



Finite-Time Guarantees for Byzantine-Resilient Distributed State Estimation With Noisy Measurements

Lili Su  and Shahin Shahrampour 

Abstract—This article considers resilient cooperative state estimation in unreliable multiagent networks. A network of agents aim to collaboratively estimate the value of an unknown vector parameter, while an *unknown* subset of agents suffer Byzantine faults. We refer to the faulty agents as Byzantine agents. Byzantine 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 agents is far from trivial. In this article, we propose a computationally efficient algorithm that is provably robust to Byzantine agents. At each iteration of the algorithm, a good agent performs a gradient descent update based on noisy local measurements, exchanges its update with other agents in its neighborhood, and 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.

Index Terms—Agents and autonomous systems, cooperative control, parameter estimation, secure distributed state estimation.

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 only noisy

and even highly incomplete information about θ^* . Nevertheless, the network of agents might be able to collaboratively learn θ^* by effectively fusing the information scattered across the network in agents' local measurements.

In the absence of system adversary, the distributed state estimation problem is well studied [5], [8]. However, some practical scenarios such as IoT, microgrids, and Federated Learning are vulnerable to unstructured faults or even adversarial attacks [9]. In particular, in large distributed systems, individual computing devices/sensors may exhibit abnormal behaviors due to unreliable devices and communication channels, and even external adversarial attacks. Such abnormal behaviors are often unstructured because of the heterogeneity in hardware, software, implementation environments, and the unpredictability of external adversarial attacks.

Motivated by that, we are interested in addressing collaborative estimation in the presence of adversarial agents. Despite the wealth of literature on collaborative estimation with random failures (e.g., [10]), perhaps less well known is estimation in the presence of *adversarial* agents, especially in the *finite-time* domain.

In this article, we adopt a Byzantine fault/adversary model [11]—a canonical fault/adversary 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 (referred to as a Byzantine agent) 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 agents is highly nontrivial (see, e.g., [12] and [13]). For example, it is well known that in complete graphs, no consensus algorithms can tolerate more than one-third 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 multidimensional 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 agents without stringent assumptions on the local measurements are still lacking. In particular, the literature has mostly focused

Manuscript received June 3, 2019; accepted October 19, 2019. Date of publication November 6, 2019; date of current version August 28, 2020. The work of L. Su was supported by the National Science Foundation under Awards CCF-1461559 and CCF-0939370. Recommended by Associate Editor Z. Chen. (*Corresponding author: Shahin Shahrampour.*)

L. Su is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: lilisu@mit.edu).

S. Shahrampour is with the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: shahin@tamu.edu).

Color versions of one or more of the figures in this article are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TAC.2019.2951686

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 article.

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 performs a gradient descent update based on local measurements only, exchanges its update with other agents in its neighborhood, and robustly aggregates the received messages using coordinate-wise trimmed means.

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. To the best of our knowledge, this is the first finite-time fast convergence guarantee of Byzantine-resilient distributed state estimation for the case that the good agents can only collect noisy measurements and their local observation matrices might not be of full rank.

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

1) Adversary-Resilient State Estimation: There is a rich line of work on adversary-resilient state estimation problem, wherein the existence of a fusion center is assumed. 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] and [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. Shoukry and Tabuada [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. Secure remote estimation of a linear Gaussian process is considered in [23], which focuses on malicious sensor detection and secure estimation in the fusion center. Interested readers are referred to [24] for a comprehensive survey on security control in industrial cyber-physical systems. In contrast, in our setting (*multiagent* networks), no fusion center exists, and transmitting the locally collected measurements to one designated agent is

forbidden; the state estimation problem must be solved in a *decentralized* manner.

2) Adversary-Resilient Distributed State Estimation: In parallel to advancements on resilient centralized estimation, recent years have witnessed intensive interest in securing distributed estimation. Sundaram and Hadjicostis [25] discuss the problem of adversary-resilient consensus. Chen *et al.* [26] propose a novel adversary detection strategy under which good agents either asymptotically learn the true state or detect the existence of a system adversary. If an adversary is flagged, the system goes through some external procedure to “repair” itself. This method is satisfactory as long as the system barely needs to go through the external procedure, which is often expensive. However, for scenarios where the existence of a system adversary is the norm, which may be the case for large-scale distributed systems, instead of such an adversary detection strategy, we need to seek for securing strategies that can *tolerate* the existence of a system adversary so that the good agents can learn the true state even in the presence of Byzantine agents. Several adversary-resilient algorithms have been proposed [27]–[33] with different assumptions and performance guarantees. Chen *et al.* [27], [32] 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 adversarial. The correctness of their algorithm is shown under the assumption that an agent can *fully* observe the true state [27, Sec. II.A] and [32, eq. (1)], as opposed to our model, which deals with both observability and noisy measurement issues. Exponential convergence is proved in [33] but under the above full observation assumption [33, eq. (1)]. Mitra and Sundaram [28] consider the more general LTI systems and characterize the fundamental limits on adversary-resilient algorithms. Under a different adversarial model,¹ a concurrent work [34] presents a method, whose correctness does not depend on the network topology. However, neither [28] nor [34] considers nonasymptotic convergence.

3) Adversary-Resilient Distributed Optimization: Xu *et al.* [29] study the general dynamic optimization problem and propose a total variation norm regularization technique to mitigate the effect of malfunctioning agents. However, even in the static case, the good agents cannot learn the true minimizer (see [29, Corollary 1]). Our algorithm is similar to [31] in that we combine local gradient descent with coordinatewise message trimming. For their algorithm to work, the optimization problem needs to be separable; otherwise, [31, Lemma 1] does not hold, and the proof in [35] cannot be applied. The algorithm *Byz-Iter* (proposed in [30] originally for distributed hypothesis testing problem) works for the state estimation problem, but its performance scales poorly in d . Shahrampour and Jadbabaie [36] consider a tracking problem in the presence of adversarial noise. In their work, all of the agents are assumed to be cooperative, i.e., they truthfully report their (partial) observations. Gupta and Vaidya propose a robust distributed gradient descent method, whose convergence does not depend on any distributional assumption [37]. However, the system architecture is a master–slave architecture rather than

¹ See the second last paragraph in [34, Sec. I.B] for details.

a multiagent network considered here. The same master-slave architecture is considered in a new line of work on distributed statistical learning [9], [38], [39], where the training data are assumed to be independent identically distributed (i.i.d.) generated from some unknown distribution. In contrast, in our distributed state estimation problem, the local measurements across different agents might follow different distributions.

II. PROBLEM FORMULATION

Notation: We represent by $\text{trace}(\cdot)$ the trace operator, by $\|\cdot\|$ the norm operator, and by $\mathbb{E}[\cdot]$ the expectation operator. $\mathbb{P}\{A\}$ denotes the probability of an event A , \mathbf{I} denotes the identity matrix, and e_k captures the k th standard basis in a Euclidean space. The vectors are all in column form.

A. Network Model

We consider a multiagent network, which is a collection of n agents communicating with each other through a network $G(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, n\}$ and \mathcal{E} denote the set of agents and communication links, 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* (i.e., *adversarial*). The set \mathcal{A} is chosen by the system adversary. We assume that $n \geq 2b + 1$. 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 Byzantine agents (agents in \mathcal{A}) can adversarially affect the estimation procedure by sending *arbitrary*, *malicious*, and possibly *conflicting* messages to the good agents; see Section II-C for details.

B. Observation Model

In this article, we focus on a linear observation model. Let $y_i(t)$ denote the local measurement of agent i at time t , i.e.,

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

where $\theta^* \in \mathbb{R}^d$ is the true state, $H_i \in \mathbb{R}^{n_i \times d}$ is the local observation matrix, and $w_i(t)$ is the observation noise. In particular, the noise sequence $\{w_i(t)\}_{t \geq 1}$ are i.i.d. with $\mathbb{E}[w_i(t)] = \mathbf{0} \in \mathbb{R}^{n_i}$, $\mathbb{E}[w_i(t)w_i(t)^\top] = \Sigma_i \in \mathbb{R}^{n_i \times n_i}$, and $\mathbb{P}\{\|w_i(t)\|_2 \leq C\} = 1$ for some absolute constant $C > 0$. Moreover, the noise sequences across agents are independent, that is, $(w_i(t), t \geq 1)$ and $(w_j(t), t \geq 1)$ for $i \neq j$ are independent. In practice, the observation matrix H_i is often fat, i.e., $n_i \ll d$. Thus, to correctly estimate θ^* , each agent i must obtain information from others.

Notably, though a Byzantine agent might send out malicious messages, its local observation is still well defined.

C. Fault/Adversary Model

To formally capture the unstructured abnormal behaviors of the adversarial agents, 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 to compromise and control. Recall that this set of agents is denoted by \mathcal{A} . An agent suffering Byzantine fault is referred to as a 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 very *powerful* in the sense that it has complete knowledge of the network, including the true state θ^* , the local program that each good agent is supposed to run, the current status and the running history of the multiagent network system, etc. Hence, the Byzantine agents can behave adaptively and collude with each other 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: 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 nontrivial even in complete graphs. In particular, it is well known that in complete graphs, no consensus algorithms can tolerate more than one-third of the agents to be Byzantine [13].

D. Finite-Time Versus Asymptotic Local Functions

The Byzantine-resilient state estimation problem can be viewed through the lens of distributed online optimization, where each good agent would only asymptotically know its local function. For each agent $i \in \mathcal{V}$, define its *asymptotic* local function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ as

$$f_i(x) := \frac{1}{2} \mathbb{E} [\|H_i x - y_i\|_2^2] \quad (2)$$

where $y_i = H_i \theta^* + w_i$ [as per (1)], and the expectation of $f_i(x)$ is taken over the randomness of w_i . Note that f_i is well defined for each agent regardless of whether it is a good agent or a Byzantine agent. 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/empirical* local function $f_{i,t}$ defined as

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

whose gradient at x is

$$\begin{aligned} \nabla f_{i,t}(x) &= \frac{1}{t} \sum_{s=1}^t H_i^\top (H_i x - y_i(s)) \\ &= H_i^\top H_i (x - \theta^*) - H_i^\top \frac{1}{t} \sum_{s=1}^t w_i(s) \end{aligned} \quad (4)$$

where the last equality follows from (1).

III. BYZANTINE-RESILIENT STATE ESTIMATION

To robustify distributed state estimation against Byzantine agents, one approach is to combine the local gradient descent with multidimensional Byzantine-resilient consensus [14], [15],

[30]. However, the performance of multidimensional Byzantine-resilient consensus itself is proved to scale poorly in the dimension of the parameter d [14], [15]. This is because 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 reduce the computation complexity, instead of using multidimensional Byzantine-resilient consensus, we robustly aggregate the received messages using coordinatewise trimmed means.

We propose an algorithm, named *Byzantine-resilient state estimation*, under which each good agent iteratively aggregates the received messages by, for each coordinate, discarding the largest b and the smallest b values, and averaging the remaining. In particular, in each iteration, an agent performs the following three steps.

- 1) *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.,

$$z_i(t) = x_i(t-1) - \nabla f_{i,t}(x_i(t-1)). \quad (5)$$

Note that the step size used is 1.

- 2) *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 $j \neq j' \in \mathcal{V} \setminus \mathcal{A}$.

- 3) *Robust aggregation*: Agent i computes the trimmed mean for each coordinate $k = 1, \dots, d$ and uses the obtained d trimmed means to obtain $x_i(t)$.

The formal description of the algorithm for a good agent (i.e., $i \in \mathcal{V} \setminus \mathcal{A}$) is given in Algorithm 1.

IV. FINITE-TIME GUARANTEE FOR COMPLETE NETWORKS

In this section, we focus on complete networks in order to build some intuitions for the resilience of the proposed algorithm without forcing the readers to worry about the complication caused by the network topologies. In fact, even complete networks themselves are of practical interests: many computer networks can be viewed as complete networks, wherein efficient communication protocols are implemented, and any two computers are logically connected.

Recall that a Byzantine agent has full knowledge of the system and can send out arbitrarily adversarial messages to the good agents.² Clearly, if a good agent is not equipped with any security strategy, even a single Byzantine agent might be able to control the evolutions of the local estimates x_i at the good agents. Fortunately, it can be shown that if the good agents use the

²Recall that a Byzantine agent can even send differently-valued messages to different good agents.

Algorithm 1: Byzantine-resilient state estimation.

Input: b and T

Initialization: Set $x_i(0)$ to an arbitrary value;
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 coordinate 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: $x_i(T)$.

coordinatewise trimming strategy in Algorithm 1, the evolutions of x_i use the information provided by the *good* agents only. More importantly, each of the good agent has only limited impact on x_i , formally stated next.

Lemma 1: At each good agent $i \in \mathcal{V}/\mathcal{A}$, for each iteration t and each coordinate $k \in \{1, \dots, d\}$, there exist convex coefficients $(\beta_{ij}^k(t), j \in \mathcal{V}/\mathcal{A})$ such that

$$1) \ x_i^k(t) = \sum_{j \in \mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) \langle z_j(t), e_k \rangle;$$

$$2) \ 0 \leq \beta_{ij}^k(t) \leq \frac{1}{\phi - b} \text{ for all } j \in \mathcal{V}/\mathcal{A} \text{ and } \sum_{j \in \mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) = 1.$$

Lemma 1 is proved in Appendix A. Notice that the sets of convex coefficients for different coordinates 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 coordinate, 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 differently valued messages to different neighbors, i.e., $m_{aj} \neq m_{aj'}$ for $a \in \mathcal{A}$ and $j \neq j'$.

Remark 2: The first item in Lemma 1 implies that the messages sent by the Byzantine agents are not used by the good agents. As a result, the Byzantine agents cannot have arbitrary control over the local estimates at the good agents, and they can at most influence the “choice” of the convex combination weights β_{ij}^k . Furthermore, the second item in Lemma 1 implies that Byzantine agents cannot significantly manipulate these weights.

Recall that a good agent can only get noisy local measurements of the true state. By standard concentration argument, we know for sufficiently large t , the gradient of the empirical local function $\nabla f_{i,t}$, which is defined in (4), is close to the gradient of the asymptotic local function. However, it is unclear whether the *overall* impacts of the measurement noises at all the agents

in the network can be controlled or not. This is because that agents perform local gradient descent as a subroutine only, and agents exchange messages with their neighbors in each iteration. As a result of message exchange, the observation noises quickly get mixed among the agents—losing independence—and their impacts might cumulative over iterations. Moreover, the impact of the random observation noises interplays with the adversarial behaviors of the Byzantine agents.

It turns out that the following quantity is crucial in bounding the overall impacts of observation noises.

For a given $\lambda \in (0, 1)$, let

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

for all $j \in \mathcal{V}/\mathcal{A}$ and $t \geq 1$. The following two concentration results are two key auxiliary lemmas for our main theorem.

Lemma 2: Given a $\lambda \in (0, 1)$, it is true that

$$\lim_{t \rightarrow \infty} R_j(\lambda, t) = 0 \text{ almost surely } \forall j \in \mathcal{V}/\mathcal{A}, \text{ and } t \geq 1.$$

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

Lemma 3: Given a $\lambda \in (0, 1)$, for any $\epsilon > 0$, it holds that

$$\begin{aligned} \mathbb{P} \left\{ R_j(\lambda, t) \geq \sqrt{\text{trace}(\Sigma_j)} \sum_{m=1}^{t-1} \lambda^m \frac{1}{\sqrt{t-m}} + \epsilon \right\} \\ \leq \exp \left(\frac{-\epsilon^2(1-\lambda)^2 t}{32 C^2} \right) \quad \forall j \in \mathcal{V}/\mathcal{A}, \text{ and } t \geq 1. \end{aligned}$$

Recall that $\text{trace}(\Sigma_j)$ is the sum of all the diagonal entries of Σ_j . Lemmas 2 and 3 are proved in Appendixes B and C, respectively. The following corollary follows immediately from Lemma 3; thus, its proof is omitted.

Corollary 1: For any given $\delta \in (0, 1)$ and $\epsilon > 0$, if $t \geq \frac{32 C^2}{(1-\lambda)^2 \epsilon^2} (\log \frac{1}{\delta} + \log \phi)$, then with probability at least $1 - \delta$, it holds that

$$R_j(\lambda, t) \leq \sqrt{\text{trace}(\Sigma_j)} \sum_{m=1}^{t-1} \lambda^m \frac{1}{\sqrt{t-m}} + \epsilon \quad \forall j \in \mathcal{V}/\mathcal{A}.$$

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

Assumption 1: $\frac{1}{\phi-b} \sum_{j \in \mathcal{V}/\mathcal{A}} \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 < 1$, for $k = 1, \dots, d$.

Assumption 1 is a sufficient condition that might not be necessary for general Byzantine-resilient distributed state estimation. We present a high-level intuition of Assumption 1 as follows: e_k can be viewed as one unit estimation error in the k th coordinate. After the local gradient descent update (right before averaging) at agent j , this one unit error becomes $(\mathbf{I} - H_j^\top H_j) e_k$. In a sense, $\|(\mathbf{I} - H_j^\top H_j) e_k\|_1$ quantifies agent j 's capability in reducing the estimation error in the k th coordinate of the *previous iteration* via local gradient descent update. Similarly, $\frac{1}{\phi-b} \sum_{j \in \mathcal{V}/\mathcal{A}} \|(\mathbf{I} - H_j^\top H_j) e_k\|_1$ is the *modified* average of such capability across all the good agents.³ Assumption 1 implies that

the good agents can collectively reduce the estimation errors in each coordinate. Now, let

$$\rho := \max_{k: 1 \leq k \leq d} \frac{\sum_{j \in \mathcal{V}/\mathcal{A}} \|(\mathbf{I} - H_j^\top H_j) e_k\|_1}{\phi - b}. \quad (7)$$

Clearly, $\rho < 1$ under Assumption 1.

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

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

Moreover, for any $\epsilon > 0$, with probability at least $1 - \phi \exp(\frac{-\epsilon^2(1-\rho)^2 t}{32 C^2})$, it holds that

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

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

The following corollary follows immediately from Theorem 1.

Corollary 2: For any given $\delta \in (0, 1)$ and $\epsilon > 0$, if $t \geq \frac{32 C^2}{(1-\rho)^2 \epsilon^2} (\log \frac{1}{\delta} + \log \phi)$, then with probability at least $1 - \delta$, it holds that

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

Theorem 1 indicates that all good agents (in a complete graph) are 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 (7).

V. FINITE-TIME GUARANTEES FOR INCOMPLETE NETWORKS

In this section, we extend our results to incomplete networks. Two types of communication are discussed.

A. Incomplete Graphs: Multihop Communication

So far, our analysis of Algorithm 1 has been 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 that the networked agents are allowed to relay the messages sent by others such that multihop communication can be implemented. We can adopt cryptographic solutions to force the Byzantine agents to either refuse to relay information or faithfully relay the messages without alternation [12]. Thus, as long as the node connectivity

³Notably, $(\mathbf{I} - H_j^\top H_j)$ also rotates e_k .

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, and essentially, all-to-all communication is ensured. 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. To make this article 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 is fundamentally different [14], [15], [40]. In particular, it has been shown that any Byzantine consensus algorithm scales poorly in the input dimension [14], [15]. Our algorithm relies on Byzantine-resilient consensus with scalar inputs—though $\theta^* \in \mathbb{R}^d$ is multidimensional.

Tight topological conditions are characterized in [40], where the conditions are stated in terms of a family of subgraphs of $G(\mathcal{V}, \mathcal{E})$, referred to as *reduced graphs*.

Definition 1 (see [40]): A reduced graph \tilde{G} of $G(\mathcal{V}, \mathcal{E})$ is obtained by 1) removing all nodes in \mathcal{A} , and all the links incident on set \mathcal{A} , and 2) for each node in \mathcal{V}/\mathcal{A} , removing up to b additional incoming links.

It is easy to see that the reduced graphs of a given graph $G(\mathcal{V}, \mathcal{E})$ are not unique. This nonuniqueness arises partially from the fact that the Byzantine agents can behave adaptively and arbitrarily. In a sense, reduced graphs 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.

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

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

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

The tight network topology condition for scalar-valued consensus to be achievable is characterized in [40].

Theorem 2 (see [40]): For scalar inputs, iterative approximate Byzantine consensus is achievable among good agents if and only if every reduced graph of $G(\mathcal{V}, \mathcal{E})$ contains only one source component.

The tight condition stated in Theorem 2 is on *every* reduced graph. Intuitively, this is because that the abnormal behaviors of Byzantine agents might be time varying, and consequently, the corresponding “effective” communication network is potentially time varying. In addition, we do not know which sequence of

reduced graphs is “picked” by the Byzantine agents throughout an execution.

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

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 that an assumption that is analogous to Assumption 1 holds.

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 incoming neighbors only.

Lemma 4 (see [41, Claim 2]): Suppose that Byzantine consensus can be achieved on graph $G(\mathcal{V}, \mathcal{E})$. Then, for each iteration t , each good agent $i \in \mathcal{V}/\mathcal{A}$, and each coordinate k , there exist convex coefficients $(\beta_{ij}^k(t), j \in \mathcal{N}_i \cup \{i\})$ such that

- 1) $x_i^k(t) = \sum_{j \in \mathcal{N}_i \cup \{i\}/\mathcal{A}} \beta_{ij}^k(t) \langle z_j(t), e_k \rangle$;
- 2) 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 $j \in \mathcal{B}_i(t)$.

Different from Lemma 1, wherein an uniform upper bound on β_{ij}^k for all i, j , and k is derived, an analogous upper bound is lacking in Lemma 4. Unfortunately, for incomplete networks such a uniform upper bound, if exists at all, might be hard to characterize. Nevertheless, we are able to show that when the network satisfies the tight condition for Byzantine consensus to be reachable, then a good agent will “assign” nontrivial weights to sufficiently many good incoming neighbors. As a result of this, if a good agent together with these good incoming neighbors *collectively* has enough information, then this agent is able to gradually reduce its local estimation error. Furthermore, if this agent is a node in the source component, then it is able to “propagate” the locally learned estimate to other good agents. These conditions are formally summarized in the following assumption.

Assumption 2: For each good node $j \in \mathcal{V}/\mathcal{A}$,

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

holds for each coordinate $k = 1, \dots, d$. In addition, any reduced graph \tilde{G} contains a node i in its unique source component such that

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

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

Note that in Assumption 2, \mathcal{N}_i includes the incoming neighbors of node i in the *original graph* $G(\mathcal{V}, \mathcal{E})$. Next, we present a high-level intuition of Assumption 2 as follows: Assumption 2 requires that $\|(\mathbf{I} - H_j^\top H_j) e_k\|_1 \leq 1$ for each good agent j , i.e., a good agent would not increase the one unit error via local gradient descent update. Assumption 2 additionally requires that any reduced graph contains a node in its source component that satisfies the stated condition, which ensures that such a node can learn the true state θ^* based on its own local measurements and the information aggregated from its neighbors. Since such a node is in the source component, it can reach every good agent in the reduced graph. Thus, all the good agents in the end can

learn θ^* . Despite the possibility that the Byzantine agents might “choose” different reduced graphs across iterations, as can be seen in the proof of Theorem 3, the sufficiency of Assumption 2 still holds.

For each coordinate k , let

$$\mathcal{D}_k := \left\{ j : \|(\mathbf{I} - H_j^\top H_j) e_k\|_1 < 1, \& j \notin \mathcal{A} \right\}.$$

Define ρ_0 as

$$\rho_0 := \max_{1 \leq k \leq d} \max_{j \in \mathcal{D}_k} \|(\mathbf{I} - H_j^\top H_j) e_k\|_1. \quad (8)$$

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 that Assumption 2 holds. Then, we have

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

Let $\gamma := 1 - \frac{1-\rho_0}{(2(\phi-b))\xi\phi}$. With probability at least

$$1 - \phi \exp\left(-\frac{\epsilon^2(1-\gamma\frac{1}{\xi\phi})^2 t}{32C^2}\right), \text{ it holds that}$$

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

VI. NUMERICAL EXAMPLE: ENERGY EFFICIENCY DATASET

We now provide empirical evidence in support of our algorithm by applying it to a regression dataset on UCI Machine Learning Repository.⁴ In this dataset, the state $\theta^* \in \mathbb{R}^8$ includes eight features: relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution. The regression model aims at representing heating load of residential buildings in terms of these features [42]. Since this dataset is real world and the ground truth value θ^* is unknown, we consider the solution of the centralized problem as the baseline. Then, we consider a network of $|\mathcal{V} \setminus \mathcal{A}| = 160$ agents. Each agent i observes only one feature corrupted by a Gaussian noise $\mathcal{N}(0, 0.25)$. Also, each agent i is connected to 40 agents $i - 20, i - 19, \dots, i + 19, i + 20$.

We consider inserting $|\mathcal{A}| \in \{1, \dots, 5\}$ Byzantine agents in the network. Throughout, the Byzantine agents can send out completely arbitrary messages in lieu of true gradients. We generate these arbitrary messages using a random eight-dimensional vector, each component of which is sampled from $\mathcal{N}(0, 9)$.

Let us now define the network performance metric as

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

and plot in Fig. 1 the error for various values $|\mathcal{A}|$. We observe that increasing the number of Byzantine agents degrades the

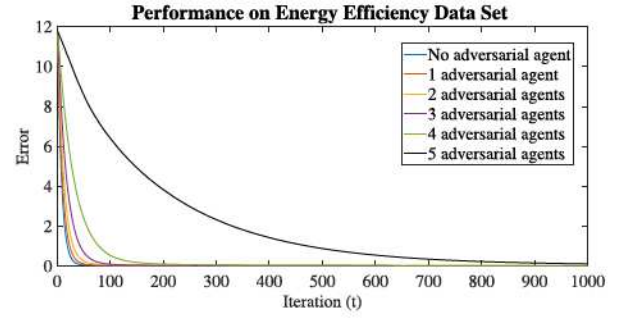


Fig. 1. Plot of error decay versus time for different number of Byzantine agents.

performance. This observation is in consistent with the result of Theorem 3, because by increasing b , the value of γ decreases, slowing down the convergence.

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: 1) Byzantine agents send out arbitrary messages to other agents; 2) good agents need to deal with noisy measurements; and 3) 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 agents' observations can be a non-linear 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. For example, Assumptions 1 and 2 are imposed in terms of ℓ_1 norm. It would be interesting to know whether it is possible to replace ℓ_1 norm by ℓ_2 norm.

APPENDIX A 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 Byzantine) does not need to know β .

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

- (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

$$\langle m_{j'i}(t), e_k \rangle \leq \langle m_{ji}(t), e_k \rangle \leq \langle m_{j''i}(t), e_k \rangle. \quad (9)$$

We consider two cases.

- 1) Case 1: $\mathcal{R}_i^k(t) \cap \mathcal{A} = \emptyset$.
- 2) Case 2: $\mathcal{R}_i^k(t) \cap \mathcal{A} \neq \emptyset$.

⁴<https://archive.ics.uci.edu/ml/datasets/Energy+efficiency>

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

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

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

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

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

and

$$|[\mathcal{R}_i^k(t)]^+ / \mathcal{A}| \geq b - |\mathcal{A}|. \quad (11)$$

By (9), 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

$$\begin{aligned} & \frac{1}{n-2b} \sum_{j \in \mathcal{R}_i^k(t)} \langle z_j(t), e_k \rangle \\ & = \frac{\alpha}{|[\mathcal{R}_i^k(t)]^- / \mathcal{A}|} \sum_{j \in [\mathcal{R}_i^k(t)]^- / \mathcal{A}} \langle z_j(t), e_k \rangle \\ & + \frac{1-\alpha}{|[\mathcal{R}_i^k(t)]^+ / \mathcal{A}|} \sum_{j \in [\mathcal{R}_i^k(t)]^+ / \mathcal{A}} \langle z_j(t), e_k \rangle. \quad (12) \end{aligned}$$

Note that

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

where equality (a) follows from (12). 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 the correctness of our proof, because our algorithm does not use these coefficients as its 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 (11) 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.

APPENDIX B 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 for any such fixed ω , $w_j(t, \omega)$ for $t = 1, \dots$ is a standard sequence of vectors. Suppose

$$\lim_{t \rightarrow \infty} \sum_{m=0}^{t-1} \lambda^m \left\| \frac{\sum_{r=1}^{t-m} w_j(r, \omega)}{t-m} \right\|_2 = 0. \quad (13)$$

Then, by the strong law of large numbers, which says that $\mathbb{P}\{\omega \in \Omega : \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{r=1}^t w_j(r, \omega) = 0\} = 1$, we conclude the lemma.

Next, we show (13). It is enough to show that for any $\epsilon > 0$, there exists $t \geq t(\epsilon, \omega)$ such that

$$\sum_{m=0}^{t-1} \lambda^m \left\| \frac{1}{t-m} \sum_{r=1}^{t-m} w_j(r) \right\|_2 \leq \epsilon. \quad (14)$$

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 (14).

APPENDIX C PROOF OF LEMMA 3

Our proof uses McDiarmid's inequality. We first bound $\mathbb{E}[R_j(\lambda, t)]$

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

where inequality (a) follows from Jensen's inequality. Recall that $w_j(r)$ s are i.i.d. with zero mean. For any $j \in \mathcal{V}/\mathcal{A}$, we have

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

Thus, we have

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

We can choose c_r as

$$c_r = 2C \sum_{m=0}^{t-r} \lambda^m \frac{1}{t-m} \quad \forall r = 1, \dots, t.$$

Let $m_0 = \frac{\log \frac{\lambda t}{2}}{\log \frac{1}{\lambda}}$. It is easy to see that $m_0 \leq \frac{t}{2}$ unless t is extremely small. For simplicity, assume that $\frac{\log \frac{\lambda t}{2}}{\log \frac{1}{\lambda}}$ is an integer. We have

$$c_1 = 2C \left(\sum_{m=0}^{m_0} \lambda^m \frac{1}{t-m} + \sum_{m=m_0+1}^{t-1} \lambda^m \frac{1}{t-m} \right) \leq \frac{8C}{(1-\lambda)t}.$$

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

$$\sum_{r=1}^t c_r^2 \leq t c_1^2 \leq \left(\frac{8C}{1-\lambda} \right)^2 \frac{1}{t}.$$

By McDiarmid's inequality, we have

$$\begin{aligned} \mathbb{P} \left\{ R_j(\lambda, t) \geq \sqrt{\text{trace}(\Sigma_j)} \sum_{m=1}^{t-1} \lambda^m \frac{1}{\sqrt{t-m}} + \epsilon \right\} \\ \leq \exp \left(\frac{-2\epsilon^2}{\sum_{r=1}^t c_r^2} \right) \leq \exp \left(\frac{-\epsilon^2(1-\lambda)^2 t}{32 C^2} \right). \end{aligned}$$

APPENDIX D 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

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

It follows from Lemma 1 that

$$\alpha_i^k(t) = \sum_{j \in \mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) \langle z_j(t), e_k \rangle - \langle \theta^*, e_k \rangle. \quad (15)$$

Recall from (4) and (5) that

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

Thus, we have

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

By Lemma 1, we have

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

For the second term, we have

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

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

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

By Assumption 1, we have

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

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

APPENDIX E PROOF OF THEOREM 3

We first show that the evolutions of $\|x_i(t) - \theta^*\|_\infty$ for all $i \in \mathcal{V}/\mathcal{A}$ —their ℓ_∞ norm of the estimation errors—collectively have a matrix representation. With this representation, to show the convergence of $\|x_i(t) - \theta^*\|_\infty$, it is enough to focus on the convergence of the obtained matrix product.

For ease of exposition, let

$$\mathcal{N}_i := (\mathcal{N}_i \cup \{i\}) \setminus \mathcal{A}.$$

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

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

For the second term, we have

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

Recalling that $0 \leq \beta_{ij}^k(t) \leq 1$ and $C_0 = \max_{j \in \mathcal{V}/\mathcal{A}} \|H_j\|_2$, for the first term, we have

$$\begin{aligned}
& \left| \sum_{j \in \mathcal{N}_i} \beta_{ij}^k(t) \left\langle H_j^\top \frac{1}{t} \sum_{r=1}^t w_j(r), e_k \right\rangle \right| \\
&\leq \sum_{j \in \mathcal{N}_i} \beta_{ij}^k(t) \left| \left\langle H_j^\top \frac{1}{t} \sum_{r=1}^t w_j(r), e_k \right\rangle \right| \\
&\leq \sum_{j \in \mathcal{N}_i} \beta_{ij}^k(t) \left\| H_j^\top \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2
\end{aligned}$$

$$\begin{aligned}
&\leq C_0 \sum_{j \in \mathcal{N}_i} \beta_{ij}^k(t) \left\| \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2 \\
&\leq C_0 \max_{j \in \mathcal{V}/\mathcal{A}} \left\| \frac{1}{t} \sum_{r=1}^t w_j(r) \right\|_2.
\end{aligned}$$

Thus, we get

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

Let $E(t) \in \mathbb{R}^\phi$ be the vector that stacks the $\|x_i(t) - \theta^*\|_\infty$ for all $i \in \mathcal{V}/\mathcal{A}$, with

$$E_i(t) = \|x_i(t) - \theta^*\|_\infty.$$

Define matrix $M(t)$ as follows: for each row $i \in \mathcal{V}/\mathcal{A}$, we have

$$M_{i,j}(t) := \beta_{i,j}^{k_i^*(t)} \|(\mathbf{I} - H_j^\top H_j) e_{k_i^*(t)}\|_1 \quad (16)$$

where $k_i^*(t)$ maximizes (over $k = 1, \dots, d$)

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

Notably, $k_i^*(t)$ might not be unique. In that case, $k_i^*(t)$ is an arbitrary such maximizer. We have

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

where the product

$$\prod_{r=1}^t M(r) = M(t)M(t-1) \cdots M(1)$$

is a backward product. Note that $M(t)$ is random, and its realization is determined by both the noises of the good agents' local observations and the Byzantine agents' adversarial behaviors. Nevertheless, 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* substochastic matrix. That is, there exists a row, say i_0 , such that

$$\sum_{j \in \mathcal{V}/\mathcal{A}} M_{i_0,j}(t) < 1.$$

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 article.

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, we have

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

In addition, we have

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

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

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

$$+ C_0 \sum_{j \in \mathcal{V}/\mathcal{A}} R_j(\gamma^{\frac{1}{\xi\phi}}, t).$$

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

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

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

APPENDIX F PROOF OF CLAIM 1

Recall from (16) that $M(t)$ (for each $t \geq 1$) is defined as

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

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 $\tilde{M}(t)$, as follows:

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

By Lemma 4, $\tilde{M}(t)$ 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{G} such that

$$\tilde{M}(t) \geq \frac{1}{2(\phi - b)} \tilde{G}(t) \quad (17)$$

where $\tilde{G}(t)$ is the adjacency matrix of the corresponding reduced graph. For ease of exposition, with a little abuse of notation, we use $\tilde{G}(t)$ to denote both the adjacency matrix and the reduced graph.⁵ We refer to $\tilde{G}(t)$ 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 $|\mathcal{G}| = \xi$, there exists at least one reduced graph in \mathcal{G} that appears at least ϕ times in the sequence of the shadow graphs. Let \tilde{G}^* be one such reduced graph. Without loss of generality, let i_0 be the node in the unique source component of \tilde{G}^* such that

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

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

For any given realization of $M(t_0 + 1), \dots, M(t_0 + \xi\phi)$, let τ_1, \dots, τ_ϕ be the first ϕ time indices at which \tilde{G}^* 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 an arbitrary t_0 . Let

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

⁵Its meaning should be clear from the context.

with $\eta_i(t)$ being the i th entry of $\eta(t)$. Note that $\eta(t) \leq 1$ as $M(r)$ is substochastic for all r .

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

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

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

B) If i is an outgoing neighbor of i_0 in the shadow graph \tilde{G}^* , then for any $j = 2, \dots, \phi$, we have

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

C) For any $j = 3, \dots, \phi$, if i_0 can reach node i in the shadow graph \tilde{G}^* 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)–(C) hold. Recall that i_0 is in the unique source component of \tilde{G}^* . At time τ_ϕ , at all $i \in \mathcal{V} \setminus \mathcal{A}$, it holds that

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

where the last inequality follows from the fact that

$$2 + \sum_{j'=3}^\phi \Delta_{j'} \leq \tau_1 + \Delta_2 + \sum_{j'=3}^\phi \Delta_{j'} = \tau_\phi \leq \xi\phi.$$

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)–(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, we have

$$\begin{aligned} \eta_{i_0}(\tau_j) &\leq \sum_{i \in \mathcal{V} \setminus \mathcal{A}} M_{i_0 i}(\tau_j) \\ &= \sum_{i \in \mathcal{V} \setminus \mathcal{A}} \beta_{i_0 i}^{k_{i_0}^*(\tau_j)} \left\| (I - H_i^\top H_i) e_{k_{i_0}^*(\tau_j)} \right\|_1 \end{aligned}$$

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

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

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

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

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

$$\begin{aligned} \eta(\tau_j) &= M(\tau_j) \eta(\tau_j - 1) \\ &= \sum_{i' \in \mathcal{V} \setminus \mathcal{A}} M_{i i'}(\tau_j) \eta_{i'}(\tau_j - 1). \end{aligned} \quad (18)$$

Recall from (16) that

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

We consider two cases.

1) $\left\| (I - H_{i_0}^\top H_{i_0}) e_{k_i^*(\tau_j)} \right\|_1 < 1$.

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

1) Suppose that $\left\| (I - H_{i_0}^\top H_{i_0}) e_{k_i^*(\tau_j)} \right\|_1 < 1$. Since $\tilde{G}_{i i_0}^* = 1$, it follows by (17) that

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

Recall the definition of ρ_0 in (8). We have by (18) and $0 < \eta_{i'}(\tau_j - 1) \leq 1$ that

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

2) Suppose that $\left\| (I - H_{i_0}^\top H_{i_0}) e_{k_i^*(\tau_j)} \right\|_1 = 1$. In this case, we have

$$M_{i i_0}(\tau_j) = \tilde{M}_{i i_0}(\tau_j) \geq \frac{1}{2(\phi - b)}.$$

Thus, we have from (18) that

$$\begin{aligned} \eta_i(\tau_j) &= M_{i i_0}(\tau_j) \eta_{i_0}(\tau_j - 1) \\ &+ \sum_{i' \in \mathcal{V} \setminus \mathcal{A} \& i' \neq i_0} M_{i i'}(\tau_j) \eta_{i'}(\tau_j - 1) \end{aligned}$$

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

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

Base case. $j = 3$: Let i be a second-order neighbor of node i_0 in the shadow graph \tilde{G}^* , i.e., there exists a directed path of length 2 such that $i_0 \rightarrow i_1 \rightarrow i$ in \tilde{G}^* .

If $\|(I - H_{i_1}^\top H_{i_1})e_{k_{i_1}^*(\tau_3)}\|_1 < 1$, 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 $\|(I - H_{i_1}^\top H_{i_1})e_{k_{i_1}^*(\tau_3)}\|_1 = 1$.

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

$$\|(I - H_{i_1}^\top H_{i_1})e_{k_{i_1}^*(r)}\|_1 < 1$$

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

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

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

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

As $\|(I - H_{i_1}^\top H_{i_1})e_{k_{i_1}^*(\tau_3)}\|_1 = 1$ and $\beta_{i_{i_0}}^{i^*(\tau_3)} \geq \frac{1}{2(\phi-b)}$, 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

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

i.e., $M_{i_{i_1}}(r) = \tilde{M}_{i_{i_1}}(r)$ 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

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

and

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

Induction step: Suppose that 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 neighbors of node i_0 in the shadow graph \tilde{G}^* , where $h = 2, \dots, j - 1$.

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

REFERENCES

- [1] A. Speranzon, C. Fischione, and K. H. Johansson, "Distributed and collaborative estimation over wireless sensor networks," in *Proc. IEEE Conf. Decis. Control*, 2006, pp. 1025–1030.
- [2] L. Xie, D.-H. Choi, S. Kar, and H. V. Poor, "Fully distributed state estimation for wide-area monitoring systems," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1154–1169, Sep. 2012.
- [3] B. Sinopoli, C. Sharp, L. Schenato, S. Schaffert, and S. S. Sastry, "Distributed control applications within sensor networks," *Proc. IEEE*, vol. 91, no. 8, pp. 1235–1246, Aug. 2003.
- [4] R. Olfati-Saber, "Distributed Kalman filtering for sensor networks," in *Proc. IEEE Conf. Decis. Control*, 2007, pp. 5492–5498.
- [5] S. Kar, J. M. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3575–3605, Jun. 2012.
- [6] F. Bullo, J. Cortes, and S. Martinez, *Distributed Control of Robotic Networks: A Mathematical Approach to Motion Coordination Algorithms*. Princeton, NJ, USA: Princeton Univ. Press, 2009, vol. 27.
- [7] Y. Chen, S. Kar, and J. M. Moura, "The Internet of Things: Secure distributed inference," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 64–75, Sep. 2018.
- [8] S. S. Stankovic, M. S. Stankovic, and D. M. Stipanovic, "Decentralized parameter estimation by consensus based stochastic approximation," *IEEE Trans. Autom. Control*, vol. 56, no. 3, pp. 531–543, Mar. 2011.
- [9] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," in *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 2, 2017, Art no. 44.
- [10] S. Kar and J. M. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, Jan. 2009.
- [11] N. A. Lynch, *Distributed Algorithms*. San Francisco, CA, USA: Morgan Kaufmann, 1996.
- [12] M. Pease, R. Shostak, and L. Lamport, "Reaching agreement in the presence of faults," *J. ACM*, vol. 27, no. 2, pp. 228–234, 1980.

- [13] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 3, pp. 382–401, 1982.
- [14] H. Mendes and M. Herlihy, "Multidimensional approximate agreement in Byzantine asynchronous systems," in *Proc. 46th Annu. ACM Symp. Theory Comput.*, 2013, pp. 391–400.
- [15] N. H. Vaidya and V. K. Garg, "Byzantine vector consensus in complete graphs," in *Proc. ACM Symp. Princ. Distrib. Comput.*, 2013, pp. 65–73.
- [16] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 645–658, Dec. 2011.
- [17] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Trans. Smart Grid*, vol. 2, no. 2, pp. 326–333, Jun. 2011.
- [18] K. C. Sou, H. Sandberg, and K. H. Johansson, "On the exact solution to a smart grid cyber-security analysis problem," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 856–865, Jun. 2013.
- [19] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.
- [20] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1454–1467, Jun. 2014.
- [21] J. Mattingley and S. Boyd, "Real-time convex optimization in signal processing," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 50–61, May 2010.
- [22] Y. Shoukry and P. Tabuada, "Event-triggered state observers for sparse sensor noise/attacks," *IEEE Trans. Autom. Control*, vol. 61, no. 8, pp. 2079–2091, Aug. 2016.
- [23] A. Chattopadhyay and U. Mitra, "Security against false data injection attack in cyber-physical systems," *IEEE Trans. Control Netw. Syst.*, 2019, *arXiv:1807.11624*. [Online]. Available: <http://arxiv.org/abs/1807.11624>
- [24] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0925231217316351>
- [25] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Trans. Autom. Control*, vol. 56, no. 7, pp. 1495–1508, Jul. 2011.
- [26] Y. Chen, S. Kar, and J. M. Moura, "Resilient distributed estimation through adversary detection," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2455–2469, May 2018.
- [27] Y. Chen, S. Kar, and J. M. Moura, "Attack resilient distributed estimation: A consensus+innovations approach," in *Proc. Annu. Amer. Control Conf.*, 2018, pp. 1015–1020.
- [28] A. Mitra and S. Sundaram, "Byzantine-resilient distributed observers for LTI systems," *Automatica*, vol. 108, 2019, Art. no. 108487.
- [29] W. Xu, Z. Li, and Q. Ling, "Robust decentralized dynamic optimization at presence of malfunctioning agents," *Signal Process.*, vol. 153, pp. 24–33, 2018.
- [30] L. Su and N. H. Vaidya, "Non-Bayesian learning in the presence of Byzantine agents," in *Proc. Int. Symp. Distrib. Comput.*, 2016, pp. 414–427.
- [31] Z. Yang and W. U. Bajwa, "ByRDIE: Byzantine-resilient distributed coordinate descent for decentralized learning," *IEEE Trans. Signal Inf. Process. Netw.*, 2017, *arXiv:1708.08155*.
- [32] Y. Chen, S. Kar, and J. M. Moura, "Resilient distributed estimation: Sensor attacks," *IEEE Trans. Autom. Control*, vol. 64, no. 9, pp. 3772–3779, Sep. 2019.
- [33] Y. Chen, S. Kar, and J. M. Moura, "Resilient distributed estimation: Exponential convergence under sensor attacks," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 7275–7282.
- [34] Y. Chen, S. Kar, and J. M. Moura, "Topology free resilient distributed estimation," 2018, *arXiv:1812.08902*.
- [35] L. Su and N. H. Vaidya, "Fault-tolerant multi-agent optimization: Optimal iterative distributed algorithms," in *Proc. ACM Symp. Princ. Distrib. Comput.*, 2016, pp. 425–434.
- [36] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Trans. Autom. Control*, vol. 63, no. 3, pp. 714–725, Mar. 2018.
- [37] N. Gupta and N. H. Vaidya, "Byzantine fault tolerant distributed linear regression," 2019, *arXiv:1903.08752*. [Online]. Available: <http://arxiv.org/abs/1903.08752>
- [38] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Int. Conf. Machine Learning*, 2018, pp. 5636–5645, *arXiv:1803.01498*.
- [39] C. Xie, O. Koyejo, and I. Gupta, "Phocas: Dimensional byzantine-resilient stochastic gradient descent," 2018, *arXiv:1805.09682*.
- [40] N. H. Vaidya, L. Tseng, and G. Liang, "Iterative approximate byzantine consensus in arbitrary directed graphs," in *Proc. ACM Symp. Princ. Distrib. Comput.*, 2012, pp. 365–374.
- [41] N. Vaidya, "Matrix representation of iterative approximate byzantine consensus in directed graphs," 2012, *arXiv:1203.1888*.
- [42] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy Buildings*, vol. 49, pp. 560–567, 2012.



Lili Su received the B.S. degree from Nankai University, Tianjin, China, in 2011, and the Ph.D. degree (under the supervision of Prof. N. H. Vaidya) from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2017.

She is currently a Postdoctoral Researcher (hosted by Prof. N. Lynch) with the Computer Science and Artificial Intelligence Laboratory, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. Her master's research was focused on ordinal data processing at CSL Communication Group from August 2012 to May 2014. Her research interests include distributed systems, brain computing, security, optimization, and learning.

Dr. Su was one of the three nominees for the 2016 International Symposium on distributed Computing Best Student Paper Award. She received the 2015 International Symposium on Stabilization Safety and Security of Distributed Systems Best Student Paper Award. She also received the Sundaram Seshu International Student Fellowship for the academic year 2016–2017, conferred by UIUC. She was on the Program Committee for the 2018 IEEE International Conference on Distributed Computing Systems, SCNSD 2018, and Workshop on Storage, Control, Networking in Dynamic Systems (SCNDS) 2017.



Shahin Shahrampour received the B.Sc. degree from the Sharif University of Technology, Tehran, Iran, in 2009, and the M.S.E. degree in electrical engineering, the M.A. degree in Statistics, and the Ph.D. degree in electrical and systems engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2012, 2014, and 2015, respectively.

He is currently an Assistant Professor with the Department of Industrial and Systems Engineering and the Department of Electrical and Computer Engineering (by courtesy), Texas A&M University (TAMU), College Station, TX, USA. Before joining TAMU, he was a Postdoctoral Fellow with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. His research interests include machine learning, optimization, sequential decision making, and distributed learning, with a focus on developing computationally efficient methods for data analytics.