A New Approach for Distributed Hypothesis Testing with Extensions to Byzantine-Resilience

Aritra Mitra, John A. Richards and Shreyas Sundaram

Abstract— We study a setting where a group of agents receiving partially informative private observations seek to collaboratively learn the true state (among a set of hypotheses) that explains their joint observation profiles over time. To solve this problem, we propose a distributed learning rule that differs fundamentally from existing approaches, in the sense, that it does not employ any form of "belief-averaging". Specifically, every agent maintains a local belief (on each hypothesis) that is updated in a Bayesian manner without any network influence, and an actual belief that is updated (up to normalization) as the minimum of its own local belief and the actual beliefs of its neighbors. Under minimal requirements on the signal structures of the agents and the underlying static, directed communication graph, we establish consistency of the proposed belief update rule, i.e., we show that the actual beliefs of the agents asymptotically concentrate on the true state almost surely. As one of the key benefits of our approach, we show that our learning rule naturally extends to scenarios that capture misbehavior on the part of certain agents in the network, modeled as per the Byzantine fault model. In particular, we prove that each non-faulty agent can learn the true state of the world almost surely, and identify conditions on the observation model and the network model for this to happen.

I. INTRODUCTION

Various distributed learning problems arising in social networks (such as opinion formation and spread on social and political issues), and in engineering systems (such as target recognition by a group of aerial robots) can be studied under the formal framework of distributed hypothesis testing. Within this framework, a group of agents repeatedly observe certain private signals, and aim to infer the "true state of the world" that explains their joint observations, a task that may prove impossible for any single agent to achieve in isolation. While some of the earlier works on this topic assumed the existence of a centralized fusion center for performing computational tasks [1], [2], the focus of this paper will be on scenarios where interactions among agents are captured by a communication graph. A typical belief update rule in the latter setting combines a local Bayesian update with a consensus-based opinion pooling of neighboring beliefs. Specifically, linear opinion pooling is studied in [3]–[5],

whereas the log linear form of belief aggregation is studied in the context of distributed hypothesis testing in [6]–[10], and distributed parameter estimation in [11], [12]. Notably, exponential convergence rates are achieved in [4], [6]–[9], while a finite-time analysis is presented in [10]. Extensions to time-varying graphs have also been studied in [5]–[7].

In [7, Section III], the authors explain that the commonly studied linear and log-linear forms of belief aggregation are specific instances of a more general class of opinion pooling known as g-Quasi-Linear Opinion pools (g-QLOP), introduced in [13]. The main contribution of this paper is the development of a novel belief update rule that deviates fundamentally from the broad family of g-QLOP learning rules. Specifically, the learning algorithm that we propose in Section III-A does not rely on any linear consensus-based belief aggregation protocol. Instead, each agent maintains two sets of beliefs: a local belief that is updated in a Bayesian manner based on the private observations (without neighbor interactions), and an actual belief that is updated (up to normalization) as the minimum of the agent's own local belief and the actual beliefs of its neighbors. In Section V, we establish that under minimal requirements on the agents' signal structures and the communication graph, the actual beliefs of the agents asymptotically concentrate on the true state almost surely. As we discuss in Remark 3, our approach works under graph-theoretic conditions that are in general weaker than the standard assumption of strong-connectivity.

Despite the wealth of literature on distributed inference, there is limited understanding of the impact of misbehaving agents for the problem under consideration. Such agents may represent stubborn individuals, religious extremists in the context of a social network, or model faults (either benign or malicious) in a networked control system. *In the presence of such misbehaving entities, how should the remaining agents process their private observations and the beliefs of their neighbors to eventually learn the truth?* The second contribution of this paper is to provide an answer to this question. To do so, we model misbehaving agents as per the classical Byzantine fault model, and develop a provably correct fault-tolerant version of our proposed learning rule in Section III-B. The only related work (that we are aware of) in this regard is reported in [9]. As we discuss in Remark 2, our proposed approach is significantly less computationally intensive relative to those in [9]. We identify conditions on the observation model and the network structure that guarantee applicability of our Byzantine-resilient learning rule, and argue (in Remark 6) that such conditions can be checked in polynomial time. To establish consistency of our

learning algorithms, we develop novel analysis techniques in Section V, which may be of independent interest.

II. MODEL AND PROBLEM FORMULATION

Network Model: We consider a group of agents $\mathcal{V} = \{1, 2, \ldots, n\}$ interacting over a time-invariant, directed communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. An edge $(i, j) \in \mathcal{E}$ indicates that agent i can directly transmit information to agent j. If $(i, j) \in \mathcal{E}$, then agent i will be called a neighbor of agent j, and agent j will be called an out-neighbor of agent i. The set of all neighbors of agent i will be denoted $\mathcal{N}_i$. Given two disjoint sets $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathcal{V}$, we say that $\mathcal{C}_2$ is reachable from $\mathcal{C}_1$ if for every $i \in \mathcal{C}_2$, there exists a directed path from some $j \in \mathcal{C}_1$ to agent i (note that j will in general be a function of i). We will use $|\mathcal{C}|$ to denote the cardinality of a set $\mathcal{C}$.

Observation Model: Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$ denote m possible states of the world; each $\theta_i \in \Theta$ will be called a hypothesis. Let $\mathbb{N}$ and $\mathbb{N}_+$ denote the set of non-negative integers and positive integers, respectively. Then at each time-step $t \in \mathbb{N}_+$, every agent $i \in \mathcal{V}$ privately observes a signal $s_{i,t} \in \mathcal{S}_i$, where $\mathcal{S}_i$ denotes the signal space of agent i. The joint observation profile so generated across the network is denoted $s_t = (s_{1,t}, s_{2,t}, \ldots, s_{n,t})$, where $s_t \in \mathcal{S}$, and $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \ldots \mathcal{S}_n$. The signal s_t is generated based on a conditional likelihood function $l(\cdot|\theta^\star)$, governed by the true state of the world $\theta^\star \in \Theta$. Let $l_i(\cdot|\theta^\star), i \in \mathcal{V}$ denote the i-th marginal of $l(\cdot|\theta^\star)$. The signal structure of each agent $i \in \mathcal{V}$ is then characterized by a family of parameterized marginals $\{l_i(w_i|\theta) : \theta \in \Theta, w_i \in \mathcal{S}_i\}$.[1]

We make the following standard assumptions: (i) The signal space of each agent i, namely $\mathcal{S}_i$, is finite, (ii) Each agent i has knowledge of its local likelihood functions $\{l_i(\cdot|\theta_p)\}_{p=1}^m$, and it holds that $l_i(w_i|\theta) > 0, \forall w_i \in \mathcal{S}_i$, and $\forall \theta \in \Theta$, (iii) The observation sequence of each agent is described by an i.i.d. random process over time; however, at any given time-step, the observations of the agents may potentially be correlated, (iv) there exists a fixed true state of the world $\theta^\star \in \Theta$ (that is unknown to the agents) that explains the observations of all the agents. Finally, we define a probability triple $(\Omega, \mathcal{F}, \mathbb{P}^{\theta^\star})$, where $\Omega \triangleq \{\omega : \omega = (s_1, s_2, \ldots), \forall s_t \in \mathcal{S}, \forall t \in \mathbb{N}_+\}$, $\mathcal{F}$ is the σ-algebra generated by the observation profiles, and $\mathbb{P}^{\theta^\star}$ is the probability measure induced by sample paths in Ω. Specifically, $\mathbb{P}^{\theta^\star} = \prod_{t=1}^{\infty} l(\cdot|\theta^\star)$. For the sake of brevity, we will say that an event occurs almost surely to mean that it occurs almost surely w.r.t. the probability measure $\mathbb{P}^{\theta^\star}$.

Given the above setup, the goal of each agent in the network is to discern the true state of the world $\theta^\star$. The challenge associated with such a task stems from the fact that the private signal structure of any given agent is in general only partially informative. To make this notion precise, define $\Theta_i^{\theta^\star} \triangleq \{\theta \in \Theta : l_i(w_i|\theta) = l_i(w_i|\theta^\star), \forall w_i \in \mathcal{S}_i\}$. In words, $\Theta_i^{\theta^\star}$ represents the set of hypotheses that are *observationally*

equivalent to the true state $\theta^\star$, from the perspective of agent i. In general, for any agent $i \in \mathcal{V}$, $|\Theta_i^{\theta^\star}| > 1$, necessitating collaboration among agents. While inter-agent collaboration is implicitly assumed in the distributed hypothesis testing literature, we will allow for potential misbehavior on the part of certain agents in the network.

Adversary Model: We assume that a certain fraction of the agents are adversarial, and model their behavior based on the Byzantine fault model [14]. Accordingly, Byzantine agents possess complete knowledge of the observation model, the network model, the algorithms being used, the information being exchanged, and the true state of the world. Leveraging such information, adversarial agents can behave arbitrarily, and can in particular, send incorrect, potentially inconsistent information to their out-neighbors. In terms of their distribution in the network, we will consider an f-local adversarial model, i.e., we assume that there are at most f adversaries in the neighborhood of any non-adversarial agent.[2] Finally, we emphasize that the non-adversarial agents are unaware of the identities of the adversaries in their neighborhood. As is fairly standard in the distributed fault-tolerant literature (see, for example [15]–[21]), we assume that non-adversarial agents know the upper bound f on the number of adversaries in their neighborhood. The adversarial set will be denoted by $\mathcal{A} \subset \mathcal{V}$, and the remaining agents $\mathcal{R} = \mathcal{V} \setminus \mathcal{A}$ will be called the regular agents.

Our **objective** in this paper will be to design a distributed learning rule that allows each regular agent $i \in \mathcal{R}$ to identify the true state of the world almost surely, despite (i) the partially informative signal structures of the agents, and (ii) the actions of any f-local Byzantine adversarial set. To this end, we introduce the following notion of *source agents*.

Definition 1. *(**Source agents**) An agent i is said to be a source agent for a pair of distinct hypotheses $\theta_1, \theta_2 \in \Theta$, if $D(l_i(\cdot|\theta_1)||l_i(\cdot|\theta_2)) > 0$, where $D(l_i(\cdot|\theta_1)||l_i(\cdot|\theta_2))$ represents the KL-divergence between the distributions $l_i(\cdot|\theta_1)$ and $l_i(\cdot|\theta_2)$, and is given by:*

$$D(l_i(\cdot|\theta_1)||l_i(\cdot|\theta_2)) = \sum_{w_i \in \mathcal{S}_i} l_i(w_i|\theta_1) \log \frac{l_i(w_i|\theta_1)}{l_i(w_i|\theta_2)}. \quad (1)$$

The set of all source agents for the pair θ_1, θ_2 is denoted by $\mathcal{S}(\theta_1, \theta_2)$.[3] □

In words, a source agent for a pair θ_1, θ_2 is an agent that can distinguish between the pair of hypotheses θ_1, θ_2 based on its private signal structure. In our developments, we will require the following two definitions.

Definition 2. *(**r-reachable set**) [17] For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a set $\mathcal{C} \subset \mathcal{V}$, and an integer $r \in \mathbb{N}_+$, $\mathcal{C}$ is an r-reachable set if there exists an $i \in \mathcal{C}$ such that $|\mathcal{N}_i \setminus \mathcal{C}| \geq r$,* □

Definition 3. *(**strongly r-robust graph** w.r.t. $\mathcal{S}(\theta_1, \theta_2)$) For $r \in \mathbb{N}_+$ and $\theta_1, \theta_2 \in \Theta$, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is strongly*

---

[1] Whereas $w_i \in \mathcal{S}_i$ will be used to refer to a generic element of the signal space of agent i, $s_{i,t}$ will denote the random variable (with distribution $l_i(\cdot|\theta^\star)$) that corresponds to agent i's observation at time-step t.

[2] Note that the f-local adversarial model assumed here is more general than the f-total adversarial model considered in [9], where the total number of adversaries in the entire network is upper bounded by f.

[3] Notice that $\mathcal{S}(\theta_1, \theta_2) = \mathcal{S}(\theta_2, \theta_1)$, since $D(l_i(\cdot|\theta_1)||l_i(\cdot|\theta_2)) > 0 \iff D(l_i(\cdot|\theta_2)||l_i(\cdot|\theta_1)) > 0$.

r-robust w.r.t. the set of source agents $\mathcal{S}(\theta_1, \theta_2)$, if for any non-empty subset $\mathcal{C} \subseteq \mathcal{V} \setminus \mathcal{S}(\theta_1, \theta_2)$, $\mathcal{C}$ is r-reachable. □

III. PROPOSED LEARNING RULES

A. A Novel Belief Update Rule

In this section, we propose a novel belief update rule and discuss the intuition behind it. To introduce the key ideas underlying our basic approach, we first consider a scenario where all agents are regular, i.e., $\mathcal{R} = \mathcal{V}$. Every agent i maintains and updates (at every time-step) two separate sets of belief vectors, namely, $\boldsymbol{\pi_{i,t}}$ and $\boldsymbol{\mu_{i,t}}$. Each of these vectors are probability distributions over the hypothesis set Θ. We will refer to $\boldsymbol{\pi_{i,t}}$ and $\boldsymbol{\mu_{i,t}}$ as the "local" belief vector (for reasons that will soon become obvious), and the "actual" belief vector, respectively, maintained by agent i. The **goal** of each agent $i \in \mathcal{V}$ in the network will be to use its own private signals, and the information available from its neighbors to update $\boldsymbol{\mu_{i,t}}$ sequentially, so that $\lim_{t\to\infty} \mu_{i,t}(\theta^*) = 1$ almost surely. To do so, for each $\theta \in \Theta$, and at each time-step $t + 1, t \in \mathbb{N}$, agent i first generates $\pi_{i,t+1}(\theta)$ via a local Bayesian update rule that incorporates the private observation $s_{i,t+1}$ using $\pi_{i,t}(\theta)$ as a prior. Having generated $\pi_{i,t+1}(\theta)$, agent i updates $\mu_{i,t+1}(\theta)$ (up to normalization) by setting it to be the minimum of its locally generated belief $\pi_{i,t+1}(\theta)$, and the actual beliefs $\mu_{j,t}(\theta), j \in \mathcal{N}_i$ of each of its neighbors at the previous time-step. It then reports its actual belief $\mu_{i,t+1}(\theta)$ to each of its out-neighbors.[4] Formally, for each $\theta \in \Theta$, and at each time-step $t + 1$ (where $t \in \mathbb{N}$), agent i updates $\mu_{i,t+1}(\theta)$ based on the following two steps.

- **Step 1: Update of the local beliefs**:

$$\pi_{i,t+1}(\theta) = \frac{l_i(s_{i,t+1}|\theta)\pi_{i,t}(\theta)}{\sum\limits_{p=1}^{m} l_i(s_{i,t+1}|\theta_p)\pi_{i,t}(\theta_p)}. \quad (2)$$

- **Step 2: Update of the actual beliefs**:

$$\mu_{i,t+1}(\theta) = \frac{\min\{\min_{j\in\mathcal{N}_i}\{\mu_{j,t}(\theta)\}, \pi_{i,t+1}(\theta)\}}{\sum\limits_{p=1}^{m} \min\{\min_{j\in\mathcal{N}_i}\{\mu_{j,t}(\theta_p)\}, \pi_{i,t+1}(\theta_p)\}}. \quad (3)$$

Intuition behind the learning rule: Consider the set of source agents $\mathcal{S}(\theta, \theta^*)$ who can differentiate between the hypothesis θ and the true state θ^*. Suppose for now that this set is non-empty. We ask: How do the agents in the set $\mathcal{S}(\theta, \theta^*)$ contribute to the process of collaborative learning? To answer this question, we note that the signal structures of such agents are rich enough for them to be able to eliminate θ on their own, i.e., without the support of their neighbors. Thus, the agents in $\mathcal{S}(\theta, \theta^*)$ should contribute towards driving the actual beliefs of their out-neighbors (and eventually, of all the agents in the set $\mathcal{V} \setminus \mathcal{S}(\theta, \theta^*)$) on the hypothesis θ to zero. To achieve the above objective, we are especially interested in devising a rule that ensures that the capability of the source agents $\mathcal{S}(\theta, \theta^*)$ to eliminate θ is not

---

[4]Note that based on our algorithm, agents only exchange their actual beliefs, and not their local beliefs.

diminished due to neighbor interactions. As we shall see later on, such a property will be particularly useful when certain agents in the network are adversarial. It is precisely these considerations that motivate us to employ (i) an auxiliary belief vector $\boldsymbol{\pi_{i,t+1}}$ generated via local processing (i.e., without any network influence) of the private signals, and (ii) a min-rule of the form (3). Specifically, if $i \in \mathcal{S}(\theta, \theta^*)$, then the sequence of local beliefs $\pi_{i,t+1}(\theta)$ will almost surely converge to 0 based on the update rule (2). Hence, for a source agent $i \in \mathcal{S}(\theta, \theta^*)$, $\pi_{i,t+1}(\theta)$ will play the key role of an external network-independent input in the min-rule (3). This in turn will trigger a process of belief reduction (on the hypothesis θ) originating at the source set $\mathcal{S}(\theta, \theta^*)$, and eventually propagating via the proposed min-rule to each agent in the network that is reachable from such a source set. The above discussion will be made precise in Section V.

Remark 1. *As is fairly evident, the proposed min-rule (3) does not employ any form of "belief-averaging". This feature is in stark contrast with existing approaches to distributed hypothesis testing that rely either on linear opinion pooling [3]–[5], or log-linear opinion pooling [6]–[12]. As such, the lack of linearity in our belief update rule precludes (direct or indirect) adaptation of existing analysis techniques to suit our needs. Consequently, we develop a novel sample path based proof technique in Section V to establish consistency of the proposed learning rule. As one of the main outcomes of this analysis, we argue that our learning rule works under graph-theoretic conditions that are in general weaker than strong-connectivity (see also Remark 3).* □

As pointed out in the Introduction, our primary aim is the design of a distributed learning rule that is worst-case fault-tolerant. While [9] offers a potential solution to this problem, the following discussion motivates the alternate approach developed in Section III-B.

Remark 2. *A standard way to analyze the impact of adversarial agents while designing resilient distributed consensus-based protocols (for applications in consensus [15], [17], optimization [16], [18], hypothesis testing [9], and multi-agent rendezvous [22]) is to construct an equivalent matrix representation of the linear update rule that involves only the regular agents [23]. In particular, this requires expressing the iterates of a regular agent as a convex combination of the iterates of its regular neighbors, based on appropriate filtering techniques, and under certain assumptions on the network structure. While this can indeed be achieved efficiently for scalar consensus problems, for problems requiring consensus on vectors (such as the belief vectors in our setting), this leads to the computation of Tverberg partitions. However, there is no known algorithm that can compute an exact Tverberg partition in polynomial time for a general d-dimensional finite point set [24]. Consequently, since the filtering approach developed in [9] requires each regular agent to compute a Tverberg partition at every iteration, the resulting computations are forbiddingly high. The authors in [9] do briefly discuss an alternate pairwise learning rule that requires agents to perform scalar consensus on relative*

confidence levels (instead of beliefs) of one hypothesis over another. This in turn requires each agent to maintain and update (at each time-step) a vector of dimension $O(m^2)$. In contrast, the algorithm developed in Section III-B does not require the computation of Tverberg partitions, and requires agents to maintain and update (at each time-step) two belief vectors, each of dimension m. $\square$

B. A Byzantine-Resilient Belief Update Rule

In this section, we develop a belief update rule that is able to tolerate the worst-case Byzantine adversarial model described in Section II. We achieve this via a suitable modification of the algorithm introduced in Section III-A. Specifically, each agent $i \in \mathcal{R}$ acts as follows at every time-step $t+1$ (where $t \in \mathbb{N}$).

- **Step 1: Update of the local beliefs**: The local belief $\pi_{i,t+1}(\theta)$ is updated as before, based on (2).

- **Step 2: Filtering extreme beliefs**: If $|\mathcal{N}_i| \geq (2f+1)$, then agent i performs a filtering operation as follows. It collects the actual beliefs $\mu_{j,t}(\theta)$ from each of its neighbors and sorts them from highest to lowest. It rejects the highest f and the lowest f of such beliefs (i.e., it throws away $2f$ beliefs in all).

- **Step 3: Update of the actual beliefs**: If $|\mathcal{N}_i| \geq (2f+1)$, then agent i updates $\mu_{i,t+1}(\theta)$ as follows. Let the set of neighbors whose beliefs are not rejected by agent i (based on the previous filtering step) be denoted by $\mathcal{M}_{i,t}^{\theta} \subset \mathcal{N}_i$. The actual belief $\mu_{i,t+1}(\theta)$ is then updated as follows:

$$\mu_{i,t+1}(\theta) = \frac{\min\{\min_{j \in \mathcal{M}_{i,t}^{\theta}}\{\mu_{j,t}(\theta)\}, \pi_{i,t+1}(\theta)\}}{\sum_{p=1}^{m} \min\{\min_{j \in \mathcal{M}_{i,t}^{\theta_p}}\{\mu_{j,t}(\theta_p)\}, \pi_{i,t+1}(\theta_p)\}}.$$
(4)

If $|\mathcal{N}_i| < (2f+1)$, then agent i updates $\mu_{i,t+1}(\theta)$ as follows:

$$\mu_{i,t+1}(\theta) = \pi_{i,t+1}(\theta).$$
(5)

As with the learning rule presented in Section III-A, agent i transmits $\mu_{i,t+1}(\theta)$ to each of its out-neighbors on completion of the above steps. We will refer to the above sequence of actions as the Local-Filtering based Resilient Hypothesis Elimination (LFRHE) algorithm.

IV. MAIN RESULTS

Our first result establishes the correctness of the learning rule proposed in Section III-A.

Theorem 1. *Suppose $\mathcal{R} = \mathcal{V}$, and that the following are true:*

(i) *For every pair of hypotheses $\theta_1, \theta_2 \in \Theta$, the corresponding source set $\mathcal{S}(\theta_1, \theta_2)$ is non-empty.*
(ii) *For every pair of hypotheses $\theta_1, \theta_2 \in \Theta$, $\mathcal{V} \setminus \mathcal{S}(\theta_1, \theta_2)$ is reachable from the source set $\mathcal{S}(\theta_1, \theta_2)$.*
(iii) *Every agent $i \in \mathcal{V}$ has a non-zero prior belief on each hypothesis, i.e., $\pi_{i,0}(\theta) > 0, \mu_{i,0}(\theta) > 0$ for all $i \in \mathcal{V}$, and for all $\theta \in \Theta$.*

Then, the learning rule described by equations (2) and (3) leads to collaborative learning of the true parameter, i.e., $\mu_{i,t}(\theta^\star) \to 1$ almost surely for all $i \in \mathcal{V}$. $\square$

Our second result establishes the correctness of the LFRHE algorithm proposed in Section III-B.

Theorem 2. *Suppose the following are true:*

(i) *For every pair of hypotheses $\theta_1, \theta_2 \in \Theta$, the graph $\mathcal{G}$ is strongly $(2f+1)$-robust w.r.t. the corresponding source set $\mathcal{S}(\theta_1, \theta_2)$.*
(ii) *Each regular agent $i \in \mathcal{R}$ has a non-zero prior belief on each hypothesis, i.e., $\pi_{i,0}(\theta) > 0, \mu_{i,0}(\theta) > 0$ for all $i \in \mathcal{R}$, and for all $\theta \in \Theta$.*

Then, the LFRHE algorithm described by equations (2), (4) and (5) leads to collaborative learning of the true parameter despite the actions of any f-local set of Byzantine adversaries, i.e., $\mu_{i,t}(\theta^\star) \to 1$ almost surely for all $i \in \mathcal{R}$. $\square$

In the following two remarks, we comment on the generality of the assumptions made in Theorems 1 and 2.

Remark 3. *(Assumptions in Theorem 1)* While the first condition in Theorem 1 is a basic global identifiability condition phrased differently, the second condition on the network structure is in general weaker than the standard assumption of strong-connectivity made in [3], [4], [8], [10]–[12]. To see why the latter statement is true, consider a scenario where $\Theta = \{\theta_1, \theta_2\}$. Clearly, any agent $i \in \mathcal{S}(\theta_1, \theta_2)$ can discern the true state without neighbor interactions, precluding the need for incoming edges to such agents.[5] Finally, the assumption of non-zero initial beliefs is fairly standard, and can be easily met by maintaining a uniform support over the hypotheses set initially. $\square$

Remark 4. *(Assumptions in Theorem 2)* The first condition in Theorem 2 blends requirements on the signal structures of the agents with those on the communication graph. To gain intuition about this condition, suppose $\Theta = \{\theta_1, \theta_2\}$, and let there exist at least one agent $i \in \mathcal{V} \setminus \mathcal{S}(\theta_1, \theta_2)$. To enable agent i to discern the truth despite potential adversaries in its neighborhood, one requires (i) redundancy in the signal structures of the agents,[6] and (ii) redundancy in the network structure to facilitate reliable information flow from $\mathcal{S}(\theta_1, \theta_2)$ to agent i. These precise requirements motivate condition (i), a point made apparent in the proof of Theorem 2. $\square$

Remark 5. *(Analogy with Distributed State Estimation)* Consider the problem of collaboratively estimating the state of an LTI process based on information exchanges among agents that receive partial measurements of the state. There are natural connections between the setting described above,

---

[5]For the problem under consideration, the argument that the strong connectivity assumption can be relaxed applies to more general scenarios as well, where there does not necessarily exist any one agent that can identify the true state based on just its private signal structure. The underlying reason for this stems from information heterogeneity and information redundancy among agents, features that are unique to distributed estimation and detection type problems, but lacking in a standard consensus setting.

[6]In particular, for any pair $\theta_1, \theta_2 \in \Theta$, notice that condition (i) of Theorem 2 requires $|\mathcal{S}(\theta_1, \theta_2)| \geq (2f+1)$, if $\mathcal{V} \setminus \mathcal{S}(\theta_1, \theta_2)$ is non-empty.

and the problem studied in this paper. For the state estimation scenario, one can fix an unstable mode of the process, and define source agents for that mode to be agents that can detect the eigenspaces associated with that mode. Interestingly, with source agents defined for each unstable mode in the manner described above, [25, Theorem 3] and [19, Theorem 7] (in the context of distributed state estimation) can be viewed as analogues to Theorem 1 and Theorem 2, respectively (modulo the requirements of non-zero priors). □

Remark 6. *Given a network of agents with associated signal structures, condition (i) in Theorem 2 can be checked in polynomial time. Specifically, for every pair $\theta_1, \theta_2 \in \Theta$, finding the source set $\mathcal{S}(\theta_1, \theta_2)$ can easily be done in polynomial time via inspection of the agents' signal structures. For a fixed source set $\mathcal{S}(\theta_1, \theta_2)$, checking whether $\mathcal{G}$ is strongly $(2f + 1)$-robust w.r.t. $\mathcal{S}(\theta_1, \theta_2)$ amounts to simulating a bootstrap percolation process on $\mathcal{G}$, with $\mathcal{S}(\theta_1, \theta_2)$ as the initial active set, and $(2f + 1)$ as the threshold. This too can be achieved in polynomial time, as discussed in [19].* □

Remark 7. *(Convergence Rate) Consider any false hypothesis $\theta \neq \theta^\star$. We conjecture that based on our learning rules, the actual beliefs of all the regular agents on θ will (almost surely) decay exponentially fast after a transient period, with the rate of decay lower bounded by $\min_{i \in \mathcal{S}(\theta, \theta^\star) \cap \mathcal{R}} D(l_i(\cdot|\theta^\star)||l_i(\cdot|\theta))$.* □

V. PROOFS OF THE MAIN RESULTS

We start with the following lemma that characterizes the asymptotic behavior of the local belief sequences generated based on (2). The result is standard; we provide a proof (adapted to our notation) to keep the paper self-contained.

Lemma 1. *Consider an agent $i \in \mathcal{S}(\theta, \theta^\star) \cap \mathcal{R}$. Suppose $\pi_{i,0}(\theta^\star) > 0$. Then, the update rule (2) ensures that (i) $\pi_{i,t}(\theta) \to 0$ almost surely, and (ii) $\pi_{i,\infty}(\theta^\star) \triangleq \lim_{t \to \infty} \pi_{i,t}(\theta^\star)$ exists almost surely, and satisfies $\pi_{i,\infty}(\theta^\star) \geq \pi_{i,0}(\theta^\star)$.* □

Proof. Pick an agent $i \in \mathcal{S}(\theta, \theta^\star) \cap \mathcal{R}$, and define:

$$\rho_{i,t}(\theta) \triangleq \log \frac{\pi_{i,t}(\theta)}{\pi_{i,t}(\theta^\star)}, \quad \lambda_{i,t}(\theta) \triangleq \log \frac{l_i(s_{i,t}|\theta)}{l_i(s_{i,t}|\theta^\star)}. \quad (6)$$

Then, based on (2), we obtain the following recursion:

$$\rho_{i,t+1}(\theta) = \rho_{i,t}(\theta) + \lambda_{i,t+1}(\theta), \forall t \in \mathbb{N}. \quad (7)$$

Rolling out the above equation over time yields:

$$\rho_{i,t}(\theta) = \rho_{i,0}(\theta) + \sum_{k=1}^{t} \lambda_{i,k}(\theta), \forall t \in \mathbb{N}_+. \quad (8)$$

Notice that $\{\lambda_{i,t}\}$ is a sequence of i.i.d. random variables with finite means and variances. In particular, it is easy to verify that each random variable $\lambda_{i,t}$ has mean[7] given by $-D(l_i(\cdot|\theta^\star)||l_i(\cdot|\theta))$. Thus, based on the strong law of large

---

[7]More precisely, the mean here is obtained by using the expectation operator $\mathbb{E}^{\theta^\star}[\cdot]$ associated with the measure $\mathbb{P}^{\theta^\star}$.

---

numbers, we have $\frac{1}{t} \sum_{k=1}^{t} \lambda_{i,k}(\theta) \to -D(l_i(\cdot|\theta^\star)||l_i(\cdot|\theta))$ almost surely. Dividing both sides of (8) by t, and taking the limit as t goes to infinity, we then obtain:

$$\lim_{t \to \infty} \frac{1}{t} \rho_{i,t}(\theta) = -D(l_i(\cdot|\theta^\star)||l_i(\cdot|\theta)) \text{ almost surely.} \quad (9)$$

Finally, note that based on the definition of the set $\mathcal{S}(\theta, \theta^\star)$, $D(l_i(\cdot|\theta^\star)||l_i(\cdot|\theta)) > 0$. It then follows from (9) that $\rho_{i,t}(\theta) \to -\infty$ almost surely, and hence $\pi_{i,t}(\theta) \to 0$ almost surely. Part (ii) of the lemma is obvious. □

We are now in position to prove Theorem 1.

Proof. (**Theorem 1**) Let $\bar{\Omega} \subseteq \Omega$ denote the set of sample paths along which for each agent $i \in \mathcal{V}$, the following hold: (i) for each $\theta \in \Theta \setminus \Theta_i^{\theta^\star}$, $\pi_{i,t}(\theta) \to 0$, and (ii) $\pi_{i,\infty}(\theta^\star) \triangleq \lim_{t \to \infty} \pi_{i,t}(\theta^\star)$ exists, and satisfies $\pi_{i,\infty}(\theta^\star) \geq \pi_{i,0}(\theta^\star)$. Recall that $\Theta_i^{\theta^\star}$ represents the set of hypotheses that are observationally equivalent to the true state $\theta^\star$ from the point of view of agent i. Hence, for each $\theta \in \Theta \setminus \Theta_i^{\theta^\star}$, we have $i \in \mathcal{S}(\theta, \theta^\star)$. Based on the third condition in the statement of Theorem 1, and Lemma 1, we infer that $\bar{\Omega}$ has measure 1. Thus, to prove the desired result, it suffices to confine our attention to the set $\bar{\Omega}$. Specifically, fix any sample path $\omega \in \bar{\Omega}$, and pick any $\epsilon > 0$. Our goal will be to establish that along the sample path ω, there exists $t(\epsilon)$ such that for all $t \geq t(\epsilon)$, $\mu_{i,t}(\theta) < \epsilon$ for all $i \in \mathcal{V}$, and for all $\theta \neq \theta^\star$. This would be equivalent to establishing that the actual beliefs of all the agents on the true state can be made arbitrarily close to 1 (since the proposed min-rule (3) generates a valid probability distribution over the hypothesis set at each time-step). We complete the proof in the following two steps.

Step 1: *Lower bounding the actual beliefs on the true state*: Consider the following scenario. During a transient phase, certain agents see private signals that cause them to temporarily lower their local beliefs on the true state. This in turn gets manifested via the min-rule (3) to the actual beliefs of the agents in the network. Can this transient phenomenon trigger a cascade of progressively low beliefs on the true state? For sample paths in the set $\bar{\Omega}$, we rule out this possibility. To this end, define $\gamma_1 \triangleq \min_{i \in \mathcal{V}} \pi_{i,0}(\theta^\star)$. Notice that $\gamma_1 > 0$ based on condition (iii) of the theorem. Given the choice of the sample path ω, we notice that $\pi_{i,\infty}(\theta^\star)$ exists for each $i \in \mathcal{V}$, and that $\pi_{i,\infty}(\theta^\star) \geq \gamma_1$. Pick a very small number $\delta > 0$ such that $\delta < \gamma_1$. The following statement is then immediate. For each agent $i \in \mathcal{V}$, there exists $t_i(\delta)$, such that for all $t \geq t_i(\delta)$, $\pi_{i,t}(\theta^\star) \geq \gamma_1 - \delta > 0$. Define $\bar{t}_1(\delta) \triangleq \max_{i \in \mathcal{V}} t_i(\delta)$. In words, $\bar{t}_1(\delta)$ represents the time-step beyond which the local beliefs of all the agents on the true state are lower-bounded by $\gamma_1 - \delta$. We ask: At such a time-step, what is the lowest actual belief held by an agent on the true state? More precisely, we define $\gamma_2 \triangleq \min_{i \in \mathcal{V}} \{\mu_{i,\bar{t}_1(\delta)}(\theta^\star)\}$. We claim $\gamma_2 > 0$. To see this, observe that given the assumption of non-zero prior beliefs on the true state, and the structure of the proposed min-rule (3), γ_2 can be 0 if and only if there exists some $0 < t' \leq \bar{t}_1(\delta)$ such that $\pi_{i,t'}(\theta^\star) = 0$, for some $i \in \mathcal{V}$. However, given the

structure of the local Bayesian update rule (2), we would then have $\pi_{i,t}(\theta^\star) = 0$, for all $t \geq t'$, contradicting the fact that $\pi_{i,t}(\theta^\star) > \gamma_1 - \delta > 0, \forall t \geq \bar{t}_1(\delta) \geq t', \forall i \in \mathcal{V}$ (the latter fact has already been established above). Having thus established that $\gamma_2 > 0$, define $\eta \triangleq \min\{\gamma_1 - \delta, \gamma_2\} > 0$. In words, η represents the lowest belief (considering both local and actual beliefs) on the true state $\theta^\star$ held by an agent at time-step $\bar{t}_1(\delta)$. We claim the following:

$$\mu_{i,t}(\theta^\star) \geq \eta, \forall t \geq \bar{t}_1(\delta), \forall i \in \mathcal{V}. \qquad (10)$$

To see why (10) is true, fix an agent $i \in \mathcal{V}$, and consider the following chain of inequalities:

$$
\begin{aligned}
\mu_{i,\bar{t}_1(\delta)+1}(\theta^\star) &\overset{(a)}{=} \frac{\min\{\min_{j \in \mathcal{N}_i}\{\mu_{j,\bar{t}_1(\delta)}(\theta^\star)\}, \pi_{i,\bar{t}_1(\delta)+1}(\theta^\star)\}}{\sum_{p=1}^{m} \min\{\min_{j \in \mathcal{N}_i}\{\mu_{j,\bar{t}_1(\delta)}(\theta_p)\}, \pi_{i,\bar{t}_1(\delta)+1}(\theta_p)\}} \\
&\overset{(b)}{\geq} \frac{\eta}{\sum_{p=1}^{m} \min\{\min_{j \in \mathcal{N}_i}\{\mu_{j,\bar{t}_1(\delta)}(\theta_p)\}, \pi_{i,\bar{t}_1(\delta)+1}(\theta_p)\}} \\
&\overset{(c)}{\geq} \frac{\eta}{\sum_{p=1}^{m} \pi_{i,\bar{t}_1(\delta)+1}(\theta_p)} \\
&\overset{(d)}{=} \eta,
\end{aligned}
$$
$$(11)$$

where (a) follows from (3), (b) follows from the way η is defined and by noting that $\pi_{i,t}(\theta^\star) \geq \eta, \forall t \geq \bar{t}_1(\delta), \forall i \in \mathcal{V}$, (c) follows trivially via inspection, and (d) follows by noting that the local belief vectors generated via (2) (at each time-step) are valid probability distributions over the hypothesis set Θ, and hence $\sum_{p=1}^{m} \pi_{i,\bar{t}_1(\delta)+1}(\theta_p) = 1$. Since the above reasoning applies to every agent in the network, we can keep repeating it to establish (10) via induction.

Step 2: *Upper bounding the actual beliefs on each false hypothesis*: The key observation that guides the rest of the proof is as follows. While step 1 of the proof guarantees that the beliefs (both local and actual) of each agent on the true state $\theta^\star$ are lower-bounded by η (and hence, bounded away from 0) after a finite period of time (given by $\bar{t}_1(\delta)$), Lemma 1 guarantees that the local beliefs on any false hypothesis θ will eventually become arbitrary small (and in particular, smaller than η) for each agent $i \in \mathcal{S}(\theta, \theta^\star)$.[8] In what follows, we investigate how such an event impacts the actual beliefs of the agents in the network. To this end, given an $\epsilon > 0$, pick a small $\bar{\epsilon} > 0$ such that $\bar{\epsilon} < \min\{\eta, \epsilon\}$. Fix a hypothesis $\theta \neq \theta^\star$. By virtue of condition (i) of the theorem, we know that $|\mathcal{S}(\theta, \theta^\star)| > 0$. Let $q = diam(\mathcal{G}) + 2$, where $diam(\mathcal{G})$ represents the diameter of the graph $\mathcal{G}$. Then, based on Lemma 1, for each $i \in \mathcal{S}(\theta, \theta^\star)$, there exists $t_i(\theta)$ such that for all $t \geq t_i(\theta)$, $\pi_{i,t}(\theta) \leq \bar{\epsilon}^q$. Define

$$\bar{t}_2(\theta) \triangleq \max\{\bar{t}_1(\delta), \max_{i \in \mathcal{S}(\theta, \theta^\star)}\{t_i(\theta)\}\}. \qquad (12)$$

For any agent $i \in \mathcal{S}(\theta, \theta^\star)$, we obtain the following chain of inequalities:

$$
\begin{aligned}
\mu_{i,\bar{t}_2(\theta)+1}(\theta) &\overset{(a)}{=} \frac{\min\{\min_{j \in \mathcal{N}_i}\{\mu_{j,\bar{t}_2(\theta)}(\theta)\}, \pi_{i,\bar{t}_2(\theta)+1}(\theta)\}}{\sum_{p=1}^{m} \min\{\min_{j \in \mathcal{N}_i}\{\mu_{j,\bar{t}_2(\theta)}(\theta_p)\}, \pi_{i,\bar{t}_2(\theta)+1}(\theta_p)\}} \\
&\overset{(b)}{\leq} \frac{\bar{\epsilon}^q}{\sum_{p=1}^{m} \min\{\min_{j \in \mathcal{N}_i}\{\mu_{j,\bar{t}_2(\theta)}(\theta_p)\}, \pi_{i,\bar{t}_2(\theta)+1}(\theta_p)\}} \\
&\overset{(c)}{\leq} \frac{\bar{\epsilon}^q}{\min\{\min_{j \in \mathcal{N}_i}\{\mu_{j,\bar{t}_2(\theta)}(\theta^\star)\}, \pi_{i,\bar{t}_2(\theta)+1}(\theta^\star)\}} \\
&\overset{(d)}{\leq} \frac{\bar{\epsilon}^q}{\eta} \\
&\overset{(e)}{<} \bar{\epsilon}^{(q-1)} < \bar{\epsilon} < \epsilon,
\end{aligned}
$$
$$(13)$$

where (a) follows from (3), (b) follows from the fact that for each $i \in \mathcal{S}(\theta, \theta^\star)$, we have $\pi_{i,t}(\theta) \leq \bar{\epsilon}^q, \forall t \geq \bar{t}_2(\theta)$, (c) follows trivially via inspection, (d) follows from (10) and (12), and (e) follows from the way $\bar{\epsilon}$ has been chosen. In particular, note that the above chain of reasoning used to arrive at (13) applies to subsequent time-steps as well. We thus conclude:

$$\mu_{i,t}(\theta) < \bar{\epsilon}^{(q-1)}, \forall t \geq \bar{t}_2(\theta) + 1, \forall i \in \mathcal{S}(\theta, \theta^\star). \qquad (14)$$

We now wish to investigate how the effect of (14) propagates through the rest of the network. If $\mathcal{V} \setminus \mathcal{S}(\theta, \theta^\star)$ is empty, then we have reached the desired conclusion w.r.t. the false hypothesis θ. If not, define

$$\mathcal{L}_1^{(\theta,\theta^\star)} \triangleq \{i \in \{\mathcal{V} \setminus \mathcal{S}(\theta, \theta^\star)\} : |\mathcal{N}_i \cap \mathcal{S}(\theta, \theta^\star)| > 0\} \qquad (15)$$

as the set of immediate out-neighbors of the source set $\mathcal{S}(\theta, \theta^\star)$. By virtue of condition (ii) of the theorem, if $\mathcal{V} \setminus \mathcal{S}(\theta, \theta^\star)$ is non-empty, then $\mathcal{L}_1^{(\theta,\theta^\star)}$ as defined above is also non-empty. Consider any agent $i \in \mathcal{L}_1^{(\theta,\theta^\star)}$. By definition, agent i has a neighbor in $\mathcal{S}(\theta, \theta^\star)$ satisfying (14). This observation coupled with equations (10), (12) can be used to obtain a similar chain of inequalities as the ones featuring in (13). Specifically, we obtain:

$$\mu_{i,t}(\theta) < \bar{\epsilon}^{(q-2)}, \forall t \geq \bar{t}_2(\theta) + 2, \forall i \in \mathcal{L}_1^{(\theta,\theta^\star)}. \qquad (16)$$

With $\mathcal{L}_0^{(\theta,\theta^\star)} \triangleq \mathcal{S}(\theta, \theta^\star)$, the above arguments can be repeated by successively defining the sets $\mathcal{L}_r^{(\theta,\theta^\star)}, 1 \leq r \leq diam(\mathcal{G})$ as follows:

$$\mathcal{L}_r^{(\theta,\theta^\star)} \triangleq \{i \in \mathcal{V} \setminus \{\bigcup_{c=0}^{r-1} \mathcal{L}_c^{(\theta,\theta^\star)}\} : |\mathcal{N}_i \cap \{\bigcup_{c=0}^{r-1} \mathcal{L}_c^{(\theta,\theta^\star)}\}| > 0\}. \qquad (17)$$

Whenever $\mathcal{V} \setminus \{\bigcup_{c=0}^{r-1} \mathcal{L}_c^{(\theta,\theta^\star)}\}$ is non-empty, condition (ii) of the theorem implies that $\mathcal{L}_r^{(\theta,\theta^\star)}$ will also be non-empty. One can then easily verify via induction on r that:

$$\mu_{i,t}(\theta) < \bar{\epsilon}^{(q-(r+1))}, \forall t \geq \bar{t}_2(\theta) + (r+1), \forall i \in \mathcal{L}_r^{(\theta,\theta^\star)}, \qquad (18)$$

where $1 \leq r \leq diam(\mathcal{G})$. Noting that $q = diam(\mathcal{G}) + 2$, we obtain the desired result that $\mu_{i,t}(\theta) < \bar{\epsilon} < \epsilon, \forall t \geq$

---

[8]We remind the reader that all our arguments are specific to a sample path ω picked at random from the set $\bar{\Omega}$.

$\bar{t}_2(\theta) + diam(G) + 1, \forall i \in \mathcal{V}$. An identical argument as the one presented above can be made for each false hypothesis $\theta \neq \theta^\star$. This completes the proof. $\qquad\square$

Proof. (**Theorem 2**) Consider an f-local adversarial set $\mathcal{A} \subset \mathcal{V}$, and let $\mathcal{R} = \mathcal{V} \setminus \mathcal{A}$. We study two separate cases.

Case 1: Consider a regular agent $i \in \mathcal{R}$ such that $|\mathcal{N}_i| < (2f+1)$. Based on condition (i) of the theorem, we claim that $i \in \mathcal{S}(\theta_1, \theta_2)$, for every pair $\theta_1, \theta_2 \in \Theta$. We prove this claim via contradiction. To do so, suppose there exists a pair $\theta_1, \theta_2 \in \Theta$, such that $i \in \mathcal{V} \setminus \mathcal{S}(\theta_1, \theta_2)$. As $|\mathcal{N}_i| < (2f+1)$, the set $\{i\}$ is clearly not $(2f+1)$-reachable. Thus, $\mathcal{G}$ is not strongly $(2f+1)$-robust w.r.t. the source set $\mathcal{S}(\theta_1, \theta_2)$, a fact that contradicts condition (i) of the theorem. Thus, we have established that for networks satisfying condition (i) of the theorem, regular agents with fewer than $(2f+1)$ neighbors can distinguish between every pair of hypotheses. Lemma 1 then implies that such agents can discern the true state of the world $\theta^\star$ by simply running the local Bayesian estimator (2), and updating their actual beliefs via (5).

Case 2: We now focus only on regular agents i satisfying $|\mathcal{N}_i| \geq (2f+1)$. For this case, the structure of the proof mirrors that of Theorem 1; we thus only elaborate on details that are specific to tackling the aspect of adversarial agents. A key property of the proposed LFRHE algorithm that will be used throughout the proof is as follows. For any $i \in \mathcal{R}$, and any $\theta \in \Theta$, the filtering operation of the LFRHE algorithm ensures that at each time-step $t \in \mathbb{N}$, we have:

$$\mu_{j,t}(\theta) \in Conv(\Psi_{i,t}^\theta), \forall j \in \mathcal{M}_{i,t}^\theta, \qquad (19)$$

where

$$\Psi_{i,t}^\theta \triangleq \{\mu_{j,t}(\theta) : j \in \mathcal{N}_i \cap \mathcal{R}\}, \qquad (20)$$

and $Conv(\Psi_{i,t}^\theta)$ is used to denote the convex hull formed by the points in the set $\Psi_{i,t}^\theta$. In words, any neighboring belief (on a particular hypothesis) that agent i uses in the update rule (4) lies in the convex hull of the actual beliefs of its regular neighbors (on that particular hypothesis). To see why (19) is true, partition the neighbor set $\mathcal{N}_i$ of a regular agent into three sets $\mathcal{U}_{i,t}^\theta$, $\mathcal{M}_{i,t}^\theta$, and $\mathcal{J}_{i,t}^\theta$ as follows. Sets $\mathcal{U}_{i,t}^\theta$ and $\mathcal{J}_{i,t}^\theta$ are each of cardinality f, and contain neighbors of agent i that transmit the highest f and the lowest f actual beliefs respectively, on the hypothesis θ, to agent i at time-step t. The set $\mathcal{M}_{i,t}^\theta$ contains the remaining neighbors of agent i, and is non-empty at every time-step since $|\mathcal{N}_i| \geq (2f+1)$. If $\mathcal{M}_{i,t}^\theta \cap \mathcal{A} = \emptyset$, then (19) holds trivially. Thus, consider the case when there are certain adversaries in the set $\mathcal{M}_{i,t}^\theta$, i.e., $\mathcal{M}_{i,t}^\theta \cap \mathcal{A} \neq \emptyset$. Given the f-locality of the adversarial model, and the nature of the filtering operation in the LFRHE algorithm, we infer that for each $j \in \mathcal{M}_{i,t}^\theta$, there exist regular agents $u, v \in \mathcal{N}_i \cap \mathcal{R}$, such that $u \in \mathcal{U}_{i,t}^\theta$, $v \in \mathcal{J}_{i,t}^\theta$, and $\mu_{v,t}(\theta) \leq \mu_{j,t}(\theta) \leq \mu_{u,t}(\theta)$. This establishes our claim regarding equation (19).

With the above property in hand, our goal will be to now establish each of the two steps in the proof of Theorem 1. To this end, let $\bar{\Omega} \subseteq \Omega$ denote the set of sample paths for which the same set of events as in Theorem 1 occur, with the

exception that such events occur only for the regular agent set $\mathcal{R}$. Based on condition (ii) of the theorem, and Lemma 1, we infer that $\bar{\Omega}$ has measure 1. Thus, as in Theorem 1, fix a sample path $\omega \in \bar{\Omega}$, and pick $\epsilon > 0$. Define $\gamma_1 = \min_{i \in \mathcal{R}} \pi_{i,0}(\theta^\star)$, pick a small number $\delta > 0$ satisfying $\delta < \gamma_1$, and observe that for each agent $i \in \mathcal{R}$, there exists $t_i(\delta)$, such that for all $t \geq t_i(\delta)$, $\pi_{i,t}(\theta^\star) \geq \gamma_1 - \delta > 0$. Define $\bar{t}_1(\delta) \triangleq \max_{i \in \mathcal{R}} t_i(\delta)$ and $\gamma_2 \triangleq \min_{i \in \mathcal{R}}\{\mu_{i,\bar{t}_1(\delta)}(\theta^\star)\}$. As before, we claim $\gamma_2 > 0$. To establish this claim, we need to answer the following question: Can an adversarial agent cause its out-neighbors to set their actual beliefs on $\theta^\star$ to be 0 by setting its own actual belief on $\theta^\star$ to be 0? We argue that this is impossible under the LFRHE algorithm. By way of contradiction, suppose there exists a time-step $\bar{t}$ satisfying:

$$\bar{t} = \inf\{t \in \mathbb{N} : \exists i \in \mathcal{R} \text{ with } \mu_{i,t}(\theta^\star) = 0\}. \qquad (21)$$

In words, $\bar{t}$ represents the first time-step when some regular agent i sets its actual belief on the true hypothesis to be zero. Clearly, $\bar{t} \neq 0$ based on condition (ii) of the theorem. Suppose $\bar{t}$ is some positive integer, and focus on how agent i updates $\mu_{i,\bar{t}}(\theta^\star)$ based on (4). Following similar arguments as in the proof of Theorem 1, we know that $\pi_{i,t}(\theta^\star) > 0, \forall t \in \mathbb{N}, \forall i \in \mathcal{R}$. At the same time, every belief featuring in the set $\Psi_{i,\bar{t}-1}^{\theta^\star}$ (as defined in equation (20)) is strictly positive based on the way $\bar{t}$ is defined. In light of the above arguments, and based on (19), (20), we infer:

$$\min\{\min_{j \in \mathcal{M}_{i,\bar{t}-1}^{\theta^\star}}\{\mu_{j,\bar{t}-1}(\theta^\star)\}, \pi_{i,\bar{t}}(\theta^\star)\} > 0. \qquad (22)$$

Thus, based on (4), we must have $\mu_{i,\bar{t}}(\theta^\star) > 0$, yielding the desired contradiction. With $\eta \triangleq \min\{\gamma_1 - \delta, \gamma_2\} > 0$, one can easily verify the following:

$$\mu_{i,t}(\theta^\star) \geq \eta, \forall t \geq \bar{t}_1(\delta), \forall i \in \mathcal{R}. \qquad (23)$$

In particular, (23) follows by noting that for each $i \in \mathcal{R}$, (i) $\pi_{i,\bar{t}_1(\delta)+1}(\theta^\star) \geq \eta$, (ii) each belief featuring in the set $\Psi_{i,\bar{t}_1(\delta)}^{\theta^\star}$ is lower bounded by η, (iii) leveraging (19), (20), and (iv) by using a similar string of arguments as those used to arrive at (11). This completes Step 1.

To proceed with Step 2, given an $\epsilon > 0$, pick a small $\bar{\epsilon} > 0$ such that $\bar{\epsilon} < \min\{\eta, \epsilon\}$. Fix a hypothesis $\theta \neq \theta^\star$, let $q = diam(\mathcal{G}) + 2$, and note that based on Lemma 1, for each $i \in \mathcal{S}(\theta, \theta^\star) \cap \mathcal{R}$, there exists $t_i(\theta)$ such that for all $t \geq t_i(\theta)$, $\pi_{i,t}(\theta) \leq \bar{\epsilon}^q$. Define

$$\bar{t}_2(\theta) \triangleq \max\{\bar{t}_1(\delta), \max_{i \in \mathcal{S}(\theta,\theta^\star) \cap \mathcal{R}}\{t_i(\theta)\}\}. \qquad (24)$$

For any agent $i \in \mathcal{S}(\theta, \theta^\star) \cap \mathcal{R}$, observe that

$$\min\{\min_{j \in \mathcal{M}_{i,\bar{t}_2(\theta)}^{\theta^\star}}\{\mu_{j,\bar{t}_2(\theta)}(\theta^\star)\}, \pi_{i,\bar{t}_2(\theta)+1}(\theta^\star)\} \geq \eta. \qquad (25)$$

Combining the above with a similar line of arguments as used to arrive at (13), we obtain:

$$\mu_{i,t}(\theta) < \bar{\epsilon}^{(q-1)}, \forall t \geq \bar{t}_2(\theta) + 1, \forall i \in \mathcal{S}(\theta, \theta^\star) \cap \mathcal{R}. \qquad (26)$$

If $\mathcal{V} \setminus \mathcal{S}(\theta, \theta^\star)$ is empty, then we are done. Else, define

$$\mathcal{L}_1^{(\theta,\theta^\star)} \triangleq \{i \in \{\mathcal{V} \setminus \mathcal{S}(\theta, \theta^\star)\} : |\mathcal{N}_i \cap \mathcal{S}(\theta, \theta^\star)| \geq (2f+1)\}. \qquad (27)$$

Whenever $\mathcal{V} \setminus \mathcal{S}(\theta, \theta^\star)$ is non-empty, we claim that $\mathcal{L}_1^{(\theta,\theta^\star)}$ (as defined above) is also non-empty based on condition (i) of the theorem. To see this, note that if $\mathcal{L}_1^{(\theta,\theta^\star)}$ is empty, then $\mathcal{C} = \mathcal{V} \setminus \mathcal{S}(\theta, \theta^\star)$ is not $(2f+1)$-reachable, violating the fact that $\mathcal{G}$ is strongly $(2f+1)$-robust w.r.t. $\mathcal{S}(\theta, \theta^\star)$. We claim

$$\min_{j \in \mathcal{M}^\theta_{i,\bar{t}_2(\theta)+1}} \mu_{j,\bar{t}_2(\theta)+1}(\theta) < \bar{\epsilon}^{(q-1)}, \forall i \in \mathcal{L}_1^{(\theta,\theta^\star)} \cap \mathcal{R}. \quad (28)$$

To verify the above claim, pick any agent $i \in \mathcal{L}_1^{(\theta,\theta^\star)} \cap \mathcal{R}$. When $|\mathcal{M}^\theta_{i,\bar{t}_2(\theta)+1} \cap \{\mathcal{S}(\theta,\theta^\star) \cap \mathcal{R}\}| > 0$, the claim follows immediately based on (26). Consider the case when $|\mathcal{M}^\theta_{i,\bar{t}_2(\theta)+1} \cap \{\mathcal{S}(\theta,\theta^\star) \cap \mathcal{R}\}| = 0$. Since $i \in \mathcal{L}_1^{(\theta,\theta^\star)}$, it has at least $(2f+1)$ neighbors in $\mathcal{S}(\theta,\theta^\star)$, out of which at least $f+1$ are regular based on the f-locality of the adversarial model. Since the set $\mathcal{J}^\theta_{i,\bar{t}_2(\theta)+1}$ has cardinality f, it must then be that $|\mathcal{U}^\theta_{i,\bar{t}_2(\theta)+1} \cap \{\mathcal{S}(\theta,\theta^\star) \cap \mathcal{R}\}| > 0$. Let $u \in \mathcal{U}^\theta_{i,\bar{t}_2(\theta)+1} \cap \{\mathcal{S}(\theta,\theta^\star) \cap \mathcal{R}\}$. Based on the way $\mathcal{M}^\theta_{i,\bar{t}_2(\theta)+1}$ is defined, it must be that $\mu_{j,\bar{t}_2(\theta)+1}(\theta) \leq \mu_{u,\bar{t}_2(\theta)+1}(\theta) < \bar{\epsilon}^{(q-1)}, \forall j \in \mathcal{M}^\theta_{i,\bar{t}_2(\theta)+1}$, where the last inequality follows from (26). This establishes our claim regarding (28). Consider the update of $\mu_{i,\bar{t}_2(\theta)+2}(\theta)$ based on (4). In light of the above arguments, the numerator of the fraction on the RHS of (4) is upper-bounded by $\bar{\epsilon}^{(q-1)}$, while the denominator is lower-bounded by η. This leads to the following conclusion:

$$\mu_{i,t}(\theta) < \bar{\epsilon}^{(q-2)}, \forall t \geq \bar{t}_2(\theta)+2, \forall i \in \mathcal{L}_1^{(\theta,\theta^\star)} \cap \mathcal{R}. \quad (29)$$

With $\mathcal{L}_0^{(\theta,\theta^\star)} \triangleq \mathcal{S}(\theta,\theta^\star)$, we recursively define the sets $\mathcal{L}_r^{(\theta,\theta^\star)}, 1 \leq r \leq diam(\mathcal{G})$ as follows:

$$\mathcal{L}_r^{(\theta,\theta^\star)} \triangleq \{i \in \mathcal{V} \setminus \{\bigcup_{c=0}^{r-1} \mathcal{L}_c^{(\theta,\theta^\star)}\} : |\mathcal{N}_i \cap \{\bigcup_{c=0}^{r-1} \mathcal{L}_c^{(\theta,\theta^\star)}\}| \geq (2f+1)\}. \quad (30)$$

We complete the proof by inducting on r. To this end, suppose the following holds for all $0 \leq r \leq diam(\mathcal{G}) - 1$:

$$\mu_{i,t}(\theta) < \bar{\epsilon}^{(q-(r+1))}, \forall t \geq \bar{t}_2(\theta)+(r+1), \forall i \in \mathcal{L}_r^{(\theta,\theta^\star)} \cap \mathcal{R}. \quad (31)$$

The claim extends to the case when $r = diam(\mathcal{G})$ by noting that (i) $\mathcal{L}_{diam(\mathcal{G})}^{(\theta,\theta^\star)}$ is non-empty if $\mathcal{V} \setminus \{\bigcup_{c=0}^{diam(\mathcal{G})-1} \mathcal{L}_c^{(\theta,\theta^\star)}\}$ is non-empty (based on condition (i) of the theorem), (ii) any agent $i \in \mathcal{L}_{diam(\mathcal{G})}^{(\theta,\theta^\star)} \cap \mathcal{R}$ has at least $(2f+1)$ neighbors in the set $\bigcup_{c=0}^{diam(\mathcal{G})-1} \mathcal{L}_c^{(\theta,\theta^\star)}$, of which at least $f+1$ are regular (based on the f-locality of the adversarial model), and (iii) using the induction hypothesis and arguments similar to those used for arriving at (29). Finally, note that the manner in which the sets $\mathcal{L}_r^{(\theta,\theta^\star)}$ are constructed ensures that all the agents in $\mathcal{R}$ are covered. The rest of the proof is identical to that of Theorem 1. $\square$

VI. Conclusion

In this paper, we developed a novel distributed learning rule and established its consistency under minimal requirements on the information structures of the agents, and the communication graph. In particular, our analysis revealed that the commonly made assumption of strong-connectivity can be relaxed. We then showed that the proposed learning rule can be easily and efficiently modified to account for the presence of misbehaving agents in the network.

Ongoing work involves performing a detailed convergence rate analysis to see how such rates compare with those existing in literature. Extensions to time-varying graphs are also of interest.

References

[1] V. V. Veeravalli, T. Basar, and H. V. Poor, "Decentralized sequential detection with a fusion center performing the sequential test," *IEEE Transactions on Information Theory*, vol. 39, no. 2, pp. 433–442, 1993.

[2] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors Part I. Fundamentals," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, 1997.

[3] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games and Economic Behavior*, vol. 76, no. 1, pp. 210–225, 2012.

[4] A. Jadbabaie, P. Molavi, and A. Tahbaz-Salehi, "Information heterogeneity and the speed of learning in social networks," *Columbia Bus. Sch. Res. Paper*, pp. 13–28, 2013.

[5] Q. Liu, A. Fang, L. Wang, and X. Wang, "Social learning with time-varying weights," *Journal of Systems Science and Complexity*, vol. 27, no. 3, pp. 581–593, 2014.

[6] A. Nedić, A. Olshevsky, and C. A. Uribe, "Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs," in *Proc. of the American Control Conference*, 2015, pp. 5884–5889.

[7] ——, "Fast convergence rates for distributed Non-Bayesian learning," *IEEE Trans. on Autom. Control*, vol. 62, no. 11, pp. 5538–5553, 2017.

[8] A. Lalitha, T. Javidi, and A. Sarwate, "Social learning and distributed hypothesis testing," *IEEE Trans. on Information Theory*, 2018.

[9] L. Su and N. H. Vaidya, "Defending Non-Bayesian learning against adversarial attacks," *Distributed Computing*, pp. 1–13, 2016.

[10] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed detection: Finite-time analysis and impact of network topology," *IEEE Trans. on Autom. Control*, vol. 61, no. 11, pp. 3256–3268, 2016.

[11] K. R. Rad and A. Tahbaz-Salehi, "Distributed parameter estimation in networks," in *Proceedings of the 49th IEEE Decision and Control Conference*, 2010, pp. 5050–5055.

[12] S. Shahrampour and A. Jadbabaie, "Exponentially fast parameter estimation in networks using distributed dual averaging," in *Proc. of the 52nd Decision and Control Conference*, 2013, pp. 6196–6201.

[13] G. L. Gilardoni and M. K. Clayton, "On reaching a consensus using DeGroot's iterative pooling," *The Annals of Stat.*, pp. 391–401, 1993.

[14] D. Dolev, N. A. Lynch, S. S. Pinter, E. W. Stark, and W. E. Weihl, "Reaching approximate agreement in the presence of faults," *Journal of the ACM (JACM)*, vol. 33, no. 3, pp. 499–516, 1986.

[15] N. H. Vaidya, L. Tseng, and G. Liang, "Iterative approximate Byzantine consensus in arbitrary directed graphs," in *Proc. of the ACM Symp. on Principles of Distributed Computing*, 2012, pp. 365–374.

[16] S. Sundaram and B. Gharesifard, "Distributed optimization under adversarial nodes," *IEEE Trans. on Autom. Control*, 2018.

[17] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 4, pp. 766–781, 2013.

[18] L. Su and N. H. Vaidya, "Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms," in *Proc. of the 2016 ACM Symp. on Principles of Dist. Comp.* ACM, 2016, pp. 425–434.

[19] A. Mitra and S. Sundaram, "Resilient distributed state estimation for LTI systems," *arXiv preprint arXiv:1802.09651*, 2018.

[20] C.-Y. Koo, "Broadcast in radio networks tolerating Byzantine adversarial behavior," in *Proc. of the ACM Symposium on Principles of Distributed Computing.* ACM, 2004, pp. 275–282.

[21] A. Pelc and D. Peleg, "Broadcasting with locally bounded Byzantine faults," *Information Proc. Letters*, vol. 93, no. 3, pp. 109–115, 2005.

[22] H. Park and S. A. Hutchinson, "Fault-tolerant rendezvous of multirobot systems," *IEEE Trans. on Robotics*, vol. 33, no. 3, pp. 565–582, 2017.

[23] N. Vaidya, "Matrix representation of iterative approximate Byzantine consensus in directed graphs," *arXiv preprint arXiv:1203.1888*, 2012.

[24] W. Mulzer and D. Werner, "Approximating Tverberg points in linear time for any fixed dimension," *Discrete & Computational Geometry*, vol. 50, no. 2, pp. 520–535, 2013.

[25] A. Mitra and S. Sundaram, "Distributed observers for LTI systems," *IEEE Trans. on Autom. Control*, 2018.