n\left(\sum_{t=1}^{k}\left[D_{k:t}\boldsymbol{H}(\theta_v,\theta_w)\right]_i + \left[\boldsymbol{H}(\theta_v,\theta_w)\right]_i\right)}{n\left(\left[D_{k:0}\mathbf{1}\right]_i + \phi_k^i n\right)} \\ &\quad - \frac{\left(\left[D_{k:0}\mathbf{1}\right]_i\right)k\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{n\left(\left[D_{k:0}\mathbf{1}\right]_i + \phi_k^i n\right)} + k\frac{\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{n}.\end{aligned}$$

Before finalizing the proof note that the denominator of the above function has the property $\left[D_{k:0}\mathbf{1}\right]_i + \phi_k^i n > \delta$. This follows from the fact that this term is the sum of the i-th row of the matrix $A_{k:0}$ multiplied n times [72]. Therefore by taking absolute value of the first terms we obtain

$$\begin{aligned}\mathbb{E}\left[\varphi_{k+1}^{i}(\theta_v,\theta_w)\right] &\leq \frac{1}{\delta}\left(\sum_{t=1}^{k}\left(\max_j |\left[D_{k:t}\right]_{ij}|\right)\|\boldsymbol{H}(\theta_v,\theta_w)\|_1 + \left[\boldsymbol{H}(\theta_v,\theta_w)\right]_i\right) \\ &\quad + \frac{k}{n\delta}\|\boldsymbol{H}(\theta_v,\theta_w)\|_1\left(\max_j |\left[D_{k:0}\right]_{ij}|\right)n + k\frac{\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{n}.\end{aligned}$$

Using Lemma 2 we obtain upper bounds on $|\left[D_{k:t}\right]_{ij}|$ where

$$\begin{aligned}\mathbb{E}\left[\varphi_{k+1}^{i}(\theta_v,\theta_w)\right] &\leq \frac{1}{\delta}\left(C\sum_{t=1}^{k}\lambda^{k-t}\|\boldsymbol{H}(\theta_v,\theta_w)\|_1 + \left[\boldsymbol{H}(\theta_v,\theta_w)\right]_i\right) \\ &\quad + k\frac{C}{\delta}\lambda^t\|\boldsymbol{H}(\theta_v,\theta_w)\|_1 + k\frac{\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{n} \\ &\leq \frac{2C}{\delta}\frac{1}{1-\lambda}\|\boldsymbol{H}(\theta_v,\theta_w)\|_1 + \frac{1}{\delta}\left[\boldsymbol{H}(\theta_v,\theta_w)\right]_i + k\frac{\mathbf{1}'\boldsymbol{H}(\theta_v,\theta_w)}{n}.\end{aligned}$$

The desired result follows by definition of the group confidence; thus,
$\mathbf{1}'\boldsymbol{H}(\theta_v, \theta_w) = \mathsf{C}^* - \mathsf{C}(\theta_v)$. □

Finally, note that the term $\varphi_k^i(\theta_v, \theta_w)$, as a function of a sequence of k random vectors, is of bounded variations, i.e.

$$\max_{\mathbf{s}_t \in \mathcal{S}} \varphi_{k+1}^i(\theta_v, \theta_w) - \min_{\mathbf{s}_t \in \mathcal{S}} \varphi_{k+1}^i(\theta_v, \theta_w) \leq \frac{2}{\delta}\left(\log \frac{1}{\alpha}\right).$$

We have now successfully developed the auxiliary results for the proof of Theorem 19.

Proof. (Theorem 19)

The proof procedure will be a compilation of previous lemmas. As a first step, we will show that the proposed learning rule can be expressed as a log-linear function.

Since $\mu_k^i(\theta) \in (0, 1]$ for all $i = 1, \ldots, n$, $k \geq 0$ and all $\theta \in \Theta$, we have that

$$\begin{aligned}
\mu_{k+1}^i(\theta_v) &\leq \frac{\mu_{k+1}^i(\theta_v)}{\mu_{k+1}^i(\theta_w)} \\
&= \frac{\left(\prod_{j=1}^n \mu_k^j(\theta_v)^{[A_k]_{ij} y_k^j} p_{\theta_v}^i(x_{k+1}^i)\right)^{\frac{1}{y_{k+1}^i}}}{\left(\prod_{j=1}^n \mu_k^j(\theta_w)^{[A_k]_{ij} y_k^j} p_{\theta_w}^i(x_{k+1}^i)\right)^{\frac{1}{y_{k+1}^i}}} \\
&= \left(\prod_{j=1}^n \left(\frac{\mu_k^j(\theta_v)}{\mu_k^j(\theta_w)}\right)^{[A_k]_{ij} y_k^j} \frac{p_{\theta_v}^i(x_{k+1}^i)}{p_{\theta_w}^i(x_{k+1}^i)}\right)^{\frac{1}{y_{k+1}^i}} \\
&= \exp\left(\frac{1}{y_{k+1}^i}\left(\sum_{j=1}^n [A_k]_{ij} \hat{\varphi}_k^j(\theta_v, \theta_w) + \log \frac{p_{\theta_v}^i(x_{k+1}^i)}{p_{\theta_w}^i(x_{k+1}^i)}\right)\right) \\
&= \exp\left(\varphi_{k+1}^j(\theta_v, \theta_w)\right).
\end{aligned}$$

This result alongside Lemma 22 provides the conditions for Lemma 21; thus, the following relation is valid:

$$\mathbb{P}\left(\mu_{k+1}^i(\theta_v) \geq \exp\left(-\frac{k}{2}\gamma_2 + \gamma_1^i\right)\right) \leq \exp\left(-\frac{\frac{1}{2}(k\gamma_2)^2}{\sum_{t=1}^{k+1}(c_t^i)^2}\right).$$

Specifically, we have that $c_t^i = \frac{2}{\delta}\log\frac{1}{\alpha}$. Therefore,

$$\begin{aligned}\mathbb{P}\left(\varphi_{k+1}^i\left(\theta_v,\theta_w\right)-\mathbb{E}\left[\varphi_{k+1}^i\left(\theta_v,\theta_w\right)\right]\geq\frac{k}{2}\gamma_2\right)&\leq\exp\left(-\frac{\frac{1}{2}\left(k\gamma_2\right)^2}{\sum_{t=1}^{k+1}\left(\frac{2}{\delta}\log\frac{1}{\alpha}\right)^2}\right)\\&=\exp\left(-\frac{\left(k\gamma_2\delta^2\right)^2}{8(k+1)\left(\log\frac{1}{\alpha}\right)^2}\right)\\&\leq\exp\left(-\frac{(k-1)\gamma_2^2\delta^2}{8\left(\log\alpha\right)^2}\right).\end{aligned}$$

Finally, for a given confidence level ρ, in order to have

$$\mathbb{P}\left(\mu_k^i\left(\theta_v\right)\geq\exp\left(-\frac{1}{2}k\gamma_2+\gamma_1^i\right)\right)\leq\rho,$$

we require that

$$k\geq\frac{8\left(\log\left(\alpha\right)\right)^2\log\frac{1}{\rho}}{\delta^2\gamma_2^2}+1.$$

This completes the proof. □

3.5 Acceleration of Distributed Learning over Fixed Undirected Graphs

For static undirected graphs, we propose a new belief update rule with one-step memory as follows: For each θ in Θ

$$\mu_{k+1}^i(\theta)=\frac{1}{\widetilde{Z}_{k+1}^i}\frac{\prod_{j=1}^n\mu_k^j(\theta)^{(1+\sigma)\bar{A}_{ij}}p_\theta^i(x_{k+1}^i)^{\beta_k^i}}{\prod_{j=1}^n\left(\mu_{k-1}^j(\theta)p_\theta^j(x_k^j)^{\beta_{k-1}^j}\right)^{\sigma\bar{A}_{ij}}},\tag{3.24}$$

where $\widetilde{Z}_{k+1}^i$ is the corresponding normalization factor given by

$$\widetilde{Z}_{k+1}^i=\sum_{r=1}^m\frac{\prod_{j=1}^n\mu_k^j\left(\theta_r\right)^{(1+\sigma)\bar{A}_{ij}}p_{\theta_r}^i(x_{k+1}^i)^{\beta_k^i}}{\prod_{j=1}^n\left(\mu_{k-1}^j(\theta_r)p_{\theta_r}^j(x_k^j)^{\beta_{k-1}^j}\right)^{\sigma\bar{A}_{ij}}},$$

where $\bar{A}$ is a specifically chosen matrix (called the *lazy Metropolis matrix*) and σ a constant to be set later. We initialize $\mu^i_{-1}(\theta)$ to be equal to $\mu^i_0(\theta)$ for all $i = 1, \ldots, n$ and $\theta \in \Theta$. We will show that this update rule generates a sequence of beliefs that concentrate at a rate a factor of n faster than the previous results. Note that the update rule described in Eq. (3.24) requires the communication of the product of the beliefs and likelihood functions and an additional memory since the beliefs at time $k+1$ depend on the beliefs a time k and at time $k-1$.

Our next result shows the belief concentration rate for the update rule described in Eq. (3.24).

Theorem 23. *Let Assumptions 3 and 4 hold and let $\rho \in (0,1)$. Furthermore let $U \geq n$ and let $\sigma = 1 - 2/(9U+1)$. Then, the update rule of Eq.* (3.24) *with this σ, uniform initial beliefs with the condition $\mu^i_{-1}(\theta) = \mu^i_0(\theta)$ and β^i_{-1} fixed to zero, has the following property: There is an integer $\boldsymbol{N}(\rho)$ such that, with probability $1-\rho$, for all $k \geq \boldsymbol{N}(\rho)$ and for all $\theta_v \notin \Theta^*$, it holds that*

$$\mu^i_k(\theta_v) \leq \exp\left(-\frac{k}{2}\gamma_2 + \gamma^i_1\right) \quad \textit{for all } i = 1, \ldots, n,$$

where

$$\begin{aligned} \boldsymbol{N}(\rho) &\triangleq \left\lceil \frac{1}{\gamma_2^2} 48 \left(\log \alpha\right)^2 \log\left(\frac{1}{\rho}\right) \right\rceil, \\ \gamma^i_1 &\triangleq \frac{4 \log n}{1-\lambda} \log \frac{1}{\alpha}, \\ \gamma_2 &\triangleq \frac{1}{n} \min_{\theta_v \notin \Theta^*} \left(\mathsf{C}^*_{\boldsymbol{q}} - \mathsf{C}_{\boldsymbol{q}}(\theta_v)\right), \end{aligned}$$

with α from Assumption 4(b) and $\lambda = 1 - \frac{1}{18U}$.

Note that the beliefs for $k = -1$ and $k = 0$ are defined as equal. Additionally, we assume there is no observation available for time 0; this holds if we assume $\beta^i_{-1} = 0$ with any realization of S^i_0.

The bound of Theorem 3 is an improvement by a factor of n compared to the bounds of Theorem 17. In a network of n agents where α, ρ and γ_2 are treated like constants with respect to the number of agents, we require at least $\mathcal{O}(n \log n)$ iterations for the beliefs on the incorrect hypotheses to be below certain small value epsilon (assuming U is within a constant factor of n). Following the results of [45], the best bound one could get using a Metropolis weights is $\mathcal{O}(n^2 \log n)$, as in Theorem 17 if $B = 1$.

We note, however, that the requirements of Theorem 23 are more stringent than those of Theorem 17. The network topology is fixed (i.e., a static graph) and all nodes need to

know an upper bound U on the total number of agents. This upper bound must be within a constant factor of the number of agents.

Next, we define some quantities that we use in the analysis of Eq. (3.24). Define the matrix B and a scalar σ as follows:

$$B = \begin{bmatrix} (1+\sigma)\,\bar{A} & -\sigma\bar{A} \\ I_n & \mathbf{0} \end{bmatrix}, \tag{3.25}$$

and

$$\sigma = 1 - \frac{2}{9U+1}, \tag{3.26}$$

where I_n is the identity matrix and $\mathbf{0}$ is the matrix with all entries equal to zero of the appropriate size and $\bar{A}$ is as defined in Assumption 3.

We have the following auxiliary result for the matrix B.

Lemma 24. *Consider the matrix B and the parameter σ as defined in Eqs.* (3.25) *and* (3.26) *respectively. Then*

$$\left| [[I_n \ \mathbf{0}]B^k[I_n \ I_n]']_{ij} - \frac{1}{n} \right| \leq \sqrt{2}\lambda^k \quad \forall\ k \geq 2,$$

where $\lambda = 1 - \frac{1}{18U}$.

Proof. The linear time consensus algorithm described in Eq. (1.2) can be expressed as

$$\begin{aligned} \boldsymbol{y}_{k+1} &= \bar{A}\boldsymbol{x}_k \\ \boldsymbol{x}_{k+1} &= \boldsymbol{y}_{k+1} + \sigma\left(\boldsymbol{y}_{k+1} - \boldsymbol{y}_k\right), \end{aligned}$$

which implies that $\boldsymbol{y}_{k+1} = \bar{A}\left(\boldsymbol{y}_k + \sigma\left(\boldsymbol{y}_k - \boldsymbol{y}_{k-1}\right)\right)$ with $y_1^i = x_1^i$. Therefore

$$\begin{bmatrix} \boldsymbol{y}_{k+1} \\ \boldsymbol{y}_k \end{bmatrix} = \begin{bmatrix} (1+\sigma)\,\bar{A} & -\sigma\bar{A} \\ I_n & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_k \\ \boldsymbol{y}_{k-1} \end{bmatrix} = B^k \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_0 \end{bmatrix},$$

where we assumed that $\boldsymbol{y}_0 = \boldsymbol{y}_1$. Thus,

$$\boldsymbol{y}_{k+1} = [I_n \ \mathbf{0}]B^k[I_n \ I_n]'\boldsymbol{y}_1.$$

By substituting the previous relation into Eq. (1.3) and using $\boldsymbol{x}_1 = \boldsymbol{y}_1$, we obtain

$$\|[I_n\ \mathbf{0}]B^k[I_n\ I_n]'\boldsymbol{y}_1 - \left(\frac{1}{n}\sum_{i=1}^n y_1^i\right)\mathbf{1}_n\|_2^2 \leq 2\left(1-\frac{1}{9U}\right)^k \|\boldsymbol{y}_1 - \frac{1}{n}\sum_{i=1}^n y_1^i\mathbf{1}_n\|_2^2,$$

which implies that

$$\max_i \left|[[I_n\ \mathbf{0}]B^k[I_n\ I_n]'\boldsymbol{y}_1]_i - \frac{1}{n}\sum_{i=1}^n y_1^i\right| \leq \sqrt{2}\left(\sqrt{1-\frac{1}{9U}}\right)^k \|\boldsymbol{y}_1 - \frac{1}{n}\sum_{i=1}^n y_1^i\mathbf{1}_n\|_2.$$

The preceding relation holds for any $\boldsymbol{y}_1$. In particular, if we take $\boldsymbol{y}_1 = \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is a vector whose j-th entry is equal to one and zero otherwise, we conclude that for every i and j,

$$\left|[[I_n\ \mathbf{0}]B^k[I_n\ I_n]']_{ij} - \frac{1}{n}\right| \leq \sqrt{2}\left(1-\frac{1}{18U}\right)^k.$$

This follows from the inequality $\sqrt{1-\beta} \leq 1-\beta/2$ for all $\beta \in (0,1)$ and the fact that $\|\boldsymbol{e}_j - \frac{1}{n}\mathbf{1}_n\| \leq 1$. □

Now, we are ready to prove Theorem 23.

Proof. (Theorem 23) The proof is along the lines of the proof for Theorem 17. From the definition of $\varphi_{k+1}^i(\theta_v,\theta_w)$ we have

$$\begin{aligned}
\varphi_{k+1}^i(\theta_v,\theta_w) &= \log\frac{\mu_{k+1}^i(\theta_v)}{\mu_{k+1}^i(\theta_w)} \\
&= \log\frac{\frac{\prod_{j=1}^n \mu_k^j(\theta_v)^{(1+\sigma)\overline{A}_{ij}} p_{\theta_v}^i(X_{k+1}^i)^{\beta_k^i}}{\prod_{j=1}^n (\mu_{k-1}^j(\theta_v)p_{\theta_v}^j(X_k^j)^{\beta_{k-1}^j})^{\sigma\overline{A}_{ij}}}}{\frac{\prod_{j=1}^n \mu_k^j(\theta_w)^{(1+\sigma)\overline{A}_{ij}} p_{\theta_w}^i(X_{k+1}^i)^{\beta_k^i}}{\prod_{j=1}^n (\mu_{k-1}^j(\theta_w)p_{\theta_w}^j(X_k^j)^{\beta_{k-1}^j})^{\sigma\overline{A}_{ij}}}} \\
&= \sum_{j=1}^n (1+\sigma)\,\bar{A}_{ij}\log\frac{\mu_k^j(\theta_v)}{\mu_k^j(\theta_w)} - \sum_{j=1}^n \sigma\bar{A}_{ij}\log\frac{\mu_{k-1}^j(\theta_v)}{\mu_{k-1}^j(\theta_w)} \\
&\quad + \beta_k^i\log\frac{p_{\theta_v}^i(X_{k+1}^i)}{p_{\theta_w}^i(X_{k+1}^i)} - \sum_{j=1}^n \sigma\bar{A}_{ij}\beta_{k-1}^j\log\frac{p_{\theta_v}^j(X_k^j)}{p_{\theta_w}^j(X_k^j)} \\
&= \sum_{j=1}^n (1+\sigma)\,\bar{A}_{ij}\varphi_k^j(\theta_v,\theta_w) - \sum_{j=1}^n \sigma\bar{A}_{ij}\varphi_{k-1}^j(\theta_v,\theta_w)
\end{aligned}$$

$$+ [\mathcal{L}_{k+1}^{\theta_v,\theta_w}]_i - \sum_{j=1}^{n} \sigma \bar{A}_{ij} [\mathcal{L}_k^{\theta_v,\theta_w}]_j.$$

Stacking the previous relation for all i we obtain the following vector representation for the dynamics:

$$\boldsymbol{\varphi}_{k+1}(\theta_v,\theta_w) = (1+\sigma)\,\bar{A}\boldsymbol{\varphi}_k(\theta_v,\theta_w) - \sigma\bar{A}\boldsymbol{\varphi}_{k-1}(\theta_v,\theta_w) + \mathcal{L}_{k+1}^{\theta_v,\theta_w} - \sigma\bar{A}\mathcal{L}_k^{\theta_v,\theta_w}. \tag{3.27}$$

Now, define the following auxiliary vector

$$\boldsymbol{z}_{k+1}(\theta_v,\theta_w) = \boldsymbol{\varphi}_k(\theta_v,\theta_w) + \mathcal{L}_{k+1}^{\theta_v,\theta_w},$$

where $\boldsymbol{z}_0(\theta_v,\theta_w) = 0$, since $\boldsymbol{\varphi}_{-1}(\theta_v,\theta_w) = 0$ by the assumption of uniform initial beliefs, and $\mathcal{L}_0^{\theta_v,\theta_w} = 0$ due to $\beta_{-1} = 0$, in which case we can set S_0^i to any value in $\mathcal{S}^i$.

By writing the evolution for the augmented state $[\boldsymbol{\varphi}_{k+1}(\theta_v,\theta_w)\ \boldsymbol{z}_{k+1}(\theta_v,\theta_w)]'$ we have

$$\begin{bmatrix} \boldsymbol{\varphi}_{k+1}(\theta_v,\theta_w) \\ \boldsymbol{z}_{k+1}(\theta_v,\theta_w) \end{bmatrix} = B \begin{bmatrix} \boldsymbol{\varphi}_k(\theta_v,\theta_w) \\ \boldsymbol{z}_k(\theta_v,\theta_w) \end{bmatrix} + \begin{bmatrix} \mathcal{L}_{k+1}^{\theta_v,\theta_w} \\ \mathcal{L}_{k+1}^{\theta_v,\theta_w} \end{bmatrix}$$

which implies that for all $k \geq 1$,

$$\begin{bmatrix} \boldsymbol{\varphi}_{k+1}(\theta_v,\theta_w) \\ \boldsymbol{z}_{k+1}(\theta_v,\theta_w) \end{bmatrix} = B^{k+1} \begin{bmatrix} \boldsymbol{\varphi}_0(\theta_v,\theta_w) \\ \boldsymbol{z}_0(\theta_v,\theta_w) \end{bmatrix} + \sum_{t=1}^{k} B^{k+1-t} \begin{bmatrix} \mathcal{L}_t^{\theta_v,\theta_w} \\ \mathcal{L}_t^{\theta_v,\theta_w} \end{bmatrix} + \begin{bmatrix} \mathcal{L}_{k+1}^{\theta_v,\theta_w} \\ \mathcal{L}_{k+1}^{\theta_v,\theta_w} \end{bmatrix}.$$

Then we have

$$\boldsymbol{\varphi}_k(\theta_v,\theta_w) = [I_n\ \mathbf{0}]B^k[I_n\ I_n]'\boldsymbol{\varphi}_0(\theta_v,\theta_w) + \sum_{t=1}^{k}[I_n\ \mathbf{0}]B^{k-t}[I_n\ I_n]'\mathcal{L}_t^{\theta_v,\theta_w},$$

where the assumption of uniform initial beliefs sets the first term of the above relation to zero.

The remainder of the proof follows the structure of the proof of Theorem 17, where we invoke Lemma 24 instead of Lemma 2. First, we will find a bound for the expected value of $\boldsymbol{\varphi}_k(\theta_v,\theta_w)$ and later we will show this is of bounded variations. In this case, we have

$$\mathbb{E}[\varphi_k^i(\theta_v,\theta_w)] = -\sum_{t=1}^{k}\sum_{j=1}^{n}[[I_n\ \mathbf{0}]B^{k-t}[I_n\ I_n]']_{ij}H^j(\theta_v,\theta_w).$$

By adding and subtracting $\sum_{t=1}^{k}\sum_{j=1}^{n}\frac{1}{n}H^j(\theta_v,\theta_w)$ we obtain

$$\mathbb{E}[\varphi_k^i(\theta_v,\theta_w)] = -\frac{k}{n}\sum_{j=1}^{n}H^j(\theta_v,\theta_w) + \sum_{t=1}^{k}\sum_{j=1}^{n}\left(\frac{1}{n} - [[I_n\ \mathbf{0}]B^{k-t}[I_n\ I_n]']_{ij}\right)H^j(\theta_v,\theta_w).$$

Similarly, as in the proof of Theorem 17, we bound the term in parentheses using the non-asymptotic bounds from Lemma 24 in conjunction with Lemma 4. By doing so, it can be seen that

$$\mathbb{E}[\varphi_k^i(\theta_v,\theta_w)] \leq \frac{4\log n}{1-\lambda}\log\frac{1}{\alpha} - \frac{k}{n}\sum_{j=1}^{n}H^j(\theta_v,\theta_w).$$

Now, we will show that $\varphi_k^i(\theta_v,\theta_w)$, as a function of the random variables consisting in S_t^j for $1\leq t\leq k$ to $1\leq j\leq n$, has bounded variations and we will compute the bound. First, we fix all other input random variables but $[\mathcal{L}_t^{\theta_v,\theta_w}]_j$ and we have

$$\begin{aligned}\max_{s_t^j\in\mathcal{S}^j}\varphi_k^i(\theta_v,\theta_w) - \min_{s_t^j\in\mathcal{S}^j}\varphi_k^i(\theta_v,\theta_w) &= \max_{s_t^j\in\mathcal{S}^j}[[I_n\ \mathbf{0}]B^{k-t}[I_n\ I_n]']_{ij}[\mathcal{L}_t^{\theta_v,\theta_w}]_j \\ &\quad - \min_{s_t^j\in\mathcal{S}^j}[[I_n\ \mathbf{0}]B^{k-t}[I_n\ I_n]']_{ij}[\mathcal{L}_t^{\theta_v,\theta_w}]_j \\ &\leq [[I_n\ \mathbf{0}]B^{k-t}[I_n\ I_n]']_{ij}2\log\frac{1}{\alpha}.\end{aligned}$$

Thus, the summation of the squared bounds in McDiarmid's inequality is

$$\sum_{t=1}^{k}\sum_{j=1}^{n}\left([[I_n\ \mathbf{0}]B^{k-t}[I_n\ I_n]']_{ij}2\log\frac{1}{\alpha}\right)^2.$$

Now, by adding and subtracting the term $1/n$ we have that

$$\begin{aligned}\sum_{t=1}^{k}\sum_{j=1}^{n}\left([[I_n\ \mathbf{0}]B^{k-t}[I_n\ I_n]']_{ij}2\log\frac{1}{\alpha}\right)^2 &\leq 8\left(\log\frac{1}{\alpha}\right)^2\sum_{t=1}^{k}\sum_{j=1}^{n}\left([[I_n\ \mathbf{0}]B^{k-t}[I_n\ I_n]']_{ij} - 1/n\right)^2 \\ &\quad + 8\left(\log\frac{1}{\alpha}\right)\sum_{t=1}^{k}\sum_{j=1}^{n}(1/n)^2,\end{aligned}$$

where we have used $x^2\leq 2((x-y)^2+y^2)$.

We can bound the first term in the preceding relation using Eq. (3.27) with $\boldsymbol{y}_1=\boldsymbol{e}_j$ since

Eq. (3.27) holds for any choice of $\boldsymbol{y}_1$. Specifically, we obtain that for all $j = 1, \ldots, n$

$$\sum_{i=1}^{n} \left([[I_n \ \mathbf{0}] B^{k-t} [I_n \ I_n]']_{ij} - 1/n \right)^2 \leq 2 \left(1 - \frac{1}{9U} \right)^{k-t}.$$

Additionally, note that $[I_n \ \mathbf{0}] B^k [I_n \ I_n]'$ is a symmetric matrix since it is a polynomial of $\bar{A}$ which is symmetric itself. This in turn implies that $[I_n \ \mathbf{0}] B^{k-t} [I_n \ I_n]'$ is also symmetric. Therefore, it holds that for all $i = 1, \ldots, n$

$$\sum_{j=1}^{n} \left([[I_n \ \mathbf{0}] B^{k-t} [I_n \ I_n]']_{ij} - 1/n \right)^2 \leq 2 \left(1 - \frac{1}{9U} \right)^{k-t}.$$

Finally, we have

$$\begin{aligned} \sum_{t=1}^{k} \sum_{j=1}^{n} \left([[I_n \ \mathbf{0}] B^{k-t} [I_n \ I_n]']_{ij} 2 \log \frac{1}{\alpha} \right)^2 &\leq 8 \left(\log \frac{1}{\alpha} \right)^2 \left(2k + \frac{k}{n} \right) \\ &\leq 24 \left(\log \alpha \right)^2 k. \end{aligned}$$

Now, by the McDiarmid inequality and getting the values of k such that the desired probabilistic tolerance level ρ is achieved, we obtain

$$\begin{aligned} \mathbb{P} \left(\varphi_k^i(\theta_v, \theta_w) - \mathbb{E}[\varphi_k^i(\theta_v, \theta_w)] \geq \frac{k}{2} \gamma_2 \right) &= \exp \left(- \frac{2 \left(\frac{1}{2} k \gamma_2 \right)^2}{24 \left(\log \alpha \right)^2 k} \right) \\ &= \exp \left(- \frac{k \gamma_2^2}{48 \left(\log \alpha \right)^2} \right). \end{aligned}$$

Therefore, for a given confidence level ρ, in order to have

$$\mathbb{P} \left(\mu_k^i(\theta_v) \geq \exp \left(-\frac{1}{2} k \gamma_2 + \gamma_1^i \right) \right) \leq \rho,$$

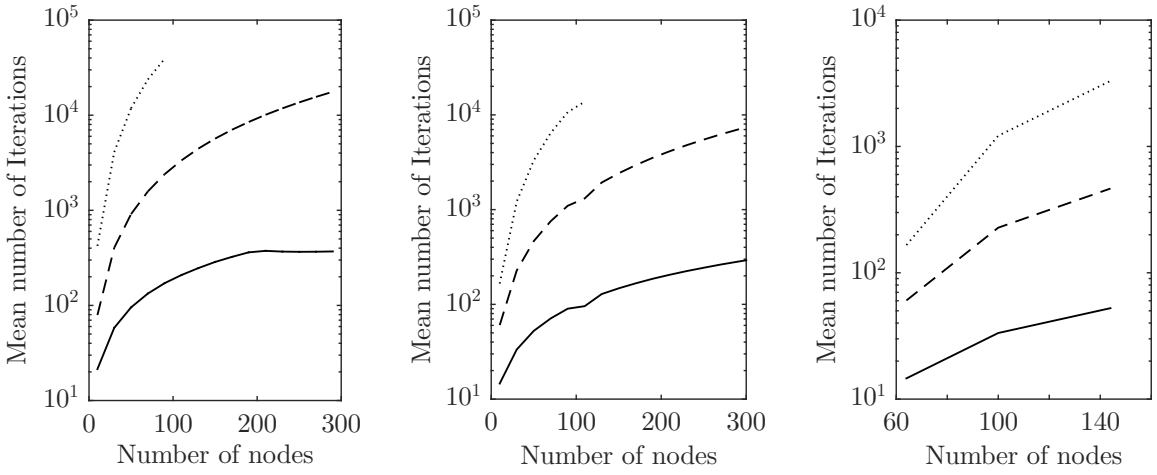
we require that

$$k \geq \frac{1}{\gamma_2^2} 48 \left(\log \alpha \right)^2 \log \frac{1}{\rho}.$$

□

Figure 3.4 presents simulation results that show how the convergence time depends on the number of agents in the network. Figure 3.4 shows the time required for a group of agents

to have a set of beliefs at a distance of $\epsilon = 0.01$ from the singleton distribution around the optimal hypothesis. For example, on a path graph, as the path grows longer, the number of iterations required to meet the desired ϵ accuracy grows rapidly. This is due to the low connectivity of the network. The time required for consensus is smaller for the circle and the grid graphs due to their better connectivity properties.



(a) Path Graph (b) Circle Graph (c) Grid Graph

Figure 3.4: Empirical mean over 50 Monte Carlo runs of the number of iterations required for $\mu_k^i(\theta) < \epsilon$ for all agents on $\theta \notin \Theta^*$. All agents but one have all their hypotheses to be observationally equivalent. Dotted line for the algorithm proposed in [1], dashed line for the procedure described in Eq. (3.9) and solid line for the procedure described in Eq. (3.24).

3.6 Generalized Non-Bayesian Learning Protocols

In this section, we discuss a general class of distributed non-Bayesian algorithms. First, we will motivate the choice of the update rules described in Eq. (3.9) and Eq. (3.24). For simplicity of exposition, we will assume that the agents always obtain observations (i.e. $\beta_k^i = 1$ in Eqs. (3.9) and (3.24) for all i and k). Then, we will provide a comparison between our algorithms and previously proposed algorithms within the generalized distributed non-Bayesian framework.

Opinion pooling or opinion aggregation has been studied before in [8, 11, 9, 10]. It is considered a traditional problem in economics, where several experts have beliefs about a hypothesis and one needs to aggregate their beliefs into a single probability distribution. Different opinion aggregation functions result from using different divergence metrics

for probability distributions (see [171]). Similarly, different opinion pool operators define different non-Bayesian distributed learning rules. A general form of opinion pooling was introduced in [11], termed *g-Quasi-Linear Opinion pools* (g-QLOP), defined as follows:

$$\tau_g^{A_k}\left(\ldots,\mu_k^j(\theta),\ldots\right)=\frac{g^{-1}\left(\sum_{j=1}^n\left[A_k\right]_{ij}g(\mu_k^j(\theta))\right)}{\sum_{r=1}^m g^{-1}\left(\sum_{j=1}^n\left[A_k\right]_{ij}g(\mu_k^j\left(\theta_r\right))\right)},$$

with $\tau_g^A:\prod_{i=1}^n\mathbb{P}\left(\Theta\right)\to\mathbb{P}\left(\Theta\right)$. The *g-QLOP* corresponds to weighted arithmetic averages when $g(x)=x$ and to weighted geometric averages when $g(x)=\log x$.

The update rules studied in this chapter can be seen as a two-step procedure. First, the beliefs of the neighbors are combined according to an opinion aggregation function. Second, the resulting aggregate distribution is updated using Bayes' rule. The proposed update rule, see Eq. (3.9), uses the logarithmic opinion pool, where

$$\tau_{\log x}^{A_k}\left(\ldots,\mu_k^j(\theta),\ldots\right)=\frac{\prod_{j=1}^n\mu_k^j(\theta)^{\left[A_k\right]_{ij}}}{\sum_{r=1}^m\prod_{j=1}^n\mu_k^j\left(\theta_r\right)^{\left[A_k\right]_{ij}}},$$

thus

$$\mu_{k+1}^i(\theta)=\frac{\tau_{\log x}^{A_k}\left(\ldots,\mu_k^j(\theta),\ldots\right)p_\theta^i(x_{k+1}^i)}{\sum_{r=1}^m\tau_{\log x}^{A_k}\left(\ldots,\mu_k^j(\theta_r),\ldots\right)p_{\theta_r}^i(x_{k+1}^i)}.$$

Logarithmic pools are externally Bayesian [8, 172], i.e., the order of aggregation of beliefs and new evidence does not influence the update rule. That is, from a learning point of view, if the function is externally Bayesian, we can interchange the innovation and diffusion steps. The order in which we aggregate opinions and make the Bayesian update does not change the update rule. The next proposition shows that the update rule in Eq. (3.9) is externally Bayesian.

Proposition 25. *Assume that $\beta_k^i=1$ for all i and k in the update rule Eq.* (3.9)*. Then, this rule is externally Bayesian, i.e. Eq.* (3.9) *is equivalent to:*

$$\mu_{k+1}^i(\theta)=\tau_{\log x}^{A_k}\left(\ldots,\frac{\mu_k^j(\theta)p_\theta^i(x_{k+1}^i)}{\sum_{r=1}^m\mu_k^j\left(\theta_r\right)p_{\theta_r}^i(x_{k+1}^i)},\ldots\right).$$

Proof. The proof of this proposition can be found in [25]. □

Consider now a linear opinion pool, where

$$\tau_x^{A_k}\left(\ldots,\mu_k^j(\theta),\ldots\right)=\sum_{j=1}^{n}[A_k]_{ij}\mu_k^j(\theta).$$

If the opinion aggregation is done first, as studied in [56], then the resulting update rule is

$$\mu_{k+1}^i(\theta)=\frac{\sum_{j=1}^{n}[A_k]_{ij}\mu_k^j(\theta)p_\theta^i(x_{k+1}^i)}{\sum_{r=1}^{m}\sum_{j=1}^{n}[A_k]_{ij}\mu_k^j(\theta_r)p_{\theta_r}^i(x_{k+1}^i)}.$$

On the other hand, if the Bayesian update is done first, then the resulting update rule is

$$\mu_{k+1}^i(\theta)=\sum_{j=1}^{n}[A_k]_{ij}\frac{\mu_k^j(\theta)p_\theta^j(x_{k+1}^j)}{\sum_{r=1}^{m}\mu_k^j(\theta_r)p_{\theta_r}^j(x_{k+1}^j)}. \tag{3.28}$$

The linear pool-based update rule is similar to the update rule proposed in [1]. The authors in [1] proposed the following rule:

$$\mu_{k+1}^i(\theta)=\tau_x^A\left(\ldots,\frac{\mu_k^i(\theta)p_\theta^i(x_{k+1}^i)}{\sum_{r=1}^{m}\mu_k^j\left(\theta_r\right)p_{\theta_r}^i(x_{k+1}^i)},\ldots\right),$$

where opinion aggregation with linear functions is performed locally with priors from the neighbors. The main difference is that in Eq. (3.28), a convex combination of the posteriors received from the neighbor set is used to generate the new individual posterior, while in [1] the update rule is a convex combination of the individual posterior and the neighbors' priors.

In [43], the authors considered the case where the randomized gossip algorithm defines the communication structure. The update protocol is based on a distributed version of the Nesterov's dual averaging with stochastic gradients corresponding to the log-likelihood models given a set of observations. In this case, the agents exchange the likelihoods of the current observations instead of the beliefs. Thus, the consensus step is performed as a geometric aggregation of the likelihoods, and the resulting update rule can be described as

$$\mu_{k+1}^i(\theta)=\frac{\mu_k^i(\theta)\tau_{\log x}^{W_k}\left(\ldots,p_\theta^j(x_{k+1}^j),\ldots\right)}{\sum_{r=1}^{m}\mu_k^i\left(\theta_r\right)\tau_{\log x}^{W_k}\left(\ldots,p_{\theta_r}^j(x_{k+1}^j),\ldots\right)}, \tag{3.29}$$

where W_k is the communication matrix coming from the gossip protocol.

The idea of communicating aggregated versions of likelihoods instead of beliefs was previously studied in the context of distributed estimation in sensor networks [53]. Approaching the problem from the point of view of the belief propagation algorithm resulted in an up-

date rule in the form of Eq. (3.29). In [53], the authors showed convergence results for primitive, rings, tree, random graphs and other extensions to the original belief propagation algorithm. Similarly, in [46], the authors propose an update rule where every agent performs local Bayesian updates before aggregating their beliefs using geometric averages, i.e.

$$\mu_{k+1}^{i}(\theta) = \tau_{\log x}^{A}\left(\ldots, \frac{\mu_k^j(\theta)p_\theta^j(x_{k+1}^j)}{\sum_{r=1}^{m}\mu_k^j(\theta_r)\,p_{\theta_r}^j(x_{k+1}^j)}, \ldots\right).$$

Convergence results for fixed communications matrices are provided, as well as asymptotic characterizations of the rates of convergence. Later in [46], the authors extended the characterization of the rate of convergence to large deviation theory, providing a statement about the existence of a random time after which the beliefs will decrease exponentially.

3.7 Numerical Example: Distributed Source Localization

As a motivating example, consider the problem of distributed source localization [17, 173]. In this scenario, a network of n agents receives noisy measurements of the distance to a source. The sensing capabilities of each sensor might be limited to a certain region. The group objective is to identify the location of the source jointly. Figure 3.5 shows a group of 7 agents (circles) seeking to localize a source (star). There is an underlying graph that indicates which nodes can exchange messages. Moreover, each node has a sensing region indicated by the dashed circle around it. Each agent observes signals proportional to the distance to the target. Since a target cannot be localized effectively from a single measure of the distance, agents must cooperate to have any hope of achieving decent localization.

In this section, we apply the proposed algorithms to the problem of distributed source localization based on differential signal amplitudes [17, 173, 174, 175, 176]. We compare the performance of our methods, Eq. (3.9) and Eq. (3.24), with the algorithms proposed in [56, 1]. For simulation purposes, we will assume the graphs are fixed, and there exists a single θ^* such that $P^i = P^i_{\theta^*}$ for all i, in which case our update rule simplifies to the learning algorithm proposed in [45].

Each agent constructs a grid of hypotheses about the possible location of the source. Figure 3.6(a) shows a 10 by 10 area partitioned in a 3 by 3 grid, which results in 9 hypotheses. Moreover, there are three agents (represented by circles), at different locations. The graph structure shows that agent 1 communicates with agent 2, and similarly, agent 3 communicates with 2. The star represents the target.

Each agent constructs likelihood functions for its hypotheses based on its sensor model.

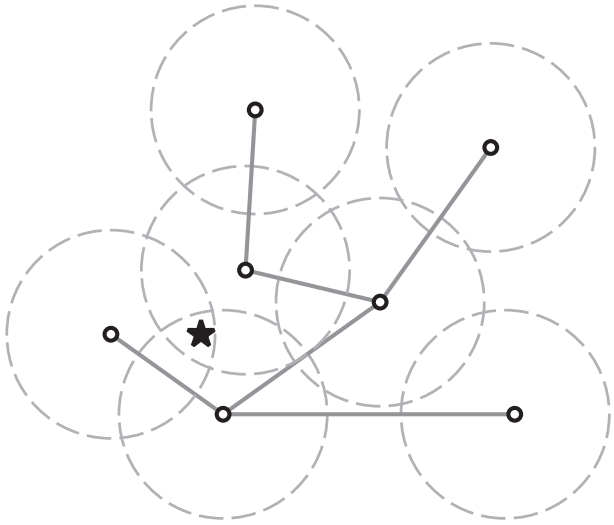


Figure 3.5: Distributed source localization example.

The observations follow a truncated normal distribution with the mean proportional to the distance between the agent and the grid point of the corresponding hypothesis. For example, assume an agent i is in a position x_a^i and the target is located at x_s. The received signals are $X_k^i = \|x_s - x_a^i\| + cW_k^i$, where c is some positive constant and W_k^i is a truncated zero-mean Gaussian noise. Now, consider that a hypothesis θ is at a point x_θ. The corresponding likelihood model under hypothesis θ assumes observations are $X_k^i|\theta = \|x_\theta - x_a\| + cW_k^i$.

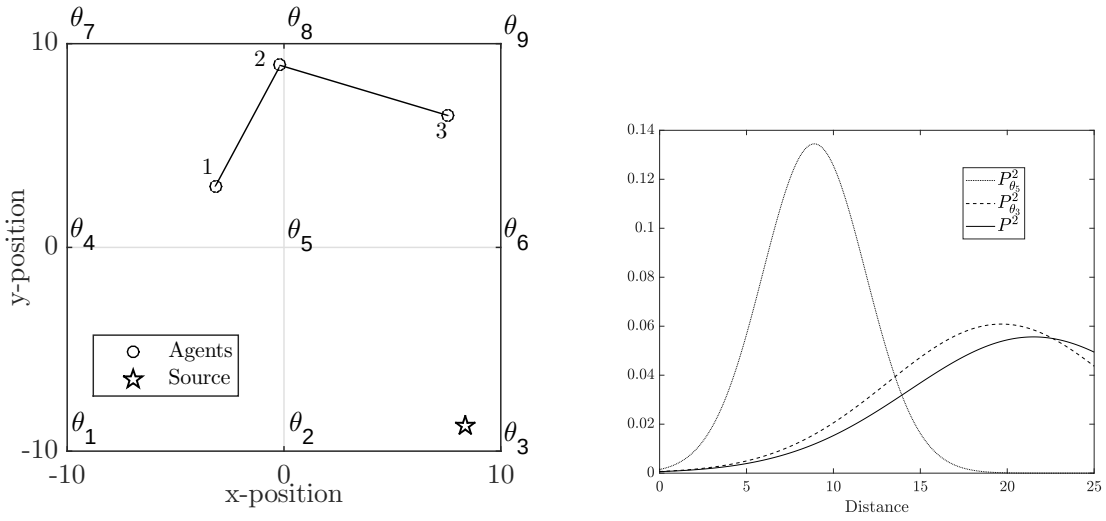
Figure 3.6(b) shows the likelihood functions for θ_5 and θ_3 of agent 2, clearly hypothesis θ_3 is closer to the true distribution of the observations P^2. Note that there is not a "true state of the world" in the sense that P^2 is not equal to any of the hypotheses in the grid.

The information each agent obtains is enough just to estimate the distance to the source, but not its complete coordinates. For instance, a single sensor can only locate the source within a circular band around it, see Fig. 3.7.

Figure 3.8(a) shows another group of 20 agents now interacting according to an appropriate network structure, see Assumptions 1 and 3. A finer grid partition has been used, where each coordinate has 100 points, resulting in 10000 hypotheses in total. Figure 3.8(b) shows the belief on the hypothesis θ^*, defined to be the grid point closest to the location of the target.

Figure 3.9 repeats the simulations presented in Figure 3.8 but including 10 agents with all their hypotheses observationally equivalent (i.e. no measurements available), and 3 conflicting agents whose observations have been modified (corrupted) such that the optimal hypothesis is the $(0, 0)$ point in the grid.

Figure 3.9(b) shows the protocols presented in Eqs. (3.9) and (3.24) concentrate the beliefs



(a) Network of 3 agents

(b) Distributions for Hypothesis θ_5 and θ_3.

Figure 3.6: Source localization on a grid with 3 agents and 9 hypotheses. (a) A group of 3 agents in a grid of 3×3 hypotheses. Each hypothesis corresponds to a possible location of the source. For example, hypothesis θ_2 locates the source at the $(-10, 0)$ point in the plane. (b) The likelihood functions for θ_2 and θ_5 and distribution of observations P^2 for agent 2.

onto the optimal hypothesis. The performance of the algorithms in [1] and [56] deteriorates if conflicting agents are present. This is evident from the lack of concentration of the beliefs around the true hypotheses.

3.8 Conclusions

We proposed two distributed cooperative learning algorithms for the problem of collaborative inference. The first algorithm focuses on general time-varying undirected graphs, and the second algorithm is specialized for fixed graphs. In both cases, we show that the beliefs converge to the hypothesis set that best describes the observations in the network. We require reasonable connectivity assumptions on the communication network over which the agents exchange information. Our results prove convergence rates that are non-asymptotic, geometric, and explicit. The bounds depend explicitly on the graph sequence properties, as well as the agent learning capabilities. Moreover, we do so in a new general setting where there might not be a true state of the world which is perfectly described by a single hypothesis, i.e., misspecified models. Additionally, we analyze networks where agents might have conflicting hypotheses, i.e., the hypothesis with the highest confidence changes if different subsets of agents are taken into account. The algorithm for fixed undirected graphs achieves

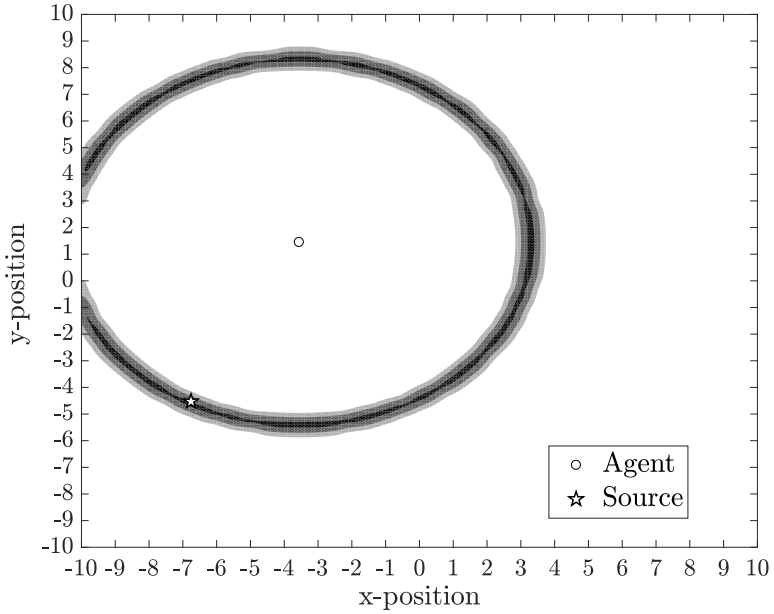
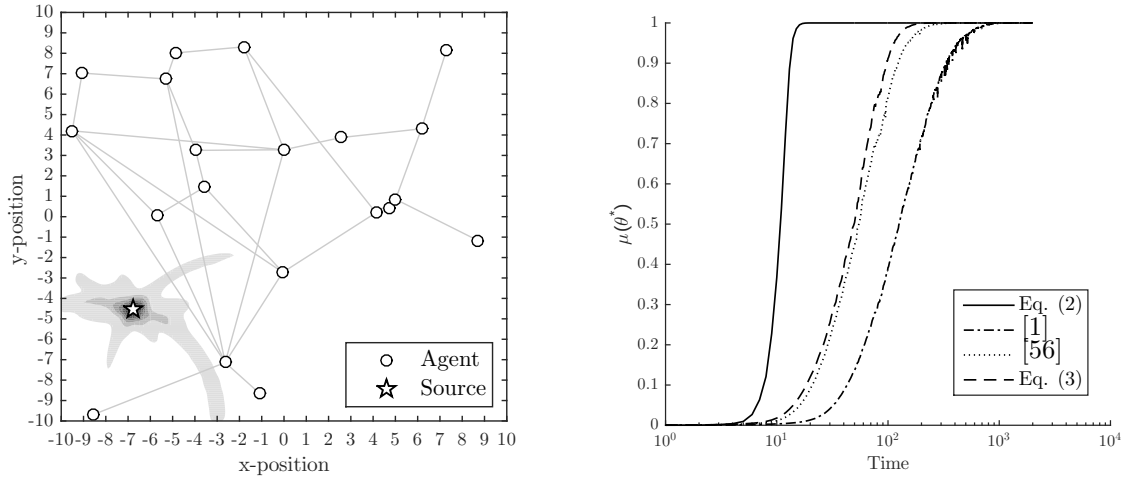


Figure 3.7: Belief distribution of one agent over the hypotheses grid. Darker shades of gray indicate higher beliefs on the corresponding hypothesis.

a factor of n improvement in the convergence rate with respect to the number of agents in comparison with that of the existing algorithms.

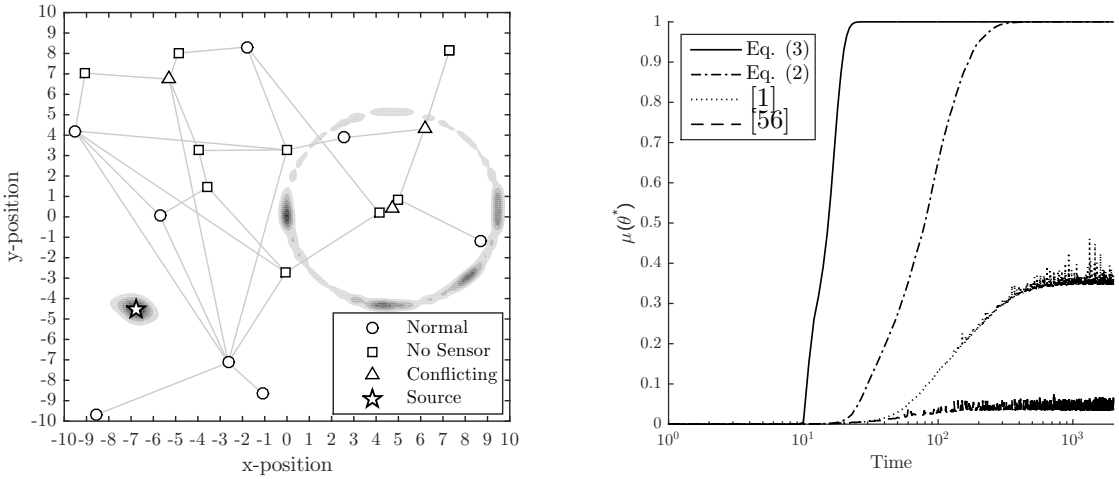
We proposed a new update rule for the problem of distributed non-Bayesian learning on time-varying directed graphs with conflicting hypotheses. We show that the beliefs of all agents concentrate around an optimal set of hypotheses explicitly characterized as the solution to an optimization problem. This optimization problem consists on finding a probability distribution (from a parametrized family of distributions) closest to the unknown distribution of the observations, and it needs to be solved by the agents interacting over a sequence of network and using the local information only. The proposed algorithm also guarantees that after a finite time, that depends on the network structure, all agents will learn at a network-independent rate that is the average of the agents' individual learning abilities. We refer to this as a "balanced" behavior since all agents are weighted equally even if its connectivity is different. This result guarantees certain robustness properties of the learning process since faulty sensors or adversarial agents will not have any advantage even if they are centrally located.



(a) Network of Agents

(b) Belief of one agent on the optimal hypothesis

Figure 3.8: Network of agents and belief of one agent on the optimal hypothesis. (a) A network of agents as well as the belief distribution over the hypothesis set (a grid in the x, y location). Darker shade of gray indicates higher beliefs on the corresponding hypothesis (point in the hypotheses grid). (b) Belief evolution on the optimal hypothesis θ^* for different belief update protocols.



(a) Network of agents

(b) Belief of one agent on the optimal hypothesis

Figure 3.9: Network of normal, faulty and no-sensor agents and belief of one agent on the optimal hypothesis. (a) A network of heterogeneous agents. $\triangle$ indicates agents whose observations have been modified such that the optimal hypothesis is the $(0, 0)$ point in the grid. $\square$ indicates agents for whom all hypotheses are observationally equivalent (i.e. no data is measured). $\circ$ indicates regular agents with correct observation models and informative hypothesis. (b) Belief evolution on the optimal hypothesis θ^* for different belief update protocols.

CHAPTER 4

DISTRIBUTED LEARNING FOR COOPERATIVE INFERENCE

In this chapter, we build upon the work in [151] on non-asymptotic behaviors of Bayesian estimators to derive new non-asymptotic concentration results for distributed learning algorithms. In contrast to Chapter 4, which assumes a finite hypothesis set, in this chapter we extend the framework to compact sets of hypotheses. Our results show that in general, the network structure will induce a transient time after which all agents learn at a network-independent rate, and this rate is geometric.

4.1 Revisit Concentration for a Finite Number of Hypotheses

We now turn to proving a concentration result when the set Θ of hypotheses is finite. We will show the exponential convergence of beliefs on a Hellinger ball around the true hypothesis θ^*. The purpose is to introduce the techniques gently. We will use the techniques later in the case of a compact set of hypotheses.

Naturally, we need some assumptions on the matrix A. For one thing, the matrix A has to be "compatible" with the underlying graph, in that information from node i should not affect node j if there is no edge from i to j in $\mathcal{G}$. At the other extreme, we want to rule out the possibility that A is the identity matrix, which in terms of Eq. (3.7) means nodes do not talk to their neighbors. Formally, we let Assumption 2 hold, which is stronger than Assumption 1 but will be sufficient to illustrate the behavior of the proposed algorithm.

We equip the set of all probability distributions $\mathscr{P}$ over the parameter set with the Hellinger distance to obtain the *metric* space $(\mathscr{P}, h)$. The metric space induces a topology, where we can define an open ball $\mathcal{B}_r(\theta)$ with a radius $r > 0$ centered at a point $\theta \in \Theta$, which we use to construct a special covering of subsets $B \subset \mathscr{P}$.

Definition 16. *Define an n-Hellinger ball of radius r centered at θ as*

$$\mathcal{B}_r(\theta) = \left\{ \hat{\theta} \in \Theta \,\middle|\, \sqrt{\frac{1}{n}\sum_{i=1}^{n} h^2\left(P_\theta^i, P_{\hat{\theta}}^i\right)} \leq r \right\}.$$

Additionally, when no center is specified, it should be assumed that it refers to θ^*, i.e. $\mathcal{B}_r = \mathcal{B}_r(\theta^*)$.

Given an n-Hellinger ball of radius r, we will use the following notation for a covering of its complement $\mathcal{B}_r^c$. Specifically, we are going to express $\mathcal{B}_r^c$ as the union of finite disjoint and concentric annuli. Let $r > 0$ and $\{r_l\}$ be a finite strictly decreasing sequence such that $r_1 = 1$ and $r_L = r$ and express the set $\mathcal{B}_r^c$ as the union of annuli generated by the sequence $\{r_l\}$ as

$$\mathcal{B}_r^c = \bigcup_{l=1}^{L-1} \mathcal{F}_l,$$

where $\mathcal{F}_l = \mathcal{B}_{r_l} \setminus \mathcal{B}_{r_{l+1}}$.

When the number of hypotheses is finite, the density update in Eq. (3.7) can be written in a simpler form for discrete beliefs over the parameter space Θ as

$$\mu_{k+1}^i(\theta) \propto p_\theta^i(x_{k+1}^i) \prod_{j=1}^{n} (\mu_k^j(\theta))^{a_{ij}}. \tag{4.1}$$

We will fix the radius r, and our goal will be to prove a concentration result for a Hellinger ball of radius r around the optimal hypothesis θ^*. We start by partitioning the complement of this ball, i.e., $\mathcal{B}_r^c$, as described above into the annuli $\mathcal{F}_l$. We introduce the notation $\mathcal{N}_l$ to denote the number of hypotheses within the annulus $\mathcal{F}_l$. We refer the reader to Fig. 4.1 which shows a set of probability distributions, represented as black dots, where the true distribution P is represented by a star.

Given that the number of hypotheses is finite, there exists an $\alpha > 0$ such that $\rho(P_{\theta_1}^i, P_{\theta_2}^i) > \alpha$ for any $\theta_1, \theta_2 \in \Theta$ and $i = 1, \dots, n$, where the separation between hypotheses is defined in terms of the Hellinger affinity between two distributions Q and P, given by

$$\rho(Q, P) = 1 - h^2(Q, P).$$

We are now ready to state our first result as a lemma that bounds concentration of aggregated log-likelihood ratios.

Lemma 26. *Let Assumption 2 hold. Given a set of independent random variables $\{X_t^i\}$ such that $X_t^i \sim P^i$ for $i = 1, \dots, n$ and $t = 1, \dots, k$, a set of distributions $\{Q^i\}$ where P^i*

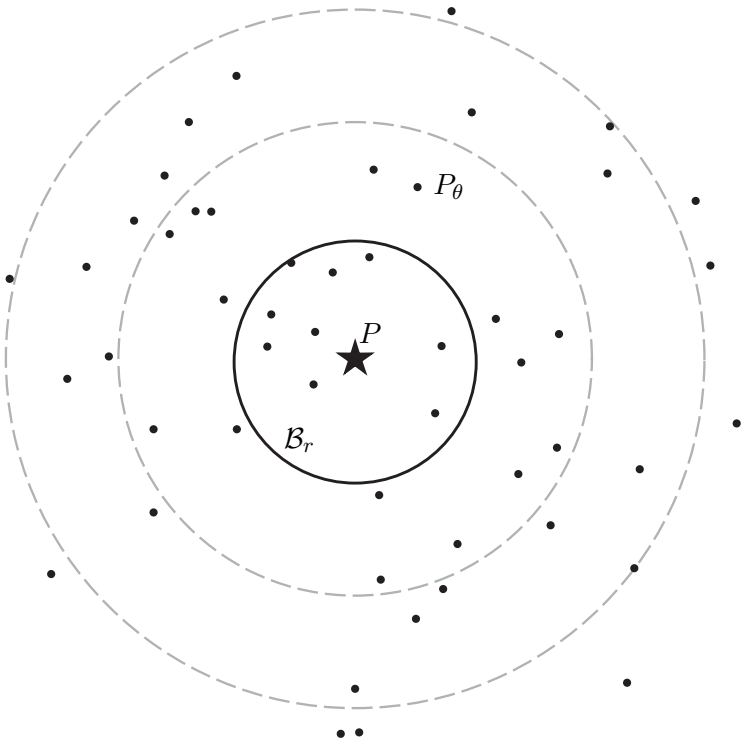


Figure 4.1: Creating a covering for a ball $\mathcal{B}_r$. ★ represents the correct hypothesis $\boldsymbol{P}_{\theta^*}$, • indicates the location of other hypotheses and the dash lines indicate the boundary of the balls $\mathcal{B}_{r_l}$.

dominates Q^i, then for all $y \in \mathbb{R}$,

$$\mathbb{P}\left[\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{dQ^j}{dP^j}(X_t^j) \geq y\right] \leq \exp\left(-\frac{y}{2} + \frac{4\log n}{1-\delta} - k\frac{1}{n}\sum_{j=1}^{n}h^2(Q^j, P^j)\right).$$

Proof. By the Markov inequality and Jensen's inequality we have

$$\begin{aligned}
\mathbb{P}\left[\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{dQ^j}{dP^j}(X_t^j) \geq y\right] &\leq \exp\left(-\frac{y}{2}\right)\mathbb{E}\left[\prod_{t=1}^{k}\prod_{j=1}^{n}\sqrt{\left(\frac{dQ^j}{dP^j}(X_t^j)\right)^{[A^{k-t}]_{ij}}}\right] \\
&\leq \exp\left(-\frac{y}{2}\right)\prod_{t=1}^{k}\prod_{j=1}^{n}\mathbb{E}\left[\sqrt{\left(\frac{dQ^j}{dP^j}(X_t^j)\right)}\right]^{[A^{k-t}]_{ij}} \\
&= \exp\left(-\frac{y}{2}\right)\prod_{t=1}^{k}\prod_{j=1}^{n}\rho(Q^j, P^j)^{[A^{k-t}]_{ij}},
\end{aligned}$$

where the last inequality follows from the definition of the Hellinger affinity function $\rho(Q, P)$.

Moreover, it follows from $\rho(Q^j, P^j) = 1 - h^2(Q^j, P^j)$ and $1 - x \leq \exp(-x)$ for $x \in [0, 1]$

that

$$\prod_{t=1}^{k}\prod_{j=1}^{n}\rho(Q^j,P^j)^{[A^{k-t}]_{ij}}\leq\exp\left(-\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}h^2(Q^j,P^j)\right).$$

Now, by adding and subtracting $\sum_{t=1}^{k}\frac{1}{n}\sum_{j=1}^{n}h^2(Q^j,P^j)$ we have

$$\begin{aligned}
&\mathbb{P}\left[\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{dQ^j}{dP^j}(X_t^j)\geq y\right]\\
&\leq\exp\left(-\frac{y}{2}-\sum_{t=1}^{k}\sum_{j=1}^{n}\left([A^{k-t}]_{ij}-\frac{1}{n}\right)h^2(Q^j,P^j)-\frac{k}{n}\sum_{j=1}^{n}h^2(Q^j,P^j)\right)\\
&\leq\exp\left(-\frac{y}{2}+\frac{4\log n}{1-\delta}-\frac{k}{n}\sum_{j=1}^{n}h^2(Q^j,P^j)\right).
\end{aligned}$$

Finally, the last line above follows from Lemma 4 applied on the second term inside the exponential. □

We are now ready to state our first main result, which bounds concentration of Eq. (4.1) around the optimal hypothesis for a finite hypothesis set Θ. The following theorem shows that the beliefs of all agents will concentrate around the Hellinger ball $\mathcal{B}_r$ at an exponential rate.

Theorem 27. *Let Assumption 2 hold, and let $\sigma\in(0,1)$ be a desired probability tolerance. Then, the belief sequences $\{\mu_k^i\}$, $i\in V$ that are generated by the update rule in Eq. (4.1), with initial beliefs such that $\mu_0^i(\theta^*)>\epsilon$ for all i, have the following property: For any radius $r>0$ with probability $1-\sigma$,*

$$\mu_{k+1}^i\left(\mathcal{B}_r\right)\geq 1-\frac{1}{\epsilon}\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-kr_{l+1}^2\right)\qquad\text{for all } i \text{ and all } k\geq N,$$

where

$$N=\inf\left\{t\geq 1\,\middle|\,\exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-tr_{l+1}^2\right)<\sigma\right\},$$

and δ as defined in Lemma 4.

Proof. We are going to focus on bounding the beliefs of a measurable set B, such that

$\theta^* \in B$. For such a set, it follows from Eq. (4.1) that

$$\begin{aligned}
\mu_k^i(B) &= \frac{1}{Z_k^i} \sum_{\theta \in B} \prod_{j=1}^{n} \mu_0^j(\theta)^{\left[A^k\right]_{ij}} \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(X_t^j)^{\left[A^{k-t}\right]_{ij}} \\
&= \left(1 + \frac{\sum_{\theta \in B^c} \prod_{j=1}^{n} \left(\frac{\mu_0^j(\theta)}{\mu_0^j(\theta^*)} \right)^{\left[A^k\right]_{ij}} \prod_{t=1}^{k} \prod_{j=1}^{n} \left(\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \right)^{\left[A^{k-t}\right]_{ij}}}{\sum_{\theta \in B} \prod_{j=1}^{n} \left(\frac{\mu_0^j(\theta)}{\mu_0^j(\theta^*)} \right)^{\left[A^k\right]_{ij}} \prod_{t=1}^{k} \prod_{j=1}^{n} \left(\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \right)^{\left[A^{k-t}\right]_{ij}}} \right)^{-1} \\
&\geq 1 - \sum_{\theta \in B^c} \prod_{j=1}^{n} \left(\frac{\mu_0^j(\theta)}{\mu_0^j(\theta^*)} \right)^{\left[A^k\right]_{ij}} \prod_{t=1}^{k} \prod_{j=1}^{n} \left(\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \right)^{\left[A^{k-t}\right]_{ij}},
\end{aligned}$$

where Z_k^i is the appropriate normalization constant.

Moreover, from the assumption that $\mu_0^i(\theta^*) > \epsilon$ for all $i = 1, \ldots, n$, it follows that

$$\mu_k^i(B) \geq 1 - \frac{1}{\epsilon} \sum_{\theta \in B^c} \prod_{t=1}^{k} \prod_{j=1}^{n} \left(\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \right)^{\left[A^{k-t}\right]_{ij}}. \tag{4.2}$$

The relation in Eq. (4.2) describes the iterative averaging of products of density functions, for which we can use Lemma 26 with $Q = P_\theta$ and $P = P_{\theta^*}$. Then,

$$\mathbb{P}\left[\sup_{\theta \in B^c} \sum_{t=1}^{k} \sum_{j=1}^{n} [A^{k-t}]_{ij} \log \frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \geq y \right] \leq \sum_{\theta \in B^c} \exp\left(-\frac{y}{2} + \frac{4 \log n}{1 - \delta} - \frac{k}{n} \sum_{j=1}^{n} h^2(P_\theta^j, P^j) \right)$$

and by setting $y = -\frac{k}{n} \sum_{j=1}^{n} h^2(P_\theta^j, P^j)$ we obtain

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\theta \in B^c} \sum_{t=1}^{k} \sum_{j=1}^{n} [A^{k-t}]_{ij} \log \frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \geq -\frac{k}{n} \sum_{j=1}^{n} h^2(P_\theta^j, P^j) \right] \\
&\qquad \leq \exp\left(\frac{4 \log n}{1 - \delta} \right) \sum_{\theta \in B^c} \exp\left(-\frac{k}{2n} \sum_{j=1}^{n} h^2(P_\theta^j, P^j) \right).
\end{aligned}$$

Now, we let the set B be the Hellinger ball of a radius r centered at θ^* and define a cover (as described above) to exploit the representation of $\mathcal{B}_r^c$ as the union of concentric Hellinger annuli, for which we have

$$\mathbb{P}\left[\sup_{\theta \in B^c} \sum_{t=1}^{k} \sum_{j=1}^{n} [A^{k-t}]_{ij} \log \frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \geq -\frac{k}{n} \sum_{j=1}^{n} h^2(P_\theta^j, P^j) \right]$$

$$\leq \exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}\sum_{\theta\in\mathcal{F}_l}\exp\left(-\frac{k}{2}\frac{1}{n}\sum_{j=1}^{n}h^2(P_\theta^j,P^j)\right)$$
$$\leq \exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-\frac{k}{2}r_{l+1}^2\right).$$

We are interested in finding a value of k large enough such that the above probability is below σ. Thus, define the value of N as

$$N=\inf\left\{t\geq 1\middle|\exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-tr_{l+1}^2\right)<\sigma\right\}.$$

It follows that for all $k\geq N$ with probability $1-\sigma$, for all $\theta\in\mathcal{B}_r^c$

$$\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)}\leq -\frac{k}{n}\sum_{j=1}^{n}h^2(P_\theta^j,P^j).$$

Thus, from Eq. (4.2) with probability $1-\sigma$ we have

$$\begin{aligned}\mu_k^i(\mathcal{B}_r)&\geq 1-\frac{1}{\epsilon}\sum_{\theta\in\mathcal{B}_r^c}\exp\left(-\frac{k}{n}\sum_{j=1}^{n}h^2(P_\theta^j,P^j)\right)\\&=1-\frac{1}{\epsilon}\sum_{l=1}^{L-1}\sum_{\theta\in\mathcal{F}_l}\exp\left(-\frac{k}{n}\sum_{j=1}^{n}h^2(P_\theta^j,P^j)\right)\\&\geq 1-\frac{1}{\epsilon}\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-kr_{l+1}^2\right).\end{aligned}$$

□

Note that in general the belief concentration rate described in Theorem 27 depends on the geometry of the hypotheses set and how are they distributed on the parameter space. Corollary 28 describes the scenario where the sequence $\{r_l\}$ is such that $L=2$, so $r_1=1$ and $r_2=r$.

Corollary 28. *Let Assumption 2 hold, and let $\sigma\in(0,1)$ be a desired probability tolerance. Then, the belief sequences $\{\mu_k^i\}$, $i\in V$ that are generated by the update rule in Eq. (4.1), with initial beliefs such that $\mu_0^i(\theta^*)>\epsilon$ for all i, have the following property: For any radius*

$r > 0$ with probability $1 - \sigma$,

$$\mu_{k+1}^{i}\left(\mathcal{B}_r\right) \geq 1 - \frac{2}{\epsilon}\left(\log\frac{\mathcal{N}}{\sigma} + \frac{4\log n}{1-\delta} - kr^2\right),$$

where $\mathcal{N}$ is the number of hypotheses outside $\mathcal{B}_r$ and δ as defined in Lemma 4.

4.2 Concentration for Compact Hypotheses Sets

Next, we consider the case when the hypothesis set Θ is a compact subset of $\mathbb{R}^d$. We will now additionally require the map from Θ to $\prod_{i=1}^{n} P_\theta^i$ be continuous (where the topology on the space of distributions comes from the Hellinger metric). This will be useful in defining coverings, which will be made clear shortly.

Definition 17. *Let (M, d) be a metric space. A subset $S \subseteq M$ is called ε-separated with $\varepsilon > 0$ if $d(x, y) \geq \varepsilon$ for any $x, y \in S$. Moreover, for a set $B \subseteq M$, let $N_B(\varepsilon)$ be the smallest number of Hellinger balls with centers in S of radius ε needed to cover the set B, i.e., such that $B \subseteq \bigcup_{m\in S} \mathcal{B}_\varepsilon(m)$.*

As before, given a decreasing sequence $1 = r_1 \geq r_2 \geq \cdots \geq r_L = r$, we will define the annulus $\mathcal{F}_l$ to be $\mathcal{F}_l = \mathcal{B}_{r_l} \setminus \mathcal{B}_{r_{l+1}}$. Furthermore, S_{ε_l} will denote maximal ε_l-separated subset of $\mathcal{F}$. Finally, $K_l = |S_{\varepsilon_l}|$.

We note that, as a consequence of our assumption that the map from Θ to $\prod_{i=1}^{n} P_\theta^i$ is continuous, we have that each K_l is finite (since the image of a compact set under a continuous map is compact). Thus, we have the following covering of $\mathcal{B}_r^c$:

$$\mathcal{B}_r^c = \bigcup_{l=1}^{L-1} \bigcup_{m\in S_{\varepsilon_l}} \mathcal{F}_{l,m},$$

where each $\mathcal{F}_{l,m}$ is the intersection of a ball centered at an element in S_{ε_l} with $\mathcal{F}_l$. Figure 4.2 shows the elements of a covering for a set $\mathcal{B}_r^c$. The cluster of circles at the top right corner represents the balls $\mathcal{B}_{\varepsilon_l}$ and, for a specific case in the left of the image, we illustrate the set $\mathcal{F}_{l,m}$.

Example 2. *We continue our previous Example 1. Suppose we are interested in analyzing the concentration of the beliefs around the true parameter θ^* on a Euclidean ball of radius 0.05; that is, we want to see the total mass on the set $[0.45, 0.55]$. This in turn represents a Hellinger ball of radius $r = 0.001254$. For this choice of r, we propose a covering where $r_1 = 1$, $r_2 = 1/2$, $r_3 = 1/4$, ..., $r_{10} = 1/512$, $r_{11} = r$.*

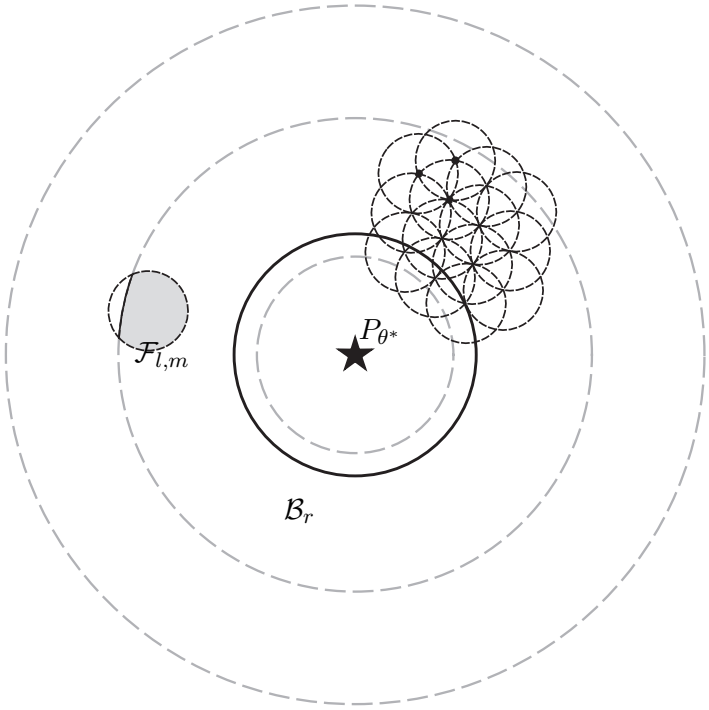


Figure 4.2: Creating a covering for a set $\mathcal{B}_r$. ★ represents the correct hypothesis $\boldsymbol{P}_{\theta^*}$.

Figure 4.3 shows the Hellinger distance between the hypotheses p_θ and the optimal one p_{θ^*}. Specifically, the x-axis is the value of θ, and the y-axis shows the Hellinger distance between the distributions. Figure 4.3 also shows the covering we defined before, as horizontal lines for each value of the sequence r_l, which in turn defines the annulus $\mathcal{F}_l$. The Hellinger ball of radius r is also shown, with the corresponding subset of Θ where we want to analyze the belief concentration.

In this example, the parameter has dimension 1. The number of balls needed to cover each annulus can be seen to be 2, i.e., we only need 2 balls of radius $r_l/2$ to cover the annulus $\mathcal{F}_l$. Thus, $K_l = 2$ for $1 \leq l \leq L-1$.

Without loss of generality, we will make the following technical assumption that will be technically convenient for the analysis of the concentration of beliefs on compact sets.

Assumption 5. *For every $i = 1, \ldots, n$ and all θ, it holds that $p^i_\theta(x) \leq 1$ almost everywhere.*

Let us give an example before explaining the reasoning behind this assertion. Let us assume there is just one agent, and say $X \sim P$ is Gaussian with mean $\theta^* = 5$ and variance 0.01. Our model is $P_\theta = \mathcal{N}(\theta, 0.01)$ for $\theta \in \Theta = [0, 10]$. Because the variance is small, the density values are larger than 1. Instead let us multiply all our observations by 10. We will

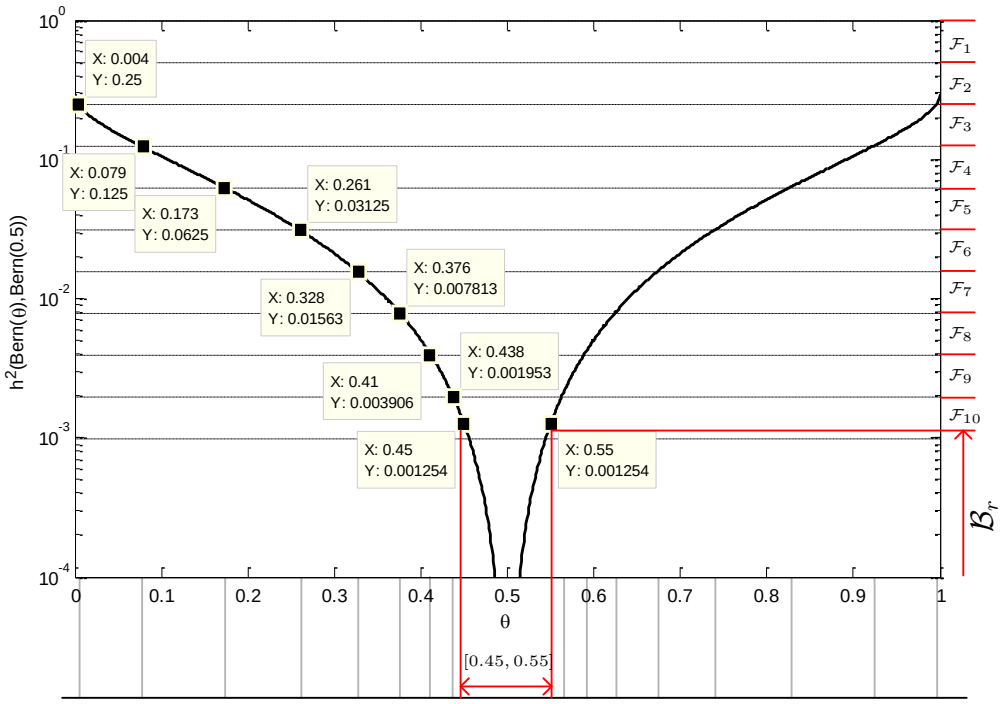


Figure 4.3: Hellinger distance of the density p_θ to the optimal density p_{θ^*}.

then have that our observations come from $10X$, which indeed has density upper bounded by one. In turn our model now should be $Q_\theta = \mathcal{N}(10\theta, 1)$ or, alternatively, $Q_\theta = \mathcal{N}(\theta, 1)$ for $\theta \in \hat{\Theta} = [0, 100]$.

We note that this modification does not come without cost. As in the case of countable hypotheses, our convergence rates will depend on α, defined to be a positive number such that $\rho(P_{\theta^1}, P_{\theta^2}) > \alpha$ for any θ^1 and θ^2. The process we have sketched out proportionally decreases the parameter α.

In the general case, if each agent observes $X_t^j \sim P^j$, then there exists a large enough constant $M > 1$ such that $MX_t^j \sim Q^j$ where the density of Q^j is at most 1. We can then have agents multiply their measurements by M and redefine the densities to account for this scaling.

We next provide a concentration result for the logarithmic likelihood of a ratio of densities, which will serve the same technical function as Lemma 26 in the countable hypothesis case. We begin by defining two measures. For a hypothesis θ and a measurable set $B \subseteq \Theta$, let $\boldsymbol{P}_B^{\otimes k}$ be the probability distribution with density (i.e., Radon-Nikodym derivative with respect to

$\lambda^{\otimes nk}$),

$$g_B(\boldsymbol{x}^k) = \frac{1}{\mu_0(B)} \int_B \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(x_t^j) d\mu_0(\theta). \tag{4.3}$$

Similarly, let $\bar{\boldsymbol{P}}_B^{\otimes k}$ be the measure with density

$$\bar{g}_B(\boldsymbol{x}^k) = \frac{1}{\mu_0(B)} \int_B \prod_{t=1}^{k} \prod_{j=1}^{n} (p_\theta^j(x_t^j))^{[A^{k-t}]_{ij}} d\mu_0(\theta). \tag{4.4}$$

Note that $\bar{\boldsymbol{P}}_B^{\otimes k}$'s are not probability distributions due to the exponential weights. Nonetheless, they are bounded and positive. The next lemma shows the concentration of the logarithmic ratio of two weighted densities, as defined in Eq. (4.4), for two different sets B_1 and B_2, in terms of the probability distribution $\boldsymbol{P}_{B_1}^{\otimes k}$.

Lemma 29. *Let Assumptions 2 and 5 hold. Consider two measurable sets $B_1, B_2 \subset \Theta$, both with positive measures, and assume that $B_1 \subset \mathcal{B}_{r^1}(\theta^1)$ and $B_2 \subset \mathcal{B}_{r^2}(\theta^2)$ where $\mathcal{B}_{r^1}(\theta^1)$ and $\mathcal{B}_{r^2}(\theta^2)$ are disjoint. Then, for all $y \in \mathbb{R}$*

$$\begin{aligned}
&\mathbb{P}_{B_1}\left[\log \frac{\bar{g}_{B_2}(\boldsymbol{X}^k)}{\bar{g}_{B_1}(\boldsymbol{X}^k)} \geq y\right] \\
&\leq \exp(-y/2) \exp\left(\log\left(\frac{1}{\alpha}\right) \frac{4\log n}{1-\delta}\right) \exp\left(-k\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n} h^2(P_{\theta^1}^j, P_{\theta^2}^j)} - r^1 - r^2\right)^2\right),
\end{aligned}$$

where $\mathbb{P}_{B_1}$ is the probability measure that gives $\boldsymbol{X}^k$ a distribution $\boldsymbol{P}_{B_1}^{\otimes k}$ with density g_{B_1} as defined in Eq. (4.3).

Proof. By the Markov inequality, it follows that

$$\begin{aligned}
\mathbb{P}_{B_1}\left[\log \frac{\bar{g}_{B_2}(\boldsymbol{X}^k)}{\bar{g}_{B_1}(\boldsymbol{X}^k)} \geq y\right] &\leq \exp(-y/2)\mathbb{E}_{B_1}\left[\sqrt{\frac{\bar{g}_{B_2}(\boldsymbol{X}^k)}{\bar{g}_{B_1}(\boldsymbol{X}^k)}}\right] \\
&= \exp(-y/2) \int_{\boldsymbol{\mathcal{X}}^k} \sqrt{\frac{\bar{g}_{B_2}(\boldsymbol{x}^k)}{\bar{g}_{B_1}(\boldsymbol{x}^k)}} g_{B_1}(\boldsymbol{x}^k) d\lambda^{\otimes kn}(\boldsymbol{x}^k).
\end{aligned}$$

Now, by Assumption 5 it follows that $g_B \leq \bar{g}_B$ almost everywhere. Thus, we have

$$\mathbb{P}_{B_1}\left[\log \frac{\bar{g}_{B_2}(\boldsymbol{X}^k)}{\bar{g}_{B_1}(\boldsymbol{X}^k)} \geq y\right] \leq \exp(-y/2) \int_{\boldsymbol{\mathcal{X}}^k} \sqrt{\bar{g}_{B_2}(\boldsymbol{x}^k)} \sqrt{\bar{g}_{B_1}(\boldsymbol{x}^k)} d\lambda^{\otimes kn}(\boldsymbol{x}^k)$$

$$\leq \exp(-y/2)\rho\left(\bar{\boldsymbol{P}}_{B_2}^{\otimes k}, \bar{\boldsymbol{P}}_{B_1}^{\otimes k}\right),$$

where we are interpreting the definition of the Hellinger affinity function $\rho(\cdot,\cdot)$ as a function of two bounded positive measures, not necessarily probability measures.

At this point, we can follow the same argument as in Lemma 2 in [177], page 477, where the Hellinger affinity of two members of the convex hull of sets of probability distributions is shown to be less than the product of the Hellinger affinity of the factors. In our particular case, the measures $\bar{\boldsymbol{P}}_B^{\otimes k}$ are not probability distributions. Nonetheless, the same disintegration argument holds. Thus, we obtain

$$\rho\left(\bar{\boldsymbol{P}}_{B_2}^{\otimes k}, \bar{\boldsymbol{P}}_{B_1}^{\otimes k}\right) \leq \prod_{t=1}^{k}\prod_{j=1}^{n} \rho\left(\bar{P}_{B_2}^{j}, \bar{P}_{B_1}^{j}\right),$$

where $\bar{P}_B^j$ is the measure with Radon-Nikodym derivative $\bar{g}_B(x) = \frac{1}{\mu_0(B)}\int\limits_B (p_\theta^j(x))^{[A^{k-t}]_{ij}} d\mu_0(\theta)$ with respect to λ.

In addition, by Jensen's inequality[1], with $x^{[A^{k-t}]_{ij}}$ being a concave function and $1/\mu_0(B)\int_B d\mu_0 = 1$, we have that

$$\bar{g}_B(x) \leq \left(\frac{1}{\mu_0(B)}\int\limits_B p_\theta^j(x) d\mu_0(\theta)\right)^{[A^{k-t}]_{ij}}.$$

Thus,

$$\mathbb{P}_{B_1}\left[\log\frac{\bar{g}_{B_2}(\boldsymbol{X}^k)}{\bar{g}_{B_1}(\boldsymbol{X}^k)} \geq y\right] \leq \exp(-y/2)\prod_{t=1}^{k}\prod_{j=1}^{n}\rho(P_{B_1}^j, P_{B_2}^j)^{[A^{k-t}]_{ij}},$$

where P_B^j is the probability distribution associated with the density $\frac{1}{\mu_0(B)}\int\limits_B p_\theta^j(x) d\mu_0(\theta)$.

The compactness of Θ guarantees that $\rho(P_{B_1}^j, P_{B_2}^j) > \alpha$ for some positive α, thus similarly as in Lemma 26, we have that

$$\begin{aligned}\mathbb{P}_{B_1}\left[\log\frac{\bar{g}_{B_2}(\boldsymbol{X}^k)}{\bar{g}_{B_1}(\boldsymbol{X}^k)} \geq y\right] &\leq \exp(-y/2)\exp\left(\log\left(\frac{1}{\alpha}\right)\frac{4\log n}{1-\delta}\right)\prod_{t=1}^{k}\prod_{j=1}^{n}\rho(P_{B_1}^j, P_{B_2}^j)^{1/n} \\ &\leq \exp(-y/2)\exp\left(\log\left(\frac{1}{\alpha}\right)\frac{4\log n}{1-\delta}\right)\exp\left(-\frac{k}{n}\sum_{j=1}^{n} h^2(P_{B_1}^j, P_{B_2}^j)\right).\end{aligned}$$

[1]For a concave function ϕ and $\int_\Omega f(x)dx = 1$, it holds that $\int_\Omega \phi(g(x))f(x)dx \leq \phi\left(\int_\Omega g(x)f(x)\right)$.

Finally, by using the metric defined for the n-Hellinger ball and the fact that for a metric $d(A,B)$ for two sets A and B $d(A,B) = \inf_{x\in A, y\in B} d(x,y)$ we have

$$\begin{aligned}
&\mathbb{P}_{B_1}\left[\log\frac{\bar{g}_{B_2}(\boldsymbol{X}^k)}{\bar{g}_{B_1}(\boldsymbol{X}^k)} \geq y\right] \\
&\leq \exp\left(-\frac{y}{2} + \frac{4\log\left(\frac{1}{\alpha}\right)\log n}{1-\delta}\right)\exp\left(-k\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n} h^2(P^j_{B_1}, P^j_{B_2})}\right)^2\right) \\
&\leq \exp\left(-\frac{y}{2} + \frac{4\log\left(\frac{1}{\alpha}\right)\log n}{1-\delta} - k\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n} h^2\left(P^j_{\theta_1}, P^j_{\theta_2}\right)} - r^1 - r^2\right)^2\right).
\end{aligned}$$

□

Lemma 29 provides a concentration result for the logarithmic ratio between two weighted densities over a pair of subsets B_1 and B_2. The terms involving the auxiliary variable y and the influence of the graph, via δ, are the same as in Lemma 26. Moreover, the rate at which this bound decays exponentially is influenced now by the radius of the two disjoint Hellinger balls where B_1 and B_2 are contained respectively.

The bound provided in Lemma 29 is defined for the random variables $\boldsymbol{X}^k$ having a distribution $\boldsymbol{P}_B^{\otimes k}$. Nonetheless, $\boldsymbol{X}^k$ are distributed according to $\boldsymbol{P}^{\otimes k}$. Therefore, we introduce a lemma that relates the Hellinger affinity of distributions defined over subsets of Θ.

Lemma 30. *Let Assumptions 2 and 5 hold. Consider $\boldsymbol{P}_B^{\otimes k}$ as the distribution with density g_B as defined in Eq. (4.3), for $B \subseteq \mathcal{B}_R$. Then $h(\boldsymbol{P}_B^{\otimes k}, \boldsymbol{P}^{\otimes k}) \leq \sqrt{nk}R$.*

Proof. By Jensen's inequality, we have that

$$\sqrt{g_B(\boldsymbol{x})} \geq \frac{1}{\mu_0(B)}\int_B \sqrt{\prod_{t=1}^{k}\prod_{j=1}^{n} p^j_\theta(x^j_t)}\, d\mu_0(\theta).$$

Then, by definition of the Hellinger affinity, it follows that

$$\rho(\boldsymbol{P}_B^{\otimes k}, \boldsymbol{P}^{\otimes k}) \geq \int_{\boldsymbol{\mathcal{X}}^{\otimes k}} \sqrt{\prod_{t=1}^{k}\prod_{j=1}^{n} p^j(x^j_t)}\left(\frac{1}{\mu_0(B)}\int_B \sqrt{\prod_{t=1}^{k}\prod_{j=1}^{n} p^j_\theta(x^j_t)}\, d\mu_0(\theta)\right) d\lambda^{\otimes nk}(\boldsymbol{x}).$$

By using the Fubini-Tonelli theorem, we obtain

$$\begin{aligned}\rho(\boldsymbol{P}_B^{\otimes k}, \boldsymbol{P}^{\otimes k}) &\geq \frac{1}{\mu_0(B)} \int_B \int_{\boldsymbol{\mathcal{X}}^{\otimes k}} \sqrt{\prod_{t=1}^{k}\prod_{j=1}^{n} p^j(x_t^j)} \sqrt{\prod_{t=1}^{k}\prod_{j=1}^{n} p_\theta^j(x_t^j)} d\lambda^{\otimes nk}(\boldsymbol{x}) d\mu_0(\theta) \\ &= \frac{1}{\mu_0(B)} \int_B \prod_{t=1}^{k}\prod_{j=1}^{n} \rho(P^j, P_\theta^j) d\mu_0(\theta) \\ &= \frac{1}{\mu_0(B)} \int_B \prod_{t=1}^{k}\prod_{j=1}^{n} \left(1 - h^2(P^j, P_\theta^j)\right) d\mu_0(\theta).\end{aligned}$$

Finally, by the Weierstrass product inequality it follows that

$$\begin{aligned}\rho(\boldsymbol{P}_B^{\otimes k}, \boldsymbol{P}^{\otimes k}) &\geq \frac{1}{\mu_0(B)} \int_B \left(1 - \sum_{t=1}^{k}\sum_{j=1}^{n} h^2(P^j, P_\theta^j)\right) d\mu_0(\theta) \\ &= \frac{1}{\mu_0(B)} \int_B \left(1 - n\frac{1}{n}\sum_{t=1}^{k}\sum_{j=1}^{n} h^2(P^j, P_\theta^j)\right) d\mu_0(\theta) \\ &\geq \frac{1}{\mu_0(B)} \int_B \left(1 - nkR^2\right) d\mu_0(\theta),\end{aligned}$$

where the last line follows by the fact that any density $\boldsymbol{P}_\theta$, inside the n-Hellinger ball defined in the statement of the lemma, is at most at a distance R to $\boldsymbol{P}$.

□

Finally, before presenting our main result for compact sets of hypotheses, we will state an assumption regarding the necessary mass all agents should have around the correct hypothesis θ^* in their initial beliefs.

Assumption 6. *The initial beliefs of all agents are equal. Moreover, they have the following property: for any constants $C \in (0, 1]$ and $r \in (0, 1]$ there exists a finite positive integer K, such that*

$$\mu_0\left(\mathcal{B}_{\frac{C}{\sqrt{k}}}\right) \geq \exp\left(-k\frac{r^2}{32}\right) \qquad \text{for all } k \geq K.$$

Assumption 6 implies that the initial beliefs should have enough mass around the correct hypothesis θ^* when we consider balls of small radius. Particularly, as we take Hellinger balls of radius decreasing as $O(1/\sqrt{k})$, the corresponding initial beliefs should not decrease faster than $O(\exp(-k))$.

The assumption can almost always be satisfied by taking initial beliefs to be uniform. The

reason is that, in any fixed dimension, the volume of a ball of radius $O(1/\sqrt{k})$ will usually scale as a polynomial in $1/\sqrt{k}$, whereas we only need to lower bound it by a decaying exponential in k. For concreteness, we show how this assumption is satisfied by an example.

Example 3. *Consider a single agent, with a uniform initial belief receiving observations from a standard Gaussian distribution, i.e.* $X_k \sim \mathcal{N}(0,1)$. *The variance is known, and the agent would like to estimate the mean. Thus the models are* $P_\theta = \mathcal{N}(\theta, 1)$. *Now, the Hellinger distance can be explicitly written as*

$$h^2(P, P_\theta) = 1 - \exp\left(-\frac{1}{4}\theta^2\right).$$

Therefore, the Hellinger balls of radius $1/\sqrt{k}$ *will correspond to Euclidean balls in the parameter space of radius*

$$2\sqrt{\log\left(\frac{1}{1-\frac{1}{k}}\right)}.$$

Uniform initial belief indicates that $\mu_0\left(\mathcal{B}_{\frac{C}{\sqrt{k}}}\right) = O(\frac{1}{\sqrt{k}})$, *which can be made larger than* $\exp(-k\frac{r^2}{32})$ *for sufficiently large* k.

We are ready now to state our main result regarding the concentration of beliefs around θ^* for compact sets of hypotheses.

Theorem 31. *Let Assumptions 2, 5 and 6 hold, and let* $\sigma \in (0,1)$ *be a given probability tolerance level. Moreover, for any* $r \in (0,1]$, *let* $\{R_k\}$ *be a decreasing sequence such that for* $k = 1, \ldots,$ $R_k \leq \min\left\{\frac{\sigma}{2\sqrt{2kn}}, \frac{r}{4}\right\}$. *Then, the beliefs* $\{\mu_k^i\}$, $i \in V$, *generated by the update rule in Eq.* (3.7) *have the following property: With probability* $1-\sigma$,

$$\mu_{k+1}^i(\mathcal{B}_r) \geq 1 - \chi \exp\left(-\frac{k}{16}r^2\right), \qquad \text{for all } i \text{ and all } k \geq \max\{N, K\},$$

where

$$N = \inf\left\{t \geq 1 \,\middle|\, \exp\left(\log\left(\frac{1}{\alpha}\right)\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1} K_l \exp\left(-\frac{t}{32}r_{l+1}^2\right) < \frac{\sigma}{2}\right\},$$

with K *as defined in Assumption 6,* $\chi = \sum_{l=1}^{L-1} \exp(-\frac{1}{16}r_{l+1}^2)$ *and* $\delta = 1 - \eta/n^2$, *where* η *is the smallest positive element of the matrix* A.

Proof. Let us start by analyzing the evolution of the beliefs on a measurable set B with $\theta^* \in B$. From Eq. (3.7) we have that

$$\mu_k^i(B) = \frac{\int_B \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}} d\mu_0(\theta)}{\int_\Theta \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}} d\mu_0(\theta)}$$
$$\geq 1 - \frac{\int_{B^c} \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}} d\mu_0(\theta)}{\int_B \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}} d\mu_0(\theta)}.$$

Now let us focus specifically on the case where B is a n-Hellinger ball of radius $r > 0$ with center at θ^*. In addition, since $R_k < r$, we get

$$\mu_k^i(\mathcal{B}_r) \geq 1 - \frac{\int_{\mathcal{B}_r^c} \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}} d\mu_0(\theta)}{\int_{\mathcal{B}_{R_k}} \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}} d\mu_0(\theta)}.$$

Our goal will be to use the concentration result in Lemma 29. Thus, we can multiply and divide the denominator on the right-hand side of the above inequality by $\mu_0(\mathcal{B}_{R_k})$ to obtain

$$\mu_k^i(\mathcal{B}_r) \geq 1 - \frac{\int_{\mathcal{B}_r^c} \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}} d\mu_0(\theta)}{\bar{g}_{\mathcal{B}_{R_k}}(\boldsymbol{X}^k)\mu_0(\mathcal{B}_{R_k})}.$$

Moreover, we use the covering of the set $\mathcal{B}_r^c$ to obtain

$$\mu_k^i(\mathcal{B}_r) \geq 1 - \frac{\sum_{l=1}^{L-1} \sum_{m=1}^{K_l} \int_{\mathcal{F}_{l,m}} \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}} d\mu_0(\theta)}{\bar{g}_{\mathcal{B}_{R_k}}(\boldsymbol{X}^k)\mu_0(\mathcal{B}_{R_k})}$$
$$\geq 1 - \frac{\sum_{l=1}^{L-1} \sum_{m=1}^{K_l} \bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)\mu_0(\mathcal{F}_{l,m})}{\bar{g}_{\mathcal{B}_{R_k}}(\boldsymbol{X}^k)\mu_0(\mathcal{B}_{R_k})}. \tag{4.5}$$

The previous relation defines a ratio between two densities, i.e. $\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)/\bar{g}_{\mathcal{B}_{R_k}}(\boldsymbol{X}^k)$, both for the weighted likelihood product of the observations, where the numerator is defined over the set $\mathcal{F}_{l,m}$ and the denominator with respect to the set $\mathcal{B}_{R_k}$.

Lemma 29 provides a way to bound term $\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)/\bar{g}_{\mathcal{B}_{R_k}}(\boldsymbol{X}^k)$ with high probability, thus

$$\begin{aligned}
&\mathbb{P}_{\mathcal{B}_{R_k}}\left(\left\{\boldsymbol{X}^k \middle| \sup_{l,m} \log \frac{\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)}{\bar{g}_{\mathcal{B}_{R_k}}(\boldsymbol{X}^k)} \geq y\right\}\right) \leq \sum_{l=1}^{L-1}\sum_{m=1}^{K_l} \mathbb{P}_{\mathcal{B}_{R_k}}\left(\log \frac{\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)}{\bar{g}_{\mathcal{B}_{R_k}}(\boldsymbol{X}^k)} \geq y\right) \\
&\leq \sum_{l=1}^{L-1}\sum_{m=1}^{K_l} \exp\left(-\frac{y}{2} + \frac{4\log\left(\frac{1}{\alpha}\right)\log n}{1-\delta} - k\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n} h^2(P_m^j, P^j)} - \delta_l - R_k\right)^2\right) \\
&\leq \sum_{l=1}^{L-1}\sum_{m=1}^{K_l} \exp\left(-\frac{y}{2} + \frac{4\log\left(\frac{1}{\alpha}\right)\log n}{1-\delta} - k\left(r_{l+1} - \delta_l - R_k\right)^2\right),
\end{aligned}$$

where p_m^j is the density of at the point $\theta = m \in S_{\varepsilon_l}$, where S_{ε_l} is the maximal ε_l separated set of $\mathcal{F}_l$ as in Definition 17.

Particularly, let's use the covering proposed in [151], where $\delta_l = r_{l+1}/2$. From this choice of covering, we have that

$$\begin{aligned}
r_{l+1} - \delta_l - R_k &> r_{l+1} - r_{l+1}/2 - r_{l+1}/4 \\
&= r_{l+1}/4,
\end{aligned}$$

where we have used the assumption that $R_k \leq r/4$ or equivalently $R_k \leq r_l/4$ for all $1 \leq l \leq L$.

Thus, we can set $y = -\frac{k}{16} r_{l+1}^2$ and it follows that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{B}_{R_k}}\left(\left\{\boldsymbol{X}^k \middle| \sup_{l,m} \log \frac{\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)}{\bar{g}_{\mathcal{B}_{R_k}}(\boldsymbol{X}^k)} \geq y\right\}\right) \\
&\leq \exp\left(\log\left(\frac{1}{\alpha}\right)\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1} K_l \exp\left(-\frac{k}{16} r_{l+1}^2\right). \qquad (4.6)
\end{aligned}$$

The probability measure in Eq. (4.6) is computed for $\boldsymbol{X}^k$ distributed according to $\boldsymbol{P}_{\mathcal{B}_{R_k}}^{\otimes k}$. Nonetheless, $\boldsymbol{X}^k$ is distributed according to the (slightly different) $\boldsymbol{P}^{\otimes k}$. Our next step is to relate these two measures.

First, we have that for any distribution $\boldsymbol{P}_\theta \in \mathcal{B}_{R_k}$, from the Definition 16 of the n-Hellinger ball, it holds that

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n} h^2(P_\theta^j, P^j)} \leq R_k,$$

and we relate the total variation distance and the Hellinger affinity as in Lemma 1 in [178];

for any measurable set A it holds that

$$\sup_A \left(\boldsymbol{P}^{\otimes k}_{\mathcal{B}_{R_k}}(A) - \boldsymbol{P}^{\otimes k}(A) \right)^2 \leq 1 - \rho^2(\boldsymbol{P}^{\otimes k}_{\mathcal{B}_{R_k}}, \boldsymbol{P}^{\otimes k}),$$

and by definition of the Hellinger affinity we have that

$$\begin{aligned} \sup_A \left(\boldsymbol{P}^{\otimes k}_{\mathcal{B}_{R_k}}(A) - \boldsymbol{P}^{\otimes k}(A) \right)^2 &= 1 - (1 - h^2(\boldsymbol{P}^{\otimes k}_{\mathcal{B}_{R_k}}, \boldsymbol{P}^{\otimes k}))^2 \\ &\leq 2h^2(\boldsymbol{P}^{\otimes k}_{\mathcal{B}_{R_k}}, \boldsymbol{P}^{\otimes k}), \end{aligned}$$

where first we have used the relation that for any $x \in \mathbb{R}$, it holds that $1 - (1 - x^2)^2 < 2x^2$. Then, from Lemma 30 we have that

$$\sup_A \left(P_{\mathcal{B}_{R_k}}(A) - \boldsymbol{P}^{\otimes k}(A) \right)^2 \leq 2knR_k^2.$$

Therefore, by considering the measurable subset

$$\Gamma^k = \left\{ \boldsymbol{X}^k \middle| \sup_{l,m} \log \frac{\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)}{\bar{g}_{\mathcal{B}_{R_k}}(\boldsymbol{X}^k)} \geq -\frac{k}{16} r_{l+1}^2 \right\},$$

we have that

$$\begin{aligned} \mathbb{P}\left(\Gamma^k\right) &< \mathbb{P}_{\mathcal{B}_{R_k}}\left(\Gamma^k\right) + \sqrt{2kn} R_k \\ &\leq \exp\left(\log\left(\frac{1}{\alpha}\right) \frac{4 \log n}{1 - \delta} \right) \sum_{l=1}^{L-1} K_l \exp\left(-\frac{k}{16} r_{l+1}^2 \right) + \frac{\sigma}{2}. \end{aligned}$$

Furthermore, we are interested in finding a large enough k such that the probability described in Eq. (4.6) is at most σ. Thus, we define

$$N \geq \inf \left\{ t \geq 1 \middle| \exp\left(\log\left(\frac{1}{\alpha}\right) \frac{4 \log n}{1 - \delta} \right) \sum_{l=1}^{L-1} K_l \exp\left(-\frac{t}{16} r_{l+1}^2 \right) < \frac{\sigma}{2} \right\}.$$

Moreover, from Eq. (4.5) we obtain that with probability $1 - \sigma$ for all $k \geq N$,

$$\begin{aligned} \mu_k^i(\mathcal{B}_r) &\geq 1 - \sum_{l=1}^{L-1} \sum_{m=1}^{K_l} \exp\left(-\frac{k}{16} r_{l+1}^2 \right) \frac{\mu_0(\mathcal{F}_{l,m})}{\mu_0(\mathcal{B}_{R_k})} \\ &= 1 - \sum_{l=1}^{L-1} \exp\left(-\frac{k}{16} r_{l+1}^2 \right) \frac{\mu_0(\mathcal{F}_l)}{\mu_0(\mathcal{B}_{R_k})} \end{aligned}$$

$$\geq 1 - \frac{1}{\mu_0(\mathcal{B}_{R_k})} \sum_{l=1}^{L-1} \exp\left(-\frac{k}{16} r_{l+1}^2\right).$$

Now, define $\chi = \sum_{l=1}^{L-1} \exp\left(-\frac{1}{16} r_{l+1}^2\right)$, then it follows that

$$\begin{aligned}
\mu_k^i(\mathcal{B}_r) &\geq 1 - \frac{1}{\mu_0(\mathcal{B}_{R_k})} \sum_{l=1}^{L-1} \exp\left(-\frac{k}{16} r_{l+1}^2\right) \\
&= 1 - \frac{1}{\mu_0(\mathcal{B}_{R_k})} \sum_{l=1}^{L-1} \exp\left(-\frac{1}{16} r_{l+1}^2\right) \exp\left(-\frac{k-1}{16} r_{l+1}^2\right) \\
&\geq 1 - \frac{1}{\mu_0(\mathcal{B}_{R_k})} \chi \exp\left(-\frac{k-1}{16} r^2\right),
\end{aligned}$$

where the last inequality follows from $r_l \geq r$ for all $L \leq l \leq 1$. Finally, by Assumption 6 we have that, for all $k \geq K$,

$$\begin{aligned}
\mu_k^i(\mathcal{B}_r) &\geq 1 - \chi \exp(-\frac{k-1}{16} r^2 + \frac{k-1}{32} r^2) \\
&= 1 - \chi \exp(-\frac{k-1}{32} r^2),
\end{aligned}$$

or equivalently $\mu_{k+1}^i(\mathcal{B}_r) \geq 1 - \chi \exp(-\frac{k}{32} r^2)$.

□

Analogous to Theorem 27, Theorem 31 provides a probabilistic concentration result for the agents' beliefs around a Hellinger ball of radius r with center at θ^* for sufficiently large k.

4.3 Cooperative Learning on the Exponential Family

We begin with the observation that, for a general class of models $\{\mathscr{P}^i\}$, the computation of the posterior beliefs μ_{k+1}^i is intractable. Indeed, computation of μ_{k+1}^i involves solving an integral of the form

$$\int_\Theta p_\theta^i(x_{k+1}^i) \prod_{j=1}^n (d\mu_k^j(\theta))^{a_{ij}}. \tag{4.7}$$

There is an entire area of research called *variational Bayes' approximations* dedicated to efficiently approximating integrals that appear in such context [179, 180, 181].

The purpose of this section is to show that for exponential family [182, 183] there are closed-form expressions for the posterior beliefs generated by the proposed distributed inference algorithm.

Definition 18. *The exponential family, for a parameter $\theta = [\theta^1, \theta^2, \dots, \theta^s]'$, is the set of probability distributions whose density can be represented as*

$$p_\theta(x) = H(x) \exp(M(\theta)' T(x))$$

for specific functions $H(\cdot)$, $M(\cdot)$ and $T(\cdot)$ where $M(\theta) = [M(\theta^1), M(\theta^2), \dots, M(\theta^s)]'$ depends on the density parameters and $T(\cdot)$ depends on the observations.

For example, consider a Normal distribution parametrized by its mean θ with known variance σ^2. Then, it holds that

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\theta}{\sigma^2} - \frac{\theta^2}{2\sigma^2}\right) \\ &= \underbrace{\frac{\exp\left(-\frac{x^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}}_{H(x)} \exp\left(\underbrace{\begin{bmatrix}\theta & \theta^2\end{bmatrix}}_{M(\theta)} \underbrace{\begin{bmatrix}\frac{x}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix}}_{T(x)}\right). \end{aligned} \tag{4.8}$$

Among the members of the exponential family, one can find the distributions such as normal, Poisson, exponential, gamma, Bernoulli, and beta, among others [184]. In our case, we will take advantage of the existence of *conjugate priors* for all members of the exponential family. The definition of the conjugate prior is given below.

Definition 19. *Assume that the prior distribution p on a parameter space Θ belongs to the exponential family. Then, the distribution p is referred to as the conjugate prior for a likelihood function $p_\theta(x)$ if the posterior distribution $p(\theta|x) \propto p_\theta(x)p(\theta)$ is in the same family as the prior.*

Definition 19 implies that, if the belief density at some time k is a conjugate prior for our likelihood model, then our belief at time $k+1$ will be of the same class as our prior. For example, if a likelihood function follows a Gaussian form, then having a Gaussian prior will

produce a Gaussian posterior. This property simplifies the structure of the belief update procedure since we can express the evolution of the beliefs generated by the proposed algorithm in Eq. (3.7) by the evolution of the natural parameters of the member of the exponential family it belongs to. Naturally, by induction, if the prior belief at time $k = 0$ is a conjugate prior of the likelihood function, the beliefs for all $k > 0$ will belong to the same exponential family.

We now proceed to provide more details. First, the conjugate prior for a member of the exponential family can be written as

$$p_\chi(M(\theta)) = f(\chi) \exp(M(\theta)'\chi),$$

which is a distribution over the natural parameters M, where χ is a vector of parameters of the conjugate prior. Going back to the example in Eq. (4.8), assume that our prior is a normal distribution on θ with mean $\hat{\theta}$ and variance $\hat{\sigma}^2$, then $\chi = [\hat{\theta} \;\; \hat{\sigma}^2]'$ and

$$\begin{aligned} p_\chi(M) &= \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(\theta-\hat{\theta})^2}{2\hat{\sigma}^2}\right) \\ &= \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{\theta^2}{2\hat{\sigma}^2} + \frac{\theta\hat{\theta}}{\hat{\sigma}^2} - \frac{\hat{\theta}^2}{2\hat{\sigma}^2}\right) \\ &= \underbrace{\frac{\exp\left(-\frac{\hat{\theta}^2}{2\hat{\sigma}^2}\right)}{\sqrt{2\pi\hat{\sigma}^2}}}_{f(\chi)} \exp\left(\underbrace{\begin{bmatrix}\theta & \theta^2\end{bmatrix}}_{M(\theta)} \underbrace{\begin{bmatrix}\frac{\hat{\theta}}{\hat{\sigma}^2} \\ -\frac{1}{2\hat{\sigma}^2}\end{bmatrix}}_{\chi}\right). \end{aligned} \tag{4.9}$$

Then, it can be shown that the posterior distribution, given some observation x, has the same exponential form as the prior with updated parameters as follows:

$$p_{\bar{\chi},\bar{\nu}}(M|x) = p_{\chi+T(x)}(M) \propto p_\theta(x) p_{\chi,\nu}(M|x). \tag{4.10}$$

Particularly, for the example in Eq. (4.8) and Eq. (4.9), the posterior distribution is still normal with parameters

$$\bar{\sigma}^2 = \frac{\sigma^2}{\bar{\nu}} \qquad \text{and} \qquad \bar{\theta} = \frac{\bar{\chi}}{\bar{\nu}}.$$

Now, we are going to exploit the structure of the exponential family of distributions to reformulate the distributed inference algorithm in Eq. (3.7) into an easy-to-implement algorithm in terms of the parametric representation of the beliefs for each agent.

Initially, consider that the set of agents have a belief at time k in the form of a distribution over the parameter space that is a member of the exponential family. That is, assume that each agent i has a belief over the natural parameter M such that

$$d\mu_k^i(M) \propto \exp\left(M'\chi_k^i\right).$$

Then, according to the first step in Eq. (3.7), an agent i needs to compute the weighted geometric average of the beliefs of its neighbors including its own. Given the parametrization in the exponential family, it holds that

$$\begin{aligned}\prod_{j=1}^{n}\left(d\mu_k^j(M)\right)^{a_{ij}} &\propto \prod_{j=1}^{n}\left(\exp\left(M'\chi_k^j\right)\right)^{a_{ij}} \\ &= \exp\left(M'\sum_{j=1}^{n}a_{ij}\chi_k^j\right).\end{aligned}$$

Now, if all agents have beliefs in the same exponential family and they are conjugate priors to their corresponding likelihood functions, then we can write the posterior of agent i as

$$\begin{aligned}d\mu_{k+1}^i(M) &\propto \exp\left(M'\sum_{j=1}^{n}a_{ij}\chi_k^j\right)p_M^i(x_{k+1}^i) \\ &= \exp\left(M'\sum_{j=1}^{n}a_{ij}\chi_k^j-\right)\exp(M'T^i(x_{k+1}^i)) \\ &= \exp\left(M'\left(\sum_{j=1}^{n}a_{ij}\chi_k^j+T^i(x_{k+1}^i)\right)\right) \\ &= \exp\left(M'\chi_{k+1}^i\right).\end{aligned}$$

As an immediate conclusion, it follows that for distributed inference problems when the observation models are members of the exponential family, one can always construct a set of beliefs using prior conjugates and the algorithm in Eq. (3.7) simplifies to updates in the parameters of the exponential family, as shown by the following proposition.

Proposition 32. *Assume the belief density $d\mu_k^i$ at time k has an exponential form with natural parameters χ_k^i and ν_k^i for all $1 \leq i \leq n$, and that these densities are conjugate priors of the likelihood models p_θ^i. Then, the belief density of agent i at time $k+1$, as computed in the update rule in Eq.* (3.7)*, has the same form as the beliefs at time k with the natural*

parameters given by

$$\chi_{k+1}^i = \sum_{j=1}^{n} a_{ij}\chi_k^j + T^i(x_{k+1}^i). \tag{4.11}$$

Proposition 32 simplifies the algorithm in Eq. (3.7) and facilitates its use in traditional estimation problems where members of the exponential family are used.

4.3.1 Additional Examples

In this subsection, we are going to state the general distributed algorithm in Eq. (4.11) presented in Proposition 32 for several distributed parameter estimation problems. Particularly, we explicitly write the definition of the vector $T^i(x_k^i)$ and χ_k^i, from which the parameters of the current beliefs for each agent can be computed. Later in Section 4.3.2 we will provide simulation results for several distributed inference problems over various graph topologies.

Distributed Gaussian filter with unknown mean and known variance

Assume each agent in the network observes a signal of the form $X_k^i = \theta^i + \epsilon_k^i$, where θ^i is finite and unknown scalar quantity, while $\epsilon^i \sim \mathcal{N}(0, 1/\tau^i)$ is a zero mean Gaussian noise with precision $\tau^i = 1/(\sigma^i)^2$ known only by agent i. The objective of the network is to agree on a single θ^* that solves the optimization problem in Eq. (3.5).

In this case, the likelihood models, the prior and the posterior are normal distributions. Thus, if the beliefs of the agents at time k are Gaussian, i.e., $\mu_k^i = \mathcal{N}(\theta_k^i, 1/\tau_k^i)$ for all $i = 1\ldots, n$, then their beliefs at time $k+1$ are also Gaussian. In particular, they are given by $\mu_k^i = \mathcal{N}(\theta_k^i, 1/\tau_k^i)$ for all $i = 1\ldots, n$, with

$$M(\theta) = \begin{bmatrix} \theta \\ \theta^2 \end{bmatrix}, \qquad T^i(x_k^i) = \begin{bmatrix} x_k^i\tau^i \\ -\frac{1}{2}\tau^i \end{bmatrix}, \qquad \chi_k^i = \begin{bmatrix} \theta_k^i\tau_k^i \\ -\frac{1}{2}\tau_k^i \end{bmatrix}.$$

We note that this specific setup is known as Gaussian learning and has been studied in [66, 185], where the expected parameter estimator is shown to converge at an $O(1/k)$ rate.

Distributed Gaussian filter with unknown variance and known mean

In this case, the agents want to cooperatively estimate the value of a variance which is the parameter for Eq. (3.5). Specifically, each agent i observes a realization of the random

variable $X_k^i = \theta^i + \epsilon_k^i$, with $\epsilon_k^i \sim \mathcal{N}(0, 1/\tau^i)$, where θ^i is known and τ^i is unknown. The beliefs of all agents are chosen to be a gGamma distribution $\mu_k^i = \text{gamma}(\alpha_k^i, \beta_k^i)$ and it follows that

$$M(\tau) = \begin{bmatrix} \tau \\ \log \tau \end{bmatrix}, \qquad T^i(x_k^i) = \begin{bmatrix} -\frac{1}{2}(x_k^i - \theta^i)^2 \\ -\frac{1}{2} \end{bmatrix}, \qquad \chi_k^i = \begin{bmatrix} -\beta_k^i \\ -(\alpha_k^i - 1) \end{bmatrix}.$$

Distributed Gaussian filter with unknown mean and variance

In the preceding examples, we have considered the cases when either the mean or the variance is known. Here, we will assume that both the mean and the variance are unknown and need to be estimated. Explicitly, we still have noise observations $X_k^i = \theta^i + \epsilon_k^i$, with $\epsilon_k^i \sim \mathcal{N}(0, 1/\tau^i)$. We are going to assume all agents have beliefs that follow the normal-gamma distribution, i.e. $\mu_k^i = \text{NormalGamma}(\theta_k^i, \lambda_k^i, \alpha_k^i, \beta_k^i)$ for $i = 1, \cdots, n$. Moreover, the it holds that

$$M(\theta, \tau) = \begin{bmatrix} \log \tau \\ \tau \\ \tau\theta \\ \tau\theta^2 \end{bmatrix}, \qquad T^i(x_k^i) = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2}(x_k^i)^2 \\ x_k^i \\ -\frac{1}{2} \end{bmatrix}, \qquad \chi_k^i = \begin{bmatrix} \alpha_k^i - \frac{1}{2} \\ -\frac{1}{2}\lambda_k^i(\theta_k^i)^2 - \beta_k^i \\ \lambda_k^i\theta_k^i \\ -\frac{1}{2}\lambda_k^i \end{bmatrix}.$$

Distributed Bernoulli filter

Here, each of the agents receives private observations of the form $X_k^i \sim \text{Bernoulli}(p^i)$, with p^i unknown. In order to estimate the network-wide parameter, each agent constructs a sequence of beliefs following a beta distribution, i.e. $\mu_k^i = \text{beta}(\alpha_k^i, \beta_k^i)$. Then, the proposed algorithm in Eq. (4.11) updates its parameters. Moreover, it holds that

$$M(p) = \begin{bmatrix} \log p \\ \log(1-p) \end{bmatrix}, \qquad T^i(x_k^i) = \begin{bmatrix} x_k^i \\ 1 - x_k^i \end{bmatrix}, \qquad \chi_k^i = \begin{bmatrix} \alpha_k^i \\ \beta_k^i \end{bmatrix}.$$

Distributed Poisson filter

Similarly as before, we consider an observation model where each agent i receives realization of a Poisson random variable with unknown parameter λ^i, i.e., $X_k^i \sim \text{Poisson}(\lambda^i)$ for all i. The conjugate prior to a Poisson likelihood model is the gamma distribution. Thus, at time k the beliefs of each agent i are given by $\mu_k^i = \text{gGamma}(\alpha_k^i, \beta_k^i)$. Moreover, it holds that

$$M(\lambda) = \begin{bmatrix} \log \lambda \\ \lambda \end{bmatrix}, \qquad T^i(x_k^i) = \begin{bmatrix} x_k^i \\ -1 \end{bmatrix}, \qquad \chi_k^i = \begin{bmatrix} \alpha_k^i - 1 \\ -\beta_k^i \end{bmatrix}.$$

Distributed exponential filter

As a final example, we consider an observation model where each agent i receives realization of an exponential random variable with unknown rate λ^i, i.e., $X_k^i \sim \text{exponential}(\lambda^i)$ for all i. The conjugate prior of an exponential likelihood model is the gamma distribution. Thus, if at time k the beliefs of each agent i are given by $\mu_k^i = \text{gamma}(\alpha_k^i, \beta_k^i)$. Moreover, it holds that

$$M(\lambda) = \begin{bmatrix} \lambda \\ \log \lambda \end{bmatrix}, \qquad T^i(x_k^i) = \begin{bmatrix} -1 \\ x_k^i \end{bmatrix}, \qquad \chi_k^i = \begin{bmatrix} \alpha_k^i - 1 \\ -\beta_k^i \end{bmatrix}.$$

4.3.2 Experimental Results

In this section, we show a number of experimental results for the problem of distributed estimation of network-wide parameters for various network topologies and various observational models. We present the experimental results with the following format.

We explore six different estimation problems:

- Figure 4.4: Distributed estimation of the network-wide mean parameter with Gaussian observations with the local knowledge of private variances.
- Figure 4.5: Distributed estimation of network-wide variance parameter with Gaussian observations with the local knowledge of private means.
- Figure 4.6: Distributed estimation of network-wide mean and variance parameters with Gaussian observations without knowledge of local means or variances.
- Figure 4.7: Distributed estimation of the network-wide parameter with heterogeneous Bernoulli observations.
- Figure 4.8: Distributed estimation of the network-wide parameter with heterogeneous Poisson observations.

- Figure 4.9: Distributed estimation of the network-wide parameter with heterogeneous Exponential observations.

For each of the figures described above, we measure the performance of the proposed algorithm using its normalized distance to optimality and the distance to consensus, defined as follows:

$$\text{Distance to Optimality: } \frac{|F(\theta_k) - F(\theta^*)|}{|F(\theta_0) - F(\theta^*)|},$$
$$\text{Distance to Consensus: } \|L\theta_k\|_2^2,$$

where $\theta_k = (\theta_k^1, \theta_k^2, \cdots, \theta_k^n)$ is the aggregation of all the current parameter estimations for each of the agents, and the function $F(\theta_k)$ is defined as

$$F(\theta_k) = \sum_{i=1}^{n} D_{KL}(P^i \| P^i_{\theta_k^i}),$$

and L is the graph Laplacian of the communication graph. We have used the graph Laplacian as a measure of distance to consensus since by definition the set where $\theta_k^1 = \theta_k^2 = \cdots = \theta_k^n$, i.e. consensus, is null space of the matrix L.

Finally, we present the results for five classes of networks, namely: complete graphs, cycle graphs, path graphs, star graphs, and Erdős-Rényi random graphs. For each of the network classes, we show the performance for 10 agents, 100 agents, and 1000 agents.

4.4 Distributed Gaussian Learning on Time-Varying Directed Graphs

In this subsection, we assume that the observations have *Gaussian distribution* and that the likelihood models are Gaussian, both with bounded second-order moments, i.e. $X_k^i \sim \mathcal{N}(\theta^i, (\sigma^i)^2)$ and $p_\theta^i(\cdot|\sigma^i) = \mathcal{N}(\theta, (\sigma^i)^2)$ where $\sigma^i > 0$ for every i. This setting corresponds to the case of having measurements of the true parameter θ^* corrupted by some Gaussian noise and the agents being informed that the noise is Gaussian with a known variance.

The Kullback-Leibler distance between two univariate Gaussian distributions P and Q, where $P = \mathcal{N}(\theta^1, (\sigma^1)^2)$ and $Q = \mathcal{N}(\theta^2, (\sigma^2)^2)$, is given by

$$D_{KL}\left(P\|Q\right) = \log\frac{\sigma^2}{\sigma^1} + \frac{(\sigma^1)^2 + (\theta^1 - \theta^2)^2}{(\sigma^2)^2} - \frac{1}{2}.$$

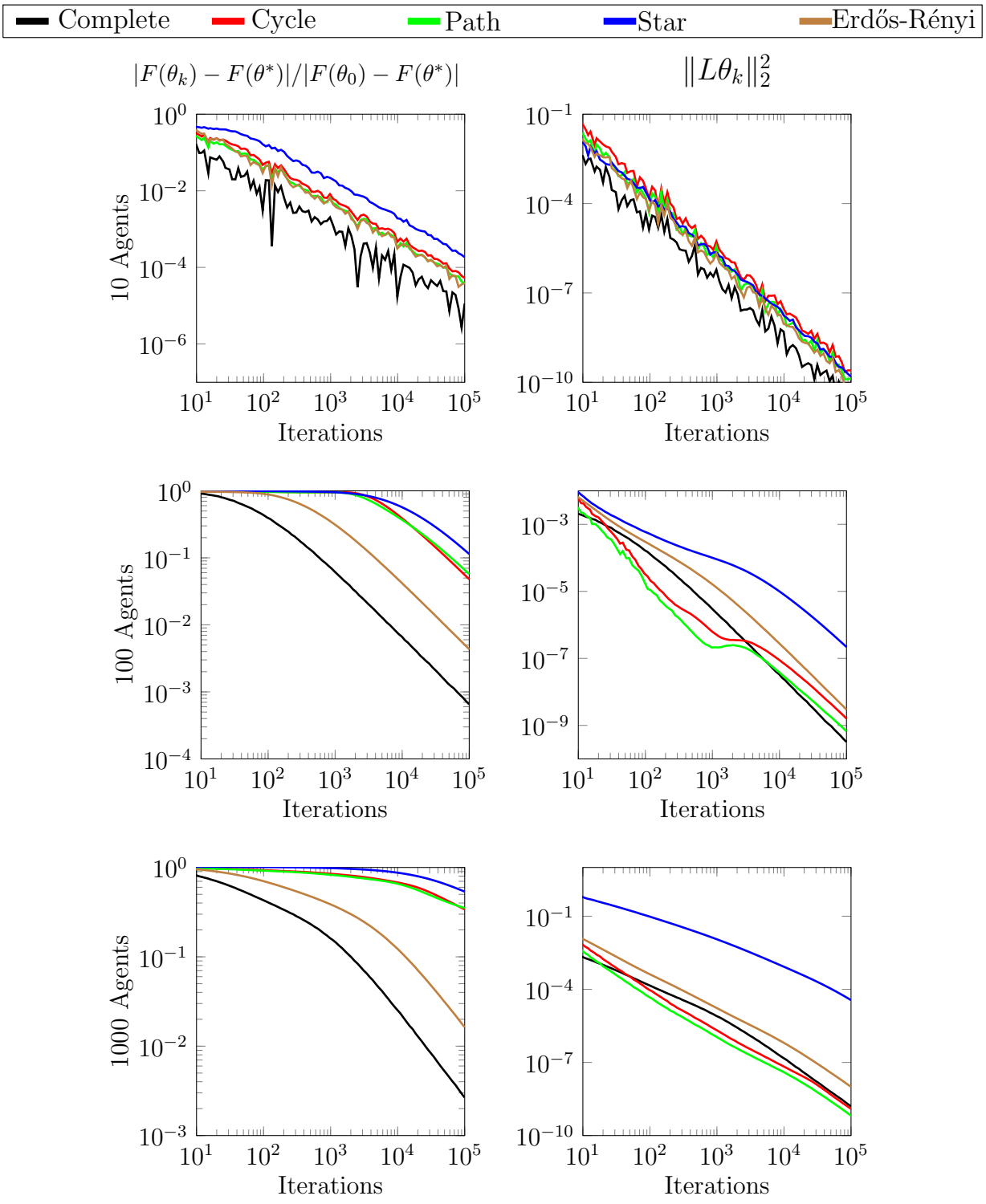


Figure 4.4: Distributed estimation of a network-wide mean from Gaussian observations. Optimality and distance to consensus for the distributed estimation of a network-wide unknown **mean** parameter, from Gaussian observations, for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

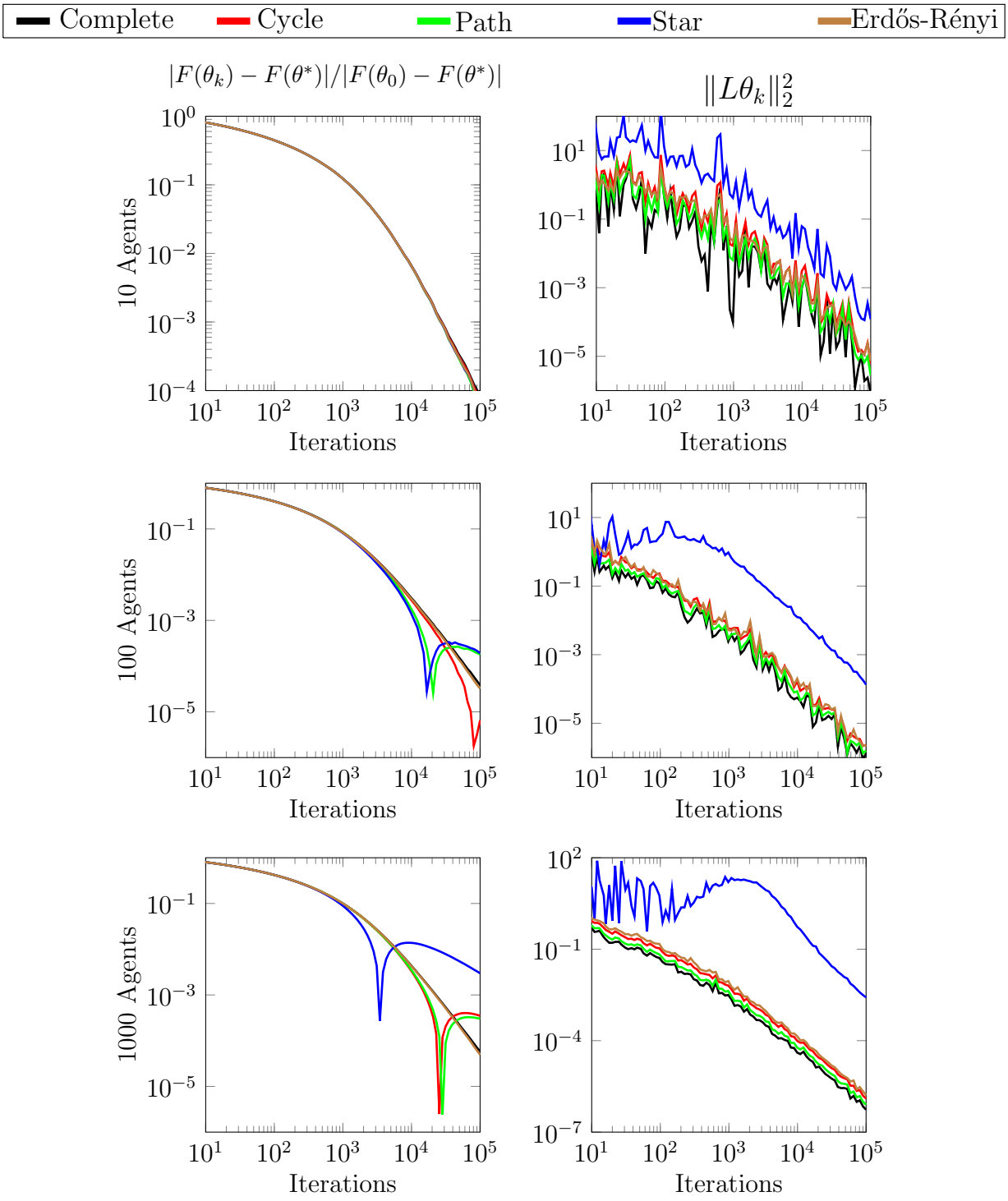


Figure 4.5: Distributed estimation of a network-wide variance from Gaussian observations. Optimality and distance to consensus for the distributed estimation of a network-wide **variance** parameter, from Gaussian observations, for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

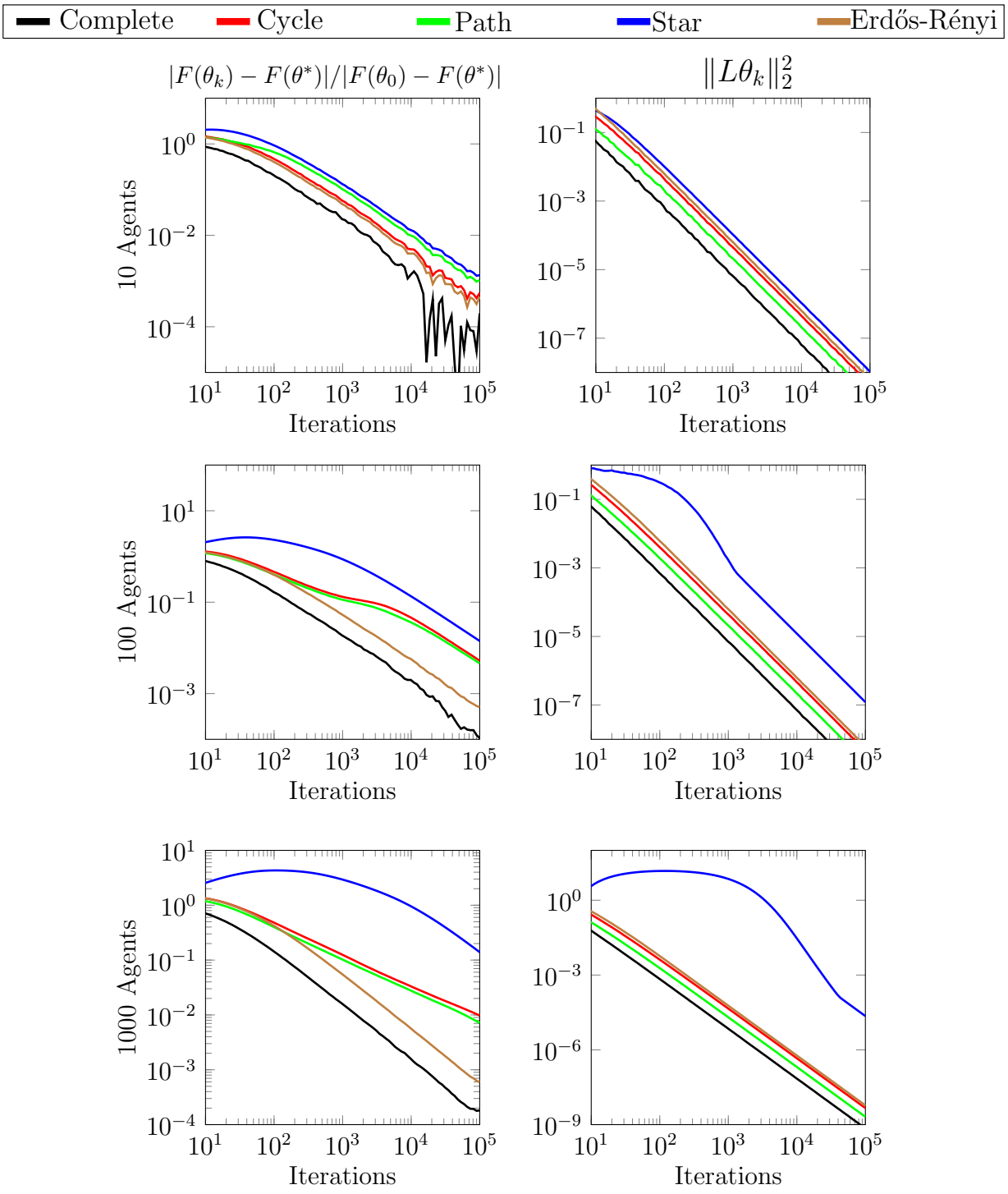


Figure 4.6: Distributed estimation of a network-wide mean and variance from Gaussian observations. Optimality and distance to consensus for the distributed estimation of network-wide **mean and variance** parameters, from Gaussian observations, for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

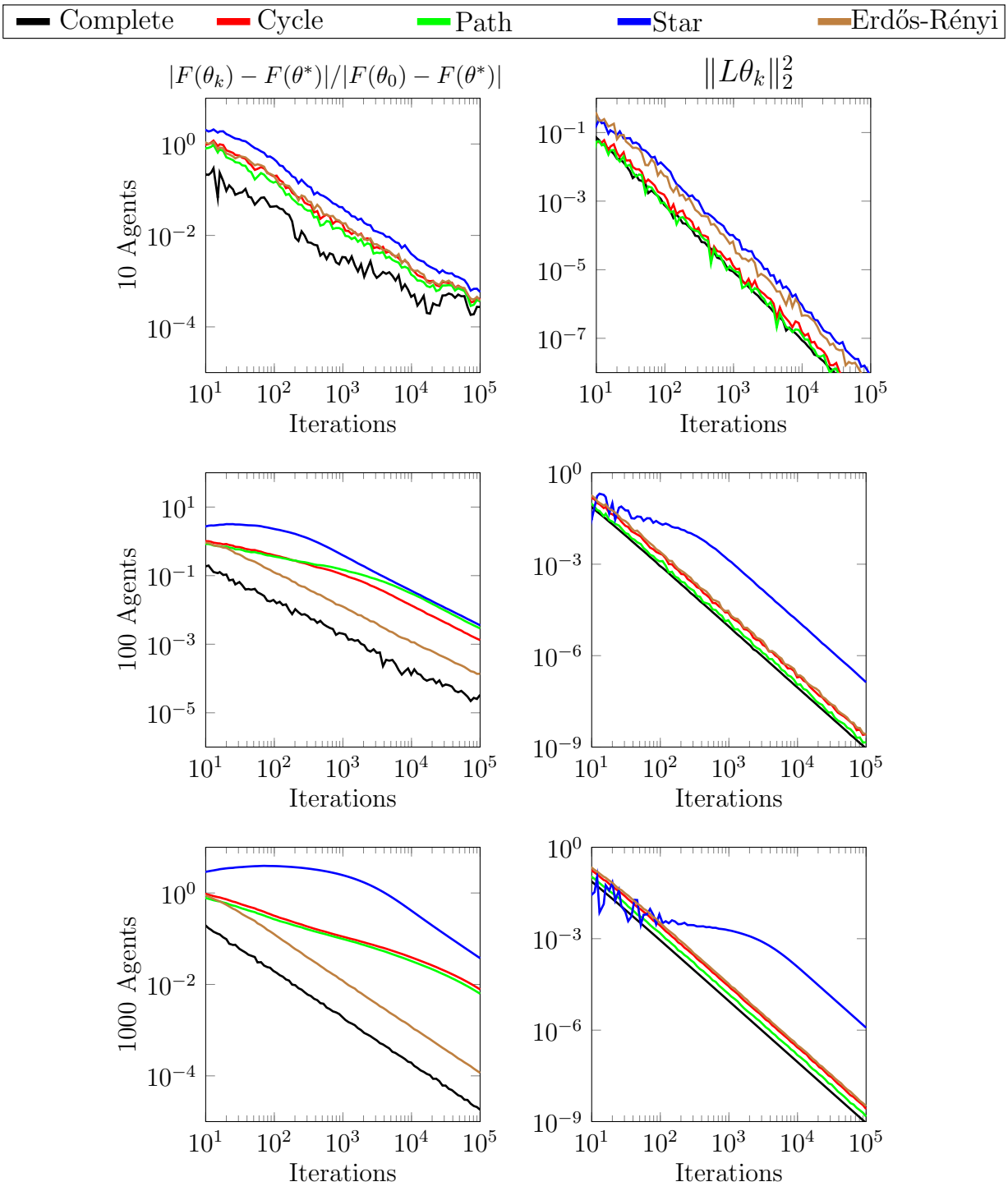


Figure 4.7: Distributed estimation of a network-wide parameter from Bernoulli observations. Optimality and distance to consensus for the distributed estimation of a network-wide parameter of Bernoulli observations for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

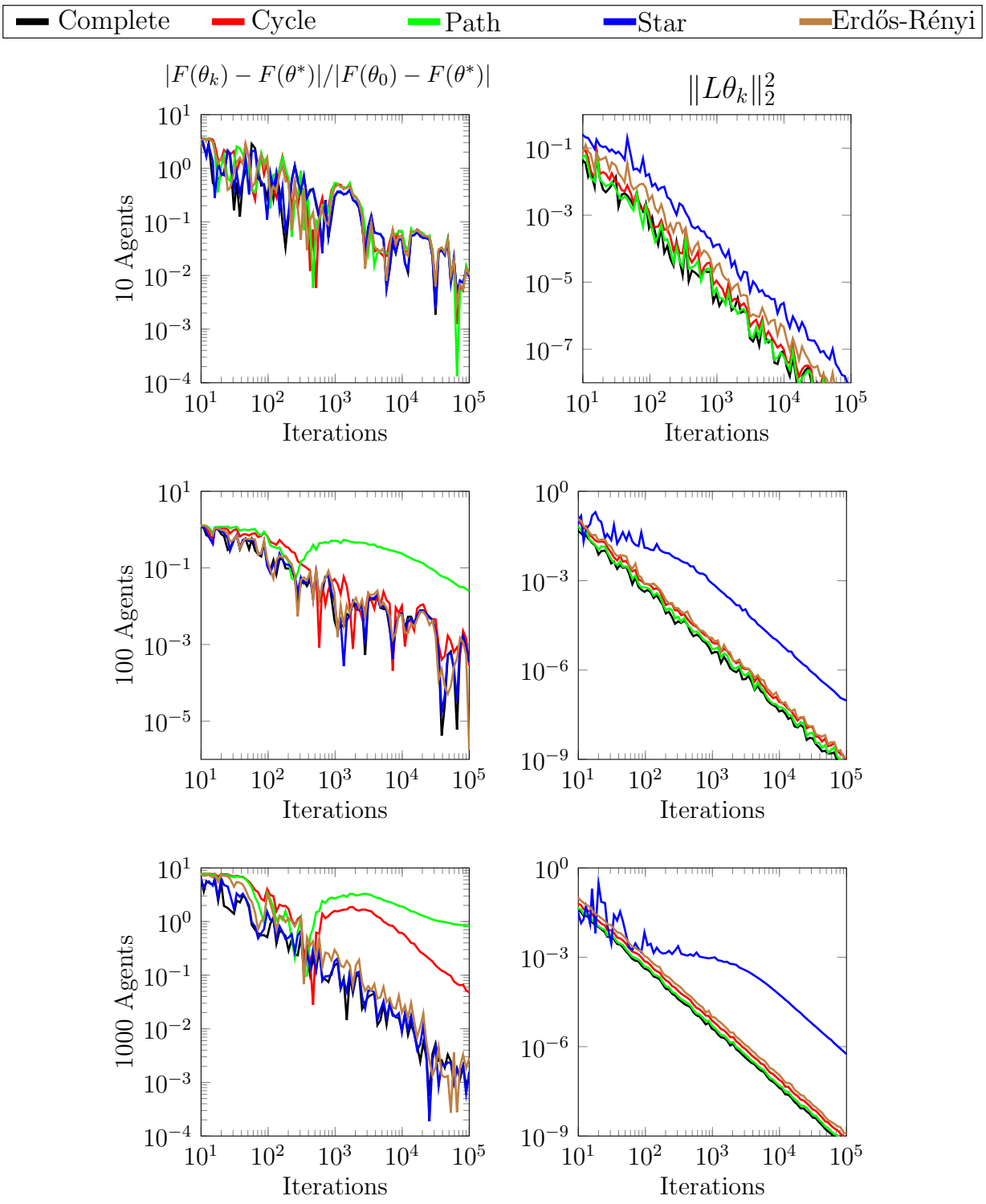


Figure 4.8: Distributed estimation of a network-wide parameter from Poisson observations. Optimality and distance to consensus for the distributed estimation of a network-wide parameter of Poisson observations for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

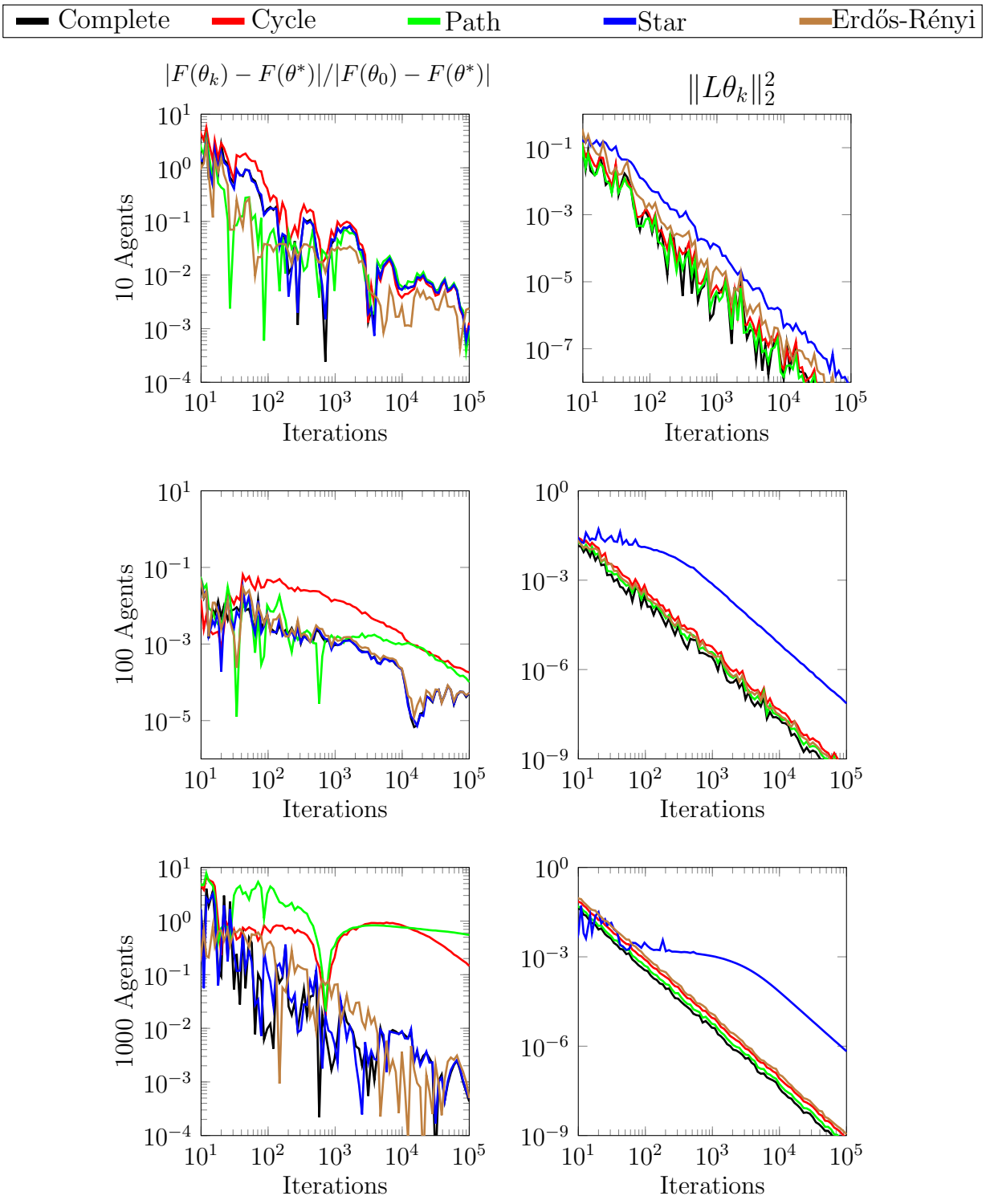


Figure 4.9: Distributed estimation of a network-wide parameter from Exponential observations. Optimality and distance to consensus for the distributed estimation of a network-wide parameter of Exponential observations for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

Thus, in this case, the problem in Eq. (3.5) is equivalent to

$$\min_{\theta\in\Theta} F(\theta) \triangleq \sum_{i=1}^{n} \frac{(\theta-\theta^i)^2}{2(\sigma^i)^2}, \tag{4.12}$$

which is convex with a unique solution

$$\theta^* = \sum_{i=1}^{n} \frac{\theta^i/(\sigma^i)^2}{\sum_{j=1}^{n} 1/(\sigma^j)^2}. \tag{4.13}$$

However, the exact value of θ^i is unknown and each agent i has access only to noisy observations of the form $X_k^i = \theta^i + \epsilon^i$, where $\epsilon^i \sim \mathcal{N}(0, (\sigma^i)^2)$. Moreover, variances are only known locally, i.e. agent i only knows σ^i.

We propose the following distributed algorithm for solving the problem in Eq. (4.12) over time-varying directed graphs:

$$\tau_{k+1}^i = \sum_{j=1}^{n} [A_k]_{ij}\, \tau_k^j + \tau^i, \tag{4.14a}$$

$$\theta_{k+1}^i = \frac{\sum_{j=1}^{n} [A_k]_{ij}\, \tau_k^j \theta_k^j + x_{k+1}^i \tau^i}{\tau_{k+1}^i}, \tag{4.14b}$$

where $\tau^i = 1/(\sigma^i)^2$ is referred to as the precision of the observations. The weights $[A_k]_{ij}$ are chosen as

$$[A_k]_{ij} = \begin{cases} \frac{1}{d_k^j+1} & \text{if } (j,i) \in E_k, \\ 0 & \text{otherwise,} \end{cases} \tag{4.15}$$

where d_k^j is the out-degree of node j at time k. Without loss of generality, we assume that $\tau_0^i = \tau^i$ for all i.

Remark 2. *It is* not *necessary for each agent to have some form of informative observations. Indeed, there might be agents with no observations working as buffers for information for which we also expect correct estimates of θ^*. These "blind" agents depend on communicating with other agents to construct their estimates.*

Remark 3. *While our focus is on the univariate Gaussian case, extensions to the multivariate are similarly possible using the results of conjugate priors for multivariate Gaussian distributions.*

The next proposition shows that the algorithm in Eq. (4.14) is a specific realization of Eq. (3.7) for the case of Gaussian distributions in the priors and likelihood models.

Proposition 33. *Let the prior belief density $\bar{\mu}_0^i$ of every agent be a Gaussian function, i.e.*

$$d\mu_0^i(\theta; \theta_0^i, \sigma^i) = \mathcal{N}(\theta_0^i, (\sigma^i)^2),$$

and let the parametric family of distributions for the likelihood models be Gaussian functions, i.e.

$$p_\theta^i(\cdot|\sigma^i) = \mathcal{N}(\theta, (\sigma^i)^2).$$

Then, for any $k \geq 1$, the posterior belief density $d\mu_k^i$, given by Eq. (3.7), is also a Gaussian function. Moreover, if the weights a_{ij} are chosen to be $1/(d_k^j + 1)$, then the mean and the standard deviation of the posterior follow Eq. (4.14).

Now, we proceed to state our two main results showing the convergence properties of the algorithm in Eq. (4.14).

Lemma 34. *The expected mean process $\{\mathbb{E}[\theta_k^i]\}$ converges to θ^* for all i with a convergence rate of $O(1/k)$. Moreover, the constant terms depend on the topology of the network, the precision of the observations and the initial guess.*

Proof. In fact, we will prove the bound

$$\left|\mathbb{E}[\theta_{k+1}^i] - \theta^*\right| \leq \frac{\tau_{max}}{\tau_{min} k \delta} \left(\|\theta_0 - \theta^* \mathbf{1}\|_1 + \frac{2C\|\theta - \theta^* \mathbf{1}\|_1}{1 - \lambda} \right), \tag{4.16}$$

with $\tau_{max} = \max_j \tau^j$, and τ_{min} is the smallest non-zero precision among all agents.

First, define a new variable as $x_k^i = \tau_k^i \theta_k^i$, then from Eq. (4.14b) it follows that

$$\begin{aligned} x_{k+1} &= A_k x_k + \text{diag}(\tau) s_{k+1} \\ &= A_{k:0} x_0 + \sum_{t=1}^{k} A_{k:t} \text{diag}(\tau) s_t + \text{diag}(\tau) s_{k+1}, \end{aligned}$$

where $\text{diag}(\tau)$ is a diagonal matrix with $[\text{diag}(\tau)]_{ii} = \tau^i$ and $x_k = [x_k^1, \ldots, x_k^n]'$, $\tau = [\tau^1, \ldots, \tau^n]'$, $s_k = [s_k^1, \ldots, s_k^n]'$.

Adding and subtracting $\sum_{t=1}^{k} \phi_k \tau' s_t$ from the preceding relation we obtain

$$x_{k+1} = A_{k:0} x_0 + \sum_{t=1}^{k} D_{k:t} \text{diag}(\tau) s_t + \text{diag}(\tau) s_{k+1} + \sum_{t=1}^{k} \phi_k \tau' s_t,$$

with $D_{k:t} = A_{k:t} - \phi_k \mathbf{1}'$, and ϕ_k is as in Lemma 2.

Following a similar procedure, from Eq. (4.14a) it holds that

$$\tau_{k+1} = A_{k:0}\tau_0 + \sum_{t=1}^{k} D_{k:t}\tau + k\phi_k \mathbf{1}'\tau + \tau.$$

Going back to the original variable θ_k, we have that

$$\mathbb{E}[\theta^i_{k+1}] = \frac{[A_{k:0}\mathrm{diag}(\tau)\theta_0]_i + \sum_{t=1}^{k}[D_{k:t}\mathrm{diag}(\tau)\theta]_i + \tau^i\theta^i + k\phi^i_k\tau'\theta}{[A_{k:0}\tau_0]_i + \sum_{t=1}^{k}[D_{k:t}\tau]_i + k\phi^i_k\mathbf{1}'\tau + \tau^i}.$$

By subtracting θ^* on both sides of the previous relation and taking the absolute value, we obtain

$$\begin{aligned}\left|\mathbb{E}[\theta^i_{k+1}] - \theta^*\right| \leq & \left|\frac{[A_{k:0}\mathrm{diag}(\tau_0)\,(\theta_0 - \theta^*\mathbf{1})]_i}{\sum_{t=1}^{k}[D_{k:t}\tau]_i + k\phi^i_k\mathbf{1}'\tau}\right| + \\ & \left|\frac{\tau^i\,(\theta^i - \theta^*)}{\sum_{t=1}^{k}[D_{k:t}\tau]_i + k\phi^i_k\mathbf{1}'\tau}\right| + \left|\frac{\sum_{t=1}^{k}\,[D_{k:t}\mathrm{diag}(\tau)\,(\theta - \theta^*\mathbf{1})]_i}{\sum_{t=1}^{k}[D_{k:t}\tau]_i + k\phi^i_k\mathbf{1}'\tau}\right|,\end{aligned}$$

where the terms involving $k\phi^i_k\tau'\theta$ cancel out and the following positive terms are removed from the denominator $[A_{k:0}\tau_0]_i + \tau^i > 0$.

Then by the fact that $[D_{k:t}\mathbf{1}]_i + \phi^i_k n > \delta$ on the denominator and using Lemma 2 on the third term it follows that

$$\left|\mathbb{E}[\theta^i_{k+1}] - \theta^*\right| \leq \left|\frac{[A_{k:0}\mathrm{diag}(\tau_0)\,(\theta_0 - \theta^*\mathbf{1})]_i}{k\delta\tau_{min}}\right| + \frac{\tau^i|\theta^i - \theta^*|}{k\delta\tau_{min}} + \frac{C\tau_{max}\|\theta - \theta^*\mathbf{1}\|_1}{k\delta\tau_{min}(1-\lambda)}.$$

Finally, the desired result follows by Hölders inequality in the first term with $\|[A_{k:0}\mathrm{diag}(\tau)]_i\|_\infty = \tau_{max}$ and grouping the second and third terms since $\frac{C}{1-\lambda} > 1$.

$$\left|\mathbb{E}[\theta^i_{k+1}] - \theta^*\right| \leq \frac{\max_j[A_{k:0}]_{ij}\tau^j\|\theta_0 - \theta^*\mathbf{1}\|_1}{k\delta\tau_{min}} + \frac{2C\tau_{max}\|\theta - \theta^*\mathbf{1}\|_1}{k\delta\tau_{min}(1-\lambda)}.$$

□

The first term in Eq. (4.16) shows the dependency on the initial estimates θ_0 while the second term depends on the heterogeneity of mean of local observations. The network topology and the number of agents are characterized by λ and δ.

We are now ready to state our main result about the almost sure convergence of the proposed algorithm.

Theorem 35. *Let the graph sequence of interactions $\{\mathcal{G}_k\}_{k=1}^{\infty}$ be B-strongly connected. Moreover, assume $X_k^i \sim \mathcal{N}(\theta^i, (\sigma^i)^2)$ and $p_\theta^i(\cdot|\sigma^i) = \mathcal{N}(\theta, (\sigma^i)^2)$ for all i. Then, the sequence $\{\theta_k^i\}$ generated by Eq. (4.14) converges almost surely to θ^*, i.e.*

$$\lim_{k\to\infty} \theta_k^i = \theta^* \qquad a.s. \quad \forall i$$

Remark 4. *The specific selection of weights as $1/(d_k^j + 1)$ is a design choice. Theorem 35 still holds for any sequence of column stochastic matrices $\{A_k\}$ with every non-zero entry bounded from below away from zero, and with positive diagonal entries.*

A specific version of the proposed problem is the case when all agents observe independent realizations of the same random variable, i.e. $X_k^i \sim \mathcal{N}(\theta^*, (\sigma^2)^*)$. Recently, authors in [185, 186] have explored this case. Specifically, in [186] the authors are concerned with the effects of the network topology on the convergence rate of the distributed mean estimation problem. They show mean square consistency of the following algorithm

$$\theta_{k+1}^i = \frac{k}{k+1} \sum_{j=1}^{n} a_{ij}\theta_k^j + \frac{1}{k+1} x_{k+1}^i, \tag{4.17}$$

and provide explicit rates for different network topologies. Note that the algorithm in Eq. (4.17) reduces to Eq. (4.14) when $\tau^i = 1$ in such a way that $\tau_k^i = k$ for all i, and the graph is static with a doubly stochastic weight matrix.

In [185], the authors proposed a new distributed Gaussian learning algorithm where communication between agents is noisy. Following the non-Bayesian learning without recall approach proposed in [47] they develop the specific realization for Gaussian random variables. Additionally, they consider the sequence of observations $\{s_k^i\}$ as coming from an agent, denoted as $n+1$, and thus a different weighting strategy is proposed. Their algorithm is

$$\tau_{k+1}^i = \tau_k^i + d_k^j \tau, \tag{4.18a}$$

$$\theta_{k+1}^i = \frac{\sum_{j=1}^{n+1} \tau_k^j a_k^j}{\tau_{k+1}^i}, \tag{4.18b}$$

with the specific condition that $\tau_k^j = \tau$ for all $j \neq i$, $a_k^j = \theta_k^i$ for $j = i$ and $a_k^j = \theta_k^j + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \tau)$, with $a_k^{n+1} = x_k^i$. The authors showed almost sure convergence of the algorithm. Moreover, a convergence rate of $O(k^{-\frac{\gamma}{2d}})$ was derived, where γ is a bound on the uniform connectivity to the truth observations and d is the maximal degree over all the networks.

One particular characteristic of the algorithm proposed in [185] is that, apart from traditional literature on distributed learning, the authors do not assume agents communicate over a *sufficiently* connected network (B-strong connectivity in Theorem 35). They replace this assumption by a so-called *truth-hearing assumption* which works as a $1/\gamma$-strong connectivity with the $n+1$ node that provides direct noisy observations of θ^*. Thus, it is required that every node receives signals from node $n+1$ at least once in every time interval of length $1/\gamma$. If all agents receive independent observations from identical distributions, connectivity of the network and truth hearing assumptions both serve the same purpose of guarantying the diffusion of the information over the network; otherwise, some form of connectivity between agents is needed.

In addition to different connectivity assumptions, one main characteristic of the algorithm in Eq. (4.18) is that agents do not differentiate the signal X_k^i coming from the observations of the parameter, and the signals $\{a_k^j\}$ coming from other agents. Every agent treats both signals similarly. The weights for observations of X_k^i and neighbor signals $\{\theta_k^i\}_{i=1}^n$ decay. In our approach in Eq. (4.14), the weight for X_k^i decays to zero and the weight for the convex combination of $\{\theta_k^i\}_{i=1}^n$ goes to one. This indeed shows that we do require the identification of signals coming from either agents or the noisy parameter observations. This extra information could explain why our approach has better performance in terms of convergence rates.

Next, we provide simulation results for our proposed algorithm, and we compare its performance with results in [185, 186]. Initially, we will consider the same scenario as in [185, 186] with static undirected graphs with all agents having identical distributions in their *noiseless beliefs sharing*. We will evaluate the performance of the algorithms for two different graph topologies, namely path/line graph and a lattice/grid graph.

Figure 4.10 shows the absolute error of the estimated value θ^* for the lattice/grid graph with 25 agents. It is assumed that $X_k^i \sim \mathcal{N}(4,1)$. An average of 500 Monte Carlo simulations is shown for one arbitrary agent. Also, the theoretical convergence rates are shown for comparison purposes. No simulation of the algorithm in Eq. (4.17) is shown since it reduces to the same algorithm as in Eq. (4.14) for the simulated scenario.

Figure 4.11 shows the simulation results for the same scenario as in Fig. 4.10 but now for a path/line graph of 15 agents. As predicted by the theoretical convergence rate bounds, the proposed algorithm in Eq. (4.14) decays as $O(1/k)$ where the topology of the network affects only the constant, whereas the proposal in Eq. (4.18) depends explicitly on the maximum degree among all graphs as $O(1/k^{1/d})$.

Next, we will show that for the case of each agent having noise with different standard deviations, by using information about the current estimate precision (i.e., τ_k^i), a better

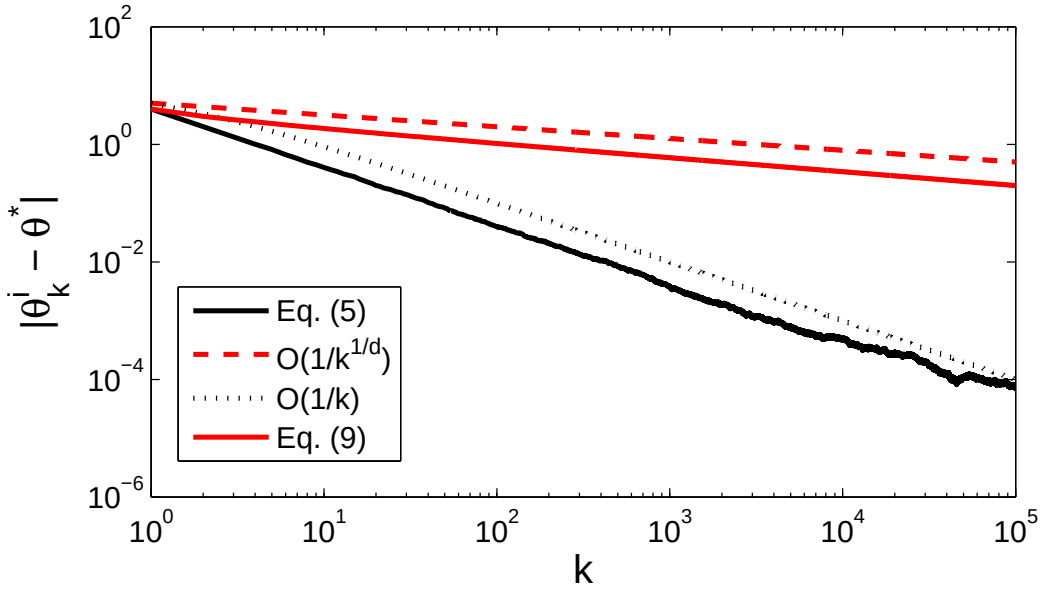


Figure 4.10: Distributed Gaussian learning on a grid graph. Simulation results of algorithms in Eq. (4.14) and Eq. (4.18) for a lattice/grid graph of 25 nodes for an average behavior over 500 Monte Carlo simulations.

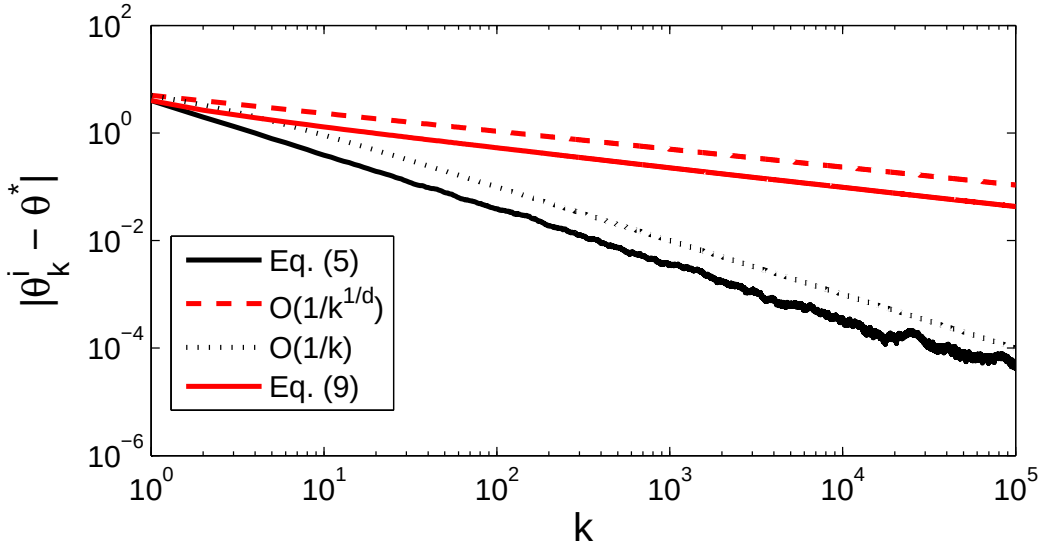


Figure 4.11: Distributed Gaussian learning on a path graph. Simulation results of algorithms in Eq. (4.14) and Eq. (4.18) for a path graph of 25 nodes. Average behavior over 500 Monte Carlo simulations.

performance is achieved. Fig. 4.12 shows the absolute error on the estimation of θ^* for the algorithm in Eq. (4.14) that uses precision information and the proposal in Eq. (4.17) that assumes uniform precision. In this simulation, agents have heterogeneous precisions such

that $X_k^i \sim \mathcal{N}(4, i)$. That is, in the path graph, the first agent has $\tau^1 = 1$; the last agent, on the other hand, has $\tau^n = n$. This implies that agent 1 has the highest variance in its observations. We have chosen to show the results for agent 1 only.

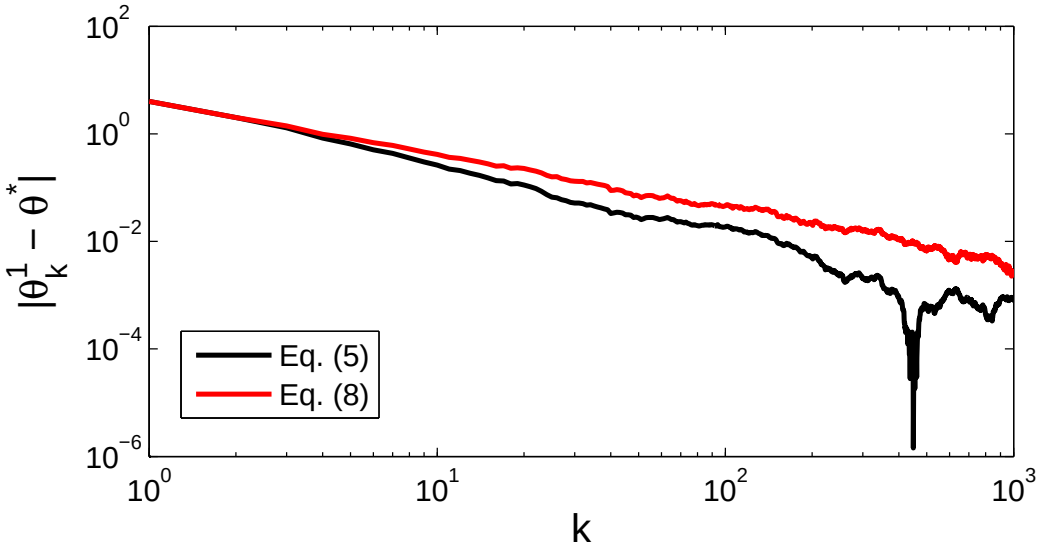


Figure 4.12: Distributed Gaussian learning on a path graph and heterogeneous variances. Simulation results of algorithms in Eq. (4.14) and Eq. (4.17) for a path graph of 25 nodes with heterogeneous precisions (i.e. $\tau's$). Average behavior over 500 Monte Carlo simulations.

Finally, we will present the simulation results for a directed static graph which has been shown to be a pathological case for the push-sum algorithm, see Fig. 4.13. Each agent receives signals of the form $X_k^i \sim \mathcal{N}(i, n-i+1)$. Thus every agent has different measurement precisions and different θ^i. The optimal θ^* is defined in Eq. (4.13).

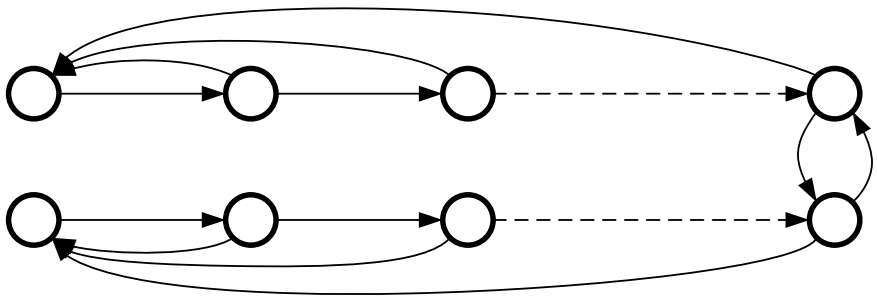


Figure 4.13: A particularly bad graph. Directed graph for simulation of the algorithm in Eq. (4.14).

Figure 4.14 shows the simulation results for the algorithm in Eq. (4.14) to the specific set of observations $S_k^i \sim \mathcal{N}(i, n-i+1)$ on the graph in Fig. 4.13. The average over 10 Monte

Carlo simulations is shown. The predicted $O(1/k)$ behavior is observed, after a transition time that depends on the number of agents in the network (i.e. the effects on n and λ in Lemma 34).

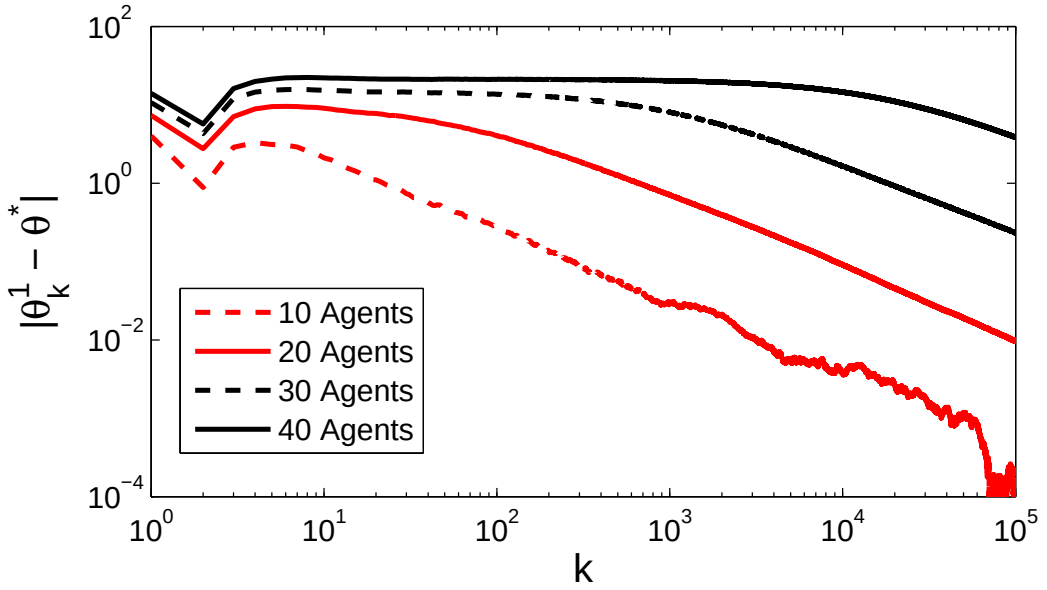


Figure 4.14: Distributed Gaussian learning on a particularly bad graph. Simulations results of algorithms in Eq. (4.14) for the graph depicted in Fig. 4.13. Four different results are shown, for 10, 20, 30 and 40 agents respectively.

4.5 Conclusions

We proposed two distributed cooperative learning algorithms for the problem of collaborative inference. We have proposed an algorithm for distributed learning with both countable and compact sets of hypotheses. Our algorithm may be viewed as a distributed version of stochastic mirror descent applied to the problem of minimizing the sum of Kullback-Leibler divergences. Our results show non-asymptotic geometric convergence rates for the belief concentration around the true hypothesis.

We developed an algorithm for distributed parameter estimation with Gaussian noise over time-varying directed graphs. The proposed algorithm is shown to be a specific case of a more general class of distributed (non-Bayesian) learning methods. Almost sure converge as well as an explicit convergence rate is shown in terms of the network topology and the number of agents. Comparisons with recently proposed approaches are presented.

CHAPTER 5

A DUAL APPROACH FOR OPTIMAL ALGORITHMS IN DISTRIBUTED OPTIMIZATION

In this chapter, we go back to our original distributed optimization problem

$$\min_{z\in\mathbb{R}^m} \sum_{i=1}^{n} f_i(z), \tag{5.1}$$

where the each f_i is convex and known by an agent i only, that represents a node in an arbitrary communication network. The problem in Eq. (5.1) is to be solved in a distributed manner by repeated interactions of a set of agents over a static network. We follow the approach in [79] by formulating a dual problem and exploit recent results in the study of convex optimization problems with affine constraints [187, 188, 189] to develop algorithms with provably optimal convergence rates for the cases where each of the objective functions f_i has one the following properties:

1. it is strongly convex and with Lipschitz continuous gradients;
2. it is strongly convex and Lipschitz continuous on a bounded set (but not necessarily smooth);
3. it is convex with Lipschitz continuous gradients;
4. it is convex and Lipschitz continuous (not necessarily smooth).

Our results match known optimal complexity bounds for centralized convex optimization (obtained by classical methods such as Nesterov's fast gradient method [190]), with an additional cost induced by the network of communication constraints. This extra cost appears in the form of a multiplicative term proportional to the square root of the spectral gap of the interaction matrix. In summary, our main results provide an algorithm that achieves ε relative accuracy on *any* fixed, connected and undirected graph according to Table 5.1, where universal constants, logarithmic terms, and dependencies on the initial conditions are hidden for simplicity. The resulting iteration complexities are given both for the optimality of the solution and the violation of the consensus constraints. Note that for distributed

Table 5.1: Iteration Complexity of Distributed Optimization Algorithms. All estimates are presented up to logarithmic factors, i.e. of the order $\tilde{O}$.

Approach	**Reference**	**μ-strongly convex and L-smooth**	**μ-strongly convex and M-Lipschitz**	**L-smooth**	**M-Lipschitz**
Centralized	[191]	$\sqrt{L/\mu}$	$M^2/(\mu\varepsilon)$	$\sqrt{L/\varepsilon}$	M^2/ε^2
Gradient Computations	[192][b]	$(L/\mu)^{5/7}n^3$	–	$1/\varepsilon^{5/7}$	–
	[73][a]	$n^2+n\sqrt{L/\mu}$	–	$1/\varepsilon$	$1/\varepsilon$
	[96]	–	–	–	nM^2/ε^2
	[193]	–	–	–	n^2M^2/ε^2
	[194]	$(L/\mu)n^2$	–	–	–
	[195]	–	–	$(L/\varepsilon)n^3$	–
	[196]	$(L/\mu)m^4$	–	$(L/\varepsilon)n^4$	–
	[75][c]	$\sqrt{L/\mu}n^2$	–	–	–
Communication Rounds	[79]	$\sqrt{L/\mu}n$	–	–	–
	[197]	–	$\sqrt{M^2/(\mu\varepsilon)}n$	–	nM/ε
	This work	$\sqrt{L/\mu}n$	$\sqrt{M^2/(\mu\varepsilon)}n$	$\sqrt{L/\varepsilon}n$	nM/ε

[a] Additionally, it is assumed functions are proximal friendly. No explicit dependence on L, M or n is provided.
[b] An iteration complexity of $\tilde{O}(\sqrt{1/\varepsilon})$ is shown if the objective is the composition of a linear map and a strongly convex and smooth function. Moreover, no explicit dependence on L and n is provided.
[c] A linear dependence on n is achieved if L is sufficiently close to μ.

algorithms based on primal iterations these estimates translate to computations of gradients of the local functions for each of the agents. On the other side, in dual based algorithms, the complexity refers to computations of the gradients of the Lagrangian dual function, which translates to the number of communication rounds in the network.

Additionally, we build upon the designed optimal algorithms for distributed optimization to propose a new class-optimal algorithm for the distributed computation of Wasserstein barycenters over networks. Assuming that each node in a graph has a probability distribution, we prove that every node can reach the barycenter of all distributions held in the network by using local interactions compliant with the topology of the graph. We show the minimum number of communication rounds required for the proposed method to achieve arbitrary relative precision both in the optimality of the solution and the consensus among all agents for undirected fixed networks.

5.1 Problem Formulation

Initially, let us introduce a stacked column vector $x = [x_1^T, x_2^T, \ldots, x_n^T]^T \in \mathbb{R}^{nm}$ to rewrite problem (5.1) in an equivalent form as follows:

$$\min_{x_1=\ldots=x_n} F(x) \qquad \text{where} \qquad F(x) \triangleq \sum_{i=1}^{n} f_i(x_i). \tag{5.2}$$

Suppose that we want to solve this problem in a distributed manner over a network. We model such a network as a fixed connected undirected graph $\mathcal{G} = (V, E)$. We assume that the graph $\mathcal{G}$ does not have self-loops. The network structure imposes information constraints; specifically, each node i has access to the function f_i only and a node can exchange information only with its immediate neighbors, i.e., a node i can communicate with node j if and only if $(i, j) \in E$.

We can represent the communication constraints imposed by the network by introducing a set of equivalent to the constraints in Eq. (5.2). To do so, we define the communication matrix (also referred to as an interaction matrix) by $W \triangleq \bar{W} \otimes I_m$, where $\otimes$ indicates the Kronecker product and $\bar{W}$ is the Laplacian matrix of the graph $\mathcal{G}$.

Throughout this chapter, *we assume that the undirected graph $\mathcal{G} = (V, E)$ is connected*. Under this assumption, the Laplacian matrix $\bar{W}$ is symmetric and positive semi-definite. Furthermore, the vector $\mathbf{1}$ is the unique (up to a scaling factor) eigenvector associated with the eigenvalue $\lambda = 0$. Given the definition $W = \bar{W} \otimes I_m$, one can verify that W inherits all the properties of $\bar{W}$, i.e., it is a symmetric positive semi-definite matrix and it satisfies the following relations:

- $Wx = 0$ if and only if $x_1 = \ldots = x_n$.
- $\sqrt{W}x = 0$ if and only if $x_1 = \ldots = x_n$.
- $\sigma_{\max}(\sqrt{W}) = \lambda_{\max}(W)$.

Therefore, one can equivalently rewrite the problem in Eq. (5.2) as follows:

$$\min_{\sqrt{W}x=0} F(x) \qquad \text{where} \qquad F(x) \triangleq \sum_{i=1}^{n} f_i(x_i). \tag{5.3}$$

Note that the constraint set $\{x \mid \sqrt{W}x = 0\}$ is the same as the set $\{x \mid x_1 = \ldots = x_n\}$, since $\ker(\sqrt{W}) = \text{span}(\mathbf{1})$ due to the connectivity of the graph $\mathcal{G}$.

Additionally, it follows that if each function $f_i(x_i)$ in Eq. (5.3) is μ_i-strongly convex in x_i, then $F(x)$ is μ-strongly convex in x, with $\mu = \min_{1 \le i \le n} \mu_i$. Also, if each $f_i(x_i)$ in Eq. (5.3)

is L_i-smooth, then $F(x)$ is L-smooth with $L = \max_{1\leq i\leq n} L_i$.

Our main algorithmic tool in this chapter will be Nesterov's fast gradient method (FGM) [198]. Equations in (5.4) state a version of the FGM method for a general μ-strongly convex and L-smooth function $f(x)$. Other variants of this method can be found in [198, 199, 200].

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k), \tag{5.4a}$$

$$y_{k+1} = x_{k+1} + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(x_{k+1} - x_k). \tag{5.4b}$$

Specifically, it holds for the iterates of Eq. (5.4) that

$$f(x_k) - f^* \leq L\left(1 - \sqrt{\mu/L}\right)^k \|x_0 - x^*\|_2^2, \tag{5.5}$$

where f^* denotes the minimum value of the function $f(x)$ over $\mathbb{R}^n$ and x^* is its minimizer.

In what follows, we will consider a generic optimization problem with linear constraints. Then, we will apply FGM and obtain some basic insights. Moreover, we will derive the results for a corresponding distributed algorithm for solving problem (5.3). To start, consider a μ-strongly convex and an L-smooth function $f(x)$ to be minimized over a set of linear constraints

$$\min_{Ax=0} f(x). \tag{5.6}$$

Assume that the problem is feasible, in which case a unique solution exists, denoted by x^*. The Lagrangian dual for the problem in Eq. (5.6) is given by

$$\min_{Ax=0} f(x) = \max_y \left\{ \min_x \left\{ f(x) - \left\langle A^T y, x \right\rangle \right\} \right\}. \tag{5.7}$$

The Lagrangian dual problem can be re-formulated as an equivalent minimization problem, as follows:

$$\min_y \varphi(y) \quad \text{where} \quad \varphi(y) \triangleq \max_x \left\{ \left\langle A^T y, x \right\rangle - f(x) \right\}. \tag{5.8}$$

The function $\varphi(y)$ is μ_φ-strongly convex on $\ker(A^T)^\perp$ with $\mu_\varphi = \sigma_{\min}^+(A)/L$. Moreover, it has L_φ-Lipschitz continuous gradients with $L_\varphi = \sigma_{\max}(A)/\mu$.

Additionally, from Demyanov-Danskin's theorem (see, for example, Proposition 4.5.1 in [123]), it follows that $\nabla\varphi(y) = Ax^*(A^T y)$ where $x^*(A^T y)$ denotes the unique solution to the

inner maximization problem

$$x^*(A^T y) = \arg\max_x \left\{ \left\langle A^T y, x \right\rangle - f(x) \right\}. \tag{5.9}$$

Note that there is no duality gap between the primal problem in Eq. (5.6) and its dual problem in (5.8). Also, the dual problem has a solution (see, for example, Proposition 6.4.2 in [123]). In view of Eq. (5.9), the primal optimal solution x^* is the same as $x^*(A^T y^*)$ where y^* is any dual optimal solution. In general, the dual problem in Eq. (5.8) can have multiple solutions of the form $y^* + \ker(A^T)$ when the matrix A does not have a full row rank. When the solution is not unique, we *will use y^* to denote the smallest norm solution*, and we let R be its norm, i.e. $R = \|y^*\|_2$. In order to find $x^*(A^T y)$ one can use optimal (randomized) numerical methods [198, 201, 202]. In the remainder of this chapter, we will assume that *we have access to $x^*(A^T y)$ explicitly for any given y.* Section 5.3 discusses possible extension when no dual solution is explicitly available.

Definition 20. *A function $f(x)$ is dual-friendly if, for any y, one has immediate access to an explicit (or efficiently computed) solution $x^*(A^T y)$ to the dual subproblem associated with the optimization problem in Eq.* (5.6)*.*

Examples of optimization problems for which Definition 20 holds can be found in the literature, i.e. the entropy-regularized optimal transport problem [203], the entropy linear programming problem [204] or the ridge regression.

Next, we will apply the bound for the FGM algorithm in Eq. (5.5) on the dual problem (5.8), which is not strongly convex in the ordinary sense (on the whole space). However, by choosing $y_0 = x_0 = 0$ in Eq. (5.4) as the initial condition, the algorithm applied to the dual problem will produce iterates that lie in the linear space of gradients $\nabla\varphi(y)$, which are of the form Ax for $x = x^*(A^T y)$. In this case, the dual function $\varphi(y)$ will be strongly convex when y is restricted to the linear space spanned by the range of the matrix A. The iterations in Eq. (5.4) for the dual problem then specialize to the following:

$$y_{k+1} = \tilde{y}_k - \frac{1}{L_\varphi} A x^*(A^T \tilde{y}_k), \tag{5.10a}$$

$$\tilde{y}_{k+1} = y_{k+1} + \frac{\sqrt{L_\varphi} - \sqrt{\mu_\varphi}}{\sqrt{L_\varphi} + \sqrt{\mu_\varphi}} (y_{k+1} - y_k). \tag{5.10b}$$

We will explore the case when the linear constraints $Ax = 0$ represent the network communication constraints as $\sqrt{W}x = 0$ and the function $f(x)$ corresponds to the network function $F(x)$ as defined in Eq. (5.3). Particularly, if we make the change of variables $\sqrt{W} y_k = z_k$

and $\sqrt{W}\tilde{y}_k = \tilde{z}_k$, then the resulting algorithm can be executed in a distributed manner. The interaction between agents is dictated by the term $Wx^*(\tilde{z}_k)$ which depends only on local information. As a result, each agent i in the network has its local variables z_k^i and $\tilde{z}_k^i$, and to compute their value at the next iteration, it only requires the information sent by the neighbors defined by the communication graph $\mathcal{G}$ as follows:

$$z_{k+1}^i = \tilde{z}_k^i - \frac{1}{L_\varphi}\sum_{j=1}^n W_{ij}x_j^*(\tilde{z}_k^j)$$
$$\tilde{z}_{k+1}^i = z_{k+1}^i + \frac{\sqrt{L_\varphi}-\sqrt{\mu_\varphi}}{\sqrt{L_\varphi}+\sqrt{\mu_\varphi}}(z_{k+1}^i - z_k^i).$$

Additionally, the dual subproblem can be computed in a distributed manner at node i as

$$x_i^*(\tilde{z}_k^i) = \arg\max_{x_i}\left\{\left\langle \tilde{z}_k^i, x_i\right\rangle - f_i(x_i)\right\}.$$

We will be interested in finding solutions to the problem in Eq. (5.6) that attain the function value arbitrarily close to the optimal value and have arbitrarily small feasibility violation of the linear constraints. For this, we introduce the following definition.

Definition 21. *[197] A point $\hat{x} \in \mathbb{R}^{nm}$ is called an $(\varepsilon, \tilde{\varepsilon})$-solution of* (5.6) *if the following conditions are satisfied*

$$f(\hat{x}) - f^* \leq \varepsilon \qquad \text{and} \qquad \|A\hat{x}\|_2 \leq \tilde{\varepsilon},$$

where f^* denotes the optimal value for the primal problem in Eq. (5.6).

Note that an $(\varepsilon, \tilde{\varepsilon})$-solution is not an optimal solution of (5.6) in the traditional sense. The point $\hat{x}$ implies only an approximate solution with $\|A\hat{x}\|_2 \leq \tilde{\varepsilon}$.

The next section presents the main results on the optimal convergence rates for different convexity and smoothness assumptions on the functions f_i.

5.2 Optimal Algorithms for Distributed Convex Optimization

Our main results provide convergence rate estimates for the solution of the problem in Eq. (5.1) for four different cases in terms of the properties of the function $F(x) = \sum_{i=1}^n f_i(x_i)$.

Assumption 7. *For a set of functions $\{f_i\}_{i=1,\cdots,n}$ assume:*

(a) Each f_i is μ_i-strongly convex and L_i-smooth, thus F is μ-strongly convex and L-smooth.

(b) Each f_i is μ_i-strongly convex and M_i-Lipschitz on a closed ball around the optimal point with radius equal to the magnitude of the optimal solution, thus F is μ-strongly convex and M-Lipschitz on that closed ball.

(c) Each f_i is convex and L_i-smooth, thus F is convex and L-smooth.

(d) Each f_i is convex and M_i-Lipschitz, thus F is convex and M-Lipschitz.

Moreover, we define $\mu = \min_{1\leq i\leq n} \mu_i$, $L = \max_{1\leq i\leq n} L_i$ and $M = \max_{1\leq i\leq n} M_i$.

Next, we provide a sequence of algorithms and theorems considering each case in Assumption 7. Under each assumption, we present the minimum number of iterations required, for the corresponding algorithm, to reach an approximate solution of the problem in Eq. (5.3).

5.2.1 Sums of Strongly Convex and Smooth Functions

Assume that each f_i in Eq. 5.3 is μ_i-strongly convex and L_i-smooth, thus F is μ-strongly convex and L-smooth. Then we propose Algorithm 1 to be executed distributedly for each of the agents in the network.

Algorithm 1 Distributed FGM for the Dual of strongly convex and smooth problems

All agents set $z_0^i = \tilde{z}_0^i = 0 \in \mathbb{R}^n$ and N.
For each agent $i \in V$
for $k = 0, 1, 2, \cdots, N$ **do**
  $x_i^*(\tilde{z}_k^i) = \arg\max_{x_i} \{\langle \tilde{z}_k^i, x_i\rangle - f_i(x_i)\}$
  Share $x_i^*(\tilde{z}_k^i)$ with neighbors, i.e. $\{j \mid (i,j) \in E\}$.
  $z_{k+1}^i = \tilde{z}_k^i - \frac{\mu}{\lambda_{\max}(W)} \sum_{j=1}^n W_{ij} x_j^*(\tilde{z}_k^j)$
  $\tilde{z}_{k+1}^i = z_{k+1}^i + \frac{\sqrt{\lambda_{\max}(W)/\mu} - \sqrt{\lambda_{\min}^+(W)/L}}{\sqrt{\lambda_{\max}(W)/\mu} + \sqrt{\lambda_{\min}^+(W)/L}} (z_{k+1}^i - z_k^i)$
end for

The next theorem presents our main result regarding the performance of Algorithm 1.

Theorem 36. *Let $F(x)$ be dual friendly and Assumption 7(a) hold. For any $\varepsilon > 0$, the output $x^*(z_N)$ of Algorithm 1 is an $(\varepsilon, \varepsilon/R)$-solution of (5.3) for*

$$N \geq 2\sqrt{\frac{L}{\mu}\chi(W)} \log\left(\frac{2\sqrt{2}\lambda_{\max}(W)R^2}{\mu \cdot \varepsilon}\right),$$

where $R = \|y^*\|_2$, *and* $\chi(W) = \lambda_{\max}(W)/\lambda^+_{\min}(W)$.

Proof. Algorithm 1 follows from the FGM in (5.4) applied to the dual problem (5.8) with the change of variables $\sqrt{W}y_k = z_k$ and $\sqrt{W}\tilde{y}_k = \tilde{z}_k$. Therefore, we are going to use the convergence results of the FGM for the dual problem in terms of the dual variables y_k and $\tilde{y}_k$ and provide an estimate of the convergence rate of in terms of the primal variables.

Initially, it follows from Theorem 2.2.2 in [198], Section 2.2.1, that the sequence of estimates generated by the iterations in (5.10) has the following property:

$$\varphi(y_k) - \varphi^* \leq L_\varphi R^2 \exp\left(-k\sqrt{\frac{\mu_\varphi}{L_\varphi}}\right). \tag{5.12}$$

Moreover, it holds that

$$\varphi^* \leq \varphi(y_{k+1}) \leq \varphi(\tilde{y}_k) - \frac{1}{2L_\varphi}\|\nabla\varphi(\tilde{y}_k)\|_2^2. \tag{5.13}$$

Thus

$$\begin{aligned}\|\nabla\varphi(\tilde{y}_k)\|_2^2 &\leq 2L_\varphi\left(\varphi(y_k) - \varphi^*\right)\\ \|\sqrt{W}x^*(\sqrt{W}\tilde{y}_k)\|_2^2 &\leq 2L_\varphi^2 R^2 \exp\left(-k\sqrt{\frac{\mu_\varphi}{L_\varphi}}\right).\end{aligned}$$

We can conclude that $\|\sqrt{W}x^*(z_k)\|_2 \leq \varepsilon/R$ if $k \geq 2\sqrt{\frac{L_\varphi}{\mu_\varphi}}\log\left(\frac{\sqrt{2}L_\varphi R^2}{\varepsilon}\right)$.
Now, by using the Cauchy–Schwarz inequality, it follows that

$$|\langle y_k, \sqrt{W}x^*(\sqrt{W}y_k)\rangle|^2 \leq \|y_k\|_2^2\|\sqrt{W}x^*(\sqrt{W}y_k)\|_2^2.$$

We can bound $\|y_k\|_2$ following ideas from [205], where it was shown that

$$\|y_k - y^*\|_2 \leq \|y_0 - y^*\|_2.$$

Thus, since we assume $y_0 = 0$, it holds that $\|y_k\|_2 \leq 2\|y^*\|_2 \leq 2R$, then

$$\begin{aligned}|\langle y_k, \sqrt{W}x^*(\sqrt{W}y_k)\rangle|^2 &\leq 4R^2\|\sqrt{W}x^*(\sqrt{W}y_k)\|_2^2,\\ &\leq 8R^4L_\varphi^2 \exp\left(-k\sqrt{\frac{\mu_\varphi}{L_\varphi}}\right).\end{aligned}$$

Therefore $f(x^*(z_k)) - f^* \leq \varepsilon$ if $k \geq 2\sqrt{\frac{L_\varphi}{\mu_\varphi}}\log\left(\frac{2\sqrt{2}L_\varphi^2 R^3}{\varepsilon}\right)$.

Finally, based on Lemma 1 in [204], Algorithm 1 will produce an $(\varepsilon, \varepsilon/R)$-solution if

$$N \geq 2\sqrt{\frac{L_\varphi}{\mu_\varphi}} \log\left(\max\left\{\frac{2\sqrt{2}L_\varphi R^2}{\varepsilon}, \frac{\sqrt{2}L_\varphi R^2}{\varepsilon}\right\}\right).$$

Following the definitions of L_φ, μ_φ, and $\chi(W)$, we obtain the desired result. □

5.2.2 Sums of Strongly Convex and M-Lipschitz Functions on a Bounded Set

Assume that each f_i in Eq. (5.3) is μ_i-strongly convex and M_i-Lipschitz on a bounded set, thus F is μ-strongly convex and M-Lipschitz on that specific set, then, we propose Algorithm 2 to be executed distributedly for each of the agents in the network.

Algorithm 2 Distributed FGM for the Dual of strongly convex and M-Lipschitz problems

1: All agents set $z_0^i = \tilde{z}_0^i = 0 \in \mathbb{R}^n$ and N.
2: For each agent $i \in V$
3: **for** $k = 0, 1, 2, \cdots, N$ **do**
4: $\quad x_i^*(\tilde{z}_k^i) = \arg\max_{x_i} \{\langle \tilde{z}_k^i, x_i\rangle - f_i(x_i)\}$
5: $\quad$ Share $x_i^*(\tilde{z}_k^i)$ with neighbors, i.e. $\{j \mid (i,j) \in E\}$.
6: $\quad z_{k+1}^i = \tilde{z}_k^i - \frac{1}{\lambda_{\max}(W)/\mu + \varepsilon/R^2}\left(\sum_{j=1}^n W_{ij} x_j^*(\tilde{z}_k^j) + \frac{\varepsilon}{R^2} z_k^i\right)$
7: $\quad \tilde{z}_{k+1}^i = z_{k+1}^i + \frac{\sqrt{\lambda_{\max}(W)/\mu + \varepsilon/R^2} - \sqrt{\varepsilon/R^2}}{\sqrt{\lambda_{\max}(W)/\mu + \varepsilon/R^2} + \sqrt{\varepsilon/R^2}}(z_{k+1}^i - z_k^i)$
8: **end for**

The next theorem presents our main result regarding the performance of Algorithm 2.

Theorem 37. *Let $F(x)$ be dual friendly and Assumption 7(b) hold. Moreover, assume $F(x)$ is M-Lipschitz in the set $\{x \mid \|x - x^*\|_2 \leq R_x\}$ with $R_x = \|x^*(0) - x^*\|_2$. For any $\varepsilon > 0$, the output $x^*(z_N)$ of Algorithm 2 is an $(\varepsilon, \varepsilon/R)$-solution of (5.3) for*

$$N \geq 2\sqrt{4\chi(W)\frac{M^2}{\mu \cdot \varepsilon} + 1} \log\left(4\chi(W)\frac{M^2}{\mu \cdot \varepsilon} + 1\right),$$

where $\chi(W) = \lambda_{\max}(W)/\lambda_{\min}^+(W)$.

Proof. Initially, consider the regularized dual function $\hat{\varphi}$ with $\hat{\mu} = \frac{\varepsilon}{4R^2}$, which is $\mu_{\hat{\varphi}}$-strongly

convex with $\mu_{\hat{\varphi}} = \frac{\varepsilon}{4R^2}$, and $L_{\hat{\varphi}}$-smooth with $L_{\hat{\varphi}} = \frac{\lambda_{\max}(W)}{\mu} + \frac{\varepsilon}{4R^2}$. Thus, similarly as in (5.12)

$$\hat{\varphi}(y_k) - \hat{\varphi}^* \leq L_{\hat{\varphi}} \hat{R}^2 \exp\left(-k\sqrt{\frac{\mu_{\hat{\varphi}}}{L_{\hat{\varphi}}}}\right) \leq L_{\hat{\varphi}} R^2 \exp\left(-k\sqrt{\frac{\mu_{\hat{\varphi}}}{L_{\hat{\varphi}}}}\right),$$

where $\hat{R} = \|\hat{y}^*\|_2$, and $\hat{y}^*$ is the smallest norm solution of the regularized dual problem. Note that by definition $\hat{R} = \|\hat{y}^*\|_2 \leq \|y^*\|_2 = R$.

Next, we provide a relation between the distance to optimality of the non-regularized primal problem and the regularized dual problem. Note that for any y it holds that

$$\hat{\varphi}(y) - \hat{\varphi}^* \geq \frac{\|\nabla\hat{\varphi}(y)\|_2^2}{2L_{\hat{\varphi}}} = \frac{\|\nabla\varphi(y) + \hat{\mu}y\|_2^2}{2L_{\hat{\varphi}}} \geq \frac{\hat{\mu}\,\langle y, \nabla\varphi(y)\rangle}{L_{\hat{\varphi}}}.$$

Therefore,

$$\langle y, \nabla\varphi(y)\rangle \leq \frac{L_{\hat{\varphi}}}{\mu_{\hat{\varphi}}}\left(\hat{\varphi}(y) - \hat{\varphi}^*\right) \leq \frac{4}{\varepsilon} L_{\hat{\varphi}}^2 R^4 \exp\left(-k\sqrt{\frac{\mu_{\hat{\varphi}}}{L_{\hat{\varphi}}}}\right).$$

Consequently, if $k \geq 2\sqrt{L_{\hat{\varphi}}/\mu_{\hat{\varphi}}}\log\left(2L_{\hat{\varphi}}R^2/\varepsilon\right)$, then $\langle y, \nabla\varphi(y)\rangle \leq \varepsilon$.

Moreover, it follows from the definition of the regularized dual function that

$$\begin{aligned}
\|\nabla\varphi(y_k)\|_2 &\leq \|\nabla\hat{\varphi}(y_k)\|_2 + \hat{\mu}\|y_k\|_2 \\
&\leq \sqrt{2L_{\hat{\varphi}}(\hat{\varphi}(y) - \hat{\varphi}^*)} + \hat{\mu}\|y_k\|_2 \\
&\leq \sqrt{2}L_{\hat{\varphi}}R\exp\left(-\frac{k}{2}\sqrt{\frac{\mu_{\hat{\varphi}}}{L_{\hat{\varphi}}}}\right) + 2\hat{\mu}\hat{R} \\
&\leq \sqrt{2}L_{\hat{\varphi}}R\exp\left(-\frac{k}{2}\sqrt{\frac{\mu_{\hat{\varphi}}}{L_{\hat{\varphi}}}}\right) + \frac{\varepsilon}{2R}.
\end{aligned}$$

Using the definition of the gradient of the dual function then we have that $\|\sqrt{W}x^*(\sqrt{W}\tilde{y}_k)\|_2 \leq \varepsilon/R$, for $k \geq 2\sqrt{L_{\hat{\varphi}}/\mu_{\hat{\varphi}}}\log\left(\sqrt{2}L_{\hat{\varphi}}R^2/\varepsilon\right)$.

We conclude, from Lemma 1 in [204], that we will have an $(\varepsilon, \varepsilon/R)$ solution of (5.3) if

$$\begin{aligned}
k &\geq 2\sqrt{\frac{L_{\hat{\varphi}}}{\mu_{\hat{\varphi}}}}\log\left(\max\left\{\frac{2L_{\hat{\varphi}}R^2}{\varepsilon}, \frac{\sqrt{2}L_{\hat{\varphi}}R^2}{\varepsilon}\right\}\right) \\
&\geq 2\sqrt{\frac{L_{\hat{\varphi}}}{\mu_{\hat{\varphi}}}}\log\left(\frac{2L_{\hat{\varphi}}R^2}{\varepsilon}\right) \\
&= 2\sqrt{\frac{\frac{\lambda_{\max}(W)}{\mu} + \frac{\varepsilon}{4R^2}}{\frac{\varepsilon}{4R^2}}}\log\left(\frac{2R^2\left(\frac{\lambda_{\max}(W)}{\mu} + \frac{\varepsilon}{4R^2}\right)}{\varepsilon}\right)
\end{aligned}$$

$$= 2\sqrt{\frac{4R^2\lambda_{\max}(W)}{\mu\cdot\varepsilon}+1}\log\left(\frac{4R^2\lambda_{\max}(W)}{\mu\cdot\varepsilon}+1\right).$$

Now, we focus our attention to find a bound on the value R such that we can provide an explicit dependency on the minimum non zero eigenvalue of the graph Laplacian. This will allow us to provide an explicit iteration complexity in terms of the condition number of the graph Laplacian.

Theorem 3 in [197] provides a bound that relates R with the magnitude of the gradient of $F(x)$ at the optimal point $x = x^*$. Particularly, it is shown that

$$R^2 = \|y^*\|_2^2 \leq \frac{\|\nabla F(x^*)\|_2^2}{\sigma_{\min}^+(A)}. \tag{5.14}$$

It was shown in [205] that the iterations generated by the FGM in (5.4) always lie inside an Euclidean ball around the optimal solution y^* (y^* is the optimal solution of the dual problem in this case), with a radius equal to $\|y_0 - y^*\|_2$ which is effectively equal to R given our initialization $z_0 = 0$. The set $\{y \mid \|y - y^*\| \leq R\}$ is defined in the dual variables. However, we seek to provide a condition on the primal variables, i.e., x. It follows from the definition of the function $x^*(\sqrt{W}y)$, that the set $\{y \mid \|y - y^*\| \leq R\}$ is mapped into an Euclidean ball centered at x^*, since the point $x^*(\sqrt{W}y^*) = x^*$. As for the radius, note that $x^*(0) = \arg\min_x F(x)$. Thus, given the assumption that $F(x)$ is M-Lipschitz in the set $\{x \mid \|x - x^*\|_2 \leq R_x\}$ with $R_x = \|x^* - x^*(0)\|$, it holds that

$$R^2 \leq \frac{M^2}{\sigma_{\min}^+(A)}.$$

Therefore to have an $(\varepsilon, \varepsilon/R)$-solution it is necessary that

$$\begin{aligned} k &\geq 2\sqrt{\frac{\lambda_{\max}(W)}{\lambda_{\min}^+(W)}\frac{M^2}{\mu\cdot\varepsilon}+1}\log\left(\frac{\lambda_{\max}(W)}{\lambda_{\min}^+(W)}\frac{M^2}{\mu\cdot\varepsilon}+1\right) \\ &\geq 2\sqrt{4\chi(W)\frac{M^2}{\mu\cdot\varepsilon}+1}\log\left(4\chi(W)\frac{M^2}{\mu\cdot\varepsilon}+1\right). \end{aligned}$$

□

5.2.3 Sums of Smooth Functions

Assume that each f_i in Eq. (5.3) is convex and L_i-smooth, thus F is convex and L-smooth. Then we propose Algorithm 3 to be executed distributedly for each of the agents in the network.

Algorithm 3 Distributed FGM for the Dual of Smooth convex functions

All agents set $z_0^i = \tilde{z}_0^i = 0 \in \mathbb{R}^n$ and N.
For each agent $i \in V$
for $k = 0, 1, 2, \cdots, N$ **do**
  $\hat{x}_i^*(\tilde{z}_k^i) = \arg\max_{x_i} \left\{ \langle \tilde{z}_k^i, x_i \rangle - f_i(x_i) - \frac{\varepsilon}{2R_x^2}\|x_i\|_2^2 \right\}$
  Share $x_i^*(\tilde{z}_k^i)$ with neighbors, i.e. $\{j \mid (i,j) \in E\}$.
  $z_{k+1}^i = \tilde{z}_k^i - \frac{1}{\lambda_{\max}(W)/(\varepsilon/R_x^2)} \sum_{j=1}^n W_{ij} \hat{x}_j^*(\tilde{z}_k^j)$
  $\tilde{z}_{k+1}^i = z_{k+1}^i + \frac{\sqrt{\frac{\lambda_{\max}(W)}{\varepsilon/R_x^2}} - \sqrt{\frac{\lambda_{\min}^+(W)}{L+\varepsilon/R_x^2}}}{\sqrt{\frac{\lambda_{\max}(W)}{\varepsilon/R_x^2}} + \sqrt{\frac{\lambda_{\min}^+(W)}{L+\varepsilon/R_x^2}}} (z_{k+1}^i - z_k^i)$
end for

The next theorem presents our main result regarding the performance of Algorithm 3.

Theorem 38. *Let $F(x)$ be dual friendly and Assumption 7(c) hold. For any $\varepsilon > 0$, the output $x^*(z_N)$ of Algorithm 3 is an $(\varepsilon, \varepsilon/R)$-solution of (5.3) for*

$$N \geq 2\sqrt{\left(\frac{2LR_x^2}{\varepsilon} + 1\right)\chi(W)} \log\left(\frac{8\sqrt{2}\lambda_{\max}(W)R^2R_x^2}{\varepsilon^2}\right).$$

where $\chi(W) = \lambda_{\max}(W)/\lambda_{\min}^+(W)$ and $R_x = \|x^* - x^*(0)\|_2$.

Proof. Initially, consider the regularized problem

$$\min_{\sqrt{W}x=0} \hat{F}(x) \qquad \text{where} \qquad \hat{F}(x) \triangleq F(x) + \frac{\varepsilon}{2R_x^2}\|x - x^*(0)\|_2^2, \tag{5.15}$$

where $F(x)$ is defined in (5.3). The function $\hat{F}(x)$ is $\hat{\mu}$-strongly convex with $\hat{\mu} = \frac{\varepsilon}{2R_x^2}$ and $\hat{L}$-smooth with $\hat{L} = L + \hat{\mu}$. Given that the regularized primal function is strongly convex and smooth, we can use the results from Theorem 36. Particularly, in order to have an $(\varepsilon/2, \varepsilon/(2R))$-solution of problem (5.15), one can use Algorithm 1 with

$$N \geq 2\sqrt{\frac{\hat{L}}{\hat{\mu}}\chi(W)} \log\left(\frac{4\sqrt{2}\lambda_{\max}(W)R^2}{\hat{\mu} \cdot \varepsilon}\right)$$

$$= 2\sqrt{\frac{L + \frac{\varepsilon}{2R_x^2}}{\frac{\varepsilon}{2R_x^2}}\chi(W)} \log\left(\frac{4\sqrt{2}\lambda_{\max}(W)R^2}{\frac{\varepsilon}{2R_x^2}\cdot\varepsilon}\right)$$
$$= 2\sqrt{\left(\frac{2LR_x^2}{\varepsilon}+1\right)\chi(W)} \log\left(\frac{8\sqrt{2}\lambda_{\max}(W)R^2R_x^2}{\varepsilon^2}\right).$$

Having an $(\varepsilon/2, \varepsilon/(2R))$-solution of problem (5.15), guarantees that $\hat{x}_N^*$ is an $(\varepsilon, \varepsilon/(R))$-solution of problem (5.3), and the desired result follows. □

5.2.4 Sums of Convex and M-Lipschitz Functions

Assume that each f_i in Eq. 5.3 is convex and M_i-Lipschitz, thus F is convex and M-Lipschitz. Then we propose Algorithm 4 to be executed distributedly for each of the agents in the network.

Algorithm 4 Distributed FGM for the Dual of M-Lipschitz functions

All agents set $z_0^i = \tilde{z}_0^i = 0 \in \mathbb{R}^n$ and N.
For each agent $i \in V$
for $k = 0, 1, 2, \cdots, N$ **do**
  $\hat{x}_i^*(\tilde{z}_k^i) = \arg\max_{x_i}\left\{\langle \tilde{z}_k^i, x_i\rangle - f_i(x_i) - \frac{\varepsilon}{2R_x^2}\|x_i\|_2^2\right\}$
  Share $x_i^*(\tilde{z}_k^i)$ with neighbors, i.e. $\{j \mid (i,j) \in E\}$.
  $z_{k+1}^i = \tilde{z}_k^i - \frac{1}{\lambda_{\max}(W)/(\varepsilon/R_x^2)+\varepsilon/R^2}\left(\sum_{j=1}^n W_{ij}x_j^*(\tilde{z}_k^j) + \frac{\varepsilon}{R^2}z_k^i\right)$
  $\tilde{z}_{k+1}^i = z_{k+1}^i + \frac{\sqrt{\frac{\lambda_{\max}(W)}{\varepsilon/R_x^2}+\varepsilon/R^2}-\sqrt{\varepsilon/R^2}}{\sqrt{\frac{\lambda_{\max}(W)}{\varepsilon/R_x^2}+\varepsilon/R^2}+\sqrt{\varepsilon/R^2}}(z_{k+1}^i - z_k^i)$
end for

The next theorem presents our main result regarding the performance of Algorithm 4.

Theorem 39. *Let $F(x)$ be dual friendly and Assumption 7(d) hold. For any $\varepsilon > 0$, the output $x^*(z_N)$ of Algorithm 4 is an $(\varepsilon, \varepsilon/R)$-solution of (5.3) for*

$$N \geq 2\sqrt{16\chi(W)\frac{M^2R_x^2}{\varepsilon^2}+1}\log\left(16\chi(W)\frac{M^2R_x^2}{\varepsilon^2}+1\right),$$

where $\chi(W) = \lambda_{\max}(W)/\lambda_{\min}^+(W)$, $R = \|y^*\|_2$, and $R_x = \|x^* - x^*(0)\|_2$.

Proof. Consider again, as in Theorem 38, the regularized problem (5.15) where $F(x)$ is defined in (5.3). The function $\hat{F}(x)$ is $\hat{\mu}$-strongly convex with $\hat{\mu} = \frac{\varepsilon}{2R_x^2}$. However, we

have assumed now that $F(x)$ is not smooth. Nevertheless, from Theorem 37, we have that Algorithm 2 will generate an $(\varepsilon/2, \varepsilon/(2R))$-solution of (5.15), namely x_N^*, for

$$\begin{aligned} N &\geq 2\sqrt{8\chi(W)\frac{M^2}{\hat{\mu}\cdot\varepsilon}+1}\log\left(8\chi(W)\frac{M^2}{\hat{\mu}\cdot\varepsilon}+1\right) \\ &= 2\sqrt{8\chi(W)\frac{M^2}{\frac{\varepsilon}{2R_x^2}\cdot\varepsilon}+1}\log\left(8\chi(W)\frac{M^2}{\frac{\varepsilon}{2R_x^2}\cdot\varepsilon}+1\right) \\ &= 2\sqrt{16\chi(W)\frac{M^2R_x^2}{\varepsilon^2}+1}\log\left(16\chi(W)\frac{M^2R_x^2}{\varepsilon^2}+1\right). \end{aligned}$$

Therefore, $x^*(z_N)$ is an $(\varepsilon, \varepsilon/R)$-solution for problem (5.3). □

5.3 Discussion and Extensions

Table 5.2 presents a summary of the results presented in Section 5.3.1. In particular, it shows the number of communication rounds required to obtain an $(\varepsilon, \varepsilon/R)$-solution for each the presented properties of the function $F(x)$.

Table 5.2: A summary of algorithmic performance.

Property of $F(x)$	**Iterations Required**
μ-strongly convex and L-smooth	$\tilde{O}\left(\sqrt{(L/\mu)\chi(W)}\right)$
μ-strongly convex and M-Lipschitz	$\tilde{O}\left(\sqrt{(M^2/(\mu\varepsilon))\chi(W)}\right)$
L-smooth	$\tilde{O}\left(\sqrt{(LR_x^2/\varepsilon)\chi(W)}\right)$
M-Lipschitz	$\tilde{O}\left(\sqrt{(M^2R_x^2/\varepsilon^2)\chi(W)}\right)$

The estimates in Table 5.2 are optimal up to logarithmic factors. Particularly in the smooth cases, where $L < \infty$, these estimates follow from classical centralized complexity estimation of the FGM algorithm. In the distributed setting, one has to perform $\sqrt{\chi(W)}\log(\varepsilon^{-1})$ additional consensus steps at each iteration. This corresponds to the number of iterations needed to solve the consensus problem

$$\min_x \frac{1}{2}\langle x, Wx\rangle, \tag{5.16}$$

where W is a communication matrix. FGM provides a direct estimate on the number of iterations required to reach consensus; particularly, we need $O(\sqrt{\chi(W)}\log\varepsilon^{-1})$, where

we have used the fact that (5.16) is $\sigma_{\min}(\sqrt{W})$-strongly convex in $x_0 + \ker(W)$ and has $\sigma_{\max}(\sqrt{W})$-Lipschitz continuous gradients. Moreover, it follows that this estimate cannot be improved up to constant factors.

The specific value of $\chi(W)$ and its dependency on the number of nodes m has been extensively studied in the literature of distributed optimization [25]. Table 1.1 provides an extensive list of *worst-case* dependencies of the spectral gap for large classes of graphs. Particularly, for fixed undirected graphs, in the worst case we have $\chi(W) = O(n^2)$ [96]. This matches the best upper bound found in the literature of consensus and distributed optimization [207, 208]. Thus, the constraint described as $\sqrt{W}x = 0$ should be preferred over the description as $Wx = 0$, even though both representations correctly describe the consensus subspace $x_1 = \ldots = x_n$. Particularly, when we pick $A = \sqrt{W}$, we have $\chi(A^T A) = \chi(W)$ instead of $\chi(W^T W) = \chi(W^2) \gg \chi(W)$.

Note that we typically do not know R or R_x. Thus, we require a method to estimate the strong convexity parameter $\hat{\mu}$ which is challenging [209, 210]. Therefore, we can apply the restarting technique on μ [210]. The payment for that is an 8 multiplicative factor in the estimation [211]. Similarly, a generalization of the FGM algorithm can be proposed when L_φ is unknown [204]. The specific details of this generalization are beyond the scope of this work.

Considering the general problem in Eq. (5.1), the condition number L/μ can be large if one of the μ_i is small. Thus, we can formulate another equivalent problem as

$$\min_{\sqrt{W}x=0} F_\alpha(x) = \sum_{i=1}^{n} f_i(x_i) + \frac{\alpha}{2} \langle x, Wx \rangle , \tag{5.17}$$

where

$$F_\alpha \text{ is } \mu \geq \min \left\{ \sum_{i=1}^{n} \mu_i, \alpha \lambda_{\min}(W) \right\} \text{-strongly convex and has}$$
$$L \leq \left(\max_{i=1,\ldots,m} L_i + \alpha \lambda_{\max}(W) \right) \text{-Lipschitz continuous gradients.}$$

Moreover, if we set $\alpha = O(\sum_{k=1}^{n} \mu_k / \lambda_{\min}(W))$, one can solve problem (5.17) with relative precision ε after

$$\tilde{O}\left(\sqrt{(L/\bar{\mu} + \chi(W))\, \chi(W)} \right),$$

communication steps, where $\bar{\mu} = \sum_{i=1}^{n} \mu_i$. This estimate shows that we can replace the smallest strong convexity constant for the sum among all of them, but we have to pay

an additive price proportional to the condition number of the graph. This result can be extended to the case when $F(x)$ is just smooth by using the regularization technique with $\mu_i = \varepsilon/(nR_{x_i}^2) = \varepsilon/(R_x^2)$.

The cases when $F(x)$ is convex or strongly convex can be generalized to p-norms, with $p \geq 1$, see [187]. Particularly, the definitions of the condition number $\chi(\cdot)$ need to be defined accordingly. Let us introduce a norm $\|x\|_p^2 = \|x_1\|_p^2 + ... + \|x_n\|_p^2$ for $p \geq 1$ and assume that $F(x)$ is μ-strongly convex and L-Lipschitz continuous gradient in this (new) norm $\|\cdot\|_p$ (in $\mathbb{R}^{mn}$), see [212] (Lemma 1), [213] (Lemma 1) and [206] (Theorem 1). Thus

$$\chi(W) = \frac{\max_{\|h\|=1} \frac{<h,Wh>}{\mu}}{\min_{\|h\|=1, h\perp\ker(W)} \frac{<h,Wh>}{L}}.$$

5.3.1 The Case When $F(x)$ is Not Dual Friendly

The results in Theorems 36, 37, 38, and 39 assume $F(x)$ is *dual-friendly*. In this section, we explore the case when no exact solution to the dual problem is available.

When the function is strongly convex and smooth or just smooth, one can solve Eq. (5.9) using FGM. Therefore, we can find a solution for the dual problem, i.e. $x^*(A^T y)$, in a logarithmic number of iterations from a desired relative precision δ. Specifically, when the function is strongly convex and smooth, we can solve the auxiliary problem in $O(\sqrt{L/\mu}\log(\delta^{-1}))$ oracle calls (i.e. calculation of $\nabla f(x)$). On the other hand, when the function is only smooth, we require $O(\sqrt{LR_x^2/\varepsilon}\log(\delta^{-1}))$ iterations. Both estimates are optimal up to logarithmic factor. Therefore, in those cases, we obtain optimal convergence rates both in the number of communication rounds (Ax, $A^T y$ multiplications) and oracle calls (computations of $\nabla f(x)$).

In the non-smooth cases, we might use another approach. Consider the case where $f(x)$ is convex and M-Lipschitz and apply Nesterov's smoothing technique [206, 214] to (5.7). Thus, we can solve the composite type mixed smooth/non-smooth type problem

$$\min_{\|x\|_2 \leq R_x} \underbrace{G_\varepsilon(Ax)}_{O(1/\varepsilon)\text{-smooth}} + \underbrace{f(x)}_{M\text{-Lipschitz}}, \tag{5.18}$$

where

$$G_\varepsilon(Ax) = \max_y \left\{ \langle y, Ax\rangle - \frac{\varepsilon}{2R^2}\|y\|_2^2 \right\} = \frac{R^2}{2\varepsilon}\|Ax\|_2^2.$$

It holds that $G_\varepsilon(Ax)$ is $(\sigma_{\max}(A^T)R^2/\varepsilon)$-smooth. Thus, using Lan's accelerated gradient sliding [215], one can find an ε-solution (in function value) of (5.18) without any auxiliary

dual problem, after

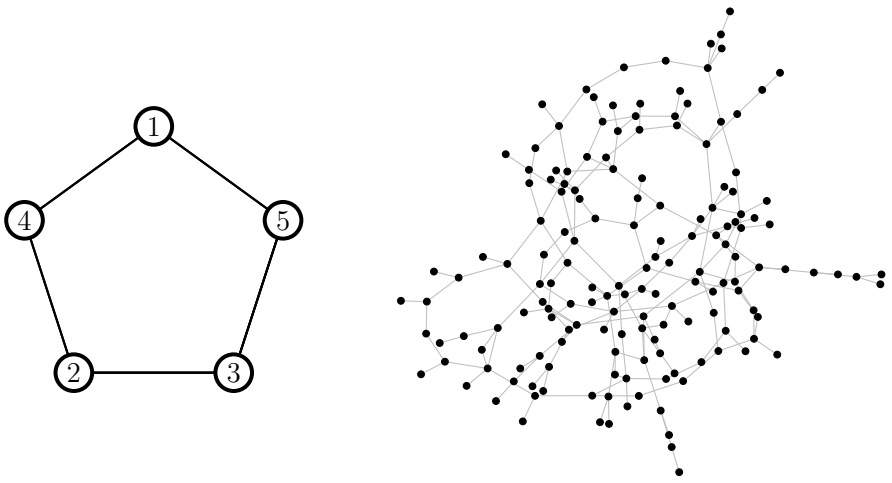
$$O(\sqrt{(M^2R_x^2/\varepsilon^2)\chi(A^TA)}) \text{ and } O\left(M^2R_x^2/\varepsilon^2\right)$$

communication rounds and gradient computations respectively. Unfortunately, in this approach we can guarantee $\|Ax_N\| \leq \varepsilon/R$ only in the best case [216].

Using the restart technique [211, 217] one can extend Lan's accelerated gradient sliding for $f(x)$ being μ-strongly convex. At the k-th restart the number of communication rounds is $O(\sqrt{\sigma_{\max}(A)R^2/(\mu\varepsilon)})$ and the number of gradient computations is $O\left((2^kM^2)/(\varepsilon^2R_x^2)\right)$. This allows to improve estimates for (5.18) to $\tilde{O}(\sqrt{M^2/(\mu\varepsilon)\chi(A^TA)})$ communication rounds and $O\left(M^2/\mu\varepsilon\right)$ gradient computations. These estimates are optimal up to logarithmic factors. Moreover, one can extend these results to stochastic optimization problems and the estimations will not change [197].

5.4 Simulation Results

In this section, we will provide experimental results that show the performance of the optimal distributed algorithms presented in the previous section for cycle and Erdős-Rényi random graph of various sizes. We choose the cycle graph ($\chi(W) = O(n^2)$) and the Erdős-Rényi random graph ($\chi(W) = O(\log(n))$), see Fig. 5.1. Moreover, we show the scalability properties of the algorithms for networks of increasing size.



(a) Cycle graph (b) Erdős-Rényi random graph

Figure 5.1: Two examples of networks of agents. (a) A cycle graph with 5 agents. (b) An Erdős-Rényi random graph with 160 agents.

Particularly for the cycle graph network of 5 agents shown in Fig. 5.1(a) agent 1 can share information with agents 4 and 5, agent 5 shares information with agents 1 and 3, and similarly for the other agents. Thus, the corresponding interaction matrix $\bar{W}$ is

$$\bar{W} = \begin{bmatrix} 2 & -1 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

Initially, consider the *regression* (strongly convex and smooth) problem

$$\min_{z \in \mathbb{R}^m} \frac{1}{2nl}\|b - Hz\|_2^2 + \frac{1}{2}c\|z\|_2^2, \tag{5.19}$$

to be solved distributedly over a network. Each entry of the data matrix $H \in \mathbb{R}^{nl \times m}$ is generated as an independent identically distributed random variable $H_{ij} \sim \mathcal{N}(0,1)$; the vector of associated values $b \in \mathbb{R}^{nl}$ is generated as a vector of random variables where $b = Hx^* + \epsilon$ for some predefined $x^* \in \mathbb{R}^m$ and $\epsilon \sim \mathcal{N}(0, 0.1)$. The columns of the data matrix H and the output vector b are evenly distributed among the agents with a total of l data points per agent. The regularization constant is set to $c = 0.1$. Thus, each agent has access to a subset of points such that

$$b^T = [\underbrace{b_1^T}_{\text{Agent } 1} \mid \underbrace{b_2^T}_{\text{Agent } 2} \mid \cdots \mid \underbrace{b_n^T}_{\text{Agent } n}] \quad \text{and} \quad H^T = [\underbrace{H_1^T}_{\text{Agent } 1} \mid \underbrace{H_2^T}_{\text{Agent } 2} \mid \cdots \mid \underbrace{H_n^T}_{\text{Agent } n}],$$

where $b_i \in \mathbb{R}^l$ and $H_i \in \mathbb{R}^{l \times m}$ for each i. In this setup, each agent i has a private local function

$$f_i(x_i) \triangleq \frac{1}{2nl}\|b_i - H_i x_i\|_2^2 + \frac{1}{2}\frac{c}{m}\|x_i\|_2^2.$$

Moreover, the optimization problem in Eq. 5.19 is equivalent to

$$\min_{\sqrt{W}x = 0} \sum_{i=1}^{n} \left(\frac{1}{2}\frac{1}{nl}\|b_i - H_i x_i\|_2^2 + \frac{1}{2}\frac{c}{m}\|x_i\|_2^2 \right),$$

where $W = \bar{W} \otimes I_m$.

Figure 5.2 shows experimental results for the ridge regression problem for a cycle graph and

an Erdős-Rényi random graph. For each type of graph we show the distance to optimality as well as the distance to consensus for a fixed graph with $n = 100$, $m = 10$ and $l = 100$. Additionally, the scalability of the algorithm is shown by plotting the required number of steps to reach an accuracy of $\epsilon = 1 \cdot 10^{-10}$ versus the number of nodes in the graph. We compare the performance of the proposed algorithm with some of the state-of-the-art methods for distributed optimization. ***Dist-Opt*** refers to Algorithm 1. ***NonAcc-Dist*** refers to the non-accelerated version of Algorithm 1. ***FGM*** is the centralized FGM. ***Acc-DNGD*** refers to the algorithm proposed in [192] with parameter $\eta = 0.1$ and $\alpha = \sqrt{\mu\eta}$. ***EXTRA*** refers to the algorithm proposed in [76] with parameter $\alpha = 1$. ***DIGing*** refers to the algorithm proposed in [24] with parameter $\alpha = 0.1$. Figure 5.2 shows linear convergence rate with faster performance than other algorithms and linear scalability with respect to the size of the cycle graphs.

Now, consider the Kullback-Leibler (KL) barycenter computation problem (strongly convex and M-Lipschitz)

$$\min_{z \in S_1(m)} \sum_{i=1}^{n} D_{KL}(z \| q_i) \triangleq \sum_{i=1}^{n} \sum_{j=1}^{m} z^i \log \left(z_i / [q_i]_j \right),$$

where $S_1(m)$ is the unit simplex in $\mathbb{R}^m$ and $q_i \in S_1(m)$ for all $i \in V$. Each agent has a private probability distribution q^i and seek to compute the a probability distribution that minimizes the average KL distance to the distributions $\{q_i\}_{i=1,\ldots,n}$. Figure 5.3 shows the results for the KL barycenter problem for a cycle graph with $n = 100$, $m = 10$ and various values of the regularization parameter. We show the distance to optimality as well as the distance to consensus and the scalability of the algorithm.

In Eq. (5.19), if we assume $c = 0$ and H_i is a *wide* matrix where $m \gg l$ (i.e., the dimension of the data points is much larger than the number of data points per agent), then the resulting problem is smooth but no longer strongly convex. Figure 5.4 compares the performance of the proposed method with the distributed accelerated method proposed in [192] for non-strongly convex functions (Acc-DNGD-NSC) for a fixed regularization value $\hat{\mu} = 1 \cdot 10^{-6}$. The bottom two plots in Figure 5.4 show the distance to optimality and distance to consensus over an Erdős-Rényi random graph of two different sizes, namely $n = 20$ and $n = 100$. The top two plots in Figure 5.4 show the same comparison for a cycle network. Figure 5.5 shows the performance of the proposed algorithm over an Erdős-Rényi random graph with $n = 50$, $m = 20$ and $l = 10$, for different values of the regularization parameter. As expected, smaller values of the regularization parameter increase the precision of the algorithm but hinder its rate. As presented in Table 5.1, the algorithms have similar convergence rates, as shown by

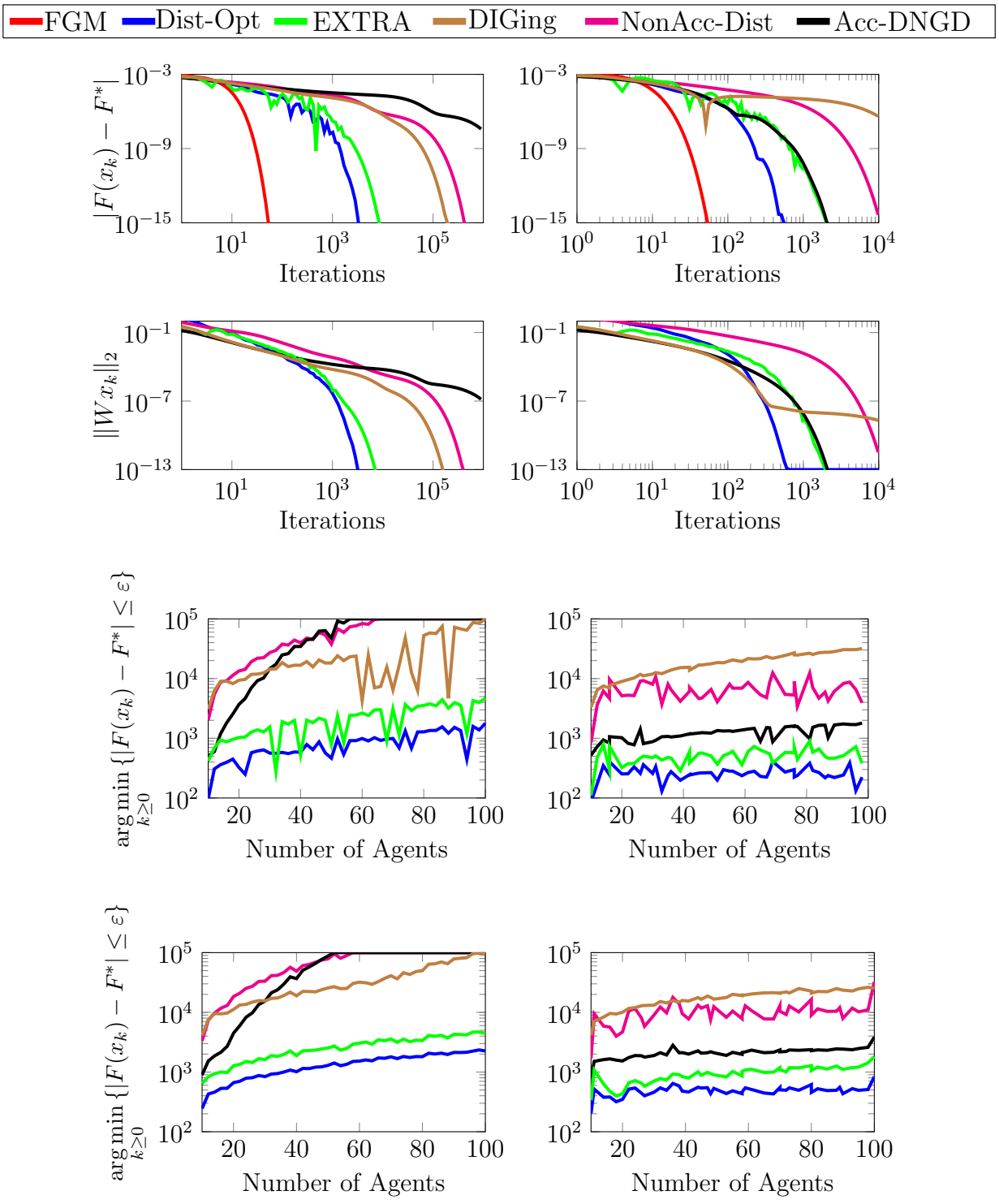


Figure 5.2: Distance to optimality and consensus, and network scalability for a strongly convex and smooth problem. Left plots correspond to cycle graphs and right plots correspond to Erdős-Rényi random graphs.

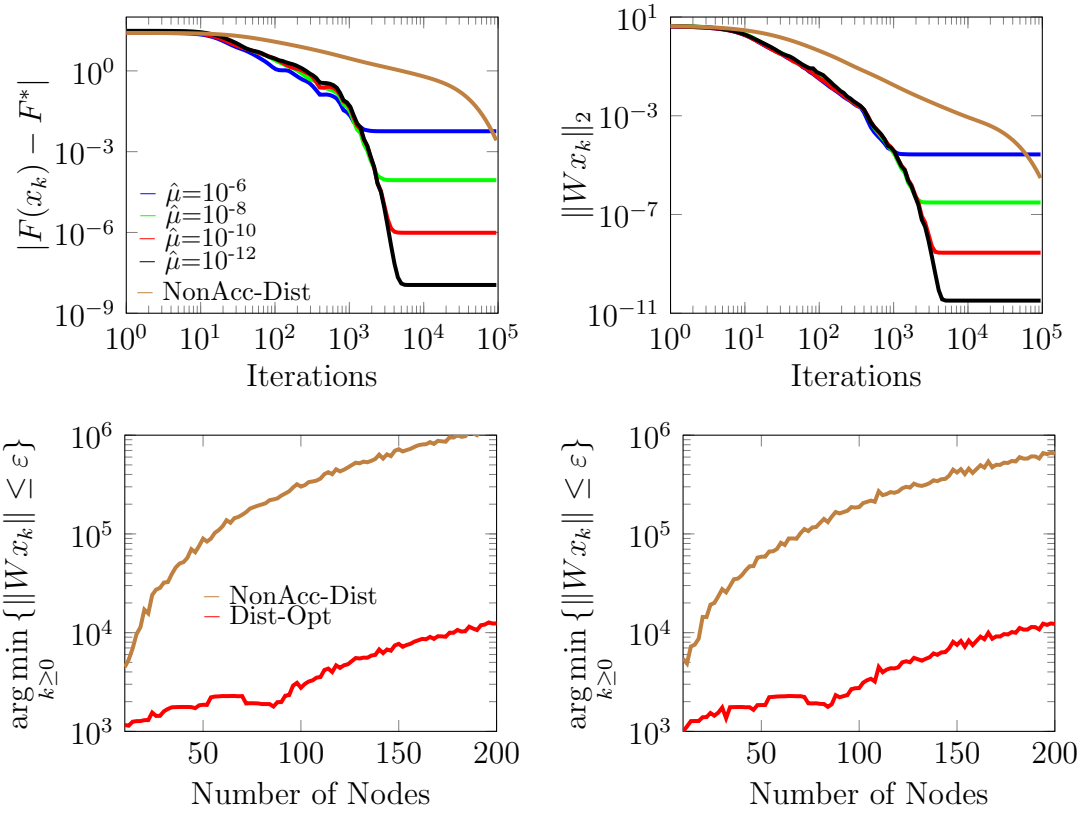


Figure 5.3: Distance to optimality and consensus, and network scalability for a strongly convex and M-Lipschitz problem over a cycle graph with $n = 100$s, $m = 10$ and various values of the regularization parameter $\hat{\mu}$. The brown line shows the performance for the non-accelerated distributed gradient descent of the dual problem.

the intersection of the curves around the accuracy point corresponding to the regularization parameter. Nevertheless, as seen in Figure 5.2, Acc-DNDG-NSC has the worst scalability with respect to the number of nodes, which is particularly evident for the cycle graph.

5.5 Distributed Computation of Wasserstein Barycenters

One of the common uses of the Wasserstein distance is the aggregation of distributions by considering their barycenter [218], which itself is another distribution [219]. Wasserstein barycenters has been shown superior to traditional Euclidean-based methods in a range of application such as image processing [218], economics and finance [220] and condensed matter physics [221]. Figure 5.6 shows a sample of 100 images of the digit 7 from the MNIST dataset [222], and their respective Euclidean mean and Wasserstein mean. The Wasserstein barycenter better captures the structural features of the input images.

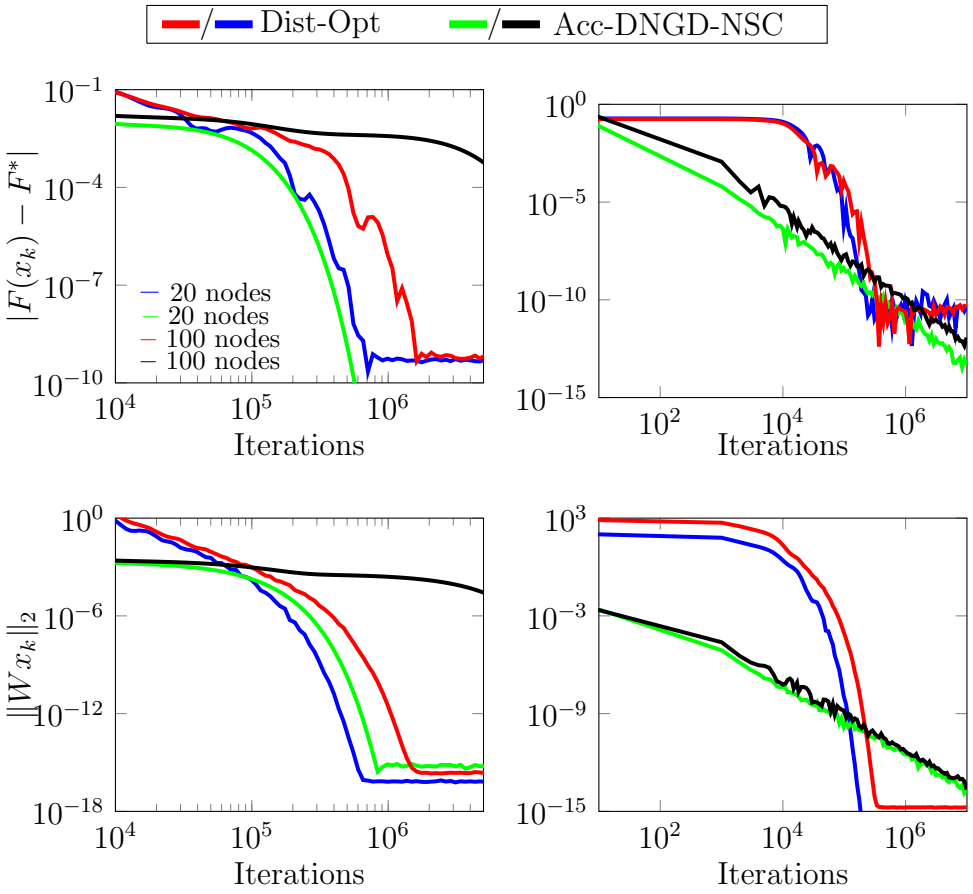


Figure 5.4: Distance to optimality and consensus for Erdős-Rényi random graph (right) and cycle graph (left) with $n = 20$ and $n = 100$.

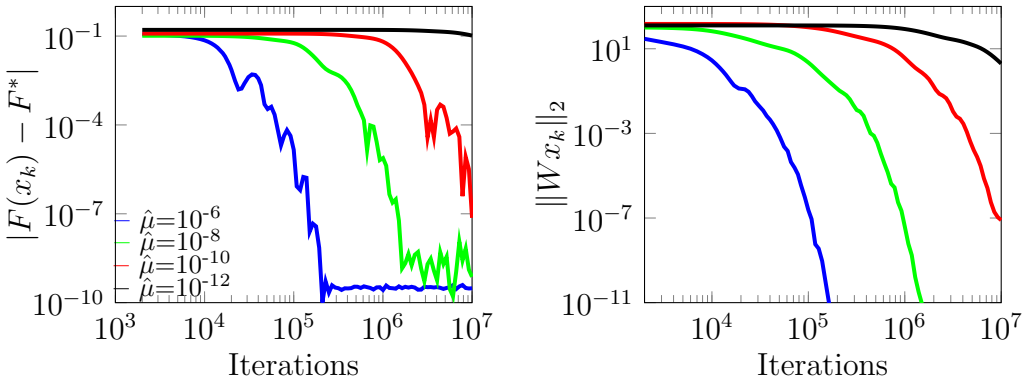


Figure 5.5: Distance to optimality and consensus for a smooth problem over a Erdős-Rényi random graph with $n = 50$, $m = 20$, $l = 10$ and various values of the regularization parameter $\hat{\mu}$.

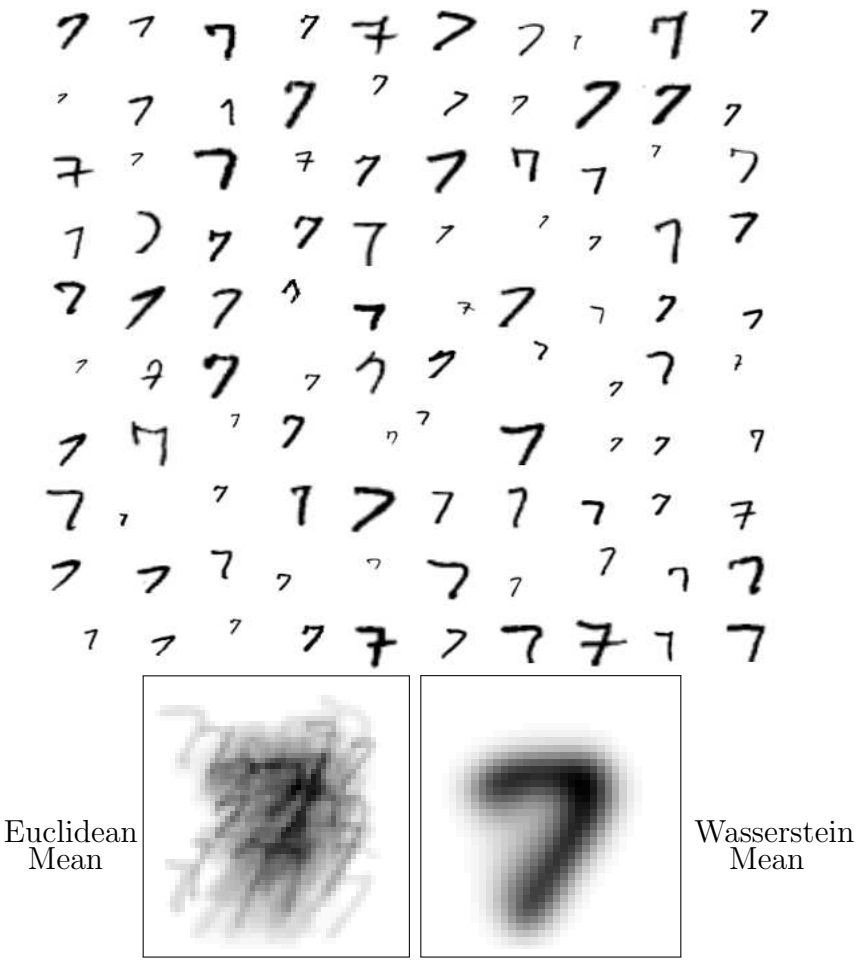


Figure 5.6: Samples of the digit 7 from the MNIST dataset and comparison of their Euclidean and Wasserstein Barycenters.

For discrete and finite distributions, the Wasserstein barycenter can be efficiently computed by solving a large linear program [223] or using regularization to approximate a solution efficiently and exploit its convenient algebraic properties [218, 219, 224]. In this section, we consider the problem of computation of Wasserstein barycenters over networks. The flexibilities induced by the distributed setup make it suitable for problems involving large quantities of data with no centralized storage [22, 24, 20, 225]. Particularly, we assume a group of agents is connected over a network, and each agent locally holds a probability distribution with finite support. The group seeks to compute the Wasserstein barycenter of all distributions in the network cooperatively. Figure 5.7 shows an Erdős-Rényi random graph with 160 agents where each agent holds a sample of the digit 7 from the MNIST dataset.

Distributed consensus with the Wasserstein metric was introduced in [87, 226]. In [226], the authors showed asymptotic convergence to the Wasserstein barycenter of the initial distri-

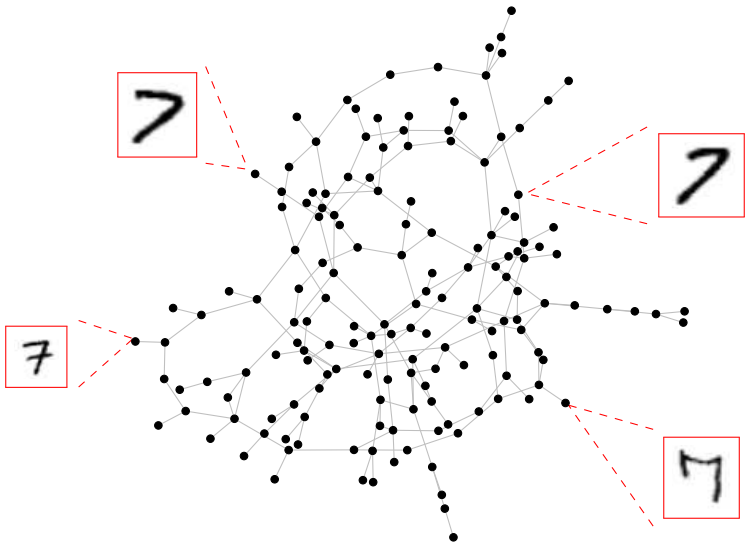


Figure 5.7: Erdős-Rényi random graph where each agent privately holds a sample of the digit 7 from the MNIST dataset.

butions given some weak connectivity assumptions on the graph over which agents exchange information. Nevertheless, the proposed algorithm requires that each agent computes an exact Wasserstein barycenter of local distributions at each iteration. Although one can have closed form solutions for some families of continuous distributions, in general, the problem can be intractable. On the other hand, a recent approach [227] explores the computational advantages of a dual formulation of the Wasserstein barycenter and exploits the parallelizable structure of the problem to propose a scalable, communication-efficient algorithm for its computation on arbitrary continuous distributions. Nevertheless, it requires a central fusion center that coordinates the actions of the parallel machines.

In contrast with existing literature [226, 227], we propose a first-order algorithm that can be executed distributedly over a network with unknown topology. We derive an explicit convergence rate of the order $O(1/k^2)$ with an additional cost that depends on the condition number of the graph over which the agents interact. Additionally, we present two numerical examples to illustrate and validate our results. First, we show some basic properties of the algorithm for the problem of computing the Wasserstein barycenter of univariate, discrete and truncated Gaussian distributions. Then, we show the result of applying our algorithm to a subset of the MNIST digit database on a large-scale network of 1000 agents.

5.5.1 Problem Statement

Consider two probability distributions $p, q \in S_1(m)$ with support on a finite set of points $\{x_i \in \mathbb{R}^d\}_{i=1}^m$ such that $p(x_i) = p_i$ and $q(x_i) = q_i$. Moreover, consider a non-negative symmetric matrix M, where $[M]_{ij} \in \mathbb{R}_+$ accounts for the cost of moving mass from p_i to bin q_j. Without loss of generality, in the numerical example we will consider the Euclidean costs where $[M]_{ij} = \|x_i - x_j\|_2^2$. Additionally, define the set of couplings or *transportation polytope* $U(p, q)$ as

$$U(p, q) \triangleq \left\{X \in \mathbb{R}_+^{m \times m} \mid X\mathbf{1} = p, X^T\mathbf{1} = q\right\}.$$

The entropy-regularized optimal transport problem [228] seeks to minimize the transportation costs while maximizing the entropy (maximum-entropy principle, $\gamma > 0$) and is defined as

$$\mathcal{W}_\gamma(p, q) \triangleq \min_{X \in U(p,q)} \left\{\langle M, X\rangle - \gamma E(X)\right\}, \tag{5.20}$$

where

$$\langle M, X\rangle \triangleq \sum_{i=1}^m \sum_{j=1}^m M_{ij}X_{ij} \quad \text{and} \quad E(X) \triangleq -\sum_{i=1}^m \sum_{j=1}^m h(X_{ij}),$$

and $\forall x > 0, h(x) \triangleq x \log x$ and $h(0) \triangleq 0$. The solution $\mathcal{W}_0(p, q)$ is called the Wasserstein distance between p and q and if $\gamma > 0$ $\mathcal{W}_\gamma(p, q)$ is known as regularized (or smoothed) Wasserstein distance. For $\gamma > 0$, problem (5.20) is strongly convex and admits a unique solution X^*.

For simplicity, let us introduce the notation $\mathcal{W}_{\gamma,q}(p)$ for fixed probability distribution $q \in S_1(m)$

$$\mathcal{W}_{\gamma,q}(p) \triangleq \mathcal{W}_\gamma(p, q).$$

One particular advantage of entropy-regularizing the Wasserstein distance is that there exist closed-form representations for the dual problem and its gradients [218, 224] where the Fenchel-Legendre transform of (5.20) is defined as

$$\mathcal{W}^*_{\gamma,q}(y) \triangleq \max_{p \in S_1(m)} \left\{\langle y, p\rangle - \mathcal{W}_{\gamma,q}(p)\right\}. \tag{5.21}$$

In [89], other regularization functions were explored. The squared 2-norm was shown to

produce sparse transportation plans. In this chapter, we will use the entropy regularization. Nevertheless, our results extend naturally to regularization functions, especially those that admit closed-form solution of dual gradients. The next theorem states the closed-form solutions of the dual problem and the gradient of the entropy regularized Wasserstein barycenter problem.

Theorem 40 (Theorem 2.4 in [224]). *For $\gamma > 0$, the Fenchel-Legendre dual function $\mathcal{W}^*_{\gamma,q}(p,q)$ is differentiable and its gradient $\nabla\mathcal{W}^*_{\gamma,q}(y)$ is $1/\gamma$-Lipschitz in the 2-norm with*

$$\begin{aligned}\mathcal{W}^*_{\gamma,q}(y) &= \gamma\left(E(q) + \langle q, \log K\alpha\rangle\right) \quad \textit{and}\\ \nabla\mathcal{W}^*_{\gamma,q}(y) &= \alpha \circ (K \cdot q/(K\alpha)) \in S_1(m),\end{aligned}$$

where $y \in \mathbb{R}^m$, $\alpha = \exp(y/\gamma)$ and $K = \exp(-M/\gamma)$.

We will use the results of Theorem 40 to design an algorithm for the computation of the Wasserstein barycenter on graphs based on recent ideas of dual approaches for convex optimization problems with affine constraints [229, 230] and optimal algorithms for distributed optimization [231].

The Wasserstein barycenter [218, 219] of a family of discrete distributions $(q_1, q_2, \cdots, q_n)$ in $S_1(m)$ is defined as the solution to the following optimization problem

$$\min_{p\in S_1(m)} \sum_{i=1}^{n} \lambda_i \mathcal{W}_{\gamma,q_i}(p), \tag{5.22}$$

where $\{\lambda\}_{i=1}^n$ is a set of weights that describe the relative importance of each distribution. Without loss of generality we assume that $\lambda_1 = \cdots = \lambda_n$.

The Wasserstein barycenter is an extension of the Euclidean barycenter to nonlinear metric spaces corresponding to the empirical Fréchet mean [232]. The existence and uniqueness of a Wasserstein barycenter has been studied in the literature [233]. Problem (5.22) is strictly convex and admits a unique solution, denoted by p^* [224].

For the distributed computation of Wasserstein barycenters, let us introduced stacked the column vectors $\mathtt{p} = [p_1^T, \cdots, p_n^T]^T$ and $\mathtt{q} = [q_1^T, \cdots, q_n^T]^T$, where $\forall\ i \in V$, $p_i, q_i \in S_1(m)$, and rewrite the problem (5.22) in an equivalent form

$$\min_{\substack{p_1=\cdots=p_n\\ p_1,\ldots,p_n\in S_1(m)}} \quad \sum_{i=1}^{n} \mathcal{W}_{\gamma,q_i}(p_i). \tag{5.23}$$

We denote the unique solution of (5.23) by $\mathtt{p}^* = [(p_1^*)^T, \cdots, (p_n^*)^T]^T$ with $p_1^* = \cdots = p_n^* =$

p^*. We seek to solve problem (5.23) in a distributed manner over a network, where each distribution q_i is held by an agent i on a network.

Therefore, one can equivalently rewrite problem (5.23) as

$$\min_{\substack{p_1,\dots,p_n \in S_1(m) \\ \sqrt{W}\mathtt{p}=0}} \mathcal{W}_{\gamma,\mathtt{q}}(\mathtt{p}) = \sum_{i=1}^{n} \mathcal{W}_{\gamma,q_i}(p_i). \tag{5.24}$$

Note that the constraint set $\{p_1, \dots, p_n \in S_1(m) \mid \sqrt{W}\mathtt{p} = 0\}$ is the same as the set $\{p_1, \dots, p_n \in S_1(m) \mid p_1 = \cdots = p_n\}$, since $\ker(\sqrt{W}) = \operatorname{span}(\mathbf{1})$ due to the connectivity of the graph $\mathcal{G}$.

Next, we state the proposed algorithm for solving the optimization problem in Eq. (5.24) and analyze its convergence rate.

5.5.2 Algorithm and Results

In [187], the authors proposed a novel analysis for the minimization of strongly convex functions with affine constraints of the form

$$\min_{Ax=0} f(x), \tag{5.25}$$

where $f(x)$ is 1-strongly convex with respect to the p-norm with the corresponding dual problem is defined as

$$\min_{y} g(y) \quad \text{where} \quad g(y) = \max_{x} \{\langle A^T y, x\rangle - f(x)\}. \tag{5.26}$$

We denote $x^*(A^T y)$ the solution to the problem defining $g(y)$. The dual function $g(y)$ is L-smooth with $L = \|A\|_{L^1 \to L^2} = \max_{i=1,\cdots,n} \|A_i\|_2^2$, where A_i is the i-th column of A. Additionally, from Demyanov-Danskin's theorem (see, for example, Proposition 4.5.1 in [123]), it follows that $\nabla g(y) = Ax^*(A^T y)$. Thus, one can use accelerated first order methods such as Nesterov's fast gradient [190] or one of its recent reformulations to obtain an approximate solution. For example, the linear coupling method presented in [234], for problem (5.26), can be written as

$$y_{k+1} = \tau_k z_k + (1 - \tau_k) w_k, \tag{5.27a}$$

$$w_{k+1} = y_{k+1} - \frac{1}{L} Ax^* \left(A^T y_{k+1}\right), \tag{5.27b}$$

$$z_{k+1} = z_k - \alpha_{k+1} Ax^* \left(A^T y_{k+1}\right), \tag{5.27c}$$

where $\alpha_{k+1} = (k+2)/(2L)$ and $\tau_k = 2/(k+2)$. Note that the update rules in (5.27), as proposed in [234], are defined for the primal problem; thus, y_k here should be understood as x_k in [234], and similarly w_k here should be understood as y_k in [234].

The novelty in [187] lies in the statement of the convergence rate of the accelerated methods in terms of the duality gap and the constraint violation. Additionally, it was shown that for the linear coupling accelerated algorithm [229] one can guarantee that the solutions will remain in a closed ball around the optimal solution with a radius proportional to the distance between the initial point of the algorithm and the optimal solution. Next, we state a technical result based on [187] that will help us in the design and analysis of our proposed algorithm for the distributed computation of the Wasserstein barycenters.

Theorem 41. *The fast gradient method based on linear coupling in Eq.* (5.27) *applied to problem* (5.26)*, with* $w_0 = y_0 = z_0 = 0$*, has the following properties:* $\forall\ k \geq N$ *it holds that*

$$g(w_k) + f(\breve{x}_k) \leq \varepsilon \qquad \textit{and} \qquad \|A\breve{x}_k - b\|_2 \leq \varepsilon/R,$$

where $\breve{x}_k = \sum_{t=0}^{k-1} \frac{(t+2)}{k(k+3)} x^*(A^T y_{t+1})$*,* $N \triangleq \sqrt{16LR^2/\varepsilon}$*,* $R = \|y^*\|_2 < \infty$ *and* y^* *is the optimal point of* $g(\cdot)$ *with minimal norm.*

Proof. The proof consists in combining Theorem 2 in [235] and proof of Theorem 1 in [236]. □

Problem in Eq. (5.24) can be equivalently reformulated as the maximization problem

$$\max_{\substack{p_1,\ldots,p_n \in S_1(m) \\ \sqrt{W}\mathtt{p}=0}} -\mathcal{W}_{\gamma,\mathtt{q}}(\mathtt{p}) = \sum_{i=1}^{n} \mathcal{W}_{\gamma,q_i}(p_i),$$

with its corresponding Lagrangian dual problem

$$\min_{\mathtt{y}} \max_{p_1,\ldots,p_n \in S_1(m)} \left\{ \langle \mathtt{y}, \sqrt{W}\mathtt{p} \rangle - \mathcal{W}_{\gamma,\mathtt{q}}(\mathtt{p}) \right\},$$

where $\mathtt{y} = [y_1^T \cdots y_n^T]^T$.

Moreover, the Fenchel-Legendre transform of $\mathcal{W}_{\gamma,q_i}(p_i)$ is

$$\mathcal{W}^*_{\gamma,q_i}([\sqrt{W}\mathtt{y}]_i) = \max_{p_i \in S_1(m)} \left\{ \left\langle [\sqrt{W}\mathtt{y}]_i, p_i \right\rangle - \mathcal{W}_{\gamma,q_i}(p_i) \right\}.$$

where $[\sqrt{W}y]_i$ is equivalent form of $\sum_j^n \sqrt{W}_{ij} y_j$, where $\sqrt{W}_{ij} = [\sqrt{\bar{W}}]_{ij} \otimes I_m$.

Therefore, we can rewrite the problem (5.24) as follows:

$$\min_{\mathtt{y}} \mathcal{W}^*_{\gamma,\mathtt{q}}(\sqrt{W}\mathtt{y}) = \sum_{i=1}^{n} \mathcal{W}^*_{\gamma,q_i}([\sqrt{W}\mathtt{y}]_i). \tag{5.28}$$

Additionally, from Theorem 40 the gradient can be expressed in closed form as

$$\nabla \mathcal{W}^*_{\gamma,q_i}\left([\sqrt{W}y]_i\right) = \sum_{j=1}^{n} \sqrt{W}_{ij} p^*_j\left([\sqrt{W}y]_j\right),$$

where

$$p^*_j(\tilde{y}) = \alpha(\tilde{y}) \circ \left(K \cdot \frac{q_j}{(K\alpha(\tilde{y}))}\right).$$

Moreover, it holds that one can recover the solution $\mathtt{p}^*$ to the primal problem (5.24) from a solution $\mathtt{y}^*$ to the dual problem (5.28) as

$$\mathtt{p}^* = \mathtt{p}^*\left(\sqrt{W}\mathtt{y}^*\right).$$

The optimality relation between the dual and the primal problem follows from Theorem 3.1 in [224]. In general, the dual problem (5.28) can have multiple solutions of the form $\mathtt{y}^* + \ker(\sqrt{W})$ when the matrix $\sqrt{W}$ does not have a full row rank. When the solution is not unique, we *will use y^* to denote the smallest norm solution*, and we let R be its norm, i.e. $R = \|\mathtt{y}^*\|_2$.

The entropy regularization term is γ-strongly convex with respect to the 1-norm over the probability simplex $S_1(m)$. As a consequence, the computation of the Wasserstein barycenter of a set of discrete probability distributions $\{q_i\}_{i=1}^n$ is equivalent to solving the dual decomposable L-smooth (with respect to the 2-norm) optimization problem (5.28) with $L = \|\sqrt{W}\|^2_{L^1\to L^2}/\gamma$ [237]. Specifically, in this setup it holds that

$$\begin{aligned}\|\sqrt{W}\|^2_{L^1\to L^2} &= \max_{i=1,\cdots,n} \|\sqrt{W}_i\|_2^2 = \max_{i=1,\cdots,n} \sqrt{W}_i^T \sqrt{W}_i \\ &= \max_{i=1,\cdots,n} [W]_{ii} = d_{\max}.\end{aligned}$$

We can explicitly write the Nesterov's accelerated gradient method (FGM) [198] for smooth functions. Particularly, we follow the linear coupling approach recently proposed

in [234]. Setting $\hat{\mathtt{w}}_k = \hat{\mathtt{z}}_k = \hat{\mathtt{y}}_k = \mathbf{0}$, the FGM generates iterates according to:

$$\hat{\mathtt{y}}_{k+1} = \tau_k \hat{\mathtt{z}}_k + (1-\tau_k)\hat{\mathtt{w}}_k \tag{5.29a}$$

$$\hat{\mathtt{w}}_{k+1} = \hat{\mathtt{y}}_{k+1} - \frac{1}{L}\sqrt{W}\mathtt{p}^*\left(\sqrt{W}\hat{\mathtt{y}}_{k+1}\right) \tag{5.29b}$$

$$\hat{\mathtt{z}}_{k+1} = \hat{\mathtt{z}}_k - \alpha_{k+1}\sqrt{W}\mathtt{p}^*\left(\sqrt{W}\hat{\mathtt{y}}_{k+1}\right) \tag{5.29c}$$

where $\alpha_{k+1} = (k+2)/(2L)$ and $\tau_k = 2/(k+2)$.

Unfortunately, algorithm (5.29) cannot be executed in a distributed manner. Although the entries of local gradient vectors can be computed independently by each node, the sparsity pattern of the matrix $\sqrt{W}$ need not be the same as the communication constraints induced by the graph $\mathcal{G}$. Thus, the variables $\hat{\mathtt{w}}_k$ and $\hat{\mathtt{z}}_k$ cannot be computed on the network. This problem is solved by a change of variables such that $\tilde{\mathtt{y}} = \sqrt{W}\hat{\mathtt{y}}$, $\tilde{\mathtt{w}} = \sqrt{W}\hat{\mathtt{w}}$ and $\tilde{\mathtt{z}} = \sqrt{W}\hat{\mathtt{z}}$.

Algorithm 5 presents the resulting distributed accelerated gradient method for the dual problem of the Wasserstein barycenter problem.

Algorithm 5 Distributed Computation of Wasserstein Barycenters

Require: Each agent $i \in V$ is assigned its distribution q_i.

All agents set $\tilde{w}_0^i = \tilde{y}_0^i = \tilde{z}_0^i = \mathbf{0} \in \mathbb{R}^n$ and N

Set $K = \exp(-M/\gamma)$

For each agent $i \in V$:

for $k = 0, 1, 2, \cdots, N-1$ **do**

  $\tau_k = \frac{2}{k+2}$ and $\alpha_{k+1} = \frac{k+2}{2}\frac{1}{L}$

  $\tilde{y}_{k+1}^i = \tau_k \tilde{z}_k^i + (1-\tau_k)\tilde{w}_k^i$

  $p_i^*(\tilde{y}_{k+1}^i) = \exp\left(\frac{\tilde{y}_{k+1}^i}{\gamma}\right) \circ \left(K \cdot \frac{q_i}{K\exp\left(\frac{\tilde{y}_{k+1}^i}{\gamma}\right)}\right)$

  Share $p_i^*(\tilde{y}_{k+1}^i)$ with $\{j \mid (i,j) \in E\}$

  $\tilde{w}_{k+1}^i = \tilde{y}_{k+1}^i - \frac{1}{L}\sum_{j=1}^n W_{ij} p_j^*(\tilde{y}_{k+1}^j)$

  $\tilde{z}_{k+1}^i = \tilde{z}_k^i - \alpha_{k+1}\sum_{j=1}^n W_{ij} p_j^*(\tilde{y}_{k+1}^j)$

end for

Set $(y_N^*)_i = \tilde{w}_N^i$, $\forall i \in V$

Set $(p_N^*)_i = \sum_{k=0}^{N-1} \frac{(k+2)}{N(N+3)} p_i^*(\tilde{y}_{k+1}^i)$, $\forall i \in V$

Based on [187], we can guarantee that Algorithm 5 generates sequences of vectors $\{\tilde{\mathtt{y}}_k, \tilde{\mathtt{w}}_k, \tilde{\mathtt{z}}_k\}$ which remain in a ball $B_R(0)$ with $R = \|\tilde{\mathtt{y}}_0 - \tilde{\mathtt{y}}^*\|_2 = \|\tilde{\mathtt{y}}^*\|_2$. Now, we are ready to state our main result that provides a convergence rate for Algorithm 5 with explicit dependencies on the problem parameters and the topology of the network.

Theorem 42. *Assume that* $\|\nabla \mathcal{W}^*_{\gamma,\mathfrak{q}}(\tilde{\mathfrak{y}})\|_2 \leq G$ *on a ball* $B_R(0)$,. *Then, it holds that that after*

$$N \geq \sqrt{\frac{16G^2}{\gamma \cdot \varepsilon}\chi(W)}$$

iterations, the outputs of Algorithm 5; i.e.,

$$\mathfrak{p}^*_N = [(p^*_N)^T_1, \cdots, (p^*_N)^T_n]^T \qquad \textit{and} \qquad \mathfrak{y}^*_N = [(y^*_N)^T_1, \cdots, (y^*_N)^T_n]^T$$

have the following properties:

$$\mathcal{W}_{\gamma,\mathfrak{q}}(\mathfrak{p}^*_N) + \mathcal{W}^*_{\gamma,\mathfrak{q}}(\mathfrak{y}^*_N) \leq \varepsilon \quad \textit{and} \quad \|\sqrt{W}\mathfrak{p}^*_N\|_2 \leq \varepsilon/R,$$

where $\chi(W) = d_{\max}/d_{\min}$.

Proof. The dual function $\mathcal{W}^*_{\gamma,\mathfrak{q}}(\mathfrak{y})$ is $(d_{\max}/\gamma)$-smooth. Thus, from Theorem 41 that

$$\mathcal{W}_{\gamma,\mathfrak{q}}(\mathfrak{p}^*_N) + \mathcal{W}^*_{\gamma,\mathfrak{q}}(\mathfrak{y}^*_N) \leq \varepsilon \;\text{ and }\; \|\sqrt{W}\mathfrak{p}^*_N\|_2 \leq \varepsilon/R$$

holds for $k \geq \sqrt{16 d_{\max} R^2/(\gamma\varepsilon)}$. Moreover, considering the boundedness of the gradients of the dual function, we can estimate the radius as

$$R^2 \leq \frac{\|\nabla \mathcal{W}^*_{\gamma,\mathfrak{q}}(\mathfrak{y}^*)\|_2^2}{\min\limits_{\substack{x \in E^\perp \ker(\sqrt{W}), u \in H^* \\ \|x\|_E = 1, \|u\|_{H^*} = 1}} \{\langle u, \sqrt{W}x\rangle\}} = \frac{G^2}{d_{\min}}.$$

Thus, we require $k \geq \sqrt{\frac{16G^2}{\gamma\cdot\varepsilon}\frac{d_{\max}}{d_{\min}}}$ and the desired result follows from the definition of $\chi(W)$. □

Theorem 42 provides the minimum number of iterations required for the proposed algorithm to reach some arbitrary relative accuracy in the solution of the distributed Wasserstein Barycenter problem. The convergence rate is shown to be of the order $O(1/k^2)$ which has been shown to be optimal for smooth convex optimization problems [198] with an additional cost proportional to the square root of the number of agents in the network in the worst case.

In general, one might be interested in finding a Wasserstein barycenter for the original Wasserstein distance with no regularization term, that is, solving the problem in Eq. (5.22) with $\gamma = 0$. The next theorem explains a choice of γ that provides converge rate result with respect to the non-regularized optimal transport based on the iterates of Algorithm 5.

Theorem 43. *Assume that* $\|\nabla \mathcal{W}^*_{\gamma,\mathtt{q}}(\tilde{\mathtt{y}})\|_2 \leq G$ *on a ball* $B_R(0)$*, and set* $\gamma = \varepsilon/(4n\log m)$*. Then, it holds that after*

$$N \geq \sqrt{\frac{128G^2 n \log m}{\varepsilon^2}\chi(W)}$$

iterations the outputs of Algorithm 5, i.e.,

$$\mathtt{p}^*_N = [(p^*_N)^T_1, \cdots, (p^*_N)^T_n]^T \qquad \textit{and} \qquad \mathtt{y}^*_N = [(y^*_N)^T_1, \cdots, (y^*_N)^T_n]^T,$$

have the following properties:

$$\mathcal{W}_{0,\mathtt{q}}(\mathtt{p}^*_N) - \mathcal{W}_{0,\mathtt{q}}(\mathtt{p}^*) \leq \varepsilon \quad \textit{and} \quad \|\sqrt{W}\mathtt{p}^*_N\|_2 \leq \varepsilon/(2R),$$

where $\chi(W) = d_{\max}/d_{\min}$*.*

Proof. Considering weak duality $\mathcal{W}^*_{\gamma,\mathtt{q}}(\mathtt{y}^*_N) \geq -\mathcal{W}_{\gamma,\mathtt{q}}(\mathtt{p}^*)$ and Theorem 42 for $\varepsilon \to \varepsilon/2$, obtain

$$\mathcal{W}_{\gamma,\mathtt{q}}(\mathtt{p}^*_N) - \mathcal{W}_{\gamma,\mathtt{q}}(\mathtt{p}^*) \leq \varepsilon/2. \tag{5.30}$$

By the choice of γ, for $\forall\ i = 1, ..., n$, it holds that

$$\begin{aligned} \mathcal{W}_{\gamma,q_i}(p^*_i) - \mathcal{W}_{0,q_i}(p^*_i) &\leq \varepsilon/(2n), \\ \mathcal{W}_{\gamma,\mathtt{q}_i}((p^*_N)_i) &\geq \mathcal{W}_{0,q_i}((p^*_N)_i). \end{aligned}$$

Summing these inequalities and combining the result with (5.30), the desired result follows. □

5.5.3 Numerical Experiments

In this section, we show two numerical experiments to validate the results from Theorem 42. We explore the problem of computing Wasserstein barycenter of univariate, discrete and truncated Gaussian densities and the computation of the Wasserstein barycenter of a subsample of 1000 digit images from the MNIST dataset.

Barycenter of Gaussian Distributions

Initially, we explore the computation of Wasserstein barycenters for sets of univariate, discretized and truncated Gaussian densities [218]. We consider a network of agents where

each agent i holds an univariate, discretized and truncated Gaussian distribution, with mean $\mu_i \in [-5, 5]$, standard deviation $\sigma \in [0.1, 2]$ and equally spaced support of 100 points in $[-5, 5]$. The entropy regularization parameter is set to $\gamma = 0.1$. Figure 5.8 shows the distance to optimality and the distance to consensus of the outputs of Algorithm 5 when the agents are connected over various graph topologies and a fixed size of 50 nodes. Also, Figure 5.8 shows the scalability of the algorithm, i.e., the number of iterations required to reach an ε accuracy in the distance to optimality and consensus for networks of increasing size.

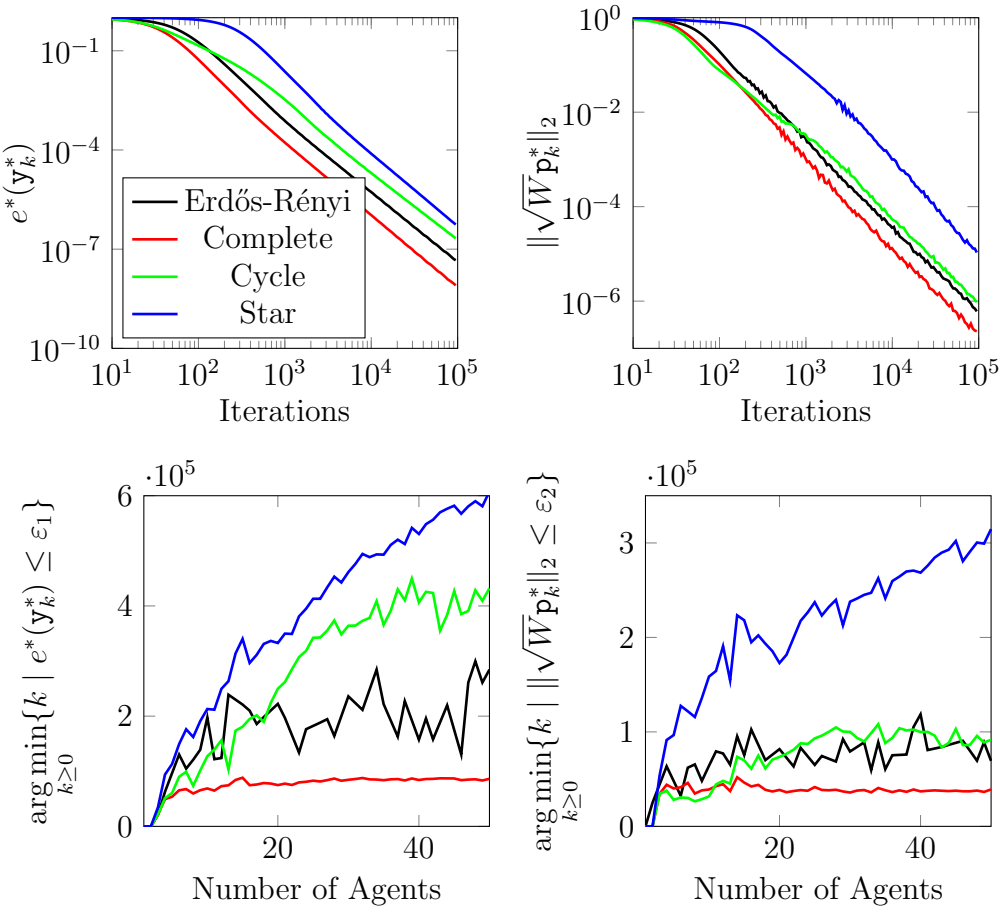


Figure 5.8: Optimality and scalability of Algorithm 5 for various graphs.
$e^*(\mathrm{y}_k^*) = (\mathcal{W}^*_{\gamma,\mathrm{q}}(\mathrm{y}_k^*) - \mathcal{W}^*_{\gamma,\mathrm{q}}(\mathrm{y}^*))/(\mathcal{W}^*_{\gamma,\mathrm{q}}(\mathrm{y}_0^*) - \mathcal{W}^*_{\gamma,\mathrm{q}}(\mathrm{y}^*))$, $\varepsilon_1 = 1 \cdot 10^{-8}$ and $\varepsilon_2 = 1 \cdot 10^{-6}$.

MNIST Dataset

We randomly sample 1000 images for each digit of the MNIST dataset [222, 238]. Each image has 28×28 pixels and is scaled uniformly at random between 0.5 and 2 of its size

and randomly located on a larger 56×56 blank image. The pixel values of the image are normalized to add up to 1. We assign one sample from each digit to each agent on a group of 1000 agents, and the objective is to jointly compute the Wasserstein barycenter for each digit of the 1000 samples present in the network. Each agent owns only one image, and these images are different. In total, the number of images assigned to each agent is equal to the number of digits. The agents are connected over an Erdős-Rényi random graph with 1000 nodes and connectivity parameter 4/1000. The entropy regularization parameter is set to $\gamma = 0.01$. Figure 5.9 shows the local barycenter of the 9 digits for a subset of 3 agents in the network at various number of iterations.

5.6 Conclusions

We have provided convergence rate estimates for the solution of convex optimization problems in a distributed manner. The provided complexity bounds depend explicitly on the properties of the function to be optimized. If $F(x)$ is smooth, then our estimates are optimal up to logarithmic factors; otherwise, our estimates are optimal up to constant factors. The inclusion of the graph properties in terms of $\sqrt{\chi(W)}$ shows the additional price to be paid in contrast with classical (centralized/non-distributed) optimal estimates. The authors recognize that the proposed algorithms required, to some extent, global knowledge about the graph properties and the condition number of the network function. Nevertheless, we aimed to provide a theoretical foundation for the performance limits of the distributed algorithms. The cases where global information is not available require additional study.

As an application example, we developed a novel algorithm for the distributed computation of Wasserstein barycenters over networks where a group of agents connected over a network and each agent holds some local probability distribution with finite support. Our results provably guarantee that all agents in the network will converge to the Wasserstein barycenter of all distribution on the network. We provide an explicit and non-asymptotic convergence rate of the order $O(1/k^2)$ with an additional cost proportional to the condition number of the graph over which the agents exchange information.

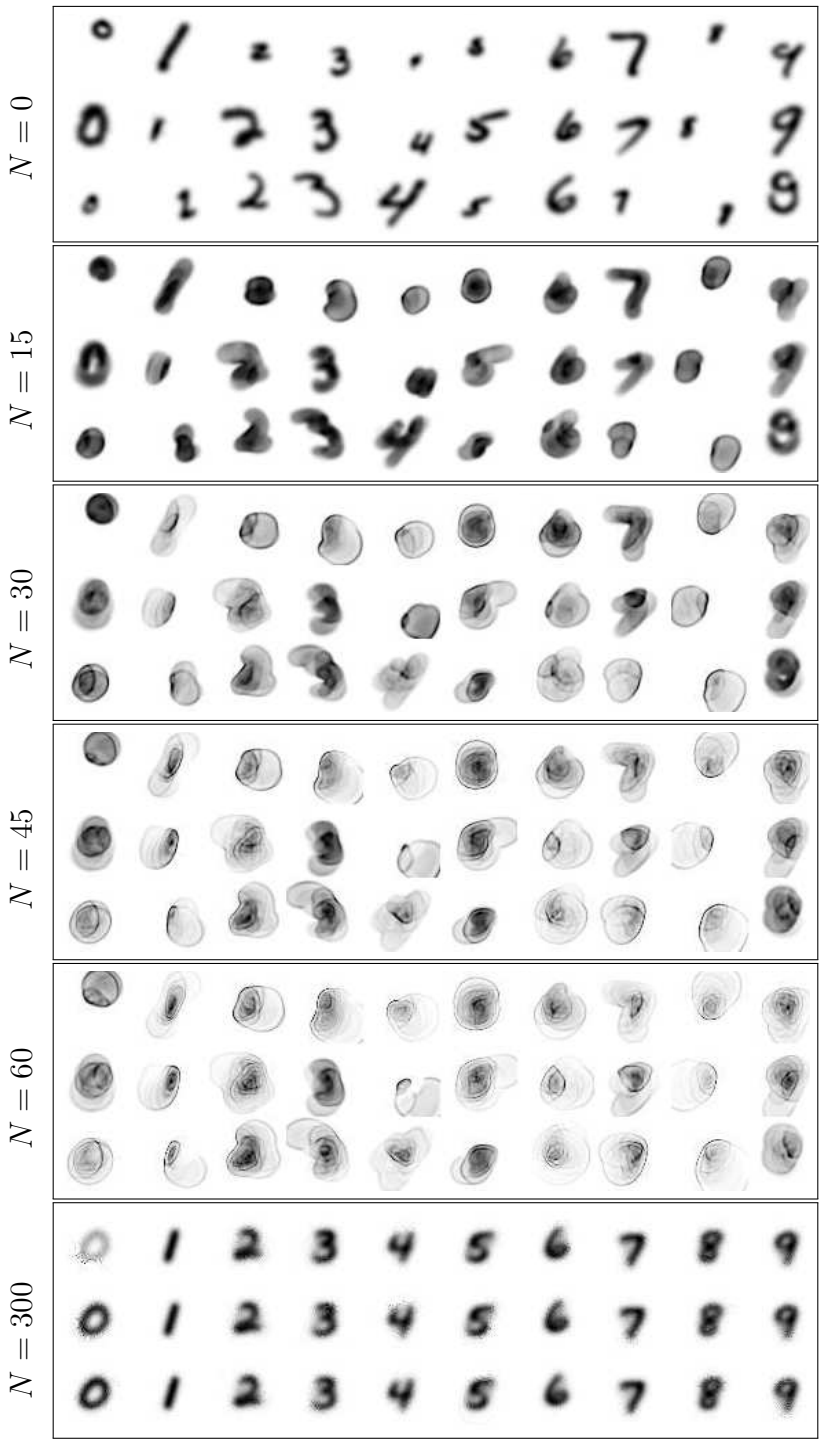


Figure 5.9: Local Wasserstein barycenter of the digits of the MNIST dataset for a subset of 3 agents out of 1000 on an Erdős-Rényi random graph. A video of the evolution of the local barycenters for 10 agents is available at `http://bit.ly/2t9fn0Y`.

CHAPTER 6

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

6.1 Conclusions

The distributed setup, where one needs to make global decisions using the local information only, is particularly well-suited for modern applications of data science, where data are sparse, hard to transmit, or stored distributedly. The main idea and contribution of this dissertation is the strengthening of connections between the challenges of processing massive data sets and the mathematical foundations of optimization, statistics, and network science. Particularly, we ought to bridge the theory of network science and mathematical programming with applications of statistical inference and belief systems over networks, and its connection to inherently distributed systems.

In this dissertation, we have mainly focused on the questions of *efficiency* and *scalability* of algorithms that can be executed in a distributed manner over a network. We have presented our results grouped in three particular problems.

The first one is concerned with the graph-theoretical analysis of the convergence properties of belief systems with logic constraints. We have provided novel graph-theoretical analysis of the questions of convergence, convergence rate and the limit value of such belief systems with a special interest in understanding the explicit influence of the topology of the social network involved and the network of induced by the logic constraints.

Second, we shifted our attention to the problem of distributed statistical inference. We have provided a novel connection between Bayesian posteriors and stochastic approximation algorithms that allowed us to propose a series of new algorithms for statistical estimation problems over networks where agents or nodes are oblivious to the topology of the network. For finite hypotheses sets, we have provided three new algorithms for large classes of networks, namely, time-varying undirected graphs, time-varying directed graphs and fixed undirected graphs. In each of these network classes, we have provided explicit, non-asymptotic convergence rates for the proposed algorithms for worst case networks. Also, we have studied the distributed statistical inference problem when the hypothesis sets are compact subsets of

$\mathbb{R}^d$. We have shown non-asymptotic belief concentration rates for a new distributed learning algorithm for static undirected graphs. Moreover, we provided a general distributed learning algorithm for distributed parameter estimation problems when observations come from members of the exponential family of distributions. For the case of Gaussian observations, we extended our results to time-varying directed graphs.

Finally, we studied the fundamental properties of solving convex optimization problems over networks. Particularly, we follow a dual approach for the design of distributed optimal algorithms for convex optimization with various convexity and smoothness properties. For the case of static undirected graphs, we propose optimal algorithms for the minimization of the sum of strongly convex and smooth functions, either strongly convex or smooth functions, or just convex functions. We show that these optimal algorithms can be executed over arbitrary static and undirected networks and they achieve the same convergence rates as their centralized counterparts. However, there is an additional multiplicative factor proportional to the topology of the network where the problem is being solved. We show that this dependency is linear in the number of nodes in the network in the worst case graph in a Euclidean setting. Then, we use these results to develop a novel algorithm for the distributed computation of Wasserstein barycenters over networks. We show the proposed algorithm achieves optimal convergence rates for the entropy regularized optimal transport problem as well as the non-regularized one. Given the geometry of the optimal transport problem, the dependency on the network topology is shown to be the ratio between the maximum degree and minimum degree among all nodes in the network.

6.2 Directions for Future Research

Each of results presented in this dissertation opens up a number of problems that require further study as we will discuss next.

- The problem of tracking optimal hypothesis when its distributions are changing with time requires further study. Ideas from social sampling can also be incorporated in this framework, where the dimension of the beliefs is large and only partial beliefs are transmitted. Moreover, studying the influence of corrupted measurements or malicious agents is also of interest, especially in the setting of social networks.

- The exploration about how variations in stochastic approximation algorithms will produce new non-Bayesian update rules for more general problems remains to be explored. Promising directions include acceleration results for proximal methods, other Bregman

distances or constraints within the space of probability distributions. Furthermore, we have modeled interactions between agents as exchanges of local probability distributions (i.e., beliefs) between neighboring nodes in a graph. The interesting open question is to understand to what extent this can be reduced when agents transmit only an approximate summary of their beliefs. We anticipate that future work will additionally consider the effect of parametric approximations allowing nodes to communicate only a finite number of parameters coming from, say, Gaussian mixture models or particle filters.

- Future work should consider nonlinear observations of the parameter θ, that is, $X_k^i \sim \mathcal{N}(g^i(\theta), (\sigma^i)^2)$ for some function $g : \theta \to \mathbb{R}$. Ongoing work develops similar parameter estimation approaches for the larger case of the exponential family of distributions on the natural parameter space. A particularly interesting case is when the parameter θ^* is changing with time, either arbitrarily, on some form of Markov process or other dependencies. This case renders observations to be neither identically distributed nor independent.

- In Chapter 4, we derived concentration results for an infinite number of hypotheses, particularly, for parametric models where $\Theta \subset \mathbb{R}^d$ is a compact set. In order to simplify the analysis, it is assumed that the networks where agents are interacting are undirected and fixed. Nonetheless, as seen in Chapter 3 for the case of finite hypotheses, we can provide algorithms for time-varying directed graphs in Eq. (3.24) and acceleration on fixed graphs in Eq. (3.18).

  We conjecture that similar algorithms can be derived for the infinite hypotheses case. For example, for time-varying directed graphs we propose an algorithm of the following form:

$$y_{k+1}^i = \sum_{j \in N_k^i} \frac{y_k^j}{d_k^j} \tag{6.1a}$$

$$d\mu_{k+1}^i(\theta) \propto \left(\prod_{j \in N_k^i} d\mu_k^j(\theta)^{\frac{y_k^j}{d_k^j}} p_\theta^i\left(x_{k+1}^i\right) \right)^{\frac{1}{y_{k+1}^i}}. \tag{6.1b}$$

  Similarly, for fixed graphs we propose that an algorithm can be derived with a linear

scalability with respect to the number of nodes, explicitly

$$d\mu_{k+1}^i(\theta) \propto \frac{\prod_{j=1}^{n} d\mu_k^j(\theta)^{(1+\sigma)\bar{A}_{ij}} p_\theta^i(x_{k+1}^i)^{\beta_k^i}}{\prod_{j=1}^{n} \left(d\mu_{k-1}^j(\theta) p_\theta^j(x_k^j)^{\beta_{k-1}^j} \right)^{\sigma \bar{A}_{ij}}}. \tag{6.2}$$

In turn, these two new protocols generate a set of new algorithms that can be made explicit for members of the exponential family.

- In Chapter 4 we develop an algorithm for the case where the observations as well as the parametric model are Gaussian distributions. It assumes that the observations are of the form: $X_k^i = \theta^i + \epsilon_k^i$ with $\epsilon_k^i \sim \mathcal{N}(0, (\sigma^i)^2)$. Therefore, $P^i = \mathcal{N}(\theta^*, \sigma)$ and $P_\theta^i = \mathcal{N}(\theta, \sigma)$.

  The problem is the estimation of a parameter $\theta^* \in \theta \subseteq \mathbb{R}$ that solves

$$\min_{\theta \in \Theta} F(\theta) \triangleq \sum_{i=1}^{n} \frac{(\theta - \theta^i)^2}{2(\sigma^i)^2}.$$

  If instead the parameter space is $\Theta \subseteq \mathbb{R}^m$, then $X_k^i = \theta^i + \epsilon_k^i$ with $\epsilon_k^i \sim \mathcal{N}(0, (\Sigma^i)^2)$, where Σ^i is now a covariance matrix in $\mathbb{R}^{m \times m}$. Therefore, $P^i = \mathcal{N}(\theta^*, \Sigma^i)$ and $P_\theta^i = \mathcal{N}(\theta, \Sigma^i)$, which implies the estimation problem consists in solving

$$\min_{\theta \in \Theta} F(\theta) \triangleq \sum_{i=1}^{n} (\theta - \theta^i)'(\Sigma^i)^{-1}(\theta - \theta^i)$$

  since for two multivariate Gaussian distributions $P = \mathcal{N}(\theta_0, \Sigma_0)$ and $Q = \mathcal{N}(\theta_1, \Sigma_1)$

$$D_{KL}(P, Q) = \frac{1}{2} \left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) - (\theta_1 - \theta_0)'\Sigma_1^{-1}(\theta_1 - \theta_0) - k + \ln \left(\frac{\det \Sigma_1}{\det \Sigma_1} \right) \right).$$

  Now, assume that the observations are in the form of $X_k^i = C^{i\prime}\theta + \epsilon_k^i$, where $\theta \in \mathbb{R}^m$, $C^i \in \mathbb{R}^m$ and $\epsilon_k^i \sim \mathcal{N}(0, \Sigma)$. Moreover, we can stack all C^i into a single matrix C, then $P^i = \mathcal{N}(C^{i\prime}\theta^*, \Sigma)$ and $P_\theta^i = \mathcal{N}(C^{i\prime}\theta, \Sigma)$. The optimization problem to be solved is then

$$\min_{\theta} \|\theta - \theta^*\|^2_{C\Sigma^{-1}C'} \tag{6.3}$$

and the resulting algorithm is

$$(\Sigma_{k+1}^i)^{-1} = \sum_{j=1}^n a_{ij}(\Sigma_k^j)^{-1} + C^i(\Sigma^i)^{-1}C^{i\prime} \tag{6.4a}$$

$$\theta_{k+1}^i = \Sigma_{k+1}^i \left(\sum_{j=1}^n a_{ij}(\Sigma_k^j)^{-1}\theta_k^j + C^{i\prime}(\Sigma^i)^{-1}x_{k+1}^i \right). \tag{6.4b}$$

It is clear from Eq. (6.3) that there is an immediate connection between the Kullback-Leibler minimization problem and the least squares or linear regression problem. Several questions can be asked about this setup:

  - How does the rates of convergence on Eq. (6.4) compare with other distributed optimization approaches for the solution of least squares problems?
  - The algorithm in Eq. (6.4) requires each agent i to transmit θ_k^i and also Σ_k^i, which can be communication intensive since the communication of a square matrix is needed. Can we have similar behavior by only transmitting certain entries? What is a good approach to select which entries to send? How do the rates of convergence get affected?
  - What happens if the observation matrices C^i are changing with time?
  - Can we provide a rate of convergence if the observations are nonlinear, i.e. $X_k^i = f^i(\theta) + \epsilon_k^i$?

- The optimal algorithms proposed in Chapter 5 require a certain amount of information regarding the spectral properties of the graph and the convexity and smoothness parameters of the functionals. The case where spectral information of the network is not available requires further study, for example, by using restarting techniques. Additionally, it is still an open problem whether one can show the optimal performance of distributed algorithms for time-varying graphs or directed graphs.

- The use of mathematical programming tools for large-scale optimal transport problems has taken on importance in recent years. Particularly for the problem of the computation of Wasserstein barycenters, one can explore the use of incremental curvature-aided information to get better performance. Also, recently proposed stochastic approaches can provide more efficient algorithms [239, 227].

- One can try to generalize the results of Chapter 5 to intermediate levels of smoothness. That is, try to propose the method for arbitrary Hölder parameter $\nu \in [0, 1]$. For

example, one can use universal Nesterov's method by skipping the adaptation and proper choosing of $\delta(\nu, \epsilon)$. This is another way to obtain results in the non-smooth case as a special situation $\nu = 0$. In the dual space, we will not have classical strong convexity but just uniform convexity. However, it can be studied by introducing inexact oracle as in [240].

- One can further extend the results in Chapter 5 to obtain the same rates of convergence when the graphs change with time by using restarting techniques [241, 242]. Nevertheless, we require additional assumptions. Particularly, the network changes should not happen often, and nodes must be able to detect when these changes occur. The condition number of the sequence of graphs $\chi(W_k)$ then is the worst one among all the graphs in the execution of the algorithm. Additionally, it is still an open research question whether these optimal convergence rates can be achieved over *directed* networks.

REFERENCES

[1] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games and Economic Behavior*, vol. 76, no. 1, pp. 210–225, 2012.

[2] K. R. Rad and A. Tahbaz-Salehi, "Distributed parameter estimation in networks," in *49th IEEE Conference on Decision and Control (CDC)*, Dec 2010, pp. 5050–5055.

[3] M. Alanyali, S. Venkatesh, O. Savas, and S. Aeron, "Distributed Bayesian hypothesis testing in sensor networks," in *Proceedings of the American Control Conference*, 2004, pp. 5369–5374.

[4] 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*. Springer, 2006, pp. 169–182.

[5] R. J. Aumann, "Agreeing to disagree," *The Annals of Statistics*, vol. 4, no. 6, pp. 1236–1239, 1976.

[6] V. Borkar and P. P. Varaiya, "Asymptotic agreement in distributed estimation," *IEEE Transactions on Automatic Control*, vol. 27, no. 3, pp. 650–655, 1982.

[7] J. N. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," *IEEE Transactions on Automatic Control*, vol. 29, no. 1, pp. 42–50, 1984.

[8] C. Genest and J. V. Zidek, "Combining probability distributions: A critique and an annotated bibliography," *Statistical Science*, vol. 1, no. 1, pp. 114–135, 1986. [Online]. Available: http://www.jstor.org/stable/2245510

[9] R. Cooke, "Statistics in expert resolution: A theory of weights for combining expert opinion," in *Statistics in Science*, ser. Boston Studies in the Philosophy of Science, R. Cooke and D. Costantini, Eds. Springer Netherlands, 1990, vol. 122, pp. 41–72.

[10] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[11] G. L. Gilardoni and M. K. Clayton, "On reaching a consensus using Degroot's iterative pooling," *The Annals of Statistics*, vol. 21, no. 1, pp. 391–401, 1993.

[12] J. A. Gubner, "Distributed estimation and quantization," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1456–1459, 1993.

[13] Y. Zhu, E. Song, J. Zhou, and Z. You, "Optimal dimensionality reduction of sensor data in multisensor estimation fusion," *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1631–1639, 2005.

[14] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors i. fundamentals," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, 1997.

[15] S.-L. Sun and Z.-L. Deng, "Multi-sensor optimal information fusion Kalman filter," *Automatica*, vol. 40, no. 6, pp. 1017–1023, 2004.

[16] L. Xiao and S. Boyd, "Optimal scaling of a gradient method for distributed resource allocation," *Journal of Optimization Theory and Applications*, vol. 129, no. 3, pp. 469–488, 2006.

[17] M. Rabbat and R. Nowak, "Decentralized source localization and tracking wireless sensor networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 2004, pp. 921–924.

[18] J. Konečnỳ, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *arXiv preprint arXiv:1511.03575*, 2015.

[19] T. Kraska, A. Talwalkar, J. C. Duchi, R. Griffith, M. J. Franklin, and M. I. Jordan, "Mlbase: A distributed machine-learning system." in *CIDR*, vol. 1, 2013, pp. 2–1.

[20] A. Nedić, A. Olshevsky, and C. A. Uribe, "Distributed learning for cooperative inference," *arXiv preprint arXiv:1704.02718*, 2017.

[21] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[22] 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*, vol. 3, no. 1, pp. 1–122, 2011.

[23] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems. corr, abs/1603.04467," in *Conference on Language Resources and Evaluation (LREC08)*, 2016, pp. 3243–3249.

[24] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[25] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-Bayesian learning," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5538–5553, Nov 2017.

[26] B. Golub and M. O. Jackson, "Naive learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, pp. 112–149, 2010.

[27] P. E. Converse and D. E. Apter, *Ideology and its Discontents.* Free Press, 1964.

[28] S. Feldman, "Structure and consistency in public opinion: The role of core beliefs and values," *American Journal of Political Science*, pp. 416–440, 1988.

[29] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *The Review of Economic Studies*, vol. 78, no. 4, pp. 1201–1236, 2011.

[30] M. O. Jackson, *Social and Economic Networks.* Princeton University Press, 2010.

[31] R. Hegselmann and U. Krause, "Opinion dynamics driven by various ways of averaging," *Computational Economics*, vol. 25, no. 4, pp. 381–405, 2005.

[32] A. Mirtabatabaei and F. Bullo, "Opinion dynamics in heterogeneous networks: convergence conjectures and theorems," *SIAM Journal on Control and Optimization*, vol. 50, no. 5, pp. 2763–2785, 2012.

[33] N. E. Friedkin, "The problem of social control and coordination of complex systems in sociology: A look at the community cleavage problem," *IEEE Control Systems*, vol. 35, no. 3, pp. 40–51, 2015.

[34] D. E. Cartwright, *Studies in Social Power.* Univer. Michigan, 1959.

[35] N. E. Friedkin and E. C. Johnsen, *Social Influence Network Theory: A Sociological Examination of Small Group Dynamics.* Cambridge University Press, 2011, vol. 33.

[36] R. P. Abelson, "Mathematical models of the distribution of attitudes under controversy," *Contributions to Mathematical Psychology*, vol. 14, pp. 1–160, 1964.

[37] N. E. Friedkin, A. V. Proskurnikov, R. Tempo, and S. E. Parsegov, "Network science on belief system dynamics under logic constraints," *Science*, vol. 354, no. 6310, pp. 321–326, 2016.

[38] S. E. Parsegov, A. V. Proskurnikov, R. Tempo, and N. E. Friedkin, "Novel multidimensional models of opinion dynamics in social networks," *IEEE Transactions on Automatic Control*, 2016.

[39] C. T. Butts, "Why I know but don't believe," *Science*, vol. 354, no. 6310, pp. 286–287, 2016.

[40] F. Amblard and G. Deffuant, "The role of network topology on extremism propagation with the relative agreement opinion dynamics," *Physica A: Statistical Mechanics and its Applications*, vol. 343, pp. 725–738, 2004.

[41] S. Fortunato, "Damage spreading and opinion dynamics on scale-free networks," *Physica A: Statistical Mechanics and its Applications*, vol. 348, pp. 683–690, 2005.

[42] S. van der Linden, "Determinants and measurement of climate change risk perception, worry, and concern," in *The Oxford Encyclopedia of Climate Change Communication*, M. Nisbet, M. Schafer, E. Markowitz, S. Ho, S. O'Neill, and J. Thaker, Eds. Oxford University Press, Oxford, UK, 2017.

[43] S. Shahrampour and A. Jadbabaie, "Exponentially fast parameter estimation in networks using distributed dual averaging," in *52nd IEEE Conference on Decision and Control (CDC)*, Dec 2013, pp. 6196–6201.

[44] S. Shahrampour, M. Rahimian, and A. Jadbabaie, "Switching to learn," in *Proceedings of the American Control Conference*, 2015, pp. 2918–2923.

[45] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed detection: Finite-time analysis and impact of network topology," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3256–3268, Nov 2016.

[46] A. Lalitha, A. Sarwate, and T. Javidi, "Social learning and distributed hypothesis testing," in *2014 IEEE International Symposium on Information Theory*, June 2014, pp. 551–555.

[47] M. A. Rahimian, S. Shahrampour, and A. Jadbabaie, "Learning without recall by random walks on directed graphs," *preprint arXiv:1509.04332*, 2015.

[48] D. Gale and S. Kariv, "Bayesian learning in social networks," *Games and Economic Behavior*, vol. 45, no. 2, pp. 329–346, 2003.

[49] E. Mossel, N. Olsman, and O. Tamuz, "Efficient Bayesian learning in social networks with Gaussian estimators," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2016, pp. 425–432.

[50] M. Mueller-Frank, "A general framework for rational learning in social networks," *Theoretical Economics*, vol. 8, no. 1, pp. 1–40, 2013.

[51] L. G. Epstein, J. Noor, and A. Sandroni, "Non-Bayesian learning," *The BE Journal of Theoretical Economics*, vol. 10, no. 1, 2010, article 3.

[52] A. Jadbabaie, P. Molavi, and A. Tahbaz-Salehi, "Information heterogeneity and the speed of learning in social networks," *Columbia Business School Research Paper*, no. 13-28, 2013.

[53] V. Saligrama, M. Alanyali, and O. Savas, "Distributed detection in sensor networks with packet losses and finite capacity links," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4118–4132, 2006.

[54] R. Rahman, M. Alanyali, and V. Saligrama, "Distributed tracking in multihop sensor networks with communication delays," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4656–4668, 2007.

[55] S. Bandyopadhyay and S.-J. Chung, "Distributed estimation using Bayesian consensus filtering," in *Proceedings of the American Control Conference*, 2014, pp. 634–641.

[56] L. Qipeng, Z. Jiuhua, and W. Xiaofan, "Distributed detection via Bayesian updates and consensus," in *2015 34th Chinese Control Conference (CCC)*, July 2015, pp. 6992–6997.

[57] L. Qipeng, F. Aili, W. Lin, and W. Xiaofan, "Non-bayesian learning in social networks with time-varying weights," in *30th Chinese Control Conference (CCC)*, 2011, pp. 4768–4771.

[58] A. Nedić, A. Olshevsky, and C. A. Uribe, "Distributed learning with infinitely many hypotheses," in *55th IEEE Conference on Decision and Control (CDC)*, Dec 2016, pp. 6321–6326.

[59] H. Salami, B. Ying, and A. Sayed, "Social learning over weakly-connected graphs," *IEEE Transactions on Signal and Information Processing over Networks*, vol. PP, no. 99, pp. 1–1, 2017.

[60] A. Tsiligkaridis and T. Tsiligkaridis, "Distributed probabilistic bisection search using social learning," *arXiv preprint arXiv:1608.06007*, 2016.

[61] L. Su and N. H. Vaidya, "Asynchronous distributed hypothesis testing in the presence of crash failures," *arXiv preprint arXiv:1606.03418*, 2016.

[62] L. Su and N. H. Vaidya, "Defending non-Bayesian learning against adversarial attacks," *arXiv preprint arXiv:1606.08883*, 2016.

[63] L. Su and N. H. Vaidya, "Non-Bayesian learning in the presence of byzantine agents," in *International Symposium on Distributed Computing*. Springer, 2016, pp. 414–427.

[64] A. Nedić, A. Olshevsky, and C. A. Uribe, "Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs," in *Proceedings of the American Control Conference*, 2015, pp. 5884–5889.

[65] A. Nedić, A. Olshevsky, and C. A. Uribe, "Network independent rates in distributed learning," in *Proceedings of the American Control Conference*, 2016, pp. 1072–1077.

[66] A. Nedić, A. Olshevsky, and C. A. Uribe, "Distributed Gaussian learning over time-varying directed graphs," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov 2016, pp. 1710–1714.

[67] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Distributed detection and estimation in wireless sensor networks," in *Academic Press Library in Signal Processing: Communications and Radar Signal Processing*, ser. Academic Press Library in Signal Processing, R. C. Nicholas D. Sidiropoulos, Fulvio Gini and S. Theodoridis, Eds. Elsevier, 2014, vol. 2, pp. 329 – 408.

[68] A. Nedić, A. Olshevsky, and C. A. Uribe, "A tutorial on distributed (non-bayesian) learning: Problem, algorithms and results," in *55th IEEE Conference on Decision and Control (CDC)*, Dec 2016, pp. 6795–6801.

[69] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2506–2517, 2009.

[70] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[71] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optimization Theory and Applications*, vol. 147, no. 3, pp. 516–545, 2010.

[72] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2015.

[73] A. Nedić, A. Olshevsky, and W. Shi, "Improved convergence rates for distributed resource allocation," *arXiv preprint arXiv:1706.05441*, 2017.

[74] A. Sundararajan, B. Hu, and L. Lessard, "Robust convergence analysis of distributed optimization algorithms," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2017, pp. 1206–1212.

[75] D. Jakovetic, "A unification, generalization, and acceleration of exact distributed first order methods," *arXiv preprint arXiv:1709.01317*, 2017.

[76] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[77] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, 2017.

[78] A. Nedić, A. Olshevsky, W. Shi, and C. A. Uribe, "Geometrically convergent distributed optimization with uncoordinated step-sizes," in *American Control Conference (ACC), 2017*. IEEE, 2017, pp. 3950–3955.

[79] K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," in *International Conference on Machine Learning*, 2017, pp. 3027–3036.

[80] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[81] L. V. Kantorovich, "On the translocation of masses," in *Dokl. Akad. Nauk. USSR (NS)*, vol. 37, 1942, pp. 199–201.

[82] B. Lévy and E. L. Schwindt, "Notions of optimal transport theory and how to implement them on a computer," *Computers & Graphics*, 2018.

[83] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a wasserstein loss," in *Advances in Neural Information Processing Systems*, 2015, pp. 2053–2061.

[84] J. Rabin, G. Peyré, J. Delon, and M. Bernot, "Wasserstein barycenter and its application to texture mixing," in *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 2011, pp. 435–446.

[85] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas, "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, p. 66, 2015.

[86] S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson, "WASP: Scalable Bayes via barycenters of subset posteriors," in *Artificial Intelligence and Statistics*, 2015, pp. 912–920.

[87] A. N. Bishop and A. Doucet, "Distributed nonlinear consensus in the space of probability measures," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 8662–8668, 2014.

[88] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, "Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm," *arXiv:1802.04367*, 2018.

[89] M. Blondel, V. Seguy, and A. Rolet, "Smooth and sparse optimal transport," *arXiv:1710.06276*, 2017.

[90] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel, "Large-scale optimal transport and mapping estimation," *arXiv:1711.02283*, 2017.

[91] G. Aude, M. Cuturi, G. Peyré, and F. Bach, "Stochastic optimization for large-scale optimal transport," *arXiv:1605.08527*, 2016.

[92] C. Villani, *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

[93] J. Solomon, "Computational optimal transport," *Mathematisches Forschungsinstitut Oberwolfach*, no. 8, 2017.

[94] G. Peyré and M. Cuturi, "Computational optimal transport," *arXiv:1803.00567*, 2018.

[95] B. Touri, *Product of Random Stochastic Matrices and Distributed Averaging*. Springer Science & Business Media, 2012.

[96] A. Olshevsky, "Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control," *preprint arXiv:1411.4186*, 2014.

[97] J. S. Rosenthal, "Convergence rates for markov chains," *SIAM Review*, vol. 37, no. 3, pp. 387–405, 1995.

[98] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*. American Mathematical Soc., 2009.

[99] P. Diaconis and D. Stroock, "Geometric bounds for eigenvalues of Markov chains," *The Annals of Applied Probability*, pp. 36–61, 1991.

[100] S. Ikeda, I. Kubo, and M. Yamashita, "The hitting and cover times of random walks on finite graphs using local degree information," *Theoretical Computer Science*, vol. 410, no. 1, pp. 94–100, 2009.

[101] R. Kannan, L. Lovász, and R. Montenegro, "Blocking conductance and mixing in random walks," *Combinatorics, Probability and Computing*, vol. 15, no. 4, pp. 541–570, 2006.

[102] J. Fulman, "Mixing time for a random walk on rooted trees," *Electronic Journal of Combinatorics*, vol. 16, 2009.

[103] A. Beveridge and J. Youngblood, "The best mixing time for random walks on trees," *Graphs and Combinatorics*, vol. 32, no. 6, pp. 2211–2239, 2016.

[104] L. Lovasz, "Random walks on graphs: A survey," *Combinatorics, Paul Erdos is Eighty*, vol. 2, 1993.

[105] C. Avin and G. Ercal, "Bounds on the mixing time and partial cover of ad-hoc and sensor networks." in *EWSN*, 2005, pp. 1–12.

[106] A. K. Chandra, P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari, "The electrical resistance of a graph captures its commute and cover times," *Computational Complexity*, vol. 6, no. 4, pp. 312–340, 1996.

[107] N. Berestycki, *Lectures on Mixing Times*. Cambridge University, 2014.

[108] J. Komjáthy, J. Miller, Y. Peres et al., "Uniform mixing time for random walk on lamplighter graphs," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 50, no. 4. Institut Henri Poincaré, 2014, pp. 1140–1160.

[109] O. Denysyuk and L. Rodrigues, "Random walks on evolving graphs with recurring topologies," in *International Symposium on Distributed Computing*. Springer, 2014, pp. 333–345.

[110] R. Montenegro, "The simple random walk and max-degree walk on a directed graph," *Random Structures & Algorithms*, vol. 34, no. 3, pp. 395–407, 2009.

[111] L. Boczkowski, Y. Peres, and P. Sousi, "Sensitivity of mixing times in Eulerian digraphs," *arXiv preprint arXiv:1603.05639*, 2016.

[112] D. Aldous and J. Fill, "Reversible Markov chains and random walks on graphs," 2002.

[113] S. P. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Mixing times for random walks on geometric random graphs." in *ALENEX/ANALCO*, 2005, pp. 240–249.

[114] C. Avin and G. Ercal, "On the cover time and mixing time of random geometric graphs," *Theoretical Computer Science*, vol. 380, no. 1-2, pp. 2–22, 2007.

[115] I. Benjamini, G. Kozma, and N. Wormald, "The mixing time of the giant component of a random graph," *Random Structures & Algorithms*, vol. 45, no. 3, pp. 383–407, 2014.

[116] N. Fountoulakis and B. Reed, "The evolution of the mixing rate," *arXiv preprint math/0701474*, 2007.

[117] J. Ding, J. H. Kim, E. Lubetzky, and Y. Peres, "Anatomy of a young giant component in the random graph," *Random Structures & Algorithms*, vol. 39, no. 2, pp. 139–178, 2011.

[118] J. Ding, E. Lubetzky, Y. Peres et al., "Mixing time of near-critical random graphs," *The Annals of Probability*, vol. 40, no. 3, pp. 979–1008, 2012.

[119] A. Nachmias and Y. Peres, "Critical random graphs: diameter and mixing time," *The Annals of Probability*, pp. 1267–1286, 2008.

[120] L. Addario-Berry and T. Lei, "The mixing time of the Newman-Watts small-world model," *Advances in Applied Probability*, vol. 47, no. 1, pp. 37–56, 2015.

[121] R. Durrett, *Random Graph Dynamics*. Cambridge University Press, UK, 2007.

[122] S. Bhamidi, G. Bresler, and A. Sly, "Mixing time of exponential random graphs," in *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*. IEEE, 2008, pp. 803–812.

[123] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Athena Scientific, 2003.

[124] C. McDiarmid, "On the method of bounded differences," *Surveys in combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[125] P. M. Weichsel, "The Kronecker product of graphs," *Proceedings of the American Mathematical Society*, vol. 13, no. 1, pp. 47–52, 1962.

[126] A. V. Proskurnikov and R. Tempo, "A tutorial on modeling and analysis of dynamic social networks. part i," *Annual Reviews in Control*, vol. 43, no. Supplement C, pp. 65 – 79, 2017.

[127] R. Tarjan, "Depth-first search and linear graph algorithms," *SIAM Journal on Computing*, vol. 1, no. 2, pp. 146–160, 1972.

[128] M. H. McAndrew, "On the product of directed graphs," *Proceedings of the American Mathematical Society*, vol. 14, no. 4, pp. 600–606, 1963.

[129] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Third Edition*, 3rd ed. The MIT Press, 2009.

[130] T. Lindvall, *Lectures on the Coupling Method*. Courier Corporation, 2002.

[131] A. Olshevsky and J. N. Tsitsiklis, "Degree fluctuations and the convergence time of consensus algorithms," in *Proc. 50th IEEE Conf. Decision and Control and European Control Conf*, Dec. 2011, pp. 6602–6607.

[132] P. Lancaster and H. Farahat, "Norms on direct sums and tensor products," *mathematics of computation*, vol. 26, no. 118, pp. 401–414, 1972.

[133] B. Gerencsér, "Markov chain mixing time on cycles," *Stochastic Processes and Their Applications*, vol. 121, no. 11, pp. 2553–2570, 2011.

[134] A. Tahbaz-Salehi and A. Jadbabaie, "Small world phenomenon, rapidly mixing Markov chains, and average consensus algorithms," in *46th IEEE Conference on Decision and Control*. IEEE, 2007, pp. 276–281.

[135] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 1999.

[136] A.-L. Barabasi, *Linked: How Everything is Connected to Everything Else and What It Means*. Plume, 2003.

[137] N. Ganguly, A. Deutsch, and A. Mukherjee, *Dynamics On and Of Complex Networks: Applications to Biology, Computer Science, and the Social Sciences*. Springer, 2009.

[138] S. Bornholdt and H. G. Schuster, *Handbook of Graphs and Networks: from the Genome to the Internet*. John Wiley & Sons, 2006.

[139] M. Penrose, *Random Geometric Graphs*. Oxford University Press, 2003.

[140] P. Erdos and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[141] M. E. Newman and D. J. Watts, "Renormalization group analysis of the small-world network model," *Physics Letters A*, vol. 263, no. 4, pp. 341–346, 1999.

[142] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, June 2014.

[143] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proceedings of the SIGCHI Conference on Human fFactors in Computing Systems*. ACM, 2010, pp. 1361–1370.

[144] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 2, 2007.

[145] J. Leskovec and J. J. McAuley, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 539–547.

[146] A. Mohaisen, A. Yun, and Y. Kim, "Measuring the mixing time of social graphs," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*. ACM, 2010, pp. 383–389.

[147] S. Kirkland, "Fastest expected time to mixing for a Markov chain on a directed graph," *Linear Algebra and Its Applications*, vol. 433, no. 11-12, pp. 1988–1996, 2010.

[148] S. Ghosal, "A review of consistency and convergence of posterior distribution," in *Varanashi Symposium in Bayesian Inference, Banaras Hindu University*, 1997.

[149] L. Schwartz, "On Bayes procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 4, no. 1, pp. 10–26, 1965.

[150] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart, "Convergence rates of posterior distributions," *Annals of Statistics*, pp. 500–531, 2000.

[151] L. Birgé, "About the non-asymptotic behaviour of Bayes estimators," *Journal of Statistical Planning and Inference*, vol. 166, pp. 67–77, 2015.

[152] S. Ghosal, A. Van Der Vaart et al., "Convergence rates of posterior distributions for noniid observations," *The Annals of Statistics*, vol. 35, no. 1, pp. 192–223, 2007.

[153] V. Rivoirard, J. Rousseau et al., "Posterior concentration rates for infinite dimensional exponential families," *Bayesian Analysis*, vol. 7, no. 2, pp. 311–334, 2012.

[154] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[155] A. Nedić and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 84–107, 2014.

[156] B. Dai, N. He, H. Dai, and L. Song, "Scalable Bayesian inference via particle mirror descent," *preprint arXiv:1506.03101*, 2015.

[157] M. Rabbat, "Multi-agent mirror descent for decentralized stochastic optimization," in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec 2015, pp. 517–520.

[158] A. Zellner, "Optimal information processing and Bayes's theorem," *The American Statistician*, vol. 42, no. 4, pp. 278–280, 1988.

[159] S. G. Walker, "Bayesian inference via a minimization rule," *Sankhyā: The Indian Journal of Statistics*, vol. 68, no. 4, pp. 542–553, 2006.

[160] T. P. Hill and M. Dall'Aglio, "Bayesian posteriors without Bayes' theorem," *preprint arXiv:1203.0251*, 2012.

[161] A. Juditsky, P. Rigollet, A. B. Tsybakov et al., "Learning by mirror averaging," *The Annals of Statistics*, vol. 36, no. 5, pp. 2183–2206, 2008.

[162] G. Lan, A. Nemirovski, and A. Shapiro, "Validation analysis of mirror descent stochastic approximation method," *Mathematical Programming*, vol. 134, no. 2, pp. 425–458, 2012.

[163] J. Li, G. Li, Z. Wu, and C. Wu, "Stochastic mirror descent method for distributed multi-agent optimization," *Optimization Letters*, pp. 1–19, 2016.

[164] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[165] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *Proceedings of the IEEE Symposium on Foundations of Computer Science*, 2003, pp. 482–491.

[166] F. Bénézit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, "Weighted gossip: Distributed averaging using non-doubly stochastic matrices," in *Proceedings of the IEEE International Symposium on Information Theory*, 2010, pp. 1753–1757.

[167] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Push-sum distributed dual averaging for convex optimization," in *IEEE 51st IEEE Conference on Decision and Control (CDC)*, Dec 2012, pp. 5453–5458.

[168] F. Iutzeler, P. Ciblat, and W. Hachem, "Analysis of sum-weight-like algorithms for averaging in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2802–2814, 2013.

[169] M. G. Rabbat and K. I. Tsianos, "Asynchronous decentralized optimization in heterogeneous systems," in *53rd IEEE Conference on Decision and Control (CDC)*, Dec 2014, pp. 1125–1130.

[170] B. Gerencsér and J. M. Hendrickx, "Push sum with transmission failures," *preprint arXiv:1504.08193*, 2015.

[171] A. G. Jayram, A. Garg, T. S. Jayram, S. Vaithyanathan, and H. Zhu, "Generalized opinion pooling," in *In Proceedings of the 8th Intl. Symp. on Artificial Intelligence and Mathematics*, 2004, pp. 79–86.

[172] A. Madansky, "Externally Bayesian groups," The Rand Corporation, Santa Monica, CA, memorandum no. RM-4141-PR, 1964.

[173] M. Rabbat, R. Nowak, and J. Bucklew, "Robust decentralized source localization via averaging," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, March 2005, pp. v/1057–v/1060 Vol. 5.

[174] X. Wang, M. Fu, and H. Zhang, "Target tracking in wireless sensor networks based on the combination of kf and mle using distance measurements," *IEEE Transactions on Mobile Computing*, vol. 11, no. 4, pp. 567–576, 2012.

[175] G. Mao, B. Fidan, and B. D. Anderson, "Wireless sensor network localization techniques," *Computer Networks*, vol. 51, no. 10, pp. 2529–2553, 2007.

[176] K. Langendoen and N. Reijers, "Distributed localization in wireless sensor networks: a quantitative comparison," *Computer Networks*, vol. 43, no. 4, pp. 499–518, 2003.

[177] L. LeCam, *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1986.

[178] L. LeCam, "Convergence of estimates under dimensionality restrictions," *The Annals of Statistics*, pp. 38–53, 1973.

[179] C. W. Fox and S. J. Roberts, "A tutorial on variational Bayesian inference," *Artificial Intelligence Review*, vol. 38, no. 2, pp. 85–95, 2012.

[180] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*. University of London, United Kingdom, 2003.

[181] B. Dai, N. He, H. Dai, and L. Song, "Provable Bayesian inference via particle mirror descent," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016, pp. 985–994.

[182] B. O. Koopman, "On distributions admitting a sufficient statistic," *Transactions of the American Mathematical society*, vol. 39, no. 3, pp. 399–409, 1936.

[183] G. Darmois, "Sur les lois de probabilitéa estimation exhaustive," *CR Acad. Sci. Paris*, vol. 260, no. 1265, p. 85, 1935.

[184] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman & Hall/CRC Boca Raton, FL, USA, 2014, vol. 2.

[185] C. Wang and B. Chazelle, "Gaussian learning-without-recall in a dynamic social network," *arXiv preprint arXiv:1609.05990*, 2016.

[186] G. Biau, K. Bleakley, and B. Cadre, "The statistical performance of collaborative inference," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2200–2228, 2016.

[187] A. Anikin, A. Gasnikov, P. Dvurechensky, A. Tyurin, and A. Chernov, "Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints," *Computational Mathematics and Mathematical Physics*, vol. 57, no. 8, pp. 1262–1276, 2017.

[188] A. Chernov, P. Dvurechensky, and A. Gasnikov, *Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints*. Cham: Springer International Publishing, 2016, pp. 391–403.

[189] A. Gasnikov, S. Kabanikhin, A. Mohamed, and M. Shishlenin, "Convex optimization in hilbert space with applications to inverse problems," *arXiv preprint arXiv:1703.00267*, 2017.

[190] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$," in *Soviet Mathematics Doklady*, vol. 27, no. 2, 1983, pp. 372–376.

[191] A. Nemirovskii and Yudin, *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[192] G. Qu and N. Li, "Accelerated distributed Nesterov gradient descent," *arXiv preprint arXiv:1705.07176*, 2017.

[193] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, Mar. 2012.

[194] T. T. Doan and A. Olshevsky, "Distributed resource allocation on dynamic networks in quadratic time," *Systems & Control Letters*, vol. 99, pp. 57–63, 2017.

[195] H. Lakshmanan and D. P. De Farias, "Decentralized resource allocation in dynamic networks of agents," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 911–940, 2008.

[196] I. Necoara, "Random coordinate descent algorithms for multi-agent convex optimization over networks," *IEEE Transactions on Automatic Control*, vol. 58, no. 8, pp. 2001–2012, Aug. 2013.

[197] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *arXiv preprint arXiv:1701.03961*, 2017.

[198] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2013, vol. 87.

[199] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[200] G. Lan, Z. Lu, and R. D. C. Monteiro, "Primal-dual first-order methods with $o(1/\varepsilon)$ iteration-complexity for cone programming," *Mathematical Programming*, vol. 126, no. 1, pp. 1–29, Jan 2011.

[201] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

[202] S. Bubeck, "Convex optimization: Algorithms and complexity," *Found. Trends Mach. Learn.*, vol. 8, no. 3-4, pp. 231–357, Nov. 2015.

[203] M. Cuturi and G. Peyré, "A smoothed dual approach for variational Wasserstein problems," *SIAM Journal on Imaging Sciences*, vol. 9, no. 1, pp. 320–343, 2016.

[204] A. V. Gasnikov, E. Gasnikova, Y. E. Nesterov, and A. Chernov, "Efficient numerical methods for entropy-linear programming problems," *Computational Mathematics and Mathematical Physics*, vol. 56, no. 4, pp. 514–524, 2016.

[205] O. Devolder, F. Glineur, and Y. Nesterov, "Double smoothing technique for large-scale linearly constrained convex optimization," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 702–727, 2012.

[206] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

[207] B. N. Oreshkin, M. J. Coates, and M. G. Rabbat, "Optimization and analysis of distributed averaging with short node memory," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2850–2865, 2010.

[208] J. Liu, B. D. Anderson, M. Cao, and A. S. Morse, "Analysis of accelerated gossip algorithms," *Automatica*, vol. 49, no. 4, pp. 873–883, 2013.

[209] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, Aug 2013.

[210] B. O'Donoghue and E. Candès, "Adaptive restart for accelerated gradient schemes," *Foundations of Computational Mathematics*, vol. 15, no. 3, pp. 715–732, Jun 2015.

[211] A. Juditsky and Y. Nesterov, "Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization," *Stochastic Systems*, vol. 4, no. 1, pp. 44–80, 2014.

[212] Y. Nesterov, "Universal gradient methods for convex optimization problems," *Mathematical Programming*, vol. 152, no. 1-2, pp. 381–404, 2015.

[213] P. Dvurechensky, "Gradient method with inexact oracle for composite non-convex optimization," *arXiv preprint arXiv:1703.09180*, 2017.

[214] N. M. Nam, N. T. An, R. B. Rector, and J. Sun, "Nonsmooth algorithms and Nesterov's smoothing technique for generalized Fermat–Torricelli problems," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 1815–1839, 2014.

[215] G. Lan, "Gradient sliding for composite optimization," *Mathematical Programming*, vol. 159, no. 1-2, pp. 201–235, 2016.

[216] A. Anikin, P. Dvurechensky, A. Gasnikov, A. Golov, A. Gornov, Y. Maximov, M. Mendel, and V. Spokoiny, "Efficient numerical algorithms for regularized regression problem with applications to traffic matrix estimations," *arXiv preprint arXiv:1508.00858*, 2015.

[217] A. Juditsky and A. Nemirovski, "First order methods for nonsmooth convex large-scale optimization, I: General purpose methods," *Optimization for Machine Learning*, pp. 121–148, 2011.

[218] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[219] M. Cuturi and A. Doucet, "Fast computation of Wasserstein barycenters," in *International Conference on Machine Learning*, 2014, pp. 685–693.

[220] M. Beiglböck, P. Henry-Labordere, and F. Penkner, "Model-independent bounds for option pricesa mass transport approach," *Finance and Stochastics*, vol. 17, no. 3, pp. 477–501, 2013.

[221] G. Buttazzo, L. De Pascale, and P. Gori-Giorgi, "Optimal-transport formulation of electronic density-functional theory," *Physical Review A*, vol. 85, no. 6, p. 062502, 2012.

[222] Y. LeCun, "The MNIST database of handwritten digits," http://yann. lecun. com/exdb/mnist/, 1998.

[223] E. Anderes, S. Borgwardt, and J. Miller, "Discrete Wasserstein barycenters: optimal transport for discrete data," *Mathematical Methods of Operations Research*, vol. 84, no. 2, pp. 389–409, 2016.

[224] M. Cuturi and G. Peyré, "A smoothed dual approach for variational Wasserstein problems," *SIAM Journal on Imaging Sciences*, vol. 9, no. 1, pp. 320–343, 2016.

[225] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-Bayesian learning," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5538–5553, Nov. 2017.

[226] A. N. Bishop and A. Doucet, "Consensus in the Wasserstein metric space of probability measures," *arXiv:1404.0145*, 2014.

[227] M. Staib, S. Claici, J. M. Solomon, and S. Jegelka, "Parallel streaming Wasserstein barycenters," in *Advances in Neural Information Processing Systems*, 2017, pp. 2644–2655.

[228] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, 2013, pp. 2292–2300.

[229] Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson, "Much faster algorithms for matrix scaling," *arXiv preprint arXiv:1704.02315*, 2017.

[230] A. V. Gasnikov, E. B. Gasnikova, Y. E. Nesterov, and A. V. Chernov, "Efficient numerical methods for entropy-linear programming problems," *Computational Mathematics and Mathematical Physics*, vol. 56, no. 4, p. 514, Apr. 2016.

[231] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić, "Optimal algorithms for distributed optimization," *arXiv:1712.00232*, 2017.

[232] M. Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié," *Ann. Inst. H. Poincaré*, vol. 10, no. 3, pp. 215–310, 1948.

[233] J. Bigot, E. Cazelles, and N. Papadakis, "Regularization of barycenters in the wasserstein space," *arXiv:1606.01025*, 2016.

[234] Z. Allen-Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," *arXiv preprint arXiv:1407.1537*, 2014.

[235] P. Dvurechensky, A. Gasnikov, E. Gasnikova, S. Matsievsky, A. Rodomanov, and I. Usik, "Primal-dual method for searching equilibrium in hierarchical congestion population games," in *Supplementary Proceedings of the 9th International Conference on Discrete Optimization and Operations Research and Scientific School (DOOR 2016) Vladivostok, Russia, September 19 - 23, 2016*, 2016, arXiv:1606.08988. pp. 584–595.

[236] A. Chernov, P. Dvurechensky, and A. Gasnikov, *Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints*. Cham: Springer International Publishing, 2016, pp. 391–403, arXiv:1605.02970.

[237] S. Kakade, S. Shalev-Shwartz, and A. Tewari, "On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization," 2009, Unpublished Manuscript.

[238] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[239] S. Claici, E. Chien, and J. Solomon, "Stochastic Wasserstein Barycenters," *arXiv:1802.05757*, 2018.

[240] A. Gasnikov, P. Dvurechensky, and D. Kamzolov, "Gradient and gradient-free methods for stochastic convex optimization with inexact oracle," *arXiv preprint arXiv:1502.06259*, 2015.

[241] O. Fercoq and Z. Qu, "Restarting accelerated gradient methods with a rough strong convexity estimate," *arXiv preprint arXiv:1609.07358*, 2016.

[242] N. Bansal and A. Gupta, "Potential-function proofs for first-order methods," *arXiv preprint arXiv:1712.04581*, 2017.