

Finite-time Guarantees for Byzantine-Resilient Distributed State Estimation with Noisy Measurements

Lili Su and Shahin Shahrampour

Abstract—This work considers resilient, cooperative state estimation in unreliable multi-agent networks. A network of agents aims to collaboratively estimate the value of an unknown vector parameter, while an *unknown* subset of agents suffer Byzantine faults. Faulty agents malfunction arbitrarily and may send out *highly unstructured* messages to other agents in the network. As opposed to fault-free networks, reaching agreement in the presence of Byzantine faults is far from trivial. In this paper, we propose a computationally-efficient algorithm that is provably robust to Byzantine faults. At each iteration of the algorithm, a good agent (1) performs a gradient descent update based on noisy local measurements, (2) exchanges its update with other agents in its neighborhood, and (3) robustly aggregates the received messages using coordinate-wise trimmed means. Under mild technical assumptions, we establish that good agents learn the true parameter asymptotically in almost sure sense. We further complement our analysis by proving (high probability) *finite-time* convergence rate, encapsulating network characteristics.

I. INTRODUCTION

Collaborative state/parameter estimation has attracted a considerable attention due to a wide range of applications in internet of things (IoT), wireless networks, power grids, sensor networks, and robotic networks [1]–[7]. In these applications, a network of (connected) agents collect information in a distributed fashion and share an overarching goal to learn the common *unknown* truth $\theta^* \in \mathbb{R}^d$. Local measurements obtained by each individual agent contain noisy and highly incomplete information about θ^* . Nevertheless, the network of agents might be able to collaboratively learn θ^* by effectively fusing the information contained in their local measurements.

In the absence of system adversary, the state estimation problem is well-studied [5], [8]. However, some practical scenarios such as IoT, micro-grids, and Federated Learning are vulnerable to faults [9]. Motivated by that, we are interested in addressing collaborative estimation in the presence of malicious agents. The existence of malicious agents might arise when some of

the networked agents are compromised by a system adversary. Despite the wealth of literature on collaborative estimation with random link failures, packet-dropping failures, and crash failures (e.g. [10]), perhaps less well-known is estimation in the presence of highly *unstructured* failures or even *adversarial* agents, especially in *finite-time* domain.

In this work, to formally capture the *unstructured* system threat, we adopt Byzantine fault model [11] – a canonical fault model in distributed computing. In this model, there exists a system adversary that can choose up to a *constant* fraction of agents to compromise and control. An agent suffering Byzantine fault behaves arbitrarily badly by sending out unstructured malicious messages to the good agents. In addition, Byzantine agents may give conflicting messages to different agents in the system. Tolerating Byzantine faults is highly non-trivial (see e.g. [12], [13]). For example, it is well-known that in complete graphs, no algorithm can tolerate more than $1/3$ of the agents to be Byzantine [13]. This difficulty arises partially from the system asymmetry caused by the conflicting messages sent by the Byzantine agents. In fact, Byzantine consensus with vector multi-dimensional inputs in the complete graphs had not been solved until only recently [14], [15].

Despite intensive efforts on securing distributed learning (see Section I-B for details), to the best of the authors’ knowledge, efficient algorithms that are provably resilient to Byzantine faults with less stringent assumptions on *noisy* local measurements are still lacking. In particular, the literature has mostly focused on the asymptotic analysis, leaving the *finite-time* guarantees for such algorithms a complementary direction to pursue, which is the main focal point of this work.

A. Our Contributions

We propose a computationally-efficient algorithm that is provably robust to Byzantine faults. At each iteration of our algorithm, a good agent (1) performs a gradient descent update based on local measurements only, (2) exchanges its update with other agents in its neighborhood, and (3) robustly aggregates the received messages using coordinate-wise trimmed means.

For ease of exposition, we first present our results for fully connected networks (complete graphs), and then generalize the obtained results to general networks

Lili Su is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 (lilisu@mit.edu).

Shahin Shahrampour is with the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843 (shahin@tamu.edu).

(incomplete graphs) assuming that the networks satisfy the necessary conditions such that Byzantine-resilient consensus with scalar inputs can be achievable. For both cases, we establish that every good agent learns the true parameter asymptotically in the almost sure sense. Most importantly, we characterize the *finite-time* convergence rate (in high-probability sense), encapsulating network characteristics. We finally provide numerical simulations for our method to verify our theoretical results.

B. Related Literature

Resilient estimation, detection, and learning has attracted a great deal of attention in the past few years, and many researchers in the fields of control, signal processing, and network science have addressed the problem by adopting different notions of *resilience* or *robustness*.

In [16]–[18], resilience has been discussed in the context of smart power grid systems using cardinality minimization and its ℓ_1 relaxations. On the other hand, the focus of [19], [20] is on estimation in Linear Time-Invariant (LTI) systems. In [19], an interesting approach is proposed for fault detection using monitors, and fundamental monitoring limitations have been characterized using tools from system theory and game theory. Furthermore, the approach of [20] is inspired from the areas of error-correction over the reals and compressed sensing. In [21], robust Kalman filtering is discussed, where the estimate updates are derived using a convex ℓ_1 optimization problem. Authors of [22] consider a model where the observation noise is sparse, in the sense that the faulty sensors have noisy measurements, while other sensors measurements are noiseless. An event triggered projected gradient descent is then proposed to reconstruct the state. In our setting, though the state is fixed, we deal with *multi-agent* networks, i.e., the problem must be solved in a *distributed* manner since each agent has local (noisy) measurements from the state, and message passing schemes (e.g. consensus) are required to learn the state.

In parallel to advancements on resilient centralized estimation, recent years have witnessed intensive interest in securing distributed estimation. The authors of [23] discuss reaching consensus in the presence of malicious agents, assuming a *broadcast* model of communication. Chen et al. [24] propose a novel adversary detection strategy under which good agents either asymptotically learn the true state or detect the existence of a system fault. If a fault is flagged, the system goes through some external procedure to “repair” itself. As a result, the method does not perform estimation under system adversary (which is the focus of this paper). Furthermore, other resilient algorithms have been proposed [25]–[29] with different assumptions and performance guarantees.

Chen et al. [25] propose an algorithm under which *all* of the agents’ estimates converge to the true state as long as less than one half of the agents are faulty. However, this algorithm works under the assumption that an agent can *fully* observe the true state in the non-faulty condition [25, Section II.A], as opposed to our model which deals with both observability and noisy measurement issues. Mitra and Sundaram [26] consider the more general LTI systems and characterize the fundamental limits on adversary-resilient algorithms. However, unlike our work, [26] deals with noiseless observations and the focus is on asymptotic analysis. Xu et al. [27] study the general dynamic optimization problem. They propose a total variation (TV) norm regularization technique to mitigate the effect of malfunctioning agents, but unfortunately, in the static case, the good agents cannot learn the true minimizer (see Corollary 1 in [27]). In fact, [27, Assumption 4] might not hold in the sense that under some strategies of the adversary agents, some good agents may appear to be bad to others, and the outgoing links from those agents might be cut off by the good agents. The lack of convergence in this case is consistent with the lower bound result in [30].

Another relevant work is the distributed hypothesis testing of [28] where the algorithm *Byz-Iter* is proposed. Though this algorithm may work for the state estimation problem, it scales poorly in dimension. Our algorithm is similar to [29] in that we both combine local gradient descent with coordinate-wise message trimming. Although [29] considers a more general optimization framework, it is implicitly assumed that the optimization problem can be separated into independent optimization problems (of the size of unknown parameter); otherwise, [29, Lemma 1] does not hold and the proof in [30] cannot be applied.

II. PROBLEM FORMULATION

A. Network Model

We consider a multi-agent network which is a collection of n agents/nodes communicating with each other through a communication network $G(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, n\}$ and $\mathcal{E}$ denote the set of nodes and edges, respectively. We denote by $\mathcal{N}_i$ the set of incoming neighbors of agent i . An unknown subset of agents of size at most b , denoted by $\mathcal{A}$, might be *bad* or *adversarial*. The set $\mathcal{A}$ is chosen by the system adversary. For ease of exposition, let

$$|\mathcal{V}/\mathcal{A}| = \phi.$$

Clearly, $\phi \geq n - b$.

Good agents (agents in $\mathcal{V}/\mathcal{A}$) aim to estimate the unknown parameter collaboratively, but bad agents (agents in $\mathcal{A}$) can adversarially affect the estimation procedure by sending *completely arbitrary, malicious*, and possibly *conflicting* messages to the good agents.

B. Observation Model

In this work, we focus on a linear observation model, where $y_i(t)$ represents the local measurement of agent i at time t as follows

$$y_i(t) = H_i \theta^* + w_i(t), \quad (1)$$

and $H_i \in \mathbb{R}^{n_i \times d}$ is the local observation matrix. The noise sequence $w_i(t)$ is *i.i.d.* with $\mathbb{E}[w_i(t)] = \mathbf{0}$ and $\mathbb{E}[w_i(t)w_i(t)^\top] = \Sigma_i$. The sequences are bounded for all agents, i.e., there exists constant $C > 0$ such that $\mathbb{P}\{\|w_i(t)\|_2 \leq C\} = 1$ for $i \in \mathcal{V}$. Moreover, the noise sequences across good agents are independent. That is, $(w_i(t), t \geq 1)$ and $(w_j(t), t \geq 1)$ for $i \neq j$ are independent. As in practice the observation matrix H_i is often fat, i.e., $n_i \ll d$, each agent i must obtain information from others to correctly estimate θ^* .

C. Fault Model

To formally capture the system threat, we adopt the Byzantine fault model [11] – a canonical fault model in distributed computing. In this model, there exists a system adversary that can choose up to b of the n agents (where $b < n$) to compromise and control. Recall that this set of agents is denoted by $\mathcal{A}$. An agent suffering Byzantine fault is referred to as Byzantine agent. While the set $\mathcal{A}$ is unknown to good agents, a standard assumption in the literature is that the value of b is common knowledge [11].

The system adversary is extremely *powerful* in the sense that it has complete knowledge of the network, including the local program that each good agent is supposed to run, the true value of the parameter θ^* , the current status and running history of the multi-agent network system, the running history, etc. Hence, the Byzantine agents can collude with each other and deviate from their pre-specified local programs to *arbitrarily* misrepresent information to the good agents. In particular, Byzantine agents can mislead each of the good agents in a unique fashion, i.e., letting $m_{ij}(t) \in \mathbb{R}^d$ be the message sent from agent $i \in \mathcal{A}$ to agent $j \in \mathcal{V} \setminus \mathcal{A}$ at time t , it is possible that $m_{ij}(t) \neq m_{ij'}(t)$ for $j \neq j' \in \mathcal{V} \setminus \mathcal{A}$.

Remark 1. *Note that due to the extreme freedom given to Byzantine agents and the system asymmetry caused by them, a resilient distributed solution to the estimation problem is highly non-trivial even in complete graphs. In particular, it is well-known that in complete graphs, no algorithm can tolerate more than 1/3 of the agents to be Byzantine [13].*

D. Finite-time vs. Asymptotic Local Functions

The Byzantine-resilient state estimation problem can be viewed with an optimization lens, where each good

agent would only asymptotically know its local function. For each agent $i \in \mathcal{V}$, define the *asymptotic* local function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ as

$$f_i(x) \triangleq \mathbb{E} \left[\frac{1}{2} \|H_i x - y_i\|_2^2 \right], \quad (2)$$

where the expectation is taken over the randomness of w_i . Note that f_i is well-defined for each $i \in \mathcal{V}$ regardless of whether it is suffering Byzantine faults or not. Since the distribution of w_i is unknown to agent i , at any finite t , function f_i is not accessible to agent i . However, the agent has access to the *finite-time* or *empirical* local function

$$f_{i,t}(x) \triangleq \frac{1}{t} \sum_{s=1}^t \frac{1}{2} \|H_i x - y_i(s)\|_2^2, \quad (3)$$

whose gradient at x is

$$\nabla f_{i,t}(x) = H_i^\top H_i (x - \theta^*) - H_i^\top \frac{1}{t} \sum_{r=1}^t w_i(r). \quad (4)$$

III. BYZANTINE-RESILIENT STATE ESTIMATION

To robustify distributed state estimation against Byzantine faults, one approach may be to combine the local gradient descent with multi-dimensional Byzantine-resilient consensus [14], [15], [28] (which typically relies on using Tverberg points). However, the performance of any such algorithm is proved to scale poorly in the dimension of the parameter d [14], [15], [28]. This is partially due to the fact that different dimensions of the inputs strongly interfere with each other, and the Byzantine agents can inject wrong information with both extreme magnitudes and directions.

To improve the scalability with respect to d and to improve the computation complexity, instead of multi-dimensional Byzantine-resilient consensus, we robustly aggregate the received messages using coordinate-wise trimmed means.

A. Algorithm

We propose an algorithm, named *Byzantine-resilient state estimation*, where each good agent iteratively aggregates the received messages. To robustify, the agent discards the largest b and the smallest b values for each component. In particular, in each iteration, an agent performs the following three steps:

- *Local gradient descent:* Agent i first computes the noisy local gradient $\nabla f_{i,t}(x_i(t-1))$, and performs local gradient descent to obtain $z_i(t)$, i.e.,

$$z_i(t) = x_i(t-1) - \nabla f_{i,t}(x_i(t-1)).$$

Note that the step-size used in this update is 1.

- *Information exchange:* It exchanges $z_i(t)$ with other agents in its local neighborhood. Recall that $m_{ij}(t) \in \mathbb{R}^d$ is the message sent from agent i to agent j at time t . It relates to $z_i(t)$ as follows:

$$m_{ij}(t) = \begin{cases} z_i(t) & \text{if } i \in (\mathcal{V}/\mathcal{A}); \\ \star & \text{if } i \in \mathcal{A}, \end{cases}$$

where $\star$ denotes an arbitrary value. Byzantine agents can mislead good agents differently, i.e., if $i \in \mathcal{A}$, it might hold that $m_{ij}(t) \neq m_{ij'}(t)$ for $j \neq j' \in \mathcal{V} \setminus \mathcal{A}$.

- *Robust aggregation:* For each component $k = 1, \dots, d$, the agent computes the trimmed mean and uses them to obtain $x_i(t)$.

The formal description of the algorithm for agent $i \in \mathcal{V} \setminus \mathcal{A}$ is given in Algorithm 1.

Algorithm 1: Byzantine-resilient state estimation

Input: b and T

Initialization: Set $x_i(0)$ to an arbitrary value for each agent $i \in \mathcal{V}$

for $t = 1, \dots, T$ **do**

- Obtain a new measurement $y_i(t)$;
- Compute the local noisy gradient $\nabla f_{i,t}(x_i(t-1))$ according to (4);
- Compute $z_i(t) = x_i(t-1) - \nabla f_{i,t}(x_i(t-1))$;
- Send $z_i(t)$ to its outgoing neighbors;

for $k = 1, \dots, d$ **do**

- Sort the k -th component of the received messages $m_{ji}(t)$ for $j \in \mathcal{N}_i \cup \{i\}$ in a non-decreasing (increasing) order;
- Remove the largest b values and the smallest b values;
- Denote the remained “agent” indices set as $\mathcal{R}_i^k(t)$ and set

$$x_i^k(t) = \frac{1}{|\mathcal{R}_i^k(t)|} \sum_{j \in \mathcal{R}_i^k(t)} \langle m_{ji}(t), e_k \rangle.$$

end

- Set $(x_i(t))^\top = (x_i^1(t), \dots, x_i^d(t))$.

end

Output: $x_i(T)$.

IV. FINITE-TIME GUARANTEE FOR COMPLETE NETWORKS

In this section, we provide results for the case that $G(\mathcal{V}, \mathcal{E})$ is a complete graph. Beside the fact that the technical analysis of complete graphs would be different from that of incomplete graphs (in terms of assumptions), the former is particularly interesting in computer networks. In fact, in many computer networks

efficient communication protocols (such as TCP/IP) can be implemented such that any two computer are logically connected.

It can be shown that the update of x_i uses the information provided by the *good* agents only. In addition, each of the good agent has limited impact on x_i , formally stated next.

Lemma 1. *For each iteration t , each good agent $i \in \mathcal{V}/\mathcal{A}$, and each k , there exist coefficients $(\beta_{ij}^k(t), j \in \mathcal{V}/\mathcal{A})$ such that*

- $x_i^k(t) = \sum_{j \in \mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) \langle z_j(t), e_k \rangle$;
- $0 \leq \beta_{ij}^k(t) \leq \frac{1}{\phi - b}$ for all $j \in \mathcal{V}/\mathcal{A}$ and $\sum_{j \in \mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) = 1$.

Notice that the sets of convex coefficients for different values of k might be different, i.e., $(\beta_{ij}^k(t), j \in \mathcal{V}/\mathcal{A}) \neq (\beta_{ij}^{k'}(t), j \in \mathcal{V}/\mathcal{A})$ for $k \neq k'$. Moreover, even for the same k , the convex coefficients might be different for different good agents, i.e., $(\beta_{ij}^k(t), j \in \mathcal{V}/\mathcal{A}) \neq (\beta_{i'j}^k(t), j \in \mathcal{V}/\mathcal{A})$ for $i \neq i'$. This stems from the freedom of Byzantine agents in sending different messages across agents, i.e., $m_{aj} \neq m_{aj'}$ if $a \in \mathcal{A}$ and $j \neq j'$.

To prove the convergence of Algorithm 1, we use the following assumption.

Assumption 1. *For all $k = 1, \dots, d$, we have that*

$$\frac{1}{\phi - b} \sum_{j \in \mathcal{V}/\mathcal{A}} \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 < 1.$$

Note that $\|(\mathbf{I} - H_j^\top H_j) e_k\|_1$ is the ℓ_1 norm of the k -th column of matrix $\mathbf{I} - H_j^\top H_j$. It can well be the case that $\|(\mathbf{I} - H_j^\top H_j) e_k\|_1 \geq 1$ for some good agents. However, Assumption 1 implies that for each $k = 1, \dots, d$, there exists at least $b + 1$ good agents such that

$$\|(\mathbf{I} - H_j^\top H_j) e_k\|_1 < 1.$$

The above assumption is imposed for the d components individually. None of the agents are required to satisfy $\|(\mathbf{I} - H_j^\top H_j) e_k\|_1 < 1$ simultaneously for all $k = 1, \dots, d$. Now, let

$$\rho \triangleq \max_{k: 1 \leq k \leq d} \frac{\sum_{j \in \mathcal{V}/\mathcal{A}} \|(\mathbf{I} - H_j^\top H_j) e_k\|_1}{\phi - b}. \quad (5)$$

Clearly, $\rho < 1$ under Assumption 1. For ease of exposition, for each $j \in \mathcal{V}/\mathcal{A}$ and for any $\lambda \in (0, 1)$, let

$$R_j(\lambda, t) \triangleq \sum_{m=0}^{t-1} \lambda^m \left\| \frac{\sum_{r=1}^{t-m} w_j(r)}{t-m} \right\|_2. \quad (6)$$

The following two concentration results are two key auxiliary lemmas for our main theorem.

Lemma 2. Suppose Assumption 1 holds. Then, for each $j \in \mathcal{V}/\mathcal{A}$ and for any $\lambda \in (0, 1)$

$$\lim_{t \rightarrow \infty} R_j(\lambda, t) = 0 \text{ almost surely.}$$

In addition, we characterize the *finite-time* convergence rate of $R_j(\lambda, \cdot)$ for any fixed λ .

Lemma 3. Suppose Assumption 1 holds. Then for each $j \in \mathcal{V}/\mathcal{A}$ and for any $\lambda \in (0, 1)$

$$\begin{aligned} & \mathbb{P} \left\{ R_j(\lambda, t) \geq \sqrt{\text{trace}(\Sigma_j)} \sum_{m=1}^{t-1} \lambda^m \frac{1}{\sqrt{t-m}} + \epsilon \right\} \\ & \leq \exp \left(\frac{-\epsilon^2(1-\lambda)^2 t}{8C^2} \right), \end{aligned}$$

Lemma 3 implies that $\forall j \in \mathcal{V}/\mathcal{A}$, with probability at least $1 - \delta$, $R_j(t) = O \left(\sqrt{(\log \frac{1}{\delta})/t} \right)$.

Theorem 1. Suppose Assumption 1 holds and the graph $G(\mathcal{V}, \mathcal{E})$ is complete. Then

$$\max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(t) - \theta^*\|_\infty \xrightarrow{\text{a.s.}} 0.$$

Moreover, with probability at least $1 - \phi \exp \left(\frac{-\epsilon^2(1-\rho)^2 t}{8C^2} \right)$, it holds that

$$\begin{aligned} & \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(t) - \theta^*\|_\infty \leq \rho^t \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(0) - \theta^*\|_\infty \\ & + C_0 \left(\sum_{i \in \mathcal{V}/\mathcal{A}} \sqrt{\text{trace}(\Sigma_j)} \right) \sum_{m=1}^{t-1} \frac{\rho^m}{\sqrt{t-m}} + \phi\epsilon, \end{aligned}$$

where $C_0 \triangleq \max_{i \in \mathcal{V}/\mathcal{A}} \|H_i\|_2$.

The theorem indicates that all good agents (in a complete graph) are eventually able to learn the true parameter θ^* almost surely. Also, with high probability the rate can be characterized as above, providing a *finite-time* guarantee for resilient estimation. The finite-time bound captures the performance, in terms of Σ_j , the noise covariance for agent $j \in \mathcal{V}/\mathcal{A}$, as well as ρ , which can crudely serve as a measure of observability in view of (5).

V. FINITE-TIME GUARANTEES FOR INCOMPLETE NETWORKS

A. Incomplete Graphs: Multihop Communication

So far, our analysis of Algorithm 1 has focused on complete graphs.

For computer networks, this is a reasonable assumption as computers are connected to each other through some communication (routing) protocols. Our results are also applicable to wireless networks under some implementation assumptions.

Concretely, let $G(\mathcal{V}, \mathcal{E})$ be the physical network that is not fully connected. Suppose the networked agents

are allowed to relay the messages sent by others such that multi-hop communication can be implemented. We can adopt coding to force the Byzantine agents to either refuse to relay information or faithfully relay the messages without alteration [12]. Thus, as long as the node-connectivity of $G(\mathcal{V}, \mathcal{E})$ is at least $b+1$, each good agent can reliably receive messages from other good agents in the network. We can use our algorithm to robustly aggregate the received messages and perform one-step update. Similar analysis applies.

B. Incomplete Graphs: Local Communication

Message forwarding might be costly or even infeasible for some wireless networks. Algorithms that rely solely on local communication are still highly desirable. Fortunately, with reasonable assumptions, Algorithm 1 works. Our algorithm is a consensus-based algorithm, so to make the paper self-contained, we briefly review relevant existing results on Byzantine consensus.

1) *Byzantine Consensus with Scalar Inputs:* Note that, in contrast to fault-free consensus, Byzantine-resilient consensus with scalar inputs and with multidimensional inputs are fundamentally different [14], [15], [31]. Our algorithm relies on Byzantine-resilient consensus with scalar inputs.

Tight topological conditions are characterized in [31], where the conditions are stated in terms of a family of subgraphs of $G(\mathcal{V}, \mathcal{E})$. Those subgraphs capture the “real” information flow under the message trimming strategy. Informally speaking, trimming certain messages can be viewed as ignoring (or removing) incoming links that carry the outliers. The non-uniqueness of the subgraph arises partially from the fact that the Byzantine agents can behave adaptively and arbitrarily. Such subgraphs are referred to as *reduced graphs*, defined as follows.

Definition 1. [31] A *reduced graph* H of $G(\mathcal{V}, \mathcal{E})$ is obtained by (i) removing all faulty nodes $\mathcal{A}$, and all the links incident on the faulty nodes $\mathcal{A}$; and (ii) for each non-faulty node (nodes in $\mathcal{V}/\mathcal{A}$), removing up to b additional incoming links.

It is important to note that the non-faulty agents *do not* know the identities of the faulty agents. Let $\mathcal{H}$ be the collection of all reduced graphs of $G(\mathcal{V}, \mathcal{E})$, and let

$$\xi := |\mathcal{H}|.$$

Definition 2. A *source component* in a given reduced graph is a strongly connected component, which does not have any incoming links from outside of that component.

It turns out that the effective communication network is potentially time-varying (partly) due to time-varying behaviors of Byzantine agents. The tight network

topology condition for scalar valued consensus to be achievable is characterized in [31].

Theorem 2. [31] *For scalar inputs, iterative approximate Byzantine consensus is achievable among non-faulty agents if and only if every reduced graph of $G(\mathcal{V}, \mathcal{E})$ contains only one source component.*

Under the condition in Theorem 2, it follows that in any reduced graph, a node in the source component can reach every other nodes.

2) *Correctness of Algorithm 1 for Incomplete Graphs:* We will show the correctness of our Algorithm 1 assuming that Byzantine consensus with scalar inputs is achievable over $G(\mathcal{V}, \mathcal{E})$, and the following assumption holds.

Assumption 2. *For each non-faulty node $j \in \mathcal{V}/\mathcal{A}$ and each $k = 1, \dots, d$,*

$$\|(\mathbf{I} - H_j^\top H_j) e_k\|_1 \leq 1.$$

In addition, any reduced graph H contains a node in its unique source component such that for all $k = 1, \dots, d$,

$$\left| (\mathcal{N}_i \cup \{i\}/\mathcal{A}) \cap \left\{ j : \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 < 1 \right\} \right| \geq b + 1.$$

Note that in Assumption 2, $\mathcal{N}_i$ is the incoming neighbors of node i in the original graph $G(\mathcal{V}, \mathcal{E})$.

Define ρ_0 as

$$\rho_0 := \max_{1 \leq k \leq d} \max_{j: \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 < 1} \|(\mathbf{I} - H_j^\top H_j) e_k\|_1. \quad (7)$$

In (7), the maximization

$$\max_{j: \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 < 1} \|(\mathbf{I} - H_j^\top H_j) e_k\|_1$$

is taken over the non-faulty nodes only.

Similar to the analysis for the complete graphs, it can be shown that the update of x_i uses the information provided by its good neighbors only.

Lemma 4. [32, Claim 2] *For each iteration t , each good agent $i \in \mathcal{V}/\mathcal{A}$, and each k , there exist coefficients $(\beta_{ij}^k(t), j \in \mathcal{N}_i \cup \{i\})$ such that*

- $x_i^k(t) = \sum_{j \in \mathcal{N}_i \cup \{i\}/\mathcal{A}} \beta_{ij}^k(t) \langle z_j(t), e_k \rangle$;
- *There exists a subset of $\mathcal{B}_i(t) \subseteq \mathcal{N}_i \cup \{i\}/\mathcal{A}$ such that $|\mathcal{B}_i(t)| \geq |\mathcal{N}_i \cup \{i\}/\mathcal{A}| - b$ and $\beta_{ij}^k(t) \geq \frac{1}{2(|\mathcal{N}_i \cup \{i\}/\mathcal{A}| - b)}$ for each $j \in \mathcal{B}_i(t)$.*

In the next theorem, we establish that (under the assumption above) the estimates of all agents are consistent almost surely, and furthermore, we characterize the (high probability) *finite-time* convergence rate of these estimates.

Theorem 3. *Suppose that every reduced graph of $G(\mathcal{V}, \mathcal{E})$ contains a single source component, and Assumption 2 holds. Then*

$$\max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(t) - \theta^*\|_\infty \xrightarrow{\text{a.s.}} 0.$$

Let $\gamma := 1 - \frac{1-\rho_0}{(2(\phi-b))^{\xi\phi}}$. With probability at least $1 - \phi \exp\left(\frac{-\epsilon^2(1-\gamma^{\frac{1}{\xi\phi}})^2 t}{8C^2}\right)$, it holds that

$$\begin{aligned} \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(t) - \theta^*\|_\infty &\leq \gamma^{\frac{t}{\xi\phi}} \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(0) - \theta^*\|_\infty \\ &+ C_0 \left(\sum_{i \in \mathcal{V}/\mathcal{A}} \sqrt{\text{trace}(\Sigma_j)} \right)^{t-1} \sum_{m=1}^{t-1} \frac{\gamma^{\frac{m}{\xi\phi}}}{\sqrt{t-m}} + \phi\epsilon. \end{aligned}$$

VI. NUMERICAL EXPERIMENTS

We now provide empirical evidence in support of our algorithm. We consider a complete graph of $|\mathcal{V} \setminus \mathcal{A}| = 30$ agents. Each component of the unknown parameter $\theta^* \in \mathbb{R}^{50}$ is generated randomly within the interval $[-1, 1]$ and is fixed thereafter during the estimation process. Moreover, the observation matrices $H_i \in \mathbb{R}^{20 \times 50}$ for each i are chosen such that Assumption 1 holds.

Evidently, in this example, $n_i = 20$ for all i . Throughout, the adversarial agents can send out completely arbitrary messages in lieu of true gradients. We generate these arbitrary messages using a random 50-dimensional vector, each component of which is sampled from $\mathcal{N}(0, 9)$.

Let us now define the network performance metric as

$$\text{Error}(t) \triangleq \frac{1}{\phi} \sum_{i \in \mathcal{V} \setminus \mathcal{A}} \|\theta^* - x_i(t)\|$$

and plot in Fig. 1 the error for various values of adversarial agents $|\mathcal{A}| \in \{4, 5, \dots, 10\}$. We observe a dichotomy, where for $|\mathcal{A}| < 7$ the error converges to zero, whereas for $|\mathcal{A}| \geq 7$ the convergence does not occur. Moreover, increase in the number of adversarial agents degrades the performance.

VII. CONCLUSION

We studied resilient distributed estimation, where a network of agents want to learn the value of an unknown parameter in the presence of Byzantine agents. The main challenges in the problem are as follows: (i) Byzantine agents send out arbitrary messages to other agents, (ii) good agents need to deal with noisy measurements, and (iii) the parameter is not locally observable. We proposed an algorithm that allows agents to collectively learn the true parameter asymptotically in almost sure sense, and we further complemented our results with *finite-time* analysis. Future directions include resilient estimation and learning in a more general setting, where

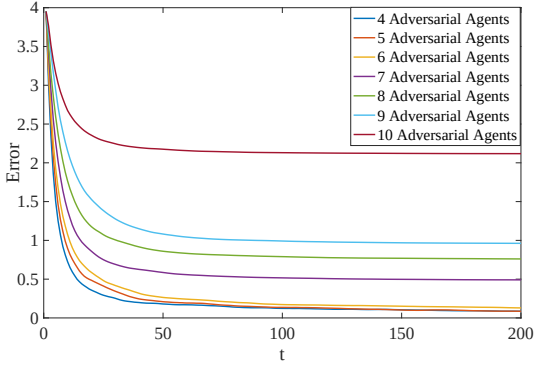


Fig. 1: The plot of error decay versus time for different number of adversarial agents.

agents observations can be a nonlinear function of the unknown parameter. Another interesting direction is to investigate the minimal condition needed on the local observation matrices of the good agents for the problem to be solvable.

APPENDIX

PROOF OF LEMMA 1

We prove this lemma by construction. Note that this construction is only used in the algorithm analysis rather than an algorithm input. That is, to run the algorithm, each agent (either good or faulty) does not need to know β .

For ease of exposition, let $[\mathcal{R}_i^k(t)]^+$ and $[\mathcal{R}_i^k(t)]^-$ be the non-overlapping subsets of $\mathcal{V}$ whose gradient's k -th entry are trimmed away by agent i . Precisely,

- (a) $|[\mathcal{R}_i^k(t)]^-| = b = |[\mathcal{R}_i^k(t)]^+|$;
- (b) $[\mathcal{R}_i^k(t)]^-$, $[\mathcal{R}_i^k(t)]^+$ and $\mathcal{R}_i^k(t)$ partition set $\mathcal{V}$;
- (c) $\forall j' \in [\mathcal{R}_i^k(t)]^-$, $j \in \mathcal{R}_i^k(t)$, and $j'' \in [\mathcal{R}_i^k(t)]^+$ it holds that

$$\langle m_{j'i}(t), e_k \rangle \leq \langle m_{ji}(t), e_k \rangle \leq \langle m_{j''i}(t), e_k \rangle. \quad (8)$$

We consider two cases: (1) $\mathcal{R}_i^k(t) \cap \mathcal{A} = \emptyset$; and (2) $\mathcal{R}_i^k(t) \cap \mathcal{A} \neq \emptyset$.

Case 1: Suppose that $\mathcal{R}_i^k(t) \cap \mathcal{A} = \emptyset$. We construct the convex coefficients as follows:

Case 1-1: When $|\mathcal{A}| = b$, we have $\phi - b = n - 2b$. We choose the convex coefficients as

$$\beta_{ij}^k(t) = \begin{cases} \frac{1}{n-2b}, & \forall j \in \mathcal{R}_i^k(t), \text{ and} \\ 0, & \forall j \notin \mathcal{R}_i^k(t). \end{cases}$$

Clearly, in this construction, $\beta_{ij}^k(t) \leq \frac{1}{\phi-b}$.

Case 1-2: When $|\mathcal{A}| < b$, it holds that

$$|[\mathcal{R}_i^k(t)]^-/\mathcal{A}| \geq b - |\mathcal{A}|, \quad (9)$$

and

$$|[\mathcal{R}_i^k(t)]^+/\mathcal{A}| \geq b - |\mathcal{A}|. \quad (10)$$

By (8), we have

$$\begin{aligned} & \frac{1}{|[\mathcal{R}_i^k(t)]^-/\mathcal{A}|} \sum_{j \in [\mathcal{R}_i^k(t)]^-/\mathcal{A}} \langle z_j(t), e_k \rangle \\ & \leq \frac{1}{n-2b} \sum_{j \in \mathcal{R}_i^k(t)} \langle z_j(t), e_k \rangle \\ & \leq \frac{1}{|[\mathcal{R}_i^k(t)]^+/\mathcal{A}|} \sum_{j \in [\mathcal{R}_i^k(t)]^+/\mathcal{A}} \langle z_j(t), e_k \rangle. \end{aligned}$$

Thus, there exists $\alpha \in [0, 1]$ such that

$$\begin{aligned} & \frac{1}{n-2b} \sum_{j \in \mathcal{R}_i^k(t)} \langle z_j(t), e_k \rangle \\ & = \frac{\alpha}{|[\mathcal{R}_i^k(t)]^-/\mathcal{A}|} \sum_{j \in [\mathcal{R}_i^k(t)]^-/\mathcal{A}} \langle z_j(t), e_k \rangle \\ & \quad + \frac{1-\alpha}{|[\mathcal{R}_i^k(t)]^+/\mathcal{A}|} \sum_{j \in [\mathcal{R}_i^k(t)]^+/\mathcal{A}} \langle z_j(t), e_k \rangle. \quad (11) \end{aligned}$$

Note that

$$\begin{aligned} & \frac{1}{n-2b} \sum_{j \in \mathcal{R}_i^k(t)} \langle z_j(t), e_k \rangle \\ & = \frac{1}{\phi-b} \left(1 + \frac{b-|\mathcal{A}|}{n-2b} \right) \sum_{j \in \mathcal{R}_i^k(t)} \langle z_j(t), e_k \rangle \\ & = \frac{1}{\phi-b} \sum_{j \in \mathcal{R}_i^k(t)} \langle z_j(t), e_k \rangle \\ & \quad + \frac{1}{\phi-b} \frac{b-|\mathcal{A}|}{n-2b} \sum_{j \in \mathcal{R}_i^k(t)} \langle z_j(t), e_k \rangle \\ & \stackrel{(a)}{=} \frac{1}{\phi-b} \sum_{j \in \mathcal{R}_i^k(t)} \langle z_j(t), e_k \rangle \\ & \quad + \frac{\alpha(b-|\mathcal{A}|)}{(\phi-b)|[\mathcal{R}_i^k(t)]^-/\mathcal{A}|} \sum_{j \in [\mathcal{R}_i^k(t)]^-/\mathcal{A}} \langle z_j(t), e_k \rangle \\ & \quad + \frac{(1-\alpha)(b-|\mathcal{A}|)}{(\phi-b)|[\mathcal{R}_i^k(t)]^+/\mathcal{A}|} \sum_{j \in [\mathcal{R}_i^k(t)]^+/\mathcal{A}} \langle z_j(t), e_k \rangle. \end{aligned}$$

where equality (a) follows from (11). Choose the convex coefficients for the good agents as follows:

$$\beta_{ij}^k(t) = \begin{cases} \frac{1}{\phi-b}, & \forall j \in \mathcal{R}_i^k(t), \\ \frac{\alpha(b-|\mathcal{A}|)}{(\phi-b)|[\mathcal{R}_i^k(t)]^-/\mathcal{A}|}, & \forall j \in [\mathcal{R}_i^k(t)]^-/\mathcal{A}, \\ \frac{(1-\alpha)(b-|\mathcal{A}|)}{(\phi-b)|[\mathcal{R}_i^k(t)]^+/\mathcal{A}|}, & \forall j \in [\mathcal{R}_i^k(t)]^+/\mathcal{A}. \end{cases}$$

The fact that α is unknown does not affect our correctness proof – as our algorithm not use these coefficients as input. We use the existence of α for analysis. It is easy to see that the above coefficients are valid convex

coefficients. It remains to check that $\beta_{ij}^k(t) \leq \frac{1}{\phi-b}$ for all $j \in \mathcal{V}/\mathcal{A}$. For all good in $\mathcal{R}_i^k(t)$, clearly $\beta_{ij}^k(t) \leq \frac{1}{\phi-b}$. For $j \in [\mathcal{R}_i^k(t)]^-/\mathcal{A}$, by (10) and the fact that $\alpha \leq 1$, we have

$$\beta_{ij}^k(t) \leq \frac{\alpha(b-|\mathcal{A}|)}{(\phi-b)(b-|\mathcal{A}|)} \leq \frac{1}{\phi-b},$$

Similarly, we can show $\beta_{ij}^k(t) \leq \frac{1}{\phi-b}$ for $j \in [\mathcal{R}_i^k(t)]^+/\mathcal{A}$.

Case 2 can be proved similarly.

PROOF OF LEMMA 2

Let ω be any sample path such that $\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{r=1}^t w_j(r, \omega) = 0$. Note that fixing ω , $w_j(t, \omega)$ for $t = 1, \dots$ is a standard sequence of vectors. We will show that

$$\lim_{t \rightarrow \infty} \sum_{m=0}^{t-1} \lambda^m \left\| \frac{\sum_{r=1}^{t-m} w_j(r, \omega)}{t-m} \right\|_2 = 0. \quad (12)$$

By Strong Law of Large Number we know that

$$\mathbb{P} \left\{ \omega \in \Omega : \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{r=1}^t w_j(r, \omega) = 0 \right\} = 1.$$

Thus, if (12) holds, then

$$\mathbb{P} \left\{ \omega \in \Omega : \lim_{t \rightarrow \infty} \sum_{m=0}^{t-1} \lambda^m \left\| \frac{\sum_{r=1}^{t-m} w_j(r, \omega)}{t-m} \right\|_2 = 0 \right\} = 1, \text{ So we get}$$

proving the lemma.

Next we show (12). It is enough to show that for any $\epsilon > 0$, there exists $t \geq t(\epsilon, \omega)$ such that

$$\sum_{m=0}^{t-1} \lambda^m \left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_2 \leq \epsilon. \quad (13)$$

Since $\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{r=1}^t w_j(r, \omega) = 0$, for any $\frac{(1-\lambda)\epsilon}{2}$, there exists $t_0(\epsilon, \omega)$ such that for any $t \geq t_0(\epsilon, \omega)$,

$$\left\| \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2 \leq \frac{(1-\lambda)\epsilon}{2}.$$

In addition, for any $t \geq t_0(\epsilon, \omega)$, it holds that

$$\begin{aligned} & \sum_{m=0}^{t-1} \lambda^m \left\| \frac{\sum_{r=1}^{t-m} w_j(r)}{t-m} \right\|_2 \\ & \leq \sum_{m=0}^{t-t_0(\epsilon, \omega)} \lambda^m \frac{(1-\lambda)\epsilon}{2} + C \sum_{m=t-t_0(\epsilon, \omega)+1}^{t-1} \lambda^m \\ & \leq \frac{\epsilon}{2} + C \frac{\lambda^{t-t_0(\epsilon, \omega)+1}}{1-\lambda}. \end{aligned}$$

There exists a sufficiently large $t(\epsilon, \omega)$ such that $C \frac{\lambda^{t-t_0(\epsilon, \omega)+1}}{1-\lambda} \leq \frac{\epsilon}{2}$. Thus, it holds that for this fixed

sample path ω , for any $\epsilon > 0$, there exists $t(\epsilon, \omega)$ such that for all $t \geq t(\epsilon, \omega)$

$$\sum_{m=0}^{t-1} \lambda^m \left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_1 \leq \epsilon,$$

proving (13).

PROOF OF LEMMA 3

Our proof uses the McDiarmid's inequality.

We first bound the expectation of $R_j(\lambda, t)$.

$$\begin{aligned} \mathbb{E}[R_j(\lambda, t)] &= \sum_{m=0}^{t-1} \lambda^m \mathbb{E} \left[\left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_2 \right] \\ &\stackrel{(a)}{\leq} \sum_{m=0}^{t-1} \lambda^m \sqrt{\mathbb{E} \left[\left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_2^2 \right]} \end{aligned}$$

where equality (a) follows from Jensen's inequality. Recall that $w_j(r)$ for $r = 1, \dots, t-m$ are independent and $\mathbb{E}[w_j(r)] = \mathbf{0}$ for each $r = 1, \dots, t-m$. Thus, for any $j \in \mathcal{V}/\mathcal{A}$, we have

$$\mathbb{E} \left[\left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_2^2 \right] = \frac{1}{t-m} \text{trace}(\Sigma_j).$$

$$\mathbb{E}[R_j(\lambda, t)] \leq \sqrt{\text{trace}(\Sigma_j)} \sum_{m=1}^{t-1} \lambda^m \frac{1}{\sqrt{t-m}}.$$

We choose h as

$$h(\{w_j(r)\}_{r=1}^t) \triangleq \sum_{m=0}^{t-1} \lambda^m \left\| \frac{\sum_{r=1}^{t-m} w_j(r)}{t-m} \right\|_2.$$

It can be shown that we can choose c_r to be

$$c_r = C \sum_{m=0}^{t-r} \lambda^m \frac{1}{t-m}, \quad \forall r = 1, \dots, t.$$

Let $m_0 = \frac{\log \frac{\lambda t}{\lambda}}{\log \frac{\lambda}{\lambda}}$. It is easy to see that $m_0 \leq \frac{t}{2}$ unless t is extremely small. For simplicity, assume that $\frac{\log \frac{\lambda t}{\lambda}}{\log \frac{\lambda}{\lambda}}$ is an integer. So we have

$$c_1 = C \left(\sum_{m=0}^{m_0} \lambda^m \frac{1}{t-m} + \sum_{m=m_0+1}^{t-1} \lambda^m \frac{1}{t-m} \right) \leq \frac{4C}{(1-\lambda)t}.$$

It is easy to see that $c_r \leq c_1$ for all $r = 1, \dots, t$. So we have

$$\sum_{r=1}^t c_r^2 \leq t c_1^2 \leq \left(\frac{4C}{1-\lambda} \right)^2 \frac{1}{t}.$$

By McDiarmid's Inequality we have

$$\mathbb{P} \left\{ R_j(\lambda, t) \geq \sqrt{\text{trace}(\Sigma_j)} \sum_{m=1}^{t-1} \lambda^m \frac{1}{\sqrt{t-m}} + \epsilon \right\} \\ \leq \exp \left(\frac{-2\epsilon^2}{\sum_{r=1}^t c_r^2} \right) \leq \exp \left(\frac{-\epsilon^2(1-\lambda)^2 t}{8C^2} \right).$$

PROOF OF THEOREM 1

For each t , $x_i(t)$ can be uniquely rewritten as

$$x_i(t) = \theta^* + \sum_{k=1}^d \alpha_i^k(t) e_k,$$

where $\alpha_i^k(t), k = 1, \dots, d$ is a linear coefficients. At time t , for each $k = 1, \dots, d$, it holds that

$$\alpha_i^k(t) = \frac{1}{|\mathcal{R}_i^k(t)|} \sum_{j \in \mathcal{R}_i^k(t)} \langle m_{ji}(t), e_k \rangle - \langle \theta^*, e_k \rangle.$$

It follows from Lemma 1 that

$$\alpha_i^k(t) = \sum_{j \in \mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) \langle z_j(t), e_k \rangle - \langle \theta^*, e_k \rangle. \quad (14)$$

Recall from (3) and (4), for each $k = 1, \dots, d$, we have

$$\langle z_i(t), e_k \rangle = \langle \theta^*, e_k \rangle + \left\langle H_i^\top \frac{1}{t} \sum_{r=1}^t w_i(r), e_k \right\rangle \\ + \left\langle \sum_{k'=1}^d \alpha_i^{k'}(t-1) (\mathbf{I} - H_i^\top H_i) e_{k'}, e_k \right\rangle.$$

Thus, (14) becomes

$$\alpha_i^k(t) = \sum_{j \in \mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) \left\langle H_j^\top \frac{1}{t} \sum_{r=1}^t w_j(r), e_k \right\rangle \\ + \sum_{j \in \mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) \left\langle \sum_{k'=1}^d \alpha_j^{k'}(t-1) (\mathbf{I} - H_j^\top H_j) e_{k'}, e_k \right\rangle.$$

By Lemma 1, we have

$$|\alpha_i^k(t)| \leq \frac{\sum_{j \in \mathcal{V}/\mathcal{A}} \left| \left\langle H_j^\top \frac{1}{t} \sum_{r=1}^t w_j(r), e_k \right\rangle \right|}{\phi - b} \\ + \frac{\sum_{j \in \mathcal{V}/\mathcal{A}} \left| \left\langle \sum_{k'=1}^d \alpha_j^{k'}(t-1) (\mathbf{I} - H_j^\top H_j) e_{k'}, e_k \right\rangle \right|}{\phi - b}.$$

For the second term, we have

$$\left| \left\langle \sum_{k'=1}^d \alpha_j^{k'}(t-1) (\mathbf{I} - H_j^\top H_j) e_{k'}, e_k \right\rangle \right| \\ \leq \left(\max_{j \in \mathcal{V}/\mathcal{A}} \max_{1 \leq k' \leq d} |\alpha_j^{k'}(t-1)| \right) \|e_k^\top (\mathbf{I} - H_j^\top H_j)\|_1 \\ = \left(\max_{j \in \mathcal{V}/\mathcal{A}} \|x_j(t-1) - \theta^*\|_\infty \right) \|(\mathbf{I} - H_j^\top H_j) e_k\|_1,$$

where the last equality follows from the fact that $(\mathbf{I} - H_j^\top H_j)$ is symmetric. For the first term, we have

$$\max_{1 \leq k \leq d} \left| \left\langle H_j^\top \frac{1}{t} \sum_{r=1}^t w_j(r), e_k \right\rangle \right| \leq \left\| H_j^\top \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2 \\ \leq C_0 \left\| \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2.$$

By Assumption 1, we have

$$\max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(t) - \theta^*\|_\infty \\ \leq \rho \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(t-1) - \theta^*\|_\infty + \max_{i \in \mathcal{V}/\mathcal{A}} C_0 \left\| \frac{1}{t} \sum_{r=1}^t w_i(r) \right\|_2 \\ \leq \rho \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(t-1) - \theta^*\|_\infty + \sum_{i \in \mathcal{V}/\mathcal{A}} C_0 \left\| \frac{1}{t} \sum_{r=1}^t w_i(r) \right\|_2 \\ \leq \rho^t \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(0) - \theta^*\|_\infty + C_0 \sum_{j \in \mathcal{V}/\mathcal{A}} R_j(\rho, t).$$

By Lemmas 2 and 3 with $\lambda = \rho$, we complete the proof.

PROOF OF THEOREM 3

We first show that the evolution of $\|x_i(t) - \theta^*\|_\infty$ – the ℓ_∞ norm of the estimation errors – for all $i \in \mathcal{V}/\mathcal{A}$ collectively have a matrix representation. Then we bound the convergence rate of the obtained matrix product.

Similar to the proof of Theorem 1, for any $i \in \mathcal{V}/\mathcal{A}$ and any k , we have

$$|\alpha_i^k(t)| \leq \left| \sum_{j \in \mathcal{N}_i/\mathcal{A}} \beta_{ij}^k(t) \left\langle H_j^\top \frac{1}{t} \sum_{r=1}^t w_j(r), e_k \right\rangle \right| \\ + \sum_{j \in \mathcal{N}_i/\mathcal{A}} \beta_{ij}^k(t) \left| \left\langle (\mathbf{I} - H_j^\top H_j) I \left(\sum_{k'=1}^d \alpha_j^{k'}(t-1) e_{k'} \right), e_k \right\rangle \right|.$$

For the second term, we have

$$\sum_{j \in \mathcal{N}_i/\mathcal{A}} \beta_{ij}^k(t) \left| \left\langle (\mathbf{I} - H_j^\top H_j) \left(\sum_{k'=1}^d \alpha_j^{k'}(t-1) e_{k'} \right), e_k \right\rangle \right| \\ \leq \sum_{j \in \mathcal{N}_i/\mathcal{A}} \beta_{ij}^k(t) \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 \|x_j(t-1) - \theta^*\|_\infty.$$

For the first term, we have

$$\left| \sum_{j \in \mathcal{N}_i/\mathcal{A}} \beta_{ij}^k(t) \left\langle H_j^\top \frac{1}{t} \sum_{r=1}^t w_j(r), e_k \right\rangle \right| \\ \leq C_0 \max_{j \in \mathcal{V}/\mathcal{A}} \left\| \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2.$$

Thus, we get

$$\begin{aligned} \|x_i(t) - \theta^*\|_\infty &= \max_{1 \leq k \leq d} |\alpha_i^k(t)| \\ &\leq \max_{1 \leq k \leq d} \sum_{j \in \mathcal{N}_i/\mathcal{A}} \beta_{ij}^k(t) \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 \|x_j(t-1) - \theta^*\|_\infty \\ &\quad + C_0 \max_{j \in \mathcal{V}/\mathcal{A}} \left\| \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2. \end{aligned}$$

Let $\mathbf{E}(t) \in \mathbb{R}^\phi$ be the vector that stacks the ℓ_∞ norm of the errors $x_i(t) - \theta^*$ for all $i \in \mathcal{V}/\mathcal{A}$. For each $i \in \mathcal{V}/\mathcal{A}$, define matrix $M(t)$ as follows:

$$M_{i,j}(t) = \beta_{i,j}^{k_i^*(t)} \|(\mathbf{I} - H_j^\top H_j) e_{k_i^*(t)}\|_1,$$

where $k_i^*(t)$ is an arbitrary maximizer of

$$\sum_{j \in \mathcal{N}_i/\mathcal{A}} \beta_{ij}^k(t) \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 \|x_j(t-1) - \theta^*\|_\infty$$

over $k = 1, \dots, d$. With this rewriting, we have

$$\begin{aligned} \mathbf{E}(t) &\leq M(t) \mathbf{E}(t-1) + C_0 \max_{j \in \mathcal{V}/\mathcal{A}} \left\| \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2 \mathbf{1} \\ &\leq \left(\prod_{r=1}^t M(r) \right) \mathbf{E}(0) + C_0 \max_{j \in \mathcal{V}/\mathcal{A}} \left\| \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2 \mathbf{1} \\ &\quad + C_0 \sum_{m=1}^{t-1} \max_{j \in \mathcal{V}/\mathcal{A}} \left\| \frac{\sum_{r=1}^{t-m} w_j(r)}{t-m} \right\|_2 \left(\prod_{r=t-m+1}^t M(r) \right) \mathbf{1}. \end{aligned}$$

Note that $M(t)$ is random, and its realization is determined by both the noises in the good agents' local observations and the Byzantine agents' adversarial behaviors. Nevertheless, this does not complicate our analysis because our analysis works for every realization of $M(t)$. Henceforth, with a little abuse of notation, we use $M(t)$ to denote both the random matrix and its realization.

By Lemma 4 and Assumption 2, we know that for every t , the matrix $M(t)$ is a *strict* sub-stochastic matrix. In particular, under the assumptions in Theorem 3, the following claim is true.

Claim 1. *For any t_0 and for any sequence of realization of the matrices $M(t)$ for $t = t_0 + 1, \dots, t_0 + \xi\phi$, the following holds*

$$\left(\prod_{t=t_0+1}^{t_0+\xi\phi} M(t) \right) \mathbf{1} \leq \gamma \mathbf{1}, \text{ where } \gamma := 1 - \frac{1 - \rho_0}{(2(\phi - b))^{\xi\phi}}.$$

For ease of exposition, the proof of Claim 1 is deferred to the end of this paper.

With Claim 1, for any fixed t_0 and for sufficiently large $t - t_0$, we have

$$\begin{aligned} &\left(\prod_{r=t_0+1}^t M(r) \right) \mathbf{1} \\ &= \left(\prod_{r=t_0+\xi\phi+1}^t M(r) \right) \left(\prod_{r=t_0+1}^{t_0+\xi\phi} M(r) \right) \mathbf{1} \\ &\leq \gamma \left(\prod_{r=t_0+\xi\phi+1}^t M(r) \right) \mathbf{1} \\ &\leq \gamma^{\lfloor \frac{t-t_0}{\xi\phi} \rfloor} \left(\prod_{r=\lfloor \frac{t-t_0}{\xi\phi} \rfloor \xi\phi+1}^t M(r) \right) \mathbf{1} \\ &\leq \gamma^{\lfloor \frac{t-t_0}{\xi\phi} \rfloor} \mathbf{1}. \end{aligned}$$

Thus,

$$\begin{aligned} \left(\prod_{r=1}^t M(r) \right) \mathbf{E}(0) &\leq \left(\prod_{r=1}^t M(r) \right) \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(0) - \theta^*\|_\infty \mathbf{1} \\ &\leq \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(0) - \theta^*\|_\infty \gamma^{\lfloor \frac{t}{\xi\phi} \rfloor} \mathbf{1}. \end{aligned}$$

In addition,

$$\begin{aligned} &\sum_{m=0}^{t-1} \max_{j \in \mathcal{V}/\mathcal{A}} \left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_2 \left(\prod_{r=t-m+1}^t M(r) \right) \mathbf{1} \\ &\leq \sum_{m=0}^{t-1} \max_{j \in \mathcal{V}/\mathcal{A}} \left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_2 \gamma^{\lfloor \frac{m}{\xi\phi} \rfloor} \mathbf{1} \\ &\leq \sum_{m=0}^{t-1} \sum_{j \in \mathcal{V}/\mathcal{A}} \left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_2 \gamma^{\lfloor \frac{m}{\xi\phi} \rfloor} \mathbf{1}. \end{aligned}$$

For ease of exposition, we assume that $\lfloor \frac{m}{\xi\phi} \rfloor$ is an integer for any m . Note that this simplification does not affect the order of convergence.

$$\begin{aligned} \mathbf{E}(t) &\leq \left(\max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(0) - \theta^*\|_\infty \right) \gamma^{\frac{t}{\xi\phi}} \mathbf{1} \\ &\quad + C_0 \sum_{j \in \mathcal{V}/\mathcal{A}} \sum_{m=0}^{t-1} \left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_2 \gamma^{\frac{m}{\xi\phi}} \mathbf{1} \\ &\leq \left(\max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(0) - \theta^*\|_\infty \right) \gamma^{\frac{t}{\xi\phi}} \mathbf{1} \\ &\quad + C_0 \sum_{j \in \mathcal{V}/\mathcal{A}} R_j(\gamma^{\frac{1}{\xi\phi}}, t). \end{aligned}$$

Applying Lemma 2 with $\lambda = \gamma^{\frac{1}{\xi\phi}}$, we have

$$0 \leq \lim_{t \rightarrow \infty} \mathbf{E}(t) \leq 0 + 0 + 0 = 0, \text{ almost surely.}$$

In addition, by applying Lemma 3 with $\lambda = \gamma^{\frac{1}{\xi\phi}}$, we complete the proof.

PROOF OF CLAIM 1

Recall that $M(t)$ (for each $t \geq 1$) is defined as

$$M_{i,j}(t) = \beta_{i,j}^{k_i^*(t)} \|(\mathbf{I} - H_j^\top H_j) e_{k_i^*(t)}\|_1,$$

where $k_i^*(t)$ is an arbitrary maximizer of

$$\sum_{j \in \mathcal{N}_i/\mathcal{A}} \beta_{i,j}^{k_i^*(t)} \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 \|x_j(t-1) - \theta^*\|_\infty$$

over $k = 1, \dots, d$.

For any sequence of realization of the matrices $M(t)$ for $t = t_0 + 1, \dots, t_0 + \xi\phi$, we construct a sequence of auxiliary *stochastic* matrices, denoted by $\widetilde{M}(t)$, as follows:

$$\widetilde{M}_{ij}(t) := \beta_{ij}^{k_i^*(t)}, \quad \forall i, j \in \mathcal{V}/\mathcal{A}.$$

By Lemma 4, $\widetilde{M}(t)$ is row-stochastic for $t = t_0 + 1, \dots, t_0 + \xi\phi$. By Definition 1 and Lemma 4, for each t there exists a reduced graph in $\mathcal{H}$ such that

$$\widetilde{M}(t) \geq \frac{1}{2(\phi - b)} H(t),$$

where $H(t)$ is the adjacency matrix of the corresponding reduced graph. For ease of exposition, with a little abuse of notation, we use $H(t)$ to denote both the adjacency matrix and the reduced graph.¹ We refer to $H(t)$ as the *shadow graph* at time t .

Since the matrix product $\prod_{t=t_0+1}^{t_0+\xi\phi} M(t)$ consists of $\xi\phi$ shadow graphs and $|\mathcal{H}| = \xi$, there exists at least one reduced graph in $\mathcal{H}$ that appears at least ϕ times in the sequence of shadow graphs. Let $\widetilde{H}$ be one such reduced graph. Without loss of generality, let i_0 be the node in the unique source component of $\widetilde{H}$ such that

$$\begin{aligned} & \left| (\mathcal{N}_{i_0} \cup \{i_0\})/\mathcal{A} \cap \left\{ j : \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 < 1 \right\} \right| \\ & \geq b + 1. \end{aligned}$$

Since i_0 is in the unique source component of $\widetilde{H}$, it follows that node i_0 can reach every other good agents within $\phi - 1$ hops using the edges in $\widetilde{H}$ only.

For any given realization of $M(t_0 + 1), \dots, M(t_0 + \xi\phi)$, let $\tau_1, \dots, \tau_\phi$ be the first ϕ time indices at which H is the shadow graph. In addition, let

$$\Delta_j := \tau_j - \tau_{j-1}, \quad \forall j = 2, \dots, \phi.$$

For ease of exposition, in the remainder of this proof, we assume $t_0 = 0$. The proof can be easily generalized to arbitrary t_0 . Let

$$\eta(t) := \left(\prod_{r=1}^t M(r) \right) \mathbf{1}, \quad \forall t.$$

Note that $\eta(t) \leq \mathbf{1}$ as $M(r)$ is sub-stochastic for all r .

¹Its meaning should be clear from the context.

To show Claim 1, it is enough to show the following three claims.

(A) For any $j = 1, \dots, \phi$,

$$\eta_{i_0}(\tau_j) \leq 1 - \frac{1 - \rho_0}{2(\phi - b)};$$

(B) If i is an outgoing neighbor of i_0 in the shadow graph $\widetilde{H}$, i.e., $\widetilde{H}_{ii_0} = 1$, then for any $j = 2, \dots, \phi$,

$$\eta_i(\tau_j) \leq 1 - \frac{1 - \rho_0}{(2(\phi - b))^2}.$$

(C) For any $j = 3, \dots, \phi$, if i_0 can reach node i in the shadow graph $\widetilde{H}$ with h hops, where $2 \leq h \leq j - 1$, then

$$\eta_i(\tau_j) \leq 1 - \frac{1 - \rho_0}{(2(\phi - b))^{2 + \sum_{j'=j+2-h}^j \Delta_{j'}}}.$$

Suppose Claims (A), (B), and (C) hold. Recall that i_0 is in the unique source component of $\widetilde{H}$. At time τ_ϕ , at all $i \in \mathcal{V}/\mathcal{A}$, it holds that

$$\begin{aligned} \eta_i(\tau_\phi) & \leq 1 - \frac{1 - \rho_0}{(2(\phi - b))^{2 + \sum_{j'=3}^\phi \Delta_{j'}}} \\ & \leq 1 - \frac{1 - \rho_0}{(2(\phi - b))^{\xi\phi}}. \end{aligned}$$

Therefore, we conclude that

$$\begin{aligned} \eta(\xi\phi) & = \left(\prod_{r=\tau_\phi+1}^{\xi\phi} M(r) \right) \eta(\tau_\phi) \\ & \leq \left(1 - \frac{1 - \rho_0}{(2(\phi - b))^{\xi\phi}} \right) \left(\prod_{r=\tau_\phi+1}^{\xi\phi} M(r) \right) \mathbf{1} \\ & \leq \left(1 - \frac{1 - \rho_0}{(2(\phi - b))^{\xi\phi}} \right) \mathbf{1}, \end{aligned}$$

proving Claim 1.

In the remainder of the proof, we prove Claims (A), (B), and (C), individually.

a) *We first show (A):* For any $j = 1, \dots, \phi$, we have

$$\eta(\tau_j) \leq M(\tau_j) \mathbf{1}.$$

Thus

$$\begin{aligned}
\eta_{i_0}(\tau_j) &\leq \sum_{i \in \mathcal{V}/\mathcal{A}} M_{i_0 i}(\tau_j) \\
&= \sum_{i \in \mathcal{V}/\mathcal{A}} \beta_{i_0 i}^{k_{i_0}^*(\tau_j)} \left\| (\mathbf{I} - H_i^\top H_{i_0}) e_{k_{i_0}^*(\tau_j)} \right\|_1 \\
&\leq \sum_{i \in \mathcal{V}/\mathcal{A} \text{ \& } \left\| (\mathbf{I} - H_i^\top H_{i_0}) e_{k_{i_0}^*(\tau_j)} \right\|_1 < 1} \beta_{i_0 i}^{k_{i_0}^*(\tau_j)} \rho_0 \\
&\quad + \sum_{i \in \mathcal{V}/\mathcal{A} \text{ \& } \left\| (\mathbf{I} - H_i^\top H_{i_0}) e_{k_{i_0}^*(\tau_j)} \right\|_1 = 1} \beta_{i_0 i}^{k_{i_0}^*(\tau_j)}.
\end{aligned}$$

By Lemma 4, Assumption 2, and the choice of i_0 , we know that

$$\begin{aligned}
&\sum_{i \in \mathcal{V}/\mathcal{A} \text{ \& } \left\| (\mathbf{I} - H_i^\top H_{i_0}) e_{k_{i_0}^*(\tau_j)} \right\|_1 < 1} \beta_{i_0 i}^{k_{i_0}^*(\tau_j)} \\
&\geq \frac{1}{2(|\mathcal{N}_{i_0} \cup \{i_0\}|/\mathcal{A}| - b)} \\
&\geq \frac{1}{2(\phi - b)}.
\end{aligned}$$

Thus, we have

$$\eta_{i_0}(\tau_j) \leq 1 - \frac{1 - \rho_0}{2(\phi - b)}.$$

b) Next we show (B): For any $j = 2, \dots, \nu$,

$$\eta(\tau_j) = M(\tau_j)\eta(\tau_j - 1) = \sum_{i' \in \mathcal{V}/\mathcal{A}} M_{ii'}(\tau_j)\eta_{i'}(\tau_j - 1).$$

Recall that

$$M_{ii_0}(\tau_j) = \beta_{ii_0}^{k_i^*(\tau_j)} \left\| (\mathbf{I} - H_{i_0}^\top H_i) k_i^*(\tau_j) \right\|_1.$$

We consider two cases:

- (1) $\left\| (\mathbf{I} - H_{i_0}^\top H_i) k_i^*(\tau_j) \right\|_1 < 1$;
- (2) $\left\| (\mathbf{I} - H_{i_0}^\top H_i) k_i^*(\tau_j) \right\|_1 = 1$.

Suppose that $\left\| (\mathbf{I} - H_{i_0}^\top H_{i_0}) k_{i_0}^*(\tau_j) \right\|_1 < 1$. Since $\tilde{H}_{ii_0} = 1$, it follows that

$$\tilde{M}_{ii_0}(\tau_j) = \beta_{ii_0}^{k_{i_0}^*(\tau_j)} \geq \frac{1}{2(\phi - b)}.$$

Thus, we have

$$\begin{aligned}
\eta_i(\tau_j) &\leq M_{ii_0}(\tau_j) + \sum_{i' \in \mathcal{V}/\mathcal{A} \text{ \& } i' \neq i_0} M_{ii'}(\tau_j) \\
&\leq \beta_{ii_0}^{k_{i_0}^*(\tau_j)} \rho_0 + \sum_{i' \in \mathcal{V}/\mathcal{A} \text{ \& } i' \neq i_0} \beta_{ii'}^{k_{i_0}^*(\tau_j)} \\
&= 1 - \beta_{ii}^{k_{i_0}^*(\tau_j)} (1 - \rho_0) \\
&\leq 1 - \frac{1 - \rho_0}{2(\phi - b)}.
\end{aligned}$$

Suppose that $\left\| (\mathbf{I} - H_{i_0}^\top H_{i_0}) k_{i_0}^*(\tau_j) \right\|_1 = 1$. In this case

$$M_{ii_0}(\tau_j) = \tilde{M}_{ii_0}(\tau_j) \geq \frac{1}{2(\phi - b)}.$$

Thus, we have

$$\begin{aligned}
\eta_i(\tau_j) &= M_{ii_0}(\tau_j)\eta_{i_0}(\tau_j - 1) \\
&\quad + \sum_{i' \in \mathcal{V}/\mathcal{A} \text{ \& } i' \neq i_0} M_{ii'}(\tau_j)\eta_{i'}(\tau_j - 1) \\
&\leq M_{ii_0}(\tau_j) \left(1 - \frac{1 - \rho_0}{2(\phi - b)} \right) \\
&\quad + \sum_{i' \in \mathcal{V}/\mathcal{A} \text{ \& } i' \neq i_0} M_{ii'}(\tau_j) \\
&\leq \sum_{i' \in \mathcal{V}/\mathcal{A}} M_{ii'}(\tau_j) - \frac{1 - \rho_0}{2(\phi - b)} M_{ii_0}(\tau_j) \\
&\leq 1 - \frac{1 - \rho_0}{(2(\phi - b))^2}.
\end{aligned}$$

c) Finally we show (C): We prove this by induction.

Base case: $j = 3$

Let i be a 2-th order neighbor of node i_0 in the shadow graph $\tilde{H}$; there exists a directed path of length 2 such that $i_0 \rightarrow i_1 \rightarrow i$ in $\tilde{H}$.

If $\left\| (\mathbf{I} - H_{i_1}^\top H_{i_0}) k_{i_0}^*(\tau_3) \right\|_1 < 1$, similar to the proof of Claim (B), we have that

$$\eta_i(\tau_3) \leq 1 - \frac{1 - \rho_0}{2(\phi - b)}.$$

Now suppose $\left\| (\mathbf{I} - H_{i_1}^\top H_{i_0}) k_{i_0}^*(\tau_3) \right\|_1 = 1$.

If there exists r where $\tau_2 + 1 \leq r \leq \tau_3 - 1$ such that

$$\left\| (\mathbf{I} - H_{i_1}^\top H_{i_1}) k_{i_1}^*(r) \right\|_1 < 1,$$

i.e., $M_{i_1 i_1}(r) < \tilde{M}_{i_1 i_1}(r)$. Let r^* be the latest time index. Note that $\beta_{ii}^k(t) \geq \frac{1}{2(\phi - b)}$ for any $i \in \mathcal{V}/\mathcal{A}$, t and k . We have

$$\eta_{i_1}(r^*) \leq \sum_{i' \in \mathcal{V}/\mathcal{A}} M_{i_1 i'}(r^*) \leq 1 - \frac{1 - \rho_0}{2(\phi - b)}.$$

In addition, by the choice of r^* , we have

$$\left[\prod_{r=r^*+1}^{\tau_3-1} M(r) \right]_{i_1 i_1} \geq \frac{1}{(2(\phi - b))^{\tau_3 - r^* - 1}}.$$

So we get

$$\begin{aligned}
\eta_{i_1}(\tau_3 - 1) &= \left[\prod_{r=r^*+1}^{\tau_3-1} M(r) \right]_{i_1 i_1} \eta_{i_1}(r^*) \\
&\quad + \sum_{i' \in \mathcal{V}/\mathcal{A}} \left[\prod_{r=r^*+1}^{\tau_3-1} M(r) \right]_{i_1 i'} \eta_{i'}(r^*) \\
&\leq 1 - \frac{1 - \rho_0}{(2(\phi - b))^{\tau_3 - r^*}}.
\end{aligned}$$

As $\|(\mathbf{I} - H_{i_1}^\top H_{i_1}) k_i^*(\tau_3)\|_1 = 1$ and $\beta_{i_0}^{i^*}(\tau_3) \geq \frac{1}{2(\phi-b)}$, we get that

$$\eta_i(\tau_3) \leq 1 - \frac{1 - \rho_0}{(2(\phi-b))^{\tau_3-r^*+1}} \leq 1 - \frac{1 - \rho_0}{(2(\phi-b))^{\Delta_3}}.$$

To finish the proof of the base case, it remains to consider the case that

$$\|(\mathbf{I} - H_{i_1}^\top H_{i_1}) k_{i_1}^*(r)\|_1 = 1,$$

i.e., $M_{i_1 i_1}(r) = \widetilde{M}_{i_1 i_1}(r)$ for all r such that $\tau_2 + 1 \leq r \leq \tau_3 - 1$. Thus, we get

$$\left[\prod_{r=\tau_2+1}^{\tau_3-1} M(r) \right]_{i_1 i_1} \geq \frac{1}{(2(\phi-b))^{\Delta_3-1}}.$$

So

$$\begin{aligned} \eta_{i_1}(\tau_3 - 1) &= \sum_{i' \in \mathcal{V}/\mathcal{A}} \left[\prod_{r=\tau_2+1}^{\tau_3-1} M(r) \right]_{i_1, i'} \eta_{i'}(\tau_2) \\ &\leq 1 - \left[\prod_{r=\tau_2+1}^{\tau_3-1} M(r) \right]_{i_1, i_1} \frac{1}{(2(\phi-b))^2} \\ &\leq 1 - \frac{1}{(2(\phi-b))^{\Delta_3+1}}, \end{aligned}$$

and

$$\eta_i(\tau_3) \leq 1 - \frac{1}{(2(\phi-b))^{\Delta_3+2}}.$$

Induction step: Suppose the following holds for any $j = 3, \dots, \phi - 1$:

$$\eta_i(\tau_j) \leq 1 - \frac{\rho_0}{(2(\phi-b))^{2+\sum_{j'=j+2-h}^j \Delta_{j'}}}$$

for all the h -th order neighbor of node i_0 in the shadow graph $\widetilde{H}$, where $h = 2, \dots, j - 1$.

Inductive step:

The proof of the inductive step is similar to the proof of the base case, thus is omitted.

REFERENCES

- [1] A. Speranzon, C. Fischione, and K. H. Johansson, "Distributed and collaborative estimation over wireless sensor networks," in *IEEE Conference on Decision and Control (CDC)*, 2006, pp. 1025–1030. [1](#)
- [2] L. Xie, D.-H. Choi, S. Kar, and H. V. Poor, "Fully distributed state estimation for wide-area monitoring systems," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1154–1169, 2012. [1](#)
- [3] B. Sinopoli, C. Sharp, L. Schenato, S. Schaffert, and S. S. Sastry, "Distributed control applications within sensor networks," *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1235–1246, 2003. [1](#)
- [4] R. Olfati-Saber, "Distributed kalman filtering for sensor networks," in *IEEE Conference on Decision and Control (CDC)*, 2007, pp. 5492–5498. [1](#)
- [5] S. Kar, J. M. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3575–3605, 2012. [1](#)
- [6] F. Bullo, J. Cortes, and S. Martinez, *Distributed control of robotic networks: a mathematical approach to motion coordination algorithms*. Princeton University Press, 2009, vol. 27. [1](#)
- [7] Y. Chen, S. Kar, and J. M. Moura, "The internet of things: Secure distributed inference," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 64–75, 2018. [1](#)
- [8] S. S. Stankovic, M. S. Stankovic, and D. M. Stipanovic, "Decentralized parameter estimation by consensus based stochastic approximation," *IEEE Transactions on Automatic Control*, vol. 56, no. 3, pp. 531–543, 2011. [1](#)
- [9] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, p. 44, 2017. [1](#)
- [10] S. Kar and J. M. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 355–369, 2009. [1](#)
- [11] N. A. Lynch, *Distributed Algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996. [1, 3](#)
- [12] M. Pease, R. Shostak, and L. Lamport, "Reaching agreement in the presence of faults," *Journal of the ACM (JACM)*, vol. 27, no. 2, pp. 228–234, 1980. [1, 5](#)
- [13] L. Lamport, R. Shostak, and M. Pease, "The byzantine generals problem," *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 4, no. 3, pp. 382–401, 1982. [1, 3](#)
- [14] H. Mendes and M. Herlihy, "Multidimensional approximate agreement in byzantine asynchronous systems," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 391–400. [1, 3, 5](#)
- [15] N. H. Vaidya and V. K. Garg, "Byzantine vector consensus in complete graphs," in *Proceedings of the 2013 ACM symposium on Principles of distributed computing*. ACM, 2013, pp. 65–73. [1, 3, 5](#)
- [16] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 645–658, 2011. [2](#)
- [17] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Transactions on Smart Grid*, vol. 2, no. 2, pp. 326–333, 2011. [2](#)
- [18] K. C. Sou, H. Sandberg, and K. H. Johansson, "On the exact solution to a smart grid cyber-security analysis problem," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 856–865, 2013. [2](#)
- [19] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013. [2](#)
- [20] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014. [2](#)
- [21] J. Mattingley and S. Boyd, "Real-time convex optimization in signal processing," *IEEE Signal processing magazine*, vol. 27, no. 3, pp. 50–61, 2010. [2](#)
- [22] Y. Shoukry and P. Tabuada, "Event-triggered state observers for sparse sensor noise/attacks," *IEEE Transactions on Automatic Control*, vol. 61, no. 8, pp. 2079–2091, 2016. [2](#)
- [23] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1495–1508, 2011. [2](#)
- [24] Y. Chen, S. Kar, and J. M. Moura, "Resilient distributed estimation through adversary detection," *IEEE Transactions on Signal Processing*, 2018. [2](#)
- [25] —, "Attack resilient distributed estimation: A consensus+ innovations approach," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 1015–1020. [2](#)
- [26] A. Mitra and S. Sundaram, "Resilient distributed state estimation for LTI systems," *arXiv preprint arXiv:1802.09651*, 2018. [2](#)
- [27] W. Xu, Z. Li, and Q. Ling, "Robust decentralized dynamic optimization at presence of malfunctioning agents," *Signal Processing*, vol. 153, pp. 24–33, 2018. [2](#)

- [28] L. Su and N. H. Vaidya, "Non-bayesian learning in the presence of byzantine agents," in *International Symposium on Distributed Computing*. Springer, 2016, pp. 414–427. [2](#), [3](#)
- [29] Z. Yang and W. U. Bajwa, "Byrdie: Byzantine-resilient distributed coordinate descent for decentralized learning," *arXiv preprint arXiv:1708.08155*, 2017. [2](#)
- [30] L. Su and N. H. Vaidya, "Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms," in *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing*. ACM, 2016, pp. 425–434. [2](#)
- [31] N. H. Vaidya, L. Tseng, and G. Liang, "Iterative approximate byzantine consensus in arbitrary directed graphs," in *Proceedings of the 2012 ACM symposium on Principles of distributed computing*. ACM, 2012, pp. 365–374. [5](#), [6](#)
- [32] N. Vaidya, "Matrix representation of iterative approximate byzantine consensus in directed graphs," *arXiv preprint arXiv:1203.1888*, 2012. [6](#)