Resilient Distributed Optimization for Multi-Agent Cyberphysical Systems

Michal Yemini, Angelia Nedić, Andrea J. Goldsmith, Stephanie Gil

***Abstract*—Enhancing resilience in distributed networks in the face of malicious agents is an important problem for which many key theoretical results and applications require further development and characterization. This work focuses on the problem of distributed optimization in multi-agent cyberphysical systems, where a legitimate agent's dynamic is influenced both by the values it receives from potentially malicious neighboring agents, and by its own self-serving target function. We develop a new algorithmic and analytical framework to achieve resilience for the class of problems where stochastic values of trust between agents exist and can be exploited. In this case we show that convergence to the true global optimal point can be recovered, both in mean and almost surely, even in the presence of malicious agents. Furthermore, we provide expected convergence rate guarantees in the form of upper bounds on the expected squared distance to the optimal value. Finally, we present numerical results that validate the analytical convergence guarantees we present in this paper even when the malicious agents compose the majority of agents in the network.**

***Index Terms*—Distributed optimization, resilience, malicious agents, Byzantine agents, stochastic trust values, cyberphysical systems.**

I. Introduction

Distributed optimization is at the core of various multi-agent tasks including distributed control and estimation, multi-robot tasks such as mapping, and many learning tasks such as Federated Learning [2]–[4]. Owing to a long history and much attention in the research community, the theory for distributed optimization has matured, leading to several important results provide rigorous performance guarantees in the form of convergence and convergence rate for different function types, underlying graph topologies, and noise [5]–[9]. However, in the presence of malicious activity many of these known results are invalidated, requiring a new characterization of performance and equivalent theory for the adversarial case.

With the growing prevalence of multi-agent and cyberphysical systems, and their reliance on distributed optimization methods for correct functioning in the real world, it becomes critical that the vulnerability of these methods is well understood. In particular, malicious agents can greatly interfere with the result of a distributed optimization scheme, driving the convergence to a non-optimal solution or preventing convergence altogether. They can accomplish this by either not sharing key information or by manipulating key information such as the shared gradients, which are critical for the correct functioning of the distributed optimization scheme [10]–[12]. Note that while well-established stochastic optimization methods characterize the effect of noise in distributed multi-agent systems [13], [14], malicious agents have the ability to inject intentionally biased or manipulated information which can lead to a greater potential damage for these systems. As a result, recent works have increasingly turned attention to the investigation of robust and resilient versions of distributed optimization methods in the face of malicious intent and/or severe (potentially biased) noise [10]–[12], [15], [16]. These approaches can be coarsely divided into two categories, those that use the transmitted data between nodes to infer the presence of anomalies (for example see [11], [17]), and those that exploit additional side information from the network or the physicality of the underlying cyberphysical system to provide additional channels of resilience [18]–[20].

We are interested in investigating the class of problems where the physicality of the system plays an important role in achieving new possibilities of resilience for these systems. Indeed the physicality of cyberphysical systems has been shown to provide many new channels of verification and establishing *inter-agent trust* through watermarking [21], wireless signal characteristics [20], [22], side information [23], and camera or lidar data cross-validation [24]. By exploiting these physical-based measurements, agents can extract additional information about the trustworthiness of their neighbors.

We capitalize on this observation which motivates us to focus on a class of problems where the existence of this additional information in the system can be exploited to arrive at much stronger performance results in the adversarial case – what we refer to as *resilience*. We abstract this information as a value α_{ij} that indicates the likelihood with which an agent i can trust data received from another agent j. We show that under mild assumptions, when this information is available, several powerful results for distributed optimization can be recovered such as 1) ***convergence to the true optimal point*** in the case of minimizing the sum of strongly convex functions, and 2) ***characterization of convergence rate*** that depends on the network topology, the amount of trust observations acquired, and the number of legitimate and malicious agents in the system.

M. Yemini is with the Faculty of Engineering, Bar-Ilan University, Israel (michal.yemini@biu.ac.il). A. Nedić is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (Angelia.Nedich@asu.edu). A. J. Goldsmith is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (goldsmith@princeton.edu). S. Gil is with the Computer Science Department, School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02139 USA (sgil@seas.harvard.edu).

This work was partially supported by NSF awards #CNS-2147641 and #CNS-2147694. M. Yemini and A. J. Goldsmith were partially supported by AFOSR award #002484665.

Preliminary results were accepted to the IEEE Conf. Decis. Control, [1].

A. Related Work

In the absence of malicious agents, the legitimate agents can construct iterates converging to an optimal point $x^\star_{\mathcal{L}}$ by using either their gradients, or sub-gradients when their objective functions are not differentiable. Each agent i updates its data value by considering the data values of its neighbors, and its self-serving gradient direction of its objective function f_i or the directions obtained from its neighbors. Convergence to an optimal point $x^\star_{\mathcal{L}}$ can be achieved for constrained multi-agent problems in [5], [7], [14], [25]–[30] and with limited gradient information [31], [32]. Additionally, a zero-order method has been proposed in [33]. Some works, such as [5], assume that the weight matrices, which dictate how agents incorporate the data they receive from their neighbors, are doubly-stochastic. However, works such as [28] overcome this assumption by performing additional weighted averaging steps. Finally, it has been established that the convergence rate of distributed gradient algorithms with diminishing step size is at best $O(\frac{1}{T})$ where T is the algorithm running time, see for example [30].

To harm the system, a malicious agent can send falsified data to their legitimate neighbors. If the legitimate agents are unaware of their malicious neighbors, then the malicious agents will succeed in controlling the system [10], [11], [34]–[37]. To combat the harmful effect of an attack, the approach taken in [34]–[36] requires the pre-existence of a set of trusted agents such that all other agents (legitimate or malicious) are connected to at least one trusted agent. Nonetheless, this approach is unrealistic when communication is sporadic such as in robotic and ad-hoc networks. The approaches in [10], [11], [37], [38] rely on the agent data values to detect and discard malicious inputs and have an upper bound on the number of tolerable malicious agents (with a star network having the largest number - a half of the number of agents in the network) [37]. When the number of malicious agents exceeds the tolerable number, the attack succeeds and malicious agents evade detection. In contrast with the existing works, our proposed method provides a significantly stronger resilience to malicious activity by exploiting the physical aspect of the problem, i.e., the wireless medium. Thus, each legitimate agent can learn trustworthy neighbors while optimizing the system objective. Our prior work [39] studies the implications of the agents' learning ability, with regards to the trustworthiness of their neighbors, on distributed *consensus* systems. This work considers the more general case of distributed *optimization* systems where the agent's goal is to minimize the sum of their local objective function under limited information exchange.

Finally, this work also relates to stochastic optimization, see for example [40]–[47]. However, unlike the typical assumption that the stochastic gradients are unbiased and statistically independent of the weights. In this work, the stochastic gradients are biased, where the bias occurs due to the adversarial inputs of the malicious agents. Furthermore, learning of the trustworthiness of neighboring agents and adjusting agents' weight accordingly, lead to a correlation between the agents' values and the weights that are assigned to them. To this end, our analysis could not rely on previous results when analyzing the rate of convergence of the agents' dynamic.

B. Paper Organization

The rest of this paper is organized as follows: Section II presents the system model and problem formulation. Section III proposes our algorithm for resilient distributed optimization. Section IV presents our learning mechanism for detecting malicious agents and provides upper bounds on the probability of misclassifying malicious and legitimate agents. Sections V and VI, respectively, present asymptotic and finite time regime convergence results. Finally, Section VII presents numerical results that validate our analytical results, and Section VIII concludes the paper.

II. Problem Formulation

We consider a multi-agent system of n agents communicating over a network, which is represented by an undirected graph, $\mathbb{G} = (\mathbb{V}, \mathbb{E})$. The node set $\mathbb{V} = \{1, \ldots, n\}$ represents the agents and the edge set $\mathbb{E} \subset \mathbb{V} \times \mathbb{V}$ represents the set of communication links, with $\{i, j\} \in \mathbb{E}$ indicating that agents i and j are connected. We denote by $\mathcal{N}_i$ the set of neighbors of agent i, that is $\mathcal{N}_i \triangleq \{j : \{i, j\} \in \mathbb{E}\}$.

We study the case where an unknown subset of the agents is malicious and the trustworthy agents are learning which neighbors they can trust. Thus, $\mathbb{V} = \mathcal{L} \cup \mathcal{M}$ where $\mathcal{L}$ is the set of legitimate agents that execute computational tasks and share their data truthfully, while $\mathcal{M}$ denotes the set of agents that are not truthful. *The sets $\mathcal{L}$ and $\mathcal{M}$ are just modeling artifacts, and none of the legitimate agents knows if it has malicious neighbors or not, at any time.* Throughout the paper, we will use the subscripts $\mathcal{L}$ and $\mathcal{M}$ to denote the various quantities related to legitimate and malicious agents, respectively.

We are interested in a general distributed optimization problem, where the legitimate agents aim at optimizing a common objective whereas the malicious agents try to impair the legitimate agents by malicious injections of harmful data. The aim of the legitimate agents is to minimize distributively the sum of their objective functions over a constraint set $\mathcal{X} \subset \mathbb{R}^d$, i.e.,

$$x^\star_{\mathcal{L}} = \arg\min_{x \in \mathcal{X}} f_{\mathcal{L}}(x), \;\text{ with }\; f_{\mathcal{L}}(x) = \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} f_i(x). \tag{1}$$

By choosing a local update rule and exchanging some information with their neighbors, the legitimate agents want to determine the optimal solution $x^\star_{\mathcal{L}}$ in (1). In contrast, the malicious agents aim to either lead the legitimate agents to a common non-optimal value $x \in \mathcal{X}$ such that $f_{\mathcal{L}}(x) > f_{\mathcal{L}}(x^\star)$, or prevent the convergence of an optimization method employed by the legitimate agents.

A. Notation

We let x^T denote the transpose of x, where $x \in \mathbb{R}^d$. We denote by $\|x\| \triangleq \sqrt{x^T x}$ the ℓ_2 vector norm. We let $\Pi_{\mathcal{X}}(x)$ be the projection of x onto the set $\mathcal{X}$, i.e.,

$$\Pi_{\mathcal{X}}(x) = \text{argmin}_{y \in \mathcal{X}} \|y - x\|.$$

Finally, we denote by $\mathbf{E}[\cdot]$ the expectation operator.

B. Trust values

We employ a probabilistic framework of trustworthiness where we assume the availability of stochastic *observations of trust* between communicating agents. This information is abstracted in the form of a random variable α_{ij} defined below.

Definition II.1 (α_{ij})**.** For every $i \in \mathcal{L}$ and $j \in \mathcal{N}_i$, the random variable $\alpha_{ij} \in [0,1]$ represents the probability that agent j is a trustworthy neighbor of agent i. We assume the availability of such observations $\alpha_{ij}(t)$ at every instant of time $t \geq 0$ throughout the paper.

This model of inter-agent trust observations has been used in prior works [22], [39]. The focus of the current work is not on the derivation of the values α_{ij} themselves, but rather on the derivation of a theoretical framework for achieving resilient distributed optimization using this model. Indeed, we show that much stronger results of convergence are achievable by properly exploiting this information in the network. We refer to [22] for an example of such a value α_{ij}. Intuitively, a random realization $\alpha_{ij}(t)$ of α_{ij} contains useful trust information regarding the legitimacy of a transmission. We assume that a value of $\alpha_{ij}(t) > 0.5$ indicates a legitimate transmission and $\alpha_{ij}(t) < 0.5$ indicates a malicious transmission in a stochastic sense (misclassifications are possible). Note that $\alpha_{ij}(t) = 0.5$ means that the observation is completely ambiguous and contains no useful trust information for the transmission at time t.

We use the following assumptions throughout the paper:

Assumption 1. *(i)* [Sufficiently connected graph] *The subgraph $\mathbb{G}_{\mathcal{L}}$ induced by the legitimate agents is connected.*
(ii) [Homogeneity of trust variables] *There are scalars $E_{\mathcal{L}} > 0$ and $E_{\mathcal{M}} < 0$ such that*

$$E_{\mathcal{L}} \triangleq \mathbf{E}[\alpha_{ij}(t)] - 0.5, \qquad \text{for all } i \in \mathcal{L},\ j \in \mathcal{N}_i \cap \mathcal{L},$$

$$E_{\mathcal{M}} \triangleq \mathbf{E}[\alpha_{ij}(t)] - 0.5, \qquad \text{for all } i \in \mathcal{L},\ j \in \mathcal{N}_i \cap \mathcal{M}.$$

(iii) [Independence of trust observations] *The observations $\alpha_{ij}(t)$ are independent for all t and all pairs of agents i and j, with $i \in \mathcal{L}$, $j \in \mathcal{N}_i$. Moreover, for any $i \in \mathcal{L}$ and $j \in \mathcal{N}_i$, the observation sequence $\{\alpha_{ij}(t)\}$ is identically distributed.*

We note that these are standard assumptions when using the probabilistic trust framework employed here [22], [39].

C. The update rule of agents

We propose distributed methods with significantly stronger resilience compared to [11]. This is enabled by each legitimate agent learning which neighbors it can trust while optimizing a system objective.

a) The update rule of legitimate agents: Each legitimate agent i updates $x_i(t)$ by considering the values $x_j(t)$ of its neighbors and *the gradient of its own objective function f_i* similarly to the iterates described in [5][1]. This method takes the following form for every legitimate agent $i \in \mathcal{L}$,

$$c_i(t) = w_{ii}(t)x_i(t) + \sum_{j \in \mathcal{N}_i} w_{ij}(t)x_j(t),$$

$$y_i(t) = c_i(t) - \gamma(t)\nabla f_i(c_i(t)),$$
$$x_i(t+1) = \Pi_{\mathcal{X}}\left(y_i(t)\right), \tag{2}$$

where $\gamma(t) \geq 0$ is a stepsize that is common to all agents $i \in \mathcal{L}$ at each time t, and $\mathcal{N}_i$ is the set of neighbors of agent i in the communication graph. The set $\mathcal{N}_i$ is composed of both legitimate and malicious neighbors of agent $i \in \mathcal{L}$, while the weights $w_{ij}(t), j \in \mathcal{N}_i \cup \{i\}$, are nonnegative and sum to 1. For each $i \in \mathcal{L}, j \in \mathcal{L} \cup \mathcal{M}$, the choice of $w_{ij}(t)$ depends on the history of the random trust observations $(\alpha_{ij}(\tau))_{0 \leq \tau \leq t}$. As a result, the weights $w_{ij}(t)$ and the data points $x_i(t)$ are random, for every $i \in \mathcal{L}, j \in \mathcal{L} \cup \mathcal{M}$ and time instant $t > 0$.

To reduce the convergence rate, we allow $\gamma(t)$ and $w_{ij}(t)$ to depend on a parameter $T_0 \geq 0$ which dictates how many trust observations a legitimate agent collects before it decided to trust one of its neighbors. We specify $\gamma(t)$ and $w_{ij}(t)$, $i \in \mathcal{L}$, $j \in \mathcal{N}_i \cup \{i\}$, precisely later on in Section III.

b) The update rule for the malicious agents: Malicious agents $i \in \mathcal{M}$ choose values arbitrarily in the set $\mathcal{X}$. We assume that their actions are not known, and thus we do not model them. For simplicity of exposition, the dynamic (2) captures malicious inputs, where an adversarial agent $i \in \mathcal{M}$ sends all its legitimate neighbors identical copies of its chosen input $x_i(t)$ at time t. Let us denote by $x_{ij}(t)$ the input of a malicious agent i to a legitimate agent j at time t, then $x_{ij_1}(t) = x_{ij_2}(t)$ for every $j_1, j_2 \in \mathcal{L}$. Nonetheless, our analytical results also hold for byzantine inputs where an adversarial agent $i \in \mathcal{M}$ can send its legitimate neighbors different inputs at time t. In this case $x_{ij_1}(t)$ need not be equal to $x_{ij_2}(t)$ for every $j_1, j_2 \in \mathcal{L}$.

D. Assumptions on the objective functions and initial points

Assumption 2. *We assume that $\mathcal{X} \subset \mathbb{R}^d$ is compact and convex and that there exists a known value $\eta > 0$ such that*

$$\|x\| \leq \eta, \quad \forall x \in \mathcal{X}. \tag{3}$$

The η value in Assumption 2 is arbitrary, and its role is to bound the malicious agents' inputs away from infinity.

Assumption 3. *For all legitimate agents $i \in \mathcal{L}$, the function f_i is μ-strongly convex and has L-Lipschitz continuous gradients, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$.*

Note that under Assumption 2 and the strong convexity of Assumption 3, the problem (1) has a unique solution $x^*_{\mathcal{L}} \in \mathcal{X}$.

Assumption 4. *Let the stepsize sequence $\{\gamma(t)\}$ be nonnegative, monotonically nonincreasing, and such that $\sum_{t=0}^{\infty}\gamma(t) = \infty$ and $\sum_{t=0}^{\infty}\gamma^2(t) < \infty$.*

E. Objectives

The objective of this work is to arrive at strong convergence results for the distributed optimization problem in (1) in the presence of malicious agents $\mathcal{M}$. We wish to achieve this by carefully exploiting the availability of stochastic trust values $\alpha_{ij}(t)$ in the network. Specifically, we aim to achieve the following:

[1]Note, however, that [5] does not include a projection on the set $\mathcal{X}$.

Objective 1: We wish to construct weight sequences $\{w_{ij}(t)\}$, $i \in \mathcal{L}$, $j \in \mathcal{N}_i$ in the method (2) to weight the influence of neighboring nodes in each legitimate agent's update. Specifically, we wish to construct these sequences such that they converge over time to some *nominal weights* $\overline{w}_{ij}, i \in \mathcal{L},\ j \in \mathcal{N}_i$, almost surely (a.s.), where $\overline{w}_{ij} = 0$ for all malicious neighbors $j \in \mathcal{N}_i \cap \mathcal{M}$ of agent $i \in \mathcal{L}$.
Objective 2: Utilizing the proposed weights $\{w_{ij}(t)\}$, we aim to show that the iterates given by (2) converge (in some sense) to the true optimal point $x^\star_{\mathcal{L}} \in \mathcal{X}$ under Assumptions 1-4.
Objective 3: We aim to establish an upper bound on the expected value of $\|x_i(t) - x^\star_{\mathcal{L}}\|^2$, for all $i \in \mathcal{L}$, as a function of the time t, for the iterates $x_i(t)$ produced by the method.

III. The Algorithm

Next, we present an algorithm that incorporates the legitimate agents' learning of inter-agent trust values into the dynamic (2) through the choice of the time-dependent weights $w_{ij}(t)$. These weights depend on a parameter T_0 that captures the number of trust measurements a legitimate agent collects before deciding if to trust one of its neighbors. We utilize the parameter T_0 to enable faster convergence rates of the algorithm Nonetheless, as we show in Section V, the algorithm converges to the optimal point $x^\star_{\mathcal{L}}$ for any choice of nonnegative integer T_0, including the special case where $T_0 = 0$. In this case, legitimate agents have no prior trust observations to rely on when they first decide whether to trust their neighbors.

A. The weight matrix sequence

Consider the sum over a history of $\alpha_{ij}(t)$ values that we denote by $\beta_{ij}(t)$:

$$\beta_{ij}(t) = \sum_{k=0}^{t-1} (\alpha_{ij}(k) - 0.5) \ \text{ for } t \geq 1, i \in \mathcal{L}, j \in \mathcal{N}_i, \quad (4)$$

and define $\beta_{ij}(0) = 0$. We note that we explore the probabilistic characteristics of $\beta_{ij}(t)$ in Section IV.

We define a time dependent *trusted neighborhood* for agent $i \in \mathcal{L}$ as:

$$\mathcal{N}_i(t) \triangleq \{j \in \mathcal{N}_i : \ \beta_{ij}(t) \geq 0\}. \quad (5)$$

This is the subset of neighbors that legitimate agent i classifies as its legitimate neighbors at time t. For all $t \geq 0$, let

$$d_i(t) \triangleq |\mathcal{N}_i(t)| + 1 \geq 1 \qquad \text{for all } i \in \mathcal{L}.$$

At each time t, *every agent i sends the value $d_i(t)$ to its neighbors $j \in \mathcal{N}_i$ in addition to the value $x_i(t)$*. Alternatively, we can assume that agent i sends $d_i(t)$ to its neighbors only when the value $d_i(t)$ changes.

Legitimate agents are the most susceptible to making classification errors regarding the trustworthiness of their neighbors when they have a small sample size of trust value observations. Thus, we delay the updating of legitimate agents' values until time $T_0 \geq 0$. Up to time T_0, the legitimate agents only collect observations of trust values.

Let $\mathbb{1}_{\{A\}}$ denote the indicator function; it is equal to one if the event A is true and zero otherwise. We define the weight matrix $W(t)$ by choosing its entries $w_{ij}(t)$ as follows: for every $i \in \mathcal{L}$, $j \in \mathcal{N}_i$,

$$w_{ij}(t) = \begin{cases} \frac{\mathbb{1}_{\{t \geq T_0\}}}{2 \cdot \max\{d_i(t), d_j(t)\}} & \text{if } j \in \mathcal{N}_i(t), \\ 0 & \text{if } j \notin \mathcal{N}_i(t) \cup \{i\}, \\ 1 - \sum_{m \in \mathcal{N}_i} w_{im}(t) & \text{if } j = i. \end{cases} \quad (6)$$

Using the weights (6) and letting the stepsize $\gamma(k) = 0, \forall k < 0$, the dynamic in (2) is equivalent to the following dynamic where agents *only consider the data values received from their trusted neighbors at time t, i.e., $\mathcal{N}_i(t)$,* when computing their own value updates: for all $i \in \mathcal{L}$ and all $t \geq 0$,

$$\begin{aligned} c_i(t) &= w_{ii}(t)x_i(t) + \sum_{j \in \mathcal{N}_i(t) \cap \mathcal{L}} w_{ij}(t)x_j(t) + \sum_{j \in \mathcal{N}_i(t) \cap \mathcal{M}} w_{ij}(t)x_j(t), \\ y_i(t) &= c_i(t) - \gamma(t - T_0)\nabla f_i(c_i(t)), \\ x_i(t+1) &= \Pi_{\mathcal{X}}\left(y_i(t)\right). \end{aligned} \quad (7)$$

We note that though the choice of the parameter T_0 affects the weights $w_{ij}(t), i \in \mathcal{L}, j \in \mathcal{L} \cup \mathcal{M}$ and the terms $c_i(t), y_i(t)$ and $x_i(t),\ i \in \mathcal{L}$, we omit this dependence from these notations for the sake of clarity of exposition.

Furthermore, the dependence of the weights $w_{ij}(t)$ on the trust observation history $\beta_{ij}(t)$ comes in through the choice of time-dependent and random trusted neighborhood $\mathcal{N}_i(t)$ (see (5)). Consequently, some entries of the matrix $W(t)$ are also random, as seen from (6). The gradients $\nabla f_i(c_i(t))$ are stochastic due to the randomness of $c_i(t)$, however, they are not unbiased as typically assumed in stochastic approximation methods, including [48]. Thus, we cannot readily rely on prior analysis for stochastic approximation methods. However, as we show in our subsequent analysis, the variance of $\|\nabla f_i(c_i(t))\|$ decays sufficiently fast and allows convergence to the optimal point even in the presence of malicious agents.

Algorithm 1 The protocol of agent $i \in \mathcal{L}$.

> **Inputs:** T, T_0, $\mathcal{N}_i$, $x_i(0)$, $\nabla f_i(\cdot)$, $\gamma(\cdot)$.
> **Outputs:** $x_i(T)$.
> Set $\beta_{ij}(t) = 0$ for all $j \in \mathcal{N}_i$;
> **for** $t = 0, \ldots, T-1$ **do**
>   Set $\mathcal{N}_i(t) = \{j \in \mathcal{N}_i : \ \beta_{ij}(t) \geq 0\}$;
>   Set $d_i(t) = |\mathcal{N}_i(t)| + 1$;
>   Send $x_i(t)$ and $d_i(t)$ to neighbors;
>   **for** $j \in \mathcal{N}_i$ **do**
>   Receive $x_j(t)$ and $d_j(t)$;
>   Extract $\alpha_{ij}(t)$;
>   Set $\beta_{ij}(t+1) = \sum_{k=0}^{t} (\alpha_{ij}(k) - 0.5)$;
>   Set the weight $w_{ij}(t)$ based on the values of T_0, $\mathcal{N}_i(t)$, $d_i(t)$, and $d_j(t)$ as follows:
> $$w_{ij}(t) = \frac{\mathbb{1}_{\{t \geq T_0\}} \mathbb{1}_{\{j \in \mathcal{N}_i(t)\}}}{2 \max\{d_i(t), d_j(t)\}};$$
>   **end for**
>   Set $w_{ii}(t) = 1 - \sum_{m \in \mathcal{N}_i} w_{im}(t)$;
>   Set $x_i(t+1)$ according to the dynamic (7);
> **end for**

IV. Learning the sets of trusted neighbors

This section establishes key characteristics of the stochastic observations $\alpha_{ij}(t)$ that result from the model described in Section II-B and that we will subsequently use in our analysis of the convergence of the iterates produced by Algo. 1.

Recall that we consider the sum $\beta_{ij}(t)$, defined in (4), over a history of $\alpha_{ij}(t)$ values. Intuitively, following the discussion on α_{ij}'s immediately after Definition II.1, the values $\beta_{ij}(t)$ will tend towards positive values for legitimate agent transmissions $i \in \mathcal{L}$ and $j \in \mathcal{N}_i \cap \mathcal{L}$, and will tend towards negative values for malicious agent transmissions where $i \in \mathcal{L}$ and $j \in \mathcal{N}_i \cap \mathcal{M}$. We restate an important result shown in [39] regarding the exponential decay rate of misclassifications given a sum over the history of stochastic observation values that we will use extensively in the forthcoming analysis.

Lemma 1 (Lemma 2 [39])**.** *Consider the random variables $\beta_{ij}(t)$ as defined in Eq.* (4)*. Then, for every $t \geq 0$ and every $i \in \mathcal{L}$, $j \in \mathcal{N}_i \cap \mathcal{L}$,*

$$\Pr\left(\beta_{ij}(t) < 0\right) \leq \max\{\exp(-2tE_{\mathcal{L}}^2), \mathbb{1}_{\{E_{\mathcal{L}}<0\}}\},$$

while for every $t \geq 0$ and every $i \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{M}$,

$$\Pr\left(\beta_{ij}(t) \geq 0\right) \leq \max\{\exp(-2tE_{\mathcal{M}}^2), \mathbb{1}_{\{E_{\mathcal{M}}>0\}}\}.$$

In other words, the probability of misclassifying malicious agents as legitimate, or vice versa, decays exponentially in the accrued number t of observations.

We can now conclude that there is a random but finite time T_f such that there exists a legitimate agent i which misclassifies the trustworthiness of at least one of its neighbors at time $T_f - 1$, and all the legitimate agents classify the trustworthiness of their neighbors correctly at each time $t \geq T_f$. We refer to the time T_f as the "correct classification time".

Corollary 1. *There exists a random finite time T_f such that*

$$\begin{aligned} &\beta_{ij}(t) \geq 0 \textit{ for all } t \geq T_f \textit{ and all } i \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{L}, \\ &\beta_{ij}(t) < 0 \textit{ for all } t \geq T_f \textit{ and all } i \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{M}, \end{aligned} \tag{8}$$

and there exists $i \in \mathcal{L}$ such that

$$\begin{aligned} &\beta_{ij}(T_f - 1) < 0 \textit{ for some } j \in \mathcal{N}_i \cap \mathcal{L}, \textit{ or}, \\ &\beta_{ij}(T_f - 1) \geq 0 \textit{ for some } j \in \mathcal{N}_i \cap \mathcal{M}. \end{aligned} \tag{9}$$

Proof: It follows directly from [39, Proposition 1]. ∎

Let $|\mathcal{N}_i \cap \mathcal{L}|$ be the number of legitimate neighbors of agent i, and $|\mathcal{N}_i \cap \mathcal{M}|$ be the number of malicious neighbors of agent i. We define by $D_{\mathcal{L}}$ the total number of legitimate neighbors, similarly we define by $D_{\mathcal{M}}$ the total number of malicious neighbors, with respect to the legitimate agents. That is,

$$D_{\mathcal{L}} \triangleq \sum_{i \in \mathcal{L}} |\mathcal{N}_i \cap \mathcal{L}| \quad \text{and} \quad D_{\mathcal{M}} \triangleq \sum_{i \in \mathcal{L}} |\mathcal{N}_i \cap \mathcal{M}|.$$

Additionally, we define the following upper bound on the probability that at least one legitimate agent misclassifies one of its legitimate neighbors as malicious or one of its malicious neighbors as legitimate, when observing k trust values for each of its neighbors

$$p_c(k) \triangleq \mathbb{1}_{\{k \geq 0\}} \left[D_{\mathcal{L}} e^{-2kE_{\mathcal{L}}^2} + D_{\mathcal{M}} e^{-2kE_{\mathcal{M}}^2} \right].$$

Furthermore, we define the following upper bound on the probability that a legitimate agent misclassifies one of its legitimate or malicious neighbors, in one of the times after observing k trust values for each of its neighbors:

$$p_e(k) \triangleq D_{\mathcal{L}} \frac{\exp(-2kE_{\mathcal{L}}^2)}{1 - \exp(-2E_{\mathcal{L}}^2)} + D_{\mathcal{M}} \frac{\exp(-2kE_{\mathcal{M}}^2)}{1 - \exp(-2E_{\mathcal{M}}^2)}.$$

Using these quantities, we obtain some useful bounds on the probabilities of the events $(T_f = k)$ and $(T_f > k-1)$ for any $k \geq 0$, as follows.

Lemma 2. *For every $k \geq 0$*

$$\Pr(T_f = k) \leq \min\{p_c(k-1), 1\}, \textit{ and,} \tag{10}$$

$$\Pr(T_f > k-1) \leq \min\{p_e(k-1), 1\}. \tag{11}$$

We present the proof of Lemma 2 in Appendix A. Note, that (10) and (11) are well defined for $k < 0$, since $p_c(k) = 0$ and $p_c(k) > 1$ for all $k < 0$. Thus $\min\{p_c(k), 1\} = 0$ and $\min\{p_e(k), 1\} = 1$.

V. Asymptotic Convergence to the Optimal Point

This section analyzes the convergence characteristics of Algo. 1 by utilizing the almost surely finite correct classification time T_f and the upper bounds we derive in Lemma 2.

Assumptions 2 and 3 lead to the following conclusion.

Corollary 2. *When $\mathcal{X}$ is compact, Assumption 3 implies that there is a scalar G such that $\|\nabla f_i(x)\| \leq G$, $\forall x \in \mathcal{X}, i \in \mathcal{L}$.*

The following lemma is a direct consequence of [25, Lemma 8]. Nonetheless, for completeness of presentation we provide the proof.

Lemma 3. *Denote $\phi_i(t) \triangleq \Pi_{\mathcal{X}}\left(y_i(t)\right) - c_i(t)$. For every $i \in \mathcal{L}$, $T_0 \geq 0$, and $t \geq T_0$ we have that*

$$\|\phi_i(t)\| \leq \gamma(t - T_0)G,$$

where $G > 0$ is from Corrolary 2.

Proof: Since $x_i(t) \in \mathcal{X}$ for $i \in \mathcal{L} \cup \mathcal{M}$ and $t \geq 0$, and $[w_{ij}(t)]_{i \in \mathcal{L}, j \in \mathcal{L} \cup \mathcal{M}}$ is a row stochastic matrix, by the convexity of the set $\mathcal{X}$, we have that $c_i(t) \in \mathcal{X}$ for all $i \in \mathcal{L}$ and $t \geq 0$. Now, by the standard non-expansiveness property of the projection operator it follows that

$$\begin{aligned} \|\phi_i(t)\| &= \|\Pi_{\mathcal{X}}\left(y_i(t)\right) - \Pi_{\mathcal{X}}\left(c_i(t)\right)\| \\ &\leq \|y_i(t) - c_i(t)\| \\ &= \|c_i(t) - \gamma(t - T_0)\nabla f_i(c_i(t)) - c_i(t)\| \\ &\leq \gamma(t - T_0)\|\nabla f_i(c_i(t))\| \leq \gamma(t - T_0)G. \end{aligned}$$

∎

Let us denote $d_{i,\mathcal{L}} \triangleq |\mathcal{N}_i \cap \mathcal{L}| + 1$. Next, we define the doubly stochastic matrix $\overline{W}_{\mathcal{L}} \in [0,1]^{|\mathcal{L}| \times |\mathcal{L}|}$ with the entries

$[\overline{W}_{\mathcal{L}}]_{i,j}$, for every $i, j \in \mathcal{L}$:

$$[\overline{W}_{\mathcal{L}}]_{i,j} = \begin{cases} \frac{1}{2\cdot\max\{d_{i,\mathcal{L}}, d_{j,\mathcal{L}}\}} & \text{if } j \in \mathcal{N}_i, \\ 0 & \text{if } j \notin \mathcal{N}_i \cup \{i\}, \\ 1 - \sum\limits_{m \in \mathcal{N}_i \cap \mathcal{L}} \overline{w}_{im} & \text{if } j = i. \end{cases} \tag{12}$$

Note that $\overline{W}_{\mathcal{L}}$ is the *nominal* weight matrix, i.e., the value the weight matrix would take in the absence of malicious agents. Let $\sigma_2(A)$ be the second largest singular value of A, and denote $\rho_{\mathcal{L}} = \max_{k\geq 1} \sigma_2(\overline{W}_{\mathcal{L}}^k)$. Since $\mathbb{G}_{\mathcal{L}}$ is connected and $\overline{W}_{\mathcal{L}}$ is doubly stochastic, the value $\rho_{\mathcal{L}} < 1$ is equal to the second largest eigenvalue modulus of $\overline{W}_{\mathcal{L}}$. Additionally, our analysis holds for the lazy Metropolis weight as well for which by [49, Lemma 2.2], the value $\rho_{\mathcal{L}}$ can be upper bounded by $(1 - 1/(71|\mathcal{L}|^2))$, while [50] improves the constant of this bound to 4.

Next, define for $t \geq 0$ the following deterministic dynamic that excludes malicious agents,

$$r_i(t) = \overline{w}_{ii} z_i(t) + \sum_{j \in \mathcal{N}_i \cap \mathcal{L}} \overline{w}_{ij} z_j(t)$$
$$z_i(t+1) = \Pi_{\mathcal{X}} \left(r_i(t) - \gamma(t) \nabla f_i(r_i(t)) \right). \tag{13}$$

Denote for $T \geq 1$,

$$\overline{h}(T) \triangleq \frac{G^2 T}{\mu} + \frac{2G^2 T}{\mu(1-\rho_{\mathcal{L}})} + \frac{8(\mu+L)G^2}{\mu^2(1-\rho_{\mathcal{L}})^2} \ln\left(\frac{T+2}{2}\right) + \frac{2\eta G}{1-\rho_{\mathcal{L}}} + \frac{2(\mu+L)(\mu\eta+2G)^2}{\mu^2(1-\rho_{\mathcal{L}})^2} + \frac{2G^2 + 4G\eta(\mu+L)}{\mu(1-\rho_{\mathcal{L}})^3} + \frac{G^2(\mu+L)}{\mu^2(1-\rho_{\mathcal{L}})^4}, \tag{14}$$

The function $\bar{h}(T)$ grows linearly in T. Additionally, it comprises two terms: 1) the first term which captures the error rate for the centralized gradient descent optimization (see [51]) without malicious agents, and 2) the following terms that include $\rho_{\mathcal{L}}$ which capture the contribution from distributing the optimization over a decentralized network (without malicious agents) that is characterized by the second largest eigenvalue modulus of $\overline{W}_{\mathcal{L}}$.

Theorem 1. *The dynamic* (13) *converges to the optimal point for every initial point $z_i(0) \in \mathcal{X}$, $i \in \mathcal{L}$, that is,*

$$\lim_{t\to\infty} \|z_i(t) - x^\star_{\mathcal{L}}\| = 0, \ \forall\, i \in \mathcal{L},$$

whenever $\sum_{t=0}^{\infty} \gamma(t) = \infty$ and $\sum_{t=0}^{\infty} \gamma^2(t) < \infty$.

Moreover, if $\gamma(t) = \frac{2}{\mu(t+1)}$, then

$$\frac{1}{|\mathcal{L}|} \sum_{i\in\mathcal{L}} \|z_i(T) - x^\star_{\mathcal{L}}\|^2 \leq \min\left\{4\eta^2, \frac{4\overline{h}(T)}{\mu T(T+1)}\right\}, \tag{15}$$

for any initial points $z_i(0) \in \mathcal{X}$, $i \in \mathcal{L}$, and any $T \geq 1$.

The proof of this theorem is conventional. Nonetheless, for the sake of the completeness of the presentation, we prove Theorem 1 in the supplementary material.

A. Convergence to optimal value almost surely

Denote by $x_{\mathcal{L}}(t, T_f = k)$ the data values $(x_i(t))_{i\in\mathcal{L}}$ of the dynamic (7), assuming that $T_f = k$. Subsequently, for every $t \geq \max\{T_f, T_0\}$ all the legitimate agents participate in the dynamic (7) and all the malicious agents are excluded from it. Thus from time $\max\{T_f, T_0\}$ the dynamic (7) can be captured by a dynamic of the form (13), where t is is replaced with $t - \max\{T_f, T_0\}$.

Theorem 2 (Convergence a.s. to the optimal point)**.** *The sequence $\{x_i(t)\}$ converges a.s. to $x^\star_{\mathcal{L}}$ for every $i \in \mathcal{L}$ and $T_0 \geq 0$.*

Proof: From Corollary 1 there exists a finite time T_f such that every legitimate agent i classifies correctly all of its legitimate and malicious neighbors at all times $t \geq T_f$ a.s. Thus, the dynamic (7) is equivalent to the dynamic (13) with the initial inputs $z_i(0) = x_i(\max\{T_f, T_0\})$ where $i \in \mathcal{L}$.

By Theorem 1 the dynamic (13) converges to $x^\star_{\mathcal{L}}$. Additionally, by Assumption 2, $x_{\mathcal{L}}(T_f)$ is finite for every finite T_f. Applying Corollary 1 with the a.s. finiteness of T_f concludes the proof. ■

B. Convergence in mean

Next, we establish the convergence in mean of each sequence $x_i(t)$ to $x^\star_{\mathcal{L}}$, where $i \in \mathcal{L}$.

Theorem 3 (Convergence in mean to the optimal point)**.** *For every $T_0 \geq 0$, the sequence $\{x_i(t)\}$ converges in the r-th mean to $x^\star_{\mathcal{L}}$ for every $i \in \mathcal{L}$ and $r \geq 1$, i.e.,*

$$\lim_{t\to\infty} \mathbf{E}\left[\|x_i(t) - x^\star_{\mathcal{L}}\|^r\right] = 0, \text{ for all } r \geq 1.$$

We present two types of proofs for this theorem, the first relies on the almost sure convergence of Theorem 2 and Assumption 2. For the sake of completeness of presentation, we additionally establish convergence in mean by definition in Appendix B.

Proof via Dominated Convergence Theorem: First, by Assumption 2, we have $\|x\| \leq \eta$. It follows by the triangle inequality that $\|x - y\|^r \leq (2\eta)^r < \infty$ for every $x, y \in \mathcal{X}$. Recalling Theorem 2 we can apply the Dominated Convergence Theorem (see [52, Theorem 1.6.7]) to each sequence $\{\|x_i(t) - x^\star_{\mathcal{L}}\|^r\}$, $i \in \mathcal{L}$, to conclude the result. ■

VI. Finite Time Analysis: Expected Convergence Rate

This section derives analytical guarantees for the finite time regime in the form of the expected convergence rate. We present two upper bounds on the convergence rate. The first upper bound, stated in Theorem 4, relies on Lemma 2 to provide probabilistic bounds on the correct classification time, and on the convergence rate of the nominal dynamic (13) which ignores the inputs of malicious agents. We tighten this bound in Theorem 5 by analyzing the dynamic (7) directly utilizing the bounds on the error probabilities presented in Lemma 1.

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}[\|x_i(t)-x^\star\|^2]\le\min\left\{4\eta^2,\frac{4\overline{h}\,(t-T_0)}{\mu(t-T_0)(t-T_0+1)}\right\}+4\eta^2p_e\,(T_0)\,, \quad (16)$$

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}[\|x_i(t)-x^\star\|^2]\le\min\left\{4\eta^2,\frac{16\overline{h}\left(\frac{t-T_0}{2}\right)}{\mu(t-T_0)(t-T_0+2)}\right\}+4\eta^2p_e\left(\frac{t+T_0}{2}-1\right), \quad (17)$$

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}[\|x_i(t)-x^\star\|^2]\le\min\left\{4\eta^2,\frac{4\overline{h}\left(t-\lceil\frac{\ln(t)}{2\min\{E_{\mathcal{L}}^2,E_{\mathcal{M}}^2\}}\rceil\right)}{\mu\left(t-\lceil\frac{\ln(t)}{2\min\{E_{\mathcal{L}}^2,E_{\mathcal{M}}^2\}}\rceil\right)\left(t-\lceil\frac{\ln(t)}{2\min\{E_{\mathcal{L}}^2,E_{\mathcal{M}}^2\}}\rceil+1\right)}\right\}+4\eta^2\cdot\frac{D_{\mathcal{L}}+D_{\mathcal{M}}}{T}. \quad (18)$$

A. The Expected Convergence Rate via the Correct Classification Time

Utilizing the upper bounds and Theorem 1 we can upper bound the expected suboptimality gap as follows.

Theorem 4. *For every $t\ge T_0$*

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}[\|x_i(t)-x^\star\|^2]\le$$
$$\min_{m\in[T_0:t-1]}\left\{\min\left\{4\eta^2,\frac{4\overline{h}\,(t-m)}{\mu(t-m)(t-m+1)}\right\}+4\eta^2p_e(m)\right\}.$$

Before proving this theorem we observe that since $\frac{\overline{h}(t)}{t(t+1)}$ and $p_e(t)$ are monotonically decreasing for $t\ge1$, choosing the values, $m=T_0$, $m=(t+T_0)/2$, and $m=\lceil\frac{\ln(t)}{2\min\{E_{\mathcal{L}}^2,E_{\mathcal{M}}^2\}}\rceil$, leads to the following corollary.

Corollary 3. *For every $t\ge\max\{T_0,1\}$ the expected convergence rate of $\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}[\|x_i(t)-x^\star\|^2]$ is upper bounded as given in* (16) *and* (17). *Furthermore, for every t such that $T_0\le\lceil\frac{\ln(t)}{2\min\{E_{\mathcal{L}}^2,E_{\mathcal{M}}^2\}}\rceil\le t-1$, the expected convergence rate of $\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}[\|x_i(t)-x^\star\|^2]$ is upper bounded by in* (18).

Now, we proceed to prove Theorem 4.

Proof of Theorem 4: First, note that $\frac{\overline{h}(t)}{t(t+1)}$ and $p_c(t)$ are nonincreasing functions of t for $t\ge1$. Denote $M_k\triangleq\max\{k,T_0\}$. For every $t-1\ge T_0$ and $m\in[T_0:t-1]$ we have that

$$\begin{aligned}
&\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}\left[\|x_i(t)-x_{\mathcal{L}}^\star\|^2\right]\\
&\overset{(a)}{=}\frac{1}{|\mathcal{L}|}\sum_{k=0}^{t-1}\Pr(T_f=k)\sum_{i\in\mathcal{L}}\mathbf{E}\left[\|x_i(t)-x_{\mathcal{L}}^\star\|^2\middle|T_f=k\right]\\
&\quad+\Pr(T_f>t-1)\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}\left[\|x_i(t)-x_{\mathcal{L}}^\star\|^2\middle|T_f>t-1\right]\\
&\overset{(b)}{\le}\frac{1}{|\mathcal{L}|}\sum_{k=0}^{t-1}\sum_{i\in\mathcal{L}}\Pr(T_f=k)\mathbf{E}\left[\|z_i(t-M_k)-x_{\mathcal{L}}^\star\|^2\,\middle|T_f=k\right]\\
&\quad+4\eta^2\Pr(T_f>t-1)\\
&\overset{(c)}{\le}\sum_{k=0}^{t-1}\Pr(T_f=k)\min\left\{4\eta^2,\frac{4\overline{h}(t-M_k)}{\mu(t-M_k)(t-M_k+1)}\right\}\\
&\quad+4\eta^2\Pr(T_f>t-1)\\
&\le\sum_{k=0}^{t-1}\Pr(T_f=k)\min\left\{4\eta^2,\frac{4\overline{h}(t-M_k)}{\mu(t-M_k)(t-M_k+1)}\right\}\\
&\quad+4\eta^2p_e(t-1)\\
&\le\sum_{k=0}^{m}\Pr(T_f=k)\min\left\{4\eta^2,\frac{4\overline{h}(t-M_k)}{\mu(t-M_k)(t-M_k+1)}\right\}\\
&\quad+\sum_{k=m+1}^{t-1}p_c(k-1)\min\left\{4\eta^2,\frac{4\overline{h}(t-M_k)}{\mu(t-M_k)(t-M_k+1)}\right\}\\
&\quad+4\eta^2p_e(t-1)\\
&\le\min\left\{4\eta^2,\frac{4\overline{h}\,(t-m)}{\mu(t-m)(t-m+1)}\right\}\sum_{k=0}^{m}\Pr(T_f=k)\\
&\quad+4\eta^2\sum_{k=m+1}^{t-1}p_c(k-1)+4\eta^2p_e(t-1)\\
&\le\min\left\{4\eta^2,\frac{4\overline{h}\,(t-m)}{\mu(t-m)(t-m+1)}\right\}\\
&\quad+4\eta^2\left(p_e(m)-p_e(t-1)\right)+4\eta^2p_e(t-1)\\
&=\min\left\{4\eta^2,\frac{4\overline{h}\,(t-m)}{\mu(t-m)(t-m+1)}\right\}+4\eta^2p_e(m),
\end{aligned}$$

where (a) follows from the law of total expectation, (b) follows from the definition of T_f and Assumption 2, (c) follows from Theorem 1, and the remaining steps follow from Lemma 2. ∎

B. Convergence Results via Trustworthiness Misclassification Error Probabilities

This section aims to tighten the bound on the convergence rate presented in Theorem 4 and Corollary 3. To this end, we develop an alternative analytical approach that evaluates more carefully how our choice of weights $w_{ij}(t)$, stepsize $\gamma(t)$, and T_0, together with the quality of trust values, captured by the constants $E_{\mathcal{L}}$ and $E_{\mathcal{M}}$, affects the dynamic (7).

Lemma 4. *Let $r\in\{1,2\}$, $i\in\mathcal{L}$, and $t\ge0$. Then,*

$$\mathbf{E}\left[\left(\sum_{j\in\mathcal{N}_i\cap\mathcal{L}}|w_{ij}(k)-\overline{w}_{ij}|\right)^r\right]\le p_c(k),$$
$$\mathbf{E}\left[\left(\sum_{j\in\mathcal{N}_i\cap\mathcal{M}}w_{ij}(k)\right)^r\right]\le\frac{p_c(k)}{2^r},$$
$$\mathbf{E}\left[|w_{ii}(k)-\overline{w}_{ii}|^r\right]\le\frac{p_c(k)}{2^r}.$$

Proof: First, note that, $w_{ii}(t) \geq 0.5$ for every $i \in \mathcal{L}$, thus

$$|w_{ii}(t) - \overline{w}_{ii}| \leq 0.5, \text{ and } \sum_{j \in \mathcal{N}_i \cap \mathcal{M}} w_{ij}(k) \leq 0.5.$$

Additionally, the event $\sum_{j \in \mathcal{N}_i \cap \mathcal{L}} |w_{ij}(k) - \overline{w}_{ij}| > 0$ can occur only if $W_{\mathcal{L}}(t) \neq \overline{W}_{\mathcal{L}}$. It follows by the triangle inequality that

$$\sum_{j \in \mathcal{N}_i \cap \mathcal{L}} |w_{ij}(k) - \overline{w}_{ij}| \leq \sum_{j \in \mathcal{N}_i \cap \mathcal{L}} [w_{ij}(k) + \overline{w}_{ij}] \leq 1.$$

Now, for all i and $j \in \mathcal{N}_i \cap \mathcal{L}$, we have

$$\begin{aligned}\mathbf{E}\left[\left(\sum_{j \in \mathcal{N}_i \cap \mathcal{L}} |w_{ij}(k) - \overline{w}_{ij}|\right)^r\right] &\leq \mathbf{E}\left[1 \cdot \mathbb{1}_{\{W_{\mathcal{L}}(t) \neq \overline{W}_{\mathcal{L}}\}}\right] \\ &\leq p_c(k).\end{aligned}$$

The rest of the proof follows similarly. ■

Next, we present an auxiliary proposition that we utilize in upper bounding the expected distance between an agent's value and the average agents' values at time t. Here we extend [53, Lemma 11] to the case of d-dimensional vectors with *random* perturbations. Note that a naïve implementation of [53, Lemma 11] for each of the dimensions $1, \ldots, d$ scales the resulting upper bound by $\sqrt{d}$. The upper bound we next derive eliminates this scaling.

Let $A \in \mathbb{R}^{d \times |\mathcal{L}|}$ and denote by $[A]_j$ the jth column of A. The Frobenius norm of the matrix A is defined as

$$\|A\|_{\mathrm{F}} \triangleq \sqrt{\sum_{i=1}^{d} \sum_{j \in \mathcal{L}} |a_{ij}|^2} = \sqrt{\sum_{j \in \mathcal{L}} \|[A]_j\|^2} = \sqrt{\sum_{i=1}^{d} \|[A^T]_i\|^2},$$

where A^T denotes the transpose matrix of the matrix A. Additionally, we denote by $\mathbf{1}$ the all-ones column vector with $|\mathcal{L}|$ entries.

Proposition 1. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact and convex set. Additionally, let $W(t) \triangleq (w_{ij}(t))_{i,j \in \mathcal{L}}$ be deterministic doubly stochastic matrices such that $\sigma_2(W(t)) \leq \rho$ for all $t \geq 0$. Furthermore, let $\Delta_i(t) \in \mathbb{R}^{d \times 1}$ be random vectors, and let $X(t) \in \mathbb{R}^{d \times |\mathcal{L}|}$ be defined by the following dynamic*

$$X(t+1) = X(t)W^T(t) + \Delta(t), \tag{19}$$

where $\Delta(t) = (\Delta_1(t), \ldots, \Delta_{|\mathcal{L}|}(t))$ for all $t \geq 0$. Assume that there exists a non-increasing sequence $\delta(t)$ such that

$$\mathbf{E}[\|\Delta_i(t)\|^2] \leq \delta^2(t), \ \forall\, i \in \mathcal{L},$$

and let

$$\overline{X}(t) \triangleq \frac{X(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T = \frac{1}{|\mathcal{L}|}\sum_{i \in \mathcal{L}} [X(t)]_i \mathbf{1}^T.$$

Then,

$$\frac{\sum_{j \in \mathcal{L}} \mathbf{E}[\|[X(t)]_j - \overline{X}(t)\|]}{|\mathcal{L}|} \leq 2\eta\rho^t + \frac{\delta(0)\rho^{t/2}}{1-\rho} + \frac{\delta(t/2)}{1-\rho},$$

$$\frac{\sum_{j \in \mathcal{L}} \mathbf{E}[\|[X(t)]_j - \overline{X}(t)\|^2]}{|\mathcal{L}|} \leq \left[2\eta\rho^t + \frac{\delta(0)\rho^{t/2}}{1-\rho} + \frac{\delta(t/2)}{1-\rho}\right]^2.$$

We present the proof for Proposition 1 in Appendix C. We apply this proposition in establishing the upper bounds in the forthcoming lemma. Toward providing the lemma, define

$$\overline{x}_{\mathcal{L}}(t) \triangleq \frac{1}{|\mathcal{L}|}\sum_{i \in \mathcal{L}} x_i(t),$$

$$\begin{aligned}\delta_{\mathcal{M}}(t, T_0) \triangleq 2\eta\rho_{\mathcal{L}}^{t-T_0} &+ \frac{(2\eta\sqrt{p_c(T_0)} + G\gamma(0))\rho_{\mathcal{L}}^{(t-T_0)/2}}{1-\rho_{\mathcal{L}}} \\ &+ \frac{2(\eta\sqrt{p_c((t+T_0)/2)} + G\gamma((t-T_0)/2))}{1-\rho_{\mathcal{L}}},\end{aligned} \tag{20}$$

where $t/2 \triangleq \lfloor \frac{t}{2} \rfloor$.

Lemma 5. *For every $t \geq 0$*

$$\frac{1}{|\mathcal{L}|}\sum_{i \in \mathcal{L}} \mathbf{E}[\|x_i(t) - \overline{x}_{\mathcal{L}}(t)\|] \leq \delta_{\mathcal{M}}(t, T_0), \ \text{and}$$

$$\frac{1}{|\mathcal{L}|}\sum_{i \in \mathcal{L}} \mathbf{E}[\|x_i(t) - \overline{x}_{\mathcal{L}}(t)\|^2] \leq \delta_{\mathcal{M}}^2(t, T_0).$$

Proof: Recall that $\phi_i(t) \triangleq \Pi_{\mathcal{X}}[y_i(t)] - c_i(t)$. The matrices $W_{\mathcal{L}}(t) \triangleq (w_{ij}(t))_{i,j \in \mathcal{L}}$ are random, vary with time, and can be sub-stochastic. Thus we cannot naïvely apply Proposition 1 for $\Delta_i(t) = \phi_i(t)$. Instead, we substitute

$$\Delta_i(t) = c_i(t) - \overline{c}_i(t) + \phi_i(t),$$

where $\overline{c}_i(t) \triangleq \overline{w}_{ii}x_i(t) + \sum_{j \in \mathcal{N}_i \cap \mathcal{L}} \overline{w}_{ij}x_j(t)$. It follows that

$$x_i(t+1) = \overline{w}_{ii}x_i(t) + \sum_{j \in \mathcal{N}_i \cap \mathcal{L}} \overline{w}_{ij}x_j(t) + \Delta_i(t).$$

By the Cauchy–Schwarz inequality for the ℓ_2 inner product, and by the Cauchy–Schwarz inequality *for expectations*

$$\begin{aligned}\mathbf{E}\left[\|\Delta_i(t)\|^2\right] &= \mathbf{E}\left[\|c_i(t) - \overline{c}_i(t) + \phi(t)\|^2\right] \\ &\leq \mathbf{E}\left[\|c_i(t) - \overline{c}_i(t)\|^2\right] + \mathbf{E}\left[\|\phi_i(t)\|^2\right] \\ &\quad + 2\sqrt{\mathbf{E}\left[\|c_i(t) - \overline{c}_i(t)\|^2\right]}\sqrt{\mathbf{E}\left[\|\phi_i(t)\|^2\right]}.\end{aligned}$$

By Lemma 4 and Assumption 2, we further have

$$\begin{aligned}&\mathbf{E}\left[\|c_i(t) - \overline{c}_i(t)\|^2\right] = \mathbf{E}\Bigg[\|[w_{ii}(t) - \overline{w}_{ii}]x_i(t) \\ &\quad + \sum_{j \in \mathcal{N}_i \cap \mathcal{L}} [w_{ij}(t) - \overline{w}_{ij}]x_j(t) + \sum_{j \in \mathcal{N}_i \cap \mathcal{M}} w_{ij}(t)x_j(t)\|^2\Bigg] \\ &\leq 4\eta^2 p_c(t).\end{aligned}$$

Additionally, by Assumption 3 and the non-expansiveness property of the projection, since $c_i(t) \in \mathcal{X}$ for all i and all t, it follows that $\|\phi_i(t)\| \leq \|y_i(t) - c_i(t)\| \leq G\gamma(t - T_0)$. Hence,

$$\begin{aligned}&\mathbf{E}\left[\|\Delta_i(t)\|^2\right] = \mathbf{E}\left[\|c_i(t) - \overline{c}_i(t) + \phi_i(t)\|^2\right] \\ &\leq 4\eta^2 p_c(t) + G^2\gamma^2(t - T_0) + 4G\eta\sqrt{p_c(t)}\,\gamma(t - T_0) \\ &= \left(2\eta\sqrt{p_c(t)} + G\gamma(t - T_0)\right)^2.\end{aligned} \tag{21}$$

We conclude the proof by substituting $\delta(t) = \widetilde{\delta}(t + T_0) = 2\eta\sqrt{p_c(t + T_0)} + G\gamma(t)$, $W(t) = \overline{W}_{\mathcal{L}}$, and $\rho = \rho_{\mathcal{L}}$ in Proposition 1 and using the transformation $t \to t - T_0$. ■

Now, we are ready to present our tightened convergence rate guarantees for Algo. 1.

Theorem 5. *Algo. 1 converges to the optimal point $x_{\mathcal{L}}^{*}$ in the mean-squared sense for every collection $x_i(0) \in \mathcal{X}$, $i \in \mathcal{L}$, of initial points i.e.,*

$$\lim_{t\to\infty} \mathbf{E}\left[\|x_i(t) - x_{\mathcal{L}}^{\star}\|^2\right] = 0, \, \forall\, i \in \mathcal{L}, \tag{22}$$

whenever $\sum_{t=0}^{\infty} \gamma(t) = \infty$ and $\sum_{t=0}^{\infty} \gamma^2(t) < \infty$.

Moreover, let $\gamma(t) = \frac{2}{\mu(t+2)}$. Then, for every $T_0 \geq 0$ and $T \geq T_0$ there exists a function $C_{\mathcal{M}}(T_0)$ that decreases exponentially with T_0 and is independent of T such that for any collection $x_i(0) \in \mathcal{X}$, $i \in \mathcal{L}$, and for all $T \geq T_0$,

$$\begin{aligned}&\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \mathbf{E}\left[\|x_i(T) - x_{\mathcal{L}}^{\star}\|^2\right] \\ &\qquad\leq \min\left\{4\eta^2, \frac{4\overline{h}(T-T_0) + C_{\mathcal{M}}(T_0)}{\mu(T-T_0)(T-T_0+1)}\right\},\end{aligned} \tag{23}$$

where $\overline{h}(\cdot)$ is defined in (14).

For the sake of simplicity of exposition, we only characterize the function $C_{\mathcal{M}}(T_0)$ with respect to its exponential decrease in T_0. Nonetheless, we define the function $C_{\mathcal{M}}(T_0)$ in (30) as part of the proof of Theorem 5. Intuitively, the $C_{\mathcal{M}}(T_0)$ term above represents the error term contributed by the presence of malicious agents in the distributed network. It can be seen that for large enough T the entire term on the right of the inequality (23) decays on the order of $O\left(\frac{1}{T}\right)$. Finally, we can observe that Theorem 5 tightens the results presented in Theorem 4 and Corollary 3 for the regime $T \gg 1$.

Before proceeding to prove this theorem, we point out that unlike the analysis for stochastic gradient models such as [48], in our model $w_{ij}(t)$ and $x_j(t)$ are correlated. This follows by the statistical dependence of $w_{ij}(t)$ and $w_{ij}(t-1)$. Thus, we cannot use the standard analysis which requires that $\mathbf{E}[w_{ij}(t)x_j(t)] = \mathbf{E}[w_{ij}(t)]\mathbf{E}[x_j(t)]$. Finally, we observe that the nonnegativity of the variance of random variables (22) and the sandwich theorem imply that $\lim_{t\to\infty} \mathbf{E}[\|x_i(t) - x_{\mathcal{L}}^{\star}\|] = 0, \, \forall\, i \in \mathcal{L}$. This result also holds since convergence in expectation in the rth moment implies convergence in expectation in the sth moment whenever $0 < s < r$.

Proof: By the non-expansiveness property of the projection, for every $t \geq T_0$ and $i \in \mathcal{L}$,

$$\|x_i(t+1) - x_{\mathcal{L}}^{\star}\|^2 = \|\Pi_{\mathcal{X}}\left(y_i(t)\right) - x_{\mathcal{L}}^{\star}\|^2 \leq \|y_i(t) - x_{\mathcal{L}}^{\star}\|^2.$$

Denote

$$\begin{aligned}\overline{c}_i(t) &\triangleq \overline{w}_{ii}x_i(t) + \sum_{j\in\mathcal{N}_i\cap\mathcal{L}} \overline{w}_{ij}x_j(t), \text{ and} \\ \overline{g}_i(t) &\triangleq \overline{c}_i(t) - \gamma(t-T_0)\nabla f_i(\overline{c}_i(t)) - x_{\mathcal{L}}^{\star}.\end{aligned}$$

Recall that $y_i(t) = c_i(t) - \gamma(t-T_0)\nabla f_i(c_i(t))$, then by the Cauchy-Schwarz inequality for the inner product on ℓ_2, the triangle inequality, and the linearity of the expectation

$$\begin{aligned}&\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \mathbf{E}\left[\|x_i(t+1) - x_{\mathcal{L}}^{\star}\|^2\right] \\ &= \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \mathbf{E}\Big[\|\overline{c}_i(t) - \gamma(t-T_0)\nabla f_i(\overline{c}_i(t)) - x_{\mathcal{L}}^{\star} \\ &\qquad + c_i(t) - \overline{c}_i(t) + \gamma(t-T_0)[\nabla f_i(\overline{c}_i(t)) - \nabla f_i(c_i(t))]\|^2\Big] \\ &\leq \underbrace{\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \mathbf{E}\Big[\|\overline{g}_i(t)\|^2\Big]}_{\text{(I)}} + \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \underbrace{\mathbf{E}\Big[\|c_i(t) - \overline{c}_i(t)\|^2\Big]}_{\text{(II)}} \\ &+ \frac{\gamma^2(t-T_0)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \underbrace{\mathbf{E}\Big[\|\nabla f_i(\overline{c}_i(t)) - \nabla f_i(c_i(t))\|^2\Big]}_{\text{(III)}} \\ &+ \frac{2}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \underbrace{\mathbf{E}\Big[\|\overline{g}_i(t)\| \cdot \|c_i(t) - \overline{c}_i(t)\|\Big]}_{\text{(IV)}} \\ &+ \frac{2\gamma(t-T_0)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \underbrace{\mathbf{E}\Big[\|\overline{g}_i(t)\| \cdot \|\nabla f_i(\overline{c}_i(t)) - \nabla f_i(c_i(t))\|\Big]}_{\text{(V)}} \\ &+ \frac{2\gamma(t-T_0)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \underbrace{\mathbf{E}\Big[\|c_i(t) - \overline{c}_i(t)\| \cdot \|\nabla f_i(\overline{c}_i(t)) - \nabla f_i(c_i(t))\|\Big]}_{\text{(VI)}}.\end{aligned}$$

By (3), Lemma 4, and the Cauchy-Schwarz inequality,

$$\begin{aligned}\text{(II)} &= \mathbf{E}\Big[\|[w_{ii}(t) - \overline{w}_{ii}]x_i(t) + \sum_{j\in\mathcal{N}_i\cap\mathcal{L}} [w_{ij}(t) - \overline{w}_{ij}]x_j(t) \\ &\qquad + \sum_{j\in\mathcal{N}_i\cap\mathcal{M}} w_{ij}(t)x_j(t)\|^2\Big] \\ &\leq 4\eta^2 p_c(t).\end{aligned} \tag{24}$$

Since ∇f_i are L-Lipschitz continuous

$$\text{(III)} \leq L^2\mathbf{E}\Big[\|\overline{c}_i(t) - c_i(t)\|^2\Big] \leq 4L^2\eta^2 p_c(t). \tag{25}$$

Now, by Lemma 4, $\mathbf{E}\Big[\|c_i(t) - \overline{c}_i(t)\|\Big] \leq 2\eta p_c(t)$. Thus,

$$\begin{aligned}\text{(IV)} &\leq (2\eta + \gamma(t-T_0)G)\mathbf{E}\Big[\|\,\overline{c}_i(t) - c_i(t)\|\Big] \\ &\leq 2\eta\left(2\eta + \gamma(t-T_0)G\right)p_c(t),\end{aligned} \tag{26}$$

and by the L-Lipschitz continuity of ∇f_i

$$\begin{aligned}\text{(V)} &\leq 2L\eta\left(2\eta + \gamma(t-T_0)G\right)p_c(t), \\ \text{(VI)} &\leq L\mathbf{E}\Big[\|c_i(t) - \overline{c}_i(t)\|^2\Big] \leq 4L\eta^2 p_c(t).\end{aligned} \tag{27}$$

Define $h_{\mathcal{M}}(t, T_0)$ as

$$4\eta^2 p_c(t)\left[2(L+1) + \gamma^2(t-T_0)L^2 + \frac{\gamma(t-T_0)G(L+1)}{2\eta}\right],$$

and recall that $\overline{x}_{\mathcal{L}}(t) \triangleq \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} x_i(t)$. Therefore,

$$\begin{aligned}&\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \mathbf{E}\left[\|x_i(t+1) - x_{\mathcal{L}}^{\star}\|^2\right] \leq \text{(I)} + h_{\mathcal{M}}(t, T_0) \\ &\leq h_{\mathcal{M}}(t, T_0) + \frac{(1 - \mu\gamma(t-T_0))}{|\mathcal{L}|}\sum_{i\in\mathcal{L}} \mathbf{E}\left[\|x_i(t) - x_{\mathcal{L}}^{\star}\|^2\right]\end{aligned}$$

$$+\gamma^2(t-T_0)G^2+\frac{2\gamma(t-T_0)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\Bigg[G\mathbf{E}\left[\|x_i(t)-\overline{x}_{\mathcal{L}}(t)\|\right] + \frac{\mu+L}{2}\mathbf{E}\left[\|x_i(t)-\overline{x}_{\mathcal{L}}(t)\|^2\right]\Bigg],$$

where the last inequality follows from Assumption 3, the convexity of $\|\cdot\|^2$ and the double stochasticity of $\overline{W}_{\mathcal{L}}$.

Recall (20) and denote

$$\widetilde{h}_{\mathcal{M}}(t,T_0)\triangleq\gamma(t-T_0)G^2+2G\delta_{\mathcal{M}}(t,T_0)+(\mu+L)\delta^2_{\mathcal{M}}(t,T_0).$$

Here, the term $\widetilde{h}_{\mathcal{M}}(t,T_0)$ is affected by the distributed nature of our optimization process and the presence of malicious agents. We utilize Lemma 5 to conclude that

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}\left[\|x_i(t+1)-x^\star_{\mathcal{L}}\|^2\right]\le\gamma(t-T_0)\widetilde{h}_{\mathcal{M}}(t,T_0) + h_{\mathcal{M}}(t,T_0)+\frac{(1-\mu\gamma(t-T_0))}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}\left[\|x_i(t)-x^\star_{\mathcal{L}}\|^2\right].$$

Thus, since $|\mathcal{L}|<\infty$

$$\lim_{t\to\infty}\mathbf{E}\left[\|x_i(t)-x^\star_{\mathcal{L}}\|^2\right]=0,\ \forall\, i\in\mathcal{L},\tag{28}$$

whenever $\sum_{t=0}^{\infty}\gamma(t)=\infty$ and $\sum_{t=0}^{\infty}\gamma^2(t)<\infty$.

To prove the second part of the theorem, we let $\gamma(t)=\frac{2}{\mu(t+2)}$ as proposed in [51]. It follows that

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}\left[\|x_i(t+1)-x^\star_{\mathcal{L}}\|^2\right]\le\frac{2\widetilde{h}_{\mathcal{M}}(t,T_0)}{\mu(t-T_0+2)} + h_{\mathcal{M}}(t,T_0)+\frac{t}{t-T_0+2}\cdot\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}\left[\|x_i(t)-x^\star_{\mathcal{L}}\|^2\right].$$

Multiplying both sides by $(t-T_0+1)(t-T_0+2)$ and summing over the set $t\in\{T_0,T_0+1,\ldots,T-1\}$ yield the upper bound

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{E}[\|x_i(T)-x^\star_{\mathcal{L}}\|^2]\le\frac{2\sum_{t=T_0}^{T-1}(t-T_0+1)\widetilde{h}_{\mathcal{M}}(t,T_0)}{\mu(T-T_0)(T-T_0+1)} + \frac{\sum_{t=T_0}^{T-1}(t-T_0+1)(t-T_0+2)h_{\mathcal{M}}(t,T_0)}{(T-T_0)(T-T_0+1)}.$$

Using the identity $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$, we deduce that

$$\sqrt{p_c(t)}\le\sqrt{D_{\mathcal{L}}}e^{-tE^2_{\mathcal{L}}}+\sqrt{D_{\mathcal{M}}}e^{-tE^2_{\mathcal{M}}}.\tag{29}$$

In addition, we utilize the identities for $|v|\in(0,1)$

$$\sum_{t=0}^{\infty}tv^t=\frac{v}{(1-v)^2}\ \text{and}\ \sum_{t=0}^{\infty}t^2v^t=\frac{2v}{(1-v)^3}.$$

Denote

$$\widetilde{C}_1(T_0,E,D)\triangleq\frac{16\eta e^{-T_0E^2}\sqrt{D}}{1-\rho_{\mathcal{L}}}\Bigg[\frac{G}{(1-e^{-E^2})^2}+ \frac{G+(\eta+\frac{4}{\mu})(\mu+L)}{(1-\rho_{\mathcal{L}})^2}+\frac{(\mu+L)(G+2\mu\sqrt{D}e^{-T_0E^2})}{\mu(1-\rho_{\mathcal{L}})^3}\Bigg],$$

and $C_1(T_0)\triangleq\widetilde{C}_1(T_0,E_{\mathcal{L}},D_{\mathcal{L}})+\widetilde{C}_1(T_0,E_{\mathcal{M}},D_{\mathcal{M}})$.

Further algebra yields that

$$\sum_{t=T_0}^{T-1}(t-T_0+1)\widetilde{h}_{\mathcal{M}}(t,T_0)\le2\overline{h}(t-T_0)+C_1(T_0),$$

here, the added term $C_1(T_0)$ captures the influence of the malicious agents on the term (I). Additionally, denote

$$\widetilde{C}_2(T_0,E,D)\triangleq\frac{4\eta(L+1)De^{-2T_0E^2}}{\left(1-e^{-2E^2}\right)} \cdot\left[\frac{4\eta e^{-2E^2}}{\left(1-e^{-2E^2}\right)^2}+\frac{(6\eta+\frac{G}{\mu})e^{-2E^2}}{1-e^{-2E^2}}+4\eta+\frac{4\eta L}{\mu^2}+\frac{G}{\mu}\right],$$

and

$$C_2(T_0)\triangleq\widetilde{C}_2(T_0,E_{\mathcal{L}},D_{\mathcal{L}})+\widetilde{C}_2(T_0,E_{\mathcal{M}},D_{\mathcal{M}}).$$

Then,

$$\sum_{t=T_0}^{T-1}(t-T_0+1)(t-T_0+2)h_{\mathcal{M}}(t,T_0) =8\eta^2(L+1)\sum_{t=0}^{T-T_0-1}(t+1)(t+2)p_c(t+T_0) +4\eta^2L^2\sum_{t=0}^{T-T_0-1}(t+1)(t+2)\gamma^2(t)p_c(t+T_0) +2\eta G(L+1)\sum_{t=0}^{T-T_0-1}(t+1)(t+2)\gamma(t)p_c(t+T_0)\le C_2(T_0).$$

We conclude the proof by letting

$$C_{\mathcal{M}}(T_0)=2C_1(T_0)+\mu C_2(T_0).\tag{30}$$

■

Thus, we have shown that indeed we are able to recover convergence *to the optimal value* of the original distributed optimization problem, given in (1) even in the presence of malicious agents. Further, we have established an upper bound on the expected value of $\|x_i(t)-x^\star_{\mathcal{L}}\|^2$, for all $i\in\mathcal{L}$, as a function of the time t as given by (23) in Theorem 5.

VII. Numerical Results

This section presents numerical results that validate the convergence results we derived for Algo. 1. As a benchmark, we compare the performance of our proposed Algo. 1 to that of [11] which adapts the W-MSR consensus algorithm [54] to the case of distributed optimization. This W-MSR based algorithm is applicable only for the one-dimensional case, i.e. $d=1$. To this end, we compare our results to [55] which extends [11] to multi-dimensional data values. Following the notations in [11], we denote by F the maximal number of highest values and lowest values that each legitimate agent discards, overall a legitimate agent may ignore no more than $2F$ values for $d=1$.

We consider a distributed network with $|\mathcal{L}|=15$ legitimate agents and $|\mathcal{M}|\in\{15,30\}$ malicious agents. To maximize the malicious agents' impact every malicious agent is connected to all the legitimate agents. The legitimate agent's connectivity is captured by Fig. 1.

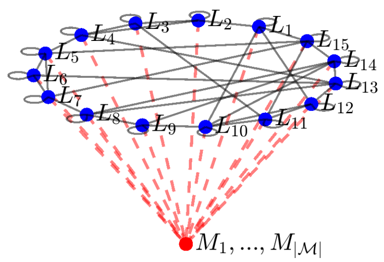

Figure 1. Undirected graph $\mathbb{G}$. Two agents are neighbors if they are connected by an edge. Legitimate and malicious agents are depicted by blue and red nodes, respectively. Edges between legitimate agents are depicted by black solid lines. Edges between legitimate and malicious agents are depicted by red dashed lines.

The trust values are generated as follows. Let $\mathbf{E}[\alpha_{ij}] = 0.55$ if $j \in \mathcal{N}_i \cap \mathcal{L}$, and $\mathbf{E}[\alpha_{ij}] = 0.45$ if $j \in \mathcal{N}_i \cap \mathcal{M}$. The random variable α_{ij} is uniformly distributed on the interval $\left[\mathbf{E}[\alpha_{ij}] - \frac{\ell}{2}, \mathbf{E}[\alpha_{ij}] + \frac{\ell}{2}\right]$, for every $i \in \mathcal{L}$ and $j \in \mathcal{L} \cup \mathcal{M}$. We consider the values ℓ: $0.6, 0.8$, in both scenarios $|E_{\mathcal{L}}| = |E_{\mathcal{M}}| = 0.05$, however, the variance of the trust values when $\ell = 0.8$ are higher. We remark that the legitimate agents are ignorant regarding the values $\mathbf{E}[\alpha_{ij}]$ and ℓ. Due to the stochasticity of the trust values, we average the numerical results across 100 system realizations.

We use the following stepsize

$$\gamma(t) = \frac{1}{t+2} \cdot \mathbb{1}_{\{t \geq 0\}}.$$

Denote for every client i, $a_i \in \mathbb{R}^d, b_i \in \mathbb{R}$. Additionally, denote $\lambda \geq 0$, and define for every agent $i \in \mathcal{L}$ the following strongly convex loss with ℓ_2 regularizer:

$$f_i(x) = \frac{1}{2}(a_i^T x - b_i)^2 + \frac{\lambda}{2}\|x\|^2.$$

We constrain the legitimate agents' values to lie in the d-dimensional box $[-\eta, \eta]^d$, i.e., $\mathcal{X} = [-\eta, \eta]^d$, where $\eta = 50$.

It follows that the global optimization problem the legitimate agents aim to solve distributively is

$$\min_{x \in [-\eta,\eta]^d} \left\{ \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} \frac{1}{2}(a_i^T x - b_i)^2 + \frac{\lambda}{2}\|x\|^2 \right\}. \tag{31}$$

Furthermore, denote by $x^{\star \mathrm{UC}}$ the optimal point of the unconstrained counterpart of (31), that is

$$x^{\star \mathrm{UC}} = \left(\lambda I + \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} a_i a_i^T \right)^{-1} \left(\frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} a_i b_i \right).$$

Then, the optimal point of (31) is,

$$[x^\star]_i = [x^{\star \mathrm{UC}}]_i \mathbb{1}_{\{[x^{\star \mathrm{UC}}]_i \in \mathcal{X}\}} + \mathrm{sign}([x^{\star \mathrm{UC}}]_i)\eta \mathbb{1}_{\{[x^{\star \mathrm{UC}}]_i \notin \mathcal{X}\}}.$$

Let $[y]_i$ refer to the ith entry of the vector y. In this setup

$$\nabla f_i(x) = a_i(a_i^T x - b_i) + \lambda x.$$

Additionally, for every $i \in \{1, \ldots, d\}$ and $y \in \mathbb{R}^d$ let

$$[\Pi_{\mathcal{X}}(y)]_i = [y]_i \mathbb{1}_{\{[y]_i \in [-\eta,\eta]\}} + \mathrm{sign}([y]_i)\eta \mathbb{1}_{\{[y]_i \notin [-\eta,\eta]\}}.$$

Hereafter, we utilize the following choice $(\tilde{b}_i)_{i=1}^{15} = (115.7, 163.3, -81.7, 127.2, -63.7, 58.4, -3.1, 62.9, 54.5, 144.9, -121.1, 9.3, -2.6, -124.5, 131)$. Note that setting $d = 1$, $a_i = 1$, and $\lambda = 0$ results in a one-dimensional constrained consensus problem. Next, we consider this special case and then a multi-dimensional setup.

To evaluate the performance of Algo. 1 we denote the average error at time t by

$$\overline{e}(t) \triangleq \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} \|x_i(t) - x_{\mathcal{L}}^\star\|.$$

A. Consensus with Constraints

Here, we consider the following consensus problem with constraints. The legitimate agents aim to minimize the function

$$\min_{x \in [-50,50]} \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} f_i(x), \text{ where } f_i = \frac{1}{2}(x - b_i)^2,$$

where $b_i = \tilde{b}_i$ for all $i \in \mathcal{L}$. The optimal nominal value is $x_{\mathcal{L}}^\star = x_{\mathcal{L}}^{\star \mathrm{UC}} \approx 31.367$ for our choice of $(b_i)_{i=1}^{15}$, where we round the solution to the second digit after the decimal point. Consequently, the dynamic (7) can be written as follows

$$\begin{aligned} c_i(t) &= \sum_{j \in \mathcal{N}_i(t) \cup \{i\}} w_{ij}(t) x_j(t) \\ y_i(t) &= c_i(t) - \frac{\mathbb{1}_{\{t - T_0 \geq 0\}}}{t - T_0 + 2} \cdot (c_i(t) - b_i), \\ x_i(t+1) &= \mathbb{1}_{\{y_i(t) \in \mathcal{X}\}} y_i(t) + \mathrm{sign}(y_i(t))\eta \mathbb{1}_{\{y_i(t) \notin \mathcal{X}\}}. \end{aligned} \tag{32}$$

The initial values of the legitimate agents are chosen randomly and uniformly in the interval $[-\eta, \eta]$. Note that in this setup the optimal point lies in the constraint set $\mathcal{X}$. Nonetheless, the set $\mathcal{X}$ affects the update rule (32) and limits the data values at all times to be in the set $\mathcal{X}$.

To maximize the harmful impact of malicious agents on our analytical results, we choose the malicious agents' values to be equal to -50, i.e., $-\eta$, at all times. We choose this easy-to-spot malicious attack since it maximizes the deviation of the malicious inputs from the optimal nominal solution. Nonetheless, our algorithm can tolerate *arbitrary* malicious node inputs including the time-varying case or small deviation case that is usually much harder to detect.

Figs. 2 and 3 capture the average value of the distance of each legitimate agent from the optimal point $x_{\mathcal{L}}^\star$ normalized by the average of this initial distance, i.e., the average value of $\frac{\overline{e}(t)}{\overline{e}(0)}$, for each time t. We can see that the W-MSR algorithm fails to converge to the optimal solution. This occurs due to the high number of malicious agents, which is higher in this case than the tolerance threshold[2] in [11]. Additionally, the W-MSR algorithm is not guaranteed to converge to the optimal value $x_{\mathcal{L}}^\star$ but to a value in the convex hull of $\Pi_{[-\eta,\eta]}(b_i)$, $i \in \mathcal{L}$. In our case this convex hull is exactly the interval $[-\eta, \eta]$, thus the W-MSR algorithm cannot guarantee the reduction of the distance to the optimal value with respect to the interval $[-\eta, \eta]$. In contrast, Algo. 1 provides resilience to malicious activity and can tolerate even $30 = 2|\mathcal{L}|$ malicious agents, as evident in Figs. 2 and 3. Furthermore, we can see from

[2] We can upper bound the tolerance threshold for this setup by 2 following our argument in [39].

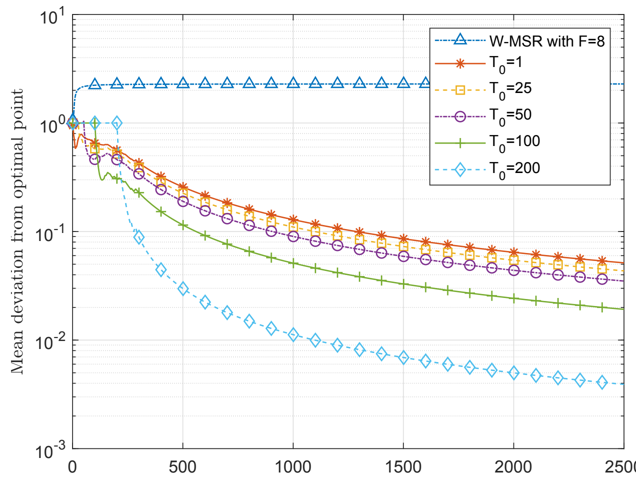

Figure 2. Average $\frac{\overline{e}(t)}{\overline{e}(0)}$ of as a function of t for $|\mathcal{M}|=15$ and the noise in the stochastic trust value is chosen as $\ell=0.8$.

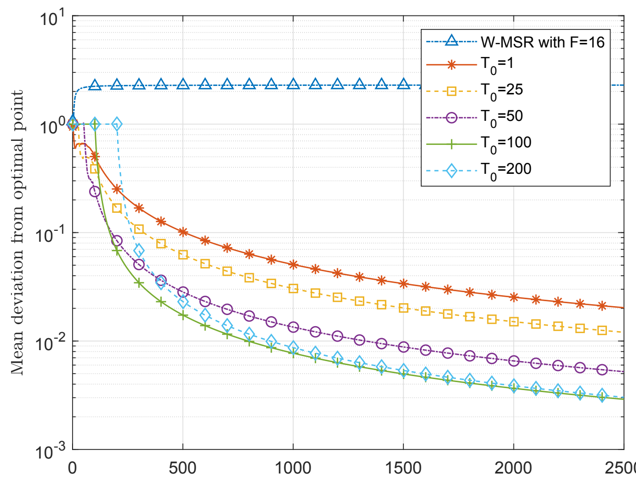

Figure 3. Average $\frac{\overline{e}(t)}{\overline{e}(0)}$ as a function of t for $|\mathcal{M}|=30$, and the noise in the stochastic trust value is chosen as $\ell=0.6$.

Figs. 2 and 3 that Algo. 1 is robust to small values of $|E_{\mathcal{L}}|$ and $|E_{\mathcal{M}}|$. Finally, Figs. 2 and 3 show that the variance of the trust values has more impact on T_0 values that are smaller than 100. This occurs since the higher variance of the trust values increases the misclassification errors. Since the probability of these errors decreases with T_0, they are less impactful when T_0 is 100 or higher. Note that regardless of the value of the observation window T_0, our algorithm eventually recovers the global optimum as predicted by theory. Thus the value of T_0 mostly dictates *the rate* of recovery of the global optimum.

B. Strongly Convex Loss with ℓ_2 Regularizer

Here, we examine a multi-dimensional setup where both the optimal solution and the updates are affected by the constraint set $\mathcal{X}$. Since the W-MSR algorithm is only valid for one-dimensional data values, we compare our results to the multi-dimensional extension of the W-MSR algorithm proposed in [55]. We note that in this case the number of tolerated malicious agents for the W-MSR algorithm is reduced by a factor of d, i.e., the dimension of the data values.

We examine the case where $d=5$ and $\lambda=1$. Additionally, we set $b_i = 2\tilde{b}_i$ for all $i \in \mathcal{L}$, and

$$\left[(a_i)_{i=1}^{15}\right]^T = \begin{pmatrix} -0.87 & -1.05 & -2.81 & -0.4 & -1.76 \\ -0.88 & -0.34 & 0.34 & -2.46 & 0.44 \\ -0.25 & 0.47 & -0.09 & -0.99 & -2.33 \\ -0.27 & -0.61 & -2.5 & -0.79 & 0.46 \\ -0.23 & 1.83 & 0.89 & -0.83 & -0.67 \\ -1.6 & 0.27 & -0.81 & -2.77 & -0.21 \\ -1.42 & -1.11 & -1.63 & -0.66 & -1.54 \\ -1.19 & -0.3 & -1.97 & -1.42 & -1.21 \\ -1.43 & -1.64 & 0.17 & -2.11 & -2.11 \\ -0.73 & 0.46 & -0.42 & -1.75 & 0.22 \\ -0.97 & -0.12 & -2.35 & -2.51 & -1.63 \\ -1.18 & -1.42 & -0.13 & -1.66 & 0.36 \\ -0.63 & -2.19 & -1.15 & -1.65 & -2.02 \\ 0.59 & -2.08 & 0.26 & -0.74 & -2.66 \\ -3.05 & -0.7 & 0.2 & -1.94 & -1.4 \end{pmatrix}. \quad (33)$$

Thus, the optimal points, rounded to the second digit after the decimal point, are

$$x_{\mathcal{L}}^{\star\mathrm{UC}} \approx (-61.67, -16.54, -21.19, -19.64, 60.4)^T, \text{ and}$$
$$x_{\mathcal{L}}^{\star} \approx (-50, -16.54, -21.19, -19.64, 50)^T.$$

Additionally, the inputs of the malicious agents at all times are $(50, 50, 50, 50, -50)^T$.

Figs. 4 and 5 capture the average value of the distance of each legitimate agent from the optimal point $x_{\mathcal{L}}^{\star}$ normalized by the average of this initial distance, i.e., the average value of $\frac{\overline{e}(t)}{\overline{e}(0)}$, for each time t. The plots included in Figs. 4 and 5 for our higher dimensional setup are consistent with the one-dimensional setup which is captured in Figs. 2 and 3. Moreover, we can see that Algo. 1 continues to perform well and mitigate the harmful effect of malicious inputs even in higher dimensions. This is in contrast to the multi-dimensional W-MSR algorithm [55] which is more vulnerable to malicious attacks as the dimension of the data values, i.e., d, increases.

VIII. Conclusions

This work studies the problem of resilient distributed optimization in the presence of malicious activity with an emphasis on cyberphysical systems. We consider the case where additional information in the form of stochastic inter-agent trust values is available. Under this model, we propose a mechanism for exploiting these trust values where legitimate agents learn to distinguish between their legitimate and malicious neighbors. We incorporate this mechanism to arrive at resilient distributed optimization where strong performance guarantees can be recovered. Specifically, we prove that our algorithm converges to the optimal solution of the nominal distributed optimization system with no malicious agents, both in expectation and almost surely. Additionally, we present two mechanisms to derive upper bounds on the expected distance of the agents' iterates from the optimal solution. The first is based on the correct classification of all malicious and legitimate agents. The second approach utilizes the dynamic of the attacked system to carefully tighten the upper bound on the expected convergence rate. Finally, we present numerical results that demonstrate the performance of our proposed distributed optimization framework.

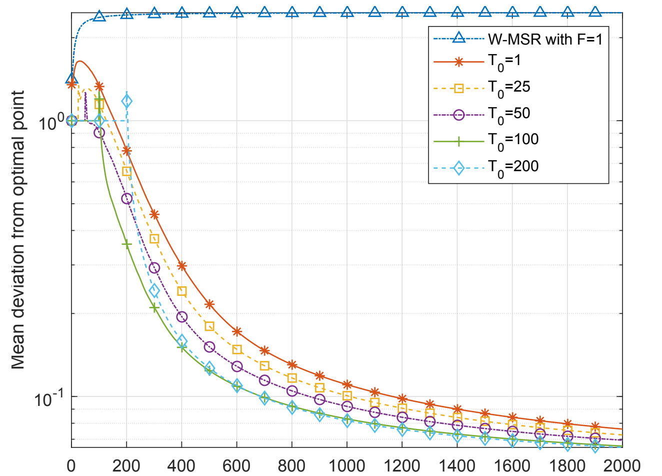

Figure 4. Average $\frac{\overline{e}(t)}{\overline{e}(0)}$ of as a function of t for $|\mathcal{M}| = 15$, $\ell = 0.8$.

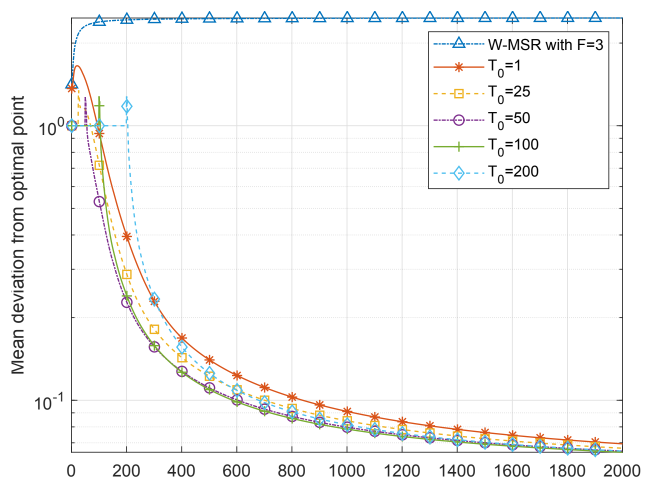

Figure 5. Average of $\frac{\overline{e}(t)}{\overline{e}(0)}$ as a function of t for $|\mathcal{M}| = 30$, $\ell = 0.6$.

The use of stochastic trust values allows us to recover convergence to the *global optimum* in distributed optimization problems even when more than half of the network is malicious. This represents a very challenging case where not many strong performance results are currently available, particularly in the case of distributed optimization problems. Thus the results of this paper strengthen our characterization of achievable performance and provide novel tools for, resilient trust-centered optimization in mutli-agent systems.

Appendix A
Proof of Lemma 2

First, since we initialize $\beta_{ij} = 0$, at time $t = 0$ all the agents are classified as legitimate by their neighbors. Thus, by the decision rule at time t we are guaranteed to make a classification mistake at time 0 whenever there are malicious agents in the system.

Denote

$$\mathcal{E}(k) \triangleq \bigcup_{\substack{i\in\mathcal{L},\\ j\in\mathcal{N}_i\cap\mathcal{L}}} \{\beta_{ij}(k) < 0\} \bigcup_{\substack{i\in\mathcal{L},\\ j\in\mathcal{N}_i\cap\mathcal{M}}} \{\beta_{ij}(k) \geq 0\}.$$

By Lemma 1, for all $t \geq 1$ we have that

$$\begin{aligned}
&\Pr(T_f = k) \\
&\overset{(a)}{\leq} \Pr\left(\mathcal{E}(k-1)\right) \\
&\overset{(b)}{\leq} \sum_{\substack{i\in\mathcal{L},\\ j\in\mathcal{N}_i\cap\mathcal{L}}} \Pr(\beta_{ij}(k-1) < 0) + \sum_{\substack{i\in\mathcal{L},\\ j\in\mathcal{N}_i\cap\mathcal{M}}} \Pr(\beta_{ij}(k-1) \geq 0) \\
&\overset{(c)}{\leq} \sum_{i\in\mathcal{L}} |\mathcal{N}_i \cap \mathcal{L}| \exp(-2(k-1)^+ E_{\mathcal{L}}^2) \\
&\qquad + \sum_{i\in\mathcal{L}} |\mathcal{N}_i \cap \mathcal{M}| \exp(-2(k-1)^+ E_{\mathcal{M}}^2), \qquad (34)
\end{aligned}$$

where (a) follows from the definition of T_f in Corollary 1 which implies that if $T_f = k$ there must be a misclassification error in the legitimacy of agents at time $k-1$, (b) follows from the union bound, and (c) follows from Lemma 1.

From the definition of T_f in Corollary 1 $\Pr(T_f > k-1) = 0$ for $k = 0$. Additionally, for all $k \geq 1$

$$\begin{aligned}
&\Pr(T_f > k-1) = \Pr\left(\bigcup_{t\geq k} \mathcal{E}(t)\right) \leq \sum_{t=k}^{\infty} \Pr\left(\mathcal{E}(t)\right) \\
&\leq \sum_{t=k}^{\infty}\sum_{i\in\mathcal{L}} |\mathcal{N}_i \cap \mathcal{L}| \exp(-2(t-1)^+ E_{\mathcal{L}}^2) \\
&\qquad + \sum_{t=k}^{\infty}\sum_{i\in\mathcal{L}} |\mathcal{N}_i \cap \mathcal{M}| \exp(-2(t-1)^+ E_{\mathcal{M}}^2) \\
&= D_{\mathcal{L}} \frac{\exp(-2(k-1)E_{\mathcal{L}}^2)}{1-\exp(-2E_{\mathcal{L}}^2)} + D_{\mathcal{M}} \frac{\exp(-2(k-1)E_{\mathcal{M}}^2)}{1-\exp(-2E_{\mathcal{M}}^2)}. \qquad (35)
\end{aligned}$$

Note that $\Pr(T_f > k-1)$ vanishes as t tends to infinity. ∎

Appendix B
Proof by Definition of Theorem 3

Let us assume that $t \geq 2T_0$. Denote $M_k \triangleq \max\{k, T_0\}$ and $t/2 \triangleq \lfloor \frac{t}{2} \rfloor$. Next, we utilize the law of total expectation as follows:

$$\begin{aligned}
&\mathbf{E}\left[\|x_i(t) - x_{\mathcal{L}}^\star\|^r\right] = \mathbf{E}\left[\mathbf{E}[\|x_i(t) - x_{\mathcal{L}}^\star\|^r \mid T_f]\right] \\
&= \sum_{k=0}^{t-1} \Pr(T_f = k)\mathbf{E}\left[\|x_i(t) - x_{\mathcal{L}}^\star\|^r \;\middle|\; T_f = k\right] \\
&\quad + \Pr(T_f > t-1)\mathbf{E}\left[\|x_i(t) - x_{\mathcal{L}}^\star\|^r \;\middle|\; T_f > t-1\right] \\
&= \sum_{k=0}^{t-1} \Pr(T_f = k)\mathbf{E}\left[\|x_{\mathcal{L}}(t, T_f = k) - x_{\mathcal{L}}^\star\|^r \;\middle|\; T_f = k\right] \\
&\quad + \Pr(T_f > t-1)\mathbf{E}\left(\|x_i(t) - x_{\mathcal{L}}^\star\|^r \;\middle|\; T_f > t-1\right) \\
&\leq \sum_{k=0}^{t-1} \Pr(T_f = k)\mathbf{E}\left[\|z_i(t - M_k) - x_{\mathcal{L}}^\star\|^r \;\middle|\; T_f = k\right] \\
&\quad + \Pr(T_f > t-1)\left(2\eta\right)^r.
\end{aligned}$$

Next, we upper bound the term $\sum_{k=0}^{t-1}\Pr(T_f = k)\mathbf{E}\left[\|z_i(t - M_k) - x_{\mathcal{L}}^\star\|^r \;\middle|\; T_f = k\right]$ utilizing the upper bounds (10) and (15) which hold for every initial point $z_i(0) \in \mathcal{X}$:

$$\begin{aligned}
&\sum_{k=0}^{t-1} \Pr(T_f = k)\mathbf{E}\left[\|z_i(t-M_k) - x^\star_{\mathcal{L}}\|^r \;\middle|\; T_f = k\right] \\
&= \sum_{k=0}^{t/2} \Pr(T_f = k)\left[\frac{4\overline{h}(t-M_k)|\mathcal{L}|}{\mu(t-M_k)(t-M_k+1)}\right]^{\frac{r}{2}} \\
&\quad + \sum_{k=t/2+1}^{t-1} \Pr(T_f = k)\left[\frac{4\overline{h}(t-M_k)|\mathcal{L}|}{\mu(t-M_k)(t-M_k+1)}\right]^{\frac{r}{2}}.
\end{aligned}$$

Now,

$$\begin{aligned}
&\sum_{k=0}^{t/2} \Pr(T_f = k)\left[\frac{4\overline{h}(t-M_k)|\mathcal{L}|}{\mu(t-M_k)(t-M_k+1)}\right]^{\frac{r}{2}} \\
&\overset{(a)}{\leq} \left[\frac{4\overline{h}(t/2)|\mathcal{L}|}{\mu(t/2)(t/2+1)}\right]^{\frac{r}{2}} \sum_{k=1}^{t/2} \Pr(T_f = k) \\
&\leq \left[\frac{4\overline{h}(t/2)|\mathcal{L}|}{\mu(t/2)(t/2+1)}\right]^{\frac{r}{2}},
\end{aligned}$$

where (a) follows since $t \geq 2T_0$. Furthermore,

$$\begin{aligned}
&\sum_{k=t/2+1}^{t-1} \Pr(T_f = k)\left[\frac{4\overline{h}(t-M_k)|\mathcal{L}|}{\mu(t-M_k)(t-M_k+1)}\right]^{\frac{r}{2}} \\
&\leq \sum_{k=t/2+1}^{t-1} D_{\mathcal{L}} \exp(-2(k-1)E^2_{\mathcal{L}}) \left[\frac{4\overline{h}(t/2)|\mathcal{L}|}{\mu(t/2)(t/2+1)}\right]^{\frac{r}{2}} \\
&\leq \left[\frac{4\overline{h}(t/2)|\mathcal{L}|}{\mu(t/2)(t/2+1)}\right]^{\frac{r}{2}} D_{\mathcal{L}} \frac{\exp(-2(t/2)E^2_{\mathcal{L}})}{1-|\mathcal{N}_i \cap \mathcal{L}|\exp(-2E^2_{\mathcal{L}})}.
\end{aligned}$$

Additionally by (11)

$$\begin{aligned}
&\Pr(T_f > t-1)\,[2\eta]^r \leq \\
&[2\eta]^r \left(D_{\mathcal{L}} \frac{\exp(-2(t-1)E^2_{\mathcal{L}})}{1-\exp(-2E^2_{\mathcal{L}})} + D_{\mathcal{M}} \frac{\exp(-2(t-1)E^2_{\mathcal{M}})}{1-\exp(-2E^2_{\mathcal{M}})}\right).
\end{aligned}$$

Consequently, $\lim_{t\to\infty} \mathbf{E}\left[\|x_i(t) - x^\star_{\mathcal{L}}\|^r\right] = 0,\ \forall\, r \geq 1.$ ■

APPENDIX C

Proof of Proposition 1: First, we utilize the identity

$$\sum_{j\in\mathcal{L}} \|[A]_j\|^2 = \|A\|^2_{\mathrm{F}}, \tag{36}$$

and the upper bound $\frac{1}{2}(\|[A]_i\|^2 + \|[A]_j\|^2) \geq \|[A]_i\| \cdot \|[A]_j\|$ to deduce that

$$\sum_{j\in\mathcal{L}} \|[A]_j\| \leq \sqrt{|\mathcal{L}|} \cdot \|A\|_{\mathrm{F}}. \tag{37}$$

Thus, in what follows, we focus our efforts on upper bounding the Frobenius norms $\|X(t) - \overline{X}(t)\|_{\mathrm{F}}$ and $\|X(t) - \overline{X}(t)\|^2_{\mathrm{F}}$.

Observe that

$$\begin{aligned}
\left\|\Delta(t) - \frac{\Delta(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|_{\mathrm{F}} &= \sqrt{\sum_{j\in\mathcal{L}} \left\|[\Delta(t)]_j - \frac{1}{|\mathcal{L}|}\sum_{k\in\mathcal{L}} [\Delta(t)]_k\right\|^2} \\
&\leq \sqrt{\sum_{j\in\mathcal{L}} \|[\Delta(t)]_j\|^2} = \|\Delta(t)\|_{\mathrm{F}}.
\end{aligned} \tag{38}$$

Additionally,

$$\begin{aligned}
&\left\|\left(X(t) - \frac{X(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right) W^T(t)\right\|_{\mathrm{F}} \\
&= \left\|W(t)\left(X(t) - \frac{X(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right)^T\right\|_{\mathrm{F}} \\
&\leq \sqrt{\sum_{i=1}^{d} \left\|W(t)\left([X^T(t)]_i - \mathbf{1}\frac{\mathbf{1}^T[X^T(t)]_i}{|\mathcal{L}|}\right)\right\|^2} \\
&\leq \sqrt{\sum_{i=1}^{d} \rho^2 \left\|\left([X^T(t)]_i - \mathbf{1}\frac{\mathbf{1}^T[X^T(t)]_i}{|\mathcal{L}|}\right)\right\|^2} \\
&= \rho \left\|X(t) - \frac{X(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|_{\mathrm{F}}.
\end{aligned} \tag{39}$$

It follows that

$$\begin{aligned}
&\|X(t+1) - \overline{X}(t+1)\|_{\mathrm{F}} \\
&= \left\|X(t+1) - \frac{X(t+1)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|_{\mathrm{F}} \\
&= \left\|X(t)W^T(t) + \Delta(t) - \frac{[X(t)+\Delta(t)]\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|_{\mathrm{F}} \\
&\overset{(a)}{\leq} \left\|X(t)W^T(t) - \frac{X(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|_{\mathrm{F}} + \left\|\Delta(t) - \frac{\Delta(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|_{\mathrm{F}} \\
&\overset{(b)}{\leq} \left\|X(t)W^T(t) - \frac{X(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|_{\mathrm{F}} + \|\Delta(t)\|_{\mathrm{F}} \\
&\overset{(c)}{=} \left\|\left(X(t) - \frac{X(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right) W^T(t)\right\|_{\mathrm{F}} + \|\Delta(t)\|_{\mathrm{F}} \\
&\overset{(d)}{\leq} \rho\left\|X(t) - \frac{X(t)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|_{\mathrm{F}} + \|\Delta(t)\|_{\mathrm{F}},
\end{aligned} \tag{40}$$

where (a) follows from the triangle inequality, (b) follows from (38), (c) follows from the double stochasticity of $W(t)$, and (d) follows from (39).

Thus,

$$\|X(t) - \overline{X}(t)\| \leq \rho^t \left\|X(0) - \frac{X(0)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\| + \sum_{k=0}^{t-1} \rho^{t-1-k} \|\Delta(k)\|.$$

Now, since $\mathbf{E}[\|\Delta_i(t)\|^2] \leq \delta^2(t)$ for all $i \in \mathcal{L}$, and by the non-negativity of the variance of $\|\Delta(k)\|_{\mathrm{F}}$

$$\begin{aligned}
\mathbf{E}[\|\Delta(k)\|_{\mathrm{F}}] &\leq \sqrt{\mathbf{E}\left[\|\Delta(k)\|^2_{\mathrm{F}}\right]} = \sqrt{\sum_{j\in\mathcal{L}} \mathbf{E}\left[\|\Delta_j(k)\|^2\right]} \\
&\leq \sqrt{|\mathcal{L}|\delta^2(k)} = \sqrt{|\mathcal{L}|}\delta(k).
\end{aligned} \tag{41}$$

then $\mathbf{E}[\|\Delta_i(t)\|] \leq \delta(t)$ for all $i \in \mathcal{L}$.

It follows from (3)

$$\begin{aligned}
&\left\|X(0) - \frac{X(0)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|_{\mathrm{F}} \\
&= \sqrt{\sum_{j\in\mathcal{L}} \left\|\left[X(0) - \frac{X(0)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right]_j\right\|^2} \\
&\leq \sqrt{\sum_{j\in\mathcal{L}} \left(\|[X(0)]_j\| + \left\|\left[\frac{X(0)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right]_j\right\|\right)^2}
\end{aligned}$$

$$\leq \sqrt{\sum_{j\in\mathcal{L}} (2\eta)^2} = 2\eta\sqrt{|\mathcal{L}|}. \tag{42}$$

It follows that

$$\begin{aligned}
&\frac{\mathbf{E}[\|X(t)-\overline{X}(t)\|_{\mathrm{F}}]}{|\mathcal{L}|}\\
&\leq \frac{2\eta}{\sqrt{|\mathcal{L}|}}\rho^t + \frac{1}{\sqrt{|L|}}\sum_{k=0}^{t-1}\rho^{t-1-k}\delta(k)\\
&\leq \frac{2\eta}{\sqrt{|\mathcal{L}|}}\rho^t + \frac{\delta(0)}{\sqrt{|L|}}\cdot\frac{\rho^{t/2}}{1-\rho} + \frac{\delta(t/2)}{\sqrt{|L|}(1-\rho)}. \qquad (43)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\|X(t)-\overline{X}(t)\|_{\mathrm{F}}^2\\
&\leq \rho^{2t}\left\|X(0)-\frac{X(0)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|^2\\
&+2\rho^t\left\|X(0)-\frac{X(0)\mathbf{1}}{|\mathcal{L}|}\mathbf{1}^T\right\|\sum_{k=0}^{t-1}\rho^{t-1-k}\|\Delta(k)\|_{\mathrm{F}}\\
&+\sum_{k_1=0}^{t-1}\sum_{k_2=0}^{t-1}\rho^{t-1-k_1}\rho^{t-1-k_2}\|\Delta(k_1)\|_{\mathrm{F}}\cdot\|\Delta(k_2)\|_{\mathrm{F}}\\
&\leq \rho^{2t}4\eta^2|\mathcal{L}| + 4\rho^t\eta\sqrt{|\mathcal{L}|}\sum_{k=1}^{t-1}\rho^{t-1-k}\|\Delta(k)\|_{\mathrm{F}}\\
&+\sum_{k_1=0}^{t-1}\sum_{k_2=0}^{t-1}\rho^{t-1-k_1}\rho^{t-1-k_2}\|\Delta(k_1)\|_{\mathrm{F}}\cdot\|\Delta(k_2)\|_{\mathrm{F}}. \qquad (44)
\end{aligned}$$

By the Cauchy–Schwarz inequality *for expectations*

$$\begin{aligned}
\mathbf{E}\left[\|\Delta(k_1)\|_{\mathrm{F}}\cdot\|\Delta(k_2)\|_{\mathrm{F}}\right] &\leq \sqrt{\mathbf{E}\left[\|\Delta(k_1)\|_{\mathrm{F}}^2\right]\mathbf{E}\left[\|\Delta(k_2)\|_{\mathrm{F}}^2\right]}\\
&\leq |\mathcal{L}|\delta(k_1)\delta(k_2). \qquad (45)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{\mathbf{E}\left[\|X(t)-\overline{X}(t)\|_{\mathrm{F}}^2\right]}{|\mathcal{L}|}\\
&\leq 4\eta^2\rho^{2t}+4\rho^t\eta\left[\frac{\delta(0)\rho^{t/2}}{1-\rho}+\frac{\delta(t/2)}{(1-\rho)}\right]\\
&\quad+\left[\frac{\delta(0)\rho^{t/2}}{1-\rho}+\frac{\delta(t/2)}{(1-\rho)}\right]^2\\
&=\left[2\eta\rho^t+\frac{\delta(0)\rho^{t/2}}{1-\rho}+\frac{\delta(t/2)}{(1-\rho)}\right]^2. \qquad (46)
\end{aligned}$$

We conclude the proof by using the upper bounds (36), (37). ■

References

[1] M. Yemini, A. Nedić, S. Gil, and A. J. Goldsmith, "Resilience to malicious activity in distributed optimization for cyberphysical systems," 2022, accpted to the IEEE Conference on Decision and Control (CDC).

[2] T. Halsted, O. Shorinwa, J. Yu, and M. Schwager, "A survey of distributed optimization methods for multi-robot systems," 2021. [Online]. Available: https://arxiv.org/pdf/2103.12840.pdf

[3] M. Zhu and S. Martínez, *Distributed optimization-based control of multi-agent networks in complex environments*. Springer, 2015.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020.

[5] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automat. Contr.*, vol. 54, no. 1, pp. 48–61, 2009.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[7] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. Automat. Contr.*, vol. 60, no. 3, pp. 601–615, 2015.

[8] G. Lan and Y. Zhou, "Asynchronous decentralized accelerated stochastic gradient descent," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 2, pp. 802–811, 2021.

[9] G. Lan, Y. Ouyang, and Y. Zhou, "Graph topology invariant gradient and sampling complexity for decentralized and stochastic optimization," 2021, https://arxiv.org/pdf/2101.00143.pdf.

[10] N. Ravi, A. Scaglione, and A. Nedić, "A case of distributed optimization in adversarial environment," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 5252–5256.

[11] S. Sundaram and B. Gharesifard, "Distributed optimization under adversarial nodes," *IEEE Trans. Automat. Contr.*, vol. 64, no. 3, pp. 1063–1076, 2019.

[12] A.-Y. Lu and G.-H. Yang, "Distributed secure state estimation in the presence of malicious agents," *IEEE Trans. Automat. Contr.*, vol. 66, no. 6, pp. 2875–2882, 2021.

[13] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Automat. Contr.*, vol. 31, no. 9, pp. 803–812, 1986.

[14] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Trans. Automat. Contr.*, vol. 61, no. 12, pp. 3936–3947, 2016.

[15] M. Zhu and S. Martínez, "On distributed constrained formation control in operator–vehicle adversarial networks," *Automatica*, vol. 49, no. 12, pp. 3571–3582, 2013.

[16] K. Saulnier, D. Saldana, A. Prorok, G. J. Pappas, and V. Kumar, "Resilient flocking for mobile robot teams," *IEEE Robotics and Automation letters*, vol. 2, no. 2, pp. 1039–1046, 2017.

[17] F. Pasqualetti, A. Bicchi, and F. Bullo, "Consensus computation in unreliable networks: A system theoretic approach," *IEEE Trans. Automat. Contr.*, vol. 57, no. 1, pp. 90–104, 2012.

[18] A. A. Cárdenas, T. Roosta, G. Taban, and S. Sastry, *Cyber Security: Basic Defenses and Attack Trends*, 2008, pp. 73–101.

[19] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *International Conference on Distributed Computing Systems Workshops*, 2008, pp. 495–500.

[20] J. Xiong and K. Jamieson, "Securearray: Improving wifi security with fine-grained physical-layer information," in *Annual International Conference on Mobile Computing and Networking*, ser. MobiCom, 2013.

[21] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93–109, 2015.

[22] S. Gil, S. Kumar, M. Mazumder, D. Katabi, and D. Rus, "Guaranteeing spoof-resilient multi-robot networks," *Autonomous Robots*, vol. 41, pp. 1383–1400, 2017.

[23] C. E. Shannon, "Channels with side information at the transmitter," *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 289–293, 1958.

[24] C. Pippin and H. Christensen, "Trust modeling in multi-robot patrolling," in *IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2014, pp. 59–66.

[25] A. Nedić, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. Automat. Contr.*, vol. 55, no. 4, pp. 922–938, 2010.

[26] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, no. 3, pp. 516–545, 2010.

[27] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Trans. Automat. Contr.*, vol. 57, no. 3, pp. 592–606, 2012.

[28] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Push-sum distributed dual averaging for convex optimization," in *IEEE Conf. Decision Control (CDC)*, 2012, pp. 5453–5458.

[29] ——, "Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning," in *Allert. Conf. Commun. Control Comput.*, 2012, pp. 1543–1550.

[30] K. I. Tsianos and M. G. Rabbat, "Distributed strongly convex optimization," in *Allert. Conf. Commun. Control Comput.*, 2012, pp. 593–600.

[31] S. Magnússon, C. Enyioha, N. Li, C. Fischione, and V. Tarokh, "Convergence of limited communication gradient methods," *IEEE Trans. Automat. Contr.*, vol. 63, no. 5, pp. 1356–1371, 2018.
[32] R. Saha, S. Rini, M. Rao, and A. J. Goldsmith, "Decentralized optimization over noisy, rate-constrained networks: Achieving consensus by communicating differences," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 449–467, 2022.
[33] Y. Tang, J. Zhang, and N. Li, "Distributed zero-order algorithms for nonconvex multiagent optimization," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 1, pp. 269–281, 2021.
[34] W. Abbas, Y. Vorobeychik, and X. Koutsoukos, "Resilient consensus protocol in the presence of trusted nodes," in *International Symposium on Resilient Control Systems (ISRCS)*, 2014, pp. 1–7.
[35] J. S. Baras and X. Liu, "Trust is the cure to distributed consensus with adversaries," in *2019 27th Mediterranean Conference on Control and Automation (MED)*, 2019, pp. 195–202.
[36] C. Zhao, J. He, and Q.-G. Wang, "Resilient distributed optimization algorithm against adversarial attacks," *IEEE Trans. Automat. Contr.*, vol. 65, no. 10, pp. 4308–4315, 2020.
[37] B. Turan, C. A. Uribe, H.-T. Wai, and M. Alizadeh, "Resilient primal-dual optimization algorithms for distributed resource allocation," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 1, pp. 282–294, 2021.
[38] T. Ding, Q. Xu, S. Zhu, and X. Guan, "A convergence-preserving data integrity attack on distributed optimization using local information," in *IEEE Conf. Decision Control (CDC)*, 2020, pp. 3598–3603.
[39] M. Yemini, A. Nedić, A. J. Goldsmith, and S. Gil, "Characterizing trust and resilience in distributed consensus for cyberphysical systems," *IEEE Trans. Robot.*, vol. 38, no. 1, pp. 71–91, 2022.
[40] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research*, vol. 12, no. 7, 2011.
[41] S. Ghadimi and G. Lan, "Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, i: a generic algorithmic framework," *SIAM J. Optim*, vol. 22, no. 4, pp. 1469–1492, 2012.
[42] ——, "Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms," *SIAM J. Optim*, vol. 23, no. 4, pp. 2061–2089, 2013.
[54] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 4, pp. 766–781, 2013.
[43] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems, (Eds.) C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger*, 2013, pp. 315–323.
[44] A. Jalilzadeh, A. Nedić, U. V. Shanbhag, and F. Yousefian, "A variable sample-size stochastic quasi-Newton method for smooth and nonsmooth stochastic convex optimization," in *IEEE Conf. Decision Control (CDC)*, 2018, pp. 4097–4102.
[45] C. Wilson, V. Veeravalli, and A. Nedić, "Adaptive sequential stochastic optimization," *IEEE Trans. Automat. Contr.*, vol. 64, no. 2, pp. 496–509, 2019.
[46] A. Nedić and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," *SIAM J. Optim*, vol. 24, no. 1, pp. 84–107, 2014.
[47] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Mathematical Programming*, vol. 187, pp. 409–457, 2021.
[48] M. O. Sayin, N. D. Vanli, S. S. Kozat, and T. Başar, "Stochastic subgradient algorithms for strongly convex optimization over distributed networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 4, no. 4, pp. 248–260, 2017.
[49] A. Olshevsky, "Linear time average consensus and distributed optimization on fixed graphs," *SIAM Journal on Control and Optimization*, vol. 55, no. 6, pp. 3990–4014, 2017.
[50] B. Charron-Bost, "Geometric bounds for convergence rates of averaging algorithms," 2020. [Online]. Available: https://arxiv.org/pdf/2007.04837.pdf
[51] S. Lacoste-Julien, M. Schmidt, and F. R. Bach, "A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method," 2012. [Online]. Available: https://arxiv.org/pdf/1212.2002.pdf
[52] R. Durrett, *Probability: theory and examples*, 5th ed. Belmont, CA: Duxbury Press, 2019.
[53] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication - computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.
[55] K. Kuwaranancharoen, L. Xin, and S. Sundaram, "Byzantine-resilient distributed optimization of multi-dimensional functions," in *2020 American Control Conference (ACC)*, 2020, pp. 4399–4404.

Supplementary Material: Resilient Distributed Optimization for Multi-Agent Cyberphysical Systems

Michal Yemini, Angelia Nedić, Andrea J. Goldsmith, Stephanie Gil

I. Proof of Theorem 1

Here, we aim to explore the dynamic (2) when $T_0 = 0$, $\mathcal{M} = \emptyset$ and the legitimate agents are aware that all the agents are legitimate. In this scenario the dynamic (2) is equivalent to

$$\begin{aligned} r_i(t) &= \overline{w}_{ii} z_i(t) + \sum_{j \in \mathcal{N}_i \cap \mathcal{L}} \overline{w}_{ij} z_j(t) \\ y_i(t) &= r_i(t) - \gamma(t) \nabla f_i(r_i(t)), \\ \phi_i(t) &= \Pi_{\mathcal{X}}\left(y_i(t)\right) - r_i(t), \\ z_i(t+1) &= r_i(t) + \phi_i(t) = \Pi_{\mathcal{X}}\left(y_i(t)\right). \end{aligned} \tag{1}$$

Denote $Z(t) \triangleq (z_1(t), \ldots, z_{|\mathcal{L}|}(t))$ and $\Phi(t) \triangleq (\phi_1(t), \ldots, \phi_{|\mathcal{L}|}(t))$, then

$$Z(t+1) = \overline{W}_{\mathcal{L}} Z(t) + \Phi(t),$$

where by Lemma 3, $\|\phi_i(t)\| \leq \gamma(t) G$. Thus we can utilize Proposition 1 to upper bound the distance of an agent value from that of the average. Denote

$$\overline{z}(t) \triangleq \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} z_i(t),$$

and

$$g(t) \triangleq \min \left\{ 2\eta, \rho_{\mathcal{L}}^t 2\eta + \frac{\rho_{\mathcal{L}}^{t/2} G \gamma(0)}{1 - \rho_{\mathcal{L}}} + \frac{G \gamma(t/2)}{1 - \rho_{\mathcal{L}}} \right\}. \tag{2}$$

Observe that by the double stochasticity of $\overline{W}_{\mathcal{L}}$ we have that $\overline{z}(t) = \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} r_i(t)$.

Corollary (Supplementary) 1. *For every $t \in |\mathcal{L}|$*

$$\begin{aligned} \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} \|z_i(t) - \overline{z}(t)\| \leq g(t), \text{ and} \\ \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} \|z_i(t) - \overline{z}(t)\|^2 \leq g^2(t). \end{aligned} \tag{3}$$

Proof: This is a direct result of (3) and of Proposition 1 (in the main manuscript), for deterministic $\Delta(t)$ where

$$[\Delta(t)]_i = \phi_i(t),$$

and

$$\delta(t) = \gamma(t) G.$$

■

We are now ready to prove Theorem 1.

Proof of Theorem 1: Under the dynamic (1), for every $i \in \mathcal{L}$ and $t \in |\mathcal{L}|$

$$\begin{aligned}
&\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t+1)-x_{\mathcal{L}}^{\star}\|^2 \\
&= \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|\Pi_{\mathcal{X}}(y_i(t))-x_{\mathcal{L}}^{\star}\|^2 \\
&= \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|\Pi_{\mathcal{X}}(y_i(t))-\Pi_{\mathcal{X}}(x_{\mathcal{L}}^{\star})\|^2 \\
&\le \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|y_i(t)-x_{\mathcal{L}}^{\star}\|^2 \\
&= \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|r_i(t)-\gamma(t)\nabla f_i(r_i(t))-x_{\mathcal{L}}^{\star}\|^2 \\
&= \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|r_i(t)-x_{\mathcal{L}}^{\star}\|^2+\frac{1}{|\mathcal{L}|}\gamma^2(t)\sum_{i\in\mathcal{L}}\|\nabla f_i(r_i(t))\|^2-\frac{2}{|\mathcal{L}|}\gamma(t)\sum_{i\in\mathcal{L}}[\nabla f_i(r_i(t))]^T(r_i(t)-x_{\mathcal{L}}^{\star}).
\end{aligned}$$

The convexity of $\|\cdot\|^2$ and the double stochasticity of $\overline{W}_{\mathcal{L}}$ yields that

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|r_i(t)-x_{\mathcal{L}}^{\star}\|^2 \le \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t)-x_{\mathcal{L}}^{\star}\|,$$

and

$$\|\overline{z}(t)-x_{\mathcal{L}}^{\star}\|^2 = \left\|\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}r_i(t)-x_{\mathcal{L}}^{\star}\right\|^2 \le \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|r_i(t)-x_{\mathcal{L}}^{\star}\|^2.$$

Additionally, we can conclude from Corollary 2 that

$$\frac{1}{|\mathcal{L}|}\gamma^2(t)\sum_{i\in\mathcal{L}}\|\nabla f_i(r_i(t))\|^2 \le \gamma^2(t)G^2.$$

Thus, by the μ-strong convexity of f_i, $\forall\, i \in \mathcal{L}$

$$\begin{aligned}
&-2\frac{\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}[\nabla f_i(r_i(t))]^T(r_i(t)-x_{\mathcal{L}}^{\star}) \\
&= \frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}[\nabla f_i(r_i(t))]^T(x_{\mathcal{L}}^{\star}-r_i(t)) \\
&\le \frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\left[f_i(x_{\mathcal{L}}^{\star})-f_i(r_i(t))-\frac{\mu}{2}\|r_i(t)-x_{\mathcal{L}}^{\star}\|^2\right] \\
&\le \frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}[f_i(x_{\mathcal{L}}^{\star})-f_i(r_i(t))]-\frac{2\gamma(t)\mu}{2}\|\overline{z}(t)-x_{\mathcal{L}}^{\star}\|^2 \\
&= \frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\left[f_i(x_{\mathcal{L}}^{\star})-f_i(\overline{z}(t))-\frac{\mu}{2}\|z_i(t)-x_{\mathcal{L}}^{\star}\|^2\right]+\frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}[f_i(\overline{z}(t))-f_i(r_i(t))]+\frac{\gamma(t)\mu}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t)-\overline{z}(t)\|^2.
\end{aligned}$$

Next, we upper bound the terms in the last line by utilizing the L Lipschitz continuity of ∇f_i:

$$\begin{aligned}
&\frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}[f_i(\overline{z}(t))-f_i(r_i(t))] \\
&\le \frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\left[(\nabla f_i(r_i(t)))^T(\overline{z}(t)-r_i(t))+\frac{L}{2}\|z_i(t)-\overline{z}(t)\|^2\right] \\
&\le \frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\left[G\|r_i(t)-\overline{z}(t)\|+\frac{L}{2}\|z_i(t)-\overline{z}(t)\|^2\right] \\
&\le \frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\left[G\|z_i(t)-\overline{z}(t)\|+\frac{L}{2}\|z_i(t)-\overline{z}(t)\|^2\right].
\end{aligned}$$

It follows that,

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t+1)-x^\star_{\mathcal{L}}\|^2$$
$$\leq \frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t)-x^\star_{\mathcal{L}}\|^2+\gamma^2(t)G^2+\frac{2\gamma(t)}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\left[G\|z_i(t)-\overline{z}(t)\|+\frac{\mu+L}{2}\|z_i(t)-\overline{z}(t)\|^2\right].$$

Denote,

$$h(t)=G^2\gamma(t)+2Gg(t)+(\mu+L)g^2(t).$$

We use Corollary (Supplementary) 1 to deduce that

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t+1)-x^\star_{\mathcal{L}}\|^2\leq(1-\mu\gamma(t))\,\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t)-x^\star_{\mathcal{L}}\|^2+\gamma(t)h(t). \tag{4}$$

Consequently,

$$\lim_{t\to\infty}\|z_i(t)-x^\star_{\mathcal{L}}\|=0,\ \forall\, i\in\mathcal{L}, \tag{5}$$

whenever $\sum_{t=0}^{\infty}\gamma(t)=\infty$ and $\sum_{t=0}^{\infty}\gamma^2(t)<\infty$.

Motivated by [51] we let $\gamma(t)=\frac{2}{\mu(t+2)}$. It follows that

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t+1)-x^\star_{\mathcal{L}}\|^2\leq\frac{t}{t+2}\cdot\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t)-x^\star_{\mathcal{L}}\|^2+\frac{2h(t)}{\mu(t+2)}. \tag{6}$$

Multiplying both sides by $(t+1)(t+2)$ yields the following upper bound

$$(t+1)(t+2)\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t+1)-x^\star_{\mathcal{L}}\|^2$$
$$\leq t(t+1)\cdot\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(t)-x^\star_{\mathcal{L}}\|^2+\frac{2(t+1)h(t)}{\mu}. \tag{7}$$

Summing both sides over the set $t=0,1,\ldots,T-1$ yields the upper bound:

$$\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\|z_i(T)-x^\star_{\mathcal{L}}\|^2\leq\frac{2\sum_{t=0}^{T-1}(t+1)h(t)}{\mu T(T+1)}. \tag{8}$$

To conclude the proof we upper bound the term $\sum_{t=0}^{T-1}(t+1)h(t)$. To this end, we utilize the identity

$$\sum_{t=0}^{\infty}(t+1)x^t=(1-x)^{-2},$$

for all $|x|\in(0,1)$ as follows

$$\sum_{t=0}^{T-1}(t+1)h(t)=\sum_{t=0}^{T-1}(t+1)\left(G^2\gamma(t)+2Gg(t)+(\mu+L)g^2(t)\right). \tag{9}$$

Now,

$$\sum_{t=0}^{T-1}(t+1)G^2\gamma(t)=\sum_{t=0}^{T-1}(t+1)G^2\frac{2}{\mu(t+2)}\leq\frac{2G^2T}{\mu}. \tag{10}$$

Additionally,

$$\begin{aligned}
&\sum_{t=0}^{T-1}(t+1)2Gg(t) \\
&\le 2G\sum_{t=0}^{T-1}(t+1)\left[\rho_{\mathcal{L}}^{t}2\eta + \frac{\rho_{\mathcal{L}}^{t/2}G\gamma(0)}{1-\rho_{\mathcal{L}}} + \frac{G\gamma(t/2)}{1-\rho_{\mathcal{L}}}\right] \\
&\le 4\eta G\sum_{t=0}^{T-1}(t+1)\rho_{\mathcal{L}}^{t} + \frac{2G^2\gamma(0)}{1-\rho_{\mathcal{L}}}\sum_{t=0}^{T-1}(t+1)\rho_{\mathcal{L}}^{t/2} + \frac{4G^2T}{\mu(1-\rho_{\mathcal{L}})} \\
&\le \frac{4\eta G}{1-\rho_{\mathcal{L}}} + \frac{4G^2\gamma(0)}{1-\rho_{\mathcal{L}}}\sum_{t=0}^{T-1}(t+1)\rho_{\mathcal{L}}^{t} + \frac{4G^2T}{\mu(1-\rho_{\mathcal{L}})} \\
&= \frac{4\eta G}{1-\rho_{\mathcal{L}}} + \frac{4G^2}{\mu(1-\rho_{\mathcal{L}})^3} + \frac{4G^2T}{\mu(1-\rho_{\mathcal{L}})},
\end{aligned} \tag{11}$$

where the last inequality follows since

$$\sum_{t=0}^{T-1}(t+1)\rho_{\mathcal{L}}^{t/2} \le \sum_{t=0}^{T-1}(2t+1)\rho_{\mathcal{L}}^{t} + \sum_{t=0}^{T-1}(2t)\rho_{\mathcal{L}}^{t} \le 2\sum_{t=0}^{T-1}(t+1)\rho_{\mathcal{L}}^{t}. \tag{12}$$

Finally, since $|\rho_{\mathcal{L}}| < 1$ then $(1-\rho_{\mathcal{L}}^{x})^{-1} \le (1-\rho_{\mathcal{L}})^{-1}$ for all $x \ge 1$. It follows that

$$\begin{aligned}
&(\mu+L)\sum_{t=0}^{T-1}(t+1)g^2(t) \\
&\le (\mu+L)\sum_{t=0}^{T-1}(t+1)\left[\rho_{\mathcal{L}}^{t}2\eta + \frac{\rho_{\mathcal{L}}^{t/2}G\gamma(0)}{1-\rho_{\mathcal{L}}} + \frac{G\gamma(t/2)}{1-\rho_{\mathcal{L}}}\right]^2 \\
&\le \frac{4\eta^2(\mu+L)}{(1-\rho_{\mathcal{L}})^2} + \frac{2G^2(\mu+L)}{\mu^2(1-\rho_{\mathcal{L}})^4} + \frac{16(\mu+L)G^2\ln\left(\frac{T+2}{2}\right)}{\mu^2(1-\rho_{\mathcal{L}})^2} + \frac{8G\eta(\mu+L)}{\mu(1-\rho_{\mathcal{L}})^3} + \frac{16G\eta(\mu+L)}{\mu(1-\rho_{\mathcal{L}})^2} + \frac{16G^2(\mu+L)}{\mu^2(1-\rho_{\mathcal{L}})^2} \\
&= \frac{4(\mu+L)(\mu\eta+2G)^2}{\mu^2(1-\rho_{\mathcal{L}})^2} + \frac{2G^2(\mu+L)}{\mu^2(1-\rho_{\mathcal{L}})^4} + \frac{8G\eta(\mu+L)}{\mu(1-\rho_{\mathcal{L}})^3} + \frac{16(\mu+L)G^2}{\mu^2(1-\rho_{\mathcal{L}})^2}\ln\left(\frac{T+2}{2}\right).
\end{aligned} \tag{13}$$

Finally, by utilizing the notation (14) we can conclude that

$$\sum_{t=0}^{T-1}(t+1)h(t) \le 2\overline{h}(T).$$

■