UNIVERSITY OF CALIFORNIA SAN DIEGO

Advances in Multi-agent Decision Making Systems with Adaptive Algorithms

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

in

Electrical Engineering (Communication Theory and Systems)

by

Ashwin Verma

Committee in charge:

Professor Behrouz Touri, Chair
Professor Arya Mazumdar
Professor Piya Pal
Professor Yang Zheng

2024

The Dissertation of Ashwin Verma is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To everyone who has taught me in their own way—my family, friends, and teachers.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am deeply thankful to my advisor, Professor Behrouz Touri, for his invaluable guidance and support. His dedication to discussing research with me has been extraordinary, encouraging me to express my thoughts freely. Behrouz's patience, diligence, enthusiasm, optimism, and unwavering commitment have continually inspired me. I see him not only as a mentor but also as a guardian, as his caring and patient nature has given me a sense of security throughout this journey. His teachings have enriched my life far beyond the academic realm, and my Ph.D. has been made possible through his support in countless ways.

I have been privileged to collaborate with and receive guidance from Prof. Marcos Vasconcelos, Prof. Urbashi Mitra, and Prof. Soheil Mohajer during my Ph.D. journey. These collaborations have exposed me to different writing styles that are clear, precise, and rigorous, greatly influencing my own. My initial work with Prof. Mitra and Prof. Vasconcelos introduced me to the benefits of research collaborations, for which I am deeply grateful. Over the past three years, nearly weekly meetings with Prof. Mohajer have taught me a great deal through his quick thinking on technical aspects, inspiring my approach to problem-solving.

During my time at UC San Diego, I was fortunate to have great teachers and mentors. I am thankful to my Ph.D. committee members Prof. Arya Mazumdar, Prof. Piya Pal, and Prof. Yang Zheng for their willingness to dedicate their valuable time to serve on my committee, engage thoughtfully with my work, and offer invaluable feedback.

I would also like to thank my labmates Rohit Parasnis, Adel Aghajan for fostering an inclusive, supportive, and friendly environment. Our insightful discussions have been invaluable. I am fortunate to have been surrounded by an incredibly caring and supportive group of friends. Special thanks to my flatmates, (meme-lord) Agrim, (always-ok) Manideep, and (chef) Ish, who kept my life lively over the past three years, supported me during tough times and fed me always. I am grateful to Ayush for his advice, Roshan for the coffee, and Sourya and Saksham for being my PhD buddies from IITK. My gratitude extends to all my friends from IITK, regardless of how often we interact, especially Utkarsh, Anand, and Eeshan. Speaking of friends, I am privileged

to have known Rao-san (Srinivas Rao). He was one of the most eccentric and lovable individuals I've ever met, with a vast knowledge spanning various subjects. I will always cherish our diverse conversations, ranging from anime and tech gadget reviews to technical discussions. To all my friends—thank you for being an integral part of my journey and for creating countless cherished memories.

I extend my heartfelt thanks to Prof. Rakesh K. Bansal, my M. Tech. thesis advisor at IIT Kanpur, whose introduction to information theory ignited my passion for research and kindled my love for research. I am also sincerely grateful to all my teachers and mentors at IIT Kanpur, especially Prof. Banerjee and Prof. Potluri.

Saving the best for last, I am profoundly grateful to my parents, Nawab and Seema Verma, whose sacrifices and unwavering encouragement have made all my endeavors possible. Their support, even when they did not fully understand my choices, has enabled me to pursue the career path I desire. Their unconditional love has been one of my greatest sources of strength throughout my life.

Chapter 2, in full, is a reprint of the material as it appears in A. Verma, M. Vasconcelos, U. Mitra, B. Touri, "Maximal Dissent: a State-Dependent Way to Agree in Distributed Convex Optimization," in IEEE *Transactions on Control of Network Systems*. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in A. Verma, A. Sharbafchi, S. Mohajer, B. Touri,"Distributed Fact Checking:A Stochastic Approximation Approach," in preparation for *IEEE Transactions on Automatic Control*. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in A. Verma, A. Sharbafchi, B. Touri, S. Mohajer, "Distributed Fact Checking," in 2023 *International Symposium on Information Theory*. The dissertation author was the primary investigator and author of this paper.

Chapter 5 in full, is a reprint of the material as it appears in A. Verma, S. Mohajer, B. Touri, "Distributed Fact Checking:Learning Unreliability," in 2024 *American Control Conference*.

The dissertation author was the primary investigator and author of this paper.

Chapter 6 in full, is a reprint of the material as it appears in A. Verma, A. Sharbafchi, S. Mohajer, B. Touri,“Distributed Fact Checking:A Stochastic Approximation Approach," in preparation for *IEEE Transactions on Autamtic Control*. The dissertation author was the primary investigator and author of this paper.

Chapter 7 in full, is a reprint of the material as it appears in A. Verma, S. Mohajer, B. Touri, “Multi-Agent Fact-Checker: Adaptive Estimators," submitted in 2024 *Conference on Decision and Control*. The dissertation author was the primary investigator and author of this paper.

VITA

2018	Bachelor of Technology and Masters of Technology (Dual Degree) Electrical Engineering, Indian Institute of Technology, Kanpur
2018–2022	Master of Science, in Electrical Engineering (Communication Theory and Systems) University of California San Diego
2018–2024	Research Assistant, University of California San Diego
2024	Doctor of Philosophy, in Electrical Engineering (Communication Theory and Systems) University of California San Diego

PUBLICATIONS

A. Verma, A. Sharbafchi, S. Mohajer, B. Touri,“Distributed Fact Checking:A Stochastic Approximation Approach," in preparation for *IEEE Transactions on Automatic Control* , IEEE

A. Verma, B. Touri, “Almost-Sure Reachability" in preparation for *IEEE Control Systems Letter*, IEEE

A. Verma, S. Mohajer, B. Touri, “Multi-Agent Fact-Checker: Adaptive Estimators," submitted in 2024 *Conference on Decision and Control* (CDC), IEEE, 2024

A. Verma, S. Mohajer, B. Touri, “Distributed Fact Checking:Learning Unreliability," in 2024 *American Control Conference* (ACC), IEEE, 2024

A. Verma, M. Vasconcelos, U. Mitra, B. Touri, “Maximal Dissent: a State-Dependent Way to Agree in Distributed Convex Optimization," in IEEE *Transactions on Control of Network Systems*, IEEE, 2023

A. Verma, A. Sharbafchi, B. Touri, S. Mohajer, “Distributed Fact Checking," in 2023 *International Symposium on Information Theory* (ISIT), pp. 2649-2654, IEEE, 2023

R. Parasnis, **A. Verma**, M. Franceschetti, B. Touri, “A random adaptation perspective on distributed averaging," in IEEE *Control Systems Letters* (L-CSS), 7, pp.241-246, IEEE, 2022

A. Verma, M. Vasconcelos, U. Mitra, B. Touri, “Max-gossip subgradient method for distributed optimization," in 2021 60th IEEE *Conference on Decision and Control* (CDC), pp. 3130-3136, IEEE, 2021

A. Verma, R. K. Bansal, “Sequential change detection based on universal compression for Markov sources," In 2019 *IEEE International Symposium on Information Theory* (ISIT), pp. 2189-2193, IEEE, 2019

ABSTRACT OF THE DISSERTATION

Advances in Multi-agent Decision Making Systems with Adaptive Algorithms

by

Ashwin Verma

Doctor of Philosophy in Electrical Engineering (Communication Theory and Systems)

University of California San Diego, 2024

Professor Behrouz Touri, Chair

With the growing demand for computation and the increasing prevalence of resource-constrained agents, the importance of leveraging a network of agents to solve complex problems has become often more pronounced. A multi-agent system consists of interconnected agents with computing capabilities, working collaboratively towards a shared objective. Distributed computation using a multi-agent system provides benefits with regards to privacy, reduction of computational load and resources. In this dissertation, we study two problems that benefit from being solved with the help of a multi-agent system namely (i) distributed convex optimization

and (ii) distributed fact-checking.

In part I, we consider a set of agents collaboratively solving a distributed convex optimization problem, asynchronously, under stringent communication constraints. In such situations, when an agent is activated and is allowed to communicate with only one of its neighbors, we would like to pick the one holding the most informative local estimate. We propose new algorithms where the agents with maximal dissent average their estimates, leading to an information mixing mechanism that often displays faster convergence to an optimal solution compared to randomized gossip.

In Part II, we explore a distributed fact-checking system to detect fake news using inexpert agents. Each agent labels news as true or false based on its reliability, modeled as a Binary Symmetric Channel (BSC) with some error probability. We develop an algorithm that estimates statement validity by thresholding a linear combination of agents' labels and deriving optimal weights and thresholds to minimize error probability. Moreover, we present an adaptive algorithm to learn the agents' unreliability parameters and prove the convergence of the adaptive estimator. We also propose a broader class of adaptive estimators for the agents' unreliability parameters, providing the necessary conditions for convergence. We show that estimators for ensembles of two and three agents adhere to a consistent update rule, while hard-decoded estimates fail to converge for any number of agents.

This dissertation contributes to the theoretical aspects of distributed optimization and fact-checking in multi-agent systems, offering novel algorithms and insights for efficient and reliable distributed decision-making.

Chapter 1

Introduction

As the demand for computation grows and resource-constrained agents become more prevalent, leveraging a network of agents to solve complex problems has become increasingly important. A multi-agent system consists of interconnected agents with computing capabilities, working collaboratively towards a shared objective. These agents can be sensors, computing devices, or even humans, and they may possess different pieces of information or varying levels of capability to perform specific tasks. The interaction among agents can occur in a decentralized manner, or there may be a central entity that coordinates their outputs to make decisions for the system's overall objective.

This dissertation focuses on two critical multi-agent decision-making challenges:

1. **Distributed Convex Optimization**: This challenge involves a collective effort by agents to minimize the sum of local objective functions through information exchange over a communication network. Our goal is to develop algorithms that enable convergence to an optimal solution via local interactions facilitated by the network.

2. **Distributed Fact Checking**: In environments with multiple imperfect fact-checkers, determining the validity of a source based on their responses is a significant challenge. Understanding the reliability of these fact-checkers is crucial. We explore how to formulate and learn the reliability of various imperfect fact-checkers over time. Our proposed model for distributed fact-checking utilizes unreliable or imperfect agents to address this issue.

1.1 Maximal Dissent: Distributed Convex Optimization

In distributed convex optimization, a collection of agents collaborate to minimize the sum of local objective functions by exchanging information over a communication network. The primary goal is to design algorithms that converge to an optimal solution via local interactions dictated by the underlying communication network. A standard strategy to solve distributed optimization problems consists of each agent first combining the local estimates shared by its neighbors followed by a first-order subgradient method on its local objective function [55, 39, 40]. Of particular relevance herein are the so-called *gossip* algorithms [51], where the information mixing step consists of averaging the states of two agents connected by one of the edges selected from the network graph.

Two benefits of gossip algorithms are their simple asynchronous implementation and a reduction in communication costs. One common gossip algorithm is *randomized*, in which an agent is randomly activated and chooses one of its neighbors randomly to average its state [11, 47, 8]. The randomization mechanism used in this gossip scheme is usually *state-independent*. We consider a different approach to gossip in which the agent chooses one of its neighbors based on its state. At one extreme, we may think of agents who prefer to gossip with neighbors with similar *opinions*, for example, in an *echo-chamber* where agents may only talk to others if they reinforce their own opinions which hinders effective information mixing mechanism. On the opposite extreme, we consider agents who prefer to gossip with neighbors with maximal disagreement or dissent. In this dissertation, we focus on the concept of *max-dissent* gossip as a state-dependent information mixing mechanism for distributed optimization. We establish the convergence of the resulting distributed subgradient method under minimal assumptions on the underlying communication graph, and the local functions.

The idea of enabling a consensus protocol to use state-dependent matrices dates back to the Hegselmann and Krause [27] model for opinion dynamics. However, the literature on state-dependent averaging in distributed optimization is scarce and mostly motivated by

applications in which the state represents the physical location of mobile agents (e.g. robots, autonomous vehicles, drones, etc.). In such settings, the state-dependency arises from the fact that physically closer agents have a higher probability of successfully communicating with each other [34, 5, 4]. Unlike previous work, our model does not assume that the state of an agent necessarily represents its position in space. Moreover, we do not impose strong assumptions on the network's connectivity over time such as in [40] and [34].

Our work is closely related to state-dependent *averaging* schemes known as *Load-Balancing* [15] and *Greedy Gossip with Eavesdropping* [56]. The main idea in these methods is to accelerate averaging by utilizing the information from the most *informative* neighbor, i.e., the neighbors whose states are maximally different with respect to some norm from each agent. We refer to it as the *maximal dissent* heuristics. The challenges of convergence analysis for maximal dissent averaging are highlighted in [15, 37, 56]. However, concepts akin to max-dissent have only been explored for the specific problem of averaging [56]. Our work, on the other hand, focuses on distributed convex optimization, whose convergence is not guaranteed by the convergence of the averaging scheme alone.

As a broader impact of the results herein, we show a general result regarding schemes that possibly incorporate a mixing of information between max-dissent agents converging to a global optimizer of the underlying distributed optimization problem almost surely. Our result enables us to propose and extend the use of load-balancing, and max-dissent gossip to distributed optimization. The key property of max-dissent averaging is that it leads to a contraction of the Lyapunov function used to establish convergence. While recent work has considered similar contraction results (e.g. [31, 30]), they are not applicable to state-dependent schemes and do not establish almost sure convergence, but only convergence in expectation.

1.2 Distributed Fact Checking

Transitioning to the problem of distributed fact-checking, we confront the increasingly complex challenge of discerning the veracity of information disseminated across online platforms. As online social networks become increasingly effective in disseminating information, the task of distinguishing between true and false information becomes increasingly challenging. This growing efficiency of information dissemination has led to a number of studies on how misinformation spreads through networks [1, 2, 12, 43, 41]. Conversely, there is growing interest in the development of automated fact-checkers that can perform tasks such as document retrieval, evidence extraction, and claim validation in an automated manner [25, 26, 54].

When there are multiple imperfect *fact checkers*, determining the validity of a source based on their responses becomes a challenge. In such cases, it is important to know the reliability statistics of the fact checkers in question. As a result, a natural question arises: in the presence of multiple imperfect fact checkers, how can we formulate and learn their reliability over time? We provide a model for distributed fact-checking using unreliable or imperfect agents. A key step in our model is that we model each agent as a BSC channel where the cross-over probability of the channels models the unreliability of each agent. Given an estimate of the unreliability parameters, a weighted thresholding estimator can be used to identify the validity of the statement [52, 42, 59], where the weights are the log-odds based on the agents' unreliability estimates. We propose and study a learning rule to estimate the reliability parameters of the agents. Our algorithm provides the advantage of requiring a minimal memory and having a simplified update rule.

In our problem, we are working with a mixture of product distributions. Determining the parameters of an identifiable mixture, has been widely researched [20, 22, 23, 14, 19]. The parameter estimation problem typically involves finding a hypothetical model that produces samples with a distribution that closely resembles the true model.

1.2.1 Related Work

Given the unreliability parameters of the agents, an optimal approach to reconstructing unknown labels involves employing weighted majority voting. In this method, the weights assigned to the output provided by each agent are equal to the log-odds based on the knowledge of the workers' unreliability [52] [42]. In [59], we provide the characterization of weights which would result in the optimal estimator for labelling of validity of statements.

On the other hand the estimation of unreliability parameters is intertwined with literature on crowdsourcing labeling. In the realm of machine learning, significant attention has been devoted to the crowdsourcing of data labeling, where multiple workers are tasked with labeling data. This process is susceptible to errors arising from various factors such as task complexity, low incentive for accurate labeling, and the repetitive nature of tasks. Estimating the unreliability of workers is challenging since the true labels of the data are unknown.

The challenge of distributed fact-checking shares similarities with the extensively studied problem of crowdsourcing labeling tasks for classification of patients' diagnosis, notably explored within the Dawid-Skene model introduced through empirical studies by Dawid and Skene in 1979 [16]. Initially applied in the medical context, where multiple clinicians label a patient's state, Dawid and Skene proposed an Expectation-Maximization (EM) algorithm. Over the years, various extensions and variants of this algorithm have emerged [48, 6, 53, 28], with a notable line of work employing spectral analysis of matrices representing correlations between agents and labeling tasks [60]. Recent years have witnessed a growing body of research focused on performance guarantees for EM and its variants. Notably, Chao and Dengyong [21], as well as Zhang et al. [60], have provided performance guarantees for different versions of EM employing diverse initialization techniques. The convergence analysis of these variants of the Dawid-Skene estimator, rooted in the EM algorithm, has been explored for the offline scenario. In this context, where the sequence of statements to be verified is available as a batch, studies by Gao et al. [21] and Zhang et al., [60] have delved into the convergence aspects.

The analyses of the EM-based algorithms hinge on a sufficiently accurate initialization derived from the output of a substantial batch of statements being validated. Importantly, all these works assume access to the storage of all labels of all agents, given their focus on an offline setting. The only notable work presenting an algorithm in a streaming setting, without the necessity to store the entire dataset, is found in the work of Bonald and Combes [9]. Their proposed Triangular Estimation (TE) algorithm focuses on estimating the unreliability parameters of agents based on correlations between triplets of agents. This algorithm directly utilizes three agents, rather than the entire set, for estimating the unreliability parameter of a specific agent. The knowledge of all agents' output becomes indirectly relevant in determining which three agents to select for computing the unreliability parameter of a given agent. Our work is the first attempt at providing an online estimator that has similarities to the EM variants. In establishing convergence results for our online estimator we draw connections to stochastic approximation concepts within the literature of control theory [7].

1.3 Thesis Organization

The subsequent sections of the dissertation will be organized as follows.

(i) In Part I (Chapter 2), we consider a set of agents collaboratively solving a distributed convex optimization problem, asynchronously, under stringent communication constraints. In such situations, when an agent is activated and is allowed to communicate with only one of its neighbors, we would like to pick the one holding the most informative local estimate. We propose new algorithms where the agents with maximal dissent average their estimates, leading to an information mixing mechanism that often displays faster convergence to an optimal solution compared to randomized gossip. The core idea is that when two neighboring agents, whose local estimates have the greatest difference among all neighboring agents in the network, average their states, it results in the largest immediate reduction of the quadratic Lyapunov function. This reduction helps establish convergence

to the set of optimal solutions. As a broader contribution, we prove the convergence of max-dissent subgradient methods using a unified framework that can be used for other state-dependent distributed optimization algorithms. Our proof technique bypasses the need to establish the information flow between any two agents within a time interval of uniform length by intelligently studying the convergence properties of the Lyapunov function used in our analysis.

(ii) In Part II, we focus on the problem of distributed fact-checking. We introduce the problem formulation and novel estimators in Chapter 3. We formulate the problem of fake news detection using distributed inexpert agents. We consider the source for news/statements as a binary source (to model true vs. false statements). Upon observing news, each agent labels the news as true or false, which equals the validity of the statement with some probability depending on the agents' reliability. In other words, each agent is viewed as a Binary Symmetric Channel (BSC) that misclassifies each statement with some error probability.

(iii) In Chapter 4, we provide a solution to the problem where the agents' unreliability is known and we need an estimate for the validity of the statements. For an algorithm that estimates the validity by thresholding a linear combination of the individual agents' labels, we characterize the optimal weights and threshold to minimize the probability of error. We establish an upper bound on this probability of error and that of the naive majority rule.

(iv) In Chapter 5, we study the algorithm to learn the unreliability parameters. We focus on the two-agent case, we extensively analyze the discrete-time limit of our algorithm, and provide convergence results for the adaptive estimator.

(v) In Chapter 6, we analyze the variation of the adaptive algorithm to learn the unreliability parameters, introduced in Chapter 3. We extensively analyze the discrete-time limit of our algorithm proving the convergence of the variant adaptive algorithm to the equilibrium

points of the mean-field Ordinary Differential Equations (ODE).

(vi) In Chapter 7, we introduce a class of adaptive estimators for the unreliability parameters of the agents. For the class of estimators, we provide the necessary conditions for the adaptive estimator to converge to the true unreliability parameters. We show that the estimators for ensembles of two and three agents eventually adhere to a consistent (fixed) update rule. Furthermore, we also show that, surprisingly, the estimator for the unreliability parameters based on the hard-decoded estimate of the statement truths fails to converge to the true unreliability parameters for any number of agents.

1.4 Notations

Let $\mathbb{N}$ denote the set of all natural numbers, $\mathbb{N}_0$ denote $\mathbb{N} \cup \{0\}$, and for any $n \in \mathbb{N}$, define $[n] := \{1, 2, \ldots, n\}$. For any $i \in [n]$, we define $[n]_{-i} := [n] \setminus \{i\}$. We denote the set of real numbers by $\mathbb{R}$ and denote the n-dimensional Euclidean space by $\mathbb{R}^n$.

We use boldface letters, such as $\mathbf{x}$, to represent vectors and lower-case letters, such as x, to represent scalars. Upper-case letters, such as A, represent matrices. We use A^T to denote the transpose of a matrix A. For $i \in [n]$, we denote by $\boldsymbol{b}_i$ the i-th standard basis vector of $\mathbb{R}^n$. We denote by $\mathbf{1}$, the vector with all components equal to one, whose dimension will be clear from the context. For a vector $\boldsymbol{v}$, we denote the ℓ_2-norm by $\|\boldsymbol{v}\|$, and the average of its entries by $\bar{v}$. We say that an $n \times n$ matrix A is stochastic if it is non-negative and the elements in each of its rows add up to one. We say that A is doubly stochastic if both A and A^T are stochastic. For two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, we define $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a}^T \boldsymbol{b}$. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$ and a scalar $y \in \mathbb{R}$, we use $(\boldsymbol{x}, y)$ to denote $(x_1, x_2, \ldots, x_n, y) \in \mathbb{R}^{n+1}$.

The trace of a square matrix A is defined to be the sum of entries on the main diagonal of A and is denoted by $\mathrm{tr}(A)$. For matrices $A, B \in \mathbb{R}^{n \times m}$ we define $\langle A, B \rangle = \mathrm{tr}(A^T B)$ as the inner product and $\|A\|_F$ to denote the resulting norm, i.e., Frobenius norm of A.

For a set $\mathcal{S}$, we use $\mathcal{S}^{\mathsf{c}}$ to denote the complement of $\mathcal{S}$. Moreover, for $\boldsymbol{x} \in \mathbb{R}^n$ and $\mathcal{S} \subseteq [n]$ we define $\boldsymbol{x}(\mathcal{S}) := \sum_{i \in \mathcal{S}} x_i$. For a statement $\mathcal{A}$, we denote $\mathbb{1}_{\{\mathcal{A}\}}$ to be the indicator function for

$\mathcal{A}$, i.e., $\mathbb{1}_{\{\mathcal{A}\}} = 1$, if $\mathcal{A}$ holds true, and 0, otherwise. In Part II, we use capital letters, such as R, S, and X, to denote random variables.

Throughout each part of the dissertation, all random variables are defined with respect to an underlying probability space $(\Omega, \mathcal{F}, \Pr)$. In Part II, when the probability measure is defined through a parameter, say $\boldsymbol{x}$, we denote the probability measure by specifying $\boldsymbol{x}$ as $\Pr(\cdot; \boldsymbol{x})$. Throughout part II when the parameter $\boldsymbol{x}$ is not specified the probability measure is defined through the true parameter (described in the problem formulation in Chapter 3) $\boldsymbol{\pi}$.

In Part II, we abuse the bar notation, for a scalar $a \in [0, 1]$. We use $\bar{a}$ to denote $1 - a$. We use $\mathbf{1}, \mathbf{0}$ to denote all one and all zero vectors respectively. For a vector $\boldsymbol{x} \in \mathbb{R}^n$, x_i denotes its i-th element. For a sequence of entities such as $\{\boldsymbol{P}(t)\}$, we denote the entry at time t by $\boldsymbol{P}(t)$. However, for step-size specifically, we denote the step-size at time t by using subscripts such as η_t, ν_t. Throughout the dissertation, we use logarithm with respect to base e.

Part I

Distributed Convex Optimization

Chapter 2

Maximal Dissent

In this chapter, we study the problem of distributed convex optimization and a state-dependent class of algorithms to solve it. The main contributions of this chapter are:

- presenting state-dependent distributed optimization schemes that do not rely on or imply explicit strong connectivity conditions (such as B-connectivity).
- characterizing a general result highlighting the importance of max-dissent agents on a graph for distributed optimization, significantly simplifying the task of establishing contraction results for a large class of consensus-based subgradient methods.
- proving the convergence of state-dependent algorithms to a global optimizer for distributed optimization problems using a technique involving the aforementioned contraction property of a quadratic Lyapunov function.
- presenting numerical experiments that suggest that the proposed state-dependent subgradient methods improve the convergence rate of distributed estimation problems relative to conventional (state-independent) gossip algorithms.

The rest of the chapter is organized as follows. First, we formulate distributed optimization problems, and outline a generic state-dependent distributed subgradient method in Section 2.1. In Section 2.2 we introduce Local and Global Max-Gossip, and review Randomized Gossip and Load-Balancing distributed averaging schemes. We discuss the role of maximal dissent agents

and their selection in averaging algorithms in Section 2.3. In Section 2.4, we present our main results on the convergence of maximal dissent state-dependent distributed subgradient methods. We provide a numerical example that shows the benefit of using algorithms based on the maximal dissent averaging in Section 2.6. We conclude the chapter in Section 2.7 where we outline future research directions.

2.1 Problem Formulation

Consider a distributed system with n agents with an underlying communication network defined by a graph $\mathcal{G} = ([n], \mathcal{E})$. Each agent $i \in [n]$ has access to a *local* convex function $f_i : \mathbb{R}^d \to \mathbb{R}$. The agents can communicate only with their one-hop neighbors as dictated by the network graph $\mathcal{G}$. Our goal is to design a distributed algorithm to solve the following unconstrained optimization problem

$$F^* = \min_{\boldsymbol{w} \in \mathbb{R}^d} F(\boldsymbol{w}), \quad \text{where} \quad F(\boldsymbol{w}) \triangleq \sum_{i=1}^{n} f_i(\boldsymbol{w}). \tag{2.1}$$

We assume that the local objective function f_i is known only to node i and the nodes can only communicate by exchanging information about their local estimates of the optimal solution.

The solution set of the problem is defined as

$$\mathcal{W}^* \triangleq \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} F(\boldsymbol{w}).$$

Throughout the chapter, we make extensive use of the notion of the *subgradient* of a function.

Definition 1 (Subgradient)**.** *A subgradient of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $\boldsymbol{w}_0 \in \mathbb{R}^d$ is a vector $\boldsymbol{g} \in \mathbb{R}^d$ such that*

$$f(\boldsymbol{w}_0) + \langle \boldsymbol{g}, \boldsymbol{w} - \boldsymbol{w}_0 \rangle \leq f(\boldsymbol{w})$$

for all $\boldsymbol{w} \in \mathbb{R}^d$. We denote the set of all subgradients of f at $\boldsymbol{w}_0$ by $\partial f(\boldsymbol{w}_0)$, which is called the subdifferential of f at $\boldsymbol{w}_0$.

We make the following assumptions on the structure of the optimization problem in Eq. (2.1).

Assumption 1 (Non-empty solution set)**.** *The optimal solution set is non-empty, i.e., $\mathcal{W}^* \neq \emptyset$.*

Assumption 2 (Bounded Subgradients)**.** *Each local objective function f_i's subgradients are uniformly bounded. In other words, for each $i \in [n]$, there exists a finite constant L_i such that for all $\boldsymbol{w} \in \mathbb{R}^d$, we have $\|\boldsymbol{g}\| \leq L_i,\ \boldsymbol{g} \in \partial f_i(\boldsymbol{w})$.*

There exist many algorithms to solve the problem in Eq. (2.1). Nedic and Ozdaglar [39] introduced one of the pioneering schemes, in which each agent keeps an estimate of the optimal solution and at each time step, the agents share their local estimate with their neighbors. Then, each agent updates its estimate using a time-varying, *state-independent* convex combination of the information received from its neighbors and its own local estimate. For $t \geq 0$, let $a_{ij}(t)$ denote the coefficients of the aforementioned convex combination at time t such that $a_{ij}(t) \geq 0$, for all $i, j \in [n]$, $a_{ij}(t) = 0$ if $\{i, j\} \notin \mathcal{E}$, and $\sum_{j=1}^{n} a_{ij}(t) = 1$, for all $i \in [n]$. Let $\mathbf{x}_i(t)$ denote the i-th agent's estimate of the optimal solution at time t. The convex combination is followed by taking a step in the direction of any subgradient in the subdifferential at the local estimate, i.e.,

$$\boldsymbol{x}_i(t+1) = \sum_{j=1}^{n} a_{ij}(t)\boldsymbol{x}_j(t) - \alpha(t)\boldsymbol{g}_i(t), \tag{2.2}$$

where $\boldsymbol{g}_i(t) \in \partial f_i\big(\boldsymbol{x}_i(t)\big)$, and $\alpha(t)$ is a step-size sequence.

Herein, we generalize the algorithm in [39] by allowing the weights in the convex combination to be *state-dependent* in addition to being time-varying. Let each agent $i \in [n]$ initialize its estimate at an arbitrary point $\boldsymbol{x}_i(0) \in \mathbb{R}^d$, which is updated at discrete-time iterations

$t \geq 0$ based on its own subgradient and the estimates received from neighboring agents as follows

$$\boldsymbol{w}_i(t+1) = \sum_{j=1}^{n} a_{ij}\big(t, \boldsymbol{x}_1(t), \boldsymbol{x}_2(t), \ldots, \boldsymbol{x}_n(t)\big)\boldsymbol{x}_j(t),$$
$$\boldsymbol{x}_i(t+1) = \boldsymbol{w}_i(t+1) - \alpha(t+1)\boldsymbol{g}_i(t+1),$$

where $a_{ij}\big(t, \boldsymbol{x}_1(t), \boldsymbol{x}_2(t), \ldots, \boldsymbol{x}_n(t)\big)$ are non-negative weights, $\alpha(t)$ is a step-size sequence, and $\boldsymbol{g}_i(t) \in \partial f_i\big(\boldsymbol{w}_i(t)\big)$ for all $t \geq 0$. We can express this update rule compactly in matrix form as

$$W(t+1) = A\big(t, X(t)\big)X(t), \tag{2.3}$$
$$X(t+1) = W(t+1) - \alpha(t+1)G(t+1),$$

where

$$A\big(t, X(t)\big) \triangleq \Big[a_{ij}\big(t, \boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_n(t)\big)\Big]_{i,j\in[n]}$$

and

$$X(t) \triangleq \begin{bmatrix} \boldsymbol{x}_1^T(t) \\ \vdots \\ \boldsymbol{x}_n^T(t) \end{bmatrix}, \; W(t) \triangleq \begin{bmatrix} \boldsymbol{w}_1^T(t) \\ \vdots \\ \boldsymbol{w}_n^T(t) \end{bmatrix}, \; G(t) \triangleq \begin{bmatrix} \boldsymbol{g}_1^T(t) \\ \vdots \\ \boldsymbol{g}_n^T(t) \end{bmatrix}.$$

Note that another difference between Eqs. (2.3) and (2.2) is that agent i computes the subgradient for the local function f_i at the computed average $\boldsymbol{w}_i(t+1)$ instead of $\boldsymbol{x}_i(t)$, $t \geq 0$.

Assumption 3 (Diminishing step-size)**.** *The step-sizes $\alpha(t) > 0$ form a non-increasing sequence that satisfies*

$$\sum_{t=1}^{\infty} \alpha(t) = \infty \;\; \textit{and} \;\; \sum_{t=1}^{\infty} \alpha^2(t) < \infty.$$

For a step-size sequence that satisfies Assumption 3, if the sequence of matrices $\{A(t)\}$, where $A(t) = [a_{ij}(t)]_{i,j\in[n]}$, is doubly stochastic and sufficiently mixing, and the objective functions satisfy the regularity conditions in Assumptions 1 and 2, then the iterates in Eq. (2.2) converge to an optimal solution irrespective of the initial conditions $\boldsymbol{x}_i(0) \in \mathbb{R}^d$,

i.e., $\lim_{t\to\infty} \boldsymbol{x}_i(t) = \boldsymbol{x}^*,\ i \in [n]$, where $\boldsymbol{x}^* \in \mathcal{W}^*$ [40, Propositions 4 and 5]. Our goal for the remainder of the chapter is to establish a similar result for state-dependent maximal dissent distributed subgradient methods.

2.2 State-dependent average-consensus

In this section, we discuss three state-dependent average-consensus schemes that can potentially accelerate the existing distributed optimization methods, in so doing, we endeavor to unify the state-dependent average-consensus methodology. The first scheme, Local Max-Gossip, was studied in [56] exclusively for the average consensus problem. We provide two novel averaging schemes, the Max-Gossip and Load-Balancing averaging schemes, that provide faster convergence. The dynamics of these algorithms can be understood as the instances of Eq. (2.3) with constant local cost functions $f_i(\boldsymbol{x}) \equiv c$, $i \in [n]$, i.e.,

$$X(t+1) = A\big(t, X(t)\big)X(t).$$

We will consider three (two asynchronous and one synchronous) algorithms. The first two algorithms are related to the well-known randomized gossip algorithm [11, 51]. First, we present a brief description of Randomized Gossip.

2.2.1 Randomized Gossip

Consider a network $\mathcal{G} = ([n], \mathcal{E})$ of n agents, where each agent has an initial estimate $\boldsymbol{x}_i(0)$. At each iteration $t \geq 0$, a node i is chosen uniformly from $[n]$, independently of the earlier realizations. Then, i chooses one of its neighbors $j \in \mathcal{N}_i$, where $\mathcal{N}_i \triangleq \{j \in [n] : \{i, j\} \in \mathcal{E}\}$, with probability $P_{ij} > 0$. The two nodes exchange their current states $\boldsymbol{x}_i(t)$ and $\boldsymbol{x}_j(t)$, and update their states according to

$$\boldsymbol{x}_i(t+1) = \boldsymbol{x}_j(t+1) = \frac{1}{2}\big(\boldsymbol{x}_i(t) + \boldsymbol{x}_j(t)\big). \tag{2.4}$$

The states of the remaining agents are unchanged. The update rule in Eq. (2.4) admits a more compact matrix representation as

$$X(t+1) = B(e)X(t), \tag{2.5}$$

where $e = \{i, j\}$, and

$$B(e) \triangleq I - \frac{1}{2}(\boldsymbol{b}_i - \boldsymbol{b}_j)(\boldsymbol{b}_i - \boldsymbol{b}_j)^T. \tag{2.6}$$

It is necessary that $\sum_{\ell=1}^{n} P_{i\ell} = 1$ for all i, where $P_{i\ell} = 0$ if and only if $\{i, \ell\} \notin \mathcal{E}$. The dynamical system described in Eq. (2.5) and its convergence rate are studied in [11].

2.2.2 Global Max-Gossip

The standard gossiping algorithm described above is state-independent in the sense that the selection of the *gossiping edge* e does not depend on the states at the agents at any time. Herein, we propose *Global Max-Gossip* where we select the edge connecting the agents with the largest possible *dissent* (disagreement) among all edges in the graph $\mathcal{G} = ([n], \mathcal{E})$, i.e.,

$$e_{\max}(\mathcal{G}, X) = \arg \max_{\{i,j\} \in \mathcal{E}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{2.7}$$

In case there are multiple solutions to Eq. (2.7), we select the smallest pair of indices (i^*, j^*) based on the lexicographical order, without loss of optimality. For brevity, we use $e_{\max}(X)$ to denote the *max-edge*.

Global Max-Gossip serves as a benchmark as to what is achievable via state-dependent averaging schemes. Global Max-Gossip requires an oracle to provide the edge resulting in the largest possible Lyapunov function reduction across all network edges. Obtaining a decentralized algorithm to determine the max-dissent edge is a challenging open problem beyond the scope of this chapter.

Given an initial state matrix $X(0)$, the Max-Gossip averaging scheme admits a state-

dependent dynamics of the form

$$A\big(t, X(t)\big) = B\Big(e_{\max}\big(X(t)\big)\Big),$$

where the gossiping matrix is given by Eq. (2.6) and the max-edge is selected according to Eq. (2.7).

2.2.3 Local Max-Gossip

In *Local Max-Gossip* introduced in [56] under the moniker of *Greedy Gossip with Eavesdropping*, a random selected node gossips with the neighbor $j \in \mathcal{N}_i$ that has the largest[1] possible state discrepancy with i, i.e.,

$$j = \arg\max_{j \in \mathcal{N}_i} \|\boldsymbol{x}_j(t) - \boldsymbol{x}_i(t)\|. \tag{2.8}$$

Convergence is accelerated by gossiping with the neighbor with the largest disagreement as this leads to the largest possible immediate reduction in the Lyapunov function used to capture the variance of the states in the network.

Since the edge over which the gossiping occurs depends on the current state of the neighbors, the resulting averaging matrix is a state-dependent, random matrix. For a sequence of independently and uniformly distributed index sequence $\{s(t)\}$, the Local Max-Gossip dynamics can be written as a state-dependent averaging scheme as follows

$$A\big(t, X(t)\big) = B\Big(\big\{s(t), r_{s(t)}\big(X(t)\big)\big\}\Big),$$

where

$$r_s(X) = \arg\max_{r \in \mathcal{N}_s} \|\boldsymbol{x}_s - \boldsymbol{x}_r\|. \tag{2.9}$$

[1] In case there are multiple solutions to Eq. (2.8), we may select the agent with the smallest index, without loss of optimality.

2.2.4 Load-Balancing

Another state-dependent algorithm known as *Load-Balancing* can also be used to speed up convergence of average-consensus [38]. However, in contrast to the previous two cases, where only two nodes update at a given time, Load-Balancing is a synchronous averaging algorithm where all the agents operate simultaneously.

In the traditional Load-Balancing algorithm, the state at each agent is a scalar, which induces a total ordering amongst the agents, i.e., the neighbours of an agent are classified by having greater or smaller state values than the agent's current state. When the states at the agents are multi-dimensional vectors, a total ordering is not available and must be defined. We introduce a variant of Load-Balancing based on the Euclidean distance between the states of any two agents as follows.

At time t, each agent $i \in [n]$ carries out the following steps:

1. Agent i sends its state to its neighbors.

2. Agent i computes the distance between its state and each of its neighbors. Let $\mathcal{S}_i$ denote the subset of neighbors of agent i whose state have maximal Euclidean distance, i.e.,

$$\mathcal{S}_i \triangleq \arg\max_{j \in \mathcal{N}_i} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{2.10}$$

   Agent i sends an averaging request to the agents in $\mathcal{S}_i$.

3. Agent i receives averaging requests from its neighbors. If it receives a request from a single agent $j \in \mathcal{S}_i$, then it sends an acknowledgement to that agent. In the event that agent i receives multiple requests, it sends an acknowledgement to one of the requests uniformly at random.

4. If agent i sends and receives an acknowledgement from agent j, then it updates its state as $\boldsymbol{x}_i \leftarrow (\boldsymbol{x}_i + \boldsymbol{x}_j)/2$.

The conditions for interaction between two nodes in Load-Balancing is characterized in the following proposition.

Proposition 1. *Consider a connected graph $\mathcal{G}$ and a stochastic process $\big\{X(t), A\big(t, X(t)\big)\big\}$, where $A\big(t, X(t)\big)$ is the characterization of averaging according to the Load-Balancing algorithm, i.e. $A\big(t, X(t)\big)X(t)$ is the output of the Load-Balancing algorithm for a network with state matrix $X(t)$, $t \geq 0$. The following statements hold:*

1. *Two agents i, j such that $(i, j) \in \mathcal{E}$ average their states only if*

$$\|\boldsymbol{x}_i(t) - \boldsymbol{x}_j(t)\| \geq \max\Big\{ \max_{r \in \mathcal{N}_i \setminus \{j\}} \|\boldsymbol{x}_i(t) - \boldsymbol{x}_r(t)\|, \max_{r \in \mathcal{N}_j \setminus \{i\}} \|\boldsymbol{x}_j(t) - \boldsymbol{x}_r(t)\| \Big\}. \quad (2.11)$$

2. *Let $(i, j) \in \mathcal{E}$. If Eq.* (2.11) *holds with strict inequality, then i, j average their states.*

Proposition 1 is proven in Section 2.8.1.

2.3 On the selection of Max-edges

Consider the stochastic process $\big\{X(t), A\big(t, X(t)\big)\big\}$, where $X(t)$ is the network state matrix, and $A\big(t, X(t)\big)$ a state-dependent averaging matrix. Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration such that $\mathcal{F}_t$ is the σ-algebra generated by

$$\big\{\{X(k), A\big(k, X(k)\big) \mid k \leq t\big\} \setminus \big\{A\big(t, X(t)\big)\big\}.$$

We establish a non-zero probability that a pair of agents that constitute a max-edge will update their states for the averaging schemes discussed in Section 2.2.

Proposition 2. *Let $\big\{X(t), A\big(t, X(t)\big)\big\}_{t=0}^{\infty}$ be the random process generated by either Randomized Gossip, Local Max-Gossip, Max-Gossip, or Load-Balancing consensus schemes. Then, for the random indices $i^*, j^* \in [n]$ defined through the max-edge in Eq.* (2.7) *as $e_{\max}\big(X(t)\big) = \{i^*, j^*\}$,*

we have

$$\mathbb{E}\Big[A\big(t,X(t)\big)^T A\big(t,X(t)\big) \mid \mathcal{F}_t\Big]_{i^*j^*} \geq \delta \ \ \textit{a.s.}, \tag{2.12}$$

where $\delta = \min_{\{i,j\}\in\mathcal{E}} P_{ij}/n$ *for Randomized Gossip, such that* P_{ij} *is the probability that node* i *chooses node* $j \in \mathcal{N}_j$*;* $\delta = 1/n$ *for Local Max-Gossip;* $\delta = 1/2$ *for Global Max-Gossip; and* $\delta = 1/\big(2(n-1)^2\big)$ *for Load-Balancing.*

Proposition 2 establishes that given the knowledge until time t, in expectation, the agents comprising the max-edge based on the network state matrix $X(t)$, exchange their values with a positive weight bounded away from zero. Qualitatively, for gossip-based algorithms, this implies that there is a positive probability bounded away from zero that the agents comprising the max-edge carry out exchange of information with each other. We use Proposition 2 along with Theorem 3 to establish that the averaging matrices characterizing the algorithms discussed in Section 2.2 are contracting. Therefore, the subgradient methods based on these averaging algorithms converge to the same optimal solution almost surely as stated in Corollary 1 of Theorem 4. In other words, as long as the averaging step involves gossip over the max-edge with positive probability (bounded away from zero), we will have a contraction in the Lyapunov function capturing the sample variance, which is a key step in proving the convergence of our averaging based-subgradient methods. Proposition 2 is proven in Section 2.8.2.

2.4 Convergence of state-dependent Distributed Optimization

In the previous section, we have set the stage for studying the convergence of state-dependent averaging-based distributed optimization algorithms. Our proofs rely on two properties: double stochasticity and the *contraction property* (Theorem 3).

To state the contraction property, we define the Lyapunov function $V : \mathbb{R}^{n\times d} \to \mathbb{R}$ as

$$V(X) \triangleq \sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2, \tag{2.13}$$

where $X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T$ and $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$.

Theorem 3 (Contraction property)**.** *Consider a connected graph $\mathcal{G}$ and the stochastic process $\big\{X(t), A\big(t, X(t)\big)\big\}_{t=0}^{\infty}$ with a natural filtration $\{\mathcal{F}_t\}_{t\geq 0}$ for the dynamics given by Eq.* (2.3)*. If $A\big(t, X(t)\big) \in \mathcal{F}_{t+1}$ is doubly stochastic for all $t \geq 0$, and for the random variables $i^*, j^* \in [n]$ defined through the max-edge in Eq.* (2.7) *as $e_{\max}\big(X(t)\big) = \{i^*, j^*\}$,*

$$\mathbb{E}\Big[A\big(t, X(t)\big)^T A\big(t, X(t)\big) \mid \mathcal{F}_t\Big]_{i^*j^*} \geq \delta, \;\; \textit{a.s.}, \tag{2.14}$$

where $\delta > 0$, holds for all $t \geq 0$ and $X(0) \in \mathbb{R}^{n\times d}$, then

$$\mathbb{E}\Big[V\big(A\big(t, X(t)\big)X(t)\big) \mid \mathcal{F}_t\Big] \leq \lambda V\big(X(t)\big) \;\; \textit{a.s.}, \tag{2.15}$$

where $\lambda = 1 - 2\delta/\big((n-1)\mathrm{diam}(\mathcal{G})^2\big)$.

Theorem 3 is proven in Section 2.8.3 and provides our key new ingredient: proving a contraction result for doubly stochastic averaging matrices containing the maximally dissenting edge. The proof of Theorem 3 makes use of the double stochasticity of the matrices to characterize the exact one-step decrease in the Lyapunov function and then uses a clever trick to characterize its fractional decrease based on the fact that underlying communication graph is connected.

Remark 1. *Theorem 3 also holds for time-varying graphs provided they remain connected at each time t. More precisely, the theorem holds for a sequence of connected graphs $\{\mathcal{G}_t\}$ and at every time $t \geq 0$, for i^*, j^* defined through $e_{\max}(\mathcal{G}_t, X(t))$, the inequality in Eq.* (2.14) *holds, then the inequality in Eq.* (2.15) *will hold with scaling at time t being*

$$\lambda_t = 1 - \frac{2\delta}{(n-1)\mathrm{diam}(\mathcal{G}_t)^2} \leq 1 - \frac{2\delta}{(n-1)^3}.$$

Therefore, the contraction property for connected time-varying graphs holds with a factor of at most $\underline{\lambda} \triangleq 1 - 2\delta/(n-1)^3$.

For a connected graph $\mathcal{G}$ and the stochastic process $\big\{X(t), A\big(t, X(t)\big)\big\}_{t=0}^{\infty}$ with the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$ generated according to the dynamics in Eq. (2.3), we define a contracting averaging matrix as follows.

Definition 2 (Contracting averaging matrix)**.** *A state-dependent averaging matrix $A\big(t, X(t)\big)$ is contracting with respect to the Lyapunov function $V(\cdot)$ in Eq.* (2.13) *if there exists a $\lambda \in (0, 1)$ such that*

$$\mathbb{E}\Big[V\Big(A\big(t, X(t)\big)X(t)\Big) \mid \mathcal{F}_t\Big] \leq \lambda V\big(X(t)\big) \tag{2.16}$$

holds a.s. for all $t \geq 0$.

The main result of this work establishes convergence guarantees for these dynamics as stated below.

Theorem 4 (Almost sure convergence of state-dependent subgradient methods)**.** *Consider the distributed optimization problem in Eq.* (2.1) *and let Assumptions 1 and 2 hold. Assume a connected communication graph $\mathcal{G}$ and the subgradient method in Eq.* (2.3)*. If the random matrices $A\big(t, X(t)\big)$ in Eq.* (2.3) *are doubly stochastic and contracting, and the step-sizes $\{\alpha(t)\}$ follow Assumption 3, then for all initial conditions $X(0) \in \mathbb{R}^{n\times d}$,*

$$\lim_{t\to\infty} \boldsymbol{w}_i(t) = \boldsymbol{w}^*, \ \forall i \in [n], \ \ \textit{a.s.},$$

where $\boldsymbol{w}^* \in \mathcal{W}^*$.

Theorem 4 establishes the almost-sure convergence of the state variables to an optimal solution of Eq. (2.1), based on the consensus-based subgradient methods where the averaging matrices are doubly stochastic and contracting. Theorem 3 provides a simplified condition, the presence of averaging over the 'max-edge', which, when satisfied, implies the averaging matrix is contracting. Note that, as shown in Proposition 2, this simplified condition holds for Local Max-Gossip, Max-Gossip, and Load-Balancing averaging. Thus, we have the subsequent corollary following immediately from Proposition 2, Theorem 3, and Theorem 4.

Corollary 1. *Consider the distributed optimization problem in Eq.* (2.1) *and let Assumptions 1 and 2 hold. Assume a connected communication graph $\mathcal{G}$ and the subgradient method* (2.3) *where the averaging matrices $A\big(t, X(t)\big)$ in Eq.* (2.3) *are based solely on either the Local Max-Gossip, Max-Gossip or Load-Balancing averaging, and the step-sizes $\{\alpha(t)\}$ follow Assumption 3. Then*

$$\lim_{t\to\infty} \boldsymbol{w}_i(t) = \boldsymbol{w}^*, \ \forall i \in [n], \ \ \textit{a.s.},$$

for all initial condition $X(0) \in \mathbb{R}^{n\times d}$, and some $\boldsymbol{w}^* \in \mathcal{W}^*$.

For the remainder of this section, we provide the key steps and results that are needed to prove Theorem 4. We defer the proof of these technical results to the end of the chapter.

The proof strategy for Theorem 4 can be broken down into two main steps: (i) showing that the evolution of the dynamics followed by the average state variable $\{\bar{\boldsymbol{x}}(t)\}$ converges to a solution of the optimization problem in (2.1) and (ii) every node $i \in [n]$ tracks the dynamics of this average state variable such that the tracking error goes to zero. The first step requires the following result which establishes a bound on the accumulation of the tracking error for every agent.

Lemma 1. *Let $\mathcal{G}$ be a connected graph and consider sequences $\{W(t)\}$ and $\{X(t)\}$ generated by the subgradient method in Eq.* (2.3) *using sate-dependent, doubly stochastic and contracting averaging matrices $A\big(t, X(t)\big)$. If Assumptions 2 and 3 hold, then for any initial estimates $X(0) \in \mathbb{R}^{n\times d}$, the following hold a.s. for all $i \in [n]$*

$$\lim_{t\to\infty} \|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{x}}(t)\| = 0, \qquad \textit{and}$$
$$\sum_{t=0}^{\infty} \alpha(t+1)\mathbb{E}\left[\|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{x}}(t)\| \mid \mathcal{F}_t\right] < \infty.$$

Lemma 1, which is proven in Section 2.8.4, establishes guarantees on the consensus error for the local estimates $\boldsymbol{w}_i(t)$. Lemma 2 will be used to bound the distance of the average state $\bar{\boldsymbol{x}}(t)$ to an optimal point.

Lemma 2 (Lemma 8, [36])**.** *Suppose that Assumption 2 holds. Then, for any connected graph $\mathcal{G}$, initial condition $X(0) \in \mathbb{R}^{n\times d}$, $\boldsymbol{v} \in \mathbb{R}^d$, and $t \geq 0$, for the dynamics $\{X(t), A(t, X(t))\}$ of the subgradient method Eq.* (2.3) *where $A(t, X(t))$ are doubly stochastic, we have*

$$\mathbb{E}\big[\|\bar{\boldsymbol{x}}(t+1) - \boldsymbol{v}\|^2 \mid \mathcal{F}_t\big] \leq \|\bar{\boldsymbol{x}}(t) - \boldsymbol{v}\|^2 - \alpha(t+1)\frac{2}{n}\Big(F\big(\bar{\boldsymbol{x}}(t)\big) - F(\boldsymbol{v})\Big) + \alpha(t+1)\frac{4}{n}\sum_{i=1}^{n} L_i\mathbb{E}[\|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{x}}(t)\| \mid \mathcal{F}_t] + \alpha^2(t+1)\frac{L^2}{n^2}, \quad \textit{a.s.}$$

We note that Lemma 8 in [36] was originally intended for state independent dynamics. However, its proof only relies on the double stochasticity of the averaging matrices, convexity of the local functions, boundedness of the subgradients, and not on whether the averaging is state-dependent or not. Finally, combining the above two results implies that the distance of each agent's local estimate $\boldsymbol{x}_i(t)$ to the optimal set $\mathcal{W}^*$ will be *approximately* decreasing. The following result then will be used to show that this *approximate* decrease results in convergence to $\mathcal{W}^*$.

Lemma 3. *Consider a minimization problem $\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function. Assume that the solution set $\mathcal{X}^*$ of the problem is nonempty. Let $\{\boldsymbol{x}_t\}$ be a stochastic process such that for all $\boldsymbol{x} \in \mathcal{X}^*$ and for all $t \geq 0$,*

$$\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2 \mid \mathcal{F}_t] \leq (1 + b_t)\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - a_t\big(f(\boldsymbol{x}_t) - f(\boldsymbol{x})\big) + c_t \qquad \textit{a.s.},$$

where $b_t \geq 0$, $a_t \geq 0$, and $c_t \geq 0$ for all $t \geq 0$ and $\sum_{t=0}^{\infty} b_t < \infty$, $\sum_{t=0}^{\infty} a_t = \infty$, and $\sum_{t=0}^{\infty} c_t < \infty$ a.s. Then the sequence $\{\boldsymbol{x}_t\}$ converges to a solution $\boldsymbol{x}^* \in \mathcal{X}^*$ a.s.

This result has been proven as part of [3, Theorem 1] but due to the stand-alone significance of the result we have stated it as a lemma above and its proof is provided in Section 2.8.5. Now, we are ready to formally prove Theorem 4 by combining the aforementioned results.

Proof of Theorem 4. From Lemma 2, for $\boldsymbol{v} = \boldsymbol{w}^* \in \mathcal{W}^*$, we have

$$\mathbb{E}[\|\bar{\boldsymbol{x}}(t+1) - \boldsymbol{w}^*\|^2 \mid \mathcal{F}_t] \leq \|\bar{\boldsymbol{x}}(t) - \boldsymbol{w}^*\|^2 - \frac{2\alpha(t+1)}{n}\Big(F\big(\bar{\boldsymbol{x}}(t)\big) - F^*\Big) + \alpha^2(t+1)\frac{L^2}{n^2} + 4\frac{\alpha(t+1)}{n}\sum_{i=1}^{n} L_i \mathbb{E}[\|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{x}}(t)\| \mid \mathcal{F}_t],$$

for all $t \geq 0$. From Lemma 1, we know that

$$\sum_{t=0}^{\infty} 4\frac{\alpha(t+1)}{n}\sum_{i=1}^{n} L_i \mathbb{E}[\|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{x}}(t)\| \mid \mathcal{F}_t] = \sum_{i=1}^{n} \frac{4L_i}{n}\sum_{t=0}^{\infty} \alpha(t+1)\mathbb{E}[\|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{x}}(t)\| \mid \mathcal{F}_t] < \infty \ \ a.s.$$

Furthermore, $\alpha(t)$ is not summable and $\sum_{t=0}^{\infty} \alpha^2(t) < \infty$. Therefore, all the conditions for Lemma 3 hold with $a_t = 2\alpha(t+1)/n$, $b_t = 0$, and

$$c_t = \alpha(t+1)\frac{4}{n}\sum_{i=1}^{n} L_i \mathbb{E}[\|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{x}}(t)\| \mid \mathcal{F}_t] + \alpha^2(t+1)\frac{L^2}{n^2}.$$

Therefore, from Lemma 3, the sequence $\{\bar{\boldsymbol{x}}(t)\}$ converges to a solution $\hat{\boldsymbol{w}} \in \mathcal{W}^*$ almost surely. Finally, Lemma 1 implies that $\lim_{t\to\infty} \|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{x}}(t)\| = 0$ for all $i \in [n]$ almost surely. Therefore, the sequences $\{\boldsymbol{w}_i(t+1)\}$ converge to the same solution $\hat{\boldsymbol{w}} \in \mathcal{W}^*$ for all $i \in [n]$ almost surely. ■

2.5 Convergence Rate

In this section we discuss the convergence rate of the time-averaged version of the discussed state-dependent consensus based subgradient methods when the step size at time t is set as $1/\sqrt{t}$ for $t \geq 1$. The convergence rates for the different algorithm differ via the contraction factor λ defined for the contracting averaging matrix through (2.16).

Let λ_t be the contraction factor defined through the contracting property of the matrices at time t. More precisely, for all $t \geq 0$

$$\mathbb{E}[V(A(t, X(t))X(t))] \leq \lambda_t V(X(t)),$$

where $\lambda_t = \lambda\phi_t$ with $\phi_t \in (0, 1)$. Here, λ is the uniform bound on the contraction factor and $\lambda_t = \lambda_t(X(t))$ is a state-dependent (and possibly time-dependent) contraction factor. We refer to the tighter contraction bound to point out the improvement in convergence rate in state-dependent consensus based subgradient method. The proof of the convergence rates closely follow the proof provided in [36].

In the following lemma, we establish the convergence rate of the accumulation of error between the estimate for each agent from the mean of the estimates over all agents.

Lemma 4. *Under the assumptions of Theorem 4 with $\alpha(t) = 1/\sqrt{t}$, we have*

$$\frac{1}{\sqrt{n}} \sum_{k=0}^{t} \alpha(k+1) \sum_{i=1}^{n} \mathbb{E}[\|\boldsymbol{w}_i(k+1) - \bar{\boldsymbol{w}}_i(k+1)\|] \leq \Big(K_1 \mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + LK_2(1 + \ln t) \Big) \tag{2.17}$$

and

$$\frac{1}{\sum_{k=0}^{t} \alpha(k+1)} \sum_{k=0}^{t} \alpha(k+1) \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \mathbb{E}[\|\boldsymbol{w}_i(k+1) - \bar{\boldsymbol{w}}_i(k+1)\|] \leq \frac{1}{\sqrt{t+1}} (K_1 \mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + LK_2(1 + \ln t)), \tag{2.18}$$

where $K_1 = K_2 = \frac{\sqrt{\lambda}}{1-\sqrt{\lambda}}$.

Proof. From triangle inequality similar to (2.30), we know for all $t \geq 1$

$$\mathbb{E}[\|W(t+1) - \bar{W}(t+1)\|_F] \leq \sqrt{\lambda_t} \mathbb{E}[\|W(t) - \bar{W}(t)\|_F] + \sqrt{\lambda_t} \mathbb{E}[\|E(t) - \bar{E}(t)\|_F].$$

Repeatedly applying the above inequality and since the perturbation is bounded as $V(E(t)) \leq \frac{L^2}{t}$ for all $t \geq 1$ we get

$$\begin{aligned}\mathbb{E}[\|W(t+1) - \bar{W}(t+1)\|_F] &\leq \prod_{s=1}^{t} \sqrt{\lambda_s}\mathbb{E}[\|W(1) - \bar{W}(1)\|_F] \\ &\quad + \sum_{s=1}^{t}\prod_{k=s}^{t}\sqrt{\lambda_k}\mathbb{E}[\|E(s) - \bar{E}(s)\|_F] \\ &\leq \prod_{s=0}^{t} \sqrt{\lambda_s}\mathbb{E}[\|W(0) - \bar{W}(0)\|_F] + \sum_{s=1}^{t}\prod_{k=s}^{t}\sqrt{\lambda_k}\frac{L}{\sqrt{s}}.\end{aligned}$$

For brevity, define $\phi(t:s) = \prod_{k=s}^{t}\phi(k)$ and rewrite the above inequality as

$$\begin{aligned}\mathbb{E}[\|W(t+1) - \bar{W}(t+1)\|_F] &\leq \sqrt{\lambda}^{t+1}\phi(t:0)\mathbb{E}[\|W(0) - \bar{W}(0)\|_F] \\ &\quad + \sum_{s=1}^{t}\sqrt{\lambda}^{t-s+1}\phi(t:s)\frac{L}{\sqrt{s}} \qquad (2.19)\end{aligned}$$

To obtain the bound on accumulation of the errors, using (2.19) we get

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{k=0}^{t}\alpha(k+1)\sum_{i=1}^{n}\mathbb{E}[\|\boldsymbol{w}_i(k+1) - \bar{\boldsymbol{w}}_i(k+1)\|] \leq \sum_{k=0}^{t}\alpha(k+1)\|W(k+1) - \bar{W}(k+1)\|_F \\ &\leq \sum_{k=0}^{t}\frac{1}{\sqrt{k+1}}\sqrt{\lambda}^{k+1}\phi(k:0)\mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + L\sum_{k=1}^{t}\frac{1}{\sqrt{k+1}}\sum_{s=1}^{k}\frac{\sqrt{\lambda}^{k+1-s}\phi(k:s)}{\sqrt{s}} \\ &= c_1(t)\mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + Lc_2(t),\end{aligned}$$

where $c_1(t), c_2(t)$ are given by

$$c_1(t) := \sum_{k=0}^{t}\frac{\sqrt{\lambda}^{k+1}}{\sqrt{k+1}}\phi(k:0), \qquad c_2(t) := \sum_{k=1}^{t}\frac{1}{\sqrt{k+1}}\sum_{s=1}^{k}\frac{\sqrt{\lambda}^{k+1-s}\phi(k:s)}{\sqrt{s}}. \qquad (2.20)$$

Using the decreasing property of $\alpha(t)$, the fact that $\phi(t) \leq 1$ for all $t \geq 0$, and the expression for

a sum of a geometric series, we can uniformly bound $c_1(t)$ by $\frac{\sqrt{\lambda}}{1-\sqrt{\lambda}}$. For $c_2(t)$, note that

$$\begin{aligned} c_2(t) &\leq \sum_{k=1}^{t}\sum_{s=1}^{k}\frac{\sqrt{\lambda}^{k+1-s}\phi(k:s)}{s} \leq \sum_{k=1}^{t}\sum_{s=1}^{k}\frac{\sqrt{\lambda}^{k+1-s}}{s} \\ &= \sum_{s=1}^{t}\frac{1}{s}\sum_{k=1}^{t}\sqrt{\lambda}^{k+1-s} \leq \frac{\sqrt{\lambda}}{1-\sqrt{\lambda}}\sum_{s=1}^{t}\frac{1}{s} \leq \frac{\sqrt{\lambda}}{1-\sqrt{\lambda}}(1+\ln t), \end{aligned} \tag{2.21}$$

where the second inequality in (2.21) follows from

$$\sum_{s=1}^{t}\frac{1}{s} = 1+\sum_{s=2}^{t}\frac{1}{s} \leq 1+\int_{1}^{t}\frac{du}{u} = 1+\ln t.$$

Define $K_1 := \frac{\sqrt{\lambda}}{1-\sqrt{\lambda}}$ and $K_2 := \frac{\sqrt{\lambda}}{1-\sqrt{\lambda}}$. Therefore, we have

$$\begin{aligned} &\frac{1}{\sqrt{n}}\sum_{k=0}^{t}\alpha(k+1)\sum_{i=1}^{n}\mathbb{E}[\|\boldsymbol{w}_i(k+1)-\bar{\boldsymbol{w}}_i(k+1)\|] \\ &\qquad\qquad \leq K_1\mathbb{E}[\|X(0)-\bar{X}(0)\|_F] + LK_2(1+\ln t). \end{aligned}$$

Finally using the fact that $\sum_{k=0}^{t}\alpha(k+1) \geq \int_0^{t+1}\frac{du}{u+1} \geq \sqrt{t+1}$ we get inequality (2.18). ■

Using the accumulation of variance of the state estimates we establish an upper bound on the expected deviation of the global function at the time-averaged version of the average state estimates from the optimal value in the following lemma.

Lemma 5. *Under the assumptions of Theorem 4 with $\alpha(t) = 1/\sqrt{t}$ for all $t \geq 1$ and for any $\boldsymbol{w}^* \in \mathcal{W}^*$ we have*

$$\begin{aligned} \mathbb{E}\left[F\left(\frac{\sum_{k=0}^{t}\alpha(k+1)\bar{\boldsymbol{x}}(k)}{\sum_{k=0}^{t}\alpha(k+1)}\right) - F(\boldsymbol{w}^*)\right] &\leq \frac{n}{2}\frac{\mathbb{E}[\|\bar{\boldsymbol{x}}(0)-\boldsymbol{w}^*\|^2]}{\sqrt{t+1}} + \frac{L^2(1+\ln(t+1))}{2n\sqrt{t+1}} \\ &\quad + \frac{2L\sqrt{n}K_1}{\sqrt{t+1}}\mathbb{E}[\|X(0)-\bar{X}(0)\|_F] + 2L^2K_2\sqrt{n}\frac{1+\ln t}{\sqrt{t+1}}, \end{aligned}$$

where $K_1 = K_2 = \frac{\sqrt{\lambda}}{1-\sqrt{\lambda}}$.

Proof. By taking expectation on both sides for the inequality Lemma 2, for any $\boldsymbol{v} \in \mathbb{R}^d$ and $t \geq 0$ we have

$$\sum_{k=0}^{t} \frac{2\alpha(k+1)}{n} \mathbb{E}[F(\bar{\boldsymbol{x}}(k)) - F(\boldsymbol{v})] \leq \mathbb{E}[\|\bar{\boldsymbol{x}}(0) - \boldsymbol{v}\|^2] + \sum_{k=0}^{t} \alpha^2(k+1) \frac{L^2}{n^2} + \sum_{k=0}^{t} \frac{4\alpha(k+1)}{n} \sum_{i=1}^{n} L_i \mathbb{E}[\|\boldsymbol{w}_i(k+1) - \bar{\boldsymbol{w}}(t+1)\|],$$

since $\bar{\boldsymbol{w}}(t+1) = \bar{\boldsymbol{x}}(t)$ for all $t \geq 0$. Define $S(t+1) = \sum_{k=0}^{t} \alpha(k+1)$. Dividing the inequality above by $\frac{2S(t+1)}{n}$ we get

$$\sum_{k=0}^{t} \frac{\alpha(k+1)}{S(t+1)} \mathbb{E}[F(\bar{\boldsymbol{x}}(k)) - F(\boldsymbol{v})] \leq \frac{n}{2} \frac{\mathbb{E}[\|\bar{\boldsymbol{x}}(0) - \boldsymbol{v}\|]^2}{S(t+1)} + \frac{1}{S(t+1)} \sum_{k=0}^{t} \alpha^2(k+1) \frac{L^2}{2n} + \sum_{k=0}^{t} \frac{2\alpha(k+1)}{S(t+1)} \sum_{i=1}^{n} L_i \mathbb{E}[\|\boldsymbol{w}_i(k+1) - \bar{\boldsymbol{w}}(t+1)\|].$$

From Lemma 4 we have

$$\sum_{i=1}^{n} \sum_{k=0}^{t} \frac{\alpha(k+1)}{S(t+1)} \sum_{i=1}^{n} L_i \mathbb{E}[\|\boldsymbol{w}_i(k+1) - \bar{\boldsymbol{w}}(t+1)\|] \leq \frac{K_1\sqrt{n}}{\sqrt{t+1}} \mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + LK_2\sqrt{n} \frac{(1+\ln t)}{\sqrt{t+1}}. \tag{2.22}$$

Furthermore as $\sum_{k=0}^{t} \alpha^2(k+1) = \sum_{k=0}^{t} \frac{1}{k+1} \leq 1 + \ln(t+1)$ and $S(t+1) \geq \sqrt{t+1}$, for $\boldsymbol{v} = \boldsymbol{w}^*$ we have

$$\sum_{k=0}^{t} \frac{\alpha(k+1)}{S(t+1)} \mathbb{E}[F(\bar{\boldsymbol{x}}(k)) - F(\boldsymbol{w}^*)] \leq \frac{n}{2} \frac{\mathbb{E}[\|\bar{\boldsymbol{x}}(0) - \boldsymbol{w}^*\|^2]}{\sqrt{t+1}} + 2\frac{LK_1\sqrt{n}}{\sqrt{t+1}} \mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + L^2K_2\sqrt{n} \frac{1+\ln t}{\sqrt{t+1}} + \frac{1+\ln(t+1)}{\sqrt{t+1}} \frac{L^2}{2n}$$

which upon rearrangement gives us the result. ■

Finally, we provide a bound on the expected deviation of the global function computed at the time averaged version of the state estimates of any agent from the optimal value in the following theorem.

Theorem 5. *Consider the assumptions of Theorem 4 with* $\alpha(t) = 1/\sqrt{t}$ *for all* $t \geq 1$ *and* $\boldsymbol{w}^* \in \mathcal{W}^*$. *For* $\tilde{\boldsymbol{w}}_i(t+1) = \frac{\sum_{k=0}^{t} \alpha(k+1)\boldsymbol{w}_i(k+1)}{\sum_{k=0}^{t} \alpha(k+1)}$, *we have*

$$\begin{aligned}\mathbb{E}[F(\tilde{\boldsymbol{w}}_i(t+1)) - F(\boldsymbol{w}^*)] \leq \frac{n}{2}\frac{\mathbb{E}[\|\bar{\boldsymbol{x}}(0) - \boldsymbol{w}^*\|^2]}{\sqrt{t+1}} + \frac{L^2(1+\ln(t+1))}{2n\sqrt{t+1}} \\ + \frac{L(2\sqrt{n}+1)K_1}{\sqrt{t+1}}\mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + L^2K_2(2\sqrt{n}+1)\frac{1+\ln t}{\sqrt{t+1}},\end{aligned}$$

where $K_1 = K_2 = \frac{\sqrt{\lambda}}{1-\sqrt{\lambda}}$.

Proof. By the boundedness assumption of the subgradients we have

$$\begin{aligned}\mathbb{E}[F(\tilde{\boldsymbol{w}}_i(t+1)) - F\left(\frac{\sum_{k=0}^{t} \alpha(k+1)\bar{\boldsymbol{x}}(k)}{\sum_{k=0}^{t} \alpha(k+1)}\right) &\leq \frac{L\sum_{k=0}^{t} \alpha(k+1)\mathbb{E}[\|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{x}}(k)\|]}{\sum_{k=0}^{t} \alpha(k+1)} \\ &\leq \frac{L(K_1\mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + LK_2(1+\ln t))}{\sqrt{t+1}}\end{aligned}$$

Using the above inequality and Lemma 5 we get

$$\begin{aligned}\mathbb{E}[F(\tilde{\boldsymbol{w}}_i(t+1)) - F(\boldsymbol{w}^*)] &\leq \frac{L}{\sqrt{t+1}}(K_1\mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + LK_2(1+\ln t)) + \\ &+ \frac{n}{2}\frac{\mathbb{E}[\|\bar{\boldsymbol{x}}(0) - \boldsymbol{w}^*\|^2]}{\sqrt{t+1}} + \frac{L^2(1+\ln(t+1))}{2n\sqrt{t+1}} \\ &+ \frac{2LK_1\sqrt{n}}{\sqrt{t+1}}\mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + 2L^2K_2n\frac{1+\ln t}{\sqrt{t+1}}. \\ &= \frac{n}{2}\frac{\mathbb{E}[\|\bar{\boldsymbol{x}}(0) - \boldsymbol{w}^*\|^2]}{\sqrt{t+1}} + \frac{L^2(1+\ln(t+1))}{2n\sqrt{t+1}} \\ &+ \frac{LK_1(2\sqrt{n}+1)}{\sqrt{t+1}}\mathbb{E}[\|X(0) - \bar{X}(0)\|_F] + L^2K_2(2\sqrt{n}+1)\frac{1+\ln t}{\sqrt{t+1}}.\end{aligned}$$

■

2.5.1 Discussion

The subgradient method converges to the optimal at the rate of $O(\frac{\ln t}{\sqrt{t}})$. For randomized gossip, the convergence rate is comparable to the result that can be obtained from the result in Theorem 5 from the result in [36, Theorem 2]. However the approach in [36] cannot be directly used to state the result in Theorem 5 since the proof involves establishing inequality for every coordinate of the vector estimates and summing up the resulting inequalities. Such an approach cannot be extended to state-dependent averaging algorithms discussed in this work since the averaging step depends on the ℓ_2 norm of the difference between the nodes' estimates and cannot be decoupled to establish result on individual coordinates. Another reason behind using the contraction factor approach is the lack of B-connectivity result for the interaction between the agents when using state-dependent averaging.

The hidden constant terms of the convergence rate, $O(\frac{\ln t}{\sqrt{t}})$, are influenced by the consensus algorithm used with the subgradient descent. In Theorem 5 the consensus step of the algorithms influences the convergence rate through the constants K_1, K_2 such that the convergence becomes faster as the constants decrease. Note that K_1, K_2 are upper bounds for $c_1(t), c_2(t)$ defined through (2.20). Based on Theorem 3, the contraction factor $\lambda = 1 - \frac{2\delta}{(n-1)\mathrm{diam}(\mathcal{G})^2}$ is obtained in the following corollary, where δ for Randomized Gossip, Local Max-Gossip, Max-Gossip, and Load-Balancing are provided through Proposition 2.

Corollary 2. *In Theorem 5 the constants K_1, K_2 are given by $\frac{\sqrt{\lambda}}{1-\sqrt{\lambda}}$ which are bounded above by $n^2(n-1)\mathrm{diam}(\mathcal{G})^2$ for Randomized Gossip, $n(n-1)\mathrm{diam}(\mathcal{G})^2$ for Local Max-Gossip, $2(n-1)\mathrm{diam}(\mathcal{G})^2$ for Max-Gossip, and $(n-1)^3\mathrm{diam}(\mathcal{G})^2$ for Load-Balancing being used as the averaging scheme with the subgradient method.*

Proof. For Randomized Gossip, $1 - \sqrt{\lambda} \geq \frac{1}{2}(1-\lambda) \geq \frac{1}{n^2(n-1)\mathrm{diam}(\mathcal{G})^2} \geq \frac{1}{n^2(n-1)\mathrm{diam}(\mathcal{G})^2}$. Therefore K_1, K_2 are bounded as $\frac{\sqrt{\lambda}}{1-\sqrt{\lambda}} \leq n^2(n-1)\mathrm{diam}(\mathcal{G})^2$.

Similarly,

i. for Local Max-Gossip, $1 - \sqrt{\lambda} \geq \frac{1}{n(n-1)\mathrm{diam}(\mathcal{G})^2}$ leading to $\frac{\sqrt{\lambda}}{1-\sqrt{\lambda}} \leq n(n-1)\mathrm{diam}(\mathcal{G})^2$,

ii. for Max-Gossip, $1-\sqrt{\lambda} \geq \frac{1}{2(n-1)\mathrm{diam}(\mathcal{G})^2}$ leading to $\frac{\sqrt{\lambda}}{1-\sqrt{\lambda}} \leq 2(n-1)\mathrm{diam}(\mathcal{G})^2$,

iii. and for Load-Balancing, $1-\sqrt{\lambda} \geq \frac{1}{(n-1)^3\mathrm{diam}(\mathcal{G})^2}$ resulting in $\frac{\sqrt{\lambda}}{1-\sqrt{\lambda}} \leq (n-1)^3\mathrm{diam}(\mathcal{G})^2$.

■

Remark 2. *We may also comment that the above result uses a conservative bound on the contraction factor $\lambda > 0$. The values mentioned in Corollary 2 are upper bounds on the constants in the convergence rate. However, tighter bounds on the constants K_1, K_2 are possible. For Randomized Gossip, the contraction factor can be improved to the square of the second largest eigenvalue of the expected averaging matrix $\mathbb{E}[A(t, X(t))]$.*

In principle, in the proof of Theorem 5, for each of the state-dependent algorithm, such a contraction factor would depend on the sample path (past trajectory) of the dynamics. For example, when the consensus scheme used is Load-Balancing, we know that in practice, when the nodes do not have multiple neighbors with maximal disagreement, the constant δ in Proposition 2 is even grater than $\frac{1}{2}$, more precisely, it is $\frac{C_e(X)}{2}$, where C_e is the number of edges over which the exchange is taking place in the averaging step with the state estimate $X \in \mathbb{R}^{n\times d}$. With the improved δ, the bound on the constants K_1, K_2 can be improved to $\frac{(n-1)\mathrm{diam}(\mathcal{G})^2}{2C_e(X(t))} \leq \frac{(n-1)\mathrm{diam}(\mathcal{G})^2}{2}$.

Similarly the bounds on the convergence rate for Local Max-Gossip can be improved by using tighter contraction factor for the averaging matrices. However as seen from [56, Theorem 2], the contraction factor may take cumbersome form which cannot be readily used to establish better bounds on $c_1(t), c_2(t)$.

The problem of finding useful convergence rate for state-dependent averaging is a non-trivial open problem.

2.6 Numerical Examples

To illustrate our analytical results, we present a simulation of a distributed optimization problem where the local functions' subgradients are not restricted to be uniformly bounded. In particular, we look at the standard distributed estimation problem in a sensor network setting with

$n = 180$ agents. Here, each agent $i \in [n]$ wants to estimate an unknown parameter θ_0. Each node has access to a noisy measurement of the parameter $c_i = \theta_0 + n_i$, where n_i's are independent, zero mean Gaussian random variables with variance $\sigma_i^2 > 0$. In this setting, the Maximum Likelihood (ML) estimator [57, Theorem 5.3] is the minimizer of the separable cost function $F(w) = \sum_{i=1}^{n}(w - c_i)^2/\sigma_i^2$. Note that this problem is a distributed optimization problem with the local cost function $f_i(w) = (w - c_i)^2/\sigma_i^2$. For the variance σ_i^2, we picked $1/\sigma_i^2$ independently and uniformly over $(0, 1)$. For each node $i \in [n]$, the initial local estimates $x_i(0)$ are drawn independently from a standard Gaussian distribution.

We consider the performance for different topologies of the underlying communication graph $\mathcal{G}$ ranging from dense graphs (Complete and Barbell), moderately dense graphs (Erdös-Rényi), to sparse graphs (Line and Star). We chose a connected graph with the edge probability $p = 0.4$ for Erdös-Rényi graph. For the Barbell graph, we chose equal number of nodes for the three components – two Complete graphs and the connecting Line graph.

We ran the averaging-based subgradient optimizer with four different averaging update rules: Randomized Gossip [11], Local Max-Gossip, Max-Gossip, and Load-Balancing. For the Randomized Gossip, at each time a node in $[n]$ wakes up uniformly at random, and it chooses one of its neighbors uniformly at random for communication. To account for the stochastic nature of Randomized Gossip and Local Max-Gossip algorithms we average the error values over 10 runs keeping the initial conditions and samples at the nodes the same. The resulting plots in Fig. 2.1, show the decay of the error $\|\boldsymbol{w}(t) - w_*\mathbf{1}\|$ as a function of t, where $w_* = \sum_{i=1}^{n} \frac{c_i}{\sigma_i^2} / \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$ is the optimal solution for $F(w)$. For the Erdös-Rényi communication graph, we also plot the decay of the error with the number of bits exchanged between the nodes in Fig. 2.2 for Randomized Gossip, Local Max-Gossip, and Load-Balancing.

In the simulation, 32 bits are used for exchange of the estimates and 1 bit is used for the exchange of each acknowledgement. Therefore, the number of bits exchanged per step for Randomized Gossip is 64. For Local Max-Gossip, at time t with $s(t) \in [n]$ being the randomly chosen node, $|\mathcal{N}_{s(t)}| + 32|\mathcal{N}_{s(t)}| + 32$ bits are exchanged for waking up the neighboring nodes,

obtaining their values, and sending the neighbor with the maximum disagreement its own value. Finally, for Load-Balancing, $32\sum_{i=1}^{n}|\mathcal{N}_i| + n + \text{ACK}(t)$ bits are exchanged for sharing the values with the neighbors, sending request to the neighbour with the maximum disagreement, and sending the acknowledgement, where $\text{ACK}(t)$ is the total number of bits exchanged for sending the acknowledgement bits at time t.[2]

2.6.1 Comparison of Asynchronous Methods

From Fig. 2.1, the performance of the subgradient methods using state-dependent averaging shows an improvement in convergence rate. The convergence rates increase as we go from Randomized Gossip, Local Max-Gossip, Max-Gossip to Load-Balancing averaging based optimizers. We will refer to the subgradient methods using the state-dependent averaging by their averaging algorithm in the succeeding discussion.

In general, the performance of Max-Gossip is superior to the one of Local Max-Gossip. Clearly, Local Max-Gossip converges faster than Randomized Gossip. However convergence rate also depends on the graph topology: Local Max-Gossip applied on a Star graph has essentially the same rate as Randomized Gossip since the nodes at the periphery have only the central node as the choice to gossip with, and the probability of the first node being selected for gossiping is $n-1$ times larger to be a peripheral node as compared to the central node. Overall, we notice the increase in the performance of Max-Gossip and Local Max-Gossip as compared to Randomized Gossip with increasing connectivity. Moreover, from Fig. 2.2 we note the significantly better performance of Local Max-Gossip with respect to the number of exchanges between the nodes as opposed to that of synchronous Load-Balancing.

2.6.2 Max-Gossip vs. Load-Balancing

When comparing different state-dependent averaging schemes, it should be noted that unlike gossip, Max-Gossip, and Local Max-Gossip, Load-Balancing is a synchronous scheme

[2]In the numerical simulation, there are no cases with multiple neighbors with maximum disagreement.

where in addition to the max-edge, other local max-edges are often incorporated in the averaging scheme simultaneously. Therefore, it is only natural that the convergence rate of Load-Balancing is superior to that of Max-Gossip, since it averages not only the two nodes defined by the max-edge, but, additionally, other nodes connected by edges with large disagreement at the same time. By a similar logic, for the Complete graph, the performance of Load-Balancing and Max-Gossip are the same since all the nodes are holding scalar estimates and due to the ordering between the estimates, all the nodes send their request for averaging to either the node with the maximum or minimum estimate resulting in only the max-edge performing the updates.

We observe that the gap in performance of Load-Balancing and Max-Gossip, which has the best performance amongst the discussed asynchronous methods, increases with the diameter of the graph. Characterizing the analytical dependence of convergence rate as a function of graph topology metrics is of interest for future work.

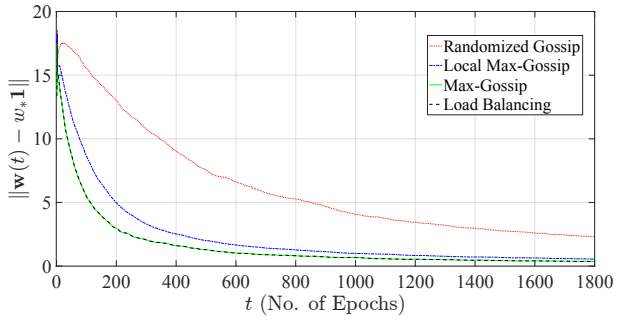
2.6.3 Logistic Regression

In order to illustrate the applicability of the results to a more general high-dimensional convex problem, we look at an example of regularized logistic regression for classification over MNIST dataset containing 56000 samples. In the experiment we train a model with the loss function defined as

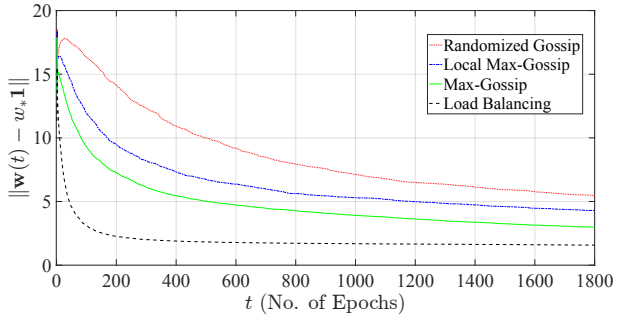
$$J(\boldsymbol{w}, b) = \frac{1}{m}\sum_{j=1}^{m}\left(-y_j \log\frac{1}{1+\exp(-(\boldsymbol{x_j}^T\boldsymbol{w})+b)} - (1-y_j)\log\frac{\exp(-(\boldsymbol{x_j}^T\boldsymbol{w})+b)}{1+\exp(-(\boldsymbol{x_j}^T\boldsymbol{w})+b)}\right) + \frac{1}{2m}\|\boldsymbol{w}\|^2 + \frac{1}{2m}|b|^2,$$

where $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^{56000}$ are the samples used for training. The samples are used to classify the digits in MNIST dataset into two classes based on whether the digits are greater than or equal to 5 or not. The experiment is run over a graph with 20 nodes with each node containing the same number of samples from the dataset. We initialize the nodes with all zero vectors.

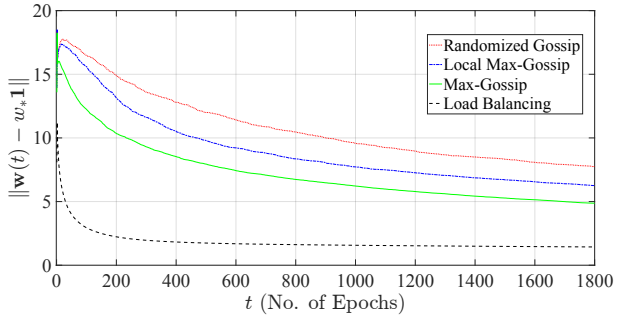
The communication graph representing the underlying connection between the nodes is a



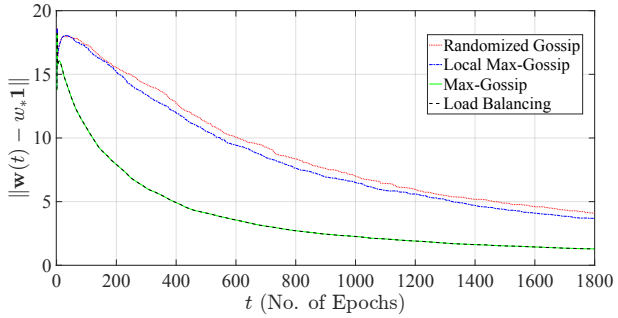
(a) Complete Graph



(b) Barbell Graph



(c) Line Graph



(d) Star Graph

Figure 2.1. Error decay for different graphs with 180 nodes

ladder graph. We consider the performance for the averaging-based subgradient optimizer with Randomized Gossip, Local Max-Gossip, Max-Gossip, and Load-Balancing. For Randomized

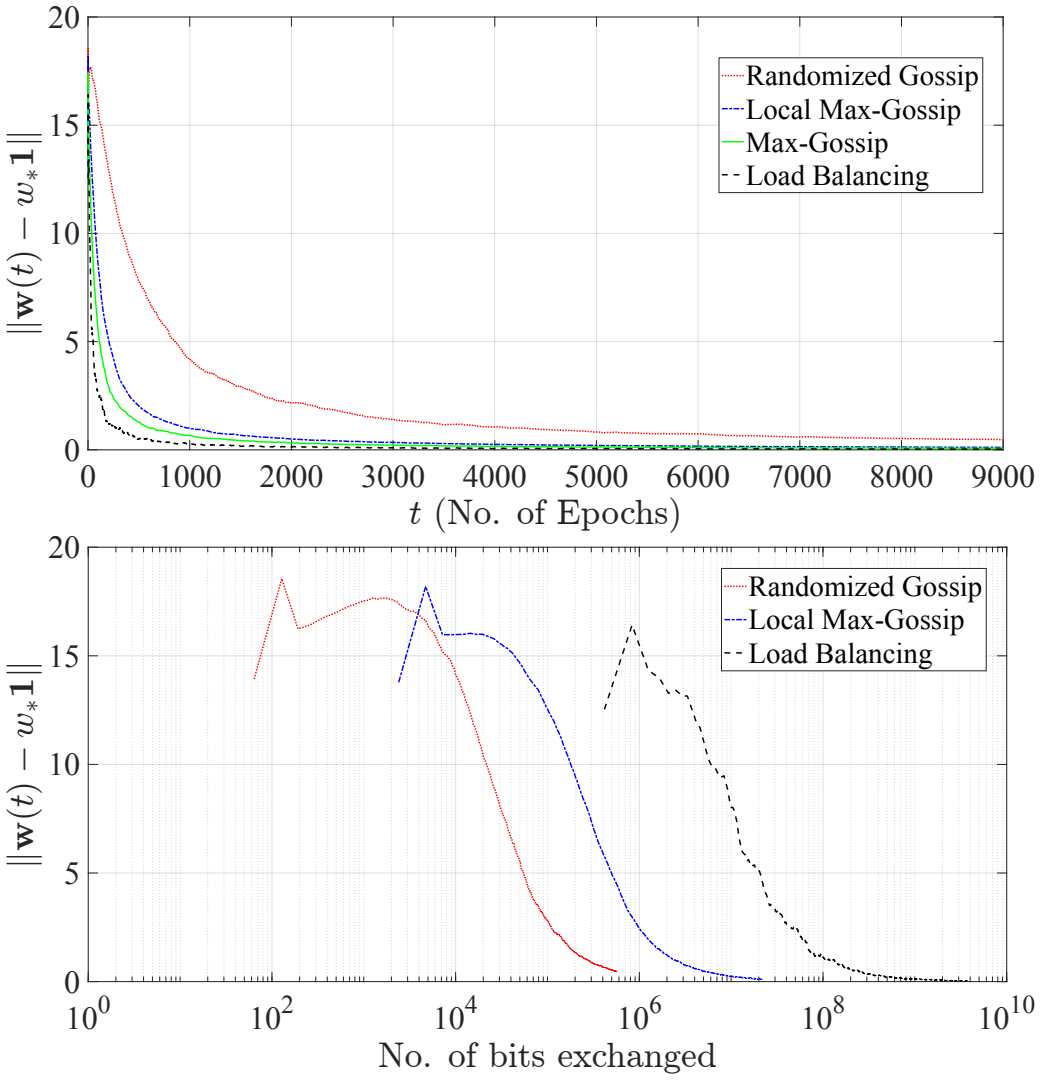


Figure 2.2. Error decay for Erdös-Rényi Graph with 180 nodes

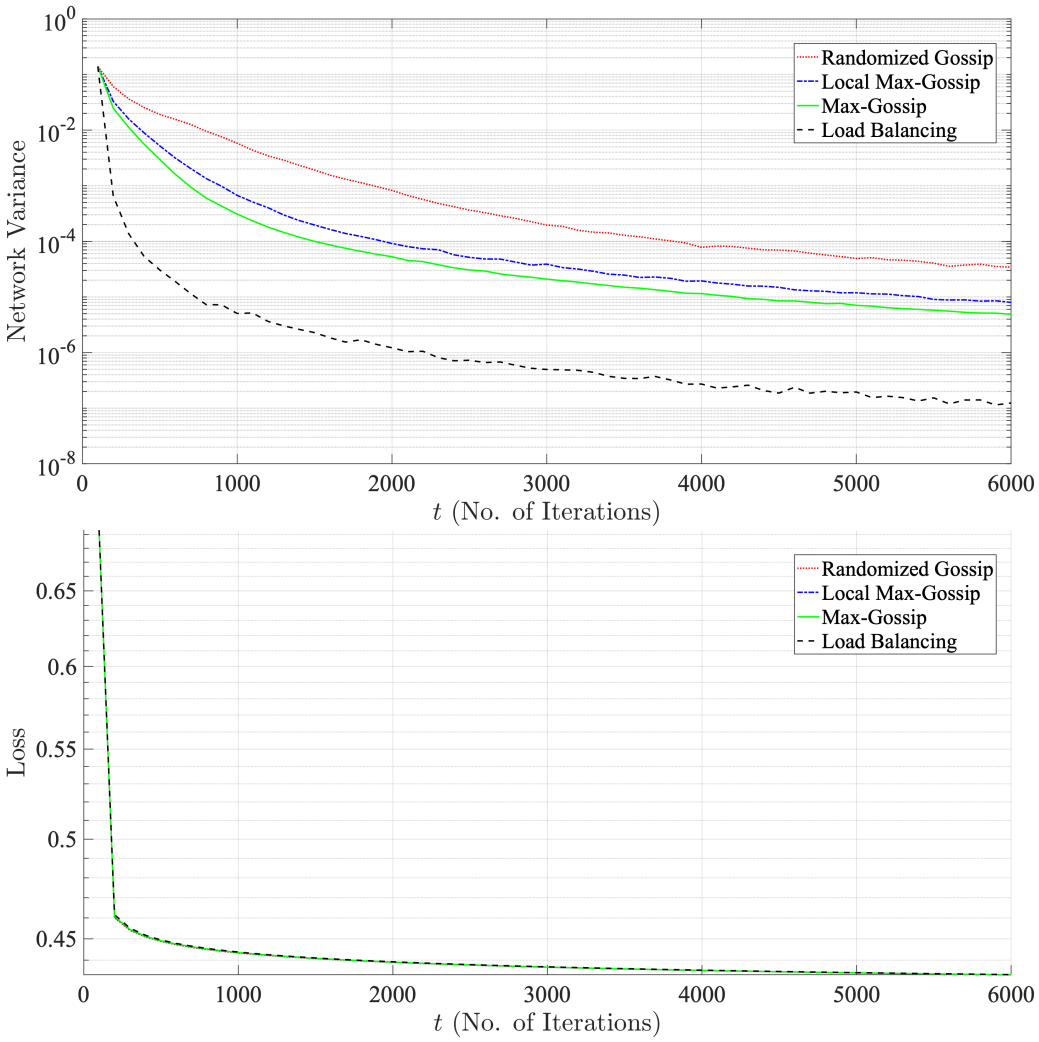


Figure 2.3. Network Variance for Ladder Graph with 20 nodes

Gossip, as in previous experiment, at each time a node in $[n]$ wakes uniformly at random and chooses one of its neighbor uniformly at random. We average the performance for Randomized Gossip and Local Max-Gossip over 3 runs. In Fig. 2.3 we plot the network variance, $\|W(t) - \frac{\mathbf{1}\mathbf{1}^T}{n} W(t)\|_F^2 + \|\boldsymbol{b}(t) - \frac{\mathbf{1}\mathbf{1}^T}{n} b(t)\|^2$ for step-size $\alpha(t) = \frac{1}{t}$ for all $t \geq 1$. We observe that the decay in the loss of the function for the consensus-based subgradient method is similar to each other. However the decay of the network variance, defined as the sum of the square of the deviation of the state estimates from their mean, over time in decreasing order of speed is observed for Load-Balancing, Max-Gossip, Local Max-Gossip, and finally Randomized Gossip.

2.7 Conclusions

We proposed, studied, and analyzed the role of maximal dissent nodes in distributed optimization schemes, leading to many exciting state-dependent consensus-based subgradient methods. The proof of our result relies on a certain contraction property of these schemes. Our result opens up avenues for synthesizing or extending the use of state-dependent averaging-schemes for distributed optimization including the Max-Gossip, Local Max-Gossip, and Load-Balancing algorithms. Finally, we compared simulation results of a distributed estimation problem for gossip-based subgradient methods and the proposed state-dependent algorithms. Our numerical experiments show the faster convergence speed of schemes that use maximal dissent between nodes compared with state-independent gossip schemes. These simulations strongly support the intuition behind our main result, i.e., mixing of information between the maximal dissent nodes is critically important for the working (and enhancing) of the consensus-based subgradient methods. Although, we have shown the convergence of such state-dependent algorithms, establishing their rate of convergence, and especially relating them to various graph quantities such as diameter and edge density of the graph remains open problems for future research endeavors. The introduction of a state-dependent element for other class of algorithms specifically those which provide linear convergence rates such as distributed gradient tracking method [46, 45] and their convergence analysis are part of future direction for the problem.

2.8 Skipped Proofs

2.8.1 Proof of Proposition 1

Proof of Proposition 1. For any $\omega \in \Omega$, consider $X(t;\omega) \in \mathbb{R}^{n\times d}$. If nodes i and j update their values to their average, that is $\big(\boldsymbol{x}_i(t;\omega) + \boldsymbol{x}_j(t;\omega)\big)/2$, then we know that during the round of Load-Balancing algorithm starting at value $X(t;\omega)$ in step 2, node i and node j have sent their averaging request to each other. Therefore, we have $j \in \arg\max_{r\in\mathcal{N}_i} \|\boldsymbol{x}_j(t;\omega) - \boldsymbol{x}_r(t;\omega)\|$ and

$i \in \arg\max_{r \in \mathcal{N}_j} \|\boldsymbol{x}_i(t;\omega) - \boldsymbol{x}_r(t;\omega)\|$. Hence, for any $\omega \in \Omega$,

$$\|\boldsymbol{x}_i(t;\omega) - \boldsymbol{x}_j(t;\omega)\| \geq \max\Big\{ \max_{r \in \mathcal{N}_i \setminus \{j\}} \|\boldsymbol{x}_i(t;\omega) - \boldsymbol{x}_r(t;\omega)\|, \max_{r \in \mathcal{N}_j \setminus \{i\}} \|\boldsymbol{x}_j(t;\omega) - \boldsymbol{x}_r(t;\omega)\| \Big\}. \tag{2.23}$$

On the other hand, if Eq. (2.23) holds with strict inequality, then node i and node j send averaging requests only to each other in step 2 and respond to each other in step 3, and carry out their averaging according to step 4. ■

2.8.2 Proof of Proposition 2

Proof. We first discuss the result for Randomized Gossip, Local Max-Gossip, and Max-Gossip averaging. The averaging matrices for the gossip algorithms where two agents update their states to their average takes the form of Eq. (2.6). Therefore, for these gossip algorithms we have

$$A\big(t, X(t)\big)^T A\big(t, X(t)\big) = A\big(t, X(t)\big)$$

and

$$\mathbb{E}\Big[A\big(t, X(t)\big)^T A\big(t, X(t)\big) \mid \mathcal{F}_t\Big] = \mathbb{E}\Big[A\big(t, X(t)\big) \mid \mathcal{F}_t\Big].$$

Consider two nodes $i, j \in [n]$ such that $\{i, j\} \in \mathcal{E}$. For Randomized Gossip, $\mathbb{E}\big[A\big(t, X(t)\big)_{ij} \mid \mathcal{F}_t\big] = (P_{ij} + P_{ji})/2n$. Moreover, since $\{i, j\} \in \mathcal{E}$, we have $P_{ij}, P_{ji} > 0$. Let $P_* = \min_{\{i,j\} \in \mathcal{E}} P_{ij}$. For the max-edge $\{i^*, j^*\}$, Eq. (2.12) holds with $\delta = P_*/n > 0$.

Let $i \in [n]$ and state estimate matrix $X(t)$. For Local Max-Gossip, let $r_i\big(X(t)\big)$ be determined according to Eq. (2.9). Consider the max-edge $e_{\max}\big(X(t)\big) = \{i^*, j^*\}$. Then, $r_{i*}\big(X(t)\big) = j^*$ and $r_{j^*}\big(X(t)\big) = i^*$. Thus,

$$\mathbb{E}\Big[A\big(t, X(t)\big)_{i^* j^*} \mid \mathcal{F}_t\Big] = \frac{1}{n}$$

and Local Max-Gossip averaging satisfies inequality Eq. (2.12) with $\delta = 1/n$.

Similarly, for the Max-Gossip averaging with state estimate $X(t)$ at time t, for the max-edge $e_{\max}\big(X(t)\big) = \{i^*, j^*\}$, we have

$$\mathbb{E}\left[A\big(t, X(t)\big)_{i^*j^*} \mid \mathcal{F}_t\right] = \frac{1}{2},$$

and Eq. (2.12) holds with $\delta = 1/2$.

Let us now discuss the presence of max-edge in the Load-Balancing averaging scheme. Consider the state estimate matrix $X(t)$ and $e_{\max}\big(X(t)\big) = \{i^*, j^*\}$ to be the max-edge with respect to $X(t)$. By the definition of a max-edge we know that nodes i^*, j^* satisfy inequality Eq. (2.11).

Consider the case when nodes i^*, j^* satisfy Eq. (2.11) with strict inequality. From Proposition 1, we know that $A(t, X(t))_{i^*j^*}, A\big(t, X(t)\big)_{j^*i^*}, A\big(t, X(t)\big)_{i^*i^*}, A\big(t, X(t)\big)_{j^*i^*}$ are equal to $1/2$, which implies that $A\big(t, X(t)\big)_{i^*\ell} = A\big(t, X(t)\big)_{\ell j^*} = 0$ for all $\ell \notin \{i^*, j^*\}$. Therefore,

$$\mathbb{E}\Big[A\big(t, X(t)\big)^T A\big(t, X(t)\big) \mid \mathcal{F}_t\Big]_{i^*j^*} = 1/2,$$

and the inequality in Eq. (2.12) holds with $\delta = 1/2$.

Finally, consider the case when there are multiple neighbors of nodes i^*, j^* with distance equal to $\|\boldsymbol{x}_{i^*}(t) - \boldsymbol{x}_{j^*}(t)\|$. Let $|\mathcal{S}_{i^*}| \geq 1$ and $|\mathcal{S}_{j^*}| \geq 1$ where $\mathcal{S}_i$ is given by Eq. (2.10). Then, according to Load-Balancing algorithm, nodes i^*, j^* update their states to their average with probability $1/(|\mathcal{S}_{i^*}| \cdot |\mathcal{S}_{j^*}|)$. Since $|\mathcal{S}_{i^*}| \leq n-1$ and $|\mathcal{S}_{j^*}| \leq n-1$, we have

$$\mathbb{E}\Big[A\big(t, X(t)\big)^T A\big(t, X(t)\big)_{i^*j^*} \mid \mathcal{F}_t\Big] \geq \frac{1}{2(n-1)^2},$$

and Eq. (2.12) holds with $\delta = 1/2(n-1)^2$. ■

2.8.3 Proof of Theorem 3

To prove Theorem 3 we must first define a few quantities related to the distance between the nodes on the graph and their relationships.

Definition 3. *Consider a connected graph $\mathcal{G}$ and a matrix $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times d}$ such that $\boldsymbol{x}_i \in \mathbb{R}^d$ is the estimate at node i in the graph $\mathcal{G}$. Let $d(X)$ denote the maximal distance between the estimates of any two nodes in the graph*

$$d(X) \triangleq \max_{i,j\in\{1,2,\ldots,n\}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{2.24}$$

Let $d_{\mathcal{G}}(X)$ denote the maximal distance between the estimates among any two connected nodes in the graph

$$d_{\mathcal{G}}(X) \triangleq \max_{\{i,j\}\in\mathcal{E}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{2.25}$$

Finally, let $\mathrm{diam}(\mathcal{G})$ denote the longest shortest path between any two nodes of the graph $\mathcal{G}$.

Proposition 6. *Given a connected graph $\mathcal{G}$ and a matrix $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times d}$, such that $\boldsymbol{x}_i \in \mathbb{R}^d$ is the solution estimate at node i in the graph $\mathcal{G}$, we have*

$$\frac{d(X)}{\mathrm{diam}(\mathcal{G})} \leq d_{\mathcal{G}}(X) \leq d(X).$$

Proof. The upper bound on $d_{\mathcal{G}}(X)$ follows from Eqs. (2.24) and (2.25) in Definition 3. To prove the lower bound on $d_{\mathcal{G}}(X)$, we assume, without loss of generality, that the rows of the matrix $X \in \mathbb{R}^{n\times d}$ are such that $d(X) = \|\boldsymbol{x}_1 - \boldsymbol{x}_n\|$. Since $\mathcal{G}$ is connected, its diameter is finite and there is a path of length $k \leq \mathrm{diam}(\mathcal{G})$, denoted by $\{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_{k-1}, v_k\}$, where $v_0 = 1$ and $v_k = n$, with $v_i \in [n]$ for $i = 0, 1, \ldots, k$. The distance $d(\boldsymbol{x})$ is bounded as

$$\|\boldsymbol{x}_1 - \boldsymbol{x}_n\| \leq \sum_{i=0}^{k-1} \|\boldsymbol{x}_{v_i} - \boldsymbol{x}_{v_{i+1}}\|, \tag{2.26}$$

where Eq. (2.26) follows from the triangle inequality. Finally, each term in the sum Eq. (2.26) is bounded above by $d_{\mathcal{G}}(\boldsymbol{x})$. Hence,

$$d(X) \leq k d_{\mathcal{G}}(X) \leq \operatorname{diam}(\mathcal{G}) d_{\mathcal{G}}(X).$$

■

Next, we state a result quantifying the decrease in the Lyapunov function defined in Eq. (2.13) that is the vector form of [38, Lemma 1].

Lemma 6. *Given a doubly stochastic matrix $A \in \mathbb{R}^{n\times n}$, let c_{ij} denote the (i,j)-th entry of the matrix A^TA. Then for all $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times d}$, we have*

$$V(AX) = V(X) - \sum_{i<j} c_{ij} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2.$$

Proof. By definition, the Lyapunov function in Eq. (2.13) can be written as

$$V(X) = \operatorname{tr}\big[(X - \bar{X})^T (X - \bar{X})\big],$$

where $\bar{X} = \frac{\mathbf{1}\mathbf{1}^T}{n} X$. The doubly stochasticity of A implies

$$\overline{AX} = \frac{\mathbf{1}\mathbf{1}^T}{n} AX = \frac{\mathbf{1}\mathbf{1}^T}{n} X = \frac{A\mathbf{1}\mathbf{1}^T}{n} X = A\bar{X}.$$

Therefore,

$$V(AX) = \operatorname{tr}[(AX - A\bar{X})^T (AX - A\bar{X})].$$

Finally,

$$V(X) - V(AX) = \operatorname{tr}[(X - \bar{X})^T (I - A^TA)(X - \bar{X})].$$

Since A^TA is a symmetric and stochastic matrix, we have $c_{ij} = c_{ji}$ and $c_{ii} = 1 - \sum_{i \neq j} c_{ij}$. Thus,

$$A^TA = I - \sum_{i<j} c_{ij}(\boldsymbol{b}_i - \boldsymbol{b}_j)(\boldsymbol{b}_i - \boldsymbol{b}_j)^T,$$

where $\boldsymbol{b}_i \in \mathbb{R}^n$ is the standard basis vector for all $i \in [n]$. Since

$$\mathrm{tr}[(X - \bar{X})^T(\boldsymbol{b}_i - \boldsymbol{b}_j)(\boldsymbol{b}_i - \boldsymbol{b}_j)^T(X - \bar{X})] = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2,$$

we have

$$V(X) - V(AX) = \sum_{i<j} c_{ij}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2.$$

■

Proof of Theorem 3. At time $t \geq 0$ consider the state estimate $X(t) = [\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_n(t)]^T \in \mathbb{R}^n$, the corresponding max-edge $e_{\max}(X(t)) = \{i^*, j^*\}$ and the doubly stochastic averaging matrix $A(t, X(t))$ such that

$$\mathbb{E}[A(t, X(t))^TA(t, X(t))_{i^*j^*} \mid \mathcal{F}_t] \geq \delta > 0 \text{ a.s.}$$

Define

$$\Omega_\delta(t) = \{\omega : \mathbb{E}[A(t, X(t))^TA(t, X(t))_{i^*j^*} \mid \mathcal{F}_t] \geq \delta\}.$$

For legibility, we drop the time index in the variables for the rest of this proof and use X, $\mathcal{F}$, $A(X)$, Ω_δ instead of $X(t), \mathcal{F}_t, \Omega_\delta(t)$, and $A(t, X(t))$.

Using arguments similar to the ones from [37, Lemma 9], for $X \in \mathcal{F}$ and doubly stochastic matrix $A(X)$ such that

$$\mathbb{E}[\big(A(X)^TA(X)\big)_{i^*j^*} \mid \mathcal{F}] \geq \delta > 0 \text{ a.s.}, \tag{2.27}$$

where $\mathcal{F}$ is a σ-field, $X \in \mathcal{F}$, and $e_{\max}(X) = \{i^*, j^*\}$. We will show that

$$\mathbb{E}\big[V(A(X)X) \mid \mathcal{F}\big] \leq \lambda V(X)$$

a.s. for some $\lambda \in (0, 1)$. From Lemma 6, the difference in the quadratic Lyapunov function V evaluated at X and $A(X)X$ is given by

$$V(X) - V(A(X)X) = \sum_{i<j} c_{ij}(X)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2,$$

where $c_{ij}(X)$ is the (i, j)-th entry of $A(X)^T A(X)$, i.e., $c_{ij}(X) = (A(X)^T A(X))_{ij}$. Taking the conditional expectation with respect to the filtration $\mathcal{F}$, we obtain

$$\begin{aligned} &V(X) - \mathbb{E}\big[V(A(X)X) \mid \mathcal{F}\big] \\ &\quad = \sum_{i<j} \big(\mathbb{E}[\big(A(X)^T A(X)\big)_{ij} \mid \mathcal{F}]\big)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \\ &\quad \geq c_{i^*j^*}(X)\|\boldsymbol{x}_{i^*} - \boldsymbol{x}_{j^*}\|^2 \geq \delta\|\boldsymbol{x}_{i^*} - \boldsymbol{x}_{j^*}\|^2 \text{ a.s.}, \end{aligned}$$

where $e_{\max}(X) = \{i^*, j^*\}$ and the first inequality follows from the non-negativity of the squared terms and the second inequality follows from Eq. (2.27). Recall that the constant δ depends on the averaging scheme.

If $V(X) = 0$, more precisely for the samples path characterized by $\omega \in \Omega_\delta(t)$ such that $V(X(t;\omega)) = 0$[3], then $X = \mathbf{1}\boldsymbol{c}^T$ for some $\boldsymbol{c} \in \mathbb{R}^d$. Therefore, $A(X)X = A(X)\mathbf{1}\boldsymbol{c}^T = \mathbf{1}\boldsymbol{c}^T$ since $A(X)$ is doubly stochastic and $V(A(X)X) = 0$. Thus, the inequality

$$\mathbb{E}\big[V(A(X)X) \mid \mathcal{F}\big] \leq \lambda V(X)$$

is satisfied.

[3]We omit the dependency on ω and t for legibility.

Let $\mathcal{L} \triangleq \{\mathbf{1}\boldsymbol{p}^T \mid \boldsymbol{p} \in \mathbb{R}^d\}$. For $X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T \notin \mathcal{L}$, more precisely for the samples path characterized by $\omega \in \Omega_\delta(t)$ such that $X(t;\omega) \notin \mathcal{L}$, the conditional expected fractional decrease in the Lyapunov function is

$$\frac{V(X) - \mathbb{E}[V(A(X)X) \mid \mathcal{F}]}{V(X)} \geq \delta \frac{\|\boldsymbol{x}_{i^*} - \boldsymbol{x}_{j^*}\|^2}{\sum_{i=1}^n \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2},$$

where $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$. Using the definition of $d_{\mathcal{G}}(X)$ and Proposition 6, we obtain the following bound

$$\frac{V(X) - \mathbb{E}[V(AX) \mid \mathcal{F}]}{V(X)} \geq \frac{\delta}{\operatorname{diam}(\mathcal{G})^2} \frac{d^2(X)}{\sum_{i=1}^n \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2}.$$

For $X \notin \mathcal{L}$, let

$$g(X) \triangleq \frac{d^2(X)}{\sum_{i=1}^n \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2}.$$

Note that $g(X)$ satisfies the following invariance relations

$$g(X + \mathbf{1}\boldsymbol{p}^T) = g(X), \;\; \boldsymbol{p} \in \mathbb{R}^d,$$

and

$$g(cX) = g(X), \;\; c \in \mathbb{R} \setminus \{0\}.$$

Therefore, for $X \notin \mathcal{L}$ the following inequality and identity hold

$$\begin{aligned} g(X) &\geq \min_{Z \in \mathbb{R}^{n\times d}: \sum_i \boldsymbol{z}_i = \mathbf{0}} \frac{d^2(Z)}{\sum_{i=1}^n \|\boldsymbol{z}_i\|^2} \\ &= \min_{Z \in \mathbb{R}^{n\times d}: \sum_i \boldsymbol{z}_i = \mathbf{0}; \sum_i \|\boldsymbol{z}_i\|^2 = 1} d^2(Z). \end{aligned}$$

Note that if $\sum_{i=1}^n \boldsymbol{z}_i = \mathbf{0}$ and $\sum_{i=1}^n \|\boldsymbol{z}_i\|^2 = 1$, then we have

$$\sum_{1 \leq i < j \leq n} \langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle = -\frac{1}{2}\sum_{i=1}^n \|\boldsymbol{z}_i\|^2 = -\frac{1}{2}. \tag{2.28}$$

By definition, $d(Z) \geq \|\boldsymbol{z}_i - \boldsymbol{z}_j\|$ for all $i, j \in [n]$. Using the fact that the maximum of a set of values is greater than its average for the set $\{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2\}_{1 \leq i < j \leq n}$, we get

$$d^2(Z) \geq \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2 = \frac{2}{n-1},$$

where the last step follows from Eq. (2.28) and the fact that $\sum_{i=1}^{n} \|\boldsymbol{z}_i\|^2 = 1$. Finally, using Eq. (2.28), we get

$$\frac{V(X) - \mathbb{E}\big[V(A(X)X) \mid \mathcal{F}\big]}{V(X)} \geq \frac{2\delta}{(n-1)\mathrm{diam}(\mathcal{G})^2}.$$

Since $\mathbb{E}\big[V(A(X)X) \mid \mathcal{F}\big] \leq \lambda V(X)$ for $X \in \mathcal{L}$ and for $X \notin \mathcal{L}$, we have $\mathbb{E}\big[V(A(X)X) \mid \mathcal{F}\big] \leq \lambda V(X)$ a.s. Thus,

$$\mathbb{E}\big[V(A(t, X(t))X(t)) \mid \mathcal{F}_t\big] \leq \lambda V\big(X(t)\big) \text{ a.s.},$$

where $\lambda = 1 - 2\delta/\big((n-1)\mathrm{diam}(\mathcal{G})^2\big)$. ■

2.8.4 Limiting properties of the Lyapunov function $V(\cdot)$

To prove Lemma 1 we will make use of the following result.

Theorem 7 (Robbins-Siegmund Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ be a sequence of sub σ-fields of $\mathcal{F}$. Let $\{u_t\}, \{v_t\}, \{q_t\}$, and $\{w_t\}$ be $\mathcal{F}_t$-measurable random variables, where $\{u_t\}$ is uniformly bounded from below, and $\{v_t\}$, $\{q_t\}$, and $\{w_t\}$ are non-negative. Let $\sum_{t=0}^{\infty} w_t < \infty$, $\sum_{t=0}^{\infty} q_t < \infty$ and*

$$\mathbb{E}[u_{t+1} \mid \mathcal{F}_t] \leq (1 + q_t)u_t - v_t + w_t, \text{ a.s.},$$

for all $t \geq 0$. Then, the sequence $\{u_t\}$ converges and $\sum_{t=0}^{\infty} v_t < \infty$ a.s.

Proof of Lemma 1. To study the convergence of $V\big(W(t)\big)$, we first derive a super-martingale like inequality for the stochastic process $\big\{V\big(W(t)\big)\big\}$. For $X(t) \in \mathcal{F}_t$ using the contracting

averaging property of $A(t, X(t))$ in Eq. (2.16), we get

$$\begin{aligned}\mathbb{E}\big[V\big(W(t+1)\big) \mid \mathcal{F}_t\big] &= \mathbb{E}\big[V\big(A(t, X(t))X(t)\big) \mid \mathcal{F}_t\big] \\ &\leq \lambda V\big(X(t)\big), \text{ a.s.,} \end{aligned} \tag{2.29}$$

where $\lambda \in (0, 1)$. We know that $X(t) = W(t) + E(t)$, so from triangle inequality on $\|W(t) - \bar{W}(t) + E(t) - \bar{E}(t)\|_F$ we have

$$V\big(X(t)\big) \leq V\big(W(t)\big) + V\big(E(t)\big) + 2\sqrt{V\big(W(t)\big)}\sqrt{V\big(E(t)\big)}. \tag{2.30}$$

Using the inequality above in Eq. (2.29), for all $t \geq 0$ we get

$$\mathbb{E}\big[V\big(W(t+1)\big) \mid \mathcal{F}_t\big] \leq \lambda\bigg(V\big(W(t)\big) + V\big(E(t)\big) + 2\sqrt{V\big(W(t)\big)}\sqrt{V\big(E(t)\big)}\bigg) \text{ a.s.}$$

Since $V\big(E(t)\big) = \|E(t) - \bar{E}(t)\|_F^2 \leq \|E(t)\|_F^2 \leq L^2\alpha^2(t)$, we get

$$\mathbb{E}[V\big(W(t+1)\big) \mid \mathcal{F}_t] \leq \lambda\left(\sqrt{V\big(W(t)\big)} + L\alpha(t)\right)^2 \text{ a.s.}$$

From Jensen's inequality, we have

$$\mathbb{E}\Big[\sqrt{V\big(W(t+1)\big)} \mid \mathcal{F}_t\Big] \leq \sqrt{\mathbb{E}\big[V\big(W(t+1)\big) \mid \mathcal{F}_t\big]} \leq \sqrt{\lambda}\Big(\sqrt{V\big(W(t)\big)} + L\alpha(t)\Big) \text{ a.s.}$$

Taking the expectation, multiplying by $\alpha(t+1)$ and using the fact that $\{\alpha(t)\}$ is non-increasing, we get

$$\alpha(t+1)\mathbb{E}\big[\sqrt{V\big(W(t+1)\big)}\big] \leq \alpha(t)\mathbb{E}\sqrt{V\big(W(t)\big)} - (1-\sqrt{\lambda})\alpha(t)\mathbb{E}\sqrt{V\big(W(t)\big)} + \alpha^2(t) \text{ a.s.}$$

Since the diminishing step sequence $\{\alpha(t)\}$ satisfies $\sum_{t=1}^{\infty} \alpha^2(t) < \infty$, Theorem 7 results in

$$\sum_{t=1}^{\infty} \alpha(t) \mathbb{E} \sqrt{V\big(W(t)\big)} < \infty,$$

and by the Monotone Convergence Theorem, we have,

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \alpha(t) \sqrt{V\big(W(t)\big)}\right] < \infty, \tag{2.31}$$

which implies that

$$\sum_{t=1}^{\infty} \alpha(t) \sqrt{V\big(W(t)\big)} < \infty, \text{ a.s.}$$

Since $V(W(t)) = \sum_{i=1}^{n} \|\boldsymbol{w}_i(t) - \bar{\boldsymbol{w}}(t)\|^2$, we know that

$$\sum_{t=1}^{\infty} \alpha(t) \|\boldsymbol{w}_i(t) - \bar{\boldsymbol{w}}(t)\| \leq \sum_{t=1}^{\infty} \alpha(t) \sqrt{V\big(W(t)\big)} < \infty,$$

for all $i \in [n]$, a.s. Since $\sum_{t=1}^{\infty} \alpha(t) \|\boldsymbol{w}_i(t) - \bar{\boldsymbol{w}}(t)\| < \infty$ and $\sum_{t=1}^{\infty} \alpha(t) = \infty$, we have

$$\liminf_{t \to \infty} \|\boldsymbol{w}_i(t) - \bar{\boldsymbol{w}}(t)\| = 0, \ \forall i \in [n], \quad \text{a.s.} \tag{2.32}$$

Further since we have,

$$\sum_{t=1}^{\infty} \alpha(t) \mathbb{E} \sqrt{V\big(W(t)\big)} = \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha(t) \mathbb{E}\left[\sqrt{V\big(W(t)\big)} \mid \mathcal{F}_t\right]\right],$$

using Monotone Convergence Theorem similar to Eq. (2.31) implies that

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \alpha(t) \mathbb{E}\left[\|\boldsymbol{w}_i(t) - \bar{\boldsymbol{w}}(t)\| \mid \mathcal{F}_t\right]\right] < \infty,$$

and so, we have

$$\sum_{t=1}^{\infty} \alpha(t) \mathbb{E}\left[\sqrt{V\big(W(t)\big)} \mid \mathcal{F}_t\right] < \infty \ \text{ a.s.},$$

and therefore,

$$\sum_{t=1}^{\infty} \alpha(t)\mathbb{E}\left[\|\boldsymbol{w}_i(t) - \bar{\boldsymbol{w}}(t)\| \mid \mathcal{F}_t\right] < \infty, \ \forall i \in [n], \text{ a.s.} \tag{2.33}$$

Further, for all $t \geq 0$, we know

$$\mathbb{E}\big[V\big(W(t+1)\big) \mid \mathcal{F}_t\big] \leq \lambda \left(V\big(W(t)\big) + 2L\alpha(t)\sqrt{V\big(W(t)\big)} + L^2\alpha^2(t) \right) \text{ a.s.}$$

Since we have

$$\sum_{t=1}^{\infty} 2\alpha(t)\sqrt{V\big(W(t)\big)} + \lambda L^2\alpha^2(t) < \infty \quad \text{a.s.},$$

Theorem 7 implies that $\{V(W(t))\}$ converges a.s. Therefore,

$$\|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{w}}(t+1)\| \text{ converges}, \forall i \in [n], \ \text{ a.s.}$$

Using (2.32) with the above result, we get

$$\lim_{t\to\infty} \|\boldsymbol{w}_i(t+1) - \bar{\boldsymbol{w}}(t+1)\| = 0, \ \forall i \in [n], \ \text{ a.s.} \tag{2.34}$$

Finally, since $\bar{\boldsymbol{w}}(t+1)^T = \frac{\mathbf{1}^T W(t+1)}{n} = \frac{\mathbf{1}^T A(t,X(t))X(t)}{n}$ from the double stochasticity of $A(t, X(t))$, we have

$$\bar{\boldsymbol{w}}(t+1)^T = \frac{\mathbf{1}^T X(t)}{n} = \bar{\boldsymbol{x}}(t)^T,$$

which from Eqs. (2.33) and (2.34) implies Lemma 1. ■

2.8.5 Proof of Lemma 3

To prove Lemma 3, we follow the proof in [3, Theorem 1].

Proof. For all $\boldsymbol{x} \in \mathcal{X}^*$ and $t \geq 0$, we have

$$\mathbb{E}\big[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2 \mid \mathcal{F}_t\big] \leq (1+b_t)\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - a_t\big(f(\boldsymbol{x}_t) - f(\boldsymbol{x})\big) + c_t \text{ a.s.}$$

For any $\boldsymbol{x} \in \mathcal{X}^*$, Theorem 7 implies that $\{\|\boldsymbol{x}_t - \boldsymbol{x}\|\}$ converges and

$$\sum_{t=0}^{\infty} a_t\big(f(\boldsymbol{x}_t) - f(\boldsymbol{x})\big) < \infty \text{ a.s.}$$

Since for any $\boldsymbol{x} \in \mathcal{X}^*$ we have $f(\boldsymbol{x}) = f^*$, the event

$$\Omega_{\boldsymbol{x}} = \left\{\omega : \lim_{t\to\infty} \|\boldsymbol{x}_t(\omega) - \boldsymbol{x}\| \text{ exists, and } \sum_{t=0}^{\infty} a_t(f(\boldsymbol{x}_t(\omega)) - f^*) < \infty\right\}$$

is such that $\mathbb{P}(\Omega_{\boldsymbol{x}}) = 1$. Note that here we denote by $\{\boldsymbol{x}_t(\omega)\}_{t\geq 0}$ the sample path for the corresponding ω.

Let $\mathcal{X}_d^* \subseteq \mathcal{X}^*$ be a countable dense subset of $\mathcal{X}^*$ and $\Omega_d = \bigcap_{\boldsymbol{x}\in\mathcal{X}_d^*} \Omega_{\boldsymbol{x}}$. We have $\mathbb{P}(\Omega_d) = 1$ since $\mathcal{X}_d^*$ is countable. For any $\omega \in \Omega_d$, since $\sum_{t=0}^{\infty} a_t = \infty$ and $\sum_{t=0}^{\infty} a_t\big(f(\boldsymbol{x}_t(\omega) - f^*\big) < \infty$, we have

$$\liminf_{t\to\infty} f\big(\boldsymbol{x}_t(\omega)\big) = f^*. \tag{2.35}$$

From Eq. (2.35) and the continuity of f, for all $\omega \in \Omega_d$, we have

$$\liminf_{t\to\infty} \|\boldsymbol{x}_t(\omega) - \boldsymbol{x}^*(\omega)\| = 0,$$

for some $\boldsymbol{x}^*(\omega) \in \mathcal{X}^*$[4]. Consider a subsequence $\{\boldsymbol{x}_{t_k}(\omega)\}_{k\geq 0}$ of $\{\boldsymbol{x}_t(\omega)\}_{t\geq 0}$ such that

$$\lim_{k\to\infty} f\big(\boldsymbol{x}_{t_k}(\omega)\big) = f^*.$$

For any $\omega \in \Omega_d$, $\lim_{t\to\infty} \|\boldsymbol{x}_t(\omega) - \hat{\boldsymbol{x}}\|$ exists for $\hat{\boldsymbol{x}} \in \mathcal{X}_d^*$. Therefore, the sequences $\{\boldsymbol{x}_t(\omega)\}_{t\geq 0}$ are bounded. Hence, $\{\boldsymbol{x}_{t_k}(\omega)\}_{k\geq 0}$ is also bounded, has a limit point $\boldsymbol{x}^*(\omega) \in \mathcal{X}^*$, and without loss of generality,

$$\lim_{k\to\infty} \boldsymbol{x}_{t_k}(\omega) = \boldsymbol{x}^*(\omega).$$

[4] $\boldsymbol{v}^*(\omega)$ may not be in $\mathcal{X}_d^*$.

Since $\mathcal{X}_d^*$ is dense, there is a sequence $\{\boldsymbol{q}_s(\omega)\}_{s\geq 0}$ in $\mathcal{X}_d^*$ such that

$$\lim_{s\to\infty} \|\boldsymbol{q}_s(\omega) - \boldsymbol{x}^*(\omega)\| = 0.$$

For $\omega \in \Omega_d$, $\lim_{t\to\infty} \|\boldsymbol{x}_t(\omega) - \boldsymbol{q}_s(\omega)\|$ exists for all $s \geq 0$, which is $\|\boldsymbol{x}^*(\omega) - \boldsymbol{q}_s(\omega)\|$. Moreover,

$$\lim_{t\to\infty} \|\boldsymbol{x}_t(\omega) - \boldsymbol{q}_s(\omega)\| \leq \liminf_{t\to\infty} \|\boldsymbol{x}_t(\omega) - \boldsymbol{x}^*(\omega)\| + \|\boldsymbol{x}^*(\omega) - \boldsymbol{q}_s(\omega)\| \leq \|\boldsymbol{x}^*(\omega) - \boldsymbol{q}_s(\omega)\|,$$

which implies that

$$\lim_{s\to\infty} \lim_{t\to\infty} \|\boldsymbol{x}_t(\omega) - \boldsymbol{q}_s(\omega)\| = 0.$$

Finally,

$$\limsup_{t\to\infty} \|\boldsymbol{x}_t(\omega) - \boldsymbol{x}^*(\omega)\| \leq \lim_{s\to\infty} \limsup_{t\to\infty} \|\boldsymbol{x}_t(\omega) - \boldsymbol{q}_s(\omega)\| + \|\boldsymbol{q}_s(\omega) - \boldsymbol{x}^*(\omega)\| = 0.$$

Therefore, for any $\omega \in \Omega_d$, we have $\lim_{t\to\infty} \boldsymbol{x}_t(\omega) = \boldsymbol{x}^*(\omega)$, where $\boldsymbol{x}^*(\omega) \in \mathcal{X}^*$. So we have, $\lim_{t\to\infty} \boldsymbol{x}_t = \boldsymbol{x}^*$ a.s. ■

Chapter 2, in full, is a reprint of the material as it appears in A. Verma, M. Vasconcelos, U. Mitra, B. Touri, "Maximal Dissent: a State-Dependent Way to Agree in Distributed Convex Optimization," in IEEE *Transactions on Control of Network Systems*. The dissertation author was the primary investigator and author of this paper.

Part II

Distributed Fact Checking

Chapter 3

Problem Formulation and Soft Estimator

In this chapter we move to the problem of Distributed Fact-Checking. The structure of the chapter is as follow:

1. *Formulation of Distributed Fact-Checking Problem*: We introduce a model for distributed *fact checking* which constitutes agents modeled as Binary Symmetric Channels with unknown reliability in Section 3.1.

2. *Online Estimator*: In Section 3.2 we propose an online estimator for the unreliability parameters of the agents which makes use of the likelihood ratio between source being fake or true given the agents' conclusion about the validity of the statement computed using the error estimate at a given time. Furthermore in Section 3.2.1 we introduce a variant of the proposed online estimator based on expanding truncation sets.

3.1 Problem Formulation

Consider a source that streams a sequence of statements. Each statement can be true or false. We use a *hidden* variable $S(t) \in \{+1, -1\}$ to denote the label (true/false) of the statement at discrete-time instance $t \in \mathbb{N}_0$. A *fact-checker* is interested in evaluating the validity of the statements using imperfect (inexpert) agents. We assume that the stream symbols are independently and identically distributed according to the Rademacher distribution, i.e., $\Pr(S(t) = +1) = \Pr(S(t) = -1) = \frac{1}{2}$, for every $t \in \mathbb{N}$.

Model for the fact-checker: We model a *fact-checker* as an overseer of multiple agents, where each agent is responsible for testing the validity of the statement provided to it. For $n \in \mathbb{N}$, let $[n]$ be the set of agents verifying the validity of the statements. At each time $t \in \mathbb{N}$, the agents observe the same statement $S(t)$ and output their evaluation regarding the validity of the statement to the fact checker, by returning their assessment about the statement. In other words, if the agent considers the statement correct it marks the statement as True, otherwise, it marks it as False. However, due to their limited expertise, the agents' assessments may be different from the actual label of the statements. Mathematically, we model agent $i \in [n]$ as a memoryless Binary Symmetric Channel (BSC) with the *error probability* or *crossover probability* $\pi_i \in (0, 1)$, takes the input $S(t)$ and outputs $R(t)$, where for every $s \in \{-1, +1\}$, the distribution of the output is given as

$$\Pr(R(t) = -s | S(t) = s) = 1 - \Pr(R(t) = s | S(t) = s) = \pi_i.$$

Therefore, agent $i \in [n]$ observes an output $R_i(t)$, which is independent of the past. Here, π_i represents the unreliability of agent i since the agent misclassifies the statement with probability π_i. Note that $\pi_i \in (0, 1/2)$ embodies the fact that the agents are not adversarial, and hence, are reliable agent on 'average'. We represent the collection of crossover probabilities by $\boldsymbol{\pi}$ and the sequence of all agents' outputs at time t by $\boldsymbol{R}(t)$.

Properties of Output distribution: Let us discuss some properties of output distribution.

i. Since the statement stream $\{S(t)\}$ is assumed to be independent and each agent is viewed as a memoryless channel, the random vector process $\{\boldsymbol{R}(t)\}$ is an independent process.

ii. At any time $t \in \mathbb{N}$, given $S(t)$, the outputs $\{R_i(t)\}_{i=1}^n$ are independent of each other. Moreover, for any $t \in \mathbb{N}$ and for every $i \in [n]$, $R_i(t)$ has the Rademacher distribution.

iii. The joint distribution of the output $\boldsymbol{R}(t)$ is given as

$$\Pr(\boldsymbol{R}(t)=\boldsymbol{r};\boldsymbol{\pi})=\frac{1}{2}\left(\prod_{i=1}^{n}\pi_i^{\frac{1+r_i}{2}}\bar{\pi}_i^{\frac{1-r_i}{2}}+\prod_{i=1}^{n}\pi_i^{\frac{1-r_i}{2}}\bar{\pi}_i^{\frac{1+r_i}{2}}\right),$$

where $\boldsymbol{r}\in\{+1,-1\}^n$, and $\bar{x}=1-x$.

We define the n-dimensional open unit hypercube as $\mathcal{X}=(0,1)^n$. For brevity, given unreliability parameters of the agents are $\boldsymbol{x}\in(0,1)^n$, we define $g_{\boldsymbol{x}}:\{+1,-1\}^n\to(0,1)$ to be the distribution of the output vector $\boldsymbol{R}\in\{-1,+1\}^n$, i.e.,

$$g_{\boldsymbol{\theta}}(\boldsymbol{r})=\Pr(\boldsymbol{R}=\boldsymbol{r};\boldsymbol{x}).$$

In this notation, $g_{\boldsymbol{\pi}}(\boldsymbol{R}(t))$ refers to the true distribution of the output vector $\boldsymbol{R}(t)$ at any time $t\in\mathbb{N}$.

If the unreliability vector is known $\boldsymbol{\pi}$, a linear thresholding estimator can be used to estimate the validity/label of the statements.

We mathematically formulate the problem of optimal estimator for truth detection/labeling. The result was obtained in [59] which characterizes the optimal set of estimators that minimize the probability of error of the labeling and will be discussed in detail later in Chapter 4. To estimate the validity of the statements based on individual agents' outputs, we focuss on the class of estimators that make the decision based on whether the linear combinations of the outputs are above or below a certain threshold.

Definition 4 (Linear Thresholding (LT) Estimator)**.** *Given outputs $\{R_i\}_{i=1}^n$ of n agents, we define a Linear Thresholding estimator with the weight vector $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_n)\in\mathbb{R}^n$ and threshold $\gamma\in\mathbb{R}$ as*

$$S_{\mathsf{L}}(\boldsymbol{\alpha},\gamma):=\mathsf{sgn}\left(\sum_{i=1}^{n}\alpha_iR_i-\gamma\right), \tag{3.1}$$

where $\mathsf{sgn}(x) = (-1)^{\mathbb{1}_{\{x<0\}}}$.

For brevity, we will refer to the LT estimator as S_{L}, and we identify an LT estimator by a vector $(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1}$ where $\boldsymbol{\alpha} \in \mathbb{R}^n$ is the weight vector and $\gamma \in \mathbb{R}$ is the threshold. Then, a fact-checker using an LT estimator with a vector $(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1}$ announces the statement to be True (or $S_{\mathsf{L}} = 1$) if $\sum_{i=1}^{n} \alpha_i R_i - \gamma \geq 0$, and False ($S_{\mathsf{L}} = -1$), otherwise.

In Chapter 4 we study the following problem regarding the optimal LT estimators for a statement using pseudo-experts (agents).

Problem 8. *Consider a fact-checker with access to n agents with known unreliability parameters $\pi_1, \pi_2, \ldots, \pi_n$ and known distribution of the statement $S \in \{-1, 1\}$ with $\Pr(S = +1) = \omega$. Determine the parameters $(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1}$ for the LT estimator S_{L}, defined through* (3.1)*, that minimizes the probability of error $\Pr(S_{\mathsf{L}} \neq S)$. distribution).*

In other words, our objective is to characterize the parameters $(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1}$ for the LT estimator to minimize the error probability $\Pr(S_{\mathsf{L}} \neq S)$. We state the solution to the problem in Theorem 10.

For distributed fact-checking another one of our main goals is to obtain reliable estimate for the unreliability parameters $\boldsymbol{\pi}$. Note that if $\{S(t)\}$ was known, $\boldsymbol{\pi}$ could be simply estimated as the fraction of time at which $R_i(t) \neq S(t)$. The challenge here is to estimate the channel parameters without the knowledge of channel input. TThe ideal problem would be to identify an estimator that converges almost surely to the true estimates of the unreliability parameter. However, some parameters result in agents' output distributions that are indistinguishable from each other. Thus, we focus on the following problem for multi-agent fact-checker.

Problem 9. *Consider a fact-checker with access to the sequence $\{\boldsymbol{R}(t)\}$ of the assessments of n agents, with unknown unreliability parameters π_i, for $i \in [n]$. Determine an online estimator for the unreliability parameters such that*

$$\lim_{t \to \infty} d(\boldsymbol{P}(t), \mathcal{S}) = 0, \textit{ almost surely (a.s.)},$$

where $\boldsymbol{P}(t)$ is the estimates of $\boldsymbol{\pi}$ based on the output of the n agents up to time t and $\mathcal{S}$ is the set of parameters that result in indistinguishable distribution for the agents' output, i.e.,

$$\mathcal{S} := \{\boldsymbol{x} \in \mathcal{X} \mid g_{\boldsymbol{x}}(\boldsymbol{r}) = g_{\boldsymbol{\pi}}(\boldsymbol{r}) \text{ for all } \boldsymbol{r} \in \{+1, -1\}^n\}. \tag{3.2}$$

In fact, we can characterize the set $\mathcal{S}$ for $n \geq 3$ as follows.

Lemma 7. *For $n \geq 3$, the set $\mathcal{S}$ defined in eq.* (3.2) *is given by $\mathcal{S} = \{\boldsymbol{\pi}, \mathbf{1} - \boldsymbol{\pi}\}$.*

Proof. For $\boldsymbol{r} \in \{+1, -1\}^n$, we have $g_{\boldsymbol{\pi}}(\boldsymbol{r}) = g_{\boldsymbol{x}}(\boldsymbol{r})$. Taking the sum over the subset where $r_1 = r_2 = +1$, i.e., $\{\boldsymbol{r} \in \{+1, -1\}^n \mid r_1 = r_2 = +1\}$ we get

$$x_1 x_2 + (1 - x_1)(1 - x_2) = \pi_1 \pi_2 + (1 - \pi_1)(1 - \pi_2).$$

Define $h(a, b) = ab + (1 - a)(1 - b)$ for $a, b \in [0, 1]$. In other words, $h(x_1, x_2) = h(\pi_1, \pi_2)$. Using $h(a, b) = 2\left(\frac{1}{2} - a\right)\left(\frac{1}{2} - b\right) + \frac{1}{2}$, we have

$$\left(\frac{1}{2} - x_1\right)\left(\frac{1}{2} - x_2\right) = \left(\frac{1}{2} - \pi_1\right)\left(\frac{1}{2} - \pi_2\right).$$

Similarly for all $i, j \in [n]$ such that $i \neq j$, we get

$$\left(\frac{1}{2} - x_i\right)\left(\frac{1}{2} - x_j\right) = \left(\frac{1}{2} - \pi_i\right)\left(\frac{1}{2} - \pi_j\right). \tag{3.3}$$

The above set of equations gives us $\left(\frac{1}{2} - x_i\right)^2 = \left(\frac{1}{2} - \pi_i\right)^2$ for all $i \in [n]$. Simplifying along with the system of equation (3.3) we get $\boldsymbol{x} = \boldsymbol{\pi}$ or $\mathbf{1} - \boldsymbol{\pi}$. Therefore $\mathcal{S} = \{\boldsymbol{\pi}, \mathbf{1} - \boldsymbol{\pi}\}$. ■

Chapter 5 provides the solution to Problem 9 for fact-checker system with two agents. The solution for the fact-checker system with $n \geq 3$ agents is provided in Chapter 6. In the following section we introduce the online soft estimator and its projection-based variant whose convergence analyses are provided in Chapters 5 and 6 respectively.

3.2 Estimator

First, let us introduce an online estimator for the unreliability parameters of the agents comprising the fact checker for any number of agents $n \geq 2$ whose convergence guarantees for $n = 2$ agents is provided in Chapter 5.

Consider the stream of output observed by the fact checker $\{\boldsymbol{R}(t)\}$. After any time $t \in \mathbb{N}$, it has an estimate $\boldsymbol{P}(t)$ of the unreliability parameters $\boldsymbol{\pi}$ obtained at time t, which is updated after observing the output $\boldsymbol{R}(t+1)$.

Recall that if $\boldsymbol{\pi}$ was known, the fact checker could evaluate the likelihood ratio $\frac{\Pr(\boldsymbol{R}(t+1)|S(t+1)=-1)}{\Pr(\boldsymbol{R}(t+1)|S(t+1)=+1)}$ to decode $S(t)$. Now, without $\boldsymbol{\pi}$, we can use its estimate $\boldsymbol{P}(t)$, to compute an approximate likelihood ratio $L(t)$ of $S(t+1) = -1$ to $S(t+1) = +1$ based on $\boldsymbol{R}(t+1)$. For this, let us define $L : \mathcal{X} \times \{+1, -1\}^n \to \mathbb{R}$ by

$$L(\boldsymbol{R}, \boldsymbol{x}) = \prod_{i=1}^{n} \left(\frac{x_i}{1 - x_i} \right)^{R_i}. \tag{3.4}$$

This represents the likelihood function of receiving $\boldsymbol{R}$ given the unreliability parameters $\boldsymbol{\pi} = \boldsymbol{x}$.

Now, for the received vector $\boldsymbol{R}(t+1)$ and an estimate $\boldsymbol{P}(t)$ of their unreliability parameters, for brevity, let

$$L(t) := L(\boldsymbol{R}(t+1), \boldsymbol{P}(t)). \tag{3.5}$$

Using $L(t)$, we can estimate $S(t+1)$ by setting

$$\hat{S}(t+1) := 2\mathbb{1}_{\{L(t)<1\}} - 1 = \begin{cases} -1 & \text{if } L(t) \geq 1, \\ +1 & \text{if } L(t) < 1, \end{cases} \tag{3.6}$$

where $\mathbb{1}_{\{\cdot\}}$ represents the indicator function.

We are ready to discuss the update rule for the unreliability parameters' estimates, given

the source symbol estimate $\hat{S}(t+1)$ and the output vector $\boldsymbol{R}(t+1)$. Note that $R_i(t+1)$ agreeing with $\hat{S}(t+1)$ suggests that it is unlikely that the agent was introducing error sat time t and hence, we average $P_i(t)$ with a value less than half to obtain $P_i(t+1)$. Similarly, if $R_i(t+1)$ disagrees with $\hat{S}(t+1)$, we average it with a value greater than half. More precisely the proposed algorithm/dynamics updates the unreliability parameters as

$$P_i(t+1) = (1-\eta_t)P_i(t) + \frac{1}{2}\eta_t \left(\frac{L(t)-1}{L(t)+1}R_i(t+1)+1\right), \tag{3.7}$$

for all $t \in \mathbb{N}_0$ and $i \in [n]$ with some initial condition (guess) $\boldsymbol{P}(0) \in (0,1/2)^n$, where $\{\eta_t\}$ is a pre-decided step-size sequence, and $L(t)$ is given in (3.5). Note that with an optimistic view of the ensemble of agents we set the initial condition in $(0,1/2)^n$, however the analysis covers all the cases, i.e., $\boldsymbol{P}(0) \in (0,1)^n$. Allow us to write the above iteration in compact form,

$$\boldsymbol{P}(t+1) = \boldsymbol{P}(t) + \eta_t \tilde{f}(\boldsymbol{R}(t+1), \boldsymbol{P}(t)), \tag{3.8}$$

where $\tilde{\boldsymbol{f}} : \{+1,-1\}^n \times \mathcal{X} \to \mathbb{R}^n$ is the vector field with its ith coordinate given by

$$\tilde{f}_i(\boldsymbol{R}, \boldsymbol{x}) := \frac{1}{2}\left(1 + \frac{L(\boldsymbol{R},\boldsymbol{x})-1}{L(\boldsymbol{R},\boldsymbol{x})+1}R_i\right) - x_i. \tag{3.9}$$

Note that (3.8) is a stochastic approximation-type iteration whose asymptotic behavior resembles the asymptotic behavior of the mean-field Ordinary Differential Equations (ODE)

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}), \tag{3.10}$$

where $\boldsymbol{f} : \mathcal{X} \to \mathcal{X}$ is defined by

$$\boldsymbol{f}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\tilde{\boldsymbol{f}}(\boldsymbol{R}, \boldsymbol{x})]. \tag{3.11}$$

We will expand on this viewpoint in Section 6.1 of Chapter 6.

Extension to $\bar{\mathcal{X}}$: Note that the likelihood function (3.4) can be extended to the vectors $\boldsymbol{x} \in [0,1]^n$, with *only* one element being 0 or 1 and such definition is not extendable to the case of more than two such elements. For this, let us first define *singly-extreme vectors* as follows.

Definition 5. *For $i \in [n]$, let us define*

$$\mathcal{X}^{(i)}_{bound} := \{\boldsymbol{x} \in [0,1]^n \mid x_i \in \{0,1\}, x_j \in (0,1) \forall j \in [n]_{-i}\}.$$

We also define the set of singly-extreme vectors *as* $\mathcal{X}_{bound} := \bigcup_{i=1}^{n} \mathcal{X}^{(i)}_{bound}$.

In fig. 3.1, we depict the sets $\mathcal{X}$, $\mathcal{X}^1_{\text{bound}}$, and $\mathcal{X}^2_{\text{bound}}$ for the case where $n = 2$. Assuming the convention $\frac{1}{0} := \lim_{p \to 0^+} \frac{1-p}{p} = \infty$, for a singly-extreme vector $\boldsymbol{x} \in \mathcal{X}^{(i)}_{\text{bound}}$, we define

$$L(\boldsymbol{R}, \boldsymbol{x}) := \begin{cases} 0 & \text{if } (-1)^{x_i} = R_i \\ \infty & \text{if } (-1)^{x_i} \neq R_i \end{cases},$$

leading to

$$\frac{L(\boldsymbol{R}, \boldsymbol{x}) - 1}{L(\boldsymbol{R}, \boldsymbol{x}) + 1} = \begin{cases} -1 & \text{if } (-1)^{x_i} = R_i \\ +1 & \text{if } (-1)^{x_i} \neq R_i \end{cases}. \tag{3.12}$$

Note that if for $i \neq j$, $x_i, x_j \in \{0,1\}$, then the likelihood ratio (3.4) cannot be defined for all vectors $\boldsymbol{R} \in \{+1,-1\}^n$. In particular, if $(-1)^{x_i} R_i \neq (-1)^{x_j} R_j$, the product in (3.4) would contain a 0 and ∞ term leading to an undefined expression $0 \times \infty$. This is in fact a fundamentally unresolvable phenomena as this is related to the case where the fact-checker is receiving two contradictory verdicts for a same statement from two fully reliable/unreliable agents.

With this discussion in mind, we can extend the definition of $\tilde{f}$ in (3.9) and $\boldsymbol{f}$ in (3.10) to the set $\bar{\mathcal{X}} := \mathcal{X} \cup \mathcal{X}_{\text{bound}}$, by considering the ratio (3.12) for $\boldsymbol{x} \in \mathcal{X}^{(i)}_{\text{bound}}$ for $i \in [n]$.

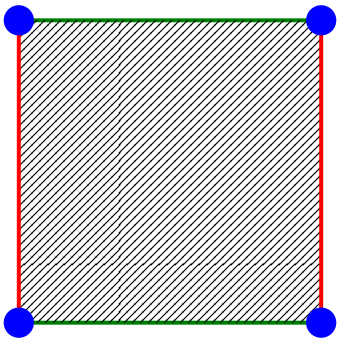


Figure 3.1. Parameter set: For $n = 2$ agents the red and green lines represent the sets $\mathcal{X}_{\text{bound}}^{(1)}$ and $\mathcal{X}_{\text{bound}}^{(2)}$ respectively. The shaded region represents $\mathcal{X}$. The box excluding the blue points represents $\bar{\mathcal{X}}$.

For the step-sizes η_t, we assume they satisfy the following stochastic approximation step-size assumption.

Assumption 4. *The step-sizes* $\{\eta_t\}$ *are positive, non-increasing, and satisfying* $\sum_{t=0}^{\infty} \eta_t = \infty$ *and* $\sum_{t=0}^{\infty} \eta_t^2 < \infty$.

One popular choice for the step-size sequence is the harmonic sequence $\eta_t = \frac{1}{t+1}$ for all $t \in \mathbb{N}_0$. To grasp the motivation behind the estimator using such a step-size sequence, examine the scenario when the fact checker knows the source sequence symbols $\{S(t)\}$. Since, at any time $t \in \mathbb{N}$, the output distribution of the agents given $S(t)$ is independent of each other, the problem of estimating the unreliability parameters of the agents is equivalent to n uncoupled problems of estimating the parameters of n Bernoulli distributions from their independent samples. Estimation of parameter for This problem is a well-studied problem and a class of estimators effective to solve it is the *add-constant* estimator [29]. For the current setting, for any $i \in [n]$, the add-β estimator, where $\beta \geq 0$ for parameter π_i at time $t \in \mathbb{N}$ is given by

$$Q_i(t) = \frac{\beta + \sum_{k=1}^{t} \mathbb{1}_{\{R_i(k) \neq S(k)\}}}{t + 2\beta}.$$

The estimator makes use of the empirical frequency of agent i misclassifying the source symbol

received and can be expressed recursively as

$$Q_i(t+1) = (1-\nu_t)Q_i(t) + \nu_t \mathbb{1}_{\{R_i(t+1)\neq S(t+1)\}}.$$

Here, $\nu_t := \frac{1}{t+1+2\beta}$ and $Q_i(0) = 1/2$. The convergence properties of estimator $Q(t)$ for different values of β and various loss functions are studied in [29]. Different values of β lead to well-known estimators, including the empirical estimator ($\beta = 0$), the Krichevsky–Trofimov (KT) estimator ($\beta = \frac{1}{2}$), and Laplace estimator ($\beta = 1$).

To see the connection to our setting, where the source symbol is unknown, consider an extreme case where $L(t) \gg 1$ (which implies $\hat{S}(t+1) = -1$). For $R_i(t+1) = +1$, we get

$$\frac{1}{2}\left(\frac{L(t)-1}{L(t)+1}R_i(t+1)+1\right) = \frac{L(t)}{L(t)+1} \approx 1,$$

whereas for $R_i(t+1) = -1$ we have $\frac{1}{L(t)+1} \approx 0$. Thus,

$$\frac{1}{2}\left(\frac{L(t)-1}{L(t)+1}R_i(t+1)+1\right) \approx \mathbb{1}_{\{R_i(t+1)\neq \hat{S}(t+1)\}}.$$

A similar situation holds when $L(t) \approx 0$. Therefore, the update rule (3.8) with $\eta_t = \frac{1}{t+1}$ can be viewed as an imperfect and adaptive version of the add-β estimator (with $\beta = 0$).

3.2.1 Estimator with resettings

With the presence of multiple agents making decisions, in order to obtain convergence guarantees, it is important to stay away from the boundary where two or more agents have estimated unreliability close to 0 or 1. In particular, in line with the concept of expanding truncations for stochastic approximation in [13, Chapter 2], we maintain a collection of growing truncation sets. These compact sets gradually converge to a superset. The estimate (but not the process) is reset to an arbitrary initial value each time it crosses the *current* truncation set.

We define $\{\mathcal{K}_t\}$ to be a sequence of increasing compact truncation sets for $\bar{\mathcal{X}}$, i.e.,

$\bigcup_{t=0}^{\infty} \mathcal{K}_t = \bar{\mathcal{X}}$ and for all $t \in \mathbb{N}_0$, $\mathcal{K}_t$ is compact and $\mathcal{K}_t \subseteq \mathcal{K}_{t+1}$.

There are many choices for the sequence of increasing truncation sets $\mathcal{K}_t$ that satisfy the criterion related to the Lyapunov functions. One such class of sets has the form $\mathcal{K}_t = \bigcup_{i=1}^{n} \{\boldsymbol{x} \in [0,1]^n | x_i \in [0,1], |x_j - \frac{1}{2}| \leq r_t, \forall j \in [n]_{-i}\}$, where r_t is a sequence of increasing positive numbers converging to $\frac{1}{2}$.

Algorithm 1. Estimator with expanding truncation

Input: $\{\mathcal{K}_t\}$, $\boldsymbol{P}_0 \in \mathcal{K}_0$,
Initialization: $\boldsymbol{P}_{\text{pr}}(0) \in \mathcal{K}_0, t = 0, \gamma(t) = 0$.
while $t \geq 0$ **do**
$\boldsymbol{y} = \boldsymbol{P}_{\text{pr}}(t) + \eta_t \tilde{f}(\boldsymbol{R}(t+1), \boldsymbol{P}_{\text{pr}}(t))$
if $\boldsymbol{y} \in \mathcal{K}_{\gamma(t)}$ **then**
$\boldsymbol{P}_{\text{pr}}(t+1) = \boldsymbol{y}$
$\gamma(t+1) = \gamma(t)$
else if $\boldsymbol{P}(t+1) \notin \mathcal{K}_{\gamma(t)}$ **then**
$\boldsymbol{P}_{\text{pr}}(t+1) = \boldsymbol{P}_0$
$\gamma(t+1) = \gamma(t) + 1$
end if
$t = t + 1$
end while

Consider the set of increasing truncation covers $\{\mathcal{K}_t\}$, an initial estimate $\boldsymbol{P}_{\text{pr}}(0) \in \mathcal{K}_0$.Let $\{\gamma(t)\}$ be a sequence of non-negative integers that keeps track of the current boundary set of the algorithm with $\gamma(0) = 0$. Then recursively, for any $t \in \mathbb{N}_0$, consider the update of the estimate similar to (3.8) at point $\boldsymbol{P}_{\text{pr}}(t)$, $\boldsymbol{y} = \boldsymbol{P}_{\text{pr}}(t) + \eta_t \tilde{\boldsymbol{f}}(\boldsymbol{R}(t+1), \boldsymbol{P}_{\text{pr}}(t))$. We reset the dynamics if $\boldsymbol{y}$ is outside the current active set $\mathcal{K}_{\gamma(t)}$, otherwise, let $\boldsymbol{P}_{\text{pr}}(t+1) = y$. In other words,

$$\boldsymbol{P}_{\text{pr}}(t+1) = \begin{cases} \boldsymbol{y}, & \text{if } \boldsymbol{y} \in \mathcal{K}_{\gamma(t)} \\ \boldsymbol{P}_0, & \text{if } \boldsymbol{y} \notin \mathcal{K}_{\gamma(t)} \end{cases}, \tag{3.13}$$

where $\boldsymbol{P}_0 \in \mathcal{K}_0$ is an arbitrarily chosen point from the set $\mathcal{K}_0$ and the counter for the truncation

set is defined as

$$\gamma(t+1) = \gamma(t) + \mathbb{1}_{\{\boldsymbol{y} \notin \mathcal{K}_{\gamma(t)}\}}. \tag{3.14}$$

Chapter 3, in in part, is a reprint of the material as it appears in A. Verma, A. Sharbafchi, S. Mohajer, B. Touri,“Distributed Fact Checking:A Stochastic Approximation Approach," in preparation for *IEEE Transactions on Autamtic Control*. The dissertation author was the primary investigator and author of this paper.

Chapter 4

LT Estimators

As stated in Chapter 3 for the distributed fact-checker, we consider the class of estimators that threshold a linear combination of the individual agents' response to estimate the validity of statements. In this chapter we provide the set of optimal weights and threshold to minimize the error probability of the estimator as the intersection of at most 2^n halfspaces, leading to uncountably many *optimal linear estimators.* For these optimal estimators, we provide an upper bound on the error probability, which provides insights regarding the decrease in the error probability of the fact-checker as the number of agents increases. We also provide an upper bound for an estimator that uses the majority rule to estimate the statements' validity and compare the two bounds. Similar problem has been studied in [52].

4.1 Problem Formulation

Let us revisit the problem formulation. For this chapter instead of a stream of statements $\{S(t)\}$ it is sufficient to study the one-shot version of the problem. Consider a source that outputs a statement, which might be True or False. We model the validity of the statement by a binary *hidden* random variable that is $S = +1$ if the statement is True and takes $S = -1$, otherwise. Further, we assume that the validity label is distributed as

$$\Pr(S = +1) = 1 - \Pr(S = -1) = \omega.$$

Model for the fact-checker: Recall that we model a *fact-checker* as an overseer of multiple agents, where each agent is responsible for labeling the validity of the statement provided to it. We consider a set of $n \in \mathbb{N}$ agents, denoted by $[n]$, each responsible for verifying the statement. The agents observe the same statement and output their evaluation regarding its validity according to their expertise and knowledge. However, due to imperfections, the label assigned to the statement by an agent is not necessarily the same as the true label S. Mathematically, we model the agents as a Binary Symmetric Channel (BSC), parameterized by the *error probability* or *crossover probability* $\pi \in [0, 1]$, which takes a binary[1] input random variable $S \in \{+1, -1\}$ and outputs a binary random variable $R \in \{+1, -1\}$ with

$$\Pr(R = -s \mid S = s) = 1 - \Pr(R = s \mid S = s) = \pi.$$

Using this, we view agent $i \in [n]$ as a BSC channel with input statement S and output R_i. Moreover, R_i is the output of a BSC with error probability π_i. In other words for $s \in \{+1, -1\}$

$$\Pr(R_i = -s \mid S = s) = 1 - \Pr(R_i = s \mid S = s) = \pi_i.$$

We represent the collection of crossover probabilities, and the output received at the fact-checker by $\boldsymbol{\pi} \coloneqq (\pi_1, \pi_2, \ldots, \pi_n)$ and $\boldsymbol{R} \coloneqq (R_1, R_2, \ldots, R_n)$, respectively. Moreover, we assume the agents' opinions are independent of each other, i.e.,

$$\Pr(\boldsymbol{R} = \boldsymbol{r} \mid S = s) = \prod_{i=1}^{n} \Pr(R_i = r_i \mid S = s),$$

for every $\boldsymbol{r} \in \{+1, -1\}^n$ and $s \in \{+1, -1\}$.

To estimate the validity of the statements based on individual agents' outputs, we focus on the class of estimators that make the decision based on whether the linear combinations of the outputs are above or below a certain threshold. Recall the class of LT estimator stated in

[1]For convenience, we use $\{+1, -1\}$ symbols instead of the default binary symbols $\{0, 1\}$.

Definition 4.

Definition (Linear Thresholding (LT) Estimator)**.** Given outputs $\{R_i\}_{i=1}^n$ of n agents, we define a *Linear Thresholding* estimator with the weight vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{R}^n$ and threshold $\gamma \in \mathbb{R}$ as

$$S_{\mathsf{L}}(\boldsymbol{\alpha}, \gamma) := \mathsf{sgn}\left(\sum_{i=1}^{n} \alpha_i R_i - \gamma\right), \tag{4.1}$$

where $\mathsf{sgn}(x) = (-1)^{\mathbb{1}\{x<0\}}$.

For brevity, we will refer to the LT estimator as S_{L}, and we identify an LT estimator by a vector $(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1}$ where $\boldsymbol{\alpha} \in \mathbb{R}^n$ is the weight vector and $\gamma \in \mathbb{R}$ is the threshold. Then, a fact-checker using an LT estimator with a vector $(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1}$ announces the statement to be True (or $S_{\mathsf{L}} = 1$) if $\sum_{i=1}^n \alpha_i R_i - \gamma \geq 0$, and False ($S_{\mathsf{L}} = -1$), otherwise.

In this work, we are interested in the study of optimal LT estimators for a statement using pseudo-experts. Recall Problem 8 stated in the Chapter 3.

Problem. Consider a fact-checker with access to n agents with known unreliability parameters $\pi_1, \pi_2, \ldots, \pi_n$ and known distribution of the statement $S \in \{-1, 1\}$ with $\Pr(S = +1) = \omega$. Determine the parameters $(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1}$ for the LT estimator S_{L}, defined through (4.1), that minimizes the probability of error $\Pr(S_{\mathsf{L}} \neq S)$.

In other words, our objective is to characterize the parameters $(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1}$ for the LT estimator to minimize the error probability $\Pr(S_{\mathsf{L}} \neq S)$.

4.2 Main Result

In order to provide the solution to Problem 8 we define a property involving two vectors that resembles the parallel condition for two vectors.

Definition 6 (Almost Parallel Vectors)**.** *Consider two vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{n+1}$. Then the vector $\boldsymbol{\alpha}$ is said to be almost parallel to $\boldsymbol{\beta}$ and denoted by $\boldsymbol{\alpha} \widetilde{\parallel} \boldsymbol{\beta}$, if for every $\boldsymbol{r} \in \{-1, +1\}^{n+1}$ with $\sum_{i=1}^{n+1} r_i \beta_i \neq 0$ we have*

$$\mathsf{sgn}\left(\sum_{i=1}^{n+1} r_i \alpha_i\right) = \mathsf{sgn}\left(\sum_{i=1}^{n+1} r_i \beta_i\right). \tag{4.2}$$

Note that if $\sum_{i=1}^{n+1} r_i \beta_i = 0$, then there is no restriction on the sign of $\sum_{i=1}^{n+1} r_i \alpha_i$.

Let $\boldsymbol{\beta}$-hyperplane be the hyperplane in $\mathbb{R}^{n+1}$ whose normal vector is $\boldsymbol{\beta}$. Then the points in $\{+1, -1\}^{n+1}$ can be partitioned into three groups, namely, those above the hyperplane, below it, and right on it. The definition above ensures that the partitions above and below $\boldsymbol{\beta}$-hyperplane are subsets of the corresponding partitions with respect to $\boldsymbol{\alpha}$-hyperplane. It is worth mentioning that the definition above ignores those points that lie on $\boldsymbol{\beta}$-hyperplanes.

Remark 3. *Note that the equality in* (4.2) *for all $\boldsymbol{r} \in \mathbb{R}^{n+1}$ (instead of $\boldsymbol{r} \in \{+1, -1\}^{n+1}$) implies that the vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are normal vectors to the same hyperplane since they separate points in $\mathbb{R}^{n+1}$ in exactly the same way. In this sense, the definition of almost parallel vectors restricts this separation condition to only points in $\{+1, -1\}^{n+1}$.*

In the following theorem, we define the set of weight vectors and thresholds that minimize the error probability for the LT estimator to be the set of all vectors almost parallel to a vector defined through the parameters $\boldsymbol{\pi}, \omega$.

Theorem 10. *For a fact-checker with agents' unreliability parameters $\pi_i \in (0, 1)$ for $i \in [n]$, the set of optimal parameters for LT estimators is given by*

$$\mathcal{A} := \left\{(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1} \middle| (\boldsymbol{\alpha}, \gamma) \widetilde{\parallel} (\boldsymbol{\ell}_{\boldsymbol{\pi}}, \ell_\omega)\right\}, \tag{4.3}$$

where $\boldsymbol{\ell}_{\boldsymbol{\pi}} := (\ell_{\pi_1}, \ldots, \ell_{\pi_n}) = \left(\log \frac{1-\pi_1}{\pi_1}, \ldots, \log \frac{1-\pi_n}{\pi_n}\right)$ and $\ell_\omega := \log \frac{1-\omega}{\omega}$. In other words, the probability of error $\Pr(S_{\mathsf{L}}(\boldsymbol{\alpha}, \gamma) \neq S)$ is minimized if and only if $(\boldsymbol{\alpha}, \gamma) \in \mathcal{A}$.

Note that LT estimators can be interpreted as classifiers characterized by separating hyperplanes: labeling vertices of $\{+1,-1\}^n$ to two labels $+1$ and -1. Theorem 10 characterizes all such separating hyperplanes that optimally bisect those vertices as those whose normal vectors are almost parallel to $(\boldsymbol{\ell_\pi}, \ell_\omega)$. We refer to Fig. 4.1 for a pictorial illustration..

Remark 4. *In Theorem 10, the value of $\ell_{\pi_i} := \log \frac{1-\pi_i}{\pi_i}$ can be interpreted as the log-likelihood ratio of the conditional probability of received output of agent i being same as source symbol to the probability of received output being flipped.*

Remark 5. *Note that $\mathcal{A}$ includes $\mathcal{B} = \bigcap_{\boldsymbol{r}\in\{-1,+1\}^n} \mathcal{A}_{\boldsymbol{r}}$ where $\mathcal{A}_{\boldsymbol{r}}$ are open sets defined by*

$$\mathcal{A}_{\boldsymbol{r}} = \begin{cases} \mathbb{R}^{n+1} & \text{if } \sum_{i=1}^n r_i \ell_{\pi_i} - \ell_\omega = 0, \\ \{(\boldsymbol{\alpha},\gamma) | \sum_{i=1}^n r_i \alpha_i > \gamma\} & \text{if } \sum_{i=1}^n r_i \ell_{\pi_i} - \ell_\omega > 0, \\ \{(\boldsymbol{\alpha},\gamma) | \sum_{i=1}^n r_i \alpha_i < \gamma\} & \text{if } \sum_{i=1}^n r_i \ell_{\pi_i} - \ell_\omega < 0. \end{cases}$$

Note that $(\boldsymbol{\ell_\pi}, \ell_\omega) \in \mathcal{B}$ and since $\mathcal{B}$ is an intersection of finitely many open sets, it is a non-empty open set contained in $\mathcal{A}$. Therefore, not only does the optimal set $\mathcal{A}$ contain infinitely many vectors, but also it has a non-zero Borel measure. Therefore, a fact-checker does not need to know the exact value of unreliability parameters (cross-over probabilities π_i) to arrive at an optimal LT estimator. For example, it is sufficient to run an estimator for the unreliability parameters until we reach an estimate $\hat{\boldsymbol{\pi}}$ such that $(\ell_{\hat{\boldsymbol{\pi}}}, \ell_\omega) \in \mathcal{A}$ where $\ell_{\hat{\boldsymbol{\pi}}} = \left(\log \frac{1-\hat{\boldsymbol{\pi}}_1}{\hat{\boldsymbol{\pi}}_1}, \ldots, \log \frac{1-\hat{\boldsymbol{\pi}}_n}{\hat{\boldsymbol{\pi}}_n}\right)$. We demonstrate the use of this fact through a simulation in Section 4.4.

Example 1. *Consider a fact-checker comprised of $n = 2$ agents with unreliability parameters $\pi_1, \pi_2 \in (0, 1/2]$ with $\pi_1 < \pi_2$ and a source with parameter $\omega = 1/2$. Then $\{(\boldsymbol{\alpha},\gamma) \in \mathbb{R}^3 \mid |\alpha_2| < \alpha_1, \gamma = 0\}$ is the subset of optimal weights with threshold $\gamma = 0$ for LT estimators and the probability of error is given by $\Pr(S_{\mathsf{L}}^\star \neq S) = \pi_1$. To show this, we note that*

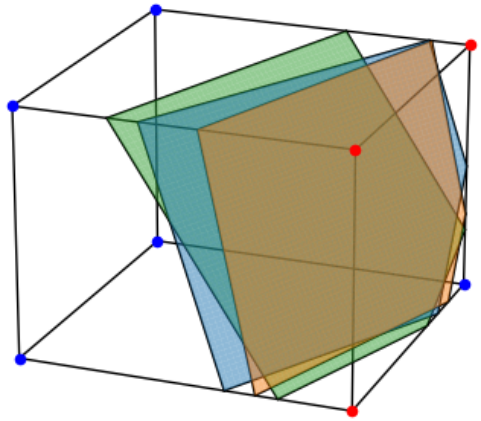


Figure 4.1. Several hyperplanes (LT estimators) leading to optimal labeling of $\{+1,-1\}^n$ for $n=3$.

when $\pi_1 < \pi_2 \leq \frac{1}{2}$, *we have*

$$\ell_{\pi_1} = \log \frac{1-\pi_1}{\pi_1} > \log \frac{1-\pi_2}{\pi_2} = \ell_{\pi_2} \geq 0.$$

Hence, $\ell_{\pi_1} - \ell_{\pi_2} > 0$ *and* $\ell_{\pi_1} + \ell_{\pi_2} > 0$. *Thus, to satisfy the almost parallel condition* (4.2), *for the threshold* $\gamma = \ell_\omega = 0$, *the optimal weights consist of vectors* $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ *with* $\alpha_1 + \alpha_2 > 0$ *and* $\alpha_1 - \alpha_2 > 0$, *i.e., any vector in* $\{\boldsymbol{\alpha} \in \mathbb{R}^2 \mid |\alpha_2| < \alpha_1\}$ *along with the threshold* $\gamma = 0$ *leads to an optimal LT estimator. Any such an optimal LT estimator announces* $S_{\mathsf{L}} = R_1$, *the output of agent* $i = 1$, *as the estimate of source symbol, and hence, the error probability would be* $\Pr(S_{\mathsf{L}}^\star \neq S) = \Pr(R_1 \neq S) = \pi_1$.

4.2.1 Upper Bound on Error Probability

In this section, we provide an upper bound for the error probability for the LT estimator as a function of the unreliability parameters $\boldsymbol{\pi}$, and the source parameter ω.

Theorem 11. *Consider a fact-checker comprised of agents with unreliability parameters* $\pi_i \in (0,1)$ *for* $i \in [n]$ *and the optimal Linear Thresholding estimator* $S_{\mathsf{L}}^\star$ *based on an optimal set of parameters* $(\boldsymbol{\alpha}^\star, \gamma^\star) \in \mathbb{R}^{n+1}$ *satisfying* (4.3). *The probability of error for* $(\boldsymbol{\alpha}^\star, \gamma^\star)$ *is upper*

bounded as

$$\Pr\left(S_{\mathsf{L}}^{\star} \neq S\right) \leq 2^{n+1} \sqrt{\omega(1-\omega) \prod_{i=1}^{n} \pi_i (1-\pi_i)}. \tag{4.4}$$

Remark 6. *The symmetry in the upper bound of error probability with respect to $\pi_i, 1-\pi_i$ is desirable because the fact-checker knows the unreliability parameters π_i. If the fact-checker flips the output of any pseudo-expert, it obtains an agent with unreliability $1-\pi_i$. Thus, π_i and $1-\pi_i$ play similar roles in the performance of the fact-checker.*

4.2.2 Majority Rule Fact-checker

A heuristic approach to the above fact-checker problem is to decide the validity of a statement based on the majority of what fact-checkers believe. Here, we provide an upper bound on the error probability of the majority-rule fact-checker.

Proposition 12. *Consider $\pi_i \in (0, 1/2)$ for $i \in [n]$ and a fact-checker with majority-based estimator $\hat{S}_{\mathsf{MJ}} := \mathsf{sgn}\left(\sum_{i=1}^{n} R_i\right)$. The probability of error, $\Pr\left(\hat{S}_{\mathsf{MJ}} \neq S\right)$, is upper bounded by*

$$\prod_{i=1}^{n} \sqrt{\pi_i(1-\pi_i)} \left[\sqrt{\prod_{j \neq i} \frac{1-\pi_j}{\pi_j}} + \sqrt{\prod_{j \neq i} \frac{\pi_j}{1-\pi_j}} \right]. \tag{4.5}$$

Remark 7. *The Arithmetic-Geometric means inequality implies $\sqrt{\prod_{j\neq i} \frac{1-\pi_j}{\pi_j}} + \sqrt{\prod_{j\neq i} \frac{\pi_j}{1-\pi_j}} \geq 2$ and $\sqrt{\omega(1-\omega)} \leq \frac{1}{2}$. Using these, it is straightforward to verify that the upper bound of error probability for the majority-rule estimator in* (4.5) *is worse than that of the optimal LT estimator in* (4.4)*.*

4.3 Proof of Main Theorems

In this section, we present the proof of the results discussed in the previous section. We begin with proving Theorem 10.

Proof of Theorem 10. Using the law of total probability, the error probability is expressed as

$$\Pr(S_{\mathsf{L}} \neq S) = \omega \Pr\left(S_{\mathsf{L}} = -1 | S = 1\right) + (1-\omega) \Pr\left(S_{\mathsf{L}} = 1 | S = -1\right).$$

Based on the definition of LT estimator, for any parameter vector $(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1}$, conditioned on the statement being true (i.e., $S = +1$), the LT estimator provides an incorrect estimate (i.e., $S_{\mathsf{L}} = -1$) when $\sum_{i=1}^{n} \alpha_i R_i < \gamma$. Therefore,

$$\Pr(S_{\mathsf{L}} = -1 \mid S = 1) = \Pr\left(\sum_{i=1}^{n} \alpha_i R_i < \gamma \middle| S = 1 \right).$$

To evaluate the latter probability, we consider all the possibilities for the received vector $\boldsymbol{R} \in \{+1, -1\}^n$ such that $\sum_{i=1}^{n} \alpha_i R_i < \gamma$, and add up the probabilities of all such disjoint events. Note that each $\boldsymbol{R} \in \{+1, -1\}^n$ corresponds to a subset $\mathcal{Q} \subseteq [n]$ where $R_i = 1$ if and only if $i \in \mathcal{Q}$. Then, we have $\sum_{i=1}^{n} \alpha_i R_i = \boldsymbol{\alpha}(\mathcal{Q}) - \boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}})$, and hence, $\boldsymbol{R}$ leads to $S_{\mathsf{L}} = -1$ if and only if $\boldsymbol{\alpha}(\mathcal{Q}) - \boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}}) < \gamma$. Therefore,

$$\begin{aligned}
&\Pr\left(\sum_{i=1}^{n} \alpha_i R_i < \gamma \middle| S = 1 \right) \\
&= \sum_{\substack{\mathcal{Q} \subseteq [n]: \\ \boldsymbol{\alpha}(\mathcal{Q}) - \boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}}) < \gamma}} \Pr\left(\{R_i = 1, \forall i \in \mathcal{Q}\} \cap \{R_i = -1, \forall i \in \mathcal{Q}^{\mathsf{c}}\} \middle| S = 1 \right) \\
&= \sum_{\mathcal{Q} \subseteq [n]} \prod_{i \in \mathcal{Q}} (1 - \pi_i) \prod_{i \in \mathcal{Q}^{\mathsf{c}}} \pi_i \cdot \mathbb{1}_{\{\boldsymbol{\alpha}(\mathcal{Q}) - \boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}}) < \gamma\}}.
\end{aligned}$$

Similarly, the conditional error probability when $S = -1$ is

$$\Pr\left(\sum_{i=1}^{n} \alpha_i R_i \geq \gamma \middle| S = -1 \right) = \sum_{\mathcal{Q} \subseteq [n]} \prod_{i \in \mathcal{Q}} \pi_i \prod_{i \in \mathcal{Q}^{\mathsf{c}}} (1 - \pi_i) \mathbb{1}_{\{\boldsymbol{\alpha}(\mathcal{Q}) - \boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}}) \geq \gamma\}}.$$

Combining the terms above and rearranging them, we get

$$\Pr\left(S_{\mathsf{L}} \neq S\right) = \sum_{\mathcal{Q}\subseteq[n]} \Bigg(\omega \prod_{i\in\mathcal{Q}} (1-\pi_i) \prod_{i\in\mathcal{Q}^{\mathsf{c}}} \pi_i \mathbb{1}_{\{\boldsymbol{\alpha}(\mathcal{Q})-\boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}})<\gamma\}} + (1-\omega) \prod_{i\in\mathcal{Q}} \pi_i \prod_{i\in\mathcal{Q}^{\mathsf{c}}} (1-\pi_i) \mathbb{1}_{\{\boldsymbol{\alpha}(\mathcal{Q})-\boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}})\geq\gamma\}} \Bigg). \tag{4.6}$$

For $\mathcal{Q} \subseteq [n]$, let us define

$$p_e(\mathcal{Q}) := \omega \prod_{i\in\mathcal{Q}} (1-\pi_i) \prod_{i\in\mathcal{Q}^{\mathsf{c}}} \pi_i \mathbb{1}_{\{\boldsymbol{\alpha}(\mathcal{Q})-\boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}})<\gamma\}} + (1-\omega) \prod_{i\in\mathcal{Q}} \pi_i \prod_{i\in\mathcal{Q}^{\mathsf{c}}} (1-\pi_i) \mathbb{1}_{\{\boldsymbol{\alpha}(\mathcal{Q})-\boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}})\geq\gamma\}}.$$

Then, the obtained expression for the probability of error (4.6) can be compactly written as $\Pr\left(S_{\mathsf{L}} \neq S\right) = \sum_{\mathcal{Q}\subseteq[n]} p_e(\mathcal{Q})$. To minimize $\Pr\left(S_{\mathsf{L}} \neq S\right)$ we use the fact that the minimum of a sum is no less than the sum of the minimum of the terms. Therefore,

$$\min_{(\boldsymbol{\alpha},\gamma)\in\mathbb{R}^{n+1}} \Pr\left(S_{\mathsf{L}} \neq S\right) \geq \sum_{\mathcal{Q}\in[n]} \min_{(\boldsymbol{\alpha},\gamma)\in\mathbb{R}^{n+1}} p_e(\mathcal{Q}).$$

Recall that $\operatorname{argmin}_{x\in\{0,1\}} ax + b(1-x) = \mathbb{1}_{\{a=\min(a,b)\}}$ for $a, b \geq 0$. Now, consider

$$a = \omega \prod_{i\in\mathcal{Q}} (1-\pi_i) \prod_{i\in\mathcal{Q}^{\mathsf{c}}} \pi_i$$

and

$$b = (1-\omega) \prod_{i\in\mathcal{Q}} \pi_i \prod_{i\in\mathcal{Q}^{\mathsf{c}}} (1-\pi_i)$$

for some fixed $\mathcal{Q} \subseteq [n]$. Since $\mathbb{1}_{\{\boldsymbol{\alpha}(\mathcal{Q})-\boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}})<\gamma\}} = 1 - \mathbb{1}_{\{\boldsymbol{\alpha}(\mathcal{Q})-\boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}})\geq\gamma\}}$, the minimizers of $p_e(\mathcal{Q})$ are those pairs of $(\boldsymbol{\alpha}, \gamma)$ in the set $\{(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^n \mid \boldsymbol{\alpha}(\mathcal{Q}) - \boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}}) \geq \gamma\}$ when

$$\frac{\omega \prod_{i\in\mathcal{Q}} (1-\pi_i) \prod_{i\in\mathcal{Q}^{\mathsf{c}}} \pi_i}{(1-\omega) \prod_{i\in\mathcal{Q}} \pi_i \prod_{i\in\mathcal{Q}^{\mathsf{c}}} (1-\pi_i)} > 1,$$

and in the set $\{(\boldsymbol{\alpha}, \gamma) \in \mathbb{R}^{n+1} \mid \boldsymbol{\alpha}(\mathcal{Q}) - \boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}}) < \gamma\}$ when

$$\frac{\omega \prod_{i\in\mathcal{Q}}(1-\pi_i)\prod_{i\in\mathcal{Q}^{\mathsf{c}}}\pi_i}{(1-\omega)\prod_{i\in\mathcal{Q}}\pi_i\prod_{i\in\mathcal{Q}^{\mathsf{c}}}(1-\pi_i)} < 1.$$

Taking the logarithm of the expression above, we get

$$\begin{aligned}\log\left(\frac{\omega \prod_{i\in\mathcal{Q}}(1-\pi_i)\prod_{i\in\mathcal{Q}^{\mathsf{c}}}\pi_i}{(1-\omega)\prod_{i\in\mathcal{Q}}\pi_i\prod_{i\in\mathcal{Q}^{\mathsf{c}}}(1-\pi_i)}\right) &= \sum_{i\in\mathcal{Q}}\log\frac{1-\pi_i}{\pi_i} - \sum_{i\in\mathcal{Q}^{\mathsf{c}}}\log\frac{1-\pi_i}{\pi_i} - \log\frac{1-\omega}{\omega} \\ &= \boldsymbol{\ell}_{\boldsymbol{\pi}}(\mathcal{Q}) - \boldsymbol{\ell}_{\boldsymbol{\pi}}(\mathcal{Q}^{\mathsf{c}}) - \ell_{\omega}.\end{aligned}$$

Thus, $\operatorname{argmin}_{(\boldsymbol{\alpha},\gamma)\in\mathbb{R}^{n+1}} p_e(\mathcal{Q})$ is given by

$$\begin{cases}\mathbb{R}^{n+1}, & \text{if } \boldsymbol{\ell}_{\boldsymbol{\pi}}(\mathcal{Q}) - \boldsymbol{\ell}_{\boldsymbol{\pi}}(\mathcal{Q}^{\mathsf{c}}) = \ell_{\omega}, \\ \{(\boldsymbol{\alpha},\gamma) | \boldsymbol{\alpha}(\mathcal{Q}) - \boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}}) \geq \gamma\}, & \text{if } \boldsymbol{\ell}_{\boldsymbol{\pi}}(\mathcal{Q}) - \boldsymbol{\ell}_{\boldsymbol{\pi}}(\mathcal{Q}^{\mathsf{c}}) > \ell_{\omega}, \\ \{(\boldsymbol{\alpha},\gamma) | \boldsymbol{\alpha}(\mathcal{Q}) - \boldsymbol{\alpha}(\mathcal{Q}^{\mathsf{c}}) < \gamma\}, & \text{if } \boldsymbol{\ell}_{\boldsymbol{\pi}}(\mathcal{Q}) - \boldsymbol{\ell}_{\boldsymbol{\pi}}(\mathcal{Q}^{\mathsf{c}}) < \ell_{\omega}.\end{cases}$$

In order to determine the minimizing weights for $\Pr(S_{\mathsf{L}} \neq S)$, it suffices to obtain $(\boldsymbol{\alpha}, \gamma)$ that minimizes $p_e(\mathcal{Q})$ for all $\mathcal{Q} \subseteq [n]$. Taking the intersection of all the minimizers of $p_e(\mathcal{Q})$ over all possible $\mathcal{Q} \subseteq [n]$, we obtain a set of $(\boldsymbol{\alpha}, \gamma)$ that minimizes the error probability $\Pr(S_{\mathsf{L}} \neq S)$ to be $\mathcal{A}$ given in (4.3). Note that $\mathcal{A}$ is a non-empty set since $(\boldsymbol{\ell}_{\boldsymbol{\pi}}, \ell_{\omega}) \widetilde{\|} (\boldsymbol{\ell}_{\boldsymbol{\pi}}, \ell_{\omega})$. ■

In Theorem 10, when we set $(\boldsymbol{\alpha}, \gamma) = (\boldsymbol{\ell}_{\boldsymbol{\pi}}, \ell_{\omega})$, the resulting estimator reduces to the Maximum A Posteriori (MAP) estimator [33, Page 20] which is given by

$$\hat{S}_{\mathsf{MAP}}(\boldsymbol{r}) = \begin{cases} +1 & \frac{\Pr(S=+1|\boldsymbol{R}=\boldsymbol{r})}{\Pr(S=-1|\boldsymbol{R}=\boldsymbol{r})} \geq 1 \\ -1 & \frac{\Pr(S=+1|\boldsymbol{R}=\boldsymbol{r})}{\Pr(S=-1|\boldsymbol{R}=\boldsymbol{r})} < 1 \end{cases},$$

where

$$\frac{\Pr(S=+1 \mid \boldsymbol{R}=\boldsymbol{r})}{\Pr(S=-1 \mid \boldsymbol{R}=\boldsymbol{r})} = \frac{\omega \prod_{i\in\mathcal{Q}}(1-\pi_i)\prod_{i\in\mathcal{Q}^{\mathrm{c}}}\pi_i}{(1-\omega)\prod_{i\in\mathcal{Q}}\pi_i \prod_{i\in\mathcal{Q}^{\mathrm{c}}}(1-\pi_i)}$$

with $\mathcal{Q} = \{i \mid r_i = +1\}$.

Next, we prove the upper bound on the error probability of the optimal LT estimator stated in Theorem 11.

Proof of Theorem 11: To determine $\Pr\left(S_{\mathsf{L}}^{\star} \neq S\right)$ we look at its components $\Pr\left(S_{\mathsf{L}}^{\star} \neq S|S=-1\right)$ and $\left(S_{\mathsf{L}}^{\star} \neq S|S=1\right)$. Note the error probability is identical for all $(\alpha^{\star}, \gamma^{\star})$ satisfying (4.3), and hence, we can focus on $\alpha_i^{\star} = \log\frac{1-\pi_i}{\pi_i}$ for every $i \in [n]$ and $\gamma^{\star} = \log\frac{1-\omega}{\omega}$. In the spirit of Chernoff bound, for any $\eta > 0$ we have

$$\begin{aligned}\Pr\left(S_{\mathsf{L}}^{\star} \neq S|S=-1\right) &= \Pr\left(\sum_{i=1}^{n}\alpha_i^{\star}R_i \geq \gamma \middle| S=-1\right) \\ &\leq \frac{\mathbb{E}\left[e^{\eta\sum_{i=1}^{n}\alpha_i^{\star}R_i} \mid S=-1\right]}{e^{\eta\gamma}} = \frac{\prod_{i=1}^{n}\mathbb{E}\left[e^{\eta\alpha_i^{\star}R_i} \mid S=-1\right]}{e^{\eta\gamma}},\end{aligned} \tag{4.7}$$

where the last equality follows from the fact that $\{R_i \mid i \in [n]\}$ are mutually independent conditioned on source symbol S. Then, for any $i \in [n]$, we have

$$\begin{aligned}\mathbb{E}[\exp\left(\eta\alpha_i^{\star}R_i\right) \mid S=-1] &= \pi_i e^{\eta\alpha_i^{\star}} + (1-\pi_i)e^{-\eta\alpha_i^{\star}} \\ &= \pi_i\left(\frac{1-\pi_i}{\pi_i}\right)^{\eta} + (1-\pi_i)\left(\frac{\pi_i}{1-\pi_i}\right)^{\eta} \\ &= \pi_i^{\eta}(1-\pi_i)^{1-\eta} + \pi_i^{1-\eta}(1-\pi_i)^{\eta}.\end{aligned} \tag{4.8}$$

Plugging (4.8) into (4.7), and setting $\eta = \frac{1}{2}$, we get

$$P\left(S_{\mathsf{L}}^{\star} \neq S|S=-1\right) \leq 2^n\sqrt{\frac{\omega}{1-\omega}\prod_{i=1}^{n}\pi_i(1-\pi_i)}.$$

Similarly, for $S = 1$ we arrive at

$$\Pr\left(S_{\mathsf{L}}^{\star} \neq S | S = 1\right) \leq 2^{n} \sqrt{\frac{1-\omega}{\omega} \prod_{i=1}^{n} \pi_i (1-\pi_i)}.$$

Using the total law of probability we get

$$\Pr\left(S_{\mathsf{L}}^{\star} \neq S\right) \leq 2^{n+1} \sqrt{\omega(1-\omega) \prod_{i=1}^{n} \pi_i (1-\pi_i)}. \qquad \blacksquare$$

Finally, we provide the proof for the upper bound of the Majority Rule Fact-checker as stated in Proposition 12.

Proof. Following along the proof of Theorem 11, we have

$$\begin{aligned}
\Pr\left(\hat{S}_{\mathsf{MJ}} \neq S \middle| S = -1\right) &= \Pr\left(\sum_{i=1}^{n} R_i \geq 0 \middle| S = -1\right) \\
&\leq \prod_{i=1}^{n} \mathbb{E}[\exp(\eta R_i) \mid S = -1] = \prod_{i=1}^{n} \left(\pi_i e^{\eta} + (1-\pi_i) e^{-\eta}\right),
\end{aligned}$$

for any $\eta > 0$. Setting $\eta = \frac{1}{2} \sum_{i=1}^{n} \log \frac{1-\pi_i}{\pi_i}$, we get

$$\begin{aligned}
&\Pr\left(\hat{S}_{\mathsf{MJ}} \neq S \middle| S = -1\right) \\
&\quad \leq \prod_{i=1}^{n} \sqrt{\pi_i (1-\pi_i)} \left[\sqrt{\prod\nolimits_{j \neq i} \frac{1-\pi_j}{\pi_j}} + \sqrt{\prod\nolimits_{j \neq i} \frac{\pi_j}{1-\pi_j}}\right].
\end{aligned}$$

Using a similar approach, we can derive the same bound for $\Pr\left(\hat{S}_{\mathsf{MJ}} \neq S \middle| S = 1\right)$, and consequently for $\Pr\left(\hat{S}_{\mathsf{MJ}} \neq S\right)$. $\blacksquare$

4.4 Simulations

This section provides some numerical experiment results for the Linear Threshold Estimator fact-checkers as discussed in previous sections. First, consider a fact-checker that

does *not* know the unreliability parameters of the agents but utilizes a reinforcement learning algorithm to learn them and use them for estimating the validity of the source statements.

For this, consider a fact-checker with $n = 9$ agents with unreliability parameters $\pi_i = \frac{i}{10}$ for $i \in [n]$. We generate $N = 10000$ i.i.d. labels $\{S(t)\}$ from the source with parameter $\omega = 0.2$. For $t = 1, \ldots, N$, we denote the output vector received by the fact-checker by $\boldsymbol{R}(t)$. For learning the unreliability parameters, consider an initial estimate of the parameters $\{\hat{\pi}_i(0) \mid i \in [n]\}$. Let us denote the estimate of the unreliability parameter at (the end of) iteration t by $\hat{\boldsymbol{\pi}}(t)$. At iteration $t+1$, based on the received vector $\boldsymbol{R}(t+1)$ and $\hat{\boldsymbol{\pi}}(t)$, we estimate $S(t+1)$ as $S_{\mathsf{L},t+1}$, which the output of an LT estimator with parameters $(\boldsymbol{\ell}_{\hat{\boldsymbol{\pi}}(t)}, \ell_\omega)$ and input $\boldsymbol{R}(t+1)$. Based on the proximity between $R_i(t+1)$ and a soft version of estimate $S_{\mathsf{L},t+1}$, the algorithm adjusts the policy estimate, $\hat{\pi}_i(t)$, for each agent $i \in [n]$. In particular, we use the following reinforcement learning dynamics

$$\hat{\pi}_i(t+1) = \frac{t}{t+1}\hat{\pi}_i(t) + \frac{1}{t+1}\frac{1 - \tanh\frac{R_i(t+1)S^{\text{soft}}_{\mathsf{L},t+1}}{2}}{2}, \tag{4.9}$$

where $S^{\text{soft}}_{\mathsf{L},t+1} = \sum_{i=1}^{n} \ell_{\hat{\pi}_i(t)} R_i(t+1) - \ell_\omega$.

In Fig. 4.2 (top), we plot the true channel parameters $\boldsymbol{\pi}$ and the estimate $\hat{\boldsymbol{\pi}}(t)$ as a function of iteration t. It can be seen that the estimates approach the true parameters as the number of iterations grows. Since the estimate of the unreliability parameter $\{\hat{\boldsymbol{\pi}}(t)\}$ appears to converge to the true parameter $\boldsymbol{\pi}$, we expect the estimate $S_{\mathsf{L},t}$ to mimic the optimum estimator $S^\star_{\mathsf{L}}$. Fig. 4.2 (bottom) illustrates the total number of mismatches between $S_{\mathsf{L},t}$ and $S^\star_{\mathsf{L}}$ at iteration t. As it is shown there, this error becomes less frequent as the number of iterations t grows, leading to only 17 mismatches in 10000 iterations.

The non-zero Borel measure of $\mathcal{A}$ allows us to have an *optimal* LT estimator in spite of having an *imperfect* estimate $\hat{\pi}$ of unreliability parameters. Define the stopping time $T = \inf\{t \geq 1 \mid (\boldsymbol{\ell}_{\hat{\boldsymbol{\pi}}(\boldsymbol{t})}, \ell_\omega) \widetilde{\|} (\boldsymbol{\ell}_{\boldsymbol{\pi}}, \ell_\omega)\}$. Fig. 4.3 shows our algorithm's histogram of T for 1000 sample paths (batches). There, $N(t)$ represents the number of runs in which the almost parallel

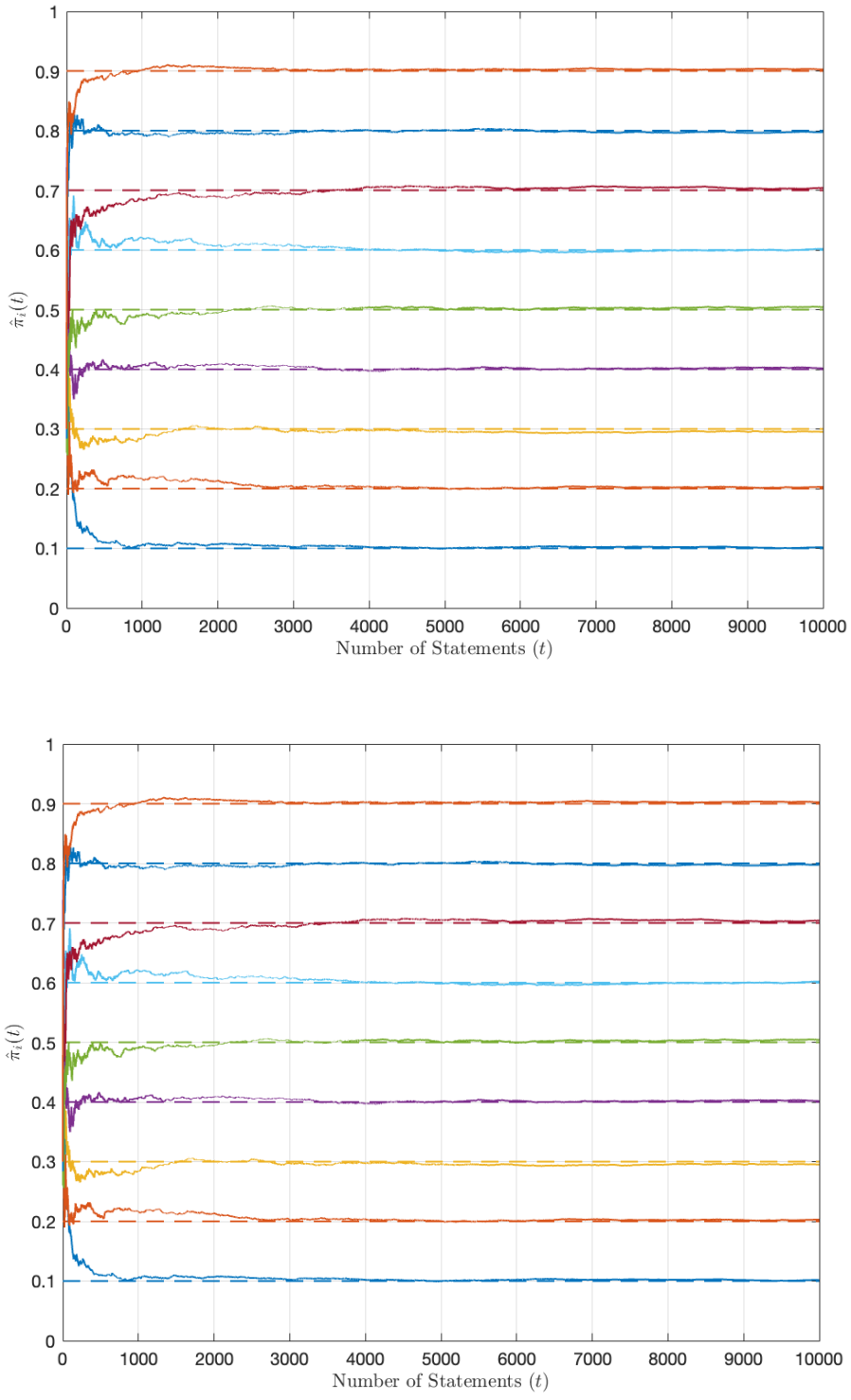


Figure 4.2. Top: Learning $\hat{\pi}_i(t)$ using (4.9) as a function of t. The horizontal lines correspond to each agent's true unreliability parameters. Bottom: The number of misclassified labels vs the number of received statements.

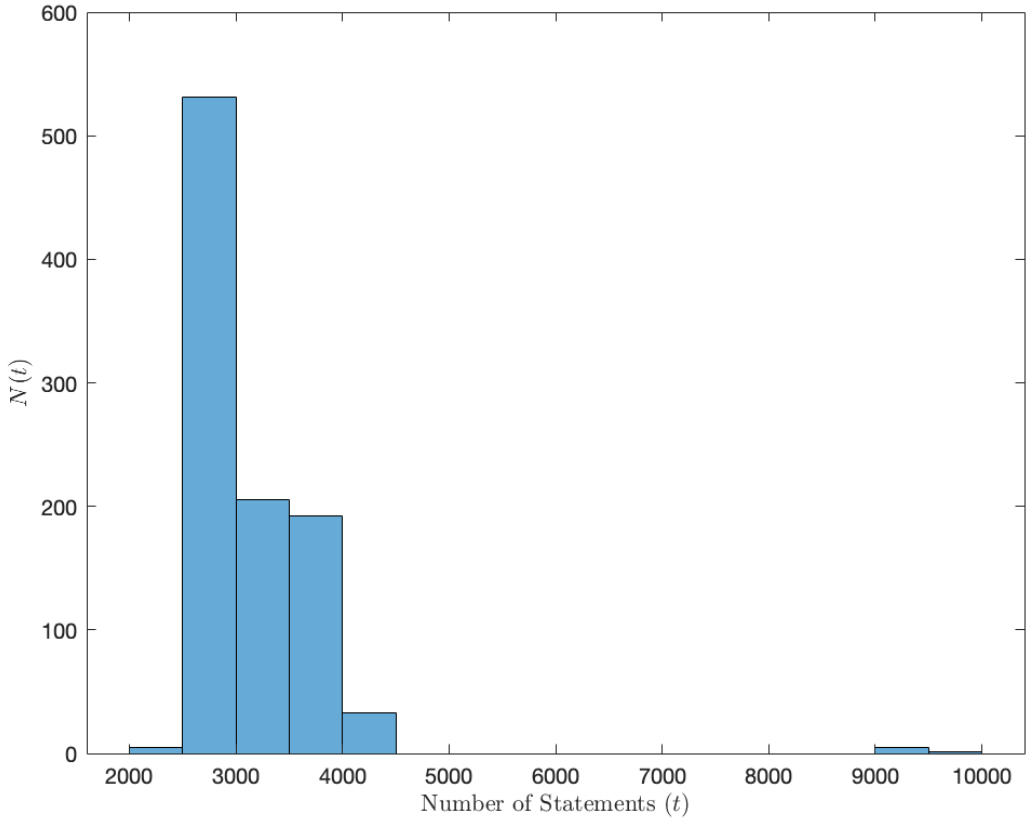


Figure 4.3. Histogram for stopping-time T

condition holds for $T = t$.

Chapter 4, in full, is a reprint of the material as it appears in A. Verma, A. Sharbafchi, B. Touri, S. Mohajer, “Distributed Fact Checking," in 2023 *International Symposium on Information Theory*. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Two-Agent Fact Checker

In this chapter, our focus is analyzing the convergence of online estimator introduced in Chapter 3 for the two-agent fact-checker ($n = 2$). An important point to recall is that since the fact checker has access only to the samples of the output vectors $\{\boldsymbol{R}(t)\}$, unreliability parameters resulting in the same output distribution, i.e., same statistics for $\boldsymbol{R}(t)$ are indistinguishable from each other from the perspective of the fact checker. In particular, it is easy to verify that for any $\boldsymbol{\mu} = (\mu_1, \mu_2)$ satisfying $\mu_1\mu_2 + \bar{\mu}_1\bar{\mu_2} = \pi_1\pi_2 + \bar{\pi}_1\bar{\pi}_2$, the probability of observing (R_1, R_2) at the output of the channel $\boldsymbol{\mu}$ is identical to observing (R_1, R_2) at the output of $\boldsymbol{\pi}$. Hence, $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$ are not distinguishable. This leads us to the following problem statement. Therefore, we focus on the following problem for two-agent fact-checker.

Problem 13. *Consider a fact-checker with access to the sequence of the assessments of two agents, with unknown unreliability parameters π_1, π_2. Determine an online estimator for the unreliability parameters such that*

$$\lim_{t\to\infty} P_1(t)P_2(t) + \bar{P}_1(t)\bar{P}_2(t) = \pi_1\pi_2 + \bar{\pi}_1\bar{\pi}_2, \textit{ a.s.},$$

where $P_1(t), P_2(t)$ are the estimates of π_1, π_2 based on the output of the two agents up to time t.

5.1 Main Result

Two-Agent Fact Checker

We focus on the study of our proposed learning dynamics (3.8) for a two-agent system. In this case, the output distribution is determined through a single parameter $\pi_1\pi_2 + \bar{\pi}_1\bar{\pi}_2$ which is the function of unreliability parameter (vector), $\boldsymbol{\pi}$, i.e., for every $s \in \{+1, -1\}$, we have

$$\Pr(\boldsymbol{R}(t) = s\mathbf{1}) = 1 - \Pr\left(\boldsymbol{R}(t) = \begin{bmatrix} s \\ \\ -s \end{bmatrix}\right) = \frac{\pi_1\pi_2 + \bar{\pi}_1\bar{\pi}_2}{2}.$$

Define $h(a, b) := ab + \bar{a}\bar{b}$ for $a, b \in [0, 1]$. Some properties of $h(a, b)$ which will be used repeatedly throughout the chapter are (i) $h(a, b) = h(\bar{a}, \bar{b})$ and consequently $h(a, \bar{b}) = h(\bar{a}, b)$, (ii) $h(a, b) + h(a, \bar{b}) = 1$, and (iii) $0 \leq h(a, b) \leq 1$ with equality only at the corner points of $[0, 1]^2$.

Remark 8. *For the update rule for the two-agent case $n = 2$, note that if $R_1(t + 1) = R_2(t + 1)$, then*

$$\frac{1}{2}\left(\frac{L(t) - 1}{L(t) + 1}R_i(t + 1) + 1\right) = \frac{P_1(t)P_2(t)}{h(P_1(t), P_2(t))}.$$

Thus, if $R_1(t + 1) = R_2(t + 1)$, from (3.8)*, $\boldsymbol{P}(t + 1)$ will be a convex combination of the current estimate $\boldsymbol{P}(t)$ and the vector $\frac{P_1(t)P_2(t)}{h(P_1(t),P_2(t))}\mathbf{1}$. Similarly, if $R_1(t) \neq R_2(t)$, then*

$$\frac{1}{2}\eta_t\left(\frac{L(t) - 1}{L(t) + 1}R_i(t+1)+1\right) = \frac{P_i(t)P_{3-i}(t)}{h(P_1(t), \bar{P}_2(t))},$$

and thus, the estimator $\boldsymbol{P}(t + 1)$ is a convex combination of the current estimate $\boldsymbol{P}(t)$ and $\frac{1}{h(P_1(t),\bar{P}_2(t))}\begin{pmatrix} P_1(t)\bar{P}_2(t) \\ \bar{P}_1(t)P_2(t) \end{pmatrix}$.

In the following theorem, we provide a close answer to Problem 13 by proving that the

proposed online estimators converge almost surely to the desired set.

Theorem 14 (Convergence of Online Estimator)**.** *For a fact-checker comprised of two agents with unreliability parameters π_i, for $i \in [2]$, under Assumption 4 on the step-sizes, the online estimator $\{\boldsymbol{P}(t)\}$, defined in* (3.8)*, converges to the set*

$$\mathcal{E} = \{\boldsymbol{x} \in [0,1]^2 \mid h(x_1, x_2) = h(\pi_1, \pi_2)\} \cup \{(0.5, 0.5)\}\,.$$

5.2 Proof of Main Result

We prove the main result using stochastic approximation techniques. To do so, we first show that the difference in the updates of the estimates in (3.8) can be decomposed into a deterministic part and a zero-difference martingale.

Lemma 8. *For $t \in \mathbb{N}_0$, the online estimator* (3.8) *satisfies*

$$\boldsymbol{P}(t+1) = \boldsymbol{P}(t) + \eta_t \left(\boldsymbol{f}(\boldsymbol{P}(t)) + \boldsymbol{M}(t+1)\right), \tag{5.1}$$

where the function $\boldsymbol{f} : (0,1)^2 \to (0,1)^2$ has coordinates

$$f_i(\boldsymbol{x}) := \frac{h(\pi_1, \pi_2) - h(x_1, x_2)}{h(x_1, x_2)(1 - h(x_1, x_2))} x_i \bar{x}_i (x_{3-i} - \bar{x}_{3-i}), \tag{5.2}$$

and the sequence $\{\boldsymbol{M}(t)\}$ is a bounded martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}_{t \geq 0}$, where $\mathcal{F}_t = \sigma(P(k), M(k) : k \leq t)$, i.e., for all $t \in \mathbb{N}_0$ we have $\mathbb{E}[\boldsymbol{M}(t+1) \mid \mathcal{F}_t] = 0$ and $\|\boldsymbol{M}(t+1)\|_\infty \leq 2$ almost surely (a.s.) (and hence, bounded in-expectation).

Proof. Let

$$\boldsymbol{M}(t+1) := \frac{1}{\eta_t}(\boldsymbol{P}(t+1) - \boldsymbol{P}(t)) - \boldsymbol{f}(\boldsymbol{P}(t)).$$

To prove the claim, first note that

$$\eta_t \mathbb{E}[\boldsymbol{M}(t+1) \mid \mathcal{F}_t] = \mathbb{E}[\boldsymbol{P}(t+1) - \boldsymbol{P}(t) | \mathcal{F}_t] - \eta_t f(\boldsymbol{P}(t)),$$

and thus, we need to show $\mathbb{E}[\boldsymbol{P}(t+1) - \boldsymbol{P}(t) | \mathcal{F}_t] = \eta_t f(\boldsymbol{P}(t))$ almost surely. From Remark 8, we know that whenever $R_1(t+1) = R_2(t+1)$, we have

$$P_i(t+1) - (1-\eta_t) P_i(t) = \eta_t \frac{P_1(t) P_2(t)}{h(P_1(t), P_2(t))},$$

for $i \in [2]$, whereas when $R_1(t+1) \neq R_2(t+1)$, we have

$$P_i(t+1) - (1-\eta_t) P_i(t) = \eta_t \frac{P_i(t) \bar{P}_{3-i}(t)}{h(P_1(t), \bar{P}_2(t))}.$$

Note that $\Pr(R_1(t) = R_2(t)) = \pi_1 \pi_2 + \bar{\pi}_1 \bar{\pi}_2 = h(\pi_1, \pi_2)$ and $\Pr(R_1(t) \neq R_2(t)) = h(\pi_1, \bar{\pi}_2)$ for all $t \in \mathbb{N}_0$. Thus,

$$\begin{aligned}
&\mathbb{E}[P_i(t+1) - P_i(t) \mid \mathcal{F}_t] \\
&= \eta_t \left(h(\pi_1, \pi_2) \frac{P_1(t) P_2(t)}{h(P_1(t), P_2(t))} + h(\pi_1, \bar{\pi}_2) \frac{P_i(t) \bar{P}_{3-i}(t)}{h(P_1(t), \bar{P}_2(t))} - P_i(t) \right),
\end{aligned}$$

for all $i \in [2]$. Therefore, $\boldsymbol{f}(\boldsymbol{x})$ is given by

$$f_i(\boldsymbol{x}) = \left(h(\pi_1, \pi_2) \frac{x_1 x_2}{h(x_1, x_2)} + h(\pi_1, \bar{\pi}_2) \frac{x_i \bar{x}_{3-i}}{h(x_1, \bar{x}_2)} - x_i \right),$$

for $i \in [2]$. It is convenient to replace the last term (i.e., $-x_1$) by $-(h(\pi_1, \pi_2) + h(\pi_1, \bar{\pi}_2)) x_1$ in

$f_1(\boldsymbol{x})$ to get

$$\begin{aligned} f_1(\boldsymbol{x}) &= \bar{x}_1 x_1 (x_2 - \bar{x}_2) \frac{h(\pi_1, \pi_2)}{h(x_1, x_2)} + x_1 \bar{x}_1 (\bar{x}_2 - x_2)) \frac{h(\pi_1, \bar{\pi}_2)}{h(x_1, \bar{x}_2)} \\ &= \bar{x}_1 x_1 (x_2 - \bar{x}_2) \left(\frac{h(\pi_1, \pi_2)}{h(x_1, x_2)} - \frac{h(\pi_1, \bar{\pi}_2)}{h(x_1, \bar{x}_2)} \right). \end{aligned}$$

Similarly, we have

$$f_2(\boldsymbol{x}) = \bar{x}_2 x_2 (x_1 - \bar{x}_1) \left(\frac{h(\pi_1, \pi_2)}{h(x_1, x_2)} - \frac{h(\pi_1, \bar{\pi}_2)}{h(x_1, \bar{x}_2)} \right).$$

Note that as $h(a, b) + h(a, \bar{b}) = 1$,

$$\frac{h(\pi_1, \pi_2)}{h(x_1, x_2)} - \frac{h(\pi_1, \bar{\pi}_2)}{h(x_1, \bar{x}_2)} = \frac{h(\pi_1, \pi_2) - h(x_1, x_2)}{h(x_1, x_2) h(x_1, \bar{x}_2)}.$$

Therefore, we have

$$\boldsymbol{f}(\boldsymbol{x}) = \frac{h(\pi_1, \pi_2) - h(x_1, x_2)}{h(x_1, x_2)(1 - h(x_1, x_2))} \begin{bmatrix} x_1 \bar{x}_1 (x_2 - \bar{x}_2) \\ x_2 \bar{x}_2 (x_1 - \bar{x}_1) \end{bmatrix}. \tag{5.3}$$

Finally, we have $\frac{1}{2}\left(R_i(t+1) \frac{L(t)-1}{L(t)+1} + 1 \right) \in (0, 1)$ for $P_i(t) \in (0, 1)$. This together with $f_i(\boldsymbol{P}(t)) \in (-1, 1)$ implies that $|M_i(t+1)| \leq 2$ for $i \in [2]$. ■

The iterates of the form (5.1) are well-known as Stochastic Approximation iterates and they were first introduced in [49] to find the zeros of scalar functions and later, were used to find the zeros of vector fields (functions from $\mathbb{R}^n$ to $\mathbb{R}^n$), using noisy measurements of the vector field $\boldsymbol{f}(\boldsymbol{x})$. Therefore, the natural next step would be to identify the zeros of the particular function $\boldsymbol{f}(\boldsymbol{x})$ given in (5.3), as stated in the following lemma.

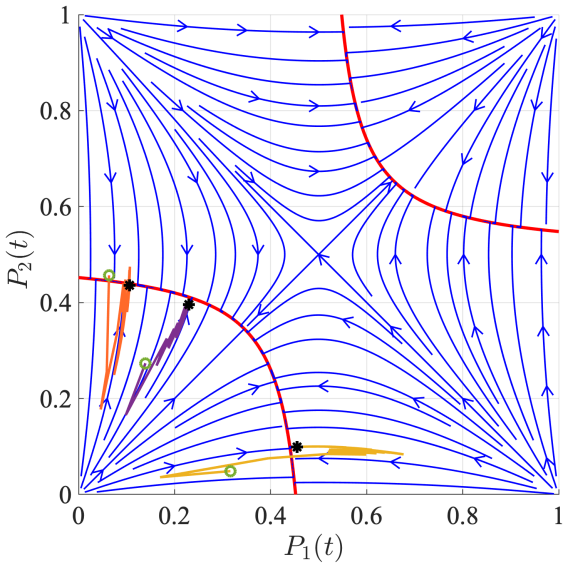


Figure 5.1. The blue with arrows and red curves represent the direction of the function $\boldsymbol{f}(\boldsymbol{x})$ at point $\boldsymbol{x}$ and level set $\{\boldsymbol{x} \mid h(\boldsymbol{x}) = h(\pi)\}$ for a $\pi = (0.32, 0.36)^T$. The other curves represent the sample paths for our estimator with different initial states (marked by o) and end states (marked by $*$).

Lemma 9. *The zero set of the function (vector-field) $\boldsymbol{f}(\boldsymbol{x})$ (defined in* (5.2)*) is the set*

$$\mathcal{E} = \{\boldsymbol{x} \in (0,1)^2 \mid h(x_1, x_2) = h(\pi_1, \pi_2)\} \cup \{(0.5, 0.5)\}. \tag{5.4}$$

Proof. Note that $\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}$ iff

$$\frac{h(\pi_1, \pi_2) - h(x_1, x_2)}{h(x_1, x_2)(1 - h(x_1, x_2))} \begin{bmatrix} x_1 \bar{x}_1 (x_2 - \bar{x}_2) \\ x_2 \bar{x}_2 (x_1 - \bar{x}_1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{5.5}$$

Therefore, $\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}$ iff either $h(\pi_1, \pi_2) - h(x_1, x_2) = 0$ or $x_1 \bar{x}_1 (x_2 - \bar{x}_2) = x_2 \bar{x}_2 (x_1 - \bar{x}_1) = 0$. For $\boldsymbol{x} \in (0,1)^2$, the second condition holds true iff $x_1 = x_2 = 0.5$. ■

Although the discrete-time process $\{\boldsymbol{P}(t)\}$ satisfies the conditions to be viewed as the Stochastic Approximation scheme for $\boldsymbol{f}(\boldsymbol{x})$, we cannot use the standard results in stochastic approximation [10, 32] to establish the convergence guarantees of our learning rule since the function $\boldsymbol{f}$ is not a Lipschitz function. We use certain potential or Lyapunov (like) functions and

their properties to show the convergence of the proposed online estimator's updates.

First, we establish a lemma showing an invariance relation for the estimates $P_1(t)$, $P_2(t)$, which we use in the analysis of the limit of the online estimator. The following lemma states that if the unreliability estimate of an agent is greater than the other, the estimates at any time for that agent will stay greater than or equal to the estimate for the other agent.

Lemma 10. *For a fact-checker with $n = 2$ agents, and for all $t \geq 0$, the online estimator $\{\boldsymbol{P}(t)\}$, given by* (3.8)*, satisfies*

$$(P_2(t) - P_1(t))(P_2(0) - P_1(0)) \geq 0.$$

Proof. It suffices to show that the sign of $P_2(t) - P_1(t)$ and $P_2(t+1) - P_1(t+1)$ always agree. If $R_1(t+1) = R_2(t+1)$, then Remark 8 implies that $\boldsymbol{P}(t+1)$ is a convex combination of $\boldsymbol{P}(t)$ and $\frac{P_1(t)P_2(t)}{h(P_1(t),P_2(t))}\mathbf{1}$, i.e., both $P_1(t)$ and $P_2(t)$ are scaled and shifted by the same amount, and hence, their order is preserved. Now, let $R_1(t+1) \neq R_2(t+1)$, and without loss of generality assume that $P_2(t) \geq P_1(t)$. Then this implies that $\bar{P}_1(t)P_2(t) \geq P_1(t)\bar{P}_2(t)$. From Remark 8 we know that $\boldsymbol{P}(t+1)$ is a convex combination of $\boldsymbol{P}(t)$ and

$$\frac{1}{h(P_1(t), \bar{P}_2(t))} \begin{pmatrix} P_1(t)\bar{P}_2(t) \\ \bar{P}_1(t)P_2(t) \end{pmatrix},$$

where in both vectors the second entry is greater than or equal is the first one. Thus, $P_2(t+1) \geq P_1(t+1)$. ■

In the following lemma, we prove the convergence of our online estimator to a superset of the zero set $\mathcal{E}$ (given by (5.4)).

Lemma 11. *Under the hypothesis of Theorem 14, the online estimator updates $\{\boldsymbol{P}(t)\}$ converges almost surely to the set*

$$\mathcal{E}_1 := \{\boldsymbol{x} \in [0,1]^2 \mid h(x_1, x_2) = h(\pi_1, \pi_2)\} \cup \left\{\boldsymbol{x} \in [0,1]^2 \mid x_1 = \frac{1}{2} \textit{ or } x_2 = \frac{1}{2}\right\}.$$

Proof. Define

$$V(t) := \left(h(P_1(t), P_2(t)) - h(\pi_1, \pi_2)\right)^2$$

for $t \geq 0$. From (5.1), we know that the update of the estimates is given by $P_i(t+1) = P_i(t) + \eta_t \tilde{f}_i(\boldsymbol{R}(t+1), \boldsymbol{P}(t))$ where

$$\tilde{f}_i(\boldsymbol{R}(t+1), \boldsymbol{P}(t)) = f_i(\boldsymbol{P}(t)) + M_i(t+1).$$

for $i \in [2]$. In order to simplify the notation, we will refer to the above expression as $\tilde{f}_i(t)$. Utilizing this expression, we can simplify the function $h(P_1(t+1), P_2(t+1))$ as

$$\begin{aligned} &h(P_1(t+1), P_2(t+1)) = h(P_1(t), P_2(t)) + 2\eta_t^2 \tilde{f}_1(t)\tilde{f}_2(t) \\ &+ \eta_t(\tilde{f}_1(t)(P_2(t) - \bar{P}_2(t)) + \tilde{f}_2(t)(P_1(t) - \bar{P}_1(t)). \end{aligned}$$

Using the fact that for all $t \in \mathbb{N}$ and $i \in [2]$, the terms $|\tilde{f}_i(t)|$ and $P_i(t)$ are bounded above by 1, for a positive constant K, we can obtain an upper bound on $V(t+1)$ as

$$\begin{aligned} &V(t+1) \leq V(t) + \eta_t^2 K + 2\eta_t(h(P_1(t), P_2(t) - h(\pi_1, \pi_2)) \\ &\times (\tilde{f}_1(t)(P_2(t) - \bar{P}_2(t)) + \tilde{f}_2(t)(P_1(t) - \bar{P}_1(t))). \end{aligned} \tag{5.6}$$

Taking the conditional expectation on the past $\mathcal{F}_t$ of the above inequality, and using the property that $\mathbb{E}[\tilde{f}_i(t) \mid \mathcal{F}_t] = f_i(\boldsymbol{P}(t))$, we get

$$\begin{aligned} &\mathbb{E}[V(t+1) \mid \mathcal{F}_t] \leq V(t) + \eta_t^2 K - 2\eta_t V(t) \times \\ &\frac{P_1(t)\bar{P}_1(t)(P_2(t) - \bar{P}_2(t))^2 + P_2(t)\bar{P}_2(t)(P_1(t) - \bar{P}_1(t))^2}{h(P_1(t), P_2(t))h(P_1(t), \bar{P}_2(t))}. \end{aligned} \tag{5.7}$$

In order to simplify the last term in (5.7) observe that

$$a\bar{a}(b-\bar{b})^2 + b\bar{b}(a-\bar{a})^2 = a\bar{a}b^2 + a\bar{a}\bar{b}^2 + b\bar{b}a^2 + b\bar{b}\bar{a}^2 - 4ab\bar{a}\bar{b}$$
$$= h(a,b)h(a,\bar{b}) - 4ab\bar{a}\bar{b}.$$

In addition, we have

$$h(a,b)h(a,\bar{b}) = (a^2+\bar{a}^2)b\bar{b} + a\bar{a}(b^2+\bar{b}^2)$$
$$\geq 2a\bar{a}b\bar{b} + a\bar{a}(b^2+\bar{b}^2) = a\bar{a}(b+\bar{b})^2 = a\bar{a}.$$

Similarly, $b\bar{b} \leq h(a,b)h(a,\bar{b})$ and hence,

$$a\bar{a}(b-\bar{b})^2 + b\bar{b}(a-\bar{a})^2 \geq h(a,b)h(a,\bar{b})(1-4h(a,b)h(a,\bar{b})).$$

Using the above inequality in (5.7), we obtain

$$\mathbb{E}[V(t+1) \mid \mathcal{F}_t] \leq V(t) + \eta_t^2 K \tag{5.8}$$
$$- 2\eta_t V(t)(1-4h(P_1(t),P_2(t))h(P_1(t),\bar{P}_2(t))).$$

Note that $1-4h(a,b)h(a,\bar{b}) = 1-4h(a,b)+4(h(a,b))^2 \geq 0$ with equality iff $h(a,b) = 1/2$, i.e., $a = \frac{1}{2}$ or $b = \frac{1}{2}$ since $2h(a,b)-1 = (1-2a)(1-2b)$. Since, $\sum_{t=0}^{\infty}\eta_t^2 < \infty$, using the Robbins-Siegmund Theorem [50] we obtain that $V(t)$ converges almost surely and

$$\sum_{t=0}^{\infty}\eta_t V(t)(1-4h(P_1(t),P_2(t))h(P_1(t),\bar{P}_2(t))) < \infty$$

almost surely, which implies that with probability one

$$\liminf_{t\to\infty} V(t)(1-4h(P_1(t),P_2(t))h(P_1(t),\bar{P}_2(t))) = 0. \tag{5.9}$$

By the above observation, the event

$$\begin{aligned}\Omega^* :=&\{\omega \in \Omega \mid V(t) \text{ converges and} \\ &\liminf_{t\to\infty} V(t)(1 - 4h(P_1(t), P_2(t))h(P_1(t), \bar{P}_2(t))) = 0\}\end{aligned}$$

happens with probability one. We can partition Ω^* into $\Omega^* = \Omega_h \cup \Omega_\pi$ where

$$\begin{aligned}\Omega_h := \{&\omega \in \Omega \mid V(t) \text{ converges and} \\ &\liminf_{t\to\infty}(1 - 4h(P_1(t), P_2(t))h(P_1(t), \bar{P}_2(t))) = 0\}\end{aligned} \tag{5.10}$$

and $\Omega_\pi = \Omega^* \setminus \Omega_h$. Note that for $\omega \in \Omega_\pi$, (5.9) implies that $\liminf_{t\to\infty} V(t) = 0$ which together with the fact that $V(t)$ converges a.s. implies that $V(t)$ converges to 0.

Since we have diminishing step-sizes, the increment in $h(P_1(t), P_2(t))$ decreases with t and the a.s. convergence of $\sqrt{V(t)} = |h(P_1(t), P_2(t)) - h(\pi_1, \pi_2)|$ implies that $h(P_1(t), P_2(t))$ converges almost surely.

To analyze the sample paths for $\omega \in \Omega_h$, recall that $1 - 4h(a, b)h(a, \bar{b}) = 0$ iff $a = 1/2$ or $b = 1/2$. Therefore, for the sample paths with

$$\liminf_{t\to\infty}(1 - 4h(P_1(t), P_2(t))h(P_1(t), \bar{P}_2(t))) = 0,$$

we know that $h(P_1(t), P_2(t))$ must converge to $1/2$. This along with the continuity of h, implies that on Ω_h, $\lim_{t\to\infty} \boldsymbol{P}(t) \in \left\{\boldsymbol{x} \in [0, 1]^2 \mid x_1 = \frac{1}{2} \text{ or } x_2 = \frac{1}{2}\right\}$. ■

With the convergence result on our online estimator, in the following lemma, we prove the convergence of sample paths corresponding to $\omega \in \Omega_h$ to the point $(1/2, 1/2)^T$ in order to establish the convergence of the updates to the zero set $\mathcal{E}$ itself instead of the superset.

Lemma 12. *Under the assumptions of Theorem 14, for $\omega \in \Omega_h$ (as defined in* (5.10)*) the online estimator updates $\{\boldsymbol{P}(t; \omega)\}$ converges to $\left(\frac{1}{2}, \frac{1}{2}\right)^T$.*

Proof. Without loss of generality assume that $P_2(0) \geq P_1(0)$, which by Lemma 10 implies that $P_2(t) \geq P_1(t)$ and hence $\Delta P(t) := P_2(t) - P_1(t) \geq 0$ for all $t \geq 0$ (surely). Since $\Pr(R_1(t+1) = R_2(t+1)) = h(\pi_1, \pi_2)$, we get

$$\begin{aligned}
\mathbb{E}[\Delta P(t+1) \mid \mathcal{F}_t] &= h(\pi_1,\pi_2)(1-\eta_t)\Delta P(t) \\
&\quad + (1 - h(\pi_1,\pi_2)) \left(1 - \eta_t + \eta_t \frac{1}{1 - h(P_1(t), P_2(t))}\right) \Delta P(t) \\
&= \Delta P(t) + \eta_t \frac{h(\pi_1,\pi_2) - h(P_1(t), P_2(t))}{h(P_1(t), P_2(t)) h(P_1(t), \bar{P}_2(t))} \\
&\quad \times (P_2(t)\bar{P}_2(t)(P_1(t) - \bar{P}_1(t)) - P_1(t)\bar{P}_1(t)(P_2(t) - \bar{P}_2(t))) \\
&= \Delta P(t) - \eta_t \Delta P(t) \frac{h(\pi_1,\pi_2) - h(P_1(t), P_2(t))}{h(P_1(t), \bar{P}_2(t))} \\
&= \Delta P(t) - \eta_t \Delta P(t) \frac{(h(\pi_1,\pi_2) - h(P_1(t), P_2(t)))^+}{h(P_1(t), \bar{P}_2(t))} \\
&\qquad + \eta_t \Delta P(t) \frac{(h(\pi_1,\pi_2) - h(P_1(t), P_2(t)))^-}{h(P_1(t), \bar{P}_2(t))},
\end{aligned}$$

where $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$. Define

$$\mathcal{C} := \{\omega \in \Omega \mid \sum_{t=0}^{\infty} \eta_t \Delta P(t) \frac{(h(\pi_1,\pi_2) - h(P_1(t), P_2(t)))^-}{h(P_1(t), \bar{P}_2(t))} < \infty\}.$$

The generalization of Robbins-Siegmund Theorem (cf. Theorem 1.3.12 in [17]) implies that for all $\omega \in \mathcal{C}$, we have

$$\sum_{t=0}^{\infty} \eta_t \Delta P(t) \frac{(h(\pi_1,\pi_2) - h(P_1(t), P_2(t)))^+}{h(P_1(t), \bar{P}_2(t))} < \infty$$

and $\Delta P(t)$ converges. Therefore, for $\omega \in \mathcal{C}$, we get

$$\sum_{t=0}^{\infty} \eta_t \Delta P(t) \frac{|h(\pi_1,\pi_2) - h(P_1(t), P_2(t))|}{h(P_1(t), \bar{P}_2(t))} < \infty.$$

On the other hand, we know that if $\omega \in \Omega_h$, then we know for any $\epsilon > 0$ there exists a time $T_\epsilon(\omega)$

such that $|h(P_1(t), P_2(t)) - 1/2| < \epsilon$ for all $t > T_\epsilon(\omega)$. Choosing $\epsilon < h(\pi_1, \pi_2) - \frac{1}{2}$ we see that

$$\liminf_{t\to\infty} h(\pi_1, \pi_2) - h(P_1(t), P_2(t)) > 0,$$

which implies that $\omega \in \mathcal{C}$, i.e., $\Omega_h \subseteq \mathcal{C}$. For $\omega \in \Omega_h$, we get that $\lim_{t\to\infty} V(t) = 0$ (since $\liminf_{t\to\infty} V(t) = 0$ from the finite-sum conclusion above and $V(t)$ converges for $\omega \in \Omega_h$) or $\lim_{t\to\infty} \Delta P(t) = 0$ (since $\liminf_{t\to\infty} \Delta P(t) = 0$ from the finite-sum conclusion above and $\Delta P(t)$ converges for $\omega \in \mathcal{C}$). By the definition of the set Ω_h for $\omega \in \Omega_h$, we know that $\lim_{t\to\infty} V(t) \neq 0$. Therefore for $\omega \in \Omega_h$, $\Delta P(t)$ converges to 0, i.e., $\boldsymbol{P}(t)$ converges to $(1/2, 1/2)$. The same arguments hold for the case when $P_2(0) \leq P_1(0)$. ■

Finally with Lemmas 11 and 12, we prove Theorem 14.

Proof of Theorem 14. Combining the results of Lemma 11 and Lemma 12, we know that for $\omega \in \Omega_\pi$ the online estimator converges to $\{\boldsymbol{x} \in [0,1]^2 \mid h(x_1, x_2) = h(\pi_1, \pi_2)\}$ and for $\omega \in \Omega_h$ the online estimator converges to $(1/2, 1/2)$. Therefore $\boldsymbol{P}(t)$ converges to $\mathcal{E}$ a.s. since $P(\Omega^*) = P(\Omega_\pi \cup \Omega_h) = 1$. ■

5.3 Conclusion

We presented a model for fact checking of binary facts involving agents modeled as memoryless binary symmetric channels and proposed an online algorithm to estimate the unreliability parameters of the agents and for $n = 2$ agents, we proved that the estimates form a dynamic process which is a stochastic approximation scheme and using results from martingale theory, we showed that it converges almost surely to the set of equilibrium points of the mean-field ODE. In particular, we characterized the zeros of the mean-field ODE, which are, interestingly, the set of unreliability parameters resulting in the same output distribution as the true unreliability parameters.

Chapter 5 in full, is a reprint of the material as it appears in A. Verma, S. Mohajer, B. Touri, “Distributed Fact Checking:Learning Unreliability," in 2024 *American Control Conference*. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Convergence in Systems with $n \geq 3$ Agents

In Section 3.2 we proposed an online estimator for the unreliability parameters of the agents which makes use of the likelihood ratio between source being fake or true given the agents' conclusion about the validity of the statement computed using the error estimate at a given time. In this chapter we study the convergence properties of the proposed online estimator involving resetting introduced in Section 3.2.1 of Chapter 3. To establish convergence, we utilize results akin to the Stochastic Approximation theorem presented in [7]. However, since the hypotheses of the theorem, specifically the assumptions on the Lyapunov function, in [7] are not met in our case, we extend the result and provide a proof tailored to our specific problem.

In the following theorem, we offer a solution to Problem 9 by demonstrating that the online estimator defined in (3.13), almost surely converge to an intended set. To articulate the theorem, recall that $h(a, b) = ab + (1 - a)(1 - b)$ for $a, b \in [0, 1]$.

Theorem 15 (Convergence of Online Estimator)**.** *For a fact-checker comprised of agents with unreliability parameters $\boldsymbol{\pi} \in \mathcal{X}$, under Assumption 4 on the step-sizes, with probability one the online estimator $\{\boldsymbol{P}_{pr}(t)\}$, defined in* (3.8)*, converges to the set $\bar{\mathcal{E}} = \mathcal{E} \cup \mathcal{G}$ where $\mathcal{E}$ is the set of equilibrium points of the mean-field ODE* (3.11)*, i.e., $\mathcal{E} = \{\boldsymbol{x} \in \mathcal{X} \mid \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}\}$ and $\mathcal{G}$ is the collection of $2n$ points $\mathcal{G} = \{\boldsymbol{x} \in \bar{\mathcal{X}} \mid x_i \in \{0, 1\}, i \in [n], x_j = h(\pi_i, x_i\pi_j + \bar{x}_i\bar{\pi}_j) \forall j \in [n]_{-i}\}$.*

In other words, $\lim_{t\to\infty} d(\boldsymbol{P}_{pr}(t), \bar{\mathcal{E}}) = 0$ *a.s.*

To understand the set $\bar{\mathcal{E}}$ to which the estimates converge, note that it is comprised of $\mathcal{E}$,

the set of zeros of the mean-field ODE $\boldsymbol{f}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\tilde{\boldsymbol{f}}(\boldsymbol{R}, \boldsymbol{x})]$ and $2n$ points at the boundary of the set $\bar{\mathcal{X}}$. We know that points $\boldsymbol{\pi}$ and $\mathbf{1} - \boldsymbol{\pi}$ are present in the set $\mathcal{E}$ and they should naturally be there as $\boldsymbol{\pi}$ and $\mathbf{1} - \boldsymbol{\pi}$ are indistinguishable from the distribution of the output vector $\boldsymbol{R}$. Moreover, we conjecture that the set $\mathcal{E}$ is comprised of these two points and the trivial point $\frac{1}{2}\mathbf{1}$.

Conjecture 1 (Characterization of $\mathcal{E}$)**.** *For $n \geq 3$, we conjecture that $\mathcal{E} = \{\boldsymbol{\pi}, \mathbf{1} - \boldsymbol{\pi}, \frac{1}{2}\mathbf{1}\}$.*

In fact for $n = 3$ we show that $\mathcal{E} = \{\boldsymbol{\pi}, \mathbf{1} - \boldsymbol{\pi}, \frac{1}{2}\mathbf{1}\}$ in Theorem 22 of Chapter 7.

On the other hand to understand the points on the boundary note that $h(\pi_i, \bar{\pi}_j)$ represents the probability of agents i and j declaring the opposite verdict regarding the validity of the statement. Given that $x_i = 0$ then $h(\pi_i, \bar{\pi}_j) = \Pr(R_j \neq R_i)$. In other words, since the fact-checker has decided that agent i is completely reliable for agent j instead of estimating $\pi_j = \Pr(R_j \neq S)$, it is estimating $h(\pi_i, \bar{\pi}_j) = \Pr(R_j \neq R_i)$. Similarly, if $x_i = 1$ the fact-checker has decided that agent i is completely unreliable, or that $1 - R_i(t)$ is the true label. Hence for $x_i = 1$, for agent j the fact-checker is estimating $\Pr(R_j \neq R_i) = \Pr(R_j = R_i) = h(\pi_i, \pi_j)$.

6.1 Proof of Main Theorem

6.1.1 Stochastic Approximation

A stochastic approximation of an Ordinary Differential Equation (ODE) is a recursive algorithm to find the zeros of a function $F : \mathcal{Z} \to \mathbb{R}^d$ from noisy observations of the function F and is commonly expressed as

$$\boldsymbol{Z}(t+1) = \boldsymbol{Z}(t) + \eta_t(F(\boldsymbol{Z}(t)) + \boldsymbol{\xi}(t+1)), \;\; t \in \mathbb{N}_0, \tag{6.1}$$

where the initial condition is $\boldsymbol{Z}(0) \in \mathcal{Z}$, $\{\eta_t\}$ is a step-size sequence that often assumed to satisfy Assumption 4. Furthermore, $\{\boldsymbol{\xi}(t)\}$ is a martingale-difference sequence, i.e., $\mathbb{E}[\boldsymbol{\xi}(t+1) \mid \mathcal{F}_t] = 0$ for all $t \in \mathbb{N}_0$, where $\mathcal{F}_t = \sigma(\boldsymbol{Z}(k), \xi(k) : k \leq t)$ is the σ-algebra generated by the past [10]. The update vectors $\boldsymbol{Z}(t)$ can be viewed as an approximation of the path taken by the so-called

mean-field ODE $\dot{\boldsymbol{z}} = \boldsymbol{F}(\boldsymbol{z})$.

To prove the main result, we first show that the difference in the updates of the estimates in (3.8) can be decomposed into a deterministic part and a zero-difference martingale.

Lemma 13. *For $t \in \mathbb{N}_0$, the online estimator* (3.8) *satisfies*

$$\boldsymbol{P}(t+1) = \boldsymbol{P}(t) + \eta_t \left(\boldsymbol{f}(\boldsymbol{P}(t)) + \boldsymbol{M}(t+1) \right), \tag{6.2}$$

where $\boldsymbol{f}$ is the mean-field ODE given in (3.11) *and $\boldsymbol{M}(t+1) := \tilde{\boldsymbol{f}}(\boldsymbol{R}(t+1), \boldsymbol{P}(t)) - \boldsymbol{f}(\boldsymbol{P}(t))$. Moreover, for the sequence $\{\boldsymbol{M}(t)\}$, for all $t \geq 0$ we have the following with probability one.*

i. $\mathbb{E}[\boldsymbol{M}(t+1) \mid \mathcal{F}_t] = 0$ *and*

ii. $\|\boldsymbol{M}(t+1)\|_\infty \leq 2$ *(and hence, bounded in-expectation),*

where the filtration $\{\mathcal{F}_t\}$ is defined as $\mathcal{F}_t = \sigma(\boldsymbol{P}(k), \boldsymbol{M}(k) : k \leq t)$ for all $t \in \mathbb{N}_0$.

In other words $\{\boldsymbol{M}(t)\}$ is a bounded martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}$.

The proof of Lemma 13 is provided in Section 6.4.1.

Although the discrete-time process $\{\boldsymbol{P}(t)\}$ satisfies the conditions to be viewed as the Stochastic Approximation scheme for $\boldsymbol{f}(\boldsymbol{x})$, we cannot use the standard results in stochastic approximation [10, 32] for convergence guarantees of our learning rule since the function $\boldsymbol{f}$ is not a Lipschitz function. We use certain Lyapunov functions and their properties to show the convergence of the proposed online estimator's updates.

In the following lemma, we provide an alternative way of expressing the probability of the output vector $g_{\boldsymbol{x}}(\boldsymbol{r})$ and the update value $\tilde{\boldsymbol{f}}(\boldsymbol{r}, \boldsymbol{x})$.

Lemma 14. *For any $\boldsymbol{r} \in \{-1, +1\}^n$ and $\boldsymbol{x} \in \mathcal{X}$, the expression of output probability for $\boldsymbol{R}$*

based on unreliability vector $\boldsymbol{x}$*,* $\Pr(\boldsymbol{R}=\boldsymbol{r};\boldsymbol{x})$*, is given as*

$$g_{\boldsymbol{x}}(\boldsymbol{r}) = \left(\frac{1}{2}\exp\frac{-\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2} + \frac{1}{2}\exp\frac{\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2}\right)\prod_{i=1}^{n}\sqrt{x_i(1-x_i)} \tag{6.3}$$
$$= \cosh\left(\frac{\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2}\right)\prod_{i=1}^{n}\sqrt{x_i(1-x_i)}.$$

Moreover, the function $\tilde{\boldsymbol{f}}(\boldsymbol{r},\boldsymbol{x})$ *can be expressed as*

$$\tilde{\boldsymbol{f}}(\boldsymbol{r},\boldsymbol{x}) = \frac{1}{2}\left(\boldsymbol{1} - \tanh\left(\frac{1}{2}\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle\right)\boldsymbol{r}\right) - \boldsymbol{x}. \tag{6.4}$$

The proof of Lemma 14 is provided in Section 6.4.2. We express the resetting based update rule (3.13) in a recursive stochastic approximation form using a correction term $\boldsymbol{\rho}(t)$ during the step of resetting. We later show that the correction term $\boldsymbol{\rho}(t)$ is non-zero finitely often in Lemma 18.

Note that the projected update rule $\{\boldsymbol{P}_{\text{pr}}(t)\}$ defined in (3.13) can also be expressed in the stochastic approximation form as

$$\boldsymbol{P}_{\text{pr}}(t+1) = \boldsymbol{P}_{\text{pr}}(t) + \eta_t[\boldsymbol{f}(\boldsymbol{P}_{\text{pr}}(t)) + \boldsymbol{M}(t+1) + \boldsymbol{\rho}(t+1)],$$

where $\boldsymbol{f}(\cdot), \boldsymbol{M}(t+1)$ are as discussed in Lemma 13, and the correction term for the event of resetting is given as

$$\boldsymbol{\rho}(t+1) := \frac{1}{\eta_t}\mathbb{1}_{\{\gamma(t+1)\neq\gamma(t)\}}\times \tag{6.5}$$
$$\left(\boldsymbol{P}_0 - (\boldsymbol{P}_{\text{pr}}(t) + \eta_t\tilde{\boldsymbol{f}}(\boldsymbol{R}(t+1), \boldsymbol{P}_{\text{pr}}(t))\right),$$

where $\boldsymbol{P}_0 \in \mathcal{K}_0$ is the reset estimate.

6.1.2 Lyapunov Function

In this section, we propose and study a Lyapunov function for the mean-field ODE (3.10), and we prove the desirable properties related to the function to establish the almost sure convergence of the projected update rule. To define the Lyapunov function, we utilize the Kullback-Leibler (KL) divergence which is defined as follows.

Definition 7 (Kullback-Leibler Divergence [44])**.** *Let μ, ν be distributions on discrete set/alphabet $\mathcal{U}$. The KL divergence between μ and ν is defined as $D_{KL}(\mu\|\nu) = \sum_{u\in\mathcal{U}} \mu(u) \log \frac{\mu(u)}{\nu(u)}$, where we use the conventions (i) $0 \cdot \log \frac{0}{0} = 0$, (ii) if there exists $u \in \mathcal{U}$ such that $\nu(u) = 0$ and $\mu(u) > 0$ then $D_{KL}(\mu\|\nu) = \infty$.*

For a Lyapunov candidate, we focus on the Kullback-Leibler divergence between the output distribution $\boldsymbol{R}$ generated by the agents having the unreliability parameters vectors $\boldsymbol{\pi}$ and $\boldsymbol{x}$. More precisely, for any $\boldsymbol{\pi} \in \mathcal{X}$ we define the Lyapunov candidate as $V : \mathcal{X} \to [0, \infty)$ with

$$V(\boldsymbol{x}) := D_{\mathrm{KL}}(g_{\boldsymbol{\pi}}\|g_{\boldsymbol{x}}). \tag{6.6}$$

Note that here the alphabet is $\mathcal{U} = \{+1, -1\}^n$. Note that this Lyapunov candidate is finite for all $\boldsymbol{x} \in \mathcal{X} = (0, 1)^n$ because $g_{\boldsymbol{x}}(\boldsymbol{r}) > 0$ for any $\boldsymbol{r} \in \{+1, -1\}^n$. Similarly, $V(\boldsymbol{x})$ is a continuously differentiable function as it is the difference between a constant and a convex combination of 2^n functions that are logarithms of polynomials $g_{\boldsymbol{x}}(\boldsymbol{r})$. More precisely,

$$V(\boldsymbol{x}) = C_{\boldsymbol{\pi}} - \sum_{\boldsymbol{r}\in\{+1,-1\}^n} g_{\boldsymbol{\pi}}(\boldsymbol{r}) \log g_{\boldsymbol{x}}(\boldsymbol{r}),$$

where $C_{\boldsymbol{\pi}} := \sum_{\boldsymbol{r}\in\{+1,-1\}^n} g_{\boldsymbol{\pi}}(\boldsymbol{r}) \log g_{\boldsymbol{\pi}}(\boldsymbol{r})$ is a finite constant for any $\boldsymbol{\pi} \in \mathcal{X}$.

Extension to $\bar{\mathcal{X}}$: Note that the Lyapunov candidate function (6.6) can be extended to the

set of singly-extreme vectors $\boldsymbol{x} \in \mathcal{X}_{\text{bound}}$. For $r \in \{+1, -1\}$ and $x \in (0, 1)$ we can rewrite

$$x^{\frac{1+r}{2}}(1-x)^{\frac{1-r}{2}} = \frac{1+r}{2}x + \frac{1-r}{2}(1-x).$$

From the above representation for $x = 0$, we adopt $0^{\frac{1+r}{2}} := \lim_{x\to 0} \frac{1+r}{2}x + \frac{1-r}{2}(1-x) = \frac{1-r}{2}$ for $r \in \{-1, +\}$. Based on the new representation, we can extend the definition of the output distribution, $g_{\boldsymbol{x}}$, for $\boldsymbol{x} \in \mathcal{X}_{\text{bound}}$. For $i \in [n]$, for a singly-extreme vector $\boldsymbol{x} \in \mathcal{X}^{(i)}_{\text{bound}}$ and vectors $\boldsymbol{r} \in \{+1, -1\}^n$, the output distribution is defined as

$$g_{\boldsymbol{x}}(\boldsymbol{r}) := \begin{cases} \prod_{k\in[n]_{-i}} x_k^{\frac{1-r_k}{2}}(1-x_k)^{\frac{1+r_k}{2}} & \text{if } r_i = (-1)^{x_i} \\ \prod_{k\in[n]_{-i}} x_k^{\frac{1+r_k}{2}}(1-x_k)^{\frac{1-r_k}{2}} & \text{if } r_i \neq (-1)^{x_i} \end{cases}. \tag{6.7}$$

For $\boldsymbol{x} \in \mathcal{X}_{\text{bound}}$ we extend the definition of $V(\boldsymbol{x}) = D_{\text{KL}}(g_{\boldsymbol{\pi}} \| g_{\boldsymbol{x}})$ using definition (6.7) for $g_{\boldsymbol{x}}$. Moreover, for $i \in [n]$, we define the gradient of $V(\boldsymbol{x})$, $\nabla V(\boldsymbol{x})$, for $\boldsymbol{x} \in \mathcal{X}^{(i)}_{\text{bound}}$, as

$$\frac{\partial V(\boldsymbol{x})}{\partial x_j} = \begin{cases} \lim_{h\to 0^+} \frac{V(\boldsymbol{x}+(-1)^{x_i}h\boldsymbol{e}_i) - V(\boldsymbol{x})}{(-1)^{x_i}h} & \text{if } j = i \\ \lim_{h\to 0} \frac{V(\boldsymbol{x}+h\boldsymbol{e}_j) - V(\boldsymbol{x})}{h} & \text{if } j \neq i \end{cases},$$

where $\boldsymbol{e}_i$ is the i-th standard basis vector of $\mathbb{R}^n$.

In the following theorem, we show that $V(\boldsymbol{x})$ satisfies the conditions to be a Lyapunov function for the ODE $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$.

Theorem 16. *For the mean-field dynamics* (3.10) *and the function* $V(\boldsymbol{x})$ *given in* (6.6)*, we have* $\langle \nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}) \rangle \leq 0$ *for all* $\boldsymbol{x} \in \mathcal{X}$*. Furthermore the equality holds iff* $\boldsymbol{x} \in \mathcal{E}$*, where* $\mathcal{E} = \{\boldsymbol{x} \in \mathcal{X} \mid \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}\}$ *is the set of equilibrium points of the ODE.*

The proof of Theorem 16 is provided in Section 6.4.2. To prove the convergence we use the Stochastic Approximation result in [7]. In [7, Theorem 2.3], it is established that under certain assumptions related to the functions $\boldsymbol{f}(\boldsymbol{x})$ and $V(\boldsymbol{x})$, if the updates stay in a compact subset $\mathcal{K}$ of

$\mathcal{X}$, and the step-sizes and error term satisfy certain boundedness conditions, then the updates converge to the set $\mathcal{K} \cap \mathcal{E}$. In doing so a key property that is being used is the compactness of the sublevel set of Lyapunov function $V(\boldsymbol{x})$ associated with the ODE $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$. However, note that the Lyapunov function $V(\boldsymbol{x}) = D_{\text{KL}}(g_{\boldsymbol{\pi}} \| g_{\boldsymbol{x}})$ defines a non-compact sublevel set for levels greater than a certain value. In other words, for large enough M $\{\boldsymbol{x} \in \mathcal{X} \mid V(\boldsymbol{x}) \leq M\}$ is not a compact set, as will be clarified as a consequence of Lemma 15 in Section 6.1.3. Consequently, if we focus only on the updates over $\mathcal{X}$, the compactness assumption essential for Theorem 2.3 in [7] would not hold. However, we can address this issue by extending the function to the set $\bar{\mathcal{X}}$ ensuring the assumptions are satisfied, as detailed in the following subsection.

6.1.3 Boundary Behavior with Extreme Unreliability

In this subsection we discuss the behavior of the Lyapunov function $V(\boldsymbol{x})$ and the mean-field function $\boldsymbol{f}(\boldsymbol{x})$ when one of the agents' estimate is at the extreme limit, *i.e.*, the unreliability vector $\boldsymbol{x} \in [0,1]^n$ is such that $x_i \in \{0,1\}$ for exactly one $i \in [n]$. In other words, we discuss the functions for $\boldsymbol{x} \in \mathcal{X}_{\text{bound}}$, where $\mathcal{X}_{\text{bound}}$ is defined in Definition 5.

A singly-extreme vector represents an unreliability parameter vector where one agent deterministically provides either the true validity or the opposite. Note that we would not have such an unreliability parameter as the true unreliability parameter of the agent since $\pi_i \in (0,1)$ for all $i \in [n]$. However, the estimates might converge to a singly-extreme point. This convergence implies that the fact-checker would solely rely on the agent with an unreliability estimate as 0 or 1 disregarding the opinions of other agents. It is worth noting that the scenario where multiple agents' estimates are in $\{0,1\}$ is not meaningful since the true unreliability for these agents lies in $(0,1)$ ensuring the existence of outputs where the agents would disagree. Such a situation would lead to an impossible estimator of the statement's validity.

To motivate the inclusion of the study of these singly-extreme points, we broaden the scope of our study by expanding the definition of the Lyapunov function V and the ODE $\boldsymbol{f}$ as the value of the functions in the limit taken along any trajectory inside $\mathcal{X}$. With these values we

show that the Lyapunov function takes finite value over $\mathcal{X}_{\text{bound}}$. Note that for $\boldsymbol{x} \in \mathcal{X}_{\text{bound}}$, the definitions of the functions for $\boldsymbol{f}(\boldsymbol{x})$ and $V(\boldsymbol{x})$ in (3.11) and (6.6) respectively remain valid and yield the expressions presented in Lemma 15.

To introduce the next lemma, let's define $H_a(x)$ for $a, x \in (0,1)$ as

$$H_a(x) := -a \log x - (1-a)\log(1-x).$$

Lemma 15. *Let $C_\pi = \mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\log g_{\boldsymbol{\pi}}(\boldsymbol{R})]$. Then for any $i \in [n]$ and $\boldsymbol{x} \in \mathcal{X}^{(i)}_{\textit{bound}}$, we have*

i. $V(\boldsymbol{x}) = C_{\boldsymbol{\pi}} + \log 2 + \sum_{k\in[n]_{-i}} H_{h(\pi_i,\pi_k)}(h(x_i,x_k))$,

ii. $\boldsymbol{f}(\boldsymbol{x})$ *is defined through* $f_i(\boldsymbol{x}) = 0$ *and*

$$f_j(\boldsymbol{x}) = h(\pi_i, h(x_i,\pi_j)) - x_j, \quad \forall j \in [n]_{-i}.$$

iii. $f_j(\boldsymbol{x}) = x_j(1-x_j)\frac{\partial V(\boldsymbol{x})}{\partial x_j}$, *for all* $j \in [n]$.

The proof of Lemma 15 is provided in Section 6.4.3.

Corollary 3. *For $\boldsymbol{x} \in \mathcal{X}_{\textit{bound}}$, $\boldsymbol{f}(\boldsymbol{x}) = 0$ iff $\boldsymbol{x} \in \mathcal{E}_{\textit{boundary}}$ where*

$$\mathcal{E}_{\textit{boundary}} := \bigcup_{i=1}^{n} \{\boldsymbol{x} \in \mathcal{X}_{\textit{bound}} \mid x_i \in \{0,1\}, x_j = h(\pi_i, x_i\pi_j + \bar{x}_i\bar{\pi}_j) \forall j \in [n]_{-i}\}. \tag{6.8}$$

Since the function $H_a(x)$ is minimized at $x = a$, we have $H_a(x) \geq H_a(a)$. Additionally, $H_a(x)$ is an unbounded function of x. Therefore,

$$\begin{aligned} V(\mathcal{X}^{(i)}_{\text{bound}}) &= \{V(\boldsymbol{x}) \mid \boldsymbol{x} \in \bar{\mathcal{X}}, x_i \in \{0,1\}\} \\ &= (C_{\boldsymbol{\pi}} + 1 + \sum_{k\in[n]_{-i}} H_{h(\pi_i,\pi_k)}(h(\pi_i,\pi_k)), \infty). \end{aligned}$$

Note that that $\mathcal{W}_M = \{\boldsymbol{x} \in \mathcal{X} \mid V(\boldsymbol{x}) \leq M\} \subset \mathcal{X} = (0,1)^n$ is not closed for any level

$$M > M_{\min} := \min_{i \in [n]} C_{\boldsymbol{\pi}} + \log 2 + \sum_{k \in [n]_{-i}} H_{h(\pi_i, \pi_k)}(h(\pi_i, \pi_k)),$$

as for such an M, $V(\boldsymbol{x}) < M$ for some $\boldsymbol{x} \in \mathcal{X}^{(i)}_{\text{bound}}$ for some $i \in [n]$ but as a sublevel set in $\mathcal{X}$, $\boldsymbol{x} \notin \mathcal{W}_M$.

We conjecture that there exists a level set of the Lyapunov function which contains the zeros of the derivative of the Lyapunov function along the trajectory of the ODE $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$.

Conjecture 2. *For the Lyapunov function $V(\boldsymbol{x}) = D_{KL}(g_{\boldsymbol{\pi}} \| g_{\boldsymbol{x}})$, there exists a positive constant $M_0 > 0$ such that*

$$\mathcal{E} = \{\boldsymbol{x} \in \mathcal{X} \mid \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}\} \subseteq \{\boldsymbol{x} \in \mathcal{X} \mid V(\boldsymbol{x}) < M_0\}. \tag{6.9}$$

Note that although we state (6.9) as a conjecture it immediately follows if $\mathcal{E}$ is a proper subset of $\mathcal{X}$. In fact if $\mathcal{E}$ has finitely many points then the condition (6.9) holds true. Since we show that $\mathcal{E} = \{\boldsymbol{\pi}, \boldsymbol{1} - \boldsymbol{\pi}, \frac{1}{2}\boldsymbol{1}\}$ in Theorem 22 of Chapter 7, condition (6.9) is known to be satisfied for $n = 3$ agent fact-checker system.

In the following lemma, we prove that the Lyapunov function $V(\boldsymbol{x})$ satisfies certain properties with the function $\boldsymbol{f}(\boldsymbol{x})$. For any $M > 0$ we denote the sublevel set of $V(\boldsymbol{x})$ over $\bar{\mathcal{X}}$ as $\bar{\mathcal{W}}_M := \{\boldsymbol{x} \in \bar{\mathcal{X}} \mid V(\boldsymbol{x}) \leq M\}$.

Lemma 16. *The function $V : \bar{\mathcal{X}} \to [0, \infty)$, defined through* (6.6)*, for the vector-field $\boldsymbol{f} : \bar{\mathcal{X}} \to \mathbb{R}^n$ satisfies the following properties.*

i. $\langle \nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}) \rangle \leq 0$ *for any $\boldsymbol{x} \in \bar{\mathcal{X}}$;*

ii. there exists $\bar{M}_0$ such that

$$\bar{\mathcal{E}} := \mathcal{E} \cup \mathcal{E}_{boundary} \subseteq \{\boldsymbol{x} \in \bar{\mathcal{X}} \mid V(\boldsymbol{x}) < \bar{M}_0\};$$

iii. for any $\bar{M}_1 \in (\bar{M}_0, \infty)$, $\bar{\mathcal{W}}_{\bar{M}_1}$ is a compact set;

iv. the closure of $V(\bar{\mathcal{E}})$ has an empty interior.

The proof of Lemma 16 is stated in Section 6.4.3.

6.1.4 Convergence of Estimator

In the following lemma, we show that the total error accumulated through the zero-difference martingale term converges almost surely.

Lemma 17. *Consider the estimates $\{\boldsymbol{P}_{pr}(t)\}$ defined through the update rule* (3.13)*. For all $t \in \mathbb{N}_0$, let $\boldsymbol{X}(t+1) = \eta_t \boldsymbol{M}(t+1)$. The series $\|\sum_{t=1}^{\infty} \boldsymbol{X}(t)\|$ converges a.s.*

The proof of Lemma 17 is provided in Section 6.4.4. We will refer to the sample paths for which $\|\sum_{t=1}^{\infty} \boldsymbol{X}(t)\|$ converges as Ω_{conv}, i.e.,

$$\Omega_{\text{conv}} = \{\omega \in \Omega \mid \| \sum_{t=0}^{\infty} \boldsymbol{X}(t)\| \text{ converges}\}. \tag{6.10}$$

From Lemma 17 we know that $\Pr(\Omega_{\text{conv}}) = 1$. Using Lemma 17, we establish that the estimates lie in a truncation set.

Lemma 18. *For the projected update rule defined through* (3.13)*, for every sample path $\omega \in \Omega_{conv}$, there exists a index $q(\omega)$ such that $\{\boldsymbol{P}_{pr}(t;\omega)\}$ lies in $\mathcal{K}_{q(\omega)}$. Moreover, $\bar{\mathcal{E}} \subseteq \mathcal{K}_{q(\omega)}$ a.s.*

The proof of Lemma 18 is provided in Section 6.4.4. Lemma 18 establishes that with probability one, $\{\boldsymbol{\rho}(t)\}$ is non-zero finitely often implying that resetting takes place finitely often.

Theorem 17. *For the projected update rule defined through* (3.13) *consider the sequence of updates associated with the sample path $\omega \in \Omega_{conv}$. Consider a compact subset $\mathcal{K}_q(\omega)$ of $\bar{\mathcal{X}}$ such that $\mathcal{K}_{q(\omega)} \cap \bar{\mathcal{E}} \neq \emptyset$. Then we have $\limsup_{t\to\infty} d(\boldsymbol{P}_{pr}(t;\omega), \mathcal{K}_{q(\omega)} \cap \bar{\mathcal{E}}) = 0$.*

The proof of Theorem 17 is provided in Section 6.4.5.

Finally, using the fact that the estimates of the projected update rule lie in a truncation set and the cumulative error term convergence almost surely, we establish the almost sure convergence result of the estimates.

Proof of Theorem 15. Through Lemma 17 we know that the error term $\sum_{t=0}^{\infty} \eta_t \boldsymbol{M}(t+1)$ converges a.s. From Lemma 18 we know that for all sample paths in the set $\omega \in \Omega_{\text{conv}}$ the sequence $\{\boldsymbol{P}_{\text{pr}}(t;\omega)\}$ stays in some compact subset $\mathcal{K}_{q(\omega)}$ of $\bar{\mathcal{X}}$. We also know that $\bar{\mathcal{E}} \subseteq \mathcal{K}_{q(\omega)}$ from Lemma 18, therefore $\mathcal{K}_{q(\omega)} \cap \bar{\mathcal{E}} = \bar{\mathcal{E}}$. So, applying Theorem 17 to every sample path in Ω_{conv}, we conclude that $d(\boldsymbol{P}_{\text{pr}}(t), \bar{\mathcal{E}})$ converges to 0 a.s. ■

6.2 Simulations

In this section we provide simulation results for both synthetic and real data, to conduct a thorough investigation into the performance of proposed estimator. In addition to analyzing the distributed fact-checking dynamics using our proposed algorithm, referred to as Soft Algorithm, we also employ the Triangular Estimation (TE) algorithm [9] as a comparative benchmark. TE algorithm estimates worker reliability by analyzing the correlations among workers' answers. At any time t we determine the estimate $\hat{S}(t)$ of the validity of the statement $S(t)$ using the current estimates of the unreliability parameters and use the LT estimator eq. (3.1). To be consistent with the plots in Chapter 4, we denote the estimates of unreliability parameter at time t by $\hat{\boldsymbol{\pi}}(t)$. For the LT estimator we use weight vector $\boldsymbol{\ell}_{\hat{\boldsymbol{\pi}}(t)}$ whose i-th coordinate is given by $\ell_{\hat{\pi}_i(t)} = \log \frac{1-\hat{\pi}_i(t)}{\hat{\pi}_i(t)}$ and threshold $\gamma = 0$.

6.2.1 Synthetic Data

In this subsection, we provide a simulation that involves a synthetic dataset comprising $n = 9$ agents. The unreliability probabilities, $\pi_i = 0.1i - 0.01$ for $i \in [n]$, incremented in steps of 0.1, representing a wide spectrum of agent behaviors. We generate $T = 10000$ i.i.d. labels $S(t)$ for $t \in [T]$ according to the Rademacher distribution. We run the proposed algorithm with

step-sizes $\eta_t = \frac{1}{t+1}$ and initialization $\pi_i(0)$ chosen independently from a uniform distribution over $(0.1, 0.4)$. In simulations, the algorithm converges even without the truncation component of the algorithm. The minimum distance of the estimates from the boundary of $(0, 1)^n$ was 0.091.

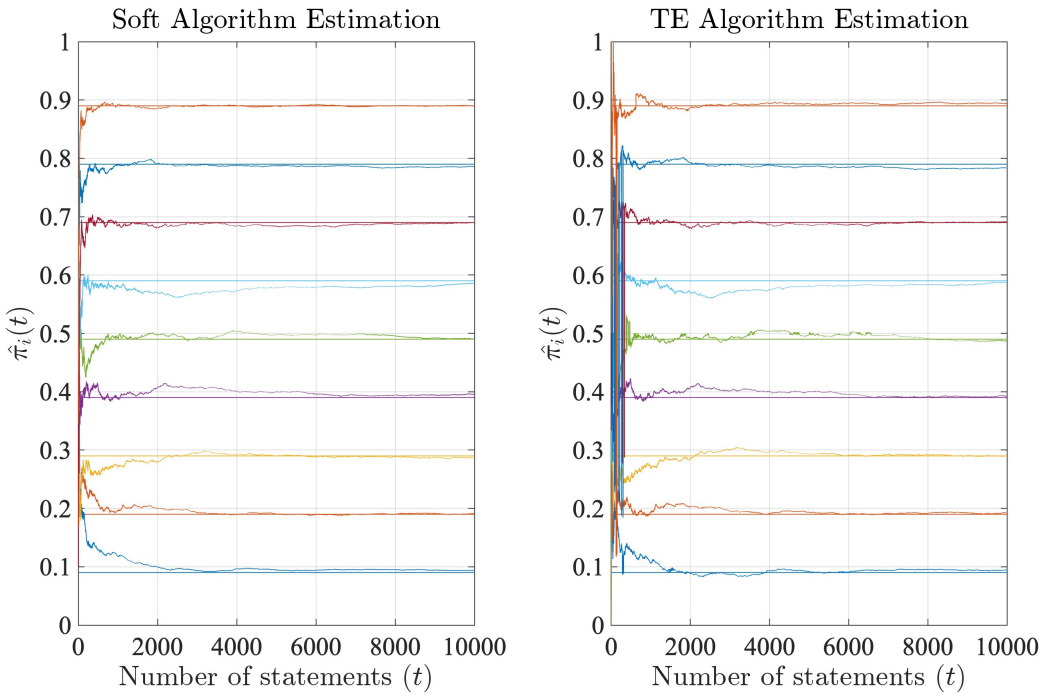


Figure 6.1. Convergence of proposed and TE Algorithms over 10000 Statements

Figure 6.1 illustrates the convergence behavior of the proposed algorithm and the TE algorithm in estimating the true unreliability parameters of agents. Each colored line corresponds to one of the nine agents, showing the trajectory of the estimated unreliability parameter as the algorithm processes more statements. The proposed algorithm's estimates approach the true unreliability values and maintain stability. On the other hand, the TE algorithm's performance, while ultimately converging to the true values, suggests a possible requirement for a larger dataset to stabilize its estimates. The early estimates of the TE algorithm estimation exhibit noticeable variance, indicating the potential need for a larger initial dataset for stabilization of the estimates. To highlight the behavior in Figure 6.2 we plot the convergence of the two algorithms over the

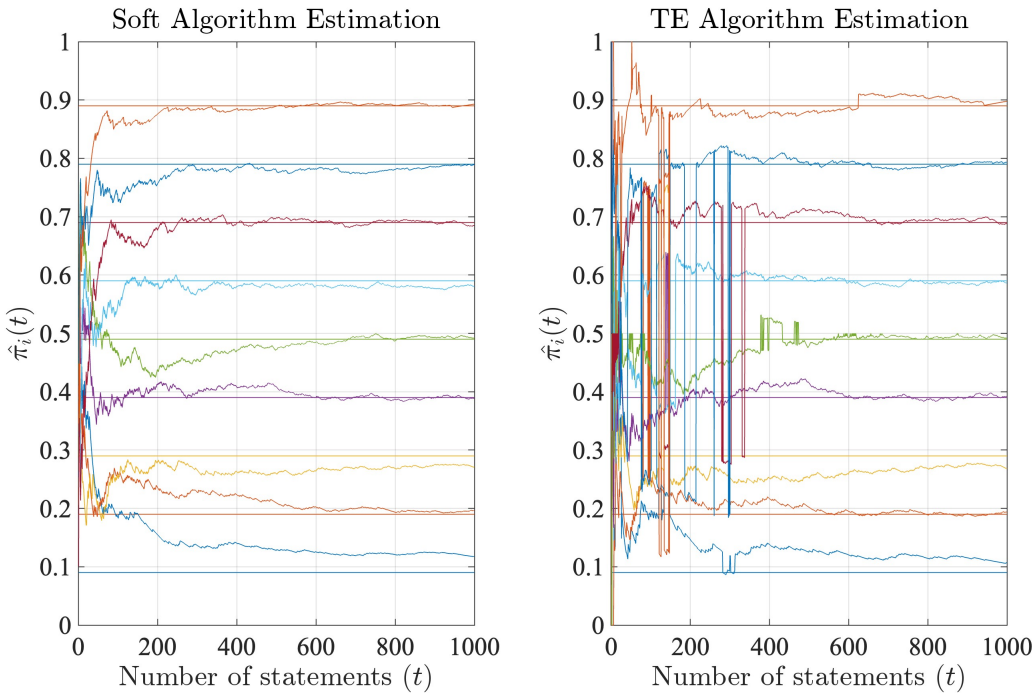


Figure 6.2. Convergence of unreliability parameters for proposed and TE Algorithms over 1000 Statements

early 1000 statements. Figure 6.3 shows a comparison of the cumulative number of mismatches between the estimates of the two algorithms. The behavior of the two algorithms is comparable in this aspect. Note that once the estimator of unreliability gets close to the true unreliability parameter, the behavior of the plot of the estimators imitates the behavior of the optimal linear threshold (LT) estimator discussed in Theorem 10.

Finally, Figure 6.4 presents the error dynamics of the proposed algorithm and the TE algorithm over the 10000 statements. As the number of statements increases, both algorithms demonstrate a decrease in the error magnitude, reflecting the convergence of the estimates to the true unreliability parameters. The figure indicates that the proposed algorithm may provide more stable convergence in estimating the unreliability parameters than the TE algorithm.

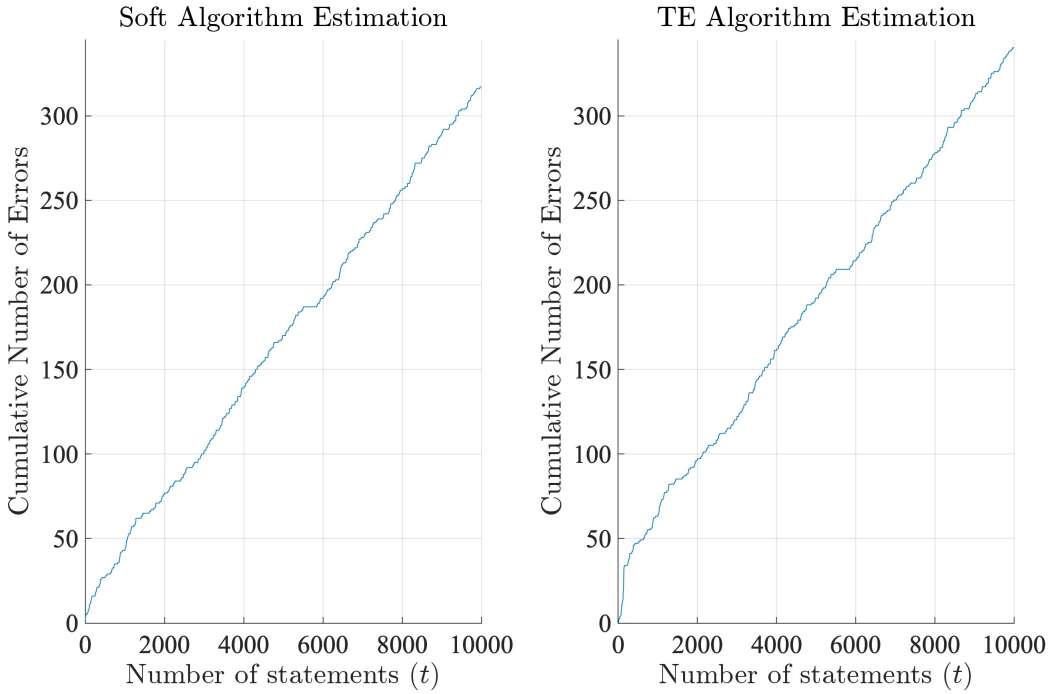


Figure 6.3. Cumulative mismatches between estimated validity $\hat{S}(t)$ and true validity $S(t)$ of proposed and TE Algorithm for synthetic dataset

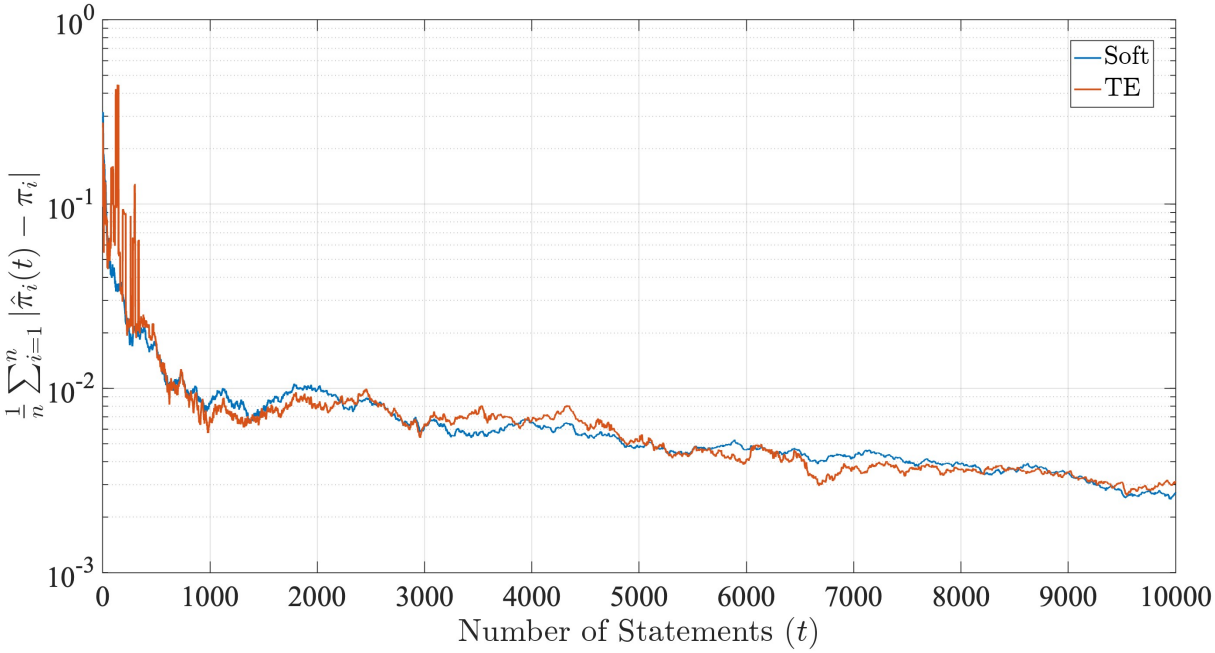


Figure 6.4. Average ℓ_1-error per agent of unreliability parameter estimates for proposed and TE Algorithm for synthetic dataset

6.2.2 Real Dataset

In this subsection we illustrate the performance of the estimators on a real-world dataset, specifically one concerning the binary classification of bird observation. This dataset, henceforth referred to as the "Blue Bird Dataset", presented in [60]. The dataset consists of a 39 number of agents and 108 statements, categorized into two distinct classes, that are labeled 1 and 2. We associate instances of label 2 with label -1, and those with label 1 with $+1$ in our setting.

The original structure of the dataset presents an ordered sequence, with all instances of label 1 positioned prior to those of label 2. Such an arrangement does not demonstrate the performance of a randomly streaming source. To address this issue we have randomized the order of the observations, thereby mitigating any potential bias that could arise from the initial ordered state. The shuffling ensures that each observation is equally likely to be sampled at any point in the analysis, adhering to the i.i.d. assumption.

In Figure 6.5 the cumulative number of mismatches is plotted for a random permutation of the data. The performance of proposed algorithm is better than the performance of TE algorithm for blue bird dataset.

6.3 Conclusion and future work

We presented a model for fact-checking of binary facts involving agents modeled as memoryless binary symmetric channels and proposed an online algorithm to estimate the unreliability parameters of the agents. We proved that the estimates form a dynamic process which is a stochastic approximation scheme and using results from stochastic approximation theory, we showed that it converges almost surely to the set of equilibrium points of the mean-field ODE over an extended domain $\bar{\mathcal{X}}$. In proving the convergence we studied the properties of the KL divergence used as the Lyapunov function $V(\boldsymbol{x})$ for mean field ODE $\boldsymbol{f}(\boldsymbol{x})$. Finally, through synthetic and real-data simulations, we showed that the proposed estimator has merits in certain cases over the streaming TE estimation algorithm. The online estimator proposed in this

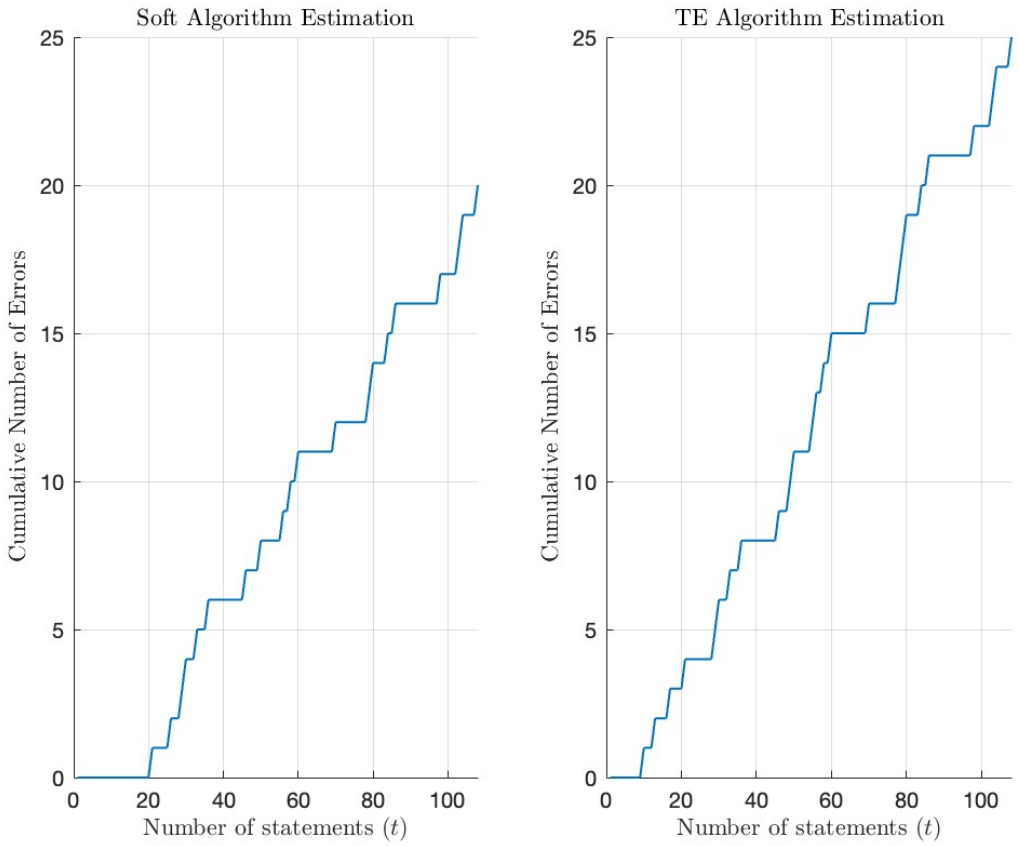


Figure 6.5. Cumulative mismatches between estimated validity $\hat{S}(t)$ and true validity $S(t)$ of proposed and TE Algorithm for Blue Bird dataset

dissertation and its analysis open up a variety of avenues for future work. We conjecture that the set to which the online estimator converges can be further reduced to a smaller set containing the stable equilibrium points such as $\boldsymbol{\pi}$ and $\mathbf{1} - \boldsymbol{\pi}$ when we have $\pi_i \neq \frac{1}{2}$ for any $i \in [n]$. Further work involves studying the convergence of variants of the proposed online estimator when not all the agents participate in the fact-checking task at every time t a new statement arrives.

6.4 Skipped Proofs

6.4.1 Proof of Stochastic Approximation Lemmas

Proof of Lemma 13. To prove the claim, first note that

$$\eta_t \mathbb{E}[\boldsymbol{M}(t+1) \mid \mathcal{F}_t] = \mathbb{E}[\boldsymbol{P}(t+1) - \boldsymbol{P}(t) | \mathcal{F}_t] - \eta_t \boldsymbol{f}(\boldsymbol{P}(t)),$$

and thus, we need to show $\mathbb{E}[\boldsymbol{P}(t+1) - \boldsymbol{P}(t) | \mathcal{F}_t] = \eta_t \boldsymbol{f}(\boldsymbol{P}(t))$ almost surely. By definition, we know that $\eta_t \tilde{\boldsymbol{f}}(\boldsymbol{R}(t+1), \boldsymbol{P}(t)) = \boldsymbol{P}(t+1) - \boldsymbol{P}(t)$ which results in the conclusion that $\boldsymbol{f}(\boldsymbol{P}(t)) = \mathbb{E}[\tilde{\boldsymbol{f}}(\boldsymbol{R}(t+1), \boldsymbol{P}(t)) \mid \mathcal{F}_t] = \mathbb{E}_{\boldsymbol{R} \sim g_{\boldsymbol{\pi}}}[\tilde{\boldsymbol{f}}(\boldsymbol{R}, \boldsymbol{P}(t))]$ since $\{\boldsymbol{R}(t)\}$ is i.i.d. according to $g_{\boldsymbol{\pi}}$.

Since $\frac{L(t)-1}{L(t)+1} R_i(t+1) \in [-1, 1]$ and $P_i(t) \in [0, 1]$, we know that the value of the coordinates satisfy $\tilde{f}_i \in [-1, 1]$. Therefore, $\|\boldsymbol{M}(t+1)\|_\infty \leq 2$ a.s. ■

6.4.2 Proof regarding Lyapunov and ODE functions

Proof of Lemma 14. For a fact-checker with unreliability vector $\boldsymbol{x} \in \mathcal{X}$, we define the log-odds value associated with agent i as $\ell_{x_i} := \log \frac{1-x_i}{x_i}$ and the log-odds vector as $\boldsymbol{\ell}_{\boldsymbol{x}} = (\ell_{x_1}, \ell_{x_2}, \ldots, \ell_{x_n})^\top$. Then the probability of the output vector being $\boldsymbol{r}$ for a set of agents with unreliability vector $\boldsymbol{x}$, $\Pr(\boldsymbol{R} = \boldsymbol{r}; \boldsymbol{x})$ is given by

$$g_{\boldsymbol{x}}(\boldsymbol{r}) = \frac{1}{2} \prod_{i=1}^{n} x_i^{\frac{1+r_i}{2}} (1 - x_i)^{\frac{1-r_i}{2}} + \frac{1}{2} \prod_{i=1}^{n} x_i^{\frac{1-r_i}{2}} (1 - x_i)^{\frac{1+r_i}{2}}. \tag{6.11}$$

Taking the exponents of the logarithm to express the probability using the log-odds vector $\boldsymbol{\ell}_{\boldsymbol{x}}$ we get

$$\begin{aligned}
g_{\boldsymbol{x}}(\boldsymbol{r}) &= \frac{1}{2}\exp\left(\sum_{i=1}^{n}\frac{1+r_i}{2}\log x_i + \frac{1-r_i}{2}\log(1-x_i)\right) \\
&\quad + \frac{1}{2}\exp\left(\sum_{i=1}^{n}\frac{1-r_i}{2}\log x_i + \frac{1+r_i}{2}\log(1-x_i)\right) \\
&= \frac{1}{2}\exp\left(\sum_{i=1}^{n}\frac{r_i}{2}\log\frac{x_i}{1-x_i} + \log x_i(1-x_i)\right) \\
&\quad + \frac{1}{2}\exp\left(\sum_{i=1}^{n}\frac{-r_i}{2}\log\frac{x_i}{1-x_i} + \log x_i(1-x_i)\right) \\
&= \left(\frac{1}{2}\exp\frac{-\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2} + \frac{1}{2}\exp\frac{\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2}\right)\prod_{i=1}^{n}\sqrt{x_i(1-x_i)}.
\end{aligned}$$

Moreover since the likelihood ratio given the unreliability vector $\boldsymbol{x}$ for output realization $\boldsymbol{r}$ is $L = \exp\left(-\langle \boldsymbol{\ell}_{\boldsymbol{x}}, \boldsymbol{r}\rangle\right)$, the function $\tilde{\boldsymbol{f}}(\boldsymbol{r}, \boldsymbol{x})$ can be rewritten as

$$\begin{aligned}
\tilde{\boldsymbol{f}}(\boldsymbol{r}, \boldsymbol{x}) &= \frac{1}{2}\left(\mathbf{1} + \frac{\exp\left(-\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle\right) - 1}{\exp\left(-\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle\right) + 1}\boldsymbol{r}\right) - \boldsymbol{x} \\
&= \frac{1}{2}\left(\mathbf{1} + \tanh\left(-\frac{1}{2}\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle\right)\boldsymbol{r}\right) - \boldsymbol{x} \\
&= \frac{1}{2}\left(\mathbf{1} - \tanh\left(\frac{1}{2}\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle\right)\boldsymbol{r}\right) - \boldsymbol{x}.
\end{aligned}$$

■

Proof of Theorem 16. Note that the KL Divergence between $g_{\boldsymbol{\pi}}$ and $g_{\boldsymbol{x}}$ can be expressed as

$$V(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\log g_{\boldsymbol{\pi}}(\boldsymbol{R})] - \mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\log g_{\boldsymbol{x}}(\boldsymbol{R})].$$

Therefore for any $i \in [n]$ the partial derivative of $V(\boldsymbol{x})$ with respect to x_i is given by

$$\frac{\partial V(\boldsymbol{x})}{\partial x_i} = -\frac{\partial \mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\log g_{\boldsymbol{x}}(\boldsymbol{R})]}{\partial x_i} = -\mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}\left[\frac{\partial \log g_{\boldsymbol{x}}(\boldsymbol{R})}{\partial x_i}\right],$$

where the second equality follows due to the finite support of random variable $\boldsymbol{R}$.

From (6.4) we know that

$$\log g_{\boldsymbol{x}}(\boldsymbol{r}) = \frac{1}{2}\sum_{i=1}^{n}\log(x_i(1-x_i)) + \log\left(\frac{1}{2}\exp\frac{-\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2} + \frac{1}{2}\exp\frac{\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2}\right). \tag{6.12}$$

Therefore the partial derivative of $\log g_{\boldsymbol{x}}(\boldsymbol{r})$ with respect to x_i can be obtained as follow.

$$\begin{aligned}
&\frac{\partial \log g_{\boldsymbol{x}}(\boldsymbol{r})}{\partial x_i} \\
&= \frac{1}{2x_i(1-x_i)}\left(1 - 2x_i - r_i\frac{\exp\frac{\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2} - \exp\frac{-\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2}}{\exp\frac{\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2} + \exp\frac{-\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2}}\right) \\
&= \frac{1}{2}\frac{1-2x_i}{x_i(1-x_i)} - \frac{r_i}{2x_i(1-x_i)}\tanh\frac{\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2} \\
&= \frac{1}{2x_i(1-x_i)}\left(1 - 2x_i - r_i\tanh\frac{\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2}\right) \\
&= \frac{1}{x_i(1-x_i)}\left(\frac{1}{2}\left(1 - r_i\tanh\frac{\langle \boldsymbol{r}, \boldsymbol{\ell}_{\boldsymbol{x}}\rangle}{2}\right) - x_i\right) \\
&= \frac{1}{x_i(1-x_i)}\tilde{f}_i(\boldsymbol{r}, \boldsymbol{x}).
\end{aligned}$$

For any $i \in [n]$, the partial derivative of V with respect to x_i is given as

$$\begin{aligned}
\frac{\partial V(\boldsymbol{x})}{\partial x_i} &= -\mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}\left[\frac{\partial \log g_{\boldsymbol{x}}(\boldsymbol{x})}{\partial x_i}\right] = -\mathbb{E}\left[\frac{1}{x_i(1-x_i)}\tilde{f}_i(\boldsymbol{R}, \boldsymbol{x})\right] \\
&= -\frac{1}{x_i(1-x_i)}f_i(\boldsymbol{x}).
\end{aligned}$$

For $\boldsymbol{x} \in \mathcal{X}$ the derivative along the trajectory of $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ is

$$\begin{aligned}
\langle \nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x})\rangle &= -\sum_{i=1}^{n}\frac{1}{x_i(1-x_i)}(f_i(\boldsymbol{x}))^2 \\
&= -\sum_{i=1}^{n} x_i(1-x_i)\left(\frac{\partial V(\boldsymbol{x})}{\partial x_i}\right)^2 \leq 0,
\end{aligned} \tag{6.13}$$

where equality holds if and only if $f_i(\boldsymbol{x}) = 0$ for all $i \in [n]$.

Therefore we have $\{\boldsymbol{x} \in \mathcal{X} \mid \langle \nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}) \rangle = 0\} = \{\boldsymbol{x} \in \mathcal{X} \mid \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}\}$.

■

6.4.3 Proof of Results on Extreme Behavior

Proof of Lemma 15. The Lyapunov function $V(\boldsymbol{x})$ is given by

$$V(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{R} \sim g_{\boldsymbol{\pi}}}[\log g_{\boldsymbol{\pi}}(\boldsymbol{R})] - \mathbb{E}_{\boldsymbol{R} \sim g_{\boldsymbol{\pi}}}[\log g_{\boldsymbol{x}}(\boldsymbol{R})].$$

For brevity, we denote the expectation as $\mathbb{E}_{\boldsymbol{R}}$.

Without loss of generality assume that $x_1 = 0$ and $x_i \in (0, 1)$ for all $i \in [n]_{-1}$. Then the probability of the output vector $\boldsymbol{R}$ is given by

$$\begin{aligned} g_{\boldsymbol{x}}(\boldsymbol{r}) &= \Pr(\{R_i = r_i, \forall i \in [n]_{-1}\} \mid S = 1 - r_1) \\ &= \frac{(1 + r_1)}{4} \prod_{i=2}^{n} x_i^{\frac{1-r_i}{2}} (1 - x_i)^{\frac{1+r_i}{2}} \\ &\quad + \frac{(1 - r_1)}{4} \prod_{i=2}^{n} x_i^{\frac{1+r_i}{2}} (1 - x_i)^{\frac{1-r_i}{2}}. \end{aligned}$$

By Theorem 4.1.13 (Law of Iterated Expectations) in [18],

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{R}}[\log g_{\boldsymbol{x}}(\boldsymbol{R})] = \mathbb{E}_{R_1}[\mathbb{E}_{\boldsymbol{R}|R_1}[\log g_{\boldsymbol{x}}(\boldsymbol{R})] \mid R_1 = +1] \\ &= \mathbb{E}_{\boldsymbol{R}|R_1=+1}[\log g_{\boldsymbol{x}}(\boldsymbol{R})] \mid R_1 = +1], \end{aligned}$$

where the last equality follows due to $g_{\boldsymbol{x}}(\boldsymbol{r}) = g_{\boldsymbol{x}}(-\boldsymbol{r})$ and $\Pr(R_1 = +1) = \frac{1}{2}$.

Therefore

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{R}|R_1=+1}[\log g_{\boldsymbol{x}}(\boldsymbol{R})] \\
&\mathbb{E}_{\boldsymbol{R}|R_1=+1}\left[\sum_{i=2}^{n}\frac{1-R_i}{2}\log x_i+\frac{1+R_i}{2}\log(1-x_i)\right]-1 \\
&=\sum_{i=2}^{n}\mathbb{E}_{\boldsymbol{R}|R_1=+1}\left[\frac{1-R_i}{2}\log x_i+\frac{1+R_i}{2}\log(1-x_i)\right]-\log 2 \\
&=-\log 2+\sum_{i=2}^{n}\frac{1}{2}\log(x_i(1-x_i))+\frac{\mathbb{E}_{\boldsymbol{R}|R_1=+1}[R_i]}{2}\log\frac{1-x_i}{x_i} \\
&=-\log 2+\sum_{i=2}^{n}\frac{\log(x_i(1-x_i))+(2h(\pi_1,\pi_i)-1])}{2}\log\frac{1-x_i}{x_i} \\
&=-\log 2+\frac{\sum_{i=2}^{n}2h(\pi_1,\pi_i)\log(1-x_i)+(2-2h(\pi_1,\pi_i)\log x_i)}{2} \\
&=-\log 2+\sum_{i=2}^{n}h(\pi_1,\pi_i)\log(1-x_i)+h(\pi_1,\bar{\pi}_i)\log x_i
\end{aligned}
$$

Finally, due to the symmetry $g_{\boldsymbol{x}}(\boldsymbol{r})=g_{1-\boldsymbol{x}}(\boldsymbol{r})$,

$$
V(\boldsymbol{x})=C_{\boldsymbol{\pi}}+\log 2-\sum_{i=2}^{n}(h(\pi_1,\pi_i)\log x_i+h(\pi_1,\bar{\pi}_i)\log(1-x_i))
$$

for any $\boldsymbol{x}\in\mathcal{X}_{\text{bound}}$ with $x_1=1$. Compactly expressed for $\boldsymbol{x}\in\mathcal{X}^{(1)}_{\text{bound}}$ we have

$$
V(\boldsymbol{x})=C_{\boldsymbol{\pi}}+\log 2+\sum_{k=2}^{n}H_{h(\pi_1,h(x_1,\pi_k))}(x_k). \tag{6.14}
$$

On the other hand, for $\boldsymbol{x}\in\mathcal{X}_{\text{bound}}$ such that $x_1=0$, by taking the limit as the argument goes to $\boldsymbol{x}$ in the set $\mathcal{X}$ we get $\boldsymbol{f}(\boldsymbol{x})=\frac{1}{2}\mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\mathbf{1}-R_1\boldsymbol{R}]-\boldsymbol{x}$. Therefore, $\tilde{f}_1(\boldsymbol{x})=0$ and for any $i\in[n]_{-1}$, we get $\frac{1}{2}\mathbb{E}[1-R_1R_i]=\Pr(R_1\neq R_i)=h(\pi_1,\bar{\pi}_i)$. Similarly for $\boldsymbol{x}\in\mathcal{X}_{\text{bound}}$ with $x_1=1$, we would have $\boldsymbol{f}(\boldsymbol{x})=\frac{1}{2}\mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\mathbf{1}+R_1\boldsymbol{R}]-\boldsymbol{x}$ and $\frac{1}{2}\mathbb{E}[1+R_1R_i]=h(\pi_1,\pi_i)$ which gives us the result.

For $\boldsymbol{x} \in \mathcal{X}^{(1)}_{\text{bound}}$, for $i \in [n]_{-1}$ we know

$$\begin{aligned}\frac{\partial V(\boldsymbol{x})}{\partial x_i} &= -\frac{h(\pi_1, h(x_1, \pi_i))}{x_i} + \frac{h(\pi_1, 1 - h(x_1, \pi_i))}{1 - x_i} \\ &= \frac{x_i - h(\pi_1, h(x_1, \pi_i))}{x_i(1 - x_i)}.\end{aligned}$$

On the other hand for $i = 1$, since $\frac{\partial V(\boldsymbol{x})}{\partial x_1}$ is finite, we have $f_1(\boldsymbol{x}) = -x_1(1-x_1)\frac{\partial V(\boldsymbol{x})}{\partial x_1} = 0$ for $\boldsymbol{x} \in \mathcal{X}^{(1)}_{\text{bound}}$.

Therefore for all $\boldsymbol{x} \in \mathcal{X}_{\text{bound}}$ we have $f_i(\boldsymbol{x}) = -x_i(1 - x_i)\frac{\partial V(\boldsymbol{x})}{\partial x_i}$ for all $i \in [n]$.

■

Proof of Lemma 16. From Theorem 16, for any $\boldsymbol{x} \in \mathcal{X}$, we know that $\langle \nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}) \rangle \leq 0$. Moreover, from Lemma 15, for $\boldsymbol{x} \in \mathcal{X}_{\text{bound}}$, we have $f_i(\boldsymbol{x}) = x_i(1 - x_i)\frac{\partial V(\boldsymbol{x})}{\partial x_i}$ for all $i \in [n]$. Therefore, (6.13) holds for $\boldsymbol{x} \in \mathcal{X}_{\text{bound}}$, $\langle \nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}) \rangle \leq 0$. Thus we have property (i).

Define the maximum value that the function $V(\boldsymbol{x})$ takes on the equilibrium points at the boundary set, $\mathcal{E}_{\text{boundary}}$ be $M_{\max} = \max\{V(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{E}_{\text{boundary}}\}$, where $\mathcal{E}_{\text{boundary}}$ is defined through (6.8)

From Conjecture 2, we know that $\mathcal{E} \subseteq \bar{\mathcal{W}}_{M_0}$. Define

$$\bar{M}_0 > \max(M_0, M_{\max}). \tag{6.15}$$

Therefore $\bar{\mathcal{E}} \subseteq \{\boldsymbol{x} \in \bar{\mathcal{X}} \mid V(\boldsymbol{x}) < \bar{M}_0\}$.

For any $C > 0$, we know that $\bar{\mathcal{W}}_C$ is a closed subset of $\bar{\mathcal{X}}$. Therefore, $\bar{\mathcal{W}}_C$ is a compact set.

Finally from Sard's theorem, we know that the closure of $V(\mathcal{E})$ has an empty interior since $\mathcal{E} = \{\boldsymbol{x} \in \mathcal{X} \mid \nabla V(\boldsymbol{x}) = 0\}$. On the other hand, $\mathcal{E}_{\text{boundary}}$ has finitely many isolated points. Therefore the closure of $V(\mathcal{E} \cup \mathcal{E}_{\text{boundary}})$ has an empty interior.

■

6.4.4 Proof of Recurrence

Proof of Lemma 17. We know that $\|\tilde{\boldsymbol{f}}(\boldsymbol{r},\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{x})\|^2 \leq 4n$, for any $\boldsymbol{r} \in \{-1,+1\}^n$ and $\boldsymbol{x} \in \mathcal{X}$. So we have

$$\mathbb{E}[\|\boldsymbol{X}(t+1)\|^2 \mid \mathcal{F}_t] \leq 4n\eta_t^2.$$

Since $\sum_{t=0}^{\infty} \eta_t^2 < \infty$, by the convergence theorem in [24, Theorem 2.17] for any coordinate $i \in [n]$, we know that $\sum_{t=0}^{\infty} X_i(t)$ converges a.s. Therefore, the series $\|\sum_{t=0}^{\infty} \boldsymbol{X}(t)\|$ converges a.s. ■

In order to prove the recurrence of estimates in a compact set, we need the following result. The following theorem ensures that after a large enough time such that the step-sizes and the accumulated error from that time forward are small enough, then the estimates stay within a sublevel set of the Lyapunov function. The following theorem is based on Theorem 2.2 in [7].

Theorem 18. *Consider the function $\boldsymbol{f} : \bar{\mathcal{X}} \to \bar{\mathcal{X}}$ and $V : \bar{\mathcal{X}} \to [0,\infty)$ defined through* (3.11) *and* (6.6) *respectively (with function definitions extended to $\bar{\mathcal{X}}$). For any $\bar{M} \in (\bar{M}_0, \bar{M}_1]$ there exist δ_0, $\lambda_0 \in \mathbb{R}^+$ such that for all $t \geq 1$, all $\boldsymbol{\theta}(0) \in \bar{\mathcal{W}}_{\bar{M}_0}$, all sequences $\{\eta_t\}$ of non-negative numbers, and all sequences $\{\boldsymbol{\xi}(t)\}$ of n-dimensional vectors satisfying*

$$\sup_{0\leq k\leq t} \eta_k \leq \lambda_0, \quad \sup_{0\leq k\leq t} \left\| \sum_{\ell=0}^{k} \eta_\ell \boldsymbol{\xi}(\ell+1) \right\| \leq \delta_0,$$

we have for $k \in [t]$, $V(\boldsymbol{\theta}(k)) \leq \bar{M}$, where $\boldsymbol{\theta}(k) = \boldsymbol{\theta}(k-1) + \eta_{k-1}(\boldsymbol{f}(\boldsymbol{\theta}(k-1)) + \boldsymbol{\xi}(k))$.

Using the above result we can prove Lemma 18.

Proof of Lemma 18. We prove the result for every sample path $\omega \in \Omega_{\text{conv}} = \{\omega \in \Omega \mid \sum_{t=0}^{\infty} \eta_t \boldsymbol{M}(t+1;\omega) < \infty\}$. For brevity, we drop the ω from the notation of random variable unless needed for clarity.

Define the supremum of the function $V(\cdot)$ over the initial set $\mathcal{K}_0$ as $C_0 := \sup\{V(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{K}_0\}$.

We know that $C_0 < \infty$ since $\mathcal{K}_0$ is a compact subset of $\bar{\mathcal{X}}$ and $V(\boldsymbol{x}) < \infty$ for all $\boldsymbol{x} \in \bar{\mathcal{X}}$.

Define $C_M = \max\{C_0, \bar{M}_0\}$, where $\bar{M}_0$ is defined through (6.15). C_M represents the constant for which the sublevel set of V contains the union of the equilibrium set $\bar{\mathcal{E}}$ and the initial truncation set $\mathcal{K}_0$. Recall that the resetting estimate $\boldsymbol{P}_0 \in \mathcal{K}_0$.

So, we have $\bar{\mathcal{E}} \cup \mathcal{K}_0 \subseteq \bar{\mathcal{W}}_{C_M}$. Recall that for any $C > 0$, $\bar{\mathcal{W}}_C$ is a compact subset of $\bar{\mathcal{X}}$ and $\cup_{t=0}^{\infty} \mathcal{K}_t = \bar{\mathcal{X}}$. So we know for any $C'_M \in (C_M, \infty)$ there exists $q \in \mathbb{N}_0$ such that $\bar{\mathcal{W}}_{C'_M} \subseteq \mathcal{K}_q$.

According to Theorem 18, there exists $\delta_0, \lambda_0 \in \mathbb{R}^+$ such that for any $t_0 \geq 0$ for all $t \geq t_0 + 1$, any $\boldsymbol{P}_{\text{pr}}(t_0) \in \mathcal{W}_{C_M}$, if we have

$$\sup_{t_0 \leq k \leq t} \eta_k \leq \lambda_0, \quad \sup_{t_0 \leq k \leq t} \left\| \sum_{\ell=t_0}^{k} \eta_\ell \boldsymbol{M}(\ell+1) \right\| \leq \delta_0, \tag{6.16}$$

then for $k \in \{t_0, \ldots, t\}$, $V(\boldsymbol{P}_{\text{pr}}(t)) \leq C'_M$.

Let us assume, contrary to the conclusion, that the resetting takes place infinitely often.

For $t \geq t_0$, define $\tau(t) := \min\{k \leq t : \gamma(k) = \gamma(t)\}$ as the most recent index at which resetting takes place.

We have $\lim_{t\to\infty} \| \sum_{k=t}^{\infty} \boldsymbol{X}(k) \| = 0$ a.s. since from Lemma 17 we know that $\| \sum_{t=0}^{\infty} \boldsymbol{X}(t) \|$ converges a.s. Therefore there exists a T_0 after which the step-sizes sequence $\{\eta_t \mid t \geq T_0\}$ and zero-difference martingale sequence $\{\boldsymbol{M}(t+1) \mid t \geq T_0\}$ satisfy (6.16).

Due to the assumption that resettings occur infinitely often, we know there exists T_1 such that $\tau(T_1 - 1) > T_0$, $T_1 = \tau(T_1)$, and $\gamma(T_1) > q$. In other words, T_1 is a time after T_0 by which two resettings have taken place. Define

$$\boldsymbol{y} = \boldsymbol{P}_{\text{pr}}(T_1 - 1) + \eta_{T_1 - 1} \tilde{\boldsymbol{f}}(\boldsymbol{R}(T_1), \boldsymbol{P}_{\text{pr}}(T_1 - 1)). \tag{6.17}$$

We know $\boldsymbol{y} \notin \mathcal{K}_q$ since $\gamma(T_1) \geq q + 1$ and $\tau(T_1) = T_1$.

Additionally $\tau(T_1 - 1)$ being a time at which resetting occurs implies that $\boldsymbol{P}_{\text{pr}}(\tau(T_1 - 1)) = \boldsymbol{P}_0 \in \mathcal{K}_0$. Since $\tau(T_1 - 1) > T_0$ we also have $\{\eta_t \mid t \geq \tau(T_1)\}$, $\{\boldsymbol{M}(t+1) \mid t \geq \tau(T_1)\}$

satisfying (6.16). From Theorem 18 we know that for $\boldsymbol{y}$ defined in (6.17) $V(\boldsymbol{y}) \leq C'_M$ which contradicts the inference that $\boldsymbol{y} \notin \mathcal{K}_q$ since $\mathcal{W}_{C'_M} \subseteq \mathcal{K}_q$. ■

6.4.5 Proof of Stochastic Approximation Result

The proof in this section closely follows the proof provided in [7]. We begin with proving the following generic properties regarding functions $\boldsymbol{f}$ and V.

Lemma 19. *Consider the functions $\boldsymbol{f} : \bar{\mathcal{X}} \to \bar{\mathcal{X}}$ and $V : \bar{\mathcal{X}} \to [0, \infty)$ defined through* (3.11) *and* (6.6) *respectively (with function definitions extended to $\bar{\mathcal{X}}$).*

i. Let $\mathcal{K} \subset \bar{\mathcal{X}}$ be a subset such that $0 < \inf_{\boldsymbol{\theta} \in \mathcal{K}} |\langle \nabla V, \boldsymbol{f} \rangle|$. For any $0 < \delta < \inf_{\boldsymbol{\theta} \in \mathcal{K}} |\langle \nabla V, \boldsymbol{f} \rangle|$, there exist $\lambda > 0$ and $\beta > 0$ such that, for any $\rho \in [0, \lambda]$, $\boldsymbol{\zeta}$, $\|\boldsymbol{\zeta}\| \leq \beta$, and $\boldsymbol{\theta} \in \mathcal{K}$, $V(\boldsymbol{\theta} + \rho \boldsymbol{f}(\boldsymbol{\theta}) + \rho \boldsymbol{\zeta}) \leq V(\boldsymbol{\theta}) - \rho\delta$.

ii. For any $\bar{M} \in (\bar{M}_0, \bar{M}_1]$, (where M_0, M_1 are defined in Lemma 16, there exist $\lambda > 0$ and $\beta > 0$ such that, for any $\rho \in [0, \lambda]$, $\boldsymbol{\zeta}$, $\|\boldsymbol{\zeta}\| \leq \beta$, and $\boldsymbol{\theta} \in \bar{\mathcal{W}}_{\bar{M}}$, $\boldsymbol{\theta} + \rho \boldsymbol{f}(\boldsymbol{\theta}) + \rho \boldsymbol{\zeta} \in \bar{\mathcal{W}}_{\bar{M}}$.

Proof. i. For any $0 < \delta < \inf_{\boldsymbol{\theta} \in \mathcal{K}} |\langle \nabla V, \boldsymbol{f} \rangle|$, there exist $\lambda > 0$ and $\beta > 0$ such that for all $\rho \in [0, \lambda]$ and $\|\boldsymbol{\zeta}\| \leq \beta$, and $t \in [0, 1]$ we have $\boldsymbol{\theta} \in \mathcal{K}$, $\boldsymbol{\theta} + \rho t \boldsymbol{f}(\boldsymbol{\theta}) + \rho t \boldsymbol{\zeta} \in \bar{\mathcal{X}}$, whose existence ensured since $\boldsymbol{f}$ will be finite over $\mathcal{K}$ when compact since $\boldsymbol{f}$ is bounded, and

$$
\begin{aligned}
&|\langle \nabla V(\boldsymbol{\theta}), \boldsymbol{f}(\boldsymbol{\theta}) \rangle - \langle \nabla V(\boldsymbol{\theta} + \rho t \boldsymbol{f}(\boldsymbol{\theta}) + \rho t \boldsymbol{\zeta}), \boldsymbol{f}(\boldsymbol{\theta}) + \boldsymbol{\zeta} \rangle \\
&\leq \inf_{\boldsymbol{\theta} \in \mathcal{K}} |\langle \nabla V, \boldsymbol{f} \rangle| - \delta.
\end{aligned}
$$

We know

$$
\begin{aligned}
&V(\boldsymbol{\theta}+\rho\boldsymbol{f}(\boldsymbol{\theta})+\rho\boldsymbol{\zeta})-V(\boldsymbol{\theta})\\
&=\rho\int_0^1\langle\nabla V(\boldsymbol{\theta}+\rho t\boldsymbol{f}(\boldsymbol{\theta})+\rho t\boldsymbol{\zeta}),\boldsymbol{f}(\boldsymbol{\theta})+\boldsymbol{\zeta}\rangle dt\\
&=\rho\langle\nabla V(\boldsymbol{\theta}),\boldsymbol{f}(\boldsymbol{\theta})\rangle\\
&+\rho\int_0^1(\langle\nabla V(\boldsymbol{\theta}+\rho t\boldsymbol{f}(\boldsymbol{\theta})+\rho t\boldsymbol{\zeta}),\boldsymbol{f}(\boldsymbol{\theta})+\boldsymbol{\zeta}\rangle-\langle\nabla V(\boldsymbol{\theta}),\boldsymbol{f}(\boldsymbol{\theta})\rangle)dt.
\end{aligned}
$$

ii. Consider $\bar{M}' \in (\bar{M}_0, \bar{M})$. There exists $\lambda_0, \beta_0 \in \mathbb{R}^+$ such that for all $\rho \in [0, \lambda_0]$ and $\|\boldsymbol{\zeta}\| \leq \beta_0$, and $\boldsymbol{\theta} \in \bar{\mathcal{W}}_{\bar{M}'}$ we have $\boldsymbol{\theta}+\rho\boldsymbol{f}(\boldsymbol{\theta})+\rho\boldsymbol{\zeta} \in \bar{\mathcal{W}}_{\bar{M}}$ since $\boldsymbol{f}$ is bounded and V is continuous.

Applying the result of $i.$ to the set

$$\mathcal{K} = \{\boldsymbol{\theta} \in \bar{\mathcal{X}} | \bar{M}' \leq V(\boldsymbol{\theta}) \leq \bar{M}\} = \mathcal{W}_{\bar{M}} \setminus \{\boldsymbol{\theta} \in \bar{\mathcal{X}} | V(\boldsymbol{\theta}) < M'\}.$$

From Lemma 16 we know that $0 < \delta < \inf_{\boldsymbol{\theta}\in\mathcal{K}} |\langle \nabla V, \boldsymbol{f}\rangle|$. Therefore there exists $\lambda_1, \beta_1 \in \mathbb{R}^+$ such that for all $\rho \in [0, \lambda_1]$ and $\|\boldsymbol{\zeta}\| \leq \beta_1$, and $\boldsymbol{\theta} \in \mathcal{K}$, we have $V(\boldsymbol{\theta}+\rho\boldsymbol{f}(\boldsymbol{\theta})+\rho\boldsymbol{\zeta}) \leq V(\boldsymbol{\theta}) \leq \bar{M}$ showing that $\boldsymbol{\theta}+\rho\boldsymbol{f}(\boldsymbol{\theta})+\rho\boldsymbol{\zeta} \in \bar{\mathcal{W}}_{\bar{M}}$.

■

Now we provide the proof for Theorem 18 following the proof of Theorem 2.2 in [7].

Proof of Theorem 18. Consider some $\bar{M}' \in (\bar{M}_0, \bar{M})$. From Lemma 19, we know that there exists $\lambda_0, \beta_0 \in \mathbb{R}^+$ such that for all $\boldsymbol{\theta}, \rho$, and $\boldsymbol{\zeta}$ satifying $V(\boldsymbol{\theta}) \leq \bar{M}'$, $\rho \in [0, \lambda_0]$, and $\|\boldsymbol{\zeta}\| \leq \beta_0$, we have

$$V(\boldsymbol{\theta}+\rho\boldsymbol{f}(\boldsymbol{\theta})+\rho\boldsymbol{\zeta}) \leq \bar{M}'.$$

By continuity of $\boldsymbol{f}$ and V there exists $\delta_0 \in (0, \beta_0]$ such that for all $\boldsymbol{\theta} \times \boldsymbol{\theta}' \in \bar{\mathcal{X}} \times \bar{\mathcal{X}}$ satisfying $V(\boldsymbol{\theta}) \leq \bar{M}$ and $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq \delta_0$, we have

$$\|\boldsymbol{f}(\boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{\theta}')\| \leq \beta_0 \text{ and } |V(\boldsymbol{\theta}) - V(\boldsymbol{\theta}')| \leq \bar{M} - \bar{M}'. \tag{6.18}$$

We will use induction to prove for all $k \in [t]$, we have $V(\boldsymbol{\theta}'(t)) \leq \bar{M}'$, and $V(\boldsymbol{\theta}(k)) \leq \bar{M}$, where the sequence $\{\boldsymbol{\theta}'(k)\}$ is defined as $\boldsymbol{\theta}'(0) = \boldsymbol{\theta}(0)$ and for all $k \in [t]$,

$$\boldsymbol{\theta}'(k) = \boldsymbol{\theta}'(k-1) + \eta_{k-1}\boldsymbol{f}(\boldsymbol{\theta}(k-1)).$$

Under the stated assumptions $V(\boldsymbol{\theta}'(0)) = V(\boldsymbol{\theta}(0)) \leq \bar{M}_0$. Since $0 \leq \eta_0 \leq \lambda_0$ and $\|\boldsymbol{\theta}'(1) - \boldsymbol{\theta}(1)\| = \|\eta_0\boldsymbol{\xi}(1)\| \leq \delta_0$, on the one hand Lemma 19 shows that $V(\boldsymbol{\theta}'(1)) = V(\boldsymbol{\theta}'(0) + \eta_0\boldsymbol{f}(\boldsymbol{\theta}(0))) \leq \bar{M}'$ and on the other hand

$$V(\boldsymbol{\theta}'(1)) = V(\boldsymbol{\theta}(0) + \eta_0\boldsymbol{f}(\boldsymbol{\theta}(0)) + \eta_0\boldsymbol{\xi}(1)) \leq \bar{M},$$

which proves the result for $t = 1$.

Assuming the result holds for $k \in [t-1]$ for $t > 1$. By construction for $j \in [k]$, $\boldsymbol{\theta}(j) - \boldsymbol{\theta}'(j) = \boldsymbol{\theta}'(j-1) - \boldsymbol{\theta}(j-1) + \eta_{j-1}\boldsymbol{\xi}(j)$, which implies that

$$\boldsymbol{\theta}(j) - \boldsymbol{\theta}'(j) = \sum_{i=1}^{j} \eta_{i-1}\boldsymbol{\xi}(i).$$

Under the stated assumptions ensuring continuity and (6.18), for $j \in [k]$, we have

$$\|\boldsymbol{\theta}(j) - \boldsymbol{\theta}'(j)\| \leq \delta_0 \text{ and } \|\boldsymbol{f}(\boldsymbol{\theta}(j)) - \boldsymbol{f}(\boldsymbol{\theta}'(j))\| \leq \beta_0.$$

On the other hand,

$$\begin{aligned}\boldsymbol{\theta}'(k+1) &= \boldsymbol{\theta}'(k) + \eta_k \boldsymbol{f}(\boldsymbol{\theta}(k)) \\ &= \boldsymbol{\theta}'(k) + \eta_k \boldsymbol{f}(\boldsymbol{\theta}'(k)) + \eta_k(\boldsymbol{f}(\boldsymbol{\theta}(k)) - \boldsymbol{f}(\boldsymbol{\theta}'(k))).\end{aligned}$$

Since $0 \leq \eta_k \leq \lambda_0$ and $V(\boldsymbol{\theta}'(k)) \leq \bar{M}'$, Lemma 19 shows $V(\boldsymbol{\theta}'(k+1)) \leq M'$. Using $\|\boldsymbol{\theta}(k+1) - \boldsymbol{\theta}'(k+1)\| \leq \delta_0$, (6.18) implies that $V(\boldsymbol{\theta}(k+1)) \leq \bar{M}$ which concludes the proof. ■

For any $A \subset \bar{\mathcal{X}}$ and $\delta > 0$ we define $A_\delta := \{\boldsymbol{\theta} \in \bar{\mathcal{X}} | d(\boldsymbol{\theta}, A) \leq \delta\}$; for any function $\phi : \bar{\mathcal{X}} \to \mathbb{R}$, we define $\|\phi\|_A := \sup_{\boldsymbol{\theta} \in A} \|\phi(\boldsymbol{\theta})\|$.

The following Lemma based on [7, Lemma 2.4] will be used in the proof of Theorem 17. We will state the lemma for any $\omega \in \Omega_{\text{conv}}$ and will drop the ω-notation from $q(\omega)$ for brevity.

Lemma 20. *Under the assumption of Theorem 17 let $\mathcal{N} \subset \bar{\mathcal{X}}$ be a neighborhood of $\bar{\mathcal{E}} \cap \mathcal{K}_q$ which satisfies $\sup_{\boldsymbol{\theta} \in \mathcal{K}_q \setminus \mathcal{N}} \langle \nabla V(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{\theta}) \rangle < 0$. There exist positive constants δ, ε, and λ (depending on the sets $\mathcal{N}$ and $\mathcal{K}_q$) such that for any $\delta' \in (0, \delta]$, $\lambda' \in (0, \lambda]$, and $\eta > 0$, one can find an integer T and a sequence $\{\hat{\boldsymbol{P}}_{pr}(j) | j \geq T\}$ satisfying*

$$\begin{aligned}&\sup_{j \geq T} \left\| \boldsymbol{P}_{pr}(j) - \hat{\boldsymbol{P}}_{pr}(j) \right\| \leq \delta', \quad \sup_{j \geq T} \eta_{j-1} \leq \lambda', \text{ and} \\ &\sup_{j \geq T} |V(\boldsymbol{P}_{pr}(j)) - V(\hat{\boldsymbol{P}}_{pr}(j))| \leq \eta,\end{aligned} \tag{6.19}$$

$$V(\hat{\boldsymbol{P}}_{pr}(j)) \leq V(\hat{\boldsymbol{P}}_{pr}(j)) - \eta_{j-1}\varepsilon + (\eta + \eta_{j-1}\varepsilon) \mathbb{1}_{\{\hat{\boldsymbol{P}}_{pr}(j-1) \in \mathcal{N}\}}), \tag{6.20}$$

for $j \geq T + 1$.

Proof. For legibility in the proof we set $\mathcal{K} = \mathcal{K}_q$. Let us choose $\delta_0 > 0$ such that the set of points in $\bar{\mathcal{X}}$ which are δ_0 away from the set $\mathcal{K}$ satisfy $\mathcal{K}_{\delta_0} \subset \bar{\mathcal{W}}_{\bar{M}_2} \subset \bar{\mathcal{X}}$, for some $\bar{M}_2 \geq \bar{M}$. The set

$\mathcal{K}_{\delta_0} \setminus \mathcal{N}$ satisfies $\sup_{\boldsymbol{\theta} \in \mathcal{K}_{\delta_0} \setminus \mathcal{N}} \langle \nabla V(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{\theta})) \rangle < 0$.

By Lemma 19, for any $\varepsilon > 0$ such that $\sup_{\boldsymbol{\theta} \in \mathcal{K}_{\delta_0} \setminus \mathcal{N}} \langle \nabla V(\boldsymbol{\theta}), \boldsymbol{f}(\boldsymbol{\theta}) \rangle < -\varepsilon$, one may choose $\lambda > 0$ and $\beta > 0$ small enough so that for any $\rho \in [0, \lambda]$ and $\|\boldsymbol{\zeta}\| \leq \beta$, and $\boldsymbol{\theta} \in \mathcal{K}_{\delta_0} \setminus \mathcal{N}$ we have

$$V(\boldsymbol{\theta} + \rho \boldsymbol{f}(\boldsymbol{\theta}) + \rho \boldsymbol{\zeta}) \leq V(\boldsymbol{\theta}) - \rho \varepsilon. \tag{6.21}$$

Note that $\boldsymbol{f}$ is bounded so $\|\boldsymbol{f}\|_{\mathcal{K}}$ is finite. So, using the uniform continuity of $\boldsymbol{f}$ on $\mathcal{K}$, for any $\eta > 0$ one may choose $\delta \in (0, \lambda \|\boldsymbol{f}\|_{\mathcal{K}})$ small enough so that for all $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \mathcal{K}_{\delta_0} \times \mathcal{K}_{\delta_0}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq \delta \leq \lambda \|\boldsymbol{f}\|_{\mathcal{K}}$,

$$\|\boldsymbol{f}(\boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{\theta}')\| \leq \beta \text{ and } |V(\boldsymbol{\theta}) - V(\boldsymbol{\theta}')| \leq \eta. \tag{6.22}$$

Under the stated conditions (regarding bounded step-size and bounded cumulative error) for all $\delta' \in (0, \delta]$ and $\lambda' \in (0, \lambda]$ there exists an integer T such that for any $t \geq T + 1$, and $\eta_t \leq \lambda'$ and $\left\| \sum_{k=T}^{t} \eta_{k-1} \boldsymbol{M}(k) \right\| \leq \delta'$.

Define recursively for $j \geq T$, the sequence $\{\hat{\boldsymbol{P}}_{\text{pr}}(j) | j \geq T\}$ as $\hat{\boldsymbol{P}}_{\text{pr}}(T) := \boldsymbol{P}_{\text{pr}}(T)$ and for $j \geq T + 1$,

$$\hat{\boldsymbol{P}}_{\text{pr}}(j) = \hat{\boldsymbol{P}}_{\text{pr}}(j-1) + \eta_{j-1} \boldsymbol{f}(\boldsymbol{P}_{\text{pr}}(j-1)).$$

By construction for $j \geq T + 1$, $\hat{\boldsymbol{P}}_{\text{pr}}(j) - \boldsymbol{P}_{\text{pr}}(j) = \sum_{i=T+1}^{j} \eta_{i-1} \boldsymbol{\xi}(i)$ which implies that $\sup_{j \geq T} \left\| \hat{\boldsymbol{P}}_{\text{pr}}(j) - \boldsymbol{P}_{\text{pr}}(j) \right\| \leq \delta'$. On the other hand for $j \geq T + 1$,

$$\begin{aligned} \hat{\boldsymbol{P}}_{\text{pr}}(j) = \hat{\boldsymbol{P}}_{\text{pr}}(j-1) + \eta_j \boldsymbol{f}(\hat{\boldsymbol{P}}_{\text{pr}}(j-1)) + \eta_{j-1}(\boldsymbol{f}(\boldsymbol{P}_{\text{pr}}(j-1)) \\ - \boldsymbol{f}(\hat{\boldsymbol{P}}_{\text{pr}}(j-1))), \end{aligned}$$

and since $\left\|\hat{\boldsymbol{P}}_{\text{pr}}(j-1) - \boldsymbol{P}_{\text{pr}}(j-1)\right\| \leq \delta' \leq \delta$, (6.22) shows that

$$\left\|\boldsymbol{f}(\boldsymbol{P}_{\text{pr}}(j-1)) - \boldsymbol{f}(\hat{\boldsymbol{P}}_{\text{pr}}(j-1))\right\| \leq \beta.$$

By (6.21) we know that whenever $\hat{\boldsymbol{P}}_{\text{pr}}(j-1) \in \mathcal{K}_\delta \setminus \mathcal{N}$ since $\mathcal{K}_\delta \subset \mathcal{W}_{M_2}$. $V(\hat{\boldsymbol{P}}_{\text{pr}}(j)) \leq V(\hat{\boldsymbol{P}}_{\text{pr}}(j-1)) - \eta_j\varepsilon$. Now (6.22) implies that $|V(\hat{\boldsymbol{P}}_{\text{pr}}(j)) - V(\hat{\boldsymbol{P}}_{\text{pr}}(j-1))| \leq \eta$ for any $\boldsymbol{P}_{\text{pr}}(j-1) \in \mathcal{K}_\delta$ and $|V(\boldsymbol{P}_{\text{pr}}(j)) - V(\hat{\boldsymbol{P}}_{\text{pr}}(j))| \leq \eta$ for any $\boldsymbol{P}_{\text{pr}}(j) \in \mathcal{K}$. ■

Finally, we need the following lemma from [7] for the proof of Theorem 17.

Lemma 21 ([7, Lemma 2.5])**.** *Assume 4. Let ε be real constant, n be an integer, and let $-\infty < a_1 < b_1 < \cdots < a_n < b_n < \infty$ be real numbers. Let $\{u_j\}$ be a bounded real sequence such that, for any $\eta > 0$, there exists an integer J such that for all $j \geq J$,*

$$u_j \leq u_{j-1} - \eta_{j-1}\varepsilon + (\eta + \eta_{j-1}\varepsilon)\mathbb{1}_{\{u_{j-1}\in A\}}, \quad A = \bigcup_{i=1}^{n}[a_i, b_i].$$

Then the limit points of the sequence $\{u_j\}$ are included in A.

With all the required lemmas we now proceed to the proof of Theorem 17. Since the statements of the proof hold for $\omega \in \Omega_{\text{conv}}$ we drop the notation of ω in the following proof.

Proof of Theorem 17. We first prove that $\lim_{j\to\infty} V(\boldsymbol{P}_{\text{pr}}(j))$ exists. For any $\alpha > 0$, define the set

$$[V(\bar{\mathcal{E}} \cap \mathcal{K}_q)]_\alpha := \{x \in \mathbb{R} : d(x, V(\bar{\mathcal{E}} \cap \mathcal{K}_q)) \leq \alpha\}.$$

Since $\|V\|_{\mathcal{K}_q} < \infty$, $V(\bar{\mathcal{E}} \cap \mathcal{K}_q)_\alpha$ is a finite union of disjoint intervals of length at least equal to 2α. By Lemma 20, there exist positive constants $\delta, \varepsilon, \lambda$ such that for any $\delta' \in (0, \delta]$, $\lambda' \in (0, \lambda]$, and $\eta > 0$, one may find an integer T and a sequence $\{\hat{\boldsymbol{P}}_{\text{pr}}(j) | j \geq T\}$ such that

$$\sup_{j\geq T}\left\|\boldsymbol{P}_{\text{pr}}(j) - \hat{\boldsymbol{P}}_{\text{pr}}(j)\right\| \leq \delta' \text{ and } \sup_{j\geq T}|V(\boldsymbol{P}_{\text{pr}}(j)) - V(\hat{\boldsymbol{P}}_{\text{pr}}(j))| \leq \eta.$$

Moreover for any $j \geq T+1$, for any $j \geq T+1$.

$$V(\hat{\boldsymbol{P}}_{\text{pr}}(j) \leq V(\hat{\boldsymbol{P}}_{\text{pr}}(j-1)) - \eta_{j-1}\varepsilon + (\eta + \eta_{j-1}\varepsilon)\mathbb{1}_{\{V(\hat{\boldsymbol{P}}_{\text{pr}}(j-1)) \in [V(\bar{\mathcal{E}} \cap \mathcal{K}_q)]_\alpha\}},$$

where we have chosen $\mathcal{N} = V^{-1}(\text{int}([V(\bar{\mathcal{E}} \cap \mathcal{K}_q)]_\alpha))$ and used $\mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{N}\}} \leq \mathbb{1}_{\{V(\boldsymbol{\theta}) \in [V(\bar{\mathcal{E}} \cap \mathcal{K}_q)]\}}$.

By Lemma 21, the limit points of the sequence $\{V(\boldsymbol{P}_{\text{pr}}(j))\}_{j \geq 0}$ are in $[V(\bar{\mathcal{E}} \cap \mathcal{K}_q)]_{\alpha'}$ for $\alpha' = \alpha + \eta$. Since α and η can be chosen arbitrarily small, this implies that the limit points of the sequence $\{V(\boldsymbol{P}_{\text{pr}}(j))\}_{j \geq 0}$ are included in $\cap_{\alpha > 0}[V(\bar{\mathcal{E}} \cap \mathcal{K}_q)]_\alpha$. We have $V(\bar{\mathcal{E}} \cap \mathcal{K}_q) = \cap_{\alpha > 0}[V(\bar{\mathcal{E}} \cap \mathcal{K}_q)]_\alpha$. Thus the limit points $\{V(\boldsymbol{P}_{\text{pr}}(j))\}$ belong to the set $V(\bar{\mathcal{E}} \cap \mathcal{K}_q)$.

On the other hand, $\limsup_{j \to \infty} |V(\boldsymbol{P}_{\text{pr}}(j) - V(\boldsymbol{P}_{\text{pr}}(j-1))| = 0$ which implies that the set of limit points of $\{V(\boldsymbol{P}_{\text{pr}}(j))\}$ is an interval. Because $V(\bar{\mathcal{E}})$ has an empty interior, the only intervals included in $V(\bar{\mathcal{E}} \cap \mathcal{K}_q)$ are isolated points, which shows that the limit, $\lim_{j \to \infty} V(\boldsymbol{P}_{\text{pr}}(j))$, exists.

We now prove that $\limsup_{j \to \infty} d(\boldsymbol{P}_{\text{pr}}(j), \bar{\mathcal{E}} \cap \mathcal{K}_q) = 0$. Let $\mathcal{N} \subset \mathcal{K}_q$ be an arbitrary neighborhood of $\bar{\mathcal{E}} \cap \mathcal{K}_q$. From Lemma 20 there exist constants $\delta, \varepsilon, \lambda \in \mathbb{R}^+$ such that for any $\delta' \in (0, \delta], \lambda' \in (0, \lambda]$, and $\eta > 0$ one may find an integer T and a sequence $\{\hat{\boldsymbol{P}}_{\text{pr}}(j)\}_{j \geq T}$ such that

$$\sup_{j \geq T} \left\| \boldsymbol{P}_{\text{pr}}(j) - \hat{\boldsymbol{P}}_{\text{pr}}(j) \right\| \leq \delta', \text{ and } \sup_{j \geq T} |V(\boldsymbol{P}_{\text{pr}}(j)) - V(\hat{\boldsymbol{P}}_{\text{pr}}(j))| \leq \eta$$

and for any $j \geq T+1$.

$$V(\hat{\boldsymbol{P}}_{\text{pr}}(j) \leq V(\hat{\boldsymbol{P}}_{\text{pr}}(j-1)) - \eta_{j-1}\varepsilon + (\eta + \eta_{j-1}\varepsilon)\mathbb{1}_{\{V(\hat{\boldsymbol{P}}_{\text{pr}}(j-1)) \in [V(\bar{\mathcal{E}} \cap \mathcal{K}_q)]_\alpha\}}.$$

For $j \geq T$, define $\tau(j) := \inf\{k \geq 0 | \hat{\boldsymbol{P}}_{\text{pr}}(k+j) \in \mathcal{N}\}$. For any integer p, define

$\tau^p(j) := \tau(j) \wedge p$, where $a \wedge b = \min(a, b)$. We have

$$V(\hat{\boldsymbol{P}}_{\text{pr}}(j + \tau^p(j))) - V(\hat{\boldsymbol{P}}_{\text{pr}}(j)) = \sum_{i=j+1}^{j+\tau^p(j)} \{V(\hat{\boldsymbol{P}}_{\text{pr}}(i)) - V(\hat{\boldsymbol{P}}_{\text{pr}}(i-1))\}$$
$$\leq -\varepsilon \sum_{i=j+1}^{j+\tau^p(j)} \eta_{i-1},$$

with the convention that, for any sequence $\{a_i\}$ and any integer l, $\sum_{i=l+1}^{l} a_i = 0$.

Therefore,

$$\begin{aligned}
&V(\boldsymbol{P}_{\text{pr}}(j + \tau^p(j))) - V(\boldsymbol{P}_{\text{pr}}(j)) \\
&= V(\boldsymbol{P}_{\text{pr}}(j + \tau^p(j))) - V(\hat{\boldsymbol{P}}_{\text{pr}}(j + \tau^p(j))) + V(\hat{\boldsymbol{P}}_{\text{pr}}(j + \tau^p(j))) - V(\hat{\boldsymbol{P}}_{\text{pr}}(j)) \\
&\quad + V(\hat{\boldsymbol{P}}_{\text{pr}}(j)) - V(\boldsymbol{P}_{\text{pr}}(j)) \\
&\leq 2\eta - \varepsilon \sum_{i=j+1}^{j+\tau^p(j)} \eta_{i-1}
\end{aligned}$$

Since $\{V(\boldsymbol{P}_{\text{pr}}(j)\}$ converges, for any $\varepsilon' > 0$ there exists $T' > T$ such that, for all $j \geq T'$,

$$-\varepsilon' < V(\boldsymbol{P}_{\text{pr}}(j + \tau^p(j))) - V(\boldsymbol{P}_{\text{pr}}(j)) \leq 2\eta - \varepsilon \sum_{i=j+1}^{j+\tau^p(j)} \eta_{i-1}$$

This implies that, for all $j \geq T'$ and all integer $p \geq 0$,

$$\sum_{i=j+1}^{j+\tau^p(j)} \eta_{i-1} \leq C(\varepsilon', \eta) := \varepsilon^{-1}(\varepsilon' + 2\eta).$$

Since $\sum_{i=j+1}^{j+\tau(j)} \eta_{i-1} = \lim_{p\to\infty} \sum_{i=j+1}^{j+\tau^p(j)} \eta_{i-1}$ and $\sum_{i=1}^{\infty} \eta_{i-1} = \infty$, the previous relation implies that, for all $j \geq T'$, $\tau(j) < \infty$, and $\sum_{i=j+1}^{j+\tau(j)} \leq C(\varepsilon', \eta)$.

For any integer p, $\boldsymbol{P}_{\text{pr}}(j+p) - \boldsymbol{P}_{\text{pr}}(j) = \sum_{i=j+1}^{j+p} \eta_{i-1} \boldsymbol{f}(\boldsymbol{P}_{\text{pr}}(i-1)) + \sum_{i=j+1}^{j+p} \eta_{i-1} \boldsymbol{M}(i)$,

which implies that

$$\|\boldsymbol{P}_{\text{pr}}(j+p) - \boldsymbol{P}_{\text{pr}}(j)\| \leq \|\boldsymbol{f}\|_{\mathcal{K}_q} \sum_{i=j+1}^{j+p} \eta_{i-1} + \left\| \sum_{i=j+1}^{j+p} \eta_{i-1} \boldsymbol{M}(i) \right\|.$$

Applying this inequality for $j \geq T'$ and $p = \tau(j)$ and using that, by definition $\hat{\boldsymbol{P}}_{\text{pr}}(j+\tau(j)) \in \mathcal{N}$,

$$\begin{aligned} d(\boldsymbol{P}_{\text{pr}}(j), \mathcal{N}) &\leq \left\| \hat{\boldsymbol{P}}_{\text{pr}}(j+\tau(j)) - \boldsymbol{P}_{\text{pr}}(j+\tau(j)) \right\| \\ &\quad + \|\boldsymbol{P}_{\text{pr}}(j+\tau(j)) = \boldsymbol{P}_{\text{pr}}(j)\| \\ &\leq \delta' + \|\boldsymbol{f}\|_{\mathcal{K}_q} C(\varepsilon', \eta) + \left\| \sum_{i=j+1}^{j+\tau(j)} \eta_{i-1} \boldsymbol{M}(i) \right\|. \end{aligned}$$

Since η, δ', and ε' can be chosen arbitrarily small, and $\limsup_{k\to\infty} \sup_{l\geq k} \left\| \sum_{i=k}^{l} \eta_{i-1} \boldsymbol{M}(i) \right\| = 0$, the latter inequality shows that $\lim_{j\to\infty} d(\boldsymbol{P}_{\text{pr}}(j), \mathcal{N}) = 0$. Since $\mathcal{N}$ is arbitrary, we thus have $\lim_{j\to\infty} d(\boldsymbol{P}_{\text{pr}}(j), \bar{\mathcal{E}} \cap \mathcal{K}_q) = 0$.

■

Chapter 6 in full, is a reprint of the material as it appears in A. Verma, A. Sharbafchi, S. Mohajer, B. Touri,"Distributed Fact Checking:A Stochastic Approximation Approach," in preparation for *IEEE Transactions on Autamtic Control*. The dissertation author was the primary investigator and author of this paper.

Chapter 7

Generalized Estimators

In this chapter (i) we move beyond ALL estimator and propose a generalized class of online estimators for the unreliability parameters of the agents. The estimators are associated with a function, of the agents' opinions and unreliability estimate. The associated function can be interpreted as an estimate for the validity of the statements. We also propose a set of axioms that a natural estimator should satisfy and hence, we call the class of functions satisfying the desired properties as the *natural functions*. (ii) We determine the class of natural functions for two and three agent fact-checker system that satisfy the desired properties and can serve as the function for the adaptive estimator. (iii) Finally we prove that a ALL estimator belongs to the class of natural functions for any n-agent fact-checker system and the hard-estimator does not belong to this class for any $n \geq 2$.

7.1 Natural Estimators

First, let us recall the online estimator for the unreliability parameters of the agents comprising the fact-checker for any number of agents $n \geq 2$ as introduced in Chapter 3. We have provided convergence guarantees for this algorithm for $n = 2$ agents in Chapter 5 and for its variant in Chapter 6.

The proposed algorithm/dynamics in (3.8) updates the unreliability parameters as

$$P_i(t+1) = (1-\eta_t)P_i(t) + \frac{1}{2}\eta_t \left(\frac{L(t)-1}{L(t)+1}R_i(t+1)+1\right),$$

for all $t \in \mathbb{N}_0$ and $i \in [n]$, with some initial condition (guess) $\boldsymbol{P}(0) \in (0,1)^n$, where $\{\eta_t\}$ is a pre-decided step-size sequence, and $L(t)$ is given in (3.5). One popular choice for the step-size sequence is the harmonic sequence $\eta_t = \frac{1}{t+1}$ for all $t \in \mathbb{N}_0$. To grasp the motivation behind the estimator using such a step-size sequence, we examined the scenario when the fact-checker knows the source sequence symbols $\{S(t)\}$. Since, at any time $t \in \mathbb{N}$, the output distribution of the agents given $S(t)$ is independent of each other, the problem of estimating the unreliability parameters of the agents is equivalent to n uncoupled problems of estimating the parameter of Bernoulli distribution from its samples. Estimation of parameter for a Bernoulli distribution from its sample is a well-studied problem and a class of estimator effective to solve it is the *add-constant* estimator [29]. For the current setting, for any $i \in [n]$, the add-β estimator, where $\beta \geq 0$ for parameter π_i at time $t \in \mathbb{N}$ is given by

$$Q_i(t) = \frac{\beta + \sum_{k=1}^{t} \mathbb{1}_{\{R_i(k) \neq S(k)\}}}{t + 2\beta}. \tag{7.1}$$

The estimator makes use of the empirical frequency of agent i misclassifying the source symbol received and can be expressed recursively as

$$Q_i(t+1) = (1-\gamma_t)Q_i(t) + \gamma_t \mathbb{1}_{\{R_i(t+1) \neq S(t+1)\}}.$$

Here, $\gamma_t := \frac{1}{t+1+2\beta}$ and $Q_i(0) = 1/2$.

Note that a central idea in describing the estimator for the unreliability parameter is to implicitly or explicitly define an estimator for the validity of the statements. Let us define a class of estimators based on functions $\mathcal{B} : \{-1,+1\}^n \times (0,1)^n \to [-1,1]$. The function $\mathcal{B}(\cdot;\cdot)$ represents a soft estimate of the statement truth S. Using the function $\mathcal{B}(\cdot;\cdot)$ we can define the

adaptive estimator for $\boldsymbol{\pi}$ as

$$\boldsymbol{P}(t+1)=(1-\eta_t)\boldsymbol{P}(t)+\eta_t\frac{1}{2}\left(\mathbf{1}-\mathcal{B}(\boldsymbol{R}(t+1);\boldsymbol{P}(t))\boldsymbol{R}(t+1)\right). \tag{7.2}$$

Let us look at three examples of the function $\mathcal{B}(\cdot;\cdot)$ representing different potential estimators.

(1)**Approximate Log-Likelihood (ALL) Estimator:** If we use the approximate log-likelihood ratio (3.5) to get an estimate of the statement validity we get the ALL estimator resulting in the online estimator defined through (3.8). It can be shown that the $\mathcal{B}$-function for this estimator is given as

$$\begin{aligned}\mathcal{B}_{\text{ALL}}(\boldsymbol{R};\boldsymbol{x}) &:= \tanh\left(\sum_{i=1}^{n}\frac{1}{2}R_i\log\frac{1-x_i}{x_i}\right)\\ &= \frac{1-L(\boldsymbol{R};\boldsymbol{x})}{1+L(\boldsymbol{R};\boldsymbol{x})},\end{aligned} \tag{7.3}$$

where $L(\boldsymbol{R};\boldsymbol{x}) := \prod_{i=1}^{n}\left(\frac{x_i}{1-x_i}\right)^{R_i}$ is the approximation of the likelihood ratio.

(2)**Hard-Thresholding (HT) Estimator:** Instead of using the approximate likelihood ration for statement validity we can use the hard estimator which uses the hard estimate of statement validity to compute the empirical frequency of misclassifying the source symbol. In other words, the HT estimator compares the output of each agent with the estimated value for $S(t+1)$ given in (3.6), if the two values agree, HT Estimator decreases the agent's unreliability parameter down, otherwise, the unreliability parameter will be increased. The $\mathcal{B}$-function for the HT estimator can be expressed as

$$\mathcal{B}_{\text{HT}}(\boldsymbol{R};\boldsymbol{x}) := \mathsf{sgn}\left(\sum_{i=1}^{n}R_i\log\frac{1-x_i}{x_i}\right), \tag{7.4}$$

where $\mathsf{sgn}(a) := -\mathbb{1}_{\{a\leq 0\}}+\mathbb{1}_{\{a\geq 0\}}$ for any $a\in\mathbb{R}$.

To corroborate the idea that $\mathcal{B}(\cdot;\cdot)$ is an estimate of the statement validity S, let us introduce the **Oracle Estimator**

$$\mathcal{B}_{\text{oracle}}(\boldsymbol{R}, S; \boldsymbol{x}) := S\mathbf{1}. \tag{7.5}$$

Using the $\mathcal{B}_{\text{oracle}}$-function results in the add-β estimator defined through (7.1). Note that the function $\mathcal{B}_{\text{oracle}}$ does not fit in our class of functions $\mathcal{B}(\cdot;\cdot)$ of interest since it takes the truth of the statement S as an argument.

7.2 Results

In this section, we state the main results of this chapter. Let us start with stating the desirable properties of the function $\mathcal{B}(\cdot;\cdot)$ that must be satisfied in order to have a feasible estimator of unreliability parameters that converges to $\boldsymbol{\pi}$.

7.2.1 Natural Estimators: Axioms and Necessary Conditions

First, let us introduce some conditions/axioms that one would expect from a reasonable estimator. Later, we will discuss why such axioms are expected from such an estimator.

Definition 8. *For any $n \in \mathbb{N}$ let us define $\mathcal{C}_n^{nat}$ as the set of all functions $\mathcal{B} : \{-1,+1\}^n \times (0,1)^n \to [-1,1]$ that satisfy*

Assumption (i) *Anti-Symmetry of reliability:*

$$\mathcal{B}(\boldsymbol{R}; \boldsymbol{x}) = -\mathcal{B}(\boldsymbol{R}; \mathbf{1} - \boldsymbol{x}). \tag{7.6}$$

Assumption (ii) *Anti-Symmetry of Opinions:*

$$\mathcal{B}(-\boldsymbol{R}; \boldsymbol{x}) = -\mathcal{B}(\boldsymbol{R}; \boldsymbol{x}). \tag{7.7}$$

Assumption (iii) *Consistency of Estimators:*

$$\mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{x}}}[\boldsymbol{R}\cdot\mathcal{B}(\boldsymbol{R};\boldsymbol{x})]=\mathbf{1}-2\boldsymbol{x}. \tag{7.8}$$

We refer to $\mathcal{C}_n^{nat}$ as the set of ***natural functions*** *for an n-agent fact-checker system.*

Assumption (i) ensures that the estimates of the statement validity for fact-checker systems with unreliability parameters $\boldsymbol{\pi}$ and $\mathbf{1}-\boldsymbol{\pi}$ takes the same absolute value but has different sign. The assumption is justified as the output of the fact-checker system with unreliability parameter vector $\mathbf{1}-\boldsymbol{\pi}$ can be seen as the flipped output of a fact-checker system with unreliability parameter $\boldsymbol{\pi}$. Similarly given a fact-checker system *Assumption (ii)* ensures that the flipping the output of all the agents' opinion flips the sign of the estimate of the statement validity. Finally, regarding *Assumption (iii)*, note that the consistency condition (7.8) is equivalent to $\boldsymbol{x}=\frac{1}{2}(\mathbf{1}-\mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{x}}}[\boldsymbol{R}\mathcal{B}(\boldsymbol{R};\boldsymbol{x})])$, which is what is expected from the mean-field dynamics of (7.2), i.e., to have $\boldsymbol{x}$ as its equilibrium point, given that the agents true reliability parameter vector is $\boldsymbol{x}$.

With the above discussion, we are ready to present our main results. The first result shows that interestingly the only natural estimator for two-agent fact-checker system is the ALL estimator (7.3).

Proposition 19 (Elements of $\mathcal{C}_2^{\text{nat}}$)**.** *For a two-agent fact-checker system the class of functions $\mathcal{B}(\cdot;\cdot)$ satisfying* Assumption (i)-Assumption (iii) *contains exclusively the function $\mathcal{B}_{ALL}(\boldsymbol{R};\boldsymbol{x})$ as defined in* (7.3)*,* i.e., $\mathcal{C}_2^{nat}=\{\mathcal{B}_{ALL}(\boldsymbol{R};\boldsymbol{x})\}$.

Remark 9. *In [58] we studied the ALL estimator for a two-agent fact-checker system whose unreliability parameter is $\boldsymbol{\pi}$ and we have shown that the estimates $\{\boldsymbol{P}(t)\}$ converge to the solution set $\mathcal{E}$ of the equation*

$$\mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\boldsymbol{R}\mathcal{B}_{ALL}(\boldsymbol{R};\boldsymbol{x})]=\mathbf{1}-2\boldsymbol{x}. \tag{7.9}$$

Note the difference in (7.6) *and* (7.9) *lies in the distribution over which the expectation is taken. Since* $\mathcal{B}_{ALL}(\cdot;\cdot)$ *satisfies* Assumption (iii) *we know that* $\boldsymbol{\pi} \in \mathcal{E}$*. However the set* $\mathcal{E}$ *is a continuum of points* $\boldsymbol{x}$ *for which* $g_{\boldsymbol{x}}(\boldsymbol{R}) = g_{\boldsymbol{\pi}}(\boldsymbol{R})$ *for all* $\boldsymbol{R} \in \{-1, +1\}^2$.

In the following proposition we identify the functions that satisfy the properties required by a natural estimator for a three-agent fact-checker system.

Proposition 20 (Elements of $\mathcal{C}_3^{\text{nat}}$)**.** *For a three-agent fact-checker system, the set of natural estimators* $\mathcal{C}_3^{nat}$ *consists of functions* $\mathcal{B}(\cdot;\cdot)$ *satisfying*

$$\mathcal{B}(\boldsymbol{R};\boldsymbol{x}) = \mathcal{B}_{ALL}(\boldsymbol{R};\boldsymbol{x}) + \frac{c_{\boldsymbol{x}} R_1 R_2 R_3}{2 g_{\boldsymbol{x}}(\boldsymbol{R})}, \tag{7.10}$$

where $\mathcal{B}_{ALL}(\boldsymbol{R};\boldsymbol{x})$ *is the ALL estimator defined in* (7.3) *and* $c_{\boldsymbol{x}}$ *is any function of the vector* $\boldsymbol{x}$ *such that* $c_{\boldsymbol{x}} = -c_{\mathbf{1}-\boldsymbol{x}}$.

In the following proposition, we show that for all $n \geq 2$ there exists $\boldsymbol{x}$ for which $\mathcal{B}_{\text{HT}}(\cdot;\cdot)$ does not satisfy *Assumption (iii)*.

Proposition 21 (Convergence for Hard-Thresholding Estimator)**.** *The function* $\mathcal{B}_{HT}(\cdot;\cdot)$ *as defined through* (7.4)*, based on the hard-thresholding estimator, does not satisfy* Assumption (iii)*. In other words,* $\mathcal{B}_{HT}(\cdot;\cdot) \notin \mathcal{C}_n^{nat}$ *for any* $n \geq 2$.

Recall that the system of equations in *Assumption (iii)* is a necessary condition for the estimates $\{\boldsymbol{P}(t)\}$ to converge to $\boldsymbol{\pi}$. However, for a fact-checker system with unreliability parameter $\boldsymbol{\pi}$ it is also important to identify the solution set $\mathcal{E}$ to the system of equation

$$\mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\boldsymbol{R}\mathcal{B}_{\text{ALL}}(\boldsymbol{R};\boldsymbol{x})] = \mathbf{1} - 2\boldsymbol{x}.$$

The solution set $\mathcal{E}$ represents the points $\boldsymbol{x} \in (0,1)^n$ that could be the points of convergence for the estimates $\{\boldsymbol{P}(t)\}$. In the following theorem, we identify the set $\mathcal{E}$ for a three-agent fact-checker

system to be the set containing the true estimate $\boldsymbol{\pi}$, the 'symmetric' estimate $\mathbf{1} - \boldsymbol{\pi}$ and the degenerate point $\frac{1}{2}\mathbf{1}$.

Theorem 22 (Fixed points of ALL-estimator for three-agent fact-checker)**.** *For a three-agent fact-checker system where the agents have unreliability parameters $\pi_i \in (0,1) \setminus \{\frac{1}{2}\}$ for $i \in [3]$, the set of solutions of the fixed-point equation $\boldsymbol{x} = \frac{1}{2}\left(\mathbf{1} - \mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\boldsymbol{R}\mathcal{B}_{ALL}(\boldsymbol{R};\boldsymbol{x})]\right)$ is $\mathcal{S} := \{\boldsymbol{\pi}, \mathbf{1}-\boldsymbol{\pi}, \frac{1}{2}\mathbf{1}\}$.*

Note that the set $\mathcal{E}$ also represents the set of convergence for the Dawid-Skene estimator [16] and the Theorem 22 is the first result to identify the exact set $\mathcal{E}$. The theorem signifies that for a three-agent system, the only points the Dawid-Skene and its variants would converge to are the relevant points $\boldsymbol{\pi}$, $\mathbf{1} - \boldsymbol{\pi}$ or the degenerate point $\frac{1}{2}\mathbf{1}$.

In the following theorem, we show that for any n-agent fact-checker system the adaptive estimator associated with the ALL estimator satisfies all the desired properties.

Theorem 23. *For $n \geq 2$, $\mathcal{B}_{ALL}(\cdot;\cdot)$, as defined in* (7.3)*, satisfies* Assumption (i)-Assumption (iii)*. In other words, $\mathcal{B}_{ALL} \in \mathcal{C}_n^{nat}$.*

7.3 Proof of Main Results

In this section, we present the proof of the results discussed in the previous section. First, let us establish a notation to impose an ordering on the 2^n distinct possibilities of the output vector $\boldsymbol{R}$.

Definition 9 (Notation)**.** *Consider the binary representation $(b_1, b_2, \ldots, b_n)$ of $N \in \{0\} \cup [2^n - 1]$. Here b_1 represents the most significant bit and b_n the least significant bit. Define the output vector $\boldsymbol{\mathcal{R}}_N$ associated with N as*

$$\boldsymbol{\mathcal{R}}_N := \begin{pmatrix} -1^{b_1} & -1^{b_2} & \ldots & -1^{b_n} \end{pmatrix}^\top.$$

Now we provide the proof for the characterization of elements in $\mathcal{C}_2^{\text{nat}}$.

Proof of Proposition 19. We show that for any fixed $\boldsymbol{x} \in (0,1)^2$, if $\mathcal{B}(\cdot;\cdot) \in \mathcal{C}_2^{\text{nat}}$, the values $\mathcal{B}(\boldsymbol{R};\boldsymbol{x})$ takes for any vector $\boldsymbol{R} \in \{-1,1\}^2$ coincides with that of $\mathcal{B}_{\text{ALL}}(\boldsymbol{R};\boldsymbol{x})$ given in (7.3). To do this, we utilize *Assumption (iii)*.

So, consider an arbitrary point $\boldsymbol{x} \in (0,1)^2$. To compute $\mathbb{E}[R_1\mathcal{B}(\boldsymbol{R};\boldsymbol{x})]$, note that $g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0) = g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3)$ and $g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1) = g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2)$. Therefore,

$$\begin{aligned}
\mathbb{E}[R_1\mathcal{B}(\boldsymbol{R};\boldsymbol{x})] &= g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)\mathcal{B}(\boldsymbol{\mathcal{R}}_0;\boldsymbol{x}) + g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)\mathcal{B}(\boldsymbol{\mathcal{R}}_1;\boldsymbol{x}) \\
&\quad - g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2)\mathcal{B}(\boldsymbol{\mathcal{R}}_3;\boldsymbol{x}) - g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3)\mathcal{B}(\boldsymbol{\mathcal{R}}_2;\boldsymbol{x}) \\
&= g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)\mathcal{B}(\boldsymbol{\mathcal{R}}_0;\boldsymbol{x}) + g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)\mathcal{B}(\boldsymbol{\mathcal{R}}_1;\boldsymbol{x}) \\
&\quad - g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)\mathcal{B}(\boldsymbol{\mathcal{R}}_3;\boldsymbol{x}) - g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)\mathcal{B}(\boldsymbol{\mathcal{R}}_2;\boldsymbol{x}) \\
&= 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)\mathcal{B}(\boldsymbol{\mathcal{R}}_0;\boldsymbol{x}) + 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)\mathcal{B}(\boldsymbol{\mathcal{R}}_1;\boldsymbol{x}),
\end{aligned}$$

where the last step follows from the *Assumption (i)*, $\mathcal{B}(-\boldsymbol{R};\boldsymbol{x}) = -\mathcal{B}(\boldsymbol{R};\boldsymbol{x})$. Similarly we have

$$\mathbb{E}[R_2\mathcal{B}(\boldsymbol{R};\boldsymbol{x})] = 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)\mathcal{B}(\boldsymbol{\mathcal{R}}_0;\boldsymbol{x}) - 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)\mathcal{B}(\boldsymbol{\mathcal{R}}_1;\boldsymbol{x}).$$

Therefore in order for $\mathcal{B}(\cdot;\cdot)$ to satisfy (7.8), we need to have

$$\begin{pmatrix} 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0) & 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1) \\ 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0) & -2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1) \end{pmatrix} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} = \begin{pmatrix} 1-2x_1 \\ 1-2x_2 \end{pmatrix},$$

where $B_i = \mathcal{B}(\boldsymbol{\mathcal{R}}_i;\boldsymbol{x})$ for $i \in \{0,1\}$. Note that for non-degenerate $\boldsymbol{x}$, i.e., if $x_i \notin \{0,1\}$, the above matrix is invertible. Solving the system of linear equations in B_0, B_1 we get

$$\begin{pmatrix} B_0 \\ B_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{4g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)} & \frac{1}{4g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)} \\ \frac{1}{4g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)} & -\frac{1}{4g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)} \end{pmatrix} \begin{pmatrix} 1-2x_1 \\ 1-2x_2 \end{pmatrix} = \begin{pmatrix} \frac{1-x_1-x_2}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)} \\ \frac{x_2-x_1}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)} \end{pmatrix}.$$

We can simplify B_0 as follow

$$B_0 = \frac{1 - x_1 - x_2}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)} = \frac{(1-x_1)(1-x_2) - x_1x_2}{x_1x_2 + (1-x_1)(1-x_2)} = \mathcal{B}_{\text{ALL}}(\boldsymbol{\mathcal{R}}_0; \boldsymbol{x}).$$

Similarly we have $B_1 = \mathcal{B}_{\text{ALL}}(\boldsymbol{\mathcal{R}}_1; \boldsymbol{x})$. ■

Next we provide the characterization of elements in $\mathcal{C}_3^{\text{nat}}$.

Proof of Proposition 20. As in the proof of Proposition 19 we show that for any fixed $\boldsymbol{x} \in (0,1)^3$, $\mathcal{B}(\cdot;\cdot) \in \mathcal{C}_3^{\text{nat}}$ iff the value $\mathcal{B}(\boldsymbol{R}; \boldsymbol{x})$ takes for any vector $\boldsymbol{R} \in \{-1, 1\}^2$ satisfies (7.10). Consider an arbitrary point $\boldsymbol{x} \in (0,1)^3$. To compute $\mathbb{E}_{\boldsymbol{R} \sim g_{\boldsymbol{x}}}[R_i \mathcal{B}(\boldsymbol{R}; \boldsymbol{x})]$ note that $g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_i) = g_{\boldsymbol{\mathcal{R}}_{7-i}}$ for any $i \in \{0,1,2,3\}$. Similar to the proof of Proposition 19, we can express the equations in terms of the values of the functions at $\boldsymbol{\mathcal{R}}_i$ for $i \in \{0,1,2,3\}$ through the equation $\mathcal{H}\boldsymbol{B} = \mathbf{1} - 2\boldsymbol{x}$, where

$$\mathcal{H} = \begin{pmatrix} 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0) & 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1) & 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2) & 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3) \\ 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0) & 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1) & -2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2) & -2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3) \\ 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0) & -2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1) & 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2) & -2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3) \end{pmatrix}$$

and $\boldsymbol{B} = \begin{pmatrix} B_0 & B_1 & B_2 & B_3 \end{pmatrix}^\top$. Here $B_i = \mathcal{B}(\boldsymbol{\mathcal{R}}_i; \boldsymbol{x})$ for $i \in \{0,1,2,3\}$. The matrix $\mathcal{H}$ in one of its row echelon form can be expressed as

$$\begin{pmatrix} 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0) & 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1) & 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2) & 2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3) \\ 0 & -2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1) & 0 & -2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3) \\ 0 & 0 & -2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2) & -2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3) \end{pmatrix}.$$

Therefore, we know that $\mathcal{H}$ is a matrix with rank 3 if $g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_i) \neq 0$ for $i \in \{0,1,2,3\}$. By the rank-nullity theorem [35, eq.(4.4.15)] the dimension of the null-space of $\mathcal{H}$ is 1. It can be seen

that the null-space of $\mathcal{H}$ is given by span($\boldsymbol{z}$), where

$$\boldsymbol{z} = \begin{pmatrix} \frac{1}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)} & -\frac{1}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)} & -\frac{1}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2)} & \frac{1}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3)} \end{pmatrix}^\top.$$

Therefore, the solution set for $\mathcal{H}\boldsymbol{B} = \mathbf{1} - 2\boldsymbol{x}$ is given by $\{\boldsymbol{b} \in \mathbb{R}^4 : \boldsymbol{b} = \boldsymbol{B}_* + c\boldsymbol{z}\}$, where $\boldsymbol{B}_*$ is one solution to the system of linear equation $\mathcal{H}\boldsymbol{B} = \mathbf{1} - 2\boldsymbol{x}$. We can choose

$$\boldsymbol{B}_* = \begin{pmatrix} \frac{(1-x_1)(1-x_2)(1-x_3)-x_1x_2x_3}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)} \\ \frac{(1-x_1)(1-x_2)x_3-x_1x_2(1-x_3)}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)} \\ \frac{(1-x_1)x_2(1-x_3)-x_1(1-x_2)x_3}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2)} \\ \frac{(1-x_1)x_2x_3-x_1(1-x_2)(1-x_3)}{2g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3)} \end{pmatrix}, \tag{7.11}$$

whose i-th element is in fact $(\boldsymbol{B}_*)_i = \mathcal{B}_{\text{ALL}}(\boldsymbol{\mathcal{R}}_{i-1}; \boldsymbol{x})$. Therefore the functions satisfying Definition 8 take the form

$$\mathcal{B}(\boldsymbol{R}; \boldsymbol{x}) = \mathcal{B}_{\text{ALL}}(\boldsymbol{R}; \boldsymbol{x}) + \frac{c_{\boldsymbol{x}} R_1 R_2 R_3}{2g_{\boldsymbol{x}}(\boldsymbol{R})},$$

where $c_{\boldsymbol{x}}$ is an arbitrary function of $\boldsymbol{x}$. Furthermore, to ensure $\mathcal{B}(\boldsymbol{R}; \boldsymbol{x}) = -\mathcal{B}(\boldsymbol{R}; 1 - \boldsymbol{x})$, we need to have

$$\mathcal{B}_{\text{ALL}}(\boldsymbol{R}; \boldsymbol{x}) + \frac{c_{\boldsymbol{x}} R_1 R_2 R_3}{2g_{\boldsymbol{x}}(\boldsymbol{R})} = -\mathcal{B}_{\text{ALL}}(\boldsymbol{R}; 1 - \boldsymbol{x}) - \frac{c_{\mathbf{1}-\boldsymbol{x}} R_1 R_2 R_3}{2g_{\mathbf{1}-\boldsymbol{x}}(\boldsymbol{R})}.$$

As $g_{\boldsymbol{x}}(\boldsymbol{R}) = g_{\mathbf{1}-\boldsymbol{x}}(\boldsymbol{R})$ and $\mathcal{B}_{\text{ALL}}(\boldsymbol{R}; \boldsymbol{x}) = -\mathcal{B}_{\text{ALL}}(\boldsymbol{R}; 1 - \boldsymbol{x})$, the above equality holds iff $c_{\boldsymbol{x}} = -c_{\mathbf{1}-\boldsymbol{x}}$. ■

Proof of Proposition 21. From Proposition 19, it readily follows that for two-agents fact-checker system $\mathcal{B}_{\text{HT}}(\cdot;\cdot) \notin \mathcal{C}_2^{\text{nat}}$. For any $n \geq 2$, we show that there exists $\boldsymbol{x} \in (0,1)^n$ such that $\mathcal{B}_{\text{HT}}(\cdot;\cdot)$ does not satisfy *Assumption (iii)*.

Consider $\boldsymbol{x}^* \in (0,1)^n$ such that

$$\log \frac{1-x_1^*}{x_1^*} > \sum_{i=2}^{n} \left| \log \frac{1-x_i^*}{x_i^*} \right|. \tag{7.12}$$

Then, for any $\boldsymbol{R} \in \{-1,+1\}^n$, we have $\mathcal{B}_{\text{HT}}(\boldsymbol{R};\boldsymbol{x}^*) = R_1$. Therefore $\mathbb{E}[R_1 \mathcal{B}_{\text{HT}}(\boldsymbol{R};\boldsymbol{x})] = \mathbb{E}[R_1^2] = 1$. However $1 - 2x_1^* < 1$. So, $\mathcal{B}_{\text{HT}}(\boldsymbol{R};\boldsymbol{x}^*)$ does not satisfy (7.8), at least for vectors $\boldsymbol{x}$ satisfying (7.12). ■

In order to prove Theorem 22 for $a, b, c \in (0,1)$, we define a function $h(a,b,c) = abc + \bar{a}\bar{b}\bar{c}$. For convenience, with an abuse of notation, we also use the same notation and define $h(a,b) = ab + \bar{a}\bar{b}$.

Proof of Theorem 22. Using the fact that $x_i \notin \{0,1\}$ for $i \in [3]$, we can perform algebraic manipulations and express the fixed-point equation $\boldsymbol{x} = \frac{1}{2}\left(\mathbf{1} - \mathbb{E}_{\boldsymbol{R}\sim g_{\boldsymbol{\pi}}}[\boldsymbol{R}\mathcal{B}_{\text{ALL}}(\boldsymbol{R};\boldsymbol{x})]\right)$ as $\mathcal{X}\boldsymbol{u} = 0$ where

$$\mathcal{X} = \begin{pmatrix} x_2+x_3-1 & x_3+x_1-1 & x_1+x_2-1 \\ 1-x_2-x_3 & x_3-x_1 & x_2-x_1 \\ x_3-x_2 & 1-x_1-x_3 & x_1-x_2 \\ x_2-x_3 & x_1-x_3 & 1-x_1-x_2 \end{pmatrix}^\top$$

and $\boldsymbol{u} = (u_0,u_1,u_2,u_3)^\top$, with $u_0 = \frac{g_{\boldsymbol{\pi}}(\boldsymbol{\mathcal{R}}_0)}{g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0)}$, $u_1 = \frac{g_{\boldsymbol{\pi}}(\boldsymbol{\mathcal{R}}_3)}{g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_3)}$, $u_2 = \frac{g_{\boldsymbol{\pi}}(\boldsymbol{\mathcal{R}}_2)}{g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_2)}$, and $u_3 = \frac{g_{\boldsymbol{\pi}}(\boldsymbol{\mathcal{R}}_1)}{g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_1)}$.

Summing equations in $\mathcal{X}\boldsymbol{u} = \mathbf{0}$ and multiplying the result by $\frac{1}{2}$, we get

$$\begin{aligned}\left(\frac{3}{2} - (x_1+x_2+x_3)\right) u_0 = &\left(\frac{1}{2} - x_1\right) u_1 \\ &+ \left(\frac{1}{2} - x_2\right) u_2 + \left(\frac{1}{2} - x_3\right) u_3.\end{aligned} \tag{7.13}$$

Case 1: Consider the case where $x_1 + x_2 + x_3 \neq \frac{3}{2}$. For $i \in [3]$ define $w_i = \frac{\frac{1}{2}-x_i}{\sum_{j=1}^{3}\frac{1}{2}-x_j}$.

Then we have

$$u_0 = w_1u_1 + w_2u_2 + w_3u_3, \tag{7.14}$$

where $w_1 + w_2 + w_3 = 1$. Replacing u_0 from (7.14) in $\mathcal{X}\boldsymbol{u} = 0$, we get

$$\begin{aligned} w_1u_1 + w_2u_2 + w_3u_3 - u_1 &= \frac{w_3 - w_2}{w_2 + w_3}(u_3 - u_2), \\ w_1u_1 + w_2u_2 + w_3u_3 - u_2 &= \frac{w_1 - w_3}{w_3 + w_1}(u_1 - u_3), \\ w_1u_1 + w_2u_2 + w_3u_3 - u_3 &= \frac{w_2 - w_1}{w_1 + w_2}(u_2 - u_1). \end{aligned} \tag{7.15}$$

We can rewrite the above system as

$$\begin{aligned} u_1 &= au_2 + (1 - a)u_3, \\ u_2 &= bu_3 + (1 - b)u_1, \\ u_3 &= cu_1 + (1 - c)u_2, \end{aligned} \tag{7.16}$$

where the coefficients a, b, and c are given as

$$\begin{aligned} a &= \frac{w_2(w_2 + w_3) + (w_3 - w_2)}{(w_1 - 1)^2}, \\ b &= \frac{w_3(w_3 + w_1) + (w_1 - w_3)}{(w_2 - 1)^2}, \\ c &= \frac{w_1(w_1 + w_2) + (w_2 - w_1)}{(w_3 - 1)^2}. \end{aligned}$$

The system of equation (7.16) is equivalent to

$$\begin{aligned}(1-a(1-b))(u_1-u_3)&=0,\\(1-b(1-c))(u_2-u_1)&=0,\\(1-c(1-a))(u_3-u_2)&=0.\end{aligned}\tag{7.17}$$

We note that $a(1-b)=1$ if and only if $w_1w_2w_3=0$. Similar conclusions hold for $b(1-c)=1$ and $c(1-a)=1$. Therefore, the system of equations in (7.17) holds only if we have either $w_1w_2w_3=0$ (Case 1-1) or $u_1=u_2=u_3$ (Case 1-2) .

Case 1-1: Note that $w_1w_2w_3=0$ implies that $x_i=\frac{1}{2}$ for some $i\in[3]$. Let us consider the case with $w_1=0$, or equivalently $x_1=\frac{1}{2}$. The system of equations $\mathcal{X}\boldsymbol{u}=\mathbf{0}$ can then be simplified to

$$\begin{aligned}(x_2+x_3-1)&\frac{(h(\pi_1,\pi_2,\pi_3)-h(\bar{\pi}_1,\pi_2,\pi_3))}{(h(\pi_1,\pi_2,\bar{\pi}_3)-h(\pi_1,\bar{\pi}_2,\pi_3))}\\&=\frac{h(x_2,x_3)}{h(x_2,\bar{x}_3)}(x_3-x_2),\\\left(x_3-\frac{1}{2}\right)&(u_0+u_1-u_2-u_3)=0,\\\left(x_2-\frac{1}{2}\right)&(u_0+u_1-u_2-u_3)=0.\end{aligned}\tag{7.18}$$

It is clear that $\boldsymbol{x}=\frac{1}{2}\mathbf{1}$ is a feasible solution for (7.18). In the following, we prove that (7.18) has no other solution.

Let $\boldsymbol{x}\neq\frac{1}{2}\mathbf{1}$, and without loss of generality, $x_2\neq\frac{1}{2}$. Hence, we should have $u_0+u_1=u_2+u_3$. However, we have $u_0+u_1=2\frac{h(\pi_2,\pi_3)}{h(x_2,x_3)}$ and $u_2+u_3=2\frac{h(\pi_2,\bar{\pi}_3)}{h(x_2,\bar{x}_3)}$. Therefore, $u_0+u_1=u_2+u_3$ holds if and only if

$$h(\pi_2,\pi_3)=h(x_2,x_3).\tag{7.19}$$

Plugging (7.19) in (7.18), we arrive at $(x_2 + x_3 - 1) = \tilde{c}(x_3 - x_2)$, or equivalently,

$$x_2 = \frac{x_3(\tilde{c} - 1) + 1}{1 + \tilde{c}}, \tag{7.20}$$

where

$$\begin{aligned}\tilde{c} &= \frac{h(\pi_2, \pi_3)}{h(\pi_2, \bar{\pi}_3)} \frac{(h(\pi_1, \pi_2, \bar{\pi}_3) - h(\pi_1, \bar{\pi}_2, \pi_3))}{h(\pi_1, \pi_2, \pi_3) - h(\bar{\pi}_1, \pi_2, \pi_3)} \\ &= \frac{h(\pi_2, \pi_3)}{h(\pi_2, \bar{\pi}_3)} \left(\frac{\pi_2 - \pi_3}{\pi_2 + \pi_3 - 1} \right) = \frac{h(\pi_2, \pi_3)}{h(\pi_2, \bar{\pi}_3)} \left(\frac{\pi_2 - \pi_3}{\pi_2 - \bar{\pi}_3} \right).\end{aligned}$$

Plugging (7.20) into (7.19), we get

$$\begin{aligned}0 &= h(x_2, x_3) - h(\pi_2, \pi_3) = 2x_2x_3 - x_2 - x_3 + 1 - h(\pi_2, \pi_3) \\ &= \frac{\tilde{c} - 1}{2(\tilde{c} + 1)}(2x_3 - 1)^2 + \frac{1}{2} - h(\pi_2, \pi_3).\end{aligned} \tag{7.21}$$

We know $h(\pi_2, \pi_3) = \frac{1}{2}(2\pi_2 - 1)(2\pi_3 - 1) + \frac{1}{2} = 2\tilde{\pi}_2\tilde{\pi}_3 + \frac{1}{2}$, where $\tilde{\pi}_i = \frac{1}{2} - \pi_i \in (-\frac{1}{2}, \frac{1}{2})$ for $i \in [3]$. Moreover, we have

$$\begin{aligned}\frac{\tilde{c} - 1}{\tilde{c} + 1} &= \frac{h(\pi_2, \pi_3)(\pi_2 - \pi_3) - (1 - h(\pi_2, \pi_3))(\pi_2 + \pi_3 - 1)}{h(\pi_2, \pi_3)(\pi_2 - \pi_3) + (1 - h(\pi_2, \pi_3))(\pi_2 + \pi_3 - 1)} \\ &= \frac{-4\tilde{\pi}_2^2\tilde{\pi}_3 + \tilde{\pi}_3}{4\tilde{\pi}_2\tilde{\pi}_3^2 - \tilde{\pi}_2} = -\frac{\tilde{\pi}_3(4\tilde{\pi}_2^2 - 1)}{\tilde{\pi}_2(4\tilde{\pi}_3^2 - 1)}.\end{aligned} \tag{7.22}$$

Using this in (7.21), we arrive at

$$\begin{aligned}0 &= -\frac{1}{2}\frac{\tilde{\pi}_3(4\tilde{\pi}_2^2 - 1)}{\tilde{\pi}_2(4\tilde{\pi}_3^2 - 1)}(2x_3 - 1)^2 - 2\tilde{\pi}_2\tilde{\pi}_3 \\ &= -\frac{\tilde{\pi}_3}{2\tilde{\pi}_2} \left(\frac{4\tilde{\pi}_2^2 - 1}{4\tilde{\pi}_3^2 - 1}(2x_3 - 1)^2 + 4\tilde{\pi}_2^2 \right).\end{aligned}$$

This last equation holds if and only if $\tilde{\pi}_3 = 0$. Plugging this in (7.22) implies $\tilde{c} = 1$, which together with (7.20) leads to $x_2 = \frac{1}{2}$, which is a contradiction. Hence, the only solution for

Case 1-1 is $\boldsymbol{x} \neq \frac{1}{2}\mathbf{1}$.

Case 1-2: Next, we study the case of $u_1 = u_2 = u_3$, which together with (7.13) leads to $u_0 = u_1 = u_2 = u_3 = K$ for some $K \in \mathbb{R}$. Equivalently, we get $g_{\boldsymbol{\pi}}(\boldsymbol{\mathcal{R}}_i) = K g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_i)$ for $i \in [3] \cup \{0\}$. Summing up the equations over i, we get $K = 1$, since $g_{\boldsymbol{\pi}}$ and $g_{\boldsymbol{x}}$ are probability mass functions. Therefore we get $g_{\boldsymbol{\pi}}(\boldsymbol{\mathcal{R}}_i) = g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_i)$ for $i \in [3] \cup \{0\}$.

From the definition of the function h we have

$$\begin{aligned} h(x_1, x_2, x_3) - h(x_1, x_2, \bar{x}_3) &= (1-2x_3)(1-x_1-x_2), \\ h(x_1, x_2, x_3) - h(x_1, \bar{x}_2, x_3) &= (1-2x_2)(1-x_3-x_1), \\ h(x_1, x_2, x_3) - h(\bar{x}_1, x_2, x_3) &= (1-2x_1)(1-x_2-x_3). \end{aligned} \tag{7.23}$$

Using (7.23) and $g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_0) - g_{\boldsymbol{x}}(\boldsymbol{\mathcal{R}}_i) = g_{\boldsymbol{\pi}}(\boldsymbol{\mathcal{R}}_0) - g_{\boldsymbol{\pi}}(\boldsymbol{\mathcal{R}}_i)$ for $i \in [3]$ we get

$$\frac{\tilde{\pi}_1(\tilde{\pi}_2 + \tilde{\pi}_3)}{\tilde{x}_1(\tilde{x}_2 + \tilde{x}_3)} = \frac{\tilde{\pi}_2(\tilde{\pi}_3 + \tilde{\pi}_1)}{\tilde{x}_2(\tilde{x}_3 + \tilde{x}_1)} = \frac{\tilde{\pi}_3(\tilde{\pi}_1 + \tilde{\pi}_2)}{\tilde{x}_3(\tilde{x}_1 + \tilde{x}_2)}, \tag{7.24}$$

where $\tilde{x}_i = \frac{1}{2} - x_i$ and $\tilde{\pi}_i = \frac{1}{2} - \pi_i$ for $i \in [3]$. Further simplifying we get the following set of equations

$$\tilde{x}_1\tilde{x}_2 = \tilde{\pi}_1\tilde{\pi}_2, \ \tilde{x}_2\tilde{x}_3 = \tilde{\pi}_2\tilde{\pi}_3, \ \tilde{x}_3\tilde{x}_1 = \tilde{\pi}_3\tilde{\pi}_1,$$

whose solution is $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = \pm(\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\pi}_3)$. Equivalently, the solution for $u_0 = u_1 = u_2 = u_3$ is $\boldsymbol{x} = \boldsymbol{\pi}$ or $\mathbf{1} - \boldsymbol{\pi}$.

Case 2: $x_1 + x_2 + x_3 = \frac{3}{2}$**.** Then, with $\tilde{x}_i = \frac{1}{2} - x_i$ for $i \in [3]$, the case condition is equivalent to $\tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 = 0$. Using this fact in (7.13), we get $\tilde{x}_1 u_1 + \tilde{x}_2 u_2 + \tilde{x}_3 u_3 = 0$. Thus, the equations

in $\mathcal{X}\boldsymbol{u} = \mathbf{0}$ can be simplified to the following

$$\begin{aligned}
\tilde{x}_1 \left(3u_1 - u_0 - u_2 - u_3\right) &= 0, \\
\tilde{x}_2 \left(3u_2 - u_0 - u_3 - u_1\right) &= 0, \\
\tilde{x}_3 \left(3u_3 - u_0 - u_1 - u_2\right) &= 0.
\end{aligned} \tag{7.25}$$

The system in (7.25) can be satisfied only if one of the following two scenarios holds: (i) if $\tilde{x}_i = 0$ or equivalently, $x_i = \frac{1}{2}$ for some $i \in [3]$. This case has been discussed under Case 1-1, and it is shown that $\boldsymbol{x} = \frac{1}{2}\mathbf{1}$ is the only solution; (ii) alternatively, if $\tilde{x}_i \neq 0$ for $i \in [3]$, we should have

$$\begin{aligned}
u_1 = \frac{u_0 + u_2 + u_3}{3} &= \frac{\sum_{i=0}^{3} u_i - u_1}{3}, \\
u_2 = \frac{u_0 + u_3 + u_1}{3} &= \frac{\sum_{i=0}^{3} u_i - u_2}{3}, \\
u_3 = \frac{u_0 + u_1 + u_2}{3} &= \frac{\sum_{i=0}^{3} u_i - u_3}{3}.
\end{aligned} \tag{7.26}$$

This set of equations leads to $u_0 = u_1 = u_2 = u_3$, which is studied under Case 1-2. It is shown that $\boldsymbol{x} = \boldsymbol{\pi}$ and $\mathbf{1} - \boldsymbol{\pi}$ are the only solutions for Case 1-2. This concludes the proof. ■

Proof of Theorem 23. For any $m \in \mathbb{N}$, for $\boldsymbol{x} \in (0,1)^m$ and $\boldsymbol{R} \in \{-1,+1\}^m$ define $\Pi(\boldsymbol{R};\boldsymbol{x})$ as

$$\Pi(\boldsymbol{R};\boldsymbol{x}) := \prod_{i=1}^{m} \left(x_i \mathbb{1}_{\{R_i=1\}} + (1-x_i)\mathbb{1}_{\{R_i=-1\}}\right).$$

Note that summing over all possible realizations of $\boldsymbol{R}$ we get

$$\sum_{\boldsymbol{R}\in\{-1,+1\}^m} \Pi(\boldsymbol{R};\boldsymbol{x}) = \prod_{i=1}^{m}(x_i + (1-x_i)) = 1.$$

For any $\boldsymbol{x} \in (0,1)^n$ and any $\boldsymbol{R} \in \{-1,+1\}^n$ we know that

$$R_i g_{\boldsymbol{x}}(\boldsymbol{R}) \mathcal{B}_{\text{ALL}}(\boldsymbol{R};\boldsymbol{x}) = \frac{1}{2}\left(\Pi(-R_i\boldsymbol{R};\boldsymbol{x}) - \Pi(R_i\boldsymbol{R};\boldsymbol{x})\right)$$
$$= \frac{1}{2}\left((1-x_i)\Pi(-\boldsymbol{R}_{-i};\boldsymbol{x}_{-i}) - x_i\Pi(\boldsymbol{R}_{-i};\boldsymbol{x}_{-i})\right),$$

where $\boldsymbol{x}_{-i} \in (0,1)^{n-1}$ and $\boldsymbol{R}_{-i} \in \{-1,+1\}^{n-1}$ are obtained by removing the ith element in $\boldsymbol{x}$ and $\boldsymbol{R}$, respectively.

Therefore, for any $i \in [n]$ we have

$$\mathbb{E}[R_i \mathcal{B}_{\text{ALL}}(\boldsymbol{R};\boldsymbol{x})] = \sum_{\boldsymbol{R}\in\{-1,+1\}^n} g_{\boldsymbol{x}}(\boldsymbol{R}) R_i \mathcal{B}_{\text{ALL}}(\boldsymbol{R};\boldsymbol{x})$$
$$= \frac{1}{2}\sum_{\boldsymbol{R}\in\{-1,+1\}^n} (1-x_i)\Pi(-\boldsymbol{R}_{-i};\boldsymbol{x}_{-i}) - x_i\Pi(\boldsymbol{R}_{-i};\boldsymbol{x}_{-i})$$
$$= (1-x_i) - x_i = 1 - 2x_i,$$

which concludes the proof. ■

Chapter 7 in full, is a reprint of the material as it appears in A. Verma, S. Mohajer, B. Touri, “Multi-Agent Fact-Checker: Adaptive Estimators," submitted in 2024 *Conference on Decision and Control*. The dissertation author was the primary investigator and author of this paper.

Bibliography

[1] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. Spread of (mis) information in social networks. *Games and Economic Behavior*, 70(2):194–227, 2010.

[2] Daron Acemoglu, Asuman Ozdaglar, and James Siderius. Misinformation: Strategic sharing, homophily, and endogenous echo chambers. Technical report, National Bureau of Economic Research, 2021.

[3] Adel Aghajan and Behrouz Touri. Distributed optimization over dependent random networks. *arXiv preprint arXiv:2010.01956*, 2020.

[4] S. Sh. Alaviani and A. G. Kelkar. Distributed convex optimization with state-dependent interactions over random networks. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3149–3153, 2021.

[5] Seyyed Shaho Alaviani and Nicola Elia. Distributed convex optimization with state-dependent (social) interactions and time-varying topologies. *IEEE Transactions on Signal Processing*, 69:2611–2624, 2021.

[6] Paul S Albert and Lori E Dodd. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2):427–435, 2004.

[7] Christophe Andrieu, Éric Moulines, and Pierre Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on control and optimization*, 44(1):283–312, 2005.

[8] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione. Broadcast gossip algorithms for consensus. *IEEE Transactions on Signal Processing*, 57(7):2748–2761, 2009.

[9] Thomas Bonald and Richard Combes. A minimax optimal algorithm for crowdsourcing. *Advances in Neural Information Processing Systems*, 30, 2017.

[10] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[11] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

[12] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 665–674, 2011.

[13] Han-Fu Chen. *Stochastic approximation and its applications*, volume 64. Springer Science & Business Media, 2005.

[14] Sitan Chen and Ankur Moitra. Beyond the low-degree algorithm: mixtures of subcubes and their applications. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 869–880, 2019.

[15] George Cybenko. Dynamic load balancing for distributed memory multiprocessors. *Journal of parallel and distributed computing*, 7(2):279–301, 1989.

[16] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *J. R. Stat. Soc.: Series C (Applied Statistics)*, 28(1):20–28, 1979.

[17] Marie Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 2013.

[18] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[19] Jon Feldman, Ryan O'Donnell, and Rocco A Servedio. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

[20] Yoav Freund and Yishay Mansour. Estimating a mixture of two product distributions. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 53–62, 1999.

[21] Chao Gao, Yu Lu, and Dengyong Zhou. Exact exponent in optimal rates for crowdsourcing. In *International Conference on Machine Learning*, pages 603–611. PMLR, 2016.

[22] Spencer Gordon, Bijan H Mazaheri, Yuval Rabani, and Leonard Schulman. Source identification for mixtures of product distributions. In *Conference on Learning Theory*, pages 2193–2216. PMLR, 2021.

[23] Spencer L Gordon and Leonard J Schulman. Hadamard extensions and the identification of mixtures of product distributions. *IEEE Transactions on Information Theory*, 2022.

[24] Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

[25] Andreas Hanselowski and Iryna Gurevych. A framework for automated fact-checking for real-time validation of emerging claims on the web. In *NIPS 2017 Workshop on Prioritising Online Content. Long Beach, USA. url: https://www. k4all. org/wp-content/uploads/2017/09/WPOC2017_paper_6. pdf*, 2017.

[26] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214*, 2019.

[27] Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.

[28] Sui L Hui and Steven D Walter. Estimating the error rates of diagnostic tests. *Biometrics*, pages 167–171, 1980.

[29] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100. PMLR, 2015.

[30] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

[31] Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, 2019.

[32] H. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer New York, 2006.

[33] Bernard C Levy. *Principles of signal detection and parameter estimation*. Springer Science & Business Media, 2008.

[34] Ilan Lobel, Asuman Ozdaglar, and Diego Feijer. Distributed multi-agent optimization with state-dependent communication. *Mathematical programming*, 129(2):255–284, 2011.

[35] Carl D Meyer. *Matrix Analysis and Applied Linear Algebra*, volume 71. SIAM, 2000.

[36] Angelia Nedic and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *52nd IEEE Conference on Decision and Control*, Dec 2013.

[37] Angelia Nedic, Alex Olshevsky, Asuman Ozdaglar, and John N Tsitsiklis. On distributed averaging algorithms and quantization effects. *IEEE Transactions on automatic control*, 54(11):2506–2517, 2009.

[38] Angelia Nedic, Alex Olshevsky, Asuman Ozdaglar, and John N Tsitsiklis. On distributed averaging algorithms and quantization effects. *IEEE Transactions on automatic control*, 54(11):2506–2517, 2009.

[39] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[40] Angelia Nedic, Asuman Ozdaglar, and Pablo A Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4):922–938, 2010.

[41] Nam P Nguyen, Guanhua Yan, My T Thai, and Stephan Eidenbenz. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 213–222, 2012.

[42] Shmuel Nitzan and Jacob Paroush. The characterization of decisive weighted majority rules. *Economics Letters*, 7(2):119–124, 1981.

[43] Yiangos Papanastasiou. Fake news propagation and detection: A sequential model. *Management Science*, 66(5):1826–1846, 2020.

[44] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

[45] Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021.

[46] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

[47] S Sundhar Ram, A Nedić, and Venugopal V Veeravalli. Asynchronous gossip algorithms for stochastic optimization. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 3581–3586. IEEE, 2009.

[48] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010.

[49] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of*

mathematical statistics, pages 400–407, 1951.

[50] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[51] Devavrat Shah. *Gossip algorithms*. Now Publishers Inc, 2009.

[52] Lloyd Shapley and Bernard Grofman. Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 43(3):329–343, 1984.

[53] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems*, 7, 1994.

[54] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[55] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

[56] Deniz Ustebay, Boris N Oreshkin, Mark J Coates, and Michael G Rabbat. Greedy gossip with eavesdropping. *IEEE Transactions on Signal Processing*, 58(7):3765–3776, 2010.

[57] Adriaan Van den Bos. *Parameter estimation for scientists and engineers*. John Wiley & Sons, 2007.

[58] A Verma, S Mohajer, and B Touri. Distributed fact checking: Estimating unreliability. In *2024 American Control Conference (ACC)*, 2024.

[59] A Verma, A Sharbafchi, B Touri, and S Mohajer. Distributed fact checking. In *2023 IEEE Int. Symp. on Inform. Theory (ISIT)*, 2023.

[60] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *The Journal of Machine Learning Research*, 17(1):3537–3580, 2016.