Byzantine-resilient Collaborative Hierarchical Non-Bayesian Learning

Connor Mclaughlin*[1], Matthew Ding*[2], Deniz Erdogmus[1], and Lili Su[1]

***Abstract*— Non-Bayesian learning is a computationally efficient approximation of Bayesian learning over multi-agent networks. As the network scale increases, existing fully distributed solutions start to lag behind real-world challenges such as slow information propagation and external adversarial attacks. In this paper, to reduce the potential information propagation delay in large systems, we consider a hierarchical system architecture in which the agents are clustered into M sub-networks, and a parameter server exists to facilitate the information exchange among sub-networks. The message exchange between any client and the parameter server is expensive; hence it needs to be carefully controlled.**

To the best of our knowledge, utilizing hierarchical structure to speed up convergence and to enhance adversarial resilience is largely under-explored, which is our focus. Byzantine resilience via consensus suffers the curse of dimensionality – no Byzantine consensus algorithms can withstand a fraction of Byzantine agents exceeding $\min\{1/3,\ 1/(d+1)\}$ where d is the input dimension. To get around this, we solve the non-Bayesian learning problem via running multiple scalar dynamics. Furthermore, we use a novel Byzantine-resilient gossiping-type rule at the parameter server to facilitate resilient information propagation across sub-networks. We show that under some technical conditions, each normal agent can asymptotically identify the underlying truth hypothesis θ^* with probability 1. Notably, our theory implies that even if there exists a sub-network whose majority of agents are Byzantine, our algorithm still enables successful learning of the normal agents in such sub-networks.

I. Introduction

Non-Bayesian learning [7], [8], [17], [18] is a "consensus + innovation" approach of social learning. It is a computational efficient approximation to Bayesian learning over networks wherein the information is scattered over different agents. Formally, social learning can be formulated as a distributed multiple hypothesis testing problem. Let $\Theta = \{\theta_1, \cdots, \theta_d\}$ be the set of d hypotheses. There is an unknown underlying truth $\theta^* \in \Theta$ that determines the joint distribution of the local measurements at individual agents. For any given $\theta \in \Theta$, the marginal distributions at the agents can be different. "Local confusion" often exists; that is, the marginal distributions of different hypotheses may appear to be the same to an agent. The goal of non-Bayesian learning is to have agents collaboratively identify the unknown θ^*.

As the scale of the multi-agent network increases, existing fully distributed solutions start to lag behind the crucial real-world challenges such as slow information propagation and external adversarial attacks. Towards scalable decentralized solutions, instead of a gigantic multi-agent network, we consider a hierarchical system architecture in which the agents are clusters into M sub-networks, and a parameter server exists to aid the information exchanges among sub-networks. The system architecture is depicted in Fig.1. Similar system architecture is adopted in the literature [6], [2], [21], [11], [10], [3], wherein centralized parameter servers can be placed at the top of the hierarchy to coordinate between client clusters. Sending messages between an agent and the parameter server is costly; hence needs to be sparse.

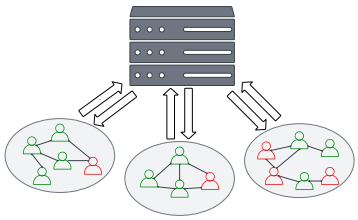

Fig. 1. A hierarchical system architecture

In this paper, we study Byzantine-resilient hierarchical non-Bayesian learning, wherein the compromised agents can send maliciously calibrated messages to others and the parameter server. Previous studies in non-Bayesian learning have explored network structures such as sparse or weakly connected graphs [16], [19], time-varying graphs [12], and more general higher-order hypergraph structures [1]. Hierarchical architectures have been explored in literature on relevant problems [6], [2], [21], [11], [10], [3]. However, existing methods are not applicable to our problem; see Section II for details. To the best of our knowledge, utilizing hierarchical structure to speed up convergence and to enhance adversarial resilience is largely under-explored, which is our focus.

Contributions. Byzantine resilience suffers curse of dimensionality – no Byzantine consensus algorithms can tolerate more than $\min\{1/3,\ 1/(d+1)\}$ fraction of agents to be Byzantine [15]. To avoid this, we solve the non-Bayesian learning problem via running multiple scalar dynamics, each of which only involves Byzantine consensus with scalar inputs. Moreover, we introduce a novel Byzantine-resilient gossiping-type to ensure the effective information propagation across networks. Our algorithm only uses sparse agent-server communication in two senses: First, the fusion among the M sub-networks occurs every other D^* rounds, where D^* is the maximal diameter of the sub-networks. Second, for each fusion, only a subset of clients are selected.

We show that our algorithm is resilient to arbitrary placement of up to F Byzantine agents provided that there exists at least $F+1$ sub-networks each of which satisfies certain condition. [1] Specifically, our algorithm enables each normal

*Student authors with equal contribution.

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston MA {mclaughlin.co, d.erdogmus, l.su}@northeastern.edu

[2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley CA matthewding@berkeley.edu

[1]Formal description of the conditions can be found in Section V.

agent to asymptotically identify θ^* with probability 1 using sparse communication with the parameter server. It is worth noting that even if there exists sub-networks whose majority of the agents are Byzantine, our algorithm still enables the normal agents in such sub-networks to learn θ^*.

II. Related Work

Hierarchical architectures have been explored to speed up the convergence of average consensus [6], [2]. In comparison to using a large single network, Epstein et al. [2] improve the speed of average consensus through hierarchical clustering of agents, followed by iteratively performing consensus on each level of the hierarchy with message passing between adjacent levels. However, their characterized final consensus errors are nonzero. Hou and Zheng [6] adopt a similar network structure with the additional challenge of time-varying communication links. Their local agent updates are based on relative intra-cluster information and relative inter-cluster "group information" when links are active. Unfortunately, such group information is often expensive to collect.

A client-edge-cloud hierarchy has also been explored in the context of edge computing [21], [11], [3], [10]. In the context of load balancing problem, Tong et al. [21] propose to use such hierarchical structure so as to efficiently utilize the cloud resources to serve the peak loads from mobile users. In their system architecture, clients are clustered based on their proximity to edge communication servers. The communication between the clusters and the corresponding edge servers are frequent yet the estimate synchronization among the edge servers is infrequent. Liu et al. [11] apply Federated Learning on such hierarchical systems and show improved convergence of Hierarchical-FedAvg [11] compared to pure cloud-based FedAvg and edge-based FedAvg [14] implementation. Subsequent works have further improved the communication efficiency through client selection based on worst-case delay [3], or an evolutionary process [10]. Departing from above existing literature, we consider Byzantine-resilience. A key technical challenge is that Byzantine agents can launch attacks via injecting adaptively calibrated messages, which destroy the commonly assumed unbiased stochastic gradients condition.

III. System and Threat Models

A. System Model

The system consists of a parameter server (PS) and M sub-networks, each of which is formally represented by graphs $G(\mathcal{V}_i, \mathcal{E}_i)$, where $\mathcal{V}_i = \{v_1^i, \cdots, v_{n_i}^i\}$ is node set and $\mathcal{E}_i$ is the set of all directed edges. Let $N := \sum_{i=1}^M n_i$, where $n_i = |\mathcal{V}_i|$. No messages can be exchanged directly between agents in different sub-networks. In addition, the PS has the freedom in querying and pushing messages to any agent. However, this type of message exchange comes at a high cost and must be limited in frequency. For an arbitrary sub-network S_i, $\mathcal{I}_j^i = \{k \mid (k,j) \in \mathcal{E}_i\}$ and $\mathcal{O}_j^i = \{k \mid (j,k) \in \mathcal{E}_i\}$ denote the sets of incoming and outgoing neighbors to agent j. For notational convenience, $d_j^i := \left|O_j^i\right|$.

Denote the diameter of sub-network i as D_i. Define

$$D^* := \max_{i \in [M]} D_i. \tag{1}$$

Intuitively, the smaller D_i, the faster the information fusion within the network. It is easy to see that, with the hierarchical structure created by the parameter server, the information fusion is expected to be faster provided that the communication cost involves the parameter server is comparable to the cost of agent-agent communication.

B. Threat Model: Byzantine Faults

We adopt Byzantine fault model [13], [9] – a canonical fault model in distributed computing. There exists a system adversary that can choose $\mathcal{A} \subset \cup_{i=1}^M \mathcal{V}_i$ such that $|\mathcal{A}| \leq F$ (where $F \leq N$) to compromise and control. Each agent in $\mathcal{A}$ is referred to as a Byzantine agent. Each normal agent (i.e., an agent in $\cup_{i=1}^M \mathcal{V}_i \setminus \mathcal{F}$) knows F but does not know the set $\mathcal{A}$ and $|\mathcal{A}|$. The system adversary has complete knowledge of the system, including the local program that each good agent is supposed to run and the problem inputs. The Byzantine agents can collude with each other and deviate from their pre-specified local programs to *arbitrarily* misrepresent information to the good agents with the only restriction that the communication channel is authenticated, i.e., a Byzantine agent cannot forge the digital signature of someone else. Moreover, Byzantine agents can use point-to-point rather than broadcast communication. Formally, let $m_{jj_1}(t)$ and $m_{jj_2}(t)$ be the messages sent by agent j to two distinct outgoing neighbors j_1 and j_2. Under point-to-point communication, it is allowed that $m_{jj_1}(t) \neq m_{jj_2}(t)$.

IV. Social Learning Problem

We following a canonical learning model in social networks/multi-agent systems [7], [8], [18]. The entire system can be in one of the d possible unknown environments $\Theta = \{\theta_1, \theta_2, \cdots, \theta_d\}$. Let $\theta^* \in \Theta$ denote the underlying environment that the normal agents try to collaboratively learn based on their locally collected signals.

For each time t, each agent *independently* obtains private signal about the environmental state θ^*, which is initially unknown to every agent in the network. Each agent j in each sub-network i knows the structure of its private signal, which is represented by a collection of parameterized distributions $\mathcal{D}^{ij} = \{\ell_{ij}(s_{ij}|\theta) | \theta \in \Theta, s_{ij} \in \mathcal{S}_{ij}\}$, where $\ell_{ij}(\cdot|\theta)$ is a distribution with parameter $\theta \in \Theta$, and $\sup_{s_{ij} \in \mathcal{S}_{ij}} \sup_{\theta, \theta' \in \Theta} \log \frac{\ell_{ij}(s_{ij}|\theta)}{\ell_{ij}(s_{ij}|\theta')} \leq L$ for some positive constant $L > 0$.

Θ: set of d hypotheses	M: number of sub-networks
$\mathcal{V}_i$: node set of sub-network S_i	$n_i = \|\mathcal{V}_i\|$
$\mathcal{E}_i$: edge set of sub-network S_i	d_j^i: incoming degree of agent j in sub-network S_i
$\mathcal{A}$: set of Byzantine agents	F: upper bound of $\|\mathcal{A}\|$
D_i: diameter of network S_i	$D^* = \max_i D_i$

TABLE I
Reference Notation Chart.

For ease of exposition, we index the agents 1 to N. Let s_t^j be the private signal observed by agent j in iteration t, and let $\mathbf{s}_t = \{s_t^1, \cdots, s_t^N\}$ be the signal profile at time t (i.e., signals observed by the agents in iteration t). Given θ^*, the signal profile $\mathbf{s}_t$ is generated according to the joint distribution $\boldsymbol{\ell}(\cdot \mid \theta^*) = \ell_1(\cdot|\theta^*) \times \cdots \times \ell_N(\cdot|\theta^*)$.

A. Multi-dimensional problems.

In a centralized setting, Bayesian learning methods can be used to identify the underlying truth θ^* via iteratively refining the posterior/belief (which is d-dimensional) based on $\boldsymbol{s}_t$. For fully distributed settings, non-Bayesian learning is a computational efficient approximation to exact Bayesian learning, wherein each agent refines its local estimate of the global posterior (which is also d-dimensional) based on local signal while running consensus update.

V. Non-Bayesian Learning: Byzantine Resilience

Due to the curse of dimensionality of Byzantine resilience, we can not directly plug in a Byzantine consensus algorithm to serve as the "consensus" component. In Algorithm 1, instead of updating any approximation to the global belief vector evolution, we run multiple scalar linear dynamics simultaneously – one for each hypothesis pair of distinct hypotheses θ and θ'. Roughly speaking, $r^j(\theta, \theta')$ is a local approximation to the log-likelihood ratio between hypotheses θ and θ'. Larger $r^j(\theta, \theta')$ implies that in the view of agent j, compared with hypothesis θ', hypothesis θ is more likely to be the underlying truth θ^*. As described in lines 1-3, since no signals s_t^j are collected prior to the algorithm execution, $r^j(\theta, \theta')$ is initialized to 0 (i.e., $r_0^j(\theta, \theta') = 0$).

If a sub-network belongs to a particular set $\mathcal{C}$ (to be specified later), each agent j in that network does a trim + consensus + innovation update as in lines 7-9. Specifically, agent j first trims away the F smallest and F largest values of the received $\widetilde{r}_{t-1}^{j'}(\theta, \theta')$. Here we use the notation $\widetilde{r}_{t-1}^{j'}(\theta, \theta')$ rather than $r_{t-1}^{j'}(\theta, \theta')$ because that when $j' \in \mathcal{A}$ (i.e., when j' is Byzantine) it may lie arbitrarily, resulting in $\widetilde{r}_{t-1}^{j'}(\theta, \theta') \neq r_{t-1}^{j'}(\theta, \theta')$. In line 9, the set $\mathcal{I}_j^*(t, \theta, \theta') \subseteq \mathcal{I}_i$ is the remained incoming neighbors whose messages $\widetilde{r}_{t-1}^{j'}(\theta, \theta')$ are not trimmed away by agent j at time t, and $\log \frac{\ell_j(s_t^j|\theta)}{\ell_j(s_t^j|\theta')}$ is the log-likelihood ratio of the newly obtained local signal s_t^j. The sparse information fusion among the M sub-networks, under of the coordination of the parameter server, is described in lines 10-20, which can be viewed as Byzantine-resilient sparse gossiping. Such information fusion is sparse in two senses: First, as is determined by the **if** command in line 10, the fusion among the M sub-networks occurs every other D^* rounds, recalling that D^* is the maximal diameter of the sub-networks (as per Eq.(1)). Second, for each fusion, only a subset of clients are selected – referred to as representatives – to send local estimates. When $M \geq 2F + 1$, the parameter server uniformly at random chooses one representative from each sub-network. If $M < 2F + 1$, the parameter server samples additional $2F + 1 - M$ agents as in line 14. Since a selected network representatives may be Byzantine, the parameter server trims away the largest F and smallest F received values (line 16) before taking the average, denoted as $\widetilde{w}$. In line 17, the set $\widetilde{\mathcal{R}}(t, \theta, \theta')$ is the set of representatives whose messages are not trimmed away. Since the number of representatives is at least $2F + 1$, $\widetilde{\mathcal{R}}(t, \theta, \theta')$ is non-empty. Via lines 19 and 20, each representative of the sub-networks not in $\mathcal{C}$, its local estimate is updated again as $r_t^j(\theta, \theta') = \widetilde{w}(t, \theta, \theta')$.

Algorithm 1: Hierarchical Byzantine-resilient Non-Bayesian Learning

1 **for** $j \in \cup_{i=1}^M \mathcal{V}_i$ **do**
2 **for** $\theta, \theta' \in \Theta$ *such that* $\theta \neq \theta'$ **do**
3 $r_0^j(\theta, \theta') \leftarrow 0$;
4 In parallel, for each hypothesis pair θ, θ' do:
5 **for** $t = 1, 2, \cdots$ **do**
6 **if** *Agent j belongs to a network in $\mathcal{C}$* **then**
7 Transmit $r_{t-1}^j(\theta, \theta')$ on all outgoing edges;
8 Filter the smallest and largest F values of the received log likelihood ratios $\widetilde{r}_{t-1}^{j'}(\theta, \theta')$
9 $r_t^j(\theta, \theta') \leftarrow \frac{\sum_{j' \in \mathcal{I}_j^*(t,\theta,\theta')} \widetilde{r}_{t-1}^{j'}(\theta,\theta') + r_{t-1}^j(\theta,\theta')}{|\mathcal{I}_j| + 1 - 2F} + \log \frac{\ell_j(s_t^j|\theta)}{\ell_j(s_t^j|\theta')}$.
10 **if** $t \mod D^* = 0$ **then**
11 **if** $M \geq 2F + 1$ **then**
12 The parameter server randomly chooses one representative from each of the M networks and queries their estimates;
13 **else**
14 For each $i \in \mathcal{C}$, randomly choose one agent in $\mathcal{V}_i$ as network representative of iteration t. Choose $(2F + 1 - |\mathcal{C}|)$ representatives from $\cup_{i \notin \mathcal{C}} \mathcal{V}_i$ uniformly at random as representatives, and queries their estimates;
15 The parameter server removes messages with the largest F and smallest F values;
16 $\widetilde{w}(t, \theta, \theta') \leftarrow \frac{1}{|\widetilde{\mathcal{R}}(t,\theta,\theta')|} \sum_{j \in \widetilde{\mathcal{R}}(t,\theta,\theta')} \widetilde{r}_t^j(\theta, \theta')$;
17 The parameter server multicasts $\widetilde{w}(t, \theta, \theta')$ to the network representatives not in $\mathcal{C}$.
18 **for** *each $i \notin \mathcal{C}$* **do**
19 the network representative updates its $r_t^j(\theta, \theta')$ to $\widetilde{w}(t, \theta, \theta')$.

VI. Convergence Results

A. Preliminaries.

Definition 1 (KL Divergence): Let P and Q be probability measures on a measurable space $\mathcal{X}$, and P is absolutely continuous with respect to Q, then the Kullback-Leibler (KL) divergence from Q to P is defined as

$$D_{\mathrm{KL}}(P \parallel Q) := \int_{\mathcal{X}} \log \left(\frac{P(dx)}{Q(dx)} \right) P(dx). \tag{2}$$

Contextualizing in our setting, $D\left(\ell_j(\cdot|\theta) \parallel \ell_j(\cdot|\theta')\right) = 0$ implies that at agent j, based on its locally collected signal $s_1^j, s_2^j, \cdots$, it cannot distinguish $\ell_j(\cdot|\theta')$ from $\ell_j(\cdot|\theta)$ no matter how many sample it collects.

Definition 2 (Reduced/information flow graph[24], [22]): Given a graph $G(\mathcal{V},\mathcal{E})$, a reduced/infomation flow graph is constructed as:

- remove all faulty nodes $\mathcal{A}$,
- remove all the links incident on the faulty nodes $\mathcal{A}$,
- for each non-faulty node, remove F additional incoming links. If there are less than F such links, remove all.

Both the malicious behaviors of the Byzantine agents and message trimming can alter the effective information flow in a network. A reduced/information flow graph as per Definition 2 captures how information flows implicitly in a network. Henceforth, for exposition clarity, we refer to such constructed graphs as information flow graphs. Notably, since there might be multiple choices of links in the third bullet of Definition 2, the information flow graph is not unique. Let $\mathcal{G}_{\text{info}}$ denote the collection of all the information flow graphs of a given graph $G(\mathcal{V},\mathcal{E})$, i.e., $\mathcal{G}_{\text{info}}(G(\mathcal{V}_i,\mathcal{E}_i))$ is the set of information flow graphs for the i-th sub-network. Let

$$\chi_i := |\mathcal{G}_{\text{info}}(G(\mathcal{V}_i,\mathcal{E}_i))|. \tag{3}$$

The information flow graphs as per Definition 2 is defined for scalar problems, i.e., the local estimate variable is a scalar [24], [22]. For multi-dimensional variables, the corresponding information flow graphs are obtained by removing dF links in the third bullet points of Definition 2 [23]. Intuitively, this is because that when the variables are multi-dimensional, to restrain the impacts of the malicious values, a normal agent trims away "extreme" messages aggressively, significantly limits the effective information flow.

Assumption 1: [24] For any given $G(\mathcal{V},\mathcal{E})$, each of its information flow graph contains only one source component.

For scalar inputs, Assumption 1 is shown to be both necessary and sufficient for Byzantine-resilient consensus to be achievable on the given network $G(\mathcal{V},\mathcal{E})$. Intuitively, under Assumption 1, agents in the source component can sufficiently fuse their local information and can propagate the fused values to each normal agent in the network.

Assumption 2: The agents in a source component can collectively distinguish hypotheses with infinitely many samples. Mathematically, for any information flow graph,

$$\sum_{j\in\mathcal{C}_s} D\left(\ell_j(\cdot|\theta)\parallel\ell_j(\cdot|\theta')\right) \neq 0, \quad \forall\,\theta\neq\theta', \tag{4}$$

where $\mathcal{C}_s$ denotes a source component.

In practice, the data sample is collected gradually rather than given all at once. Moreover, the malicious behaviors of the Byzantine agents can cause highly unstructured and intricate statistical dependency, leading to tremendous obstacles in both the algorithm design and analysis.

B. Convergence

Let $\mathcal{C}\subseteq\{1,2,\cdots,M\}$ be the set of sub-networks that satisfy Assumptions 1 and 2. We first show that each normal agent that is contained in a sub-network in $\mathcal{C}$ can identify θ^* asymptotically with probability 1. Towards this, define

$$D^*_{KL} := \min_{i\in\mathcal{C}}\min_{\theta\in\Theta\setminus\{\theta^*\}}\min_{\mathcal{C}_s\in\mathcal{G}^i_{\text{info}}}\sum_{j\in\mathcal{C}_s} D_{KL}\left(\ell_j(\cdot|\theta^*)\parallel\ell_j(\cdot|\theta)\right).$$

It is easy to see that as long as $\mathcal{C}\neq\emptyset$, $D^*_{KL}>0$.

Theorem 1: Fix $i\in\mathcal{C}$. For any non-Byzantine agent j, there exists $\widetilde{\theta}\in\Theta$ such that $\limsup_{t\to\infty} r^j_t(\widetilde{\theta},\theta)\xrightarrow{\text{a.s.}}+\infty$ and $\liminf_{t\to\infty} r^j_t(\theta,\widetilde{\theta})\xrightarrow{\text{a.s.}}-\infty$. Moreover, $\widetilde{\theta}=\theta^*$.

We prove Theorem 1 through a couple of lemmas.

Lemma 1: Fix any network i in $\mathcal{C}$. Let $\phi_i=|\mathcal{V}_i\setminus\mathcal{A}|$. Let $j\in\mathcal{V}_i\setminus\mathcal{A}$ be an arbitrary non-Byzantine agent. Then with probability 1, it holds that

$$\lim_{t\to\infty}\frac{1}{t^2}r^j_t(\theta^*,\theta)\geq\frac{1}{2}\beta^{\chi_i(n_i-\phi_i)}D^*_{KL},\ \text{and}$$
$$\lim_{t\to\infty}\frac{1}{t^2}r^j_t(\theta,\theta^*)\leq-\frac{1}{2}\beta^{\chi_i(n_i-\phi_i)}D^*_{KL},$$

where $\beta\triangleq\min_{i\in\mathcal{C}}\min_{j\in\mathcal{V}_i\setminus\mathcal{A}}\frac{1}{2(d^i_j-2F)+1}$.

Recall that d^i_j is the incoming degree of agent j that belongs to sub-network i. It is worth noting that Lemma 1 does not provide any characterization of $r^j_t(\widetilde{\theta},\theta)$ and $r^j_t(\theta,\widetilde{\theta})$ for $\widetilde{\theta}\neq\theta^*$. To enable agent j to identify θ^* by monitoring $r^j_t(\widetilde{\theta},\theta)$ and $r^j_t(\theta,\widetilde{\theta})$, we need to exclude the following possibility: there exists $\widetilde{\theta}\neq\theta^*$ such that

$$\lim_{t\to\infty} r^j_t(\widetilde{\theta},\theta)\xrightarrow{\text{a.s.}}+\infty,\ \text{and}\quad \lim_{t\to\infty} r^j_t(\theta,\widetilde{\theta})\xrightarrow{\text{a.s.}}-\infty.$$

Lemma 2: Fix a network in $i\in\mathcal{C}$. Suppose there exists $\widetilde{\theta}\in\Theta$ such that for any $\theta\neq\widetilde{\theta}$, it holds that $\lim_{t\to\infty} r^j_t(\widetilde{\theta},\theta)\xrightarrow{\text{a.s.}}+\infty$, and $\lim_{t\to\infty} r^j_t(\theta,\widetilde{\theta})\xrightarrow{\text{a.s.}}-\infty$. Then $\widetilde{\theta}=\theta^*$, where $\xrightarrow{\text{a.s.}}$ denotes "converge almost surely".

Proof. We prove this proposition by contradiction. Suppose there exists $\widetilde{\theta}\neq\theta^*\in\Theta$ such that for any $\theta\neq\widetilde{\theta}$, it holds that $\lim_{t\to\infty} r^j_t(\widetilde{\theta},\theta)\xrightarrow{\text{a.s.}}+\infty$, and $\lim_{t\to\infty} r^j_t(\theta,\widetilde{\theta})\xrightarrow{\text{a.s.}}-\infty$. Then we know that $\lim_{t\to\infty} r^j_t(\widetilde{\theta},\theta^*)\xrightarrow{\text{a.s.}}+\infty$ and $\lim_{t\to\infty} r^j_t(\theta^*,\widetilde{\theta})\xrightarrow{\text{a.s.}}-\infty$, contradicting Lemma 1. □

It remains to show the case when agent j does not belong to any network in $\mathcal{C}$.

Theorem 2: Suppose that $|\mathcal{C}|\geq F+1$. For any non-Byzantine agent j that does not belong to any of the networks in $\mathcal{C}$, there exists $\widetilde{\theta}$ such that $\forall\,\theta\neq\widetilde{\theta}$: $\limsup_{t\to\infty} r^j_t(\widetilde{\theta},\theta)\xrightarrow{\text{a.s.}}+\infty$, and $\liminf_{t\to\infty} r^j_t(\theta,\widetilde{\theta})\xrightarrow{\text{a.s.}}-\infty$. Moreover, $\widetilde{\theta}=\theta^*$.

Remark 1: Theorem 2 is non-trivial. By [20], $|\mathcal{C}|\geq F+1$ implies that at least $F+1$ networks can reach consensus individually despite different learning rates. However, since the Byzantine agents can lie arbitrarily and the local signals are non-IID and noisy, agents in $\mathcal{C}$ may not effectively propagate its local learning to agents in a different network. Particularly, in line 16, it is possible that the messages from the sample agents in $\mathcal{C}$ are all filtered out by the PS. Though we conjectured on the pairwise linear dynamics in [20], formal analysis was missing and the sketched proof does not go through. This is because the KL divergence term shows up only when one of the hypothesis involved is the truth θ^*.

Remark 2: The Byzantine agents $\mathcal{A}$ can be arbitrary subset of $\cup_{i=1}^M\mathcal{V}_i$ as long as $|\mathcal{A}|\leq F$. One interesting extreme

case is when all the Byzantine agents are located in the same sub-network. Assumption 1 implies that $F < \frac{1}{3}n_i$ for $i \in \mathcal{C}$.

It is worth noting that for a sub-network outside $\mathcal{C}$, even if the majority of the agents are Byzantine, our algorithm still enables the normal agents to learn θ^*.

Appendix I
Proof Sketch of Lemma 1.

Since the intersection of finitely many almost surely events is also almost surely and the problem is invariant to the permutation of hypothesis indices, it is enough to consider the convergence for the distinct hypotheses pair θ_1 and θ_2 and assume that $\theta^* \in \{\theta_1, \theta_2\}$.

By [22], we know that for each pair of hypotheses θ_1 and θ_2, there exists a row-stochastic matrix $\mathbf{M}^{1,2}[t] \in R^{(n_i-\phi_i)\times(n_i-\phi_i)}$ such that

$$r_t^j(\theta_1,\theta_2) = \sum_{j'=1}^{n_i-\phi_i} \mathbf{M}^{1,2}_{jj'}[t] r_{t-1}^j(\theta_1,\theta_2) + \log\frac{\ell_j(s^j_{1,t} \mid \theta_1)}{\ell_j(s^j_{1,t} \mid \theta_2)}. \quad (5)$$

Matrix $\mathbf{M}^{1,2}[t]$ depends on θ_1 and θ_2, and is time-varying. The reason of that $\mathbf{M}^{1,2}[t]$ is time-varying is two-fold: (1) The log likelihood ratio of the cumulative signals $\log\frac{\ell_j(s^j_{1,t}|\theta_1)}{\ell_j(s^j_{1,t}|\theta_2)}$ is changing over time due to the obtain of new signal and the randomness in the signal; and (2) the Byzantine agents can adaptively calibrate their malicious messages based on algorithm execution up to time t.

Let $\mathbf{r}_t(\theta_1,\theta_2) \in R^{n_i-\phi_i}$ be the vector that stacks $r_t^j(\theta_1,\theta_2)$. The evolution of $\mathbf{r}(\theta_1,\theta_2)$ can be written as

$$\begin{aligned}\mathbf{r}_t(\theta_1,\theta_2) &= \mathbf{M}^{1,2}[t]\mathbf{r}_{t-1}(\theta_1,\theta_2) + \sum_{r=1}^{t}\mathcal{L}_r(\theta_1,\theta_2)\\ &= \sum_{r=1}^{t}\mathbf{\Phi}^{1,2}(t,r+1)\sum_{k=1}^{r}\mathcal{L}_k(\theta_1,\theta_2), \quad (6)\end{aligned}$$

where $\mathbf{\Phi}^{1,2}(t,r+1) \triangleq \mathbf{M}^{1,2}[t]\mathbf{M}^{1,2}[t-1]\cdots\mathbf{M}^{1,2}[r+1]$ for $r \le t$, $\mathbf{\Phi}^{1,2}(t,t) \triangleq \mathbf{M}^{1,2}[t]$ and $\mathbf{\Phi}^{1,2}(t,t+1) \triangleq \mathbf{I}$.

Using coefficients of ergodicity [5], under Assumption 1, it has been shown [23] that

$$\lim_{t\ge r,\ t\to\infty}\mathbf{\Phi}^{1,2}(t,r) = \mathbf{1}\pi^{1,2}(r), \quad (7)$$

where $\pi(r)^{1,2} \in R^{(n_i-\phi_i)}$ is a row stochastic vector, and $\mathbf{1}$ is the column vector with each entry being 1. To prove $\lim_{t\to\infty}\frac{1}{t^2}r_t^j(\theta^*,\theta) \ge \frac{1}{2}\beta^{\chi_i(n_i-\phi_i)}D^*_{KL}$, without loss of generality, let $\theta_1 = \theta^*$. For each $j \in \mathcal{V}_i \setminus \mathcal{A}$, we have

$$\begin{aligned}r_t^j(\theta_1,\theta_2) = \sum_{r=1}^{t}\Bigg(&\underbrace{\Bigg(\sum_{j'=1}^{n_i-\phi_i}\mathbf{\Phi}^{1,2}_{jj'}(t,r+1)\sum_{k=1}^{r}\mathcal{L}_k^{j'}(\theta_1,\theta_2) - r\sum_{j'=1}^{n_i-\phi_i}\pi^{1,2}_{j'}(r+1)D^{j'}_{KL}(\theta_1,\theta_2)\Bigg)}_{(A)}\\ &+\underbrace{\sum_{r=1}^{t}r\sum_{j'=1}^{n_i-\phi_i}\pi^{1,2}_{j'}(r+1)D^{j'}_{KL}(\theta_1,\theta_2)}_{(B)}.\end{aligned}$$

By [23, Lemma 4], we know that: For any $r \ge 1$, there exists an information flow graph with source component $\mathcal{C}_s$ such that $\pi_i^{1,2}(r) \ge \beta^{\chi_i(n_i-\phi_i)}\ \forall\, j \in \mathcal{C}_s$. Thus,

$$\begin{aligned}(B) &\ge \sum_{r=1}^{t} r\beta^{\chi_i(n_i-\phi_i)}\sum_{j'\in\mathcal{C}_s}D^{j'}_{KL}(\theta_1,\theta_2)\\ &\ge \frac{t(t+1)}{2}\beta^{\chi_i(n_i-\phi_i)}D^*_{KL}.\end{aligned}$$

Thus, let $t\to\infty$, it holds that $(B)\to\infty$, and that

$$\lim_{t\to\infty}\frac{1}{t^2}(B) \ge \frac{1}{2}\beta^{\chi_i(n_i-\phi_i)}D^*_{KL}.$$

To bound (A), we first note that when $\theta_1 = \theta^*$,

$$\mathbb{E}_{s\sim\ell(\cdot|\theta^*)}\left[\mathcal{L}_k^{j'}(\theta^*,\theta_2)\right] = D^{j'}_{KL}(\theta^*,\theta_2).$$

Following [20, Lemma 3], we can show that

$$\frac{1}{t^2}\sum_{r=1}^{t}r\sum_{j'=1}^{n_i-\phi_i}\pi_{j'}(r+1)(\mathcal{L}_k^{j'}(\theta_1,\theta_2) - D^{j'}_{KL}(\theta_1,\theta_2)) \xrightarrow{a.s.} 0.$$

With Eq. (6), we conclude that

$$\lim_{t\to\infty}\frac{1}{t^2}r_t^j(\theta^*,\theta) \ \ge\ \frac{1}{2}\beta_m^{\chi_i(n_i-\phi_i)}D^*_{KL}, \quad \text{almost surely.}$$

It can be shown analogously that

$$\lim_{t\to\infty}\frac{1}{t^2}r_t^j(\theta,\theta^*) \le -\frac{1}{2}\beta_m^{\chi_i(n_i-\phi_i)}D^*_{KL}, \quad \text{almost surely.}$$

Appendix II
Proof Sketch of Theorem 2.

We focus on the scenario where $M \ge 2F+1$. The analysis can be easily adapted for the scenario where $M \le 2F$. Without loss of generality, let $j \in \mathcal{V}_1 \setminus \mathcal{A}$.

For each t such that $t \mod D^* = 0$, with probability $\frac{1}{n_1}$, agent j will be selected as a representative. Let $j_i \in \mathcal{V}_i\setminus\mathcal{A}$ be an arbitrary non-Byzantine agent for $i = 2,\cdots,M$. Let $j_1(k),\cdots,j_M(k)$ denote the representatives of networks $1,\cdots,M$ at $t = kD^*$. We define a sequence of events:

$$A_k := \{\omega : j_1(k) = j, \text{ and } j_i(k) = j_i\ \forall\ i \ne 1\}. \quad (8)$$

Let $p_k = \mathbb{P}\{A_k\}$ for $k = 1,2,\cdots$. It is easy to see that $p_k = \frac{1}{n_1}\prod_{i=2}^{M}\frac{1}{n_i} = \prod_{i=1}^{M}\frac{1}{n_i}$. Since $\sum_{k=1}^{\infty}p_k = \sum_{k=1}^{\infty}\prod_{i=1}^{M}\frac{1}{n_i} = \infty$, and $A_1, A_2, \cdots$ are mutually independent, by Borel-Cantelli lemma [4, Lemma 1.3], we know

$$\mathbb{P}\{A_k \text{ infinitely often}\} = 1, \quad (9)$$

where $A_k$infinitely often $= \cap_{n\ge1}(\cup_{k\ge n}A_k)$. That is, with probability 1 (almost surely), agent j will be selected infinitely many times. Let $\tau_1, \tau_2, \cdots$ be the time indices at which agent j is selected.

Let ω be a sample path in which each of the network in $\mathcal{C}$ learn θ^* independently, and that agent j is selected as the representative of network S_1 infinitely often. Let t^* be the time index such that for all $t \ge t^*$, $r_t^{j'} \ge \frac{1}{2}\beta^{\chi_i(n_i-\phi_i)}D^*_{KL}t^2$ for all $i \in \mathcal{C}$ and $j' \in \mathcal{V}_i\setminus\mathcal{A}$. By Theorem 1 and Eq.(9), we know that $\mathbb{P}\{\text{all such } \omega\} = 1$. Notably, t^* may change as the sample path ω changes. Henceforth, we fix one such sample path. Let $\theta_1 = \theta^*$ and let

$$\begin{aligned}c_{\min} &:= \min_{i\in\mathcal{C},j'\in\mathcal{V}_i\setminus\mathcal{A},\ t<t^*} r_t^{j'}(\theta_1,\theta_2),\\ c_{\max} &:= \min_{i\in\mathcal{C},j'\in\mathcal{V}_i\setminus\mathcal{A},\ t<t^*} r_t^{j'}(\theta_1,\theta_2).\end{aligned}$$

By definition, $\widetilde{w}(t) = \frac{1}{|\widetilde{\mathcal{R}}(t)|}\sum_{j\in\widetilde{\mathcal{R}}(t)} m_j(t)$. For any τ_r, none of the representatives are Byzantine; hence, $\widetilde{w}(\tau_r) = \frac{1}{|\widetilde{\mathcal{R}}(\tau_r)|}\sum_{j\in\widetilde{\mathcal{R}}(\tau_r)} r^j_{\tau_r}(\theta_1,\theta_2)$. Hence, we are able to rewrite (via two steps) $\widetilde{w}(\tau_r)$ in a form in which at least $M-F$ representatives have non-trivial influence on j. Let $k_1,\cdots,k_F$ be the indices of the bottom F values that are filtered out by the parameter server. Similarly, let $k'_1,\cdots,k'_F$ be the indices of the filtered top F values. For each $\ell=1,\cdots,F$, there exists $\alpha_\ell\in[0,1]$ such that

$$\widetilde{w}(\tau_r) = \alpha_\ell r^{k_\ell}_{\tau_r}(\theta_1,\theta_2) + (1-\alpha_\ell)\, r^{k'_\ell}_{\tau_r}(\theta_1,\theta_2).$$

Thus, $\widetilde{w}(\tau_r) = \frac{1}{F}\sum_{\ell=1}^{F}\alpha_\ell r^{k_\ell}_{\tau_r}(\theta_1,\theta_2) + (1-\alpha_\ell)\, r^{k'_\ell}_{\tau_r}(\theta_1,\theta_2)$. We further rewrite $\widetilde{w}(\tau_r)$ as

$$\begin{aligned}
r^{j_\ell(t)}_t(\theta_1,\theta_2) &= \widetilde{w}(\tau_r) = \frac{M-2F}{M}\widetilde{w}(\tau_r) + \left(1-\frac{M-2F}{M}\right)\widetilde{w}(\tau_r)\\
&= \frac{M-2F}{M}\frac{1}{|\widetilde{\mathcal{R}}(\tau_r)|}\sum_{j\in\widetilde{\mathcal{R}}(\tau_r)} r^j_{\tau_r}(\theta_1,\theta_2)\\
&+ \left(1-\frac{M-2F}{M}\right)\frac{1}{F}\sum_{\ell=1}^{F}\left(\alpha_\ell r^{k_\ell}_{\tau_r}(\theta_1,\theta_2) + (1-\alpha_\ell)\, r^{k'_\ell}_{\tau_r}(\theta_1,\theta_2)\right)\\
&= \frac{1}{M}\sum_{j\in\widetilde{\mathcal{R}}(\tau_r)} r^j_{\tau_r}(\theta_1,\theta_2)\\
&\quad + \frac{2}{M}\sum_{\ell=1}^{F}\left(\alpha_\ell r^{k_\ell}_{\tau_r}(\theta_1,\theta_2) + (1-\alpha_\ell)\, r^{k'_\ell}_{\tau_r}(\theta_1,\theta_2)\right).
\end{aligned}$$

Notably, either $\alpha_\ell \geq 1/2$ or $(1-\alpha_\ell)\geq 1/2$. Recall that $\left|\widetilde{\mathcal{R}}(\tau_r)\right| = M-2F$. Hence, we conclude that $\widetilde{w}(\tau_r)$ can be written as a convex combination of all the local estimates of the M representatives with at least $M-F$ representatives with weights at least $\frac{1}{M}$. Since $|\mathcal{C}|\geq F+1$, we now that at least one representative from a network in $\mathcal{C}$ will have corresponding coefficient $\geq \frac{1}{M}$. Thus, when $\theta_1=\theta^*$, by Theorem 1, we have

$$r^j_{k_\tau}(\theta_1,\theta_2) \geq \frac{1}{M}\beta^{\chi_i(n_i-\phi_i)} D^*_{KL}\,(k_\tau)^2 - \max\{|c_{\min}|,|c_{\max}|\}.$$

Let $\tau\to\infty$, we have $\lim_{\tau\to\infty} r^j_{k_\tau}(\theta_1,\theta_2) = +\infty$, i.e., $\limsup_{t\to\infty} r^j_{k_\tau}(\theta^*,\theta) = +\infty$. For $t\neq\tau_r$ for any r and $t\geq t^*$, via the same argument in [23], we are able to write $r^j_t(\theta^*,\theta) = \widetilde{w}(t)$ as a convex combination of the non-Byzantine representatives of iteration t. That is, there exists $\widetilde{\alpha}^1_t,\cdots,\widetilde{\alpha}^M_t$ such that

$$r^j_t(\theta^*,\theta) = \widetilde{w}(t) = \sum_{i=1}^{M}\widetilde{\alpha}^i_t r^{j_i(t)}_t(\theta^*,\theta).$$

Thus, we have $r^j_t(\theta^*,\theta) \geq -\max\{|c_{\min}|,|c_{\max}|\}$. Thus, $\liminf_{t\to\infty} r^j_t(\theta^*,\theta) \geq -\max\{|c_{\min}|,|c_{\max}|\}$. Since $\mathbb{P}\{\text{all such }\omega\}=1$, we conclude that with probability 1, for all $\theta\neq\theta^*$, it is true that $\limsup_{t\to\infty} r^j_t(\theta^*,\theta) = +\infty$ and $\liminf_{t\to\infty} r^j_t(\theta^*,\theta)\geq -\max\{|c_{\min}|,|c_{\max}|\}$.

Similarly, we are able to show that with probability 1, for all $\theta\neq\theta^*$, it holds that $\limsup_{t\to\infty} r^j_t(\theta,\theta^*)\leq\max\{|c_{\min}|,|c_{\max}|\}$, and $\liminf_{t\to\infty} r^j_t(\theta^*,\theta) = -\infty$.

It can be easily shown by contradiction (similar to the proof of Lemma 2) that if there exists $\widetilde{\theta}\in\Theta$ such that for any $\theta\neq\widetilde{\theta}$, it holds that $\limsup_{t\to\infty} r^j_t(\widetilde{\theta},\theta)=\infty$, $\liminf_{t\to\infty} r^j_t(\widetilde{\theta},\theta) > -\infty$, and $\limsup_{t\to\infty} r^j_t(\theta,\widetilde{\theta}) < \infty$, $\liminf_{t\to\infty} r^j_t(\widetilde{\theta},\widetilde{\theta}) = -\infty$, then $\widetilde{\theta}=\theta^*$, proving Theorem 2.

REFERENCES

[1] Q. Chen, W. Shi, D. Sui, and S. Leng. Distributed consensus algorithms in sensor networks with higher-order topology. *Entropy*, 25(8), 2023.

[2] M. Epstein, K. Lynch, K. H. Johansson, and R. M. Murray. Using hierarchical decomposition to speed up average consensus. *IFAC Proceedings Volumes*, 41(2), 2008. 17th IFAC World Congress.

[3] J. Feng, L. Liu, Q. Pei, and K. Li. Min-max cost optimization for efficient hierarchical federated learning in wireless edge networks. *IEEE Transactions on Parallel and Distributed Systems*, 2021.

[4] B. Hajek. *Random processes for engineers*. Cambridge university press, 2015.

[5] J. Hajnal and M. Bartlett. Weak ergodicity in non-homogeneous markov chains. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 54. Cambridge Univ Press, 1958.

[6] J. Hou and R. Zheng. Hierarchical consensus problem via group information exchange. *IEEE Transactions on Cybernetics*, 49(6), 2019.

[7] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi. Non-bayesian social learning. *Games and Economic Behavior*, 76(1), 2012.

[8] A. Jadbabaie, P. Molavi, and A. Tahbaz-Salehi. Information heterogeneity and the speed of learning in social networks. *Columbia Business School Research Paper*, 2013.

[9] L. Lamport, R. Shostak, and M. Pease. The byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3), July 1982.

[10] W. Y. B. Lim, J. S. Ng, Z. Xiong, J. Jin, Y. Zhang, D. Niyato, C. Leung, and C. Miao. Decentralized edge intelligence: A dynamic resource allocation framework for hierarchical federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(3), 2021.

[11] L. Liu, J. Zhang, S. Song, and K. B. Letaief. Client-edge-cloud hierarchical federated learning. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020.

[12] Q. Liu, A. Fang, L. Wang, and X. Wang. Social learning with time-varying weights. *Journal of Systems Science and Complexity*, 27, 2014.

[13] N. A. Lynch. *Distributed Algorithms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996.

[14] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, arXiv, 2017.

[15] H. Mendes, M. Herlihy, N. Vaidya, and V. K. Garg. Multidimensional agreement in byzantine systems. *Distributed Computing*, 28(6), 2015.

[16] A. Mitra, J. A. Richards, and S. Sundaram. A communication-efficient algorithm for exponentially fast non-bayesian learning in networks. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019.

[17] P. Molavi, A. Tahbaz-Salehi, and A. Jadbabaie. Foundations of non-bayesian social learning. *Columbia Business School Research Paper*, 2017.

[18] A. Nedić, A. Olshevsky, and C. A. Uribe. Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs. In *2015 American Control Conference (ACC)*. IEEE, 2015.

[19] H. Salami, B. Ying, and A. H. Sayed. Social learning over weakly connected graphs. *IEEE Transactions on Signal and Information Processing over Networks*, 3(2), 2017.

[20] L. Su and N. H. Vaidya. Defending non-bayesian learning against adversarial attacks. *Distributed Computing*, 32(4), 2019.

[21] L. Tong, Y. Li, and W. Gao. A hierarchical edge cloud architecture for mobile computing. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 2016.

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

[23] N. H. Vaidya. Iterative byzantine vector consensus in incomplete graphs. In *Distributed Computing and Networking*. Springer, 2014.

[24] N. H. Vaidya, L. Tseng, and G. Liang. Iterative approximate byzantine consensus in arbitrary directed graphs - part II: synchronous and asynchronous systems. *CoRR*, abs/1202.6094, 2012.