Byzantine-Resilient Resource Allocation Over Decentralized Networks

Runhua Wang , *Student Member, IEEE*, Yaohua Liu, and Qing Ling , *Senior Member, IEEE*

Abstract—This paper considers the resource allocation problem over a decentralized multi-agent network and at presence of Byzantine agents. Compared with its centralized counterpart, a decentralized algorithm enjoys better scalability when the network is large-scale, but is more vulnerable when some of the agents are malicious and send wrong messages during the optimization process. We utilize the classical Byzantine attack model to describe these malicious actions, and propose a novel Byzantine-resilient decentralized resource allocation algorithm, abbreviated as BREDA. At each iteration of BREDA, each honest agent receives messages from its neighbors, uses coordinate-wise trimmed mean (CTM) to aggregate these messages, and then updates its local primal and dual variables with gradient descent and ascent, respectively. Theoretical analysis indicates that BREDA converges to a neighborhood of an optimal solution. Numerical experiments demonstrate the resilience of BREDA to various Byzantine attacks.

Index Terms—Resource allocation, decentralized multi-agent network, Byzantine-resilience.

I. INTRODUCTION

RESOURCE allocation, which aims at assigning a limited amount of resources among a group of users (also known as agents) to maximize their utility or minimize their cost, is one of the key issues in network optimization. Resource allocation has been extensively investigated in recent decades, and found broad applications in various fields, such as smart grids, wireless networks, to name a few.

A. Related Works

Resource Allocation Algorithms: In general, resource allocation over a network can be modeled as maximizing the

Runhua Wang and Qing Ling are with the School of Computer Science and Engineering and Guangdong Provincial Key Laboratory of Computational Science, Sun Yat-Sen University, Guangdong 510006, China, and also with the Pazhou Lab, Guangdong 510300, China (e-mail: wangrh37@mail2.sysu.edu.cn; lingqing556@mail.sysu.edu.cn).

Yaohua Liu is with the School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: yaoh.liu@nuist.edu.cn).

average utility or minimizing the average cost of the agents, subject to global and local resource constraints [2]. Centralized algorithms require a master node to coordinate all the agents, and are unscalable to network size [3], [4]. Hence, decentralized algorithms that rely on coordination among neighboring agents become attractive alternatives. The main challenge in the decentralized algorithms is to handle the global resource constraints that couple all the agents. For instance, [5] and [6] propose decentralized, weighted gradient methods in which the global resource constraints are always satisfied under proper initializations. The underlying communication graph is assumed to be fixed in [5], and dynamic in [6]. Continuous-time and discrete-time primal-dual algorithms are developed in [7], [8], [9] and [10], [11], [12], respectively. Among the discrete-time primal-dual algorithms in which this paper is interested, [10] develops a push-pull algorithm over a fixed but unbalanced network, [11] exactly solves the primal subproblems and performs dual gradient ascent over a dynamic network, and [12] proposes a primal-dual decomposition algorithm, also over a dynamic network. Several techniques are developed to achieve faster convergence. One is using both the gradients and Hessians of the cost functions [13], [14], and another is using the gradients of the cost functions as well as the past iterates when computing the future ones [15]. An asynchronous resource allocation algorithm is proposed in [16], utilizing delayed gradient information to carry out updates and hence suitable for heterogeneous networks. Non-smooth resource allocation problems are investigated in [17], [18]. Online convex optimization approaches have been proposed in [19], [20], [21] to solve resource allocation problems with time-varying and coupled global resource constraints.

Byzantine-Resilience: Most of the aforementioned works assume that all the agents are reliable and strictly follow the algorithms. However, in the real world, some of the agents might be unreliable in either computing or communicating, and even be malfunctioning. These agents deviate from the expected optimization process and send wrong messages to their neighbors, yielding biased results. We characterize these behaviors with the classical Byzantine attack model [22], [23], in which the malicious agents (also called the Byzantine agents) can collude and arbitrarily modify the messages sent to their neighbors. Such an attack model imposes no restrictions on the Byzantine agents and is worst-case. To the best of our knowledge, so far there is no work considering Byzantine attacks in decentralized resource allocation. The work of [24] takes observation noise and communication uncertainties into

consideration, and designs a stochastic approximation algorithm to solve the decentralized resource allocation problem. Nevertheless, the noise and uncertainties in [24] are not worst-case, in contrast to the Byzantine attack model studied here. The work of [3] investigates Byzantine-resilient resource allocation, but the proposed algorithm is centralized.

Byzantine-resilient optimization is now a popular topic in federated learning, where a master node coordinates the learning process of multiple geographically distributed agents [25], [26], [27]. It has also attracted attention in decentralized consensus optimization, where decentralized agents collaboratively minimize the average cost subject to consensus constraints – namely, all the local variables must reach the same value [28], [29], [30], [31], [32], [33], [34], [35]. Below we briefly survey the existing Byzantine-resilient decentralized consensus optimization algorithms. The works of [28] and [29] propose resilient algorithms based on trimmed mean aggregation. At each iteration each honest agent chooses a coordinate, uses trimmed mean on this coordinate to aggregate the local iterates received from its neighbors, followed by coordinate gradient descent to update its own local iterate. In the trimmed mean, a given number of b largest and b smallest values are trimmed, and the rest are averaged. However, it is inefficient to screen only one coordinate of the received local iterates in each iteration. To addresses this issue, [30] devises BRIDGE, which allows each honest agent to aggregate the received local iterates in all coordinates at every iteration with coordinate-wise trimmed mean (CTM) and then perform gradient descent. A variant of CTM is proposed in [31], where for each coordinate, each honest agent trims up to b received values larger than its own value, and up to b received values smaller than its own value. In [32], each honest agent discards a given number of received local iterates that increase its own cost. Total variation regularization is adopted in [33], [34] to drive the local iterates of the honest agents to be close, but allows to tolerate Byzantine attacks. The idea of total variation regularization has been extended to decentralized stochastic consensus optimization [35]. Despite the success of Byzantine-resilient decentralized consensus optimization algorithms, the ideas therein cannot be directly applied to Byzantine-resilient decentralized resource allocation. The key is that, in the latter, the optimal local variables of the honest agents are not necessarily consensual. Therefore, naively applying the existing decentralized Byzantine-resilient aggregation rules such as trimmed mean, CTM and total variation regularization no longer works.

B. Our Contributions

We investigate the almost untouched territory of Byzantine-resilient decentralized resource allocation. The Byzantine agents send wrong messages to their neighbors so as to bias the optimization process. For example, they can collude to manipulate the optimization process so that the honest agents are allocated with less resources than needed. As we have indicated above, directly applying the decentralized Byzantine-resilient aggregation rules is infeasible here. To address these issues, we make the following contributions.

C1) We propose a novel primal-dual Byzantine-resilient decentralized resource allocation (BREDA) algorithm, where

the primal and dual variables are updated locally. We introduce a local auxiliary variable to each agent for approximating the average amount of required resources, which is used in updating the local dual variable. A first-order decentralized dynamic average consensus method equipped with CTM aggregation is then applied to update the local auxiliary variable in a Byzantine-resilient manner.

C2) We prove that BREDA converges to a neighborhood of the saddle point of a regularized Lagrangian function. We also conduct extensive numerical experiments on decentralized resource allocation problems. The experimental results show the resilience of BREDA to various Byzantine attacks.

C. Paper Organization and Notations

The rest of this paper is organized as follows. Section II describes the Byzantine-resilient decentralized resource allocation problem. A primal-dual decentralized resource allocation (DRA) algorithm is developed in Section III given that the Byzantine agents are absent, and its failure under Byzantine attacks is investigated. Section IV proposes BREDA, a Byzantine-resilient decentralized resource allocation algorithm. Section V establishes the convergence of DRA and BREDA. Numerical experiments are conducted in Section VI and conclusions are drawn in Section VII.

We use $(\cdot)^T$ to stand for matrix transposition, $\|\cdot\|$ for the ℓ_2-norm of a vector and the spectral norm of a matrix, $\rho(\cdot)$ for the spectral radius of a matrix, $\Pi_{\mathcal{C}}[\cdot]$ for the projection on a set $\mathcal{C}$, $\otimes$ for Kronecker product, $\langle \cdot, \cdot \rangle$ for inner product of vectors. We denote $\boldsymbol{I} \in \mathbb{R}^{D \times D}$ as an identity matrix and $\mathbf{1} \in \mathbb{R}^{|\mathcal{H}|}$ as an all-one column vector, where D is the dimension of local optimization variables and $|\mathcal{H}|$ is the number of honest agents.

II. PROBLEM FORMULATION

We consider a connected network of N agents, modeled as an undirected, fixed graph $\mathcal{G}(\mathcal{J}, \mathcal{E})$. The set of vertices $\mathcal{J} := \{1, \ldots, N\}$ represents the agents in the network and the set of edges $\mathcal{E}$ represents the communication links between the agents. If $(i, j) \in \mathcal{E}$, then agents i and j are neighbors and can communicate with each other. Let $\mathcal{N}_i = \{j \mid (i, j) \in \mathcal{E}\}$ be the set of neighbors of agent i. Resources are allocated among the agents, and our objective is to find an optimal allocation that minimizes the average cost under resource constraints. This decentralized resource allocation problem is given by

$$\min_{\boldsymbol{\theta}} \quad f(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{\theta}_i),$$

$$\text{s.t.} \quad \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\theta}_i \leq \boldsymbol{s}, \quad \boldsymbol{\theta}_i \in \boldsymbol{C}_i, \forall i \in \mathcal{J}. \tag{1}$$

Therein, $\boldsymbol{\theta}_i \in \mathbb{R}^D$ is the local optimization variable of agent i, representing the amount of allocated resources, and $f_i(\cdot)$ is the continuously differentiable and convex cost function of agent i. The average amount of allocated resources $\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\theta}_i$ is upper bounded by a constant vector $\boldsymbol{s} \in \mathbb{R}^D$ such that $N\boldsymbol{s}$ corresponds to the total amount of resources. Each $\boldsymbol{\theta}_i$ is also confined to a

bounded, convex set C_i. For simplicity, we collect all the local optimization variables in a vector $\boldsymbol{\theta} := [\boldsymbol{\theta}_1; \cdots; \boldsymbol{\theta}_N] \in \mathbb{R}^{ND}$.

Example 1 (Electric vehicle charging, adapted from [3]): We consider an electric vehicle charging problem with 100 electric vehicles that are connected over a decentralized network. The aim is to obtain optimal charging powers that minimize the average cost of 100 electric vehicles under charging power constraints. For electric vehicle i, $\boldsymbol{\theta}_i \in \mathbb{R}$ is its charging power and $f_i(\boldsymbol{\theta}_i) = -\log(1 + \boldsymbol{\theta}_i)$ is its cost function – here we assume all cost functions are the same just for simplicity. Due to the limit of electricity, the total charging power is upper-bounded by 1500 kW and the average is upper-bounded by 15 kW. Different electric vehicles have different acceptable maximum charging powers. We assume half of of them have maximum charging powers of 20 kW, and half have 25 kW. Therefore, the problem can be written as

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad & f(\boldsymbol{\theta}) := \frac{1}{100} \sum_{i=1}^{100} f_i(\boldsymbol{\theta}_i), \\
\text{s.t.} \quad & \frac{1}{100} \sum_{i=1}^{100} \boldsymbol{\theta}_i \le 15, \\
& 0 \le \boldsymbol{\theta}_i \le 20, \quad i = 1, \ldots, 50, \\
& 0 \le \boldsymbol{\theta}_i \le 25, \quad i = 51, \ldots, 100,
\end{aligned}
\tag{2}
$$

which falls into the form of (1). Since all the cost functions are the same, the optimal charging powers in this simple example are $\boldsymbol{\theta}_i^* = 15$ kW for all electric vehicles i.

To obtain an optimal allocation to (1), the agents must communicate with their neighbors and collaboratively optimize their local optimization variables. However, not all the agents in the decentralized network are honest. Some of them might be malfunctioning or even malicious. These agents can arbitrarily deviate from the expected optimization process and send wrong messages to their neighbors. We term them as Byzantine agents [22], [23]. Let $\mathcal{B}$ be the set of Byzantine agents and $\mathcal{H} := \mathcal{J} \setminus \mathcal{B}$ be the set of honest agents. The numbers of Byzantine and regular agents are denoted as $|\mathcal{B}|$ and $|\mathcal{H}|$, respectively. Note that the number and identities of the Byzantine agents are unknown in prior.

The objectives of Byzantine agents may vary in different scenarios. One of them is that the Byzantine agents simply disturb the ideal optimization process by sending random messages. Another is that the Byzantine agents collude to send crafted messages such that they can occupy more resources than needed. In turn, the honest agents shall be allocated with less resources. We consider the worst case in which the Byzantine agents are arbitrarily malicious. Therefore, an oracle goal for the honest agents is to solve

$$
\begin{aligned}
\min_{\boldsymbol{\Theta}} \quad & f(\boldsymbol{\Theta}) := \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} f_i(\boldsymbol{\theta}_i), \\
\text{s.t.} \quad & \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \boldsymbol{\theta}_i \le \boldsymbol{s}, \quad \boldsymbol{\theta}_i \in \boldsymbol{C}_i, \forall i \in \mathcal{H},
\end{aligned}
\tag{3}
$$

where $\boldsymbol{\Theta} \in \mathcal{R}^{|\mathcal{H}|D}$ collects all $\boldsymbol{\theta}_i$ of the regular agents $i \in \mathcal{H}$.

However, solving (3) is nontrivial when the number and identities of Byzantine agents are unknown. In this paper, we aim at developing a Byzantine-resilient decentralized resource allocation algorithm that approximately solves (3).

III. DECENTRALIZED RESOURCE ALLOCATION WITHOUT BYZANTINE ATTACKS

Algorithm Development: When there are no Byzantine attacks, we introduce a decentralized resource allocation (DRA) algorithm to solve (1) as the baseline.

Define the regularized Lagrangian function of (1) as

$$
\mathcal{L}_v(\boldsymbol{\theta}; \hat{\boldsymbol{\lambda}}) := \mathcal{L}(\boldsymbol{\theta}; \hat{\boldsymbol{\lambda}}) + \frac{v}{2} \sum_{i=1}^{N} \|\boldsymbol{\theta}_i\|^2 - \frac{v}{2} \|\hat{\boldsymbol{\lambda}}\|^2,
\tag{4}
$$

where the Lagrangian function of (1) is

$$
\mathcal{L}(\boldsymbol{\theta}; \hat{\boldsymbol{\lambda}}) := \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{\theta}_i) + \left\langle \hat{\boldsymbol{\lambda}}, \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\theta}_i - \boldsymbol{s} \right\rangle,
\tag{5}
$$

$\hat{\boldsymbol{\lambda}} \in \mathbb{R}^D$ is the dual variable, and $v > 0$ is a regularization parameter. Both functions are defined on $\{\boldsymbol{\theta}_i \in \boldsymbol{C}_i, \forall i = 1, \ldots N, \hat{\boldsymbol{\lambda}} \in \mathbb{R}_+^D\}$. Thus, $\mathcal{L}_v(\cdot)$ is v-strongly convex and v-strongly concave in $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\lambda}}$, respectively.

Remark 1: Adding strongly convex regularization terms to cost functions is a classical technique, often called as Nesterov smoothing [36]. With the regularization terms $\frac{v}{2} \sum_{i=1}^{N} \|\boldsymbol{\theta}_i\|^2$ and $\frac{v}{2} \|\hat{\boldsymbol{\lambda}}\|^2$, the primal and dual functions respectively become strongly convex and strongly concave, which facilitates the convergence analysis. It has been proven in [4], [37] that the saddle points of (4) and (5) are within a bounded neighborhood in terms of the primal and dual function values by choosing an appropriate v. Therefore, reaching a saddle point of (4) means approximately solving (1).

Let k be the iteration index and $\gamma^k > 0$ be the corresponding step size. To approximately solve (1), at iteration k, each agent i performs projected gradient descent on the primal variable and projected gradient ascent on the dual variable as

$$
\boldsymbol{\theta}_i^{k+1} = \Pi_{\boldsymbol{C}_i} \left[\boldsymbol{\theta}_i^k - \gamma^k \nabla_{\boldsymbol{\theta}_i} \mathcal{L}_v(\boldsymbol{\theta}^k; \hat{\boldsymbol{\lambda}}^k) \right],
\tag{6}
$$

$$
\hat{\boldsymbol{\lambda}}^{k+1} = \Pi_{\hat{\boldsymbol{u}}} \left[\hat{\boldsymbol{\lambda}}^k + \gamma^k \nabla_{\hat{\boldsymbol{\lambda}}} \mathcal{L}_v(\boldsymbol{\theta}^k; \hat{\boldsymbol{\lambda}}^k) \right],
\tag{7}
$$

where $\hat{\boldsymbol{u}}$ is a closed, convex and bounded set within the non-negative orthant. The gradients related to the primal and dual variables are respectively given by

$$
\nabla_{\boldsymbol{\theta}_i} \mathcal{L}_v(\boldsymbol{\theta}^k; \hat{\boldsymbol{\lambda}}^k) = \frac{1}{N}(\nabla_{\boldsymbol{\theta}_i} f_i(\boldsymbol{\theta}_i^k) + \hat{\boldsymbol{\lambda}}^k) + v\boldsymbol{\theta}_i^k,
\tag{8}
$$

$$
\nabla_{\hat{\boldsymbol{\lambda}}} \mathcal{L}_v(\boldsymbol{\theta}^k; \hat{\boldsymbol{\lambda}}^k) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\theta}_i^k - \boldsymbol{s} - v \hat{\boldsymbol{\lambda}}^k.
\tag{9}
$$

However, (6) and (7) cannot be implemented in a decentralized manner, because the dual variable is global, and the computation of dual gradient in (9) involves the average of all the local primal variables $\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\theta}_i^k$. To address these issues, we first let each agent i hold a local dual variable $\boldsymbol{\lambda}_i \in \mathbb{R}^D$. Then, we assign each agent i an auxiliary variable $\boldsymbol{x}_i \in \mathbb{R}^D$ to track the

---

Algorithm 1: Decentralized Resource Allocation (DRA).

Initialization: Agent i sets $\boldsymbol{\theta}_i^0 = \boldsymbol{x}_i^0$ and $\boldsymbol{\lambda}_i^0$.
for $k = 0, 1, 2, \ldots$ **do**
  [Communication Stage]:
  Agent i broadcasts $\boldsymbol{x}_i^k$ to $j \in \mathcal{N}_i$.
  [Computation Stage]:
  Agent i computes $\boldsymbol{\theta}_i^{k+1}$ according to (11).
  Agent i computes $\boldsymbol{\lambda}_i^{k+1}$ according to (12).
  Agent i computes $\boldsymbol{x}_i^{k+1}$ according to (13).
end for

---

value of $\frac{1}{N}\sum_{i=1}^N \boldsymbol{\theta}_i$. Motivated by the first-order decentralized dynamic average consensus method proposed in [38], $\boldsymbol{x}_i$ tracks $\frac{1}{N}\sum_{i=1}^N \boldsymbol{\theta}_i$ according to

$$\boldsymbol{x}_i^{k+1} = \sum_{j \in \mathcal{N}_i \cup \{i\}} \boldsymbol{E}_{ij}\boldsymbol{x}_j^k + \triangle\boldsymbol{\theta}_i^{k+1}, \tag{10}$$

where $\boldsymbol{E}_{ij}$ is the weight of agents i and j, and $\triangle\boldsymbol{\theta}_i^{k+1} := \boldsymbol{\theta}_i^{k+1} - \boldsymbol{\theta}_i^k$. Collecting the weights in a doubly stochastic matrix $\boldsymbol{E} = [\boldsymbol{E}_{ij}] \in \mathbb{R}^{N \times N}$, we require that $\boldsymbol{E}_{ij} > 0$ if and only if $(i, j) \in \mathcal{E}$ or $i = j$.

With these introduced local variables $\boldsymbol{\lambda}_i$ and $\boldsymbol{x}_i$, in DRA each agent i modifies (6) and (7) to

$$\boldsymbol{\theta}_i^{k+1} = \Pi_{\boldsymbol{C}_i}\left[\boldsymbol{\theta}_i^k - \gamma^k \nabla_{\boldsymbol{\theta}_i}\mathcal{L}_v\left(\boldsymbol{\theta}^k; \boldsymbol{\lambda}_i^k\right)\right] \tag{11}$$

$$= \Pi_{\boldsymbol{C}_i}\left[\boldsymbol{\theta}_i^k - \gamma^k\left(\frac{1}{N}\left(\nabla_{\boldsymbol{\theta}_i}f_i\left(\boldsymbol{\theta}_i^k\right) + \boldsymbol{\lambda}_i^k\right) + v\boldsymbol{\theta}_i^k\right)\right],$$

$$\boldsymbol{\lambda}_i^{k+1} = \Pi_{\boldsymbol{U}_i}\left[\boldsymbol{\lambda}_i^k + \gamma^k\nabla_{\boldsymbol{\lambda}_i}\mathcal{L}_v\left(\boldsymbol{\theta}^k; \boldsymbol{\lambda}_i^k\right)\mid_{\frac{1}{N}\sum_{i=1}^N \boldsymbol{\theta}_i^k = \boldsymbol{x}_i^k}\right] \tag{12}$$

$$= \Pi_{\boldsymbol{U}_i}\left[\boldsymbol{\lambda}_i^k + \gamma^k\left(\boldsymbol{x}_i^k - \boldsymbol{s} - v\boldsymbol{\lambda}_i^k\right)\right],$$

$$\boldsymbol{x}_i^{k+1} = \sum_{j \in \mathcal{N}_i \cup \{i\}} \boldsymbol{E}_{ij}\boldsymbol{x}_j^k + \triangle\boldsymbol{\theta}_i^{k+1}, \tag{13}$$

where $\boldsymbol{U}_i = \hat{\boldsymbol{u}}$ for all $i = 1, \ldots, N$. The updates of DRA are summarized in Algorithm 1.

Remark 2: We use the same regularization parameter v in (4) for both the primal and dual variables but they can be set as different values if necessary. Similarly, the primal step size in (11) and the dual step size in (12) can be different too.

Remark 3: Although DRA is developed to solve the decentralized resource allocation problem with the inequality resource constraints, it is also applicable when the resource constraints are equality, as long as we allow the projections of the dual variables in (12) to be negative.

Remark 4: As we will see later, with the introduction of the auxiliary variables $\boldsymbol{x}_i$, Byzantine-resilient aggregation becomes reasonable because the auxiliary variables are expected to eventually reach a consensus. We emphasize that DRA is not one of the main contributions of this paper, but instead a necessary intermediate for developing the ensuing Byzantine-resilient algorithm BREDA.

Failure of DRA under Byzantine Attacks: As we will demonstrate with numerical experiments, when all the agents are

honest, DRA is able to approximately solve (1) in a fully decentralized manner. However, applying DRA at the presence of Byzantine attacks will lead to undesirable, catastrophic outcomes. In iteration k of DRA, each agent i updates $\boldsymbol{x}_i^{k+1}$ to track $\frac{1}{N}\sum_{i=1}^N \boldsymbol{\theta}_i^{k+1}$ based on the messages $\boldsymbol{x}_j^k$ from its neighbors. An honest agent $j \in \mathcal{H}$ shall broadcast the true $\boldsymbol{x}_j^k$ to its neighbors. Yet, a Byzantine agent $j \in \mathcal{B}$ broadcasts an arbitrary, wrong message $\boldsymbol{q}_j^k \in \mathbb{R}^D$, instead of $\boldsymbol{x}_j^k$. We formally define the message sent by agent j as

$$\boldsymbol{m}_j^k = \begin{cases} \boldsymbol{x}_j^k, & \text{if } j \in \mathcal{H}, \\ \boldsymbol{q}_j^k, & \text{if } j \in \mathcal{B}. \end{cases} \tag{14}$$

The existence of the wrong messages will prevent the honest agents from obtaining desirable resource allocation strategies. For example, if Byzantine agent $j \in \mathcal{B}$ sends to its honest neighbor i a wrong message $\boldsymbol{q}_j^k$, which is much larger than the average amount of resources $\boldsymbol{s}$, then $\boldsymbol{x}_i^{k+1}$ computed by honest agent i from (13) is larger than $\boldsymbol{s}$. Therefore, its local dual variable $\boldsymbol{\lambda}_i^{k+2}$ will be larger than normal according to (12), and consequently, its primal variable $\boldsymbol{\theta}_i^{k+3}$ will be smaller than normal according to (11).

Example 2 (Electric vehicle charging under Byzantine attacks): For the charging task in Example 1, the electric vehicles communicate with their neighbors and collaboratively optimize their charging powers. However, some electric vehicles may be attacked by hackers and send wrong messages to their neighbors. We call them as Byzantine vehicles. For example, assume that electric vehicle 100 is Byzantine, i.e., $\mathcal{B} = \{100\}$ and $\mathcal{H} = \{1, \ldots, 99\}$. The Byzantine vehicle sends a wrong message $\boldsymbol{q}_{100}^k = 25\,\text{kW}$ to its neighbors at any time k. These wrong messages will mislead the honest vehicles to obtain wrong estimates of the global average charging power, which are larger than the upper bound 15 kW. To meet the average charging power constraint, the honest vehicles have to reduce their charging powers $\boldsymbol{\theta}_i^k$ so that they are less than the optimal ones, i.e., $\boldsymbol{\theta}_i^k < \boldsymbol{\theta}_i^* = 15\,kW, i = 1, \ldots, 99$.

IV. BYZANTINE-RESILIENT DECENTRALIZED RESOURCE ALLOCATION

As we have discussed in Section III, when there exist Byzantine agents, their wrong messages will affect the optimization process. To address this issue, we propose a novel BREDA algorithm. Instead of directly aggregating the received messages with weighted average in DRA, BREDA adopts the coordinate-wise trimmed mean (CTM) [30], which is able to tolerate Byzantine attacks, to aggregate the received messages. To be specific, CTM requires to roughly estimate an upper bound b for the number of Byzantine neighbors of each honest agent. Then for each coordinate, each honest agent eliminates the smallest b and the largest b values in the received messages and average the remaining values for aggregation. This way, at iteration k and for coordinate d, honest agent i separates its set of neighbors $\mathcal{N}_i$ into three subsets, as

$$\mathcal{N}_{i,d}^{k^{\min}} = \underset{\mathcal{X}:\{\mathcal{X} \in \mathcal{N}_i, |\mathcal{X}|=b\}}{\arg\min} \sum_{j \in \mathcal{X}}\left[\boldsymbol{m}_j^k\right]_d, \tag{15}$$

Algorithm 2: Byzantine-Resilient Decentralized Resource Allocation (BREDA).

Initialization: Agent i sets $\boldsymbol{\theta}_i^0 = \boldsymbol{x}_i^0$ and $\boldsymbol{\lambda}_i^0$.
for $k = 0, 1, 2, \ldots$ **do**
   [Communication stage]:
   Honest agent $i \in \mathcal{H}$ broadcasts $\boldsymbol{x}_i^k$ to $j \in \mathcal{N}_i$.
   Byzantine agent $i \in \mathcal{B}$ broadcasts $\boldsymbol{q}_i^k$ to $j \in \mathcal{N}_i$.
   [Computation stage]:
   Honest agent $i \in \mathcal{H}$ computes $\boldsymbol{\theta}_i^{k+1}$ according to (18).
   Honest agent $i \in \mathcal{H}$ computes $\boldsymbol{\lambda}_i^{k+1}$ according to (19).
   Honest agent $i \in \mathcal{H}$ separates $\mathcal{N}_i$ into three groups.
   Honest agent $i \in \mathcal{H}$ computes $\boldsymbol{x}_i^{k+1}$ according to (20).
end for

$$\mathcal{N}_{i,d}^{k^{\max}} = \underset{\mathcal{X}:\{\mathcal{X} \in \mathcal{N}_i, |\mathcal{X}| = b\}}{\arg\max} \sum_{j \in \mathcal{X}} \left[\boldsymbol{m}_j^k\right]_d, \quad (16)$$

$$\mathcal{N}_{i,d}^k = \mathcal{N}_i \setminus \mathcal{N}_{i,d}^{k^{\min}} \setminus \mathcal{N}_{i,d}^{k^{\max}}. \quad (17)$$

With these definitions, for each regular agent $i \in \mathcal{H}$, the updates of BREDA are given by

$$\boldsymbol{\theta}_i^{k+1} = \Pi_{\boldsymbol{C}_i}\left[\boldsymbol{\theta}_i^k - \gamma^k\left(\frac{1}{N}\left(\nabla_{\boldsymbol{\theta}_i}f_i\left(\boldsymbol{\theta}_i^k\right) + \boldsymbol{\lambda}_i^k\right) + v\boldsymbol{\theta}_i^k\right)\right], \quad (18)$$

$$\boldsymbol{\lambda}_i^{k+1} = \Pi_{\boldsymbol{U}_i}\left[\boldsymbol{\lambda}_i^k + \gamma^k\left(\boldsymbol{x}_i^k - \boldsymbol{s} - v\boldsymbol{\lambda}_i^k\right)\right], \quad (19)$$

$$\left[\boldsymbol{x}_i^{k+1}\right]_d = \frac{1}{|\mathcal{N}_i| - 2b + 1}\sum_{j \in \mathcal{N}_{i,d}^k \cup \{i\}}\left[\boldsymbol{m}_j^k\right]_d + \left[\triangle\boldsymbol{\theta}_i^{k+1}\right]_d, \quad (20)$$

in which (20) applies to all dimensions $d = 1, \ldots, D$. We summarize the updates of BREDA in Algorithm 2.

We will show with numerical experiments that in BREDA, the local optimization variables of the honest agents are close to their Byzantine-free optima. However, the Byzantine agents are still able to manipulate their local optimization variables, asking for more resources than needed. Therefore, after the agents reach their resource allocation strategies, the resource provider can collect all the local optimization variables, satisfy those with the smallest resource requirements, and partially satisfy those with the largest resource requirements. Although this extra step needs global information collection, the incurred communication burden is low and the local cost functions are kept private.

Remark 5: BREDA focuses on solving the Byzantine-resilient resource allocation problem. Nevertheless, the underlying idea of designing a Byzantine-resilient primal-dual type solver can be extended to solving a general decentralized optimization problem with global constraints, in the form of (3).

V. CONVERGENCE ANALYSIS

This section analyzes the convergence of DRA and BREDA. We have already defined $\boldsymbol{\theta} := [\boldsymbol{\theta}_1; \cdots; \boldsymbol{\theta}_N] \in \mathbb{R}^{ND}$ and $\boldsymbol{\Theta} \in \mathcal{R}^{|\mathcal{H}|D}$ to collect all $\boldsymbol{\theta}_i$ of the regular agents $i \in \mathcal{H}$. Likewise, define $\boldsymbol{\lambda} := [\boldsymbol{\lambda}_1; \cdots; \boldsymbol{\lambda}_N] \in \mathbb{R}^{ND}$ and $\boldsymbol{\lambda}$ to collect all $\boldsymbol{\lambda}_i$ of

the regular agents $i \in \mathcal{H}$; define $\boldsymbol{x} := [\boldsymbol{x}_1; \cdots; \boldsymbol{x}_N] \in \mathbb{R}^{ND}$ and $\boldsymbol{X}$ to collect all $\boldsymbol{x}_i$ of the regular agents $i \in \mathcal{H}$. Define $\boldsymbol{A} := [\boldsymbol{I}_1, \ldots, \boldsymbol{I}_N] \in \mathbb{R}^{D \times ND}$ where $\boldsymbol{I}_i = \boldsymbol{I}$ for any $i = 1, \ldots, N$. Therefore, $\bar{\boldsymbol{\theta}} := \frac{1}{N}\boldsymbol{A}\boldsymbol{\theta} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\theta}_i$. Likewise, define $\tilde{\boldsymbol{A}} := [\boldsymbol{I}_1, \ldots, \boldsymbol{I}_{|\mathcal{H}|}] \in \mathbb{R}^{D \times |\mathcal{H}|D}$ such that $\bar{\boldsymbol{\Theta}} := \frac{1}{|\mathcal{H}|}\tilde{\boldsymbol{A}}\boldsymbol{\Theta} = \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} \boldsymbol{\theta}_i$. Let $\boldsymbol{C}$ be the Cartesian product of $\boldsymbol{C}_i$ for all $i = 1, \ldots, N$ and $\tilde{\boldsymbol{C}}$ be the Cartesian product of $\boldsymbol{C}_i$ for all $i \in \mathcal{H}$. Let $\boldsymbol{U}$ be the Cartesian product of $\boldsymbol{U}_i$ for all $i = 1, \ldots, N$ and $\tilde{\boldsymbol{U}}$ be the Cartesian product of $\boldsymbol{U}_i$ for all $i \in \mathcal{H}$. Define the expanded weight matrix $\boldsymbol{W} := \boldsymbol{E} \otimes \boldsymbol{I}$.

Following these notations, we can rewrite the DRA updates (11), (12) and (13) to a compact form

$$\boldsymbol{\theta}^{k+1} = \Pi_{\boldsymbol{C}}\left[\boldsymbol{\theta}^k - \gamma^k\left(\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}^k) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\theta}^k\right)\right], \quad (21)$$

$$\boldsymbol{\lambda}^{k+1} = \Pi_{\boldsymbol{U}}\left[\boldsymbol{\lambda}^k + \gamma^k\left(\boldsymbol{x}^k - \boldsymbol{A}^T\boldsymbol{s} - v\boldsymbol{\lambda}^k\right)\right], \quad (22)$$

$$\boldsymbol{x}^{k+1} = \boldsymbol{W}\boldsymbol{x}^k + \triangle\boldsymbol{\theta}^{k+1}. \quad (23)$$

We make the following assumptions on the cost functions, the constraint sets and the underlying communication graph.

Assumption 1: The sets $\boldsymbol{C}_i$ are closed, convex and bounded. The sets $\boldsymbol{U}_i$ are closed, convex, bounded, and within the non-negative orthant. The feasible sets of (1) and (3) are non-empty. The functions $f_i(\boldsymbol{\theta}_i)$ are convex. The gradient of $f_i(\boldsymbol{\theta}_i)$ is Lipschitz continuous with constant L.

Assumption 2: The undirected graph $\mathcal{G}(\mathcal{J}, \mathcal{E})$ is connected. The weight matrix $\boldsymbol{E}$ is doubly stochastic.

Assumption 3: For any regular agent $i \in \mathcal{H}$, respectively denote $|\mathcal{N}_i|$ and $|\mathcal{B}_i|$ as its numbers of neighbors and Byzantine neighbors, and suppose $|\mathcal{B}_i| \leq b < \frac{|\mathcal{N}_i|}{3}$. Consider a graph set $\mathcal{H}_{\mathcal{G}}$ whose elements are the subgraphs of $\mathcal{G}$ by removing all edges of Byzantine agents, and removing any additional b incoming edges at each honest agent. Any subgraph $\mathcal{G}' \in \mathcal{H}_{\mathcal{G}}$ has at least one agent i^* which has a directed path to all agents in $\mathcal{G}'$. The path length is no more than $\tau_{\mathcal{G}}$, which is called as the maximum network diameter after CTM.

Assumption 1 is common in constrained optimization [4], [39] and applies to both DRA and BREDA. Assumption 2 is common in decentralized optimization and applies to DRA. Assumption 3 is for BREDA, requiring that the network of the regular agents, even after CTM, are still able to disseminate messages [40, Assumption 4].

Denote $(\boldsymbol{\theta}_v^*, \boldsymbol{\lambda}_v^*)$ as the saddle point of $\mathcal{L}_v(\boldsymbol{\theta}; \hat{\boldsymbol{\lambda}})$ in (4), where $\boldsymbol{\theta}_v^* \in \mathbb{R}^{ND}$ and $\boldsymbol{\lambda}_v^* \in \mathbb{R}^D$. The following theorem shows that DRA converges to the saddle point.

Theorem 1: Consider the DRA updates (21), (22) and (23). Define a column vector

$$\boldsymbol{z}^k := \left[\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\|; \|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|; \|\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\|\right].$$

If Assumptions 1 and 2 hold, then with a proper fixed step size $\gamma^k = \gamma$ and regularization parameter v it holds

$$\lim_{k \to \infty} \boldsymbol{z}^k = \lim_{k \to \infty} (\boldsymbol{G})^k \boldsymbol{z}^0 = \boldsymbol{0}. \quad (24)$$

Here the matrix G is defined as

$$G := \begin{bmatrix} 1 - \gamma v + \gamma L & \frac{\gamma}{N} & 0 \\ \gamma & 1 - \gamma v & \gamma \\ \gamma L + \gamma v & \frac{\gamma}{N} & \delta \end{bmatrix}, \qquad (25)$$

whose spectral radius satisfies $\rho(G) < 1$, and $\delta \in [0, 1)$ is the spectral norm of $W - \frac{1}{N} A^T A$.

The proof of Theorem 1 and the conditions on γ and v are left to Appendix A. To our particularly interest is the influence of the network connectedness. For a well-connected network with small δ, it is shown that the step size γ can be large and the regularization parameter v can be small.

Now we analyze the convergence of BREDA. We rewrite the BREDA updates (18), (19) and (20) in a compact form

$$\Theta^{k+1} = \Pi_{\tilde{C}} \left[\Theta^k - \gamma^k \left(\nabla f\left(\Theta^k\right) + \frac{1}{N} \lambda^k + v\Theta^k \right) \right], \tag{26}$$

$$\lambda^{k+1} = \Pi_{\tilde{U}} \left[\lambda^k + \gamma^k \left(X^k - \tilde{A}^T s - v\lambda^k \right) \right], \tag{27}$$

$$X_d^{k+1} = Y^k(d) X_d^k + \triangle \Theta_d^{k+1}. \tag{28}$$

In (28), $X_d^{k+1} \in \mathbb{R}^{|\mathcal{H}|}$ and $\triangle \Theta_d^{k+1} \in \mathbb{R}^{|\mathcal{H}|}$ collect the d-th elements of x_i^{k+1} and $\triangle \theta_i^{k+1}$ for all the regular agents $i \in \mathcal{H}$, respectively. The matrix $Y^k(d) \in \mathbb{R}^{|\mathcal{H}| \times |\mathcal{H}|}$ describes the trimmed mean operation on the d-th element at time k [41]. The construction of such $Y^k(d)$ can be found in [40].

Define the regularized Lagrangian function of (3) as

$$\mathcal{L}'_v(\Theta; \hat{\lambda}) := \mathcal{L}(\Theta; \hat{\lambda}) + \frac{v}{2} \sum_{i \in \mathcal{H}} \|\theta_i\|^2 - \frac{v}{2} \left\| \hat{\lambda} \right\|^2, \tag{29}$$

where the Lagrangian function of (3) is

$$\mathcal{L}(\Theta; \hat{\lambda}) := \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} f_i(\theta_i) + \left\langle \hat{\lambda}, \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \theta_i - s \right\rangle. \tag{30}$$

Denote $(\Theta_v^*, \lambda_v^*)$ as the saddle point of $\mathcal{L}'_v(\Theta; \hat{\lambda})$ in (29), where $\Theta_v^* \in \mathbb{R}^{|\mathcal{H}|D}$ and $\lambda_v^* \in \mathbb{R}^D$. The following theorem shows that BREDA converges to a neighborhood of $(\Theta_v^*, \lambda_v^*)$.

Theorem 2: Consider the BREDA updates (26), (27) and (28). Define a column vector

$$Z^k := \left[\|\Theta^k - \Theta_v^*\|; \|\lambda^k - \tilde{A}^T \lambda_v^*\| \right].$$

Suppose that Assumptions 1 and 3 hold. The step size is set as $\gamma^0 = \underline{\gamma}$ and $\gamma^k = \min\{\underline{\gamma}, \frac{\overline{\gamma}}{k^\epsilon}\}$ with $\epsilon > 1$ for $k \geq 1$, such that there exists a smallest integer $k_0 > 1$ satisfying $\underline{\gamma} \geq \frac{\overline{\gamma}}{k_0^\epsilon}$. Then, with properly chosen step size parameters $\underline{\gamma}$, $\overline{\gamma}$ and the regularization parameter v, for all $k < k_0$ it holds

$$\|Z^{k+1}\| \leq \left(1 - \eta\underline{\gamma}\right)^{k+1} \|Z^0\| + \frac{1}{\eta} \left(\underline{\gamma}\Delta_1 + \Delta_2\right), \tag{31}$$

and for all $k \geq k_0$ it holds

$$\limsup_{k \to +\infty} \|Z^{k+1}\| \leq \frac{(\epsilon - 1)k_0^{\epsilon-1}}{(\epsilon - 1)k_0^{\epsilon-1} + \eta\overline{\gamma}} \|Z^{k_0}\|$$

$$+ \frac{\overline{\gamma}\Delta_3}{(\epsilon-1)(k_0-1)^{\epsilon-1}} + \frac{\overline{\gamma}^2\Delta_4}{(2\epsilon-1)(k_0-1)^{2\epsilon-1}}, \tag{32}$$

where $\eta\underline{\gamma} \in (0, 1)$ with $\eta = \frac{2v - L - \sqrt{4 + L^2}}{2}$, while Δ_1, Δ_2, Δ_3 and Δ_4 are certain positive constants.

The proof of Theorem 2 and the conditions on $\underline{\gamma}$, $\overline{\gamma}$ and v are left to Appendix B. The step size rule is two-stage. At the first stage, with a constant step size, $\|Z^k\|$ converges to a neighborhood of 0 at a linear rate according to (31). The second stage involves a diminishing step size and the convergence error is upper-bounded according to (32). Observe that there is a tradeoff in setting the step size parameter ϵ. A small ϵ leads to a small first term but large second and third terms at the right-hand side of (32), and vise versa. In the numerical experiments, we find that using a constant step size $\gamma^k = \gamma$ works well for BREDA.

VI. NUMERICAL EXPERIMENTS

In this section, we present numerical experiments to demonstrate the resilience of BREDA to various Byzantine attacks.

A. Case 1: Synthetic Problem

Consider the one-dimensional case with $D = 1$. The upper bound of average resource is $s = 5$. The local constraint of agent i is $\theta_i \in C_i = [0, 10]$. The local cost function of agent i is in the form of $f_i(\theta_i) = -\alpha\beta_i \log(1 + \theta_i)$ where $\alpha = 300$ is a constant and different agents i have different β_i. The variance of β_i reflects the heterogeneity of local cost functions.

We generate a connected random network consisting of $N = 100$ agents, with each agent having 15 neighbors. The maximum available amount of resources is 500, meaning that $s = 5$. We randomly select $|\mathcal{B}|$ Byzantine agents, with $|\mathcal{B}| = 6$ by default. We test the performance of BREDA with a constant step size $\gamma^k = \gamma$ under two typical Byzantine attacks: max value and trimmed Gaussian attacks. For max value attacks, Byzantine agent $i \in \mathcal{B}$ sets its message as $q_i^k = 10$. For trimmed Gaussian attacks, Byzantine agent $i \in \mathcal{B}$ draws its message q_i^k from a Gaussian distribution with mean 7 and variance 9, followed by being trimmed to the range of $[0, 10]$. We set the lower and upper bounds since all the agents have the same local constraint $\theta_i \in C_i = [0, 10]$ such that the average demand of resources should also be within this range. We consider two baselines, Oracle and DRA. Oracle means that the Byzantine agents behave honestly, while DRA is subject to Byzantine attacks. In DRA, the weights $E_{ij} = \frac{1}{16}$ if and only if $(i, j) \in \mathcal{E}$ or $i = j$. The parameters γ and v are tuned to the best for Oracle, and then applied to DRA and BREDA. When β_i are randomly distributed within $[1, 2]$, $\gamma = 0.620$ and $v = 0.146$. Otherwise, when β_i follow Gaussian distribution with mean 1.5 and variance 1, $\gamma = 0.624$ and $v = 0.156$. The parameter of CTM is set as $b = 6$. The sets U_i are the non-negative orthant. Performance metrics include resource allocation strategy θ_i^k for a randomly chosen honest agent i, average cost of honest agents $f(\Theta^k)$, and constraint violation of honest agents $\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \theta_i^k - s$.

Fig. 1 depicts the performance of Oracle, DRA and BREDA when β_i are randomly distributed within $[1, 2]$. Observe that the resource allocation strategy of DRA is far from its oracle value, while that of BREDA is much closer, under both Byzantine attacks. In terms of average cost and constraint violation of
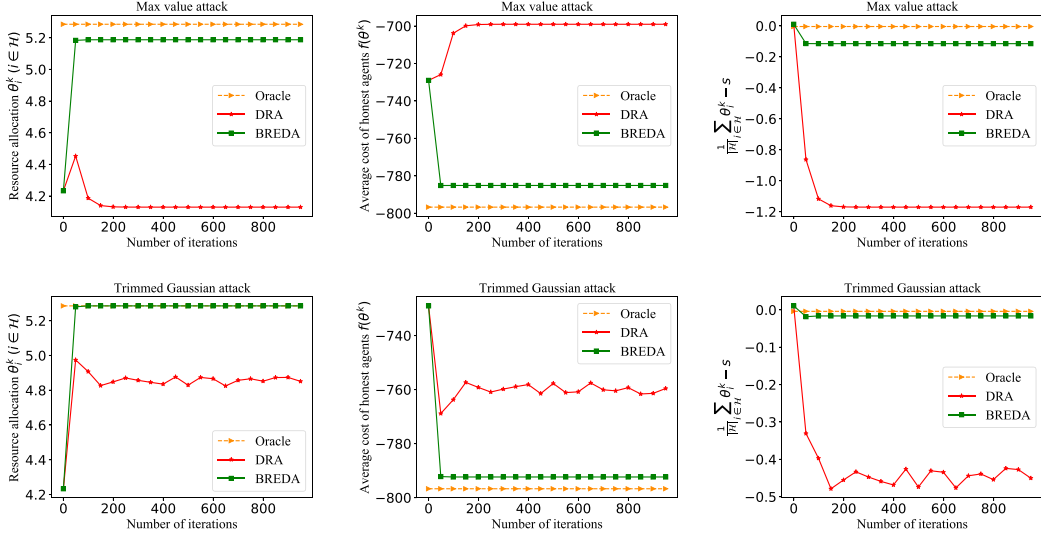
Fig. 1. Oracle without Byzantine attacks, DRA and BREDA under Byzantine attacks on synthetic problem, when β_i are randomly distributed within $[1, 2]$. From top to bottom: max value attacks and trimmed Gaussian attacks. From left to right: resource allocation of a randomly chosen honest agent i, average cost of honest agents, and constraint violation of honest agents.
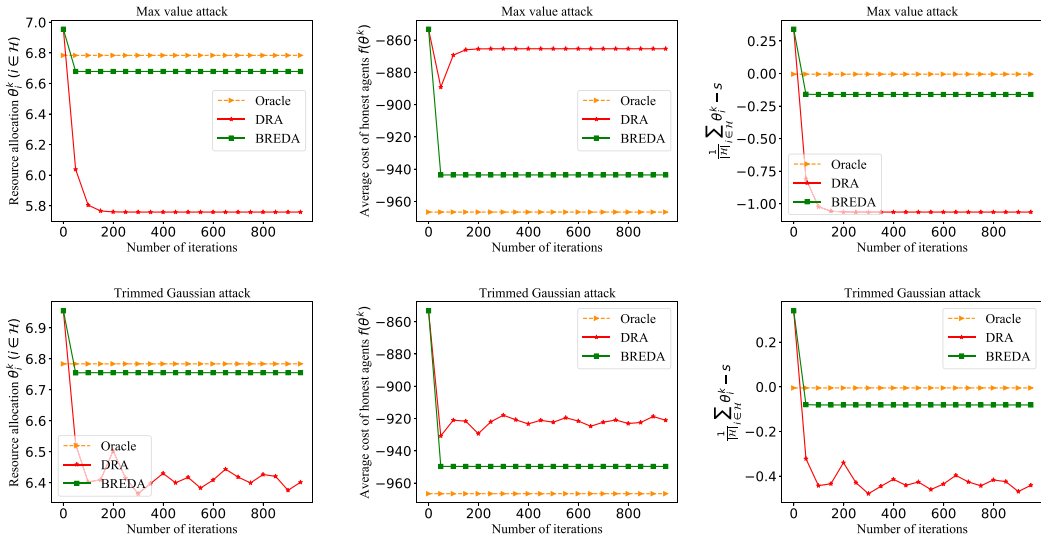

Fig. 2. Oracle without Byzantine attacks, DRA and BREDA under Byzantine attacks on synthetic problem, when β_i follow Gaussian distribution with mean 1.5 and variance 1. From top to bottom: max value attacks and trimmed Gaussian attacks. From left to right: resource allocation of a randomly chosen honest agent i, average cost of honest agents, and constraint violation of honest agents.

honest agents, the gaps between BREDA and Oracle are acceptable, while those between DRA and Oracle are significant. For the two Byzantine attacks, max value attacks are stronger than trimmed Gaussian attacks. Fig. 2 shows the performance of Oracle, DRA and BREDA when β_i follow Gaussian distribution with mean 1.5 and variance 1. Similar conclusions can be made as for Fig. 1. These two sets of numerical experiments demonstrate the vulnerability of DRA to Byzantine attacks, as well as the satisfactory Byzantine-resilience of BREDA.

In the above numerical experiments, the parameter of CTM b and the number of Byzantine agents $|\mathcal{B}|$ are both 6. In Fig. 3 we check the sensitivity of BREDA with respect to the number of Byzantine agents, by varying $|\mathcal{B}|$ as 1, 3 and 6. DRA fails in all cases, while BREDA is almost insensitive to $|\mathcal{B}|$.

B. Case 2: Economic Dispatch for IEEE 118-Bus Test System

Consider the power dispatch problem on the IEEE-118 bus test system with 54 generators [42]. Each generator i has a cost function of generated power $\boldsymbol{\theta}_i$, given by $f_i(\boldsymbol{\theta}_i) = \xi_i + \zeta_i \boldsymbol{\theta}_i + \eta_i \boldsymbol{\theta}_i^2$. The coefficients are in the ranges of $\xi_i \in [6.78, 74.33]$, $\zeta_i \in [8.3391, 37.6968]$, and $\eta_i \in [0.0024, 0.0697]$. Each $\boldsymbol{\theta}_i$ belongs to the set $[\boldsymbol{\theta}_i^{\min}, \boldsymbol{\theta}_i^{\max}]$, where $\boldsymbol{\theta}_i^{\min} \in [5, 150]$ and $\boldsymbol{\theta}_i^{\max} \in [30, 420]$. The total load is 6000 as in [11]. We let the physical connection between any two neighboring generators be bidirectional such that the graph
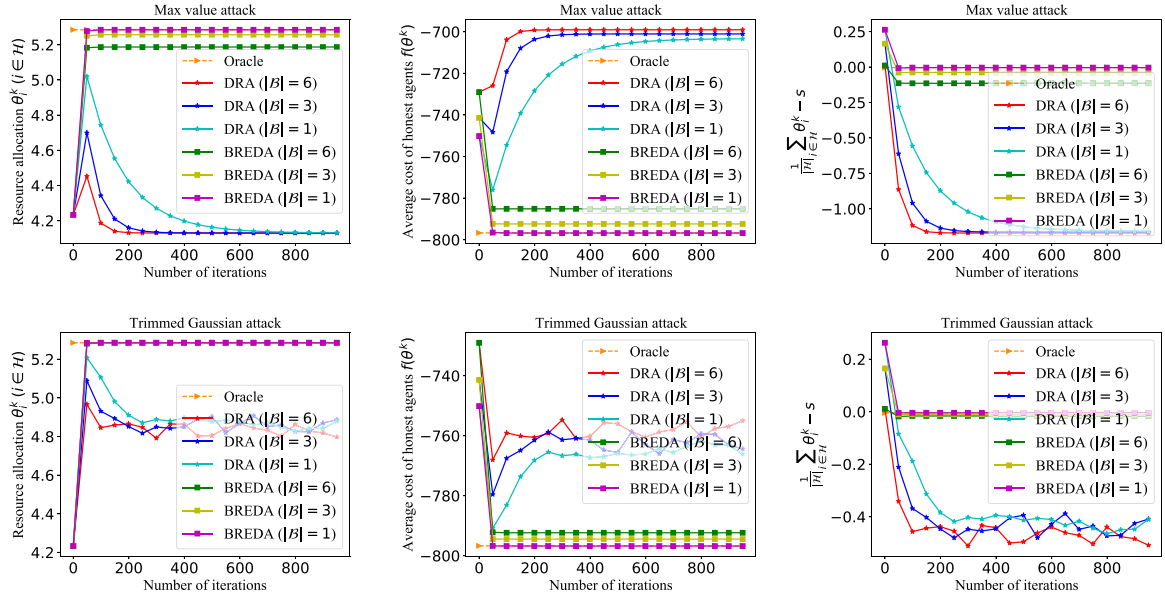
Fig. 3. Oracle without Byzantine attacks, DRA and BREDA under Byzantine attacks on synthetic problem, when β_i are randomly distributed within $[1, 2]$. The number of Byzantine agents $|\mathcal{B}|$ is set as 1, 3 and 6. From top to bottom: max value attacks and trimmed Gaussian attacks. From left to right: resource allocation of a randomly chosen honest agent i, average cost of honest agents, and constraint violation of honest agents.
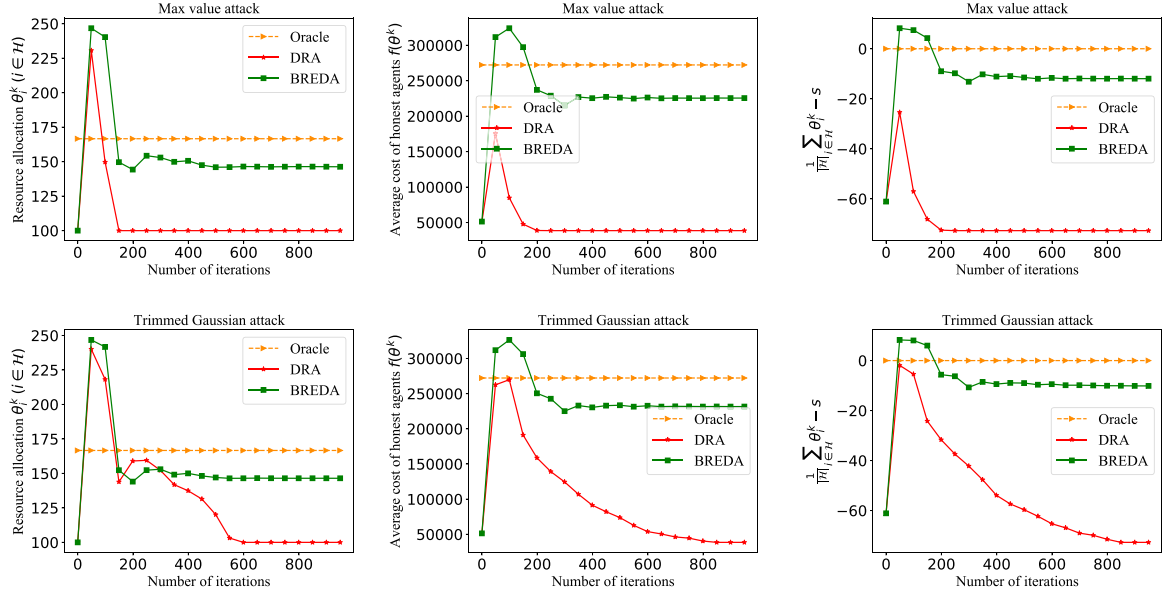

Fig. 4. Oracle without Byzantine attacks, DRA and BREDA under Byzantine attacks on economic dispatch problem. From top to bottom: max value attacks and trimmed Gaussian attacks. From left to right: resource allocation of a randomly chosen honest agent i, average cost of honest agents, and constraint violation of honest agents.

is undirected. We randomly select $|\mathcal{B}| = 1$ Byzantine generator. The parameters γ and υ are set as 0.5 and 0, respectively.

We test the performance of BREDA under max value and trimmed Gaussian attacks. For max value attacks, the Byzantine generator sets its message as 420. For trimmed Gaussian attacks, the Byzantine generator sets its message from a Gaussian distribution with mean 150 and variance 900, followed by being trimmed to the range of [5,420]. In DRA, E is generated by the Metropolis constant weight rule [43]. Fig. 4 shows the performance of Oracle, DRA and BREDA. DRA fails under Byzantine

attacks, while BREDA is close to Oracle and demonstrates to be Byzantine-resilient.

VII. CONCLUSION AND FUTURE WORKS

This paper deals with Byzantine-resilient decentralized resource allocation. We propose BREDA, a primal-dual algorithm equipped with Byzantine-resilient first-order decentralized dynamic average consensus for tracking the average resource demand, where each honest agent uses CTM to aggregate received

messages. We show that BREDA converges to a neighborhood of the saddle point of the regularized resource allocation problem. Extensive numerical experiments demonstrate the resilience of BREDA to various Byzantine attacks. In our future work, we will investigate the development of Byzantine-resilient algorithms for online and asynchronous decentralized resource allocation problems, which are of practical interest in various applications.

APPENDIX A
PROOF OF THEOREM 1

To analyze the convergence of DRA, below we successively bound the three terms $\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}_v^*\|$, $\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|$ and $\|\boldsymbol{x}^{k+1} - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^{k+1}\|$.

Step 1: Bound $\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}_v^*\|$.

According to (21), using the non-expansive property of the projection operator and the fact $\nabla_{\boldsymbol{\theta}}\mathcal{L}_v(\boldsymbol{\theta}_v^*; \boldsymbol{\lambda}_v^*) = \nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_v^*) + \frac{1}{N}\boldsymbol{A}^T\boldsymbol{\lambda}_v^* + v\boldsymbol{\theta}_v^* = \boldsymbol{0}$, we have

$$\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}_v^*\|$$

$$= \left\| \Pi_C\left[\boldsymbol{\theta}^k - \gamma^k\left(\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}^k) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\theta}^k\right)\right] \right.$$

$$\left. - \Pi_C\left[\boldsymbol{\theta}_v^* - \gamma^k\left(\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_v^*) + \frac{1}{N}\boldsymbol{A}^T\boldsymbol{\lambda}_v^* + v\boldsymbol{\theta}_v^*\right)\right] \right\| \quad (33)$$

$$\leq \left\| \boldsymbol{\theta}^k - \gamma^k\left(\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}^k) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\theta}^k\right) \right.$$

$$\left. - \left(\boldsymbol{\theta}_v^* - \gamma^k\left(\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_v^*) + \frac{1}{N}\boldsymbol{A}^T\boldsymbol{\lambda}_v^* + v\boldsymbol{\theta}_v^*\right)\right) \right\|$$

$$= \left\| (1 - \gamma^k v)(\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*) - \gamma^k(\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}^k) - \nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_v^*)) \right.$$

$$\left. - \frac{\gamma^k}{N}\left(\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\right) \right\|.$$

Applying the triangle inequality to the right-hand side of (33), we have

$$\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}_v^*\|$$

$$\leq (1 - \gamma^k v)\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\| + \gamma^k\|\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}^k) - \nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_v^*)\|$$

$$+ \frac{\gamma^k}{N}\|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|$$

$$\leq (1 - \gamma^k v)\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\| + \gamma^k L\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\| + \frac{\gamma^k}{N}\|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|$$

$$= (1 - \gamma^k v + \gamma^k L)\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\| + \frac{\gamma^k}{N}\|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|. \quad (34)$$

To derive the last inequality, we use Assumption 1 that each $f_i(\cdot)$ has Lipschitz continuous gradients with constant L, such that $f(\cdot)$ has Lipschitz continuous gradients with constant L.

Step 2: Bound $\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|$.

According to (22), using the non-expansive property of the projection operator and the fact $\nabla_{\hat{\boldsymbol{\lambda}}}\mathcal{L}_v(\boldsymbol{\theta}_v^*; \boldsymbol{\lambda}_v^*) = \frac{1}{N}\boldsymbol{A}\boldsymbol{\theta}_v^* -$

$\boldsymbol{s} - v\boldsymbol{\lambda}_v^* = \boldsymbol{0}$, we have

$$\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|$$

$$= \left\| \Pi_U[\boldsymbol{\lambda}^k + \gamma^k(\boldsymbol{x}^k - \boldsymbol{A}^T\boldsymbol{s} - v\boldsymbol{\lambda}^k)] \right.$$

$$\left. - \Pi_U\left[\boldsymbol{A}^T\boldsymbol{\lambda}_v^* + \gamma^k\boldsymbol{A}^T\left(\frac{1}{N}\boldsymbol{A}\boldsymbol{\theta}_v^* - \boldsymbol{s} - v\boldsymbol{\lambda}_v^*\right)\right] \right\|$$

$$\leq \left\| \boldsymbol{\lambda}^k + \gamma^k(\boldsymbol{x}^k - \boldsymbol{A}^T\boldsymbol{s} - v\boldsymbol{\lambda}^k) \right.$$

$$\left. - \left(\boldsymbol{A}^T\boldsymbol{\lambda}_v^* + \gamma^k\boldsymbol{A}^T\left(\frac{1}{N}\boldsymbol{A}\boldsymbol{\theta}_v^* - \boldsymbol{s} - v\boldsymbol{\lambda}_v^*\right)\right) \right\|$$

$$= \left\| (1 - \gamma^k v)(\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*) + \gamma^k\left(\boldsymbol{x}^k - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}_v^*\right) \right\|$$

$$= \left\| (1 - \gamma^k v)(\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*) \right.$$

$$\left. + \gamma^k\left(\boldsymbol{x}^k - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^k\right) + \gamma^k\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^k - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}_v^*\right) \right\|. \quad (35)$$

Applying the triangle inequality to the right-hand side of (35), we have

$$\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|$$

$$\leq (1 - \gamma^k v)\|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\| + \gamma^k\left\|\boldsymbol{x}^k - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^k\right\|$$

$$+ \gamma^k\left\|\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^k - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}_v^*\right\|$$

$$\leq (1 - \gamma^k v)\|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\| + \gamma^k\|\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\| + \gamma^k\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\|. \quad (36)$$

To derive the last inequality, we use the definition of $\bar{\boldsymbol{\theta}}^k = \frac{1}{N}\boldsymbol{A}\boldsymbol{\theta}^k$ and the fact that the eigenvalues of $\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}$ are either 0 or 1.

Step 3: Bound $\|\boldsymbol{x}^{k+1} - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^{k+1}\|$.

According to (23), we obtain

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^{k+1}\|$$

$$= \|\boldsymbol{W}\boldsymbol{x}^k + \triangle\boldsymbol{\theta}^{k+1} - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^{k+1}\|$$

$$= \|\boldsymbol{W}\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k + \triangle\boldsymbol{\theta}^{k+1} - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^{k+1} + \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\|$$

$$= \left\| \boldsymbol{W}\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k + \triangle\boldsymbol{\theta}^{k+1} - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^{k+1} + \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^k \right\|$$

$$= \left\| \boldsymbol{W}\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k + \triangle\boldsymbol{\theta}^{k+1} - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\triangle\boldsymbol{\theta}^{k+1} \right\|$$

$$\leq \|\boldsymbol{W}\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\| + \left\| \triangle\boldsymbol{\theta}^{k+1} - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\triangle\boldsymbol{\theta}^{k+1} \right\|$$

$$\leq \|\boldsymbol{W}\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\| + \|\triangle\boldsymbol{\theta}^{k+1}\|, \quad (37)$$

where the last equality holds because eigenvalues of $\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}$ is either 0 or 1.

We proceed to analyzing the two items at the right-hand side of (37). Multiplying $\frac{1}{N}\boldsymbol{A}$ at both sides of (23), we have

$$\frac{1}{N}\boldsymbol{A}\boldsymbol{x}^{k+1} = \frac{1}{N}\boldsymbol{A}\boldsymbol{W}\boldsymbol{x}^k + \frac{1}{N}\boldsymbol{A}\triangle\boldsymbol{\theta}^{k+1}$$

$$= \frac{1}{N}\boldsymbol{A}\boldsymbol{W}\boldsymbol{x}^k + \frac{1}{N}\boldsymbol{A}\boldsymbol{\theta}^{k+1} - \frac{1}{N}\boldsymbol{A}\boldsymbol{\theta}^k. \quad (38)$$

Because $\boldsymbol{W} := \boldsymbol{E}\otimes\boldsymbol{I}$ and $\boldsymbol{E}$ is doubly stochastic, we have $\frac{1}{N}\boldsymbol{A}\boldsymbol{W}\boldsymbol{x}^k = \frac{1}{N}\boldsymbol{A}\boldsymbol{x}^k = \bar{\boldsymbol{x}}^k$. By definition, $\frac{1}{N}\boldsymbol{A}\boldsymbol{\theta}^k = \bar{\boldsymbol{\theta}}^k$. Therefore, we can we rewrite (38) as

$$\bar{\boldsymbol{x}}^{k+1} = \bar{\boldsymbol{x}}^k + \bar{\boldsymbol{\theta}}^{k+1} - \bar{\boldsymbol{\theta}}^k. \quad (39)$$

Since the initialization of $\boldsymbol{x}^0$ satisfies $\boldsymbol{x}^0 = \boldsymbol{\theta}^0$, from (39) we obtain $\bar{\boldsymbol{x}}^k = \bar{\boldsymbol{\theta}}^k$ for any k. Therefore, we have

$$\|\boldsymbol{W}\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\| = \left\|\left(\boldsymbol{W} - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\right)\left(\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\right)\right\|$$

$$\leq \left\|\boldsymbol{W} - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\right\|\|\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\|. \quad (40)$$

According to Assumption 2, the underlying communication graph is bidirectionally connected such that the spectral norm δ of $\boldsymbol{W} - \frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}$ is within [0,1) [44]. Thus, we have

$$\|\boldsymbol{W}\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\| \leq \delta\|\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\|. \quad (41)$$

According to (21), using the non-expansive property of the projection operator and the fact $\nabla_{\boldsymbol{\theta}}\mathcal{L}_v(\boldsymbol{\theta}_v^*;\boldsymbol{\lambda}_v^*) = \nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_v^*) + \frac{1}{N}\boldsymbol{A}^T\boldsymbol{\lambda}_v^* + v\boldsymbol{\theta}_v^* = \mathbf{0}$, we have

$$\|\triangle\boldsymbol{\theta}^{k+1}\|$$

$$= \|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|$$

$$= \left\|\Pi_C\left[\boldsymbol{\theta}^k - \gamma^k\left(\nabla_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}^k\right) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\theta}^k\right)\right] - \Pi_C\left[\boldsymbol{\theta}^k\right]\right\|$$

$$\leq \left\|\boldsymbol{\theta}^k - \gamma^k\left(\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}^k) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\theta}^k\right) - \boldsymbol{\theta}^k\right\|$$

$$= \gamma^k\left\|\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}^k) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\theta}^k\right.$$

$$\left. - \left(\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_v^*) + \frac{1}{N}\boldsymbol{A}^T\boldsymbol{\lambda}_v^* + v\boldsymbol{\theta}_v^*\right)\right\|. \quad (42)$$

Applying the triangle inequality to the right-hand side of (42), we have

$$\|\triangle\boldsymbol{\theta}^{k+1}\|$$

$$\leq \gamma^k\|\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}^k) - \nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_v^*)\| + \frac{\gamma^k}{N}\|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|$$

$$+ \gamma^k v\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\|$$

$$\leq \gamma^k(L+v)\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\| + \frac{\gamma^k}{N}\|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|, \quad (43)$$

where the last inequality uses the fact that $f(\cdot)$ has Lipschitz continuous gradients with constant L.

Substituting (41) and (43) into (37) followed by rearranging the terms, we have

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^{k+1}\|$$

$$\leq \delta\|\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\| + \gamma^k(L+v)\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\| + \frac{\gamma^k}{N}\|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\|. \quad (44)$$

Step 4: Reach conclusion.

Finally, combining (34), (36) and (44) and choosing a fixed step size $\gamma^k = \gamma$, we have

$$\underbrace{\begin{bmatrix}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}_v^*\| \\ \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\| \\ \|\boldsymbol{x}^{k+1} - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^{k+1}\|\end{bmatrix}}_{:=\boldsymbol{z}^{k+1}\in\mathbb{R}^3}$$

$$\leq \underbrace{\begin{bmatrix}1-\gamma v+\gamma L & \frac{\gamma}{N} & 0 \\ \gamma & 1-\gamma v & \gamma \\ \gamma(L+v) & \frac{\gamma}{N} & \delta\end{bmatrix}}_{:=\boldsymbol{G}\in\mathbb{R}^{3\times3}}\underbrace{\begin{bmatrix}\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_v^*\| \\ \|\boldsymbol{\lambda}^k - \boldsymbol{A}^T\boldsymbol{\lambda}_v^*\| \\ \|\boldsymbol{x}^k - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^k\|\end{bmatrix}}_{:=\boldsymbol{z}^k\in\mathbb{R}^3}. \quad (45)$$

When $1-\gamma v+\gamma L \geq 0$ and $1-\gamma v \geq 0$, the entries of $\boldsymbol{z}^k$ and $\boldsymbol{G}$ are all nonnegative. Then, we can rewrite (45) recursively and obtain

$$\boldsymbol{z}^k \leq (\boldsymbol{G})^k\boldsymbol{z}^0. \quad (46)$$

According to Theorem 5.6.12 in [45], if the spectral radius of $\boldsymbol{G}$ satisfies $\rho(\boldsymbol{G}) < 1$, then $\lim_{k\to\infty}(\boldsymbol{G})^k = \mathbf{0}$ and DRA linearly converges to the saddle point. Below we give the conditions under which $\rho(\boldsymbol{G}) < 1$ by showing the eigenvalues of $\boldsymbol{G}$ are all inside the unit circle of the complex plane.

Gershgorin Circle Theorem [46] indicates that all eigenvalues of $\boldsymbol{G}$ lie in the union of the closed circles with centers $\boldsymbol{G}_{ii}$ and radii $\sum_{j\neq i}^3 |\boldsymbol{G}_{ij}|, i = 1,2,3$. Denote any complex eigenvalue of $\boldsymbol{G}$ as τ. According to the definition of $\boldsymbol{G}$ and Gershgorin Circle Theorem, we have

$$\begin{cases} |\tau - (1-\gamma v+\gamma L)| \leq \frac{\gamma}{N}, \\ |\tau - (1-\gamma v)| \leq 2\gamma, \\ |\tau - \delta| \leq \gamma L + \gamma v + \frac{\gamma}{N}. \end{cases} \quad (47)$$

To guarantee that the eigenvalues of $\boldsymbol{G}$ are all inside the unit circle of the complex plane, it suffices that the regularization parameter v satisfies $v > \max\{2, L + \frac{1}{N}\}$, the step size γ satisfies $\gamma < \min\{\frac{1}{2}, \frac{N(1-\delta)}{2NL+2}, \frac{N(1-\delta)}{2N+NL+1}\}$, while v and γ jointly satisfy $v < \frac{2}{\gamma} + L - \frac{1}{N}$, $v < \frac{2}{\gamma} - 2$, and $v < \frac{1-\delta}{\gamma} - L - \frac{1}{N}$. With these conditions, $1-\gamma v+\gamma L \geq 0$ and $1-\gamma v \geq 0$ that we use to derive (46) also hold true. This completes the proof.

APPENDIX B
PROOF OF THEOREM 2

To analyze the convergence of BREDA, we successively bound the three terms $\|\Theta^{k+1} - \Theta_v^*\|$, $\|\lambda^{k+1} - \tilde{A}^T \lambda_v^*\|$ and $\|X^{k+1} - \tilde{A}^T \bar{\Theta}^{k+1}\|$.

Step 1: Bound $\|\Theta^{k+1} - \Theta_v^*\|$.

According to (26), using the non-expansive property of the projection operator and the fact $\nabla_\Theta \mathcal{L}_v'(\Theta_v^*; \lambda_v^*) = \nabla_\Theta f(\Theta_v^*) + \frac{1}{|\mathcal{H}|}\tilde{A}^T \lambda_v^* + v\Theta_v^* = 0$, we have

$$
\|\Theta^{k+1} - \Theta_v^*\|
$$
$$
= \left\| \Pi_{\tilde{C}}\left[\Theta^k - \gamma^k \left(\nabla_\Theta f(\Theta^k) + \frac{1}{N}\lambda^k + v\Theta^k \right) \right] \right.
$$
$$
\left. - \Pi_{\tilde{C}}\left[\Theta_v^* - \gamma^k \left(\nabla_\Theta f(\Theta_v^*) + \frac{1}{|\mathcal{H}|}\tilde{A}^T \lambda_v^* + v\Theta_v^* \right) \right] \right\|
$$
$$
\leq \left\| \Theta^k - \gamma^k \left(\nabla_\Theta f(\Theta^k) + \frac{1}{N}\lambda^k + v\Theta^k \right) \right.
$$
$$
\left. - \left(\Theta_v^* - \gamma^k \left(\nabla_\Theta f(\Theta_v^*) + \frac{1}{|\mathcal{H}|}\tilde{A}^T \lambda_v^* + v\Theta_v^* \right) \right) \right\|
$$
$$
= \left\| (1 - \gamma^k v)(\Theta^k - \Theta_v^*) - \gamma^k (\nabla_\Theta f(\Theta^k) - \nabla_\Theta f(\Theta_v^*)) \right.
$$
$$
\left. - \left(\frac{\gamma^k}{N}\lambda^k - \frac{\gamma^k}{|\mathcal{H}|}\tilde{A}^T \lambda_v^* \right) \right\|
$$
$$
= \left\| (1 - \gamma^k v)(\Theta^k - \Theta_v^*) - \gamma^k (\nabla_\Theta f(\Theta^k) - \nabla_\Theta f(\Theta_v^*)) \right.
$$
$$
\left. - \left(\frac{\gamma^k}{N}\lambda^k - \frac{\gamma^k}{N}\tilde{A}^T \lambda_v^* \right) - \left(\frac{\gamma^k}{N}\tilde{A}^T \lambda_v^* - \frac{\gamma^k}{|\mathcal{H}|}\tilde{A}^T \lambda_v^* \right) \right\|. \tag{48}
$$

Applying the triangle inequality to the right-hand side of (48), we have

$$
\|\Theta^{k+1} - \Theta_v^*\|
$$
$$
\leq (1 - \gamma^k v)\|\Theta^k - \Theta_v^*\| + \gamma^k \|\nabla_\Theta f(\Theta^k) - \nabla_\Theta f(\Theta_v^*)\|
$$
$$
+ \frac{\gamma^k}{N}\|\lambda^k - \tilde{A}^T \lambda_v^*\| + \left(\frac{\gamma^k}{|\mathcal{H}|} - \frac{\gamma^k}{N} \right) \|\tilde{A}^T \lambda_v^*\|
$$
$$
\leq (1 - \gamma^k v)\|\Theta^k - \Theta_v^*\| + \gamma^k L\|\Theta^k - \Theta_v^*\| + \frac{\gamma^k}{N}\|\lambda^k - \tilde{A}^T \lambda_v^*\|
$$
$$
+ \sqrt{|\mathcal{H}|} \left(\frac{\gamma^k}{|\mathcal{H}|} - \frac{\gamma^k}{N} \right) \|\lambda_v^*\|
$$
$$
= (1 - \gamma^k v + \gamma^k L)\|\Theta^k - \Theta_v^*\| + \frac{\gamma^k}{N}\|\lambda^k - \tilde{A}^T \lambda_v^*\|
$$
$$
+ \gamma^k \sqrt{|\mathcal{H}|} \left(\frac{1}{|\mathcal{H}|} - \frac{1}{N} \right) \|\lambda_v^*\|. \tag{49}
$$

The last inequality comes from that $f(\cdot)$ has Lipschitz continuous gradients with constant L and the fact $\|\tilde{A}^T\| = \sqrt{|\mathcal{H}|}$. Defining $\Omega_1^k := \gamma^k \sqrt{|\mathcal{H}|}(\frac{1}{|\mathcal{H}|} - \frac{1}{N})\|\lambda_v^*\|$, we have

$$
\|\Theta^{k+1} - \Theta_v^*\|
$$
$$
\leq (1 - \gamma^k v + \gamma^k L)\|\Theta^k - \Theta_v^*\| + \frac{\gamma^k}{N}\|\lambda^k - \tilde{A}^T \lambda_v^*\| + \Omega_1^k
$$
$$
\leq (1 - \gamma^k v + \gamma^k L)\|\Theta^k - \Theta_v^*\| + \gamma^k \|\lambda^k - \tilde{A}^T \lambda_v^*\| + \Omega_1^k. \tag{50}
$$

Step 2: Bound $\|\lambda^{k+1} - \tilde{A}^T \lambda_v^*\|$.

According to (27), using the non-expansive property of the projection operator and the fact $\nabla_{\hat{\lambda}} \mathcal{L}_v(\Theta_v^*; \lambda_v^*) = \frac{1}{|\mathcal{H}|}\tilde{A}\Theta_v^* - s - v\lambda_v^* = 0$, we have

$$
\|\lambda^{k+1} - \tilde{A}^T \lambda_v^*\|
$$
$$
= \left\| \Pi_{\tilde{U}}\left[\lambda^k + \gamma^k \left(X^k - \tilde{A}^T s - v\lambda^k \right) \right] \right.
$$
$$
\left. - \Pi_{\tilde{U}}\left[\tilde{A}^T \lambda_v^* + \gamma^k \tilde{A}^T \left(\frac{1}{|\mathcal{H}|}\tilde{A}\Theta_v^* - s - v\lambda_v^* \right) \right] \right\|
$$
$$
\leq \left\| \lambda^k + \gamma^k \left(X^k - \tilde{A}^T s - v\lambda^k \right) \right.
$$
$$
\left. - \left(\tilde{A}^T \lambda_v^* + \gamma^k \tilde{A}^T \left(\frac{1}{|\mathcal{H}|}\tilde{A}\Theta_v^* - s - v\lambda_v^* \right) \right) \right\|
$$
$$
= \left\| (1 - \gamma^k v)\left(\lambda^k - \tilde{A}^T \lambda_v^* \right) + \gamma^k \left(X^k - \frac{1}{|\mathcal{H}|}\tilde{A}^T \tilde{A}\Theta_v^* \right) \right\|
$$
$$
= \left\| (1 - \gamma^k v)\left(\lambda^k - \tilde{A}^T \lambda_v^* \right) + \gamma^k \left(X^k - \frac{1}{|\mathcal{H}|}\tilde{A}^T \tilde{A}\Theta^k \right) \right.
$$
$$
\left. + \gamma^k \left(\frac{1}{|\mathcal{H}|}\tilde{A}^T \tilde{A}\Theta^k - \frac{1}{|\mathcal{H}|}\tilde{A}^T \tilde{A}\Theta_v^* \right) \right\|. \tag{51}
$$

Applying the triangle inequality to the right-hand side of (51), we have

$$
\|\lambda^{k+1} - \tilde{A}^T \lambda_v^*\|
$$
$$
\leq (1 - \gamma^k v)\|\lambda^k - \tilde{A}^T \lambda_v^*\| + \gamma^k \left\| X^k - \frac{1}{|\mathcal{H}|}\tilde{A}^T \tilde{A}\Theta^k \right\|
$$
$$
+ \gamma^k \left\| \frac{1}{|\mathcal{H}|}\tilde{A}^T \tilde{A}\Theta^k - \frac{1}{|\mathcal{H}|}\tilde{A}^T \tilde{A}\Theta_v^* \right\|
$$
$$
\leq (1 - \gamma^k v)\|\lambda^k - \tilde{A}^T \lambda_v^*\| + \gamma^k \|\Theta^k - \Theta_v^*\|
$$
$$
+ \gamma^k \|X^k - \tilde{A}^T \bar{\Theta}^k\|. \tag{52}
$$

To derive the last inequality, we use the definition of $\bar{\Theta}^k = \frac{1}{|\mathcal{H}|}\tilde{A}\Theta^k$ and the fact that the eigenvalues of $\frac{1}{|\mathcal{H}|}\tilde{A}^T \tilde{A}$ are either 0 or 1.

Step 3: Bound $\|\boldsymbol{X}^{k+1} - \tilde{\boldsymbol{A}}^T \bar{\boldsymbol{\Theta}}^{k+1}\|$.
We rewrite (28) as

$$\boldsymbol{X}_d^{k+1} = \boldsymbol{Y}^k(d)\boldsymbol{X}_d^k + \triangle\boldsymbol{\Theta}_d^{k+1}$$

$$= \bar{\boldsymbol{Y}}_0^k(d)\boldsymbol{X}_d^0 + \sum_{g=0}^k \bar{\boldsymbol{Y}}_{g+1}^k(d)\triangle\boldsymbol{\Theta}_d^{g+1}, \qquad (53)$$

where $\bar{\boldsymbol{Y}}_{k_0}^k(d) := \boldsymbol{Y}^k(d)\boldsymbol{Y}^{k-1}(d)\cdots\boldsymbol{Y}^{k_0}(d)$. The following lemma characterizes the limiting property of $\bar{\boldsymbol{Y}}_{k_0}^k(d)$.

Lemma 1 ([40], Lemma 3): If the matrices $\boldsymbol{Y}^k(d)$ are constructed following the way in [40], then the limit of $\bar{\boldsymbol{Y}}_{k_0}^k(d)$ exists and is in the form of $\mathbf{1}\boldsymbol{p}_{k_0}^T(d)$, where $\boldsymbol{p}_{k_0}(d) \in \mathbb{R}^{|\mathcal{H}|}$ is a stochastic vector and all of its elements are within the range of $\left[0, \frac{1}{\min_{i\in\mathcal{H}}\{|\mathcal{H}_i|+|\mathcal{B}_i|-2b+1\}}\right]$. Further, $\bar{\boldsymbol{Y}}_{k_0}^k(d)$ converges to the limit at a linear rate. To be specific, we have

$$\lim_{k\to\infty} \bar{\boldsymbol{Y}}_{k_0}^k(d) = \mathbf{1}\boldsymbol{p}_{k_0}^T(d), \qquad (54)$$

$$\|\bar{\boldsymbol{Y}}_{k_0}^k(d) - \mathbf{1}\boldsymbol{p}_{k_0}^T(d)\|^2 \le 4|\mathcal{H}|^2(u)^{k-k_0+1}, \qquad (55)$$

where the constant $u \in (0,1)$ monotonically increases as the maximum network diameter $\tau_{\mathcal{G}}$ increases.

Motivated by Lemma 1, we introduce an auxiliary sequence $\widehat{\boldsymbol{X}}^{k+1} := \tilde{\boldsymbol{A}}^T\boldsymbol{w}^{k+1} \in \mathbb{R}^{|\mathcal{H}|D}$, in which the d-th element of $\boldsymbol{w}^{k+1} \in \mathbb{R}^D$, denoted by $\boldsymbol{w}_d^{k+1}$, satisfies $\boldsymbol{w}_d^{k+1} = \boldsymbol{p}_0^T(d)\widehat{\boldsymbol{X}}_d^0 + \sum_{g=0}^k \boldsymbol{p}_{g+1}^T(d)\triangle\boldsymbol{\Theta}_d^{g+1}$. Therefore, $\widehat{\boldsymbol{X}}^{k+1}$ asymptotically approximates $\boldsymbol{X}^{k+1}$. For notational convenience, define

$$\widehat{\boldsymbol{X}}_d^{k+1} = \mathbf{1}\boldsymbol{w}_d^{k+1}$$

$$= \mathbf{1}\boldsymbol{p}_0^T(d)\widehat{\boldsymbol{X}}_d^0 + \mathbf{1}\sum_{g=0}^k \boldsymbol{p}_{g+1}^T(d)\triangle\boldsymbol{\Theta}_d^{g+1}. \qquad (56)$$

Utilizing $\widehat{\boldsymbol{X}}^{k+1}$ as an intermediate variable, we can bound $\|\boldsymbol{X}^{k+1} - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^{k+1}\|$ as

$$\|\boldsymbol{X}^{k+1} - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^{k+1}\|$$

$$\le \sum_{d=1}^D \|\boldsymbol{X}_d^{k+1} - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^{k+1}\|$$

$$= \sum_{d=1}^D \|\boldsymbol{X}_d^{k+1} - \widehat{\boldsymbol{X}}_d^{k+1} + \widehat{\boldsymbol{X}}_d^{k+1} - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^{k+1}\|$$

$$\le \sum_{d=1}^D \|\boldsymbol{X}_d^{k+1} - \widehat{\boldsymbol{X}}_d^{k+1}\| + \sum_{d=1}^D \|\widehat{\boldsymbol{X}}_d^{k+1} - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^{k+1}\|. \qquad (57)$$

Based on (53), (56) and the initialization $\widehat{\boldsymbol{X}}_d^0 = \boldsymbol{X}_d^0$ for $d = 1, \ldots, D$, the first term at the right-hand side of (57) can be bounded by

$$\|\boldsymbol{X}_d^{k+1} - \widehat{\boldsymbol{X}}_d^{k+1}\|$$

$$= \left\|\bar{\boldsymbol{Y}}_0^k(d)\boldsymbol{X}_d^0 + \sum_{g=0}^k \bar{\boldsymbol{Y}}_{g+1}^k(d)\triangle\boldsymbol{\Theta}_d^{g+1}\right.$$

$$\left. - \mathbf{1}\boldsymbol{p}_0^T(d)\widehat{\boldsymbol{X}}_d^0 - \mathbf{1}\sum_{g=0}^k \boldsymbol{p}_{g+1}^T(d)\triangle\boldsymbol{\Theta}_d^{g+1}\right\|$$

$$\le \left\|\left(\bar{\boldsymbol{Y}}_0^k(d) - \mathbf{1}\boldsymbol{p}_0^T(d)\right)\boldsymbol{X}_d^0\right\|$$

$$+ \sum_{g=0}^k \left\|\left(\bar{\boldsymbol{Y}}_{g+1}^k(d) - \mathbf{1}\boldsymbol{p}_{g+1}^T(d)\right)\triangle\boldsymbol{\Theta}_d^{g+1}\right\|. \qquad (58)$$

For the first term at the right-hand side of (58), since $\bar{\boldsymbol{Y}}_0^k(d)$ and $\mathbf{1}\boldsymbol{p}_0^T(d)$ both are row-stochastic matrices, $(\bar{\boldsymbol{Y}}_0^k(d) - \mathbf{1}\boldsymbol{p}_0^T(d))\mathbf{1} = \mathbf{0}$. Then, given the initialization $\widehat{\boldsymbol{X}}_d^0 = \boldsymbol{X}_d^0$ for $d = 1, \ldots, D$, we have

$$\|\left(\bar{\boldsymbol{Y}}_0^k(d) - \mathbf{1}\boldsymbol{p}_0^T(d)\right)\boldsymbol{X}_d^0\| = 0. \qquad (59)$$

For the second term at the right-hand side of (58), based on Lemma 1, we have

$$\|\left(\bar{\boldsymbol{Y}}_{g+1}^k(d) - \mathbf{1}\boldsymbol{p}_{g+1}^T(d)\right)\triangle\boldsymbol{\Theta}_d^{g+1}\|$$

$$\le \|\bar{\boldsymbol{Y}}_{g+1}^k(d) - \mathbf{1}\boldsymbol{p}_{g+1}^T(d)\|\|\triangle\boldsymbol{\Theta}_d^{g+1}\|$$

$$\le 2|\mathcal{H}|\sqrt{(u)^{k-g}}\|\triangle\boldsymbol{\Theta}_d^{g+1}\|. \qquad (60)$$

Substituting (59) and (60) into (58) and rearranging the terms, we have

$$\|\boldsymbol{X}_d^{k+1} - \widehat{\boldsymbol{X}}_d^{k+1}\| \le 2|\mathcal{H}|\sum_{g=0}^k \sqrt{(u)^{k-g}}\|\triangle\boldsymbol{\Theta}_d^{g+1}\|. \qquad (61)$$

Next, we analyze the second term at the right-hand side of (57). We rewrite as (56) as

$$\widehat{\boldsymbol{X}}_d^{k+1} = \widehat{\boldsymbol{X}}_d^k + \mathbf{1}\boldsymbol{p}_{k+1}^T(d)\triangle\boldsymbol{\Theta}_d^{k+1}. \qquad (62)$$

Therefore, based on (62) and the initialization $\widehat{\boldsymbol{X}}_d^0 = \boldsymbol{X}_d^0 = \boldsymbol{\Theta}_d^0$ for $d = 1, \ldots, D$, we have

$$\widehat{\boldsymbol{X}}_d^{k+1} - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^{k+1}$$

$$= \widehat{\boldsymbol{X}}_d^k + \mathbf{1}\boldsymbol{p}_{k+1}^T(d)\triangle\boldsymbol{\Theta}_d^{k+1} - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^k - \frac{1}{|\mathcal{H}|}\mathbf{1}\mathbf{1}^T\triangle\boldsymbol{\Theta}_d^{k+1}$$

$$= \widehat{\boldsymbol{X}}_d^k - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^k + \left(\mathbf{1}\boldsymbol{p}_{k+1}^T(d) - \frac{1}{|\mathcal{H}|}\mathbf{1}\mathbf{1}^T\right)\triangle\boldsymbol{\Theta}_d^{k+1}$$

$$= \boldsymbol{X}_d^0 - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^0 + \sum_{g=0}^k \left(\mathbf{1}\boldsymbol{p}_{g+1}^T(d) - \frac{1}{|\mathcal{H}|}\mathbf{1}\mathbf{1}^T\right)\triangle\boldsymbol{\Theta}_d^{g+1}$$

$$= \boldsymbol{\Theta}_d^0 - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^0 + \sum_{g=0}^k \left(\mathbf{1}\boldsymbol{p}_{g+1}^T(d) - \frac{1}{|\mathcal{H}|}\mathbf{1}\mathbf{1}^T\right)\triangle\boldsymbol{\Theta}_d^{g+1}. \qquad (63)$$

Consequently, it holds that

$$\|\widehat{\boldsymbol{X}}_d^{k+1} - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^{k+1}\|$$

$$\leq \|\boldsymbol{\Theta}_d^0 - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^0\| + \sum_{g=0}^{k} \|\mathbf{1}\boldsymbol{p}_{g+1}^T(d) - \frac{1}{|\mathcal{H}|}\mathbf{1}\mathbf{1}^T\| \|\triangle\boldsymbol{\Theta}_d^{g+1}\|$$

$$= \|\boldsymbol{\Theta}_d^0 - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^0\| + \sqrt{|\mathcal{H}|}\sum_{g=0}^{k} \|\boldsymbol{p}_{g+1}^T(d) - \frac{1}{|\mathcal{H}|}\mathbf{1}^T\| \|\triangle\boldsymbol{\Theta}_d^{g+1}\|$$

$$\leq \|\boldsymbol{\Theta}_d^0 - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^0\| + \sqrt{P_{\mathcal{G}}}\sum_{g=0}^{k} \|\triangle\boldsymbol{\Theta}_d^{g+1}\|, \tag{64}$$

where $P_{\mathcal{G}}$ is the degree of network unsaturation defined as $P_{\mathcal{G}} := \frac{|\mathcal{H}|}{\min_{i \in \mathcal{H}}\{|\mathcal{H}_i| + |\mathcal{B}_i| - 2b + 1\}} - 1$. The second equality utilizes the fact that $\|\mathbf{1}\| = \sqrt{|\mathcal{H}|}$. The last inequality comes from the fact $\|\boldsymbol{p}_{g+1}^T(d) - \frac{1}{|\mathcal{H}|}\mathbf{1}^T\| \leq \sqrt{\frac{P_{\mathcal{G}}}{|\mathcal{H}|}}$ since $\boldsymbol{p}_{g+1}(d)$ is a stochastic vector and all of its elements are within the range of $\left[0, \frac{1}{\min_{i \in \mathcal{H}}\{|\mathcal{H}_i| + |\mathcal{B}_i| - 2b + 1\}}\right]$.

Substituting (61) and (64) into (57), we have

$$\|\boldsymbol{X}^{k+1} - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^{k+1}\|$$

$$\leq 2|\mathcal{H}|\sum_{g=0}^{k}\sqrt{(u)^{k-g}}\sum_{d=1}^{D}\|\triangle\boldsymbol{\Theta}_d^{g+1}\| + \sum_{d=1}^{D}\|\boldsymbol{\Theta}_d^0 - \mathbf{1}\bar{\boldsymbol{\Theta}}_d^0\|$$

$$+ \sqrt{P_{\mathcal{G}}}\sum_{d=1}^{D}\sum_{g=0}^{k}\|\triangle\boldsymbol{\Theta}_d^{g+1}\|$$

$$\leq 2|\mathcal{H}|\sqrt{D}\sum_{g=0}^{k}\sqrt{(u)^{k-g}}\|\triangle\boldsymbol{\Theta}^{g+1}\| + \sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|$$

$$+ \sqrt{P_{\mathcal{G}}}\sqrt{D}\sum_{g=0}^{k}\|\triangle\boldsymbol{\Theta}^{g+1}\|. \tag{65}$$

Recall that in DRA, the term $\|\boldsymbol{x}^{k+1} - \boldsymbol{A}^T\bar{\boldsymbol{\theta}}^{k+1}\|$ vanishes asymptotically. Unfortunately, due to the Byzantine attacks and the CTM aggregation, in BREDA $\|\boldsymbol{X}^{k+1} - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^{k+1}\|$ cannot converge to 0. Even worse, it can go unbounded when we use a constant step size. This explains why we will consider an elaborated diminishing step size in the following proof.

According to (26), using the non-expansive property of the projection operator, we have

$$\|\triangle\boldsymbol{\Theta}^{k+1}\|$$

$$= \|\boldsymbol{\Theta}^{k+1} - \boldsymbol{\Theta}^k\|$$

$$= \left\|\Pi_{\tilde{C}}\left[\boldsymbol{\Theta}^k - \gamma^k\left(\nabla_{\boldsymbol{\Theta}}f\left(\boldsymbol{\Theta}^k\right) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\Theta}^k\right)\right] - \Pi_{\tilde{C}}\left[\boldsymbol{\Theta}^k\right]\right\|$$

$$\leq \left\|\boldsymbol{\Theta}^k - \gamma^k\left(\nabla_{\boldsymbol{\Theta}}f\left(\boldsymbol{\Theta}^k\right) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\Theta}^k\right) - \boldsymbol{\Theta}^k\right\|$$

$$= \gamma^k\left\|\nabla_{\boldsymbol{\Theta}}f\left(\boldsymbol{\Theta}^k\right) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\Theta}^k\right\|$$

$$\leq \gamma^k M, \tag{66}$$

where $M := \max_k \|\nabla_{\boldsymbol{\Theta}}f(\boldsymbol{\Theta}^k) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\Theta}^k\|$.

Substituting (66) into (65) and rearranging terms, we have

$$\|\boldsymbol{X}^{k+1} - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^{k+1}\|$$

$$\leq 2|\mathcal{H}|\sqrt{D}M\sum_{g=0}^{k}\sqrt{(u)^{k-g}}\gamma^g + \sqrt{P_{\mathcal{G}}}\sqrt{D}M\sum_{g=0}^{k}\gamma^g$$

$$+ \sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|. \tag{67}$$

Step 4: Reach conclusion.

Finally, combining (50), (52) and (67), we have

$$\underbrace{\begin{bmatrix} \|\boldsymbol{\Theta}^{k+1} - \boldsymbol{\Theta}_v^*\| \\ \|\boldsymbol{\lambda}^{k+1} - \tilde{\boldsymbol{A}}^T\boldsymbol{\lambda}_v^*\| \end{bmatrix}}_{:=\boldsymbol{Z}^{k+1}\in\mathbb{R}^2}$$

$$\leq \underbrace{\begin{bmatrix} 1 - \gamma^k v + \gamma^k L & \gamma^k \\ \gamma^k & 1 - \gamma^k v \end{bmatrix}}_{:=\tilde{\boldsymbol{G}}^k\in\mathbb{R}^{2\times2}} \underbrace{\begin{bmatrix} \|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}_v^*\| \\ \|\boldsymbol{\lambda}^k - \tilde{\boldsymbol{A}}^T\boldsymbol{\lambda}_v^*\| \end{bmatrix}}_{:=\boldsymbol{Z}^k\in\mathbb{R}^2} + \underbrace{\begin{bmatrix} \Omega_1^k \\ \Omega_2^k \end{bmatrix}}_{:=\boldsymbol{\Omega}^k\in\mathbb{R}^2},$$

$$\tag{68}$$

where $\Omega_1^k := \gamma^k\sqrt{|\mathcal{H}|}\left(\frac{1}{|\mathcal{H}|} - \frac{1}{N}\right)\|\boldsymbol{\lambda}_v^*\|, \Omega_2^k := 2\gamma^k|\mathcal{H}|\sqrt{D}M \sum_{g=0}^{k-1}\sqrt{(u)^{k-g-1}}\gamma^g + \gamma^k\sqrt{P_{\mathcal{G}}}\sqrt{D}M\sum_{g=0}^{k-1}\gamma^g + \gamma^k\sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|$.

Taking ℓ_2-norms on both sides of (68), we have

$$\|\boldsymbol{Z}^{k+1}\| \leq \|\tilde{\boldsymbol{G}}^k\|\|\boldsymbol{Z}^k\| + \|\boldsymbol{\Omega}^k\|$$

$$\leq \|\tilde{\boldsymbol{G}}^k\|\|\boldsymbol{Z}^k\| + \Omega_1^k + \Omega_2^k. \tag{69}$$

According to Theorem 5.6.9 in [45], for symmetric real matrix $\tilde{\boldsymbol{G}}^k$, the spectral norm and the spectral radius are equal. Below we give the conditions under which $\|\tilde{\boldsymbol{G}}^k\| = \rho(\tilde{\boldsymbol{G}}^k) < 1$.

Denote any real eigenvalue of $\tilde{\boldsymbol{G}}^k$ as $\tilde{\tau}^k$. Based on the definition of $\tilde{\boldsymbol{G}}^k$ and Gershgorin Circle Theorem, we have

$$\begin{cases} |\tilde{\tau}^k - (1 - \gamma^k v + \gamma^k L)| \leq \gamma^k, \\ |\tilde{\tau}^k - (1 - \gamma^k v)| \leq \gamma^k. \end{cases} \tag{70}$$

To guarantee that the eigenvalues of $\tilde{\boldsymbol{G}}^k$ are all inside the unit circle, it suffices that the regularization parameter v satisfies $v > 1 + L$, the step size γ^k satisfies $\gamma^k < \frac{2}{2+L}$, while v and γ^k jointly satisfy $v < \frac{2}{\gamma^k} - 1$. Under the conditions described above, the eigenvalues of $\tilde{\boldsymbol{G}}^k$ are $\tilde{\tau}_1^k = 1 - \frac{2v - L + \sqrt{4 + L^2}}{2}\gamma^k$ and $\tilde{\tau}_2^k = 1 - \frac{2v - L - \sqrt{4 + L^2}}{2}\gamma^k$, both within $(-1, 1)$ and $\tilde{\tau}_1^k \leq \tilde{\tau}_2^k$.

Only having $\|\tilde{\boldsymbol{G}}^k\| < 1$ is insufficient to guarantee bounded $\|\boldsymbol{Z}^{k+1}\|$, because the error terms Ω_1^k and Ω_2^k in (69) can accumulate over time. Therefore, we must resort to a delicate diminishing step size to bound the error.

According to the step size rule $\gamma^0 = \underline{\gamma}$ and $\gamma^k = \min\{\underline{\gamma}, \frac{\bar{\gamma}}{k^\epsilon}\}$ with $\epsilon > 1$ for $k \geq 1$, there exists a smallest integer $k_0 > 1$ satisfying $\underline{\gamma} \geq \frac{\bar{\gamma}}{k_0^\epsilon}$. When $k < k_0$, $\gamma^k = \underline{\gamma}$. When $k \geq k_0$, $\gamma^k = \frac{\bar{\gamma}}{k^\epsilon}$. With a proper step size γ^k such that $1 > \frac{2v - L + \sqrt{4 + L^2}}{2}\gamma^k$, $0 <$

$\tilde{\tau}_1^k < \tilde{\tau}_2^k$ and thus $\|\tilde{\boldsymbol{G}}^k\| = \max\{|\tilde{\tau}_1^k|, |\tilde{\tau}_2^k|\} = \tilde{\tau}_2^k = 1 - \eta\gamma^k$, where $\eta = \frac{2v - L - \sqrt{4 + L^2}}{2}$. We further require $\eta\underline{\gamma} \in (0, 1)$ and $(\epsilon - 1)(k_0 - 1)^{\epsilon-1} > \eta\overline{\gamma}$. Based on (69), for all $k < k_0$, we have

$$\|\boldsymbol{Z}^{k+1}\|$$

$$\leq (1 - \eta\underline{\gamma})\|\boldsymbol{Z}^k\| + \underline{\gamma}\sqrt{|\mathcal{H}|}\left(\frac{1}{|\mathcal{H}|} - \frac{1}{N}\right)\|\boldsymbol{\lambda}_v^*\|$$

$$+ 2\underline{\gamma}|\mathcal{H}|\sqrt{D}M\sum_{g=0}^{k-1}\sqrt{(u)^{k-g-1}}\gamma^g$$

$$+ \underline{\gamma}\sqrt{P_{\mathcal{G}}}\sqrt{D}M\sum_{g=0}^{k-1}\gamma^g + \underline{\gamma}\sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|$$

$$\leq (1 - \eta\underline{\gamma})\|\boldsymbol{Z}^k\| + (\underline{\gamma})^2\left(\frac{2|\mathcal{H}|\sqrt{D}M}{1 - \sqrt{u}} + k_0\sqrt{P_{\mathcal{G}}}\sqrt{D}M\right)$$

$$+ \underline{\gamma}\left(\sqrt{|\mathcal{H}|}\left(\frac{1}{|\mathcal{H}|} - \frac{1}{N}\right)\|\boldsymbol{\lambda}_v^*\| + \sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|\right)$$

$$= (1 - \eta\underline{\gamma})\|\boldsymbol{Z}^k\| + (\underline{\gamma})^2\Delta_1 + \underline{\gamma}\Delta_2. \tag{71}$$

Here for simplicity, denote $\Delta_1 = \frac{2|\mathcal{H}|\sqrt{D}M}{1 - \sqrt{u}} + k_0\sqrt{P_{\mathcal{G}}}\sqrt{D}M$ and $\Delta_2 = \sqrt{|\mathcal{H}|}\left(\frac{1}{|\mathcal{H}|} - \frac{1}{N}\right)\|\boldsymbol{\lambda}_v^*\| + \sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|$.

Since $\eta\underline{\gamma} \in (0, 1)$, applying telescopic cancellation to (71) through iteration 0 to $k < k_0$, we have for all $k < k_0$ that

$$\|\boldsymbol{Z}^{k+1}\| \leq (1 - \eta\underline{\gamma})^{k+1}\|\boldsymbol{Z}^0\| + \frac{1}{\eta}\left(\underline{\gamma}\Delta_1 + \Delta_2\right). \tag{72}$$

Based on (69), for all $k \geq k_0$, we have

$$\|\boldsymbol{Z}^{k+1}\|$$

$$\leq \left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right)\|\boldsymbol{Z}^k\| + \frac{\overline{\gamma}}{k^\epsilon}\sqrt{|\mathcal{H}|}\left(\frac{1}{|\mathcal{H}|} - \frac{1}{N}\right)\|\boldsymbol{\lambda}_v^*\|$$

$$+ 2\frac{\overline{\gamma}}{k^\epsilon}|\mathcal{H}|\sqrt{D}M\sum_{g=0}^{k-1}\sqrt{(u)^{k-g-1}}\gamma^g$$

$$+ \frac{\overline{\gamma}}{k^\epsilon}\sqrt{P_{\mathcal{G}}}\sqrt{D}M\sum_{g=0}^{k-1}\gamma^g + \frac{\overline{\gamma}}{k^\epsilon}\sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|$$

$$\leq \left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right)\|\boldsymbol{Z}^k\| + \frac{\overline{\gamma}}{k^\epsilon}\left[\sqrt{|\mathcal{H}|}\left(\frac{1}{|\mathcal{H}|} - \frac{1}{N}\right)\|\boldsymbol{\lambda}_v^*\|\right.$$

$$+ 2|\mathcal{H}|\sqrt{D}M\left(k_0\underline{\gamma} + \sum_{g=k_0}^{k-1}\sqrt{(u)^{k-g-1}}\gamma^g\right)$$

$$+ \sqrt{P_{\mathcal{G}}}\sqrt{D}M(k_0\underline{\gamma} + \frac{\overline{\gamma}}{(\epsilon-1)(k_0-1)^{\epsilon-1}})$$

$$\left.+ \sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|\right]. \tag{73}$$

To derive the second inequality, we use the fact that $\sum_{g=k_0}^{k-1}\gamma^g = \sum_{g=k_0}^{k-1}\frac{\overline{\gamma}}{g^\epsilon} < \int_{k_0-1}^{k-1}\frac{\overline{\gamma}}{g^\epsilon}\mathrm{d}g < \frac{\overline{\gamma}}{(\epsilon-1)(k_0-1)^{\epsilon-1}}$.

We bound $\sum_{g=k_0}^{k-1}\sqrt{(u)^{k-g-1}}\gamma^g$ in (73) using the following Lemma.

Lemma 2 ([40], Lemma 5): If for any $k \geq k_0$, step size γ^k satisfies

$$1 \leq \frac{\gamma^k}{\gamma^{k+1}} \leq \frac{2}{1 + \psi_1}, \tag{74}$$

and for some $\psi_1 \in (0, 1)$ and $\psi_2 \geq 0$, iterates $\{y^k\}$ satisfy

$$y^{k+1} \leq \psi_1 y^k + \psi_2\gamma^k \text{ and } y^{k_0} \leq \psi_2\gamma^{k_0}, \tag{75}$$

then y^k has an upper bound

$$y^k \leq \frac{2\psi_2}{1 - \psi_1}\gamma^k. \tag{76}$$

To bound $\sum_{g=k_0}^{k-1}\sqrt{(u)^{k-g-1}}\gamma^g$ in (73), we define y^k as

$$y^k := \sum_{g=k_0}^{k-1}\sqrt{(u)^{k-g-1}}\gamma^g, \tag{77}$$

which satisfy the relation $y^{k+1} = \sqrt{u}y^k + \gamma^k$, $\psi_1 = \sqrt{u} \in (0, 1)$, $\psi_2 = 1 \geq 0$ and $y^{k_0} = 0 \leq \gamma^{k_0}$. According to Lemma 2, when for $k \geq k_0$ step size $\gamma^k = \frac{\overline{\gamma}}{k^\epsilon}$ satisfies $1 \leq \frac{\gamma^k}{\gamma^{k+1}} \leq \frac{2}{1+\sqrt{u}}$, i.e., $1 \leq (\frac{k_0+1}{k_0})^\epsilon \leq \frac{2}{1+\sqrt{u}}$, we have

$$y^k \leq \frac{2}{1 - \sqrt{u}}\gamma^k. \tag{78}$$

Substituting (78) into (73) and rearranging the terms, we have

$$\|\boldsymbol{Z}^{k+1}\|$$

$$\leq \left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right)\|\boldsymbol{Z}^k\| + \frac{\overline{\gamma}}{k^\epsilon}\left[\sqrt{|\mathcal{H}|}\left(\frac{1}{|\mathcal{H}|} - \frac{1}{N}\right)\|\boldsymbol{\lambda}_v^*\|\right.$$

$$+ 2|\mathcal{H}|\sqrt{D}Mk_0\underline{\gamma} + \sqrt{P_{\mathcal{G}}}\sqrt{D}M\left(k_0\underline{\gamma} + \frac{\overline{\gamma}}{(\epsilon-1)(k_0-1)^{\epsilon-1}}\right)$$

$$\left.+ \sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|\right] + \left(\frac{\overline{\gamma}}{k^\epsilon}\right)^2\frac{4|\mathcal{H}|\sqrt{D}M}{1 - \sqrt{u}}$$

$$= \left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right)\|\boldsymbol{Z}^k\| + \frac{\overline{\gamma}\Delta_3}{k^\epsilon} + \left(\frac{\overline{\gamma}}{k^\epsilon}\right)^2\Delta_4, \tag{79}$$

in which $\Delta_3 = \sqrt{|\mathcal{H}|}\left(\frac{1}{|\mathcal{H}|} - \frac{1}{N}\right)\|\boldsymbol{\lambda}_v^*\| + 2|\mathcal{H}|\sqrt{D}Mk_0\underline{\gamma} + \sqrt{P_{\mathcal{G}}}\sqrt{D}M\left(k_0\underline{\gamma} + \frac{\overline{\gamma}}{(\epsilon-1)(k_0-1)^{\epsilon-1}}\right) + \sqrt{D}\|\boldsymbol{\Theta}^0 - \tilde{\boldsymbol{A}}^T\bar{\boldsymbol{\Theta}}^0\|$ and $\Delta_4 = \frac{4|\mathcal{H}|\sqrt{D}M}{1 - \sqrt{u}}$.

Since $\frac{\eta\overline{\gamma}}{k^\epsilon} \in (0, 1)$, applying telescopic cancellation to (79) through iteration k_0 to $k \geq k_0$, we have for all $k \geq k_0$ that

$$\|\boldsymbol{Z}^{k+1}\|$$

$$\leq \prod_{g=k_0}^{k}\left(1 - \frac{\eta\overline{\gamma}}{g^\epsilon}\right)\|\boldsymbol{Z}^{k_0}\| + \prod_{g=k_0+1}^{k}\left(1 - \frac{\eta\overline{\gamma}}{g^\epsilon}\right)\frac{\overline{\gamma}\Delta_3}{k_0^\epsilon}$$

$$+ \prod_{g=k_0+2}^{k}\left(1 - \frac{\eta\overline{\gamma}}{g^\epsilon}\right)\frac{\overline{\gamma}\Delta_3}{(k_0+1)^\epsilon} + \cdots$$

$$+ \left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right) \frac{\overline{\gamma}\Delta_3}{(k-1)^\epsilon} + \frac{\overline{\gamma}\Delta_3}{k^\epsilon}$$

$$+ \prod_{g=k_0+1}^{k} \left(1 - \frac{\eta\overline{\gamma}}{g^\epsilon}\right) \frac{\overline{\gamma}^2 \Delta_4}{k_0^{2\epsilon}} + \prod_{g=k_0+2}^{k} \left(1 - \frac{\eta\overline{\gamma}}{g^\epsilon}\right)$$

$$\times \frac{\overline{\gamma}^2 \Delta_4}{(k_0+1)^{2\epsilon}} + \cdots$$

$$+ \left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right) \frac{\overline{\gamma}^2 \Delta_4}{(k-1)^{2\epsilon}} + \frac{\overline{\gamma}^2 \Delta_4}{k^{2\epsilon}} \qquad (80)$$

$$\leq \frac{1}{1 + \sum_{g=k_0}^{k} \frac{\eta\overline{\gamma}}{g^\epsilon}} \|\boldsymbol{Z}^{k_0}\| + \left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right) \overline{\gamma}\Delta_3 \sum_{g=k_0}^{k-1} \frac{1}{g^\epsilon} + \frac{\overline{\gamma}\Delta_3}{k^\epsilon}$$

$$+ \left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right) \overline{\gamma}^2 \Delta_4 \sum_{g=k_0}^{k-1} \frac{1}{g^{2\epsilon}} + \frac{\overline{\gamma}^2 \Delta_4}{k^{2\epsilon}}$$

$$\leq \frac{1}{1 + \frac{\eta\overline{\gamma}}{\epsilon-1}\left[\frac{1}{k_0^{\epsilon-1}} - \frac{1}{(k+1)^{\epsilon-1}}\right]} \|\boldsymbol{Z}^{k_0}\| + \frac{\left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right)\overline{\gamma}\Delta_3}{(\epsilon-1)(k_0-1)^{\epsilon-1}} + \frac{\overline{\gamma}\Delta_3}{k^\epsilon}$$

$$+ \frac{\left(1 - \frac{\eta\overline{\gamma}}{k^\epsilon}\right)\overline{\gamma}^2 \Delta_4}{(2\epsilon-1)(k_0-1)^{2\epsilon-1}} + \frac{\overline{\gamma}^2 \Delta_4}{k^{2\epsilon}},$$

where the second inequality utilizes the fact that $\sum_{g=k_0}^{k} \frac{\eta\overline{\gamma}}{g^\epsilon} < \frac{\eta\overline{\gamma}}{(\epsilon-1)(k_0-1)^{\epsilon-1}} < 1$. To drive the last inequality, we use the fact that $\sum_{g=k_0}^{k} \frac{\eta\overline{\gamma}}{g^\epsilon} > \int_{k_0}^{k+1} \frac{\eta\overline{\gamma}}{g^\epsilon} \mathrm{d}g = \frac{\eta\overline{\gamma}}{\epsilon-1}\left[\frac{1}{k_0^{\epsilon-1}} - \frac{1}{(k+1)^{\epsilon-1}}\right]$. Taking $k \to +\infty$, we have

$$\limsup_{k \to +\infty} \|\boldsymbol{Z}^{k+1}\| \leq \frac{(\epsilon-1)k_0^{\epsilon-1}}{(\epsilon-1)k_0^{\epsilon-1} + \eta\overline{\gamma}} \|\boldsymbol{Z}^{k_0}\|$$

$$+ \frac{\overline{\gamma}\Delta_3}{(\epsilon-1)(k_0-1)^{\epsilon-1}} + \frac{\overline{\gamma}^2 \Delta_4}{(2\epsilon-1)(k_0-1)^{2\epsilon-1}}. \qquad (81)$$

Below we summarize the conditions on the parameters $\underline{\gamma}$, $\overline{\gamma}$ and v. The regularization parameter v satisfies $v > 1 + \underline{L}$, $\underline{\gamma}$ satisfies $\underline{\gamma} < \frac{2}{2+L}$ and guarantees $\left(\frac{k_0+1}{k_0}\right)^\epsilon \leq \frac{2}{1+\sqrt{u}}$, v and $\underline{\gamma}$ jointly satisfy $v < \frac{2}{\underline{\gamma}} - 1$ and $v < \frac{1}{\underline{\gamma}} + \frac{L-\sqrt{4+L^2}}{2}$, while v and $\overline{\gamma}$ jointly satisfy $v < \frac{2(\epsilon-1)(k_0-1)^{\epsilon-1}}{2\overline{\gamma}} + \frac{L+\sqrt{4+L^2}}{2}$. This completes the proof.

Remark 6: Recall that the constant $u \in (0,1)$ monotonically increases as $\tau_{\mathcal{G}}$, the maximum network diameter after CTM, increases. A well-connected network has small $\tau_{\mathcal{G}}$ and u. In consequence, according to the condition $\left(\frac{k_0+1}{k_0}\right)^\epsilon \leq \frac{2}{1+\sqrt{u}}$, it will lead to a large ϵ. Since (80) implies that BREDA inexactly converges at a rate of $O\left(\frac{1}{k^\epsilon}\right)$, we can see that better connectedness yields a faster convergence rate for BREDA.

Remark 7: In BREDA, the existence of Byzantine attacks and the CTM operation bring the convergence error. To guarantee bounded convergence error, we require that $\|\boldsymbol{\lambda}_v^*\|$ is bounded. It holds true as long as $v > 0$ [4], [47]. We also require that $M := \max_k \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}^k) + \frac{1}{N}\boldsymbol{\lambda}^k + v\boldsymbol{\Theta}^k\|$ is bounded, which means

that $\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}^k)\|$, $\|\frac{1}{N}\boldsymbol{\lambda}^k\|$ and $\|v\boldsymbol{\Theta}^k\|$ are bounded. Due to the bounded projections in terms of both the primal and dual variables, $\|\frac{1}{N}\boldsymbol{\lambda}^k\|$ and $\|v\boldsymbol{\Theta}^k\|$ are bounded, and so is the gradient norm $\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}^k)\|$.

REFERENCES

[1] R. Wang, Y. Liu, and Q. Ling, "Byzantine-resilient decentralized resource allocation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 5293–5297.

[2] F. P. Kelly, A. K. Maulloo, and D. K. Tan, "Rate control for communication networks: Shadow prices, proportional fairness, and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, 1998.

[3] B. Turan, C. A. Uribe, H. 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, Mar. 2021.

[4] J. Koshal, A. Nedic, and U. V. Shanbhag, "Multiuser optimization: Distributed algorithms and error analysis," *SIAM J. Optim.*, vol. 21, no. 3, pp. 1046–1081, 2011.

[5] L. Xiao and S. Boyd, "Optimal scaling of a gradient method for distributed resource allocation," *J. Optim. Theory Appl.*, vol. 129, no. 3, pp. 469–488, 2006.

[6] H. Lakshmanan and D. P. De Farias, "Decentralized resource allocation in dynamic networks of agents," *SIAM J. Optim.*, vol. 19, no. 2, pp. 911–940, 2008.

[7] S. Liang, X. Zeng, and Y. Hong, "Distributed sub-optimal resource allocation over weight-balanced graph via singular perturbation," *Automatica*, vol. 95, pp. 222–228, 2018.

[8] S. Liang, X. Zeng, G. Chen, and Y. Hong, "Distributed sub-optimal resource allocation via a projected form of singular perturbation," *Automatica*, vol. 121, 2020, Art. no. 109180.

[9] W. Lin, Y. Wang, C. Li, and X. Yu, "Distributed resource allocation: An indirect dual ascent method with an exponential convergence rate," *Nonlinear Dyn.*, vol. 102, pp. 1685–1699, 2020.

[10] J. Zhang, K. You, and K. Cai, "Distributed dual gradient tracking for resource allocation in unbalanced networks," *IEEE Trans. Signal Process.*, vol. 68, pp. 2186–2198, 2020.

[11] T. T. Doan and C. L. Beck, "Distributed resource allocation over dynamic networks with uncertainty," *IEEE Trans. Autom. Control*, vol. 66, no. 9, pp. 4378–4384, Sep. 2021.

[12] A. Camisa, F. Farina, I. Notarnicola, and G. Notarstefano, "Distributed constraint-coupled optimization via primal decomposition over random time-varying graphs," *Automatica*, vol. 131, 2021, Art. no. 109739.

[13] M. Zargham, A. Ribeiro, A. Ozdaglar, and A. Jadbabaie, "Accelerated dual descent for network flow optimization," *IEEE Trans. Autom. Control*, vol. 59, no. 4, pp. 905–920, Apr. 2014.

[14] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A distributed Newton method for network utility maximization–I: Algorithm," *IEEE Trans. Autom. Control*, vol. 58, no. 9, pp. 2162–2175, Sep. 2013.

[15] E. Ghadimi, I. Shames, and M. Johansson, "Multi-step gradient methods for networked optimization," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5417–5429, Nov. 2013.

[16] A. Bedi and K. Rajawat, "Asynchronous incremental stochastic dual descent algorithm for network resource allocation," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2229–2244, May 2018.

[17] S. A. Alghunaim, K. Yuan, and A. H. Sayed, "A proximal diffusion strategy for multiagent optimization with sparse affine constraints," *IEEE Transaction Autom. Control*, vol. 65, no. 11, pp. 4554–4567, Nov. 2020.

[18] Z. Deng, X. Nian, and C. Hu, "Distributed algorithm design for nonsmooth resource allocation problems," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3208–3217, Jul. 2019.

[19] T. Chen, Q. Ling, and G. B. Giannakis, "An online convex optimization approach to proactive network resource allocation," *IEEE Trans. Signal Process.*, vol. 65, no. 24, pp. 6350–6364, Dec. 2017.

[20] T. Chen, A. Mokhtari, X. Wang, A. Ribeiro, and G. B. Giannakis, "Stochastic averaging for constrained optimization with application to online resource allocation," *IEEE Trans. Signal Process.*, vol. 65, no. 12, pp. 3078–3093, Jun. 2017.

[21] X. Yi, X. Li, L. Xie, and K. Johansson, "Distributed online convex optimization with time-varying coupled inequality constraints," *IEEE Trans. Signal Process.*, vol. 68, pp. 731–746, 2020.

[22] L. Lamport, R. E. Shostak, and M. C. Pease, "The Byzantine generals problem," *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 3, pp. 382–401, 1982.

[23] Z. Yang, A. Gang, and W. U. Bajwa, "Adversary-resilient distributed and decentralized statistical inference and machine learning: An overview of recent advances under the byzantine threat model," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 146–159, May 2020.

[24] P. Yi, J. Lei, and Y. Hong, "Distributed resource allocation over random networks based on stochastic approximation," *Syst. Control Lett.*, vol. 114, pp. 44–51, 2018.

[25] L. Chen, H. Wang, Z. Charles, and D. Papailiopoulos, "DRACO: Byzantine-resilient distributed training via redundant gradients," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 903–912.

[26] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 1544–1551.

[27] Z. Wu, Q. Ling, T. Chen, and G. B. Giannakis, "Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks," *IEEE Trans. Signal Process.*, vol. 68, pp. 4583–4596, 2020.

[28] Z. Yang and W. U. Bajwa, "ByRDiE: Byzantine-resilient distributed coordinate descent for decentralized learning," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 4, pp. 611–627, Dec. 2019.

[29] L. Su and N. H. Vaidya, "Byzantine-resilient multiagent optimization," *IEEE Trans. Autom. Control*, vol. 66, no. 5, pp. 2227–2233, May 2021.

[30] C. Fang, Z. Yang, and W. U. Bajwa, "BRIDGE: Byzantine-resilient decentralized gradient descent," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 8, pp. 610–626, Jul. 2022.

[31] S. Sundaram and B. Gharesifard, "Distributed optimization under adversarial nodes," *IEEE Trans. Autom. Control*, vol. 64, no. 3, pp. 1063–1076, Mar. 2019.

[32] J. Li, W. Abbas, and X. Koutsoukos, "Resilient distributed diffusion in networks with adversaries," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 1–7, 2020.

[33] W. Ben-Ameur, P. Bianchi, and J. Jakubowicz, "Robust distributed consensus using total variation," *IEEE Trans. Autom. Control*, vol. 61, no. 6, pp. 1550–1564, Jun. 2015.

[34] W. Xu, Z. Li, and Q. Ling, "Robust decentralized dynamic optimization at presence of malfunctioning agents," *Signal Process.*, vol. 153, pp. 24–33, 2018.

[35] J. Peng, W. Li, and Q. Ling, "Byzantine-robust decentralized stochastic optimization over static and time-varying networks," *Signal Process.*, vol. 183, 2021, Art. no. 108020.

[36] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.

[37] A. V. Gasnikov, E. B. Gasnikova, Y. E. Nesterov, and A. V. Chernov, "Efficient numerical methods for entropy-linear programming problems," *Comput. Math. Math. Phys.*, vol. 56, no. 4, pp. 523–534, 2016.

[38] M. Zhu and S. Martinez, "Discrete-time dynamic average consensus," *Automatica*, vol. 46, no. 2, pp. 322–329, 2014.

[39] A. Koppel, F. Y. Jakubiec, and A. Ribeiro, "A saddle point algorithm for networked online convex optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5149–5164, Oct. 2015.

[40] Z. Wu, H. Shen, T. Chen, and Q. Ling, "Byzantine-resilient decentralized policy evaluation with linear function approximation," *IEEE Trans. Signal Process.*, vol. 69, pp. 3839–3853, 2021.

[41] N. Vaidya, "Matrix representation of iterative approximate byzantine consensus in directed graphs," 2012, *arXiv:1203.1888v1*.

[42] "IEEE 118 bus system." 2015. [Online]. Available: https://www.al-roomi.org/power-flow/118-bus-system

[43] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Optim.*, vol. 25, no. 2, pp. 944–966, 2015.

[44] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, Sep. 2018.

[45] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[46] S. U. Pillai, T. Suel, and S. Cha, "The Perron-Frobenius theorem: Some of its applicatons," *IEEE Signal Process. Mag.*, vol. 22, no. 2, pp. 62–75, Mar. 2005.

[47] H. Uzawa, "Iterative Methods in Concave Programming," in *Studies in Linear and Nonlinear Programming*. Stanford, CA, USA: Stanford Univ. Press, 1958, pp. 154–165.

Runhua Wang (Student Member, IEEE) received the B.E. degree in network engineering from Huaibei Normal University, Huaibei, China, and the M.E. degree in software engineering from Central South University, Changsha, China. She is currently working toward the Ph.D. degree with Sun Yat-Sen University, Guangzhou, China. Her research interests include resource allocation and distributed optimization.

Yaohua Liu received the B.E. degree in automation from the South China University of Technology, Guangzhou, China, in 2015, and the Ph.D. degree in control theory and control engineering from the University of Science and Technology of China, Hefei, China, in 2020. Since August 2020, she has been with the School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China. Her research interests include decentralized network optimization and its applications.

Qing Ling (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in control theory and control engineering from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, Michigan Technological University, Houghton, MI, USA, from 2006 to 2009, and an Associate Professor with the Department of Automation, University of Science and Technology of China, from 2009 to 2017. He is currently a Professor with the School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China. His research interests include distributed and decentralized optimization and its application in machine learning. His work was the recipient of the 2017 IEEE Signal Processing Society Young Author Best Paper Award. He is also the Senior Area Editor of IEEE SIGNAL PROCESSING LETTERS.