Byzantine Resilient Distributed Multi-Task Learning

Jiani Li, Waseem Abbas, and Xenofon Koutsoukos
Department of Electrical Engineering and Computer Science
Vanderbilt University, Nashville, TN, USA
{jiani.li, waseem.abbas, xenofon.koutsoukos}@vanderbilt.edu

Abstract

Distributed multi-task learning provides significant advantages in multi-agent networks with heterogeneous data sources where agents aim to learn distinct but correlated models simultaneously. However, distributed algorithms for learning relatedness among tasks are not resilient in the presence of Byzantine agents. In this paper, we present an approach for Byzantine resilient distributed multi-task learning. We propose an efficient online weight assignment rule by measuring the accumulated loss using an agent's data and its neighbors' models. A small accumulated loss indicates a large similarity between the two tasks. In order to ensure the Byzantine resilience of the aggregation at a normal agent, we introduce a step for filtering out larger losses. We analyze the approach for convex models and show that normal agents converge resiliently towards their true targets. Further, an agent's learning performance using the proposed weight assignment rule is guaranteed to be at least as good as in the non-cooperative case as measured by the expected regret. Finally, we demonstrate the approach using three case studies, including regression and classification problems, and show that our method exhibits good empirical performance for non-convex models, such as convolutional neural networks.

1 Introduction

Distributed machine learning models are gaining much attention recently as they improve the learning capabilities of agents distributed within a network with no central entity or server. In a distributed multi-agent system, agents interact with each other to improve their learning capabilities by leveraging the shared information via exchanging either data or models. In particular, agents that do not have enough data to build refined models or agents that have limited computational capabilities, benefit most from such cooperation. Distributed learning also addresses the single point of failure problem as well as scalability issues and is naturally suited to mobile phones, autonomous vehicles, drones, healthcare, smart cities, and many other applications [1, 2, 3, 4].

In networks with heterogeneous data sources, it is natural to consider the multi-task learning (MTL) framework, where agents aim to learn distinct but correlated models simultaneously [5]. Typically, prior knowledge of the relationships among models is assumed in MTL. The relationships among agents can be promoted via several methods, such as mean regularization, clustered regularization, low-rank and sparse structures regularization [6, 7, 8]. However, in real-world applications, such relationships are unknown beforehand and need to be estimated online from data. Learning similarities among tasks to promote effective cooperation is a primary consideration in MTL. There has been extensive work for learning the relationship matrix *centrally* by optimizing a global convex regularized function [9, 10, 11]. In contrast, this paper focuses on computationally efficient *distributed* learning of the relationship among agents that does not require optimizing a relationship matrix centrally [12, 13, 14, 15].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Although the distributed approach to learning and promoting similarities among neighbors from online data has many advantages, it is not resilient to Byzantine agents. Fault-tolerance for MTL is discussed in [5], focusing on dropped nodes that occasionally stop sending information to their neighbors. In [16], the relationship promoted by measuring the quadratic distance between two model parameters for distributed MTL is shown to be vulnerable to gradient-based attacks, and a Byzantine resilient distributed MTL algorithm is proposed for regression problems to cope with such attacks. The proposed algorithm relies on a user-defined parameter F to filter out information from F neighbors in the aggregation step and is resilient to F Byzantine neighbors, but requires exponential time with respect to the number of agents.

In this paper, we propose an *online weight adjustment rule* for MTL that is guaranteed to achieve resilient distributed MTL for every normal agent using the rule. Compared to [16], the proposed method is suited for both regression and classification problems, is resilient to an arbitrary number of Byzantine agents (without the need to select a pre-defined parameter F bounding the number of Byzantine agents), and has linear time complexity. To the best of our knowledge, this is the first solution that aims to address the Byzantine resilient cooperation in distributed MTL networks via a resilient similarity promoting method. We note that the proposed rule is not limited to the multi-task setting but can also be used for general distributed machine learning and federated learning systems to achieve resilient consensus. We list our contributions below.

- We propose an efficient Byzantine resilient online weight adjustment rule for distributed MTL. We measure similarities among agents based on the accumulated loss of an agent's data and the models of its neighbors. In each iteration, a normal agent computes the weights assigned to its neighbors in time that is linear in the size of its neighborhood and the dimension of the data.
- We show that using the proposed rule, normal agents with convex models converge resiliently towards their true target with an improved learning performance compared to the non-cooperative case as measured by the expected regret at convergence. Even when all the neighbors are Byzantine, a normal agent can still resiliently converge to its true target with the same expected regret as without any cooperation with other agents, achieving resilience to an arbitrary number of Byzantine agents.
- We conduct three experiments for both regression and classification problems and demonstrate that our approach yields good empirical performance for non-convex models, such as convolutional neural networks.

2 Related Work

Multi-Task Learning. MTL deals with the problem of learning multiple related tasks simultaneously to improve the generalization performance of the models learned by each task with the help of the other auxiliary tasks [17, 18]. The extensive literature in MTL can be broadly categorized into two categories based on how the data is collected. The *centralized* approach assumes the data is collected beforehand at a centralized entity. Many successful MTL applications with deep networks, such as in natural language processing and computer vision, fall into this category [19, 20, 21, 22]. This approach usually learns multiple objectives from a shared representation by sharing layers and splitting architecture in the deep networks. On the other hand, the *distributed* approach assumes data is collected separately by each task in a distributed manner. This approach is naturally suited to model distributed learning in multi-agent systems such as mobile phones, autonomous vehicles, and smart cities [2, 3, 4]. We focus on distributed MTL in this paper.

Relationship Learning in MTL. Although it is often assumed that a clustered, sparse, or low-rank structure among tasks is known *a priori* [6, 7, 8], such information may not be available in many real-world applications. Learning the relatedness among tasks online from data to promote effective cooperation is a principle approach in MTL when the relationships among tasks are not known *a priori*. There has been extensive work in online relationship learning that can be broadly categorized into centralized and distributed methods. The first group assumes that a centralized server collects the task models and utilizes a convex formulation of the regularized MTL optimization problem over the relationship matrix, which is learned by solving the convex optimization problem [9, 10, 11]. The second group relies on a distributed architecture in which agents learn relationships with their neighbors based on the similarities of their models and accordingly adjust weights assigned to neighbors [12, 13, 14, 15]. Typical similarity metrics, such as $\mathcal{H}$ divergence [23, 24, 25] and

Wasserstein distance [25, 26], can be used in MTL in the same way they are used in domain adaptation, transfer learning, and adversarial learning. However, such metrics are mainly designed for measuring the divergence in data distributions and are not suitable for online relationship learning due to efficiency and privacy concerns in data sharing.

Resilient Aggregation in Distributed ML. Inspired by the resilient consensus algorithms in multi-agent networks [27, 28], various resilient aggregation rules have been adapted in distributed ML, including the coordinate-wise trimmed mean [29], the coordinate-wise median [29, 30, 31], the geometric median [32, 33], and the Krum algorithm [34]. However, studies have shown that these rules are not resilient against certain attacks [35, 36, 37]. The centerpoint based aggregation rule [38] has been proposed recently that guarantees resilient distributed learning to Byzantine attacks. However, since each agent fits a distinct model in MTL, consensus-based resilient aggregation rules are not directly applicable to MTL.

3 Distributed Multi-Task Learning

Notation. In this paper, $|A|$ denotes the cardinality of a set A, $\|\cdot\|$ denotes the ℓ_2 norm, $\text{Tr}(\cdot)$ denotes the trace of a matrix, and $\mathbb{E}_\xi[\cdot]$ denotes the expected value of a random variable ξ. If the context is clear, $\mathbb{E}[\cdot]$ is used.

Background. Consider a network of m agents[1] modeled by an *undirected graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents agents and $\mathcal{E}$ represents interactions between agents. A bi-directional edge $(l, k) \in \mathcal{E}$ means that agents k and l can exchange information with each other. Since each agent also has its own information, we have $(k, k) \in \mathcal{E}, \forall k \in \mathcal{V}$. The *neighborhood* of k is the set $\mathcal{N}_k = \{l \in \mathcal{V} | (l, k) \in \mathcal{E}\}$. Each agent k has data $\left\{(x_k^i, y_k^i)\right\}$ sampled randomly from the distribution generated by the random variable ξ_k, where $x_k^i \in \mathbb{R}^{d_x}$, $y_k^i \in \mathbb{R}^{d_y}$. We consider a convex *prediction function* (model) $f_k(x_k^i) = \theta_k^\top x_k^i$, where $\theta_k \in \mathbb{R}^{d_x}$ is the model parameter. We use $\ell_k(\cdot)$ to denote a convex *loss function* associated with the prediction function for agent k. MTL is concerned with fitting separate models θ_k to the data for agent k via the *expected risk function* $r_k(\theta_k) = \mathbb{E}\left[\ell_k(\theta_k; \xi_k)\right]$. Since $r_k(\cdot)$ is also convex, we use θ_k^* to denote the global minimum (target) of $r_k(\theta_k)$. The model parameters θ_k can be optimized via the following problem:

$$\min_{\Theta} \left\{ \sum_{k=1}^{m} r_k(\theta_k) + \eta \mathcal{R}(\Theta, \Omega) \right\}, \tag{1}$$

where $\Theta = [\theta_1, \ldots, \theta_m] \in \mathbb{R}^{d \times m}$, $\mathcal{R}(\cdot)$ is a convex regularization function promoting the relationships among the agents, and $\Omega \in \mathbb{R}^{m \times m}$ models the relationships among the agents that can be assigned a priori or can be estimated from data. An example of the regularizer takes the form of $\mathcal{R}(\Theta, \Omega) = \lambda_1 \text{Tr}(\Theta\Omega\Theta^\top) + \lambda_2 \text{Tr}(\Theta\Theta^\top)$, where λ_1, λ_2 are non-negative parameters. In a *centralized* setting, where a centralized server optimizes the relationship matrix by collecting the models of agents, an optimal solution $\Omega = \frac{(\Theta^\top\Theta)^{\frac{1}{2}}}{\text{Tr}((\Theta^\top\Theta))^{\frac{1}{2}}}$ is proposed in [10] for learning the structure of clustered MTL using the above regularizer. In the *distributed* case, the task relationships Ω are not learned centrally and we can use the *adapt-then-combine (ATC) diffusion* algorithm [39] as a projection-based distributed solution of (1):

$$\hat{\theta}_{k,i} = \theta_{k,i-1} - \mu_k \nabla \ell_k(\theta_{k,i-1}; \xi_k^{i-1}), \qquad \text{(adaptation)} \tag{2}$$

$$\theta_{k,i} = \sum_{l \in \mathcal{N}_k} a_{lk} \hat{\theta}_{l,i}, \text{ subject to } \sum_{l \in \mathcal{N}_k} a_{lk} = 1, a_{lk} \geq 0, a_{lk} = 0 \text{ if } l \notin \mathcal{N}_k, \qquad \text{(combination)} \tag{3}$$

where μ_k is the step size, $\mathcal{N}_k$ is the neighborhood of agent k, a_{lk} denotes the weight[2] assigned by agent k to l, which should accurately reflect the similarity relationships among agents. $\nabla \ell_k(\theta_{k,i-1}; \xi_k^{i-1})$ is the gradient using the instantaneous realization ξ_k^{i-1} of the random variable ξ_k. At each iteration i, agent k minimizes the individual risk using stochastic gradient descent (SGD) given local data followed by a combination step that aggregates neighboring models according to the weights assigned to them. The weights $\{a_{lk}\}$ are free parameters selected by the designer and they serve the same

[1] Each agent is modeled as a separate task, thus, the terms *agent* and *task* are used interchangeably.

[2] a_{lk} can be time-dependent but when context allows, we write $a_{lk}(i)$ as a_{lk} for simplicity.

purpose as Ω in a centralized formulation. Thus, there is no need to design Ω in the case of distributed MTL that utilizes ATC diffusion algorithm for aggregation [40].

Online Weight Adjustment Rules. Without knowing the relationships *a priori*, one can assume the existence of similarities among agents and can learn these similarities online from the data. The approach is based on the distance between the model parameters of agents, where small distance indicates a large similarity [12, 13, 41, 42]. A common approach to learning similarities between two agents online is given by

$$a_{lk}(i) = \frac{\|\tilde{\theta}_k^* - \hat{\theta}_{l,i}\|^{-2}}{\sum_{p\in\mathcal{N}_k} \|\tilde{\theta}_k^* - \hat{\theta}_{p,i}\|^{-2}}, \tag{4}$$

where $\tilde{\theta}_k^*$ is an approximation of θ_k^*. Since θ_k^* is unknown, one can only estimate θ_k^* using current knowledge. Examples include using the current model $\tilde{\theta}_k^* = \theta_{k,i-1}$, and one-step ahead approximation $\tilde{\theta}_k^* = \hat{\theta}_{k,i} + \mu_k \nabla \ell_k(\hat{\theta}_{k,i}; \xi_k^{i-1})$. Although the ℓ_2 norm is widely used, this formulation of weights can be generalized to ℓ_p norm as well.

4 Problem Formulation

Byzantine agents can send different information to different neighbors usually with a malicious goal of disrupting the network's convergence by increasing the expected risk.We assume a synchronous network in which Byzantine agents send information to their neighbors in each iteration. It has been shown in [16] that normal agents assigning weights according to (4) are vulnerable to Byzantine agents. This result can be stated in the context of this paper as follows:

Lemma 1. [3] *If a normal agent k adapts weights according to* (4)*, then a single Byzantine agent can lead k away from θ_k^* if $\tilde{\theta}_k^* \neq \theta_k^*$.*

To address the vulnerabilities of the online weight adjustment rules derived from (4), this paper aims to design an efficient resilient online weight assignment rule in the presence of Byzantine agents for MTL. Let the *expected regret* $R_k(i)$ be the value of the expected difference between the risk of $\theta_{k,i}$ and the optimal decision θ_k^*, i.e., $R_k(i) = \mathbb{E}[r_k(\theta_{k,i}) - r_k(\theta_k^*)]$. As a baseline, we consider the case when every normal agent runs the SGD algorithm without cooperation, i.e., $\theta_{k,i}^{(\text{ncop})} = \hat{\theta}_{k,i}^{(\text{ncop})}$, followed by (2). We also consider the cooperative case when $\theta_{k,i}^{(\text{coop})} = \sum_{l\in\mathcal{N}_k} a_{lk}\hat{\theta}_{l,i}^{(\text{coop})}$ as indicated in (3), followed by (2). The respective expected regrets for the two methods are given as

$$R_k^{(\text{ncop})}(i) = \mathbb{E}[r_k(\theta_{k,i}^{(\text{ncop})}) - r_k(\theta_k^*)] \text{ and } R_k^{(\text{coop})}(i) = \mathbb{E}[r_k(\theta_{k,i}^{(\text{coop})}) - r_k(\theta_k^*)].$$

The weight assignment must satisfy the following three conditions:

Resilient Convergence. It must be guaranteed that using the computed weights $A_k = [a_{1k}, \ldots, a_{mk}] \in \mathbb{R}^{1\times m}$, every normal agent k resiliently converges to the true target θ_k^*, i.e.,

$$\lim_{i\to\infty} \theta_{k,i}^{(\text{coop})} = \theta_k^*, \forall k \in \mathcal{N}^+, \tag{5}$$

where $\mathcal{N}^+$ denotes the set of normal agents in the network.

Improved Expected Regret w.r.t. Non-Cooperation. Cooperation among agents is meaningful only when it results in improving the learning performance. Hence, it is important to guarantee that using the computed weights A_k, a normal agent k obtains an improved expected regret as compared to using the SGD algorithm without cooperation, even in the presence of Byzantine agents, i.e.,

$$\lim_{i\to\infty} \sup R_k^{(\text{coop})}(i) \leq \lim_{i\to\infty} \sup R_k^{(\text{ncop})}(i), \forall k \in \mathcal{N}^+. \tag{6}$$

Computational Efficiency. At each iteration, a normal agent k needs to compute the weights A_k in time that is linear in the size of the neighborhood of k and the dimension of the data, i.e., in $O(|\mathcal{N}_k|(d_x + d_y))$ time.

[3] All proofs are given in Appendix A; appendices can be found in the supplementary material.

5 Loss-based Online Weight Adjustment

Weight Optimization. We follow a typical approach of learning the optimal weight adjustment rule [12, 13, 41, 42] in which the goal is to minimize the quadratic distance between the aggregated model $\theta_{k,i}$ and the true model θ_k^* over the weights, i.e., $\min_{A_k} \|\theta_{k,i}^{(\text{coop})} - \theta_k^*\|^2$. Using (3), we get an equivalent problem:

$$\min_{A_k} \left\| \sum_{l \in \mathcal{N}_k} a_{lk} \hat{\theta}_{l,i}^{(\text{coop})} - \theta_k^* \right\|^2, \text{ subject to } \sum_{l \in \mathcal{N}_k} a_{lk} = 1, a_{lk} \geq 0, a_{lk} = 0 \text{ if } l \notin \mathcal{N}_k,$$

where $\left\| \sum_{l \in \mathcal{N}_k} a_{lk} \hat{\theta}_{l,i}^{(\text{coop})} - \theta_k^* \right\|^2 = \sum_{l \in \mathcal{N}_k} \sum_{p \in \mathcal{N}_k} a_{lk} a_{pk} (\hat{\theta}_{l,i}^{(\text{coop})} - \theta_k^*)^\top (\hat{\theta}_{p,i}^{(\text{coop})} - \theta_k^*)$. As in a typical approximation approach, we consider

$$\left\| \sum_{l \in \mathcal{N}_k} a_{lk} \hat{\theta}_{l,i}^{(\text{coop})} - \theta_k^* \right\|^2 \approx \sum_{l \in \mathcal{N}_k} a_{lk}^2 \left\| \hat{\theta}_{l,i}^{(\text{coop})} - \theta_k^* \right\|^2. \tag{7}$$

The weight assignment rule (4) is an optimal solution of (7) using the approximation of θ_k^*, which as we show above, can be easily attacked. To avoid the use of the distance between model parameters as a similarity measure, we introduce a *resilient* counterpart, which is the *accumulated loss* (or risk). If we assume risk functions r_k to be m-strongly convex[4], then it holds that

$$r_k(\hat{\theta}_{l,i}^{(\text{coop})}) - r_k(\theta_k^*) \geq \langle \nabla r_k(\theta_k^*), y - x \rangle + \frac{m}{2} \|\hat{\theta}_{l,i}^{(\text{coop})} - \theta_k^*\|^2,$$

where $r_k(\hat{\theta}_{l,i}^{(\text{coop})}) = \mathbb{E}\left[\ell_k(\hat{\theta}_{l,i}^{(\text{coop})}; \xi_k)\right]$. Since $\nabla r_k(\theta_k^*) = 0$, we obtain

$$\|\hat{\theta}_{l,i}^{(\text{coop})} - \theta_k^*\|^2 \leq \frac{2}{m} \left(r_k(\hat{\theta}_{l,i}^{(\text{coop})}) - r_k(\theta_k^*) \right). \tag{8}$$

Instead of directly minimizing the right side of (7), we consider minimizing its upper bound given in (8). Later in Section 6, we show that this alternate approach facilitates the resilient distributed MTL, which cannot be achieved by minimizing the distance between models directly. Hence, by combining (7) and (8), we consider the following minimization problem:

$$\min_{A_k} \sum_{l \in \mathcal{N}_k} a_{lk}^2 \left(r_k(\hat{\theta}_{l,i}^{(\text{coop})}) - r_k(\theta_k^*) \right) \text{ subject to } \sum_{l \in \mathcal{N}_k} a_{lk} = 1, a_{lk} \geq 0, a_{lk} = 0 \text{ if } l \notin \mathcal{N}_k.$$

This optimization problem indicates that if a neighbor l's model has a small regret on agent k's data distribution, then it should be assigned a large weight. Since θ_k^* is unknown, one can use $r_k(\theta_{k,i}^{(\text{coop})})$ to approximate $r_k(\theta_k^*)$. Alternatively, since $r_k(\theta_k^*)$ is small compared to $r_k(\theta_{l,i}^{(\text{coop})})$, we could simply assume $r_k(\theta_k^*) = 0$ and consider the following minimization problem:

$$\min_{A_k} \sum_{l \in \mathcal{N}_k} a_{lk}^2 r_k(\hat{\theta}_{l,i}^{(\text{coop})}) \text{ subject to } \sum_{l \in \mathcal{N}_k} a_{lk} = 1, a_{lk} \geq 0, a_{lk} = 0 \text{ if } l \notin \mathcal{N}_k. \tag{9}$$

Using the Lagrangian relaxation,[5] we obtain the optimal solution of (9) as

$$a_{lk}(i) = \frac{r_k(\hat{\theta}_{l,i}^{(\text{coop})})^{-1}}{\sum_{p \in \mathcal{N}_k} r_k(\hat{\theta}_{p,i}^{(\text{coop})})^{-1}}. \tag{10}$$

We can approximate $r_k(\hat{\theta}_{l,i}^{(\text{coop})})$ using the exponential moving average $\varphi_{lk}^i = (1 - \nu_k)\varphi_{lk}^{i-1} + \nu_k \ell_k(\hat{\theta}_{l,i}^{(\text{coop})}; \xi_k)$, where ν_k is the forgetting factor. Given $\mathbb{E}[\varphi_{lk}^i] = (1 - \nu_k)\mathbb{E}[\varphi_{lk}^{i-1}] + \nu_k \mathbb{E}[\ell_k(\hat{\theta}_{l,i}^{(\text{coop})}; \xi_k)]$, we obtain $\lim_{i \to \infty} \mathbb{E}[\varphi_{lk}^i] = \lim_{i \to \infty} \mathbb{E}[\ell_k(\hat{\theta}_{l,i}^{(\text{coop})}; \xi_k)] = \lim_{i \to \infty} r_k(\hat{\theta}_{l,i}^{(\text{coop})})$, which means φ_{lk}^i converges (in expectation) to $\lim_{i \to \infty} r_k(\hat{\theta}_{l,i}^{(\text{coop})})$. Hence, we can use φ_{lk}^i to approximate $r_k(\hat{\theta}_{l,i}^{(\text{coop})})$. Note that in addition to the smoothing methods, one can use the average batch loss

[4] Details of the assumptions are given in Appendix A.1.

[5] Detailed solution is given in Appendix A.3.

to approximate $r_k(\hat{\theta}_{l,i}^{(\text{coop})})$ when using the (mini-) batch gradient descent in the place of SGD for adaptation.

Filtering for Resilience. Let $\mathcal{N}_k^+$ denote the set of k's normal neighbors with $|\mathcal{N}_k^+| \geq 1$. We assume there are q Byzantine neighbors in the set $\mathcal{B} = \mathcal{N}_k \backslash \mathcal{N}_k^+$. In the following, we examine the resilience of the cooperation using (10) in the presence of Byzantine agents.

Lemma 2. *Using* (10)*, the expected regret satisfies*

$$\mathbb{E}\left[r_k(\theta_{k,i}^{(\text{coop})}) - r_k^*\right] \leq \frac{1}{|\mathcal{N}_k|} \sum_{l \in \mathcal{N}_k} \mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right].$$

Since l can be a Byzantine agent, it is possible that $\mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right]$ is a large value. Consequently, we cannot compute a useful upper bound on the value of $\mathbb{E}\left[r_k(\theta_{k,i}^{(\text{coop})}) - r_k^*\right]$ and cannot provide further convergence guarantees. To facilitate the resilient cooperation, we consider a modification of (10) as follows.

$$a_{lk}(i) = \begin{cases} \frac{r_k(\hat{\theta}_{l,i}^{(\text{coop})})^{-1}}{\sum_{p \in \mathcal{N}_k^{\leq}} r_k(\hat{\theta}_{p,i}^{(\text{coop})})^{-1}}, & \text{if } r_k(\hat{\theta}_{l,i}^{(\text{coop})}) \leq r_k(\hat{\theta}_{k,i}^{(\text{coop})}), \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where $\mathcal{N}_k^{\leq}$ denotes the set of neighbors with $r_k(\hat{\theta}_{l,i}^{(\text{coop})}) \leq r_k(\hat{\theta}_{k,i}^{(\text{coop})})$. This implies that the cooperation filters out the information coming from the neighbors incurring a larger risk and cooperate only with the remaining neighbors. In the next section, we show how this modification guarantees the resilient convergence of MTL with an improved learning performance as measured by the expected regret w.r.t. the non-cooperative case.

Computational Efficiency. It takes $\mathcal{O}(d_x)$ time to compute the predicted value $f_l(x_k^i)$ using the model $\hat{\theta}_{l,i}^{(\text{coop})}$ and data x_k^i. Similarly, it takes $\mathcal{O}(d_y)$ time to compute the loss from $f_l(x_k^i)$ and y_k^i. Hence, it takes $\mathcal{O}(d_x + d_y)$ time to compute $\ell_k(\hat{\theta}_{l,i}^{(\text{coop})}; \xi_k^i)$. Using the exponential moving average method for approximating $r_k(\hat{\theta}_{l,i}^{(\text{coop})})$, at each iteration, the total time for computing $A_k(i)$ is $\mathcal{O}(|\mathcal{N}_k|(d_x + d_y))$.

6 Byzantine Resilient Convergence Analysis

Assumptions. We make the following assumptions to facilitate the analysis.

- Normal agents share the same stepsize $\mu_k = \mu, \forall k \in \mathcal{N}^+$.
- For every normal agent k, the loss function $\ell_k(\cdot)$ and the risk function $r_k(\cdot)$ are m*-strongly convex* and have L*-Lipschitz continuous gradient*.[6]
- There are n clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ in the network (n is unrevealed to each agent). All agents in the same cluster $\mathcal{C}_j$ have the same target θ_j^*, prediction function $f_j(\cdot)$, and loss function $\ell_j(\cdot)$. In other words, if $k, l \in \mathcal{C}_j$, then $\theta_k^* = \theta_l^* = \theta_j^*$, $f_k(\cdot) = f_l(\cdot) = f_j(\cdot)$, and $\ell_k(\cdot) = \ell_l(\cdot) = \ell_j(\cdot)$.
- For $\{k, l\} \in \mathcal{C}_j$, the stochastic gradients $\nabla \ell_j(\theta; \xi_k)$ and $\nabla \ell_j(\theta; \xi_l)$ are unbiased estimates of $\nabla r_k(\theta)$ (and $\nabla r_l(\theta)$), i.e., $\mathbb{E}[\nabla \ell_j(\theta; \xi_k)] = \nabla r_k(\theta)$, and $\mathbb{E}[\nabla \ell_j(\theta; \xi_l)] = \nabla r_k(\theta)$. We also restrict the variance of $\nabla \ell_j(\theta; \xi_k)$ and $\nabla \ell_j(\theta; \xi_l)$ to be $\text{Var}[\nabla \ell_j(\theta; \xi_k)] = \mathbb{E}[\|\nabla \ell_j(\theta; \xi_k)\|^2] - \|\mathbb{E}[\nabla \ell_j(\theta; \xi_k)]\|^2 \leq \sigma_k^2$, and $\text{Var}[\nabla \ell_j(\theta; \xi_l)] = \mathbb{E}[\|\nabla \ell_j(\theta; \xi_l)\|^2] - \|\mathbb{E}[\nabla \ell_j(\theta; \xi_l)]\|^2 \leq \sigma_l^2$.

Given these assumptions, we provide the following results for the non-cooperative SGD and the cooperative SGD using rule (11).

Lemma 3. *A normal agent k which runs the SGD algorithm without cooperation converges for fixed stepsize $\mu \in (0, \frac{2}{L}]$, and the expected regret at the convergence point satisfies*

$$\lim_{i \to \infty} \sup R_k^{(\text{ncop})}(i) = \lim_{i \to \infty} \sup \mathbb{E}\left[r_k\left(\theta_{k,i}^{(\text{ncop})}\right) - r_k^*\right] = \frac{\mu L \sigma_k^2}{2m(2 - \mu L)}.$$

[6]Details of the assumptions about the loss functions are given in Appendix A.1.

Theorem 1. *A normal agent k which runs the cooperative SGD algorithm using the loss-based weights* (11) *converges in the presence of arbitrary number of Byzantine neighbors, for fixed stepsize $\mu \in (0, \frac{2}{L}]$, and the expected regret at the convergence point satisfies*

$$\lim_{i\to\infty} \sup R_k^{(\text{coop})}(i) = \lim_{i\to\infty} \sup \mathbb{E}\left[r_k\left(\theta_{k,i}^{(\text{coop})}\right) - r_k^*\right] = \frac{\mu L}{2m(2-\mu L)} \frac{1}{|\mathcal{N}_k^{\leq}|} \sum_{l\in\mathcal{N}_k^{\leq}} \sigma_l^2.$$

Furthermore, in the presence of arbitrary number of Byzantine neighbors, we have

$$\lim_{i\to\infty} \sup R_k^{(\text{coop})}(i) \leq \lim_{i\to\infty} \sup R_k^{(\text{ncop})}(i).$$

Theorem 1 indicates that the cooperative case using weights in (11) is always at least as good as the non-cooperative case, as measured by the expected regret at convergence, which satisfies the conditions in (5) and (6). Note that even when all the neighbors of a normal agent are Byzantine, one can still guarantee that the agent's learning performance as a result of cooperation with neighbors using (11) will be same as the non-cooperative case.

Discussion. We assume convex models to carry out the analysis, which is typical in the literature. However, the intuition behind the approach is — *to measure the relatedness of a neighbor to itself, a normal agent evaluates the loss of the neighbor using the neighbor's model parameters and its own data, and cuts down the cooperation if this loss is larger than the agent's own loss* — and the same idea should also apply to non-convex models. In the next section, we also test our results on non-convex models, such as CNN, which generates experimental results similar to those produced by convex models.

7 Evaluation

In this section, we evaluate the resilience of the proposed online weight adjustment rule (11) with the smoothing method discussed in Section 5, and compare it with the non-cooperative case, the average weights ($a_{lk} = \frac{1}{|\mathcal{N}_k|}$), and the quadratic distance-based weights (4) (with $\tilde{\theta}_k^* = \theta_{k,i-1}$ and use the same smoothing method $\phi_{lk}^i = (1-\nu_k)\phi_{lk}^{i-1} + \nu_k\|\tilde{\theta}_k^* - \hat{\theta}_{l,i}\|^2$ in the place of $\|\tilde{\theta}_k^* - \hat{\theta}_{l,i}\|^2$, with the same forgetting factor ν_k used for (11)). We use three distributed MTL case studies, including the regression and classification problems, with and without the presence of Byzantine agents. Although the convergence analysis in Section 6 is based on convex models and SGD, we show empirically that the weight assignment rule (11) performs well for non-convex models, such as convolutional neural networks and mini-batch gradient descent. Our code is available at `https://github.com/JianiLi/resilientDistributedMTL`.

7.1 Datasets and Simulation Setups

- **Target Localization**: Target localization is a widely-studied linear regression problem [43]. The task is to estimate the location of the target by minimizing the squared error loss of noisy streaming sensor data. We consider a network of 100 agents with four targets as shown in Figure 1a. Agents in the same color share the same target, however, they do not know this group information beforehand.
- **Human Activity Recognition**[7]: Mobile phone sensor data (accelerometer and gyroscope) is collected from 30 individuals performing one of six activities: {walking, walking-upstairs, walking-downstairs, sitting, standing, lying-down}. The goal is to predict the activities performed using 561-length feature vectors for each instance generated by the processed sensor signals [2]. We model each individual as a separate task and use a complete graph to model the network topology. We use linear model as the prediction function with cross-entropy-loss.
- **Digit Classification**: We consider a network of ten agents performing digit classification. Five of the ten agents have access to the MNIST dataset[8] [44] (group 1) and the other five have access to the synthetic dataset[9] (group 2) that is composed by generated images of digits embedded on

[7] `https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones`

[8] `http://yann.lecun.com/exdb/mnist`

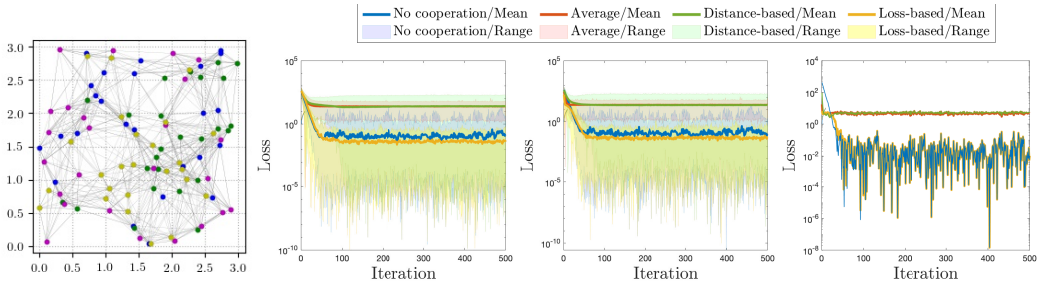
[9] `https://www.kaggle.com/prasunroy/synthetic-digits`

random backgrounds [45]. All the images are preprocessed to be 28×28 grayscale images. We model each agent as a separate task and use a complete graph to model the network topology. An agent does not know which of its neighbors are performing the same task as the agent itself. We use a CNN model of the same architecture for each agent and cross-entropy-loss.

7.2 Results[10]

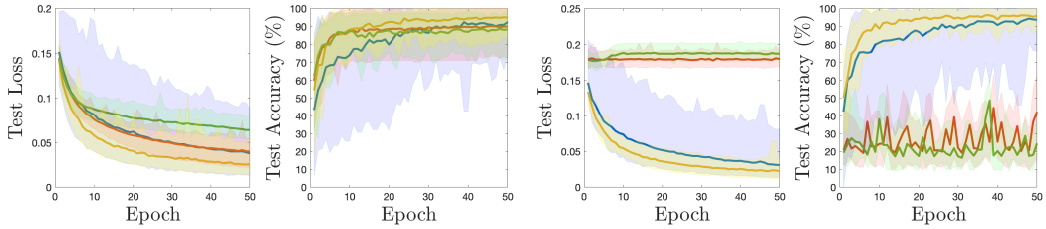
We plot the mean and range of the average loss of every normal agent using streaming data for the target localization problem in Figure 1b–d. Similarly, we plot the mean and range of the average testing loss and classification accuracy of every normal agent for human action recognition in Figure 2, and for digit classification in Figure 3 (for group 1) and Figure 4 (for group 2). At each iteration, Byzantine agents send random values (for each dimension) from the interval $[15, 16]$ for target localization, and $[0, 0.1]$ for the other two case studies.

In all of the examples, we find that the loss-based weight assignment rule (11) outperforms all the other rules and the non-cooperative case, with respect to the mean and range of the average loss and accuracy with and without the presence of Byzantine agents. Hence, our simulations validate the results indicated by (6) and imply that the loss-based weights (11) have accurately learned the relationship among agents. Moreover, normal agents having a large regret in their estimation benefit from cooperating with other agents having a small regret. We also consider the extreme case in which there is only one normal agent in the network, and all the other agents are Byzantine. In such a case, the loss-based weight assignment rule (11) has the same performance as the non-cooperative case, thus, showing that it is resilient to an arbitrary number of Byzantine agents.



(a) Network topology (b) No attack (c) 20 Byzantine agents (d) 99 Byzantine agents

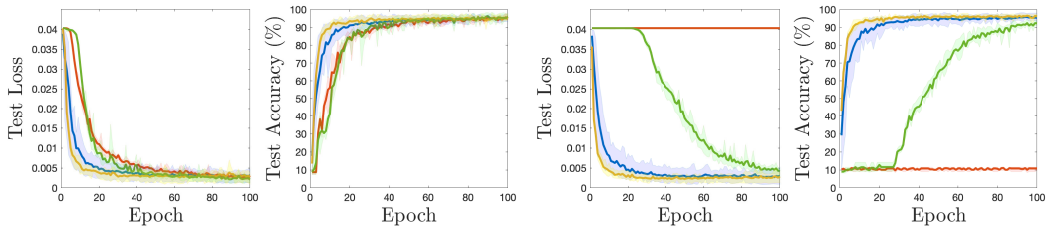
Figure 1: Target Localization: network topology and loss of streaming data for normal agents.



(a) No attack (b) 10 Byzantine agents

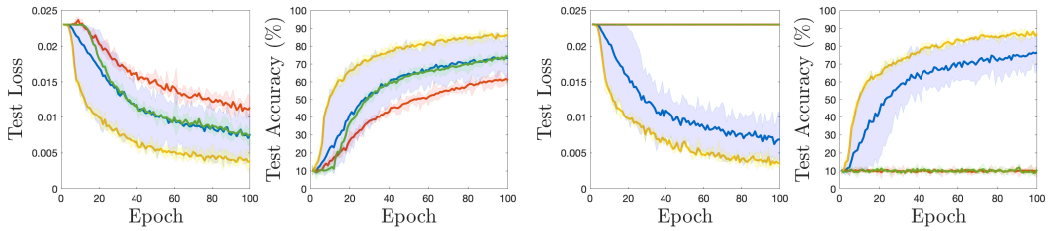
Figure 2: Human Action Recognition: average testing loss and accuracy for normal agents.

[10] Simulation details and supplementary results are given in Appendix B.



(a) No attack (b) 2 Byzantine agents

Figure 3: Digit Classification: average testing loss and accuracy for normal agents in group 1.



(a) No attack (b) 2 Byzantine agents

Figure 4: Digit Classification: average testing loss and accuracy for normal agents in group 2.

8 Conclusion

In this paper, we propose an efficient online weight adjustment rule for learning the similarities among agents in distributed multi-task networks with an arbitrary number of Byzantine agents. We show that a widely used approach of measuring the similarities based on the distance between two agents' model parameters is vulnerable to Byzantine attacks. To cope with such vulnerabilities, we propose to measure similarities based on the (accumulated) loss using an agent's data and its neighbors' models. A small loss indicates a large similarity between the agents. To eliminate the influence of Byzantine agents, a normal agent filters out the information from neighbors whose losses are larger than the agent's own loss. With filtering, the loss-based weight adjustment rule makes each normal agent converge resiliently towards its target parameter with an improved expected regret than the non-cooperative case. The experiment results validate the effectiveness of our approach.

Broader Impact

The problem of Byzantine resilient aggregation of distributed machine learning models has been actively studied in recent years; however, the issue of Byzantine resilient distributed learning in multi-task networks has received much less attention. It is a general intuition that MTL is robust and resilient to cyber-attacks since it can identify attackers by measuring similarities between neighbors. In this paper, we have shown that some commonly used similarity measures are not resilient against certain attacks. With an increase in data heterogeneity, we hope this work could highlight the security and privacy concerns in designing distributed MTL frameworks.

Acknowledgments and Disclosure of Funding

This work is supported in part by the NSA Lablet (H98230-18-D-0010). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSA.

References

[1] Jakub Konecný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *CoRR*, abs/1610.05492, 2016.

[2] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *21st European Symposium on Artificial Neural Networks, ESANN 2013, Bruges, Belgium, April 24-26, 2013*, 2013.

[3] Kai Yang, Tao Jiang, Yuanming Shi, and Zhi Ding. Federated learning via over-the-air computation. *CoRR*, abs/1812.11750, 2018.

[4] Yiqiang Chen, Jindong Wang, Chaohui Yu, Wen Gao, and Xin Qin. Fedhealth: A federated transfer learning framework for wearable healthcare. *CoRR*, abs/1907.09173, 2019.

[5] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S. Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems, Long Beach, CA, USA*, pages 4424–4434, 2017.

[6] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA*, pages 109–117, 2004.

[7] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Clustered multi-task learning via alternating structure optimization. In *Advances in Neural Information Processing Systems, Granada, Spain*, pages 702–710, 2011.

[8] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA*, pages 42–50, 2011.

[9] Laurent Jacob, Francis R. Bach, and Jean-Philippe Vert. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada*, pages 745–752, 2008.

[10] Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA*, pages 733–442, 2010.

[11] Avishek Saha, Piyush Rai, Hal Daumé III, and Suresh Venkatasubramanian. Online learning of multiple tasks and their relationships. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, USA*, pages 643–651, 2011.

[12] Xiaochuan Zhao and Ali H. Sayed. Clustering via diffusion adaptation over networks. In *3rd International Workshop on Cognitive Information Processing*, pages 1–6, May 2012.

[13] Jie Chen, Cédric Richard, and Ali H. Sayed. Diffusion LMS over multitask networks. *IEEE Transactions on Signal Processing*, 63(11):2733–2748, June 2015.

[14] Keerthiram Murugesan, Hanxiao Liu, Jaime G. Carbonell, and Yiming Yang. Adaptive smoothed online multi-task learning. In *Advances in Neural Information Processing Systems, Barcelona, Spain*, pages 4296–4304, 2016.

[15] Keerthiram Murugesan and Jaime G. Carbonell. Active learning from peers. In *Advances in Neural Information Processing Systems, Long Beach, CA, USA*, pages 7008–7017, 2017.

[16] Jiani Li, Waseem Abbas, and Xenofon Koutsoukos. Resilient distributed diffusion in networks with adversaries. *IEEE Transactions on Signal and Information Processing over Networks*, 6:1–17, 2020.

[17] Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, 1997.

[18] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098, 2017.

[19] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Philip S. Yu. Learning multiple tasks with multilinear relationship networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems, Long Beach, CA, USA*, pages 1594–1603, 2017.

[20] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA*, pages 3994–4003, 2016.

[21] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. A joint many-task model: Growing a neural network for multiple NLP tasks. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark*, pages 1923–1933, 2017.

[22] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 7482–7491. IEEE Computer Society, 2018.

[23] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, 2016.

[24] Nikola Konstantinov and Christoph Lampert. Robust learning from untrusted sources. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, USA*, pages 3488–3498, 2019.

[25] Changjian Shui, Mahdieh Abbasi, Louis-Émile Robitaille, Boyu Wang, and Christian Gagné. A principled approach for learning task similarity in multitask learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China*, pages 3446–3452, 2019.

[26] Yitong Li, Michael Murias, Geraldine Dawson, and David E. Carlson. Extracting relationships by multi-domain matching. In *Advances in Neural Information Processing Systems, Montréal, Canada*, pages 6799–6810, 2018.

[27] Reza Olfati-Saber, J. Alexander Fax, and Richard M. Murray. Consensus and cooperation in networked multi-agent systems. *Proc. IEEE*, 95(1):215–233, 2007.

[28] Heath J LeBlanc, Haotian Zhang, Xenofon Koutsoukos, and Shreyas Sundaram. Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications*, 31(4):766–781, 2013.

[29] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm, Sweden*, pages 5636–5645, 2018.

[30] Xiangyi Chen, Tiancong Chen, Haoran Sun, Zhiwei Steven Wu, and Mingyi Hong. Distributed training with heterogeneous data: Bridging median- and mean-based algorithms. *CoRR*, abs/1906.01736, 2019.

[31] Haibo Yang, Xin Zhang, Minghong Fang, and Jia Liu. Byzantine-resilient stochastic gradient descent for distributed learning: A Lipschitz-inspired coordinate-wise median approach. *CoRR*, abs/1909.04532, 2019.

[32] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(2):44:1–44:25, December 2017.

[33] Venkata Krishna Pillutla, Sham M. Kakade, and Zaïd Harchaoui. Robust aggregation for federated learning. *CoRR*, abs/1912.13445, 2019.

[34] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Annual Conference on Neural Information Processing Systems*, pages 118–128, 2017.

[35] Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In *Advances in Neural Information Processing Systems, Vancouver, BC, Canada*, pages 8632–8642, 2019.

[36] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant SGD by inner product manipulation. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence, Tel Aviv, Israel*, page 83, 2019.

[37] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local model poisoning attacks to Byzantine-robust federated learning. *CoRR*, abs/1911.11815, 2019.

[38] Jiani Li, Waseem Abbas, Mudassir Shabbir, and Xenofon Koutsoukos. Resilient distributed diffusion for multi-robot systems using centerpoint. In *Robotics: Science and Systems, Corvalis, Oregon, USA*, 2020.

[39] Roula Nassif, Stefan Vlaski, Cédric Richard, Jie Chen, and Ali H. Sayed. Multitask learning over graphs: An approach for distributed, streaming machine learning. *IEEE Signal Process. Mag.*, 37(3):14–25, 2020.

[40] Jie Chen, Cédric Richard, and Ali H. Sayed. Multitask diffusion adaptation over networks. *IEEE Transactions on Signal Processing*, 62(16):4129–4144, Aug 2014.

[41] Danqi Jin, Jie Chen, Cédric Richard, Jingdong Chen, and Ali H. Sayed. Affine combination of diffusion strategies over networks. *IEEE Transactions on Signal Processing*, 68:2087–2104, 2020.

[42] Jie Chen, Cédric Richard, Shang Kee Ting, and Ali H. Sayed. Chapter 3 - multitask learning over adaptive networks with grouping strategies. In Petar M. Djurić and Cédric Richard, editors, *Cooperative and Graph Signal Processing*, pages 107 – 129. Academic Press, 2018.

[43] Jie Chen, Cédric Richard, and Ali H. Sayed. Diffusion LMS for clustered multitask networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy*, pages 5487–5491, 2014.

[44] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[45] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018.

A Assumptions and Theoretical Results

A.1 Assumptions of the loss functions

Definition 1. *(L-Lipschitz continuous gradient). A differentiable convex function f is said to have an L-Lipschitz continuous gradient, if there exists a constant $L > 0$, such that*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y.$$

If f has an L-Lipschitz continuous gradient, then it holds that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \forall x, y.$$

Definition 2. *(m-strongly convex). A differentiable convex function f is said to be m-strongly convex if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|y - x\|^2, \forall x, y.$$

If f is m-strongly convex and $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$, then it holds that

$$\|\nabla f(x)\|^2 \geq 2m(f(x) - f(x^*)), \forall x.$$

If f is m-strongly convex and has an L-Lipschitz continuous gradient, then it is obvious that $m \leq L$.

A.2 Proof of Lemma 1

Proof. By sending $\|\hat{\theta}_{b,i} - \tilde{\theta}_k^*\| \ll \|\hat{\theta}_{k,i} - \tilde{\theta}_k^*\|$ and $\|\hat{\theta}_{b,i} - \theta_k^*\| > \|\hat{\theta}_{k,i} - \theta_k^*\|$, a Byzantine agent b can gain a large weight from k by the first condition and make $\theta_{k,i}$ move away from θ_k^* by the second condition. (The same strategy can be generalized to ℓ_p norm.) □

A.3 Optimal solution of equation (9)

Let λ be the Lagrange multiplier. We define the Lagrangian of (9) given the constraints on the weights as

$$\mathcal{L}(a_{lk}, \lambda) = \sum_{l\in\mathcal{N}_k} a_{lk}^2 r_k(\hat{\theta}_{l,i}^{(\text{coop})}) + \lambda(1 - \sum_{l\in\mathcal{N}_k} a_{lk}).$$

Set $\nabla_{a_{lk},\lambda}\mathcal{L}(a_{lk}, \lambda) = \left(\frac{\partial\mathcal{L}}{\partial a_{lk}}, \frac{\partial\mathcal{L}}{\partial\lambda}\right) = 0$, i.e.,

$$\begin{cases} 2a_{lk}r_k(\hat{\theta}_{l,i}^{(\text{coop})}) - \lambda = 0, \forall l \in \mathcal{N}_k, \\ 1 - \sum_{l\in\mathcal{N}_k} a_{lk} = 0. \end{cases}$$

Thus, $a_{lk} = \frac{\lambda}{r_k(\hat{\theta}_{l,i}^{(\text{coop})})}, \forall l \in \mathcal{N}_k$ and $\sum_{l\in\mathcal{N}_k} a_{lk} = 1$. We have $\lambda\sum_{l\in\mathcal{N}_k}\frac{1}{r_k(\hat{\theta}_{l,i}^{(\text{coop})})} = 1$ and hence $\lambda = \frac{1}{\sum_{l\in\mathcal{N}_k} r_k(\hat{\theta}_{l,i}^{(\text{coop})})^{-1}}$, and $a_{lk} = \frac{r_k(\hat{\theta}_{l,i}^{(\text{coop})})^{-1}}{\sum_{p\in\mathcal{N}_k} r_k(\hat{\theta}_{p,i}^{(\text{coop})})^{-1}}$ is the optimal solution of (9).

A.4 Proof of Lemma 2

Proof. Given (3), $r_k(\theta_{k,i}^{(\text{coop})}) = r_k\left(\sum_{l\in\mathcal{N}_k} a_{lk}(i)\hat{\theta}_{l,i}^{(\text{coop})}\right)$. Using Jensen's inequality, we have

$$r_k(\theta_{k,i}^{(\text{coop})}) \leq \sum_{l\in\mathcal{N}_k} a_{lk}(i) r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right). \tag{12}$$

Subtracting r_k^* from both sides of (12) and taking expectations over the joint distribution ξ_k, we obtain

$$\begin{aligned} \mathbb{E}\left[r_k(\theta_{k,i}^{(\text{coop})}) - r_k^*\right] &\leq \sum_{l\in\mathcal{N}_k} \mathbb{E}[a_{lk}(i)]\mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right] \\ &\leq \frac{\sum_{l\in\mathcal{N}_k}\mathbb{E}\left[r_k(\hat{\theta}_{l,i}^{(\text{coop})})\right]^{-1}\mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right]}{\sum_{p\in\mathcal{N}_k}\mathbb{E}\left[r_k(\hat{\theta}_{p,i}^{(\text{coop})})\right]^{-1}}. \end{aligned} \tag{13}$$

For succinctness, we use $\chi_{l,i}$ to denote $\mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right)\right]^{-1}$, and $\Delta_{l,i}$ to denote $\mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right]$. We next prove $\frac{\sum_{l\in\mathcal{N}_k}\chi_{l,i}\Delta_{l,i}}{\sum_{p\in\mathcal{N}_k}\chi_{p,i}} \leq \frac{1}{|\mathcal{N}_k|}\sum_{l\in\mathcal{N}_k}\Delta_{l,i}$, or equivalently, $|\mathcal{N}_k|\sum_{l\in\mathcal{N}_k}\chi_{l,i}\Delta_{l,i} \leq \sum_{p\in\mathcal{N}_k}\chi_{p,i}\sum_{l\in\mathcal{N}_k}\Delta_{l,i}$.

Assume $|\mathcal{N}_k| \geq 2$, let $l_{1,i}$ be the one with the smallest $r_k\left(\hat{\theta}^{(\text{coop})}_{l_{1,i},i}\right) = \min_{l\in\mathcal{N}_k} r_k\left(\hat{\theta}^{(\text{coop})}_{l,i}\right)$ and $l_{2,i}$ be the one with the second smallest $r_k\left(\hat{\theta}^{(\text{coop})}_{l_{2,i},i}\right) = \min_{l\in\mathcal{N}_k\setminus l_{1,i}} r_k\left(\hat{\theta}^{(\text{coop})}_{l,i}\right)$. Hence, $\chi_{l_{1,i},i} \geq \chi_{l_{2,i},i} \geq \chi_{l,i}$, and $\Delta_{l_{1,i},i} \leq \Delta_{l_{2,i},i} \leq \Delta_{l,i}$ for $l \in \mathcal{N}_k\setminus\{l_{1,i}, l_{2,i}\}$. Thus,

$$
\begin{aligned}
&|\mathcal{N}_k| \sum_{l\in\mathcal{N}_k} \chi_{l,i}\Delta_{l,i} - \sum_{p\in\mathcal{N}_k} \chi_{p,i} \sum_{l\in\mathcal{N}_k} \Delta_{l,i} \\
=&\sum_{l\in\mathcal{N}_k} \chi_{l,i} \left(|\mathcal{N}_k|\Delta_{l,i} - \sum_{p\in\mathcal{N}_k} \Delta_{p,i} \right) \\
=&\chi_{l_{1,i},i} \left((|\mathcal{N}_k|-1)\,\Delta_{l_{1,i},i} - \sum_{l\in\mathcal{N}_k\setminus l_{1,i}} \Delta_{l,i} \right) + \sum_{l\in\mathcal{N}_k\setminus l_{1,i},i} \chi_{l,i} \left(|\mathcal{N}_k|\Delta_{l,i} - \sum_{p\in\mathcal{N}_k} \Delta_{p,i} \right) \\
\leq&\chi_{l_{1,i},i} \left((|\mathcal{N}_k|-1)\,\Delta_{l_{1,i},i} - \sum_{l\in\mathcal{N}_k\setminus l_{1,i}} \Delta_{l,i} \right) + \chi_{l_{2,i},i} \left(\sum_{l\in\mathcal{N}_k\setminus l_{1,i}} |\mathcal{N}_k|\Delta_{l,i} - (|\mathcal{N}_k|-1) \sum_{p\in\mathcal{N}_k} \Delta_{p,i} \right) \\
=&\chi_{l_{1,i},i} \left((|\mathcal{N}_k|-1)\,\Delta_{l_{1,i},i} - \sum_{l\in\mathcal{N}_k\setminus l_{1,i}} \Delta_{l,i} \right) + \chi_{l_{2,i},i} \left(\sum_{l\in\mathcal{N}_k\setminus l_{1,i}} \Delta_{l,i} - (|\mathcal{N}_k|-1)\,\Delta_{l_{1,i},i} \right) \\
=&\left(\chi_{l_{1,i},i} - \chi_{l_{2,i},i}\right) \left((|\mathcal{N}_k|-1)\,\Delta_{l_{1,i},i} - \sum_{l\in\mathcal{N}_k\setminus l_{1,i}} \Delta_{l,i} \right) \\
=&\left(\chi_{l_{1,i},i} - \chi_{l_{2,i},i}\right) \left(\sum_{l\in\mathcal{N}_k\setminus l_{1,i}} \left(\Delta_{l_{1,i},i} - \Delta_{l,i}\right) \right) \leq 0.
\end{aligned}
$$

Therefore, $\frac{\sum_{l\in\mathcal{N}_k} \chi_{l,i}\Delta_{l,i}}{\sum_{p\in\mathcal{N}_k} \chi_{p,i}} \leq \frac{1}{|\mathcal{N}_k|} \sum_{l\in\mathcal{N}_k} \Delta_{l,i}$. Put it back to (13), we obtain as $i \to \infty$,

$$
\mathbb{E}\left[r_k(\theta^{(\text{coop})}_{k,i}) - r^*_k \right] \leq \frac{1}{|\mathcal{N}_k|} \sum_{l\in\mathcal{N}_k} \mathbb{E}\left[r_k\left(\hat{\theta}^{(\text{coop})}_{l,i}\right) - r^*_k \right],
$$

which completes the proof. □

A.5 Proof of Lemma 3

Proof. Given that the risk function $r_k(\cdot)$ has an L-Lipschitz continuous gradient, it holds that

$$
r_k\left(\hat{\theta}_{k,i}\right) - r_k(\theta_{k,i-1}) \leq \nabla r_k(\theta_{k,i-1})^\top (\hat{\theta}_{k,i} - \theta_{k,i-1}) + \frac{1}{2} L \|\hat{\theta}_{k,i} - \theta_{k,i-1}\|^2.
$$

Given the SGD step (2), we have

$$
r_k\left(\hat{\theta}_{k,i}\right) - r_k(\theta_{k,i-1}) \leq -\mu \nabla r_k(\theta_{k,i-1})^\top \nabla \ell_k(\theta_{k,i-1}; \xi_k^{i-1}) + \frac{1}{2}\mu^2 L \|\nabla \ell_k(\theta_{k,i-1}; \xi_k^{i-1})\|^2. \tag{14}
$$

Take the expected value of the above equation with respect to the random variable ξ_k^{i-1}. Since $\hat{\theta}_{k,i}$ depends on ξ_k^{i-1}, whereas $\theta_{k,i-1}$ does not, we obtain

$$
\begin{aligned}
&\mathbb{E}_{\xi_k^{i-1}}\left[r_k\left(\hat{\theta}_{k,i}\right) \right] - r_k(\theta_{k,i-1}) \\
\leq& -\mu \nabla r_k(\theta_{k,i-1})^\top \mathbb{E}_{\xi_k^{i-1}}[\nabla \ell_k(\theta_{k,i-1}; \xi_k^{i-1})] + \frac{1}{2}\mu^2 L \mathbb{E}_{\xi_k^{i-1}}[\|\nabla \ell_k(\theta_{k,i-1}; \xi_k^{i-1})\|^2].
\end{aligned}
$$

Given that the stochastic gradient $\nabla \ell_k(\theta_{k,i}; \xi_k)$ is an unbiased estimate of $\nabla r_k(\theta_{k,i})$, and ξ_k^i is an instantaneous realization of ξ_k, it holds that $\mathbb{E}_{\xi_k^i}[\nabla \ell_k(\theta_{k,i}; \xi_k^i)] = \nabla r_k(\theta_{k,i}), \forall i$, and the variance of $\nabla \ell_k(\theta_{k,i}; \xi_k^i)$ satisfies

$$
\text{Var}_{\xi_k^i}[\nabla \ell_k(\theta_{k,i}; \xi_k^i)] = \mathbb{E}_{\xi_k^i}[\|\nabla \ell_k(\theta_{k,i}; \xi_k^i)\|^2] - \|\mathbb{E}_{\xi_k^i}[\nabla \ell_k(\theta_{k,i}; \xi_k^i)]\|^2 \leq \sigma_k^2, \forall i.
$$

We have $\mathbb{E}_{\xi_k^i}[\|\nabla \ell_k(\theta_{k,i}; \xi_k^i)\|^2] \leq \|\mathbb{E}_{\xi_k^i}[\nabla \ell_k(\theta_{k,i}; \xi_k^i)]\|^2 + \sigma_k^2 = \|\nabla r_k(\theta_{k,i})\|^2 + \sigma_k^2$. Hence,

$$
\begin{aligned}
\mathbb{E}_{\xi_k^{i-1}}\left[r_k\left(\hat{\theta}_{k,i}\right) \right] - r_k(\theta_{k,i-1}) &\leq -\mu \|\nabla r_k(\theta_{k,i-1})\|^2 + \frac{1}{2}\mu^2 L \mathbb{E}_{\xi_k^{i-1}}[\|\nabla \ell_k(\theta_{k,i-1}; \xi_k^{i-1})\|^2] \\
&\leq -\mu \|\nabla r_k(\theta_{k,i-1})\|^2 + \frac{1}{2}\mu^2 L \left(\|\nabla r_k(\theta_{k,i-1})\|^2 + \sigma_k^2 \right) \qquad (15) \\
&= -\mu(1 - \frac{\mu L}{2}) \|\nabla r_k(\theta_{k,i-1})\|^2 + \frac{1}{2}\mu^2 L \sigma_k^2.
\end{aligned}
$$

Given the strong convexity condition of r_k, there exists $m \leq L$ such that

$$\|\nabla r_k(\theta)\|^2 \geq 2m\left(r_k(\theta) - r_k^*\right) \text{ for all } \theta.$$

Assuming $(1 - \frac{\mu L}{2}) > 0$ (which is guaranteed by the following claims), (15) turns into

$$\mathbb{E}_{\xi_k^{i-1}}\left[r_k\left(\hat{\theta}_{k,i}\right)\right] - r_k(\theta_{k,i-1}) \leq -2m\mu(1 - \frac{\mu L}{2})\left(r_k(\theta_{k,i-1}) - r_k^*\right) + \frac{1}{2}\mu^2 L\sigma_k^2. \tag{16}$$

Let $\mathbb{E}[\cdot]$ denote an expected value taken with respect to the joint distribution of all random variables, i.e.

$$\mathbb{E}\left[r_k\left(\hat{\theta}_{k,i}\right)\right] = \mathbb{E}_{\xi_k^1}\mathbb{E}_{\xi_k^2}\ldots\mathbb{E}_{\xi_k^{i-1}}\left[r_k\left(\hat{\theta}_{k,i}\right)\right].$$

Subtracting r_k^* from both sides of (16) and taking expectations over the joint distribution, we have

$$\mathbb{E}\left[r_k\left(\hat{\theta}_{k,i}\right) - r_k^*\right] \leq \left(1 - 2m\mu(1 - \frac{\mu L}{2})\right)\mathbb{E}\left[(r_k(\theta_{k,i-1}) - r_k^*)\right] + \frac{1}{2}\mu^2 L\sigma_k^2.$$

Subtracting the constant $\frac{\mu L\sigma_k^2}{2m(2-\mu L)}$ from both sides, we obtain

$$\begin{aligned}
&\mathbb{E}\left[r_k\left(\hat{\theta}_{k,i}\right) - r_k^*\right] - \frac{\mu L\sigma_k^2}{2m(2-\mu L)} \\
\leq &\left(1 - 2m\mu(1 - \frac{\mu L}{2})\right)\mathbb{E}\left[(r_k(\theta_{k,i-1}) - r_k^*)\right] + \frac{1}{2}\mu^2 L\sigma_k^2 - \frac{\mu L\sigma_k^2}{2m(2-\mu L)} \\
= &\left(1 - 2m\mu(1 - \frac{\mu L}{2})\right)\left(\mathbb{E}\left[(r_k(\theta_{k,i-1}) - r_k^*)\right] - \frac{\mu L\sigma_k^2}{2m(2-\mu L)}\right).
\end{aligned} \tag{17}$$

Note that the one-step progress of SGD given in (17) is true for both the cooperative and non-cooperative cases since they both use the SGD step (2).

In the following, we compute the expected regret for non-cooperative SGD where $\theta_{k,i} = \hat{\theta}_{k,i}$. Given (17), repeating through iteration i, we have

$$\begin{aligned}
&\mathbb{E}\left[r_k\left(\theta_{k,i}^{(\text{ncop})}\right) - r_k^*\right] \\
\leq &\frac{\mu L\sigma_k^2}{2m(2-\mu L)} + \left(1 - 2m\mu(1 - \frac{\mu L}{2})\right)^i\left(\mathbb{E}\left[\left(r_k(\theta_{k,0}^{(\text{ncop})}) - r_k^*\right)\right] - \frac{\mu L\sigma_k^2}{2m(2-\mu L)}\right).
\end{aligned} \tag{18}$$

Since $1 - 2m\mu(1 - \frac{\mu L}{2}) = 1 + m\mu(L\mu - 2) \leq 1 + L\mu(L\mu - 2) = (L\mu - 1)^2$, by selecting fixed stepsize $\mu \in (0, \frac{2}{L}]$, $\left(1 - 2m\mu(1 - \frac{\mu L}{2})\right) \in [0, 1)$, it is guaranteed that $r_k\left(\theta_{k,i}^{(\text{ncop})}\right)$ converges towards r_k^*. And the expected regret as $i \to \infty$ satisfies

$$\lim_{i\to\infty}\sup R_k^{(\text{ncop})}(i) = \lim_{i\to\infty}\sup\mathbb{E}\left[r_k\left(\theta_{k,i}^{(\text{ncop})}\right) - r_k^*\right] = \frac{\mu L\sigma_k^2}{2m(2-\mu L)},$$

which completes the proof. □

A.6 Proof of Theorem 1

Proof. Let $\mathbb{E}[\cdot]$ denote the expected value taken with respect to the joint distribution of all random variables ξ_k and ξ_l for $l \in \mathcal{N}_k^{\leq}$, i.e.

$$\mathbb{E}[\cdot] = \mathbb{E}_{\xi_k}\mathbb{E}_{\{\xi_l | l\in\mathcal{N}_k^{\leq}\}}[\cdot].$$

Similar to the proof for Lemma 2, using $\mathcal{N}_k^{\leq}$ in the place of $\mathcal{N}_k$, with rule (11), we obtain

$$\mathbb{E}\left[r_k(\theta_{k,i}^{(\text{coop})}) - r_k^*\right] \leq \frac{1}{|\mathcal{N}_k^{\leq}|}\sum_{l\in\mathcal{N}_k}\mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right]. \tag{19}$$

For every $l \in \mathcal{N}_k^{\leq}$, we have $r_k(\hat{\theta}_{l,i}^{(\text{coop})}) \leq r_k(\hat{\theta}_{k,i}^{(\text{coop})})$ and hence $\mathbb{E}\left[\left(r_k(\hat{\theta}_{l,i}^{(\text{coop})}) - r_k^*\right)\right] \leq \mathbb{E}\left[\left(r_k(\hat{\theta}_{k,i}^{(\text{coop})}) - r_k^*\right)\right]$. Put it back to (19), we obtain

$$\mathbb{E}\left[r_k\left(\theta_{k,i}^{(\text{coop})}\right) - r_k^*\right] \leq \frac{1}{|\mathcal{N}_k^{\leq}|}\sum_{l\in\mathcal{N}_k^{\leq}}\mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right] \leq \mathbb{E}\left[r_k\left(\hat{\theta}_{k,i}^{(\text{coop})}\right) - r_k^*\right]. \tag{20}$$

Inequation (20) implies that the cooperation step (3) using rule (11) reduces the expected regret as compared to the non-cooperative case. As a result, if the non-cooperative SGD converges, then the cooperative SGD using (11) also converges. Given the results of Lemma 3, when $\mu \in (0, \frac{2}{L}]$, the non-cooperative SGD converges. Hence, using $\mu \in (0, \frac{2}{L}]$, the cooperative SGD using (11) also converges, i.e., as $i \to \infty$, $\lim_{i\to\infty} \mathbb{E}\left[r_k\left(\theta_{k,i}^{(\text{coop})}\right) - r_k^*\right] = 0$, $\lim_{i\to\infty} \nabla \ell_k(\theta_{k,i}^{(\text{coop})}; \xi_k^i) = 0$, and $\lim_{i\to\infty} \hat{\theta}_{k,i}^{(\text{coop})} = \lim_{i\to\infty} \theta_{k,i}^{(\text{coop})} = \theta_k^*$, $\forall k \in \mathcal{N}^+$. For $l \in \mathcal{N}_k^{\leq}$, since $\mathbb{E}\left[\left(r_k(\hat{\theta}_{l,i}^{(\text{coop})}) - r_k^*\right)\right] \leq \mathbb{E}\left[\left(r_k(\hat{\theta}_{k,i}^{(\text{coop})}) - r_k^*\right)\right]$, it holds that $\lim_{i\to\infty} \mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right] = 0$. If $l \in \mathcal{N}^+$, then $\lim_{i\to\infty} \hat{\theta}_{l,i}^{(\text{coop})} = \lim_{i\to\infty} \theta_{l,i}^{(\text{coop})}$, thus $\lim_{i\to\infty} \mathbb{E}\left[r_k\left(\theta_{l,i}^{(\text{coop})}\right) - r_k^*\right] = 0$, and $\lim_{i\to\infty} \theta_{l,i}^{(\text{coop})} = \theta_k^*$. This implies that k and l share the same target θ_k^*, and therefore they are from the same cluster.

Note that to disrupt the convergence, a Byzantine neighbor b needs to send a large $r_k(\hat{\theta}_{b,i}^{(\text{coop})})$ and therefore it will not be included in the cooperation by rule (11). If the Byzantine agent sends $r_k(\hat{\theta}_{b,i}^{(\text{coop})}) \leq r_k(\hat{\theta}_{k,i}^{(\text{coop})})$, then it reduces the upper bound of $\mathbb{E}\left[r_k(\theta_{k,i}^{(\text{coop})}) - r_k^*\right]$ by (20) and will not do any damage to the system. Therefore, we can treat it as a normal agent within the same cluster of k in such a case. This shows the resilience of the weights (11) in the presence of an arbitrary number of Byzantine agents.

Since k and $l \in \mathcal{N}_k^{\leq}$ are in the same cluster, assume $\{k, l\} \in \mathcal{C}_j$. Similar to (14), for $l \in \mathcal{N}_k^{\leq}$, we have

$$\begin{aligned} r_k\left(\hat{\theta}_{l,i}\right) - r_k(\theta_{l,i-1}) &\leq \nabla r_k(\theta_{l,i-1})^\top (\hat{\theta}_{l,i} - \theta_{l,i-1}) + \frac{1}{2} L \|\hat{\theta}_{l,i} - \theta_{l,i-1}\|^2 \\ &= -\mu \nabla r_k(\theta_{l,i-1})^\top \nabla \ell_j(\theta_{l,i-1}; \xi_l^{i-1}) + \frac{1}{2}\mu^2 L \|\nabla \ell_j(\theta_{l,i-1}; \xi_l^{i-1})\|^2. \end{aligned}$$

Given that the stochastic gradient $\nabla \ell_j(\theta_{l,i}; \xi_l)$ is an unbiased estimate of $\nabla r_k(\theta_{l,i})$ and ξ_l^i is an instantaneous realization of ξ_l, it holds that $\mathbb{E}_{\xi_l^i}[\nabla \ell_j(\theta_{l,i}; \xi_l^i)] = \nabla r_k(\theta_{l,i})$, and the variance of $\nabla \ell_j(\theta_{l,i}; \xi_l^i)$ satisfies

$$\text{Var}_{\xi_l^i}[\nabla \ell_j(\theta_{l,i}; \xi_l^i)] = \mathbb{E}_{\xi_l^i}[\|\nabla \ell_j(\theta_{l,i}; \xi_l^i)\|^2] - \|\mathbb{E}_{\xi_l^i}[\nabla \ell_j(\theta_{l,i}; \xi_l^i)]\|^2 \leq \sigma_l^2, \forall i.$$

Following the claims in Appendix A.5, we obtain

$$\begin{aligned} &\mathbb{E}_{\xi_l}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right] \\ \leq &\frac{\mu L \sigma_l^2}{2m(2-\mu L)} + \left(1 - 2m\mu(1 - \frac{\mu L}{2})\right)\left(\mathbb{E}_{\xi_l}\left[\left(r_k(\theta_{l,i-1}^{(\text{coop})}) - r_k^*\right)\right] - \frac{\mu L \sigma_l^2}{2m(2-\mu L)}\right). \end{aligned}$$

Given (19), we obtain

$$\begin{aligned} &\mathbb{E}\left[r_k(\theta_{k,i}^{(\text{coop})}) - r_k^*\right] \leq \frac{1}{|\mathcal{N}_k^{\leq}|} \sum_{l \in \mathcal{N}_k^{\leq}} \mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right] \\ \leq &\frac{1}{|\mathcal{N}_k^{\leq}|} \sum_{l \in \mathcal{N}_k^{\leq}} \left(\frac{\mu L \sigma_l^2}{2m(2-\mu L)} + \left(1 - 2m\mu(1 - \frac{\mu L}{2})\right)\left(\mathbb{E}\left[\left(r_k(\theta_{l,i-1}^{(\text{coop})}) - r_k^*\right)\right] - \frac{\mu L \sigma_l^2}{2m(2-\mu L)}\right)\right) \\ = &\frac{1}{|\mathcal{N}_k^{\leq}|} \frac{\mu L \sum_{l \in \mathcal{N}_k^{\leq}} \sigma_l^2}{2m(2-\mu L)} + \left(1 - 2m\mu(1 - \frac{\mu L}{2})\right)\left(\frac{1}{|\mathcal{N}_k^{\leq}|} \sum_{l \in \mathcal{N}_k^{\leq}} \mathbb{E}\left[\left(r_k(\theta_{l,i-1}^{(\text{coop})}) - r_k^*\right)\right] - \frac{1}{|\mathcal{N}_k^{\leq}|} \frac{\mu L \sum_{l \in \mathcal{N}_k^{\leq}} \sigma_l^2}{2m(2-\mu L)}\right). \end{aligned} \tag{21}$$

As discussed above, the cooperative SGD using rule (11) converges. Therefore, it holds that as $t \to \infty$ and $i > t$, $\nabla \ell_j(\theta_{k,i}^{(\text{coop})}; \xi_k^i) = 0$, and $\nabla \ell_j(\theta_{l,i}^{(\text{coop})}; \xi_l^i) = 0$. Moreover, from (2), we deduce that $\hat{\theta}_{k,i}^{(\text{coop})} = \theta_{k,i-1}^{(\text{coop})}$ and $\hat{\theta}_{l,i}^{(\text{coop})} = \theta_{l,i-1}^{(\text{coop})}$. Since $\mathbb{E}\left[\left(r_k(\hat{\theta}_{l,i}^{(\text{coop})}) - r_k^*\right)\right] \leq \mathbb{E}\left[\left(r_k(\hat{\theta}_{k,i}^{(\text{coop})}) - r_k^*\right)\right]$ for $l \in \mathcal{N}_k^{\leq}$, it also holds that $\mathbb{E}\left[\left(r_k(\theta_{l,i-1}^{(\text{coop})}) - r_k^*\right)\right] \leq \mathbb{E}\left[\left(r_k(\theta_{k,i-1}^{(\text{coop})}) - r_k^*\right)\right]$. Hence, $\frac{1}{|\mathcal{N}_k^{\leq}|} \sum_{l \in \mathcal{N}_k^{\leq}} \mathbb{E}\left[\left(r_k(\theta_{l,i-1}^{(\text{coop})}) - r_k^*\right)\right] \leq \mathbb{E}\left[\left(r_k(\theta_{k,i-1}^{(\text{coop})}) - r_k^*\right)\right]$. Put it back to (21), we obtain

$$\begin{aligned} &\mathbb{E}\left[r_k(\theta_{k,i}^{(\text{coop})}) - r_k^*\right] \\ \leq &\frac{1}{|\mathcal{N}_k^{\leq}|} \frac{\mu L \sum_{l \in \mathcal{N}_k^{\leq}} \sigma_l^2}{2m(2-\mu L)} + \left(1 - 2m\mu(1 - \frac{\mu L}{2})\right)\left(\mathbb{E}\left[\left(r_k(\theta_{k,i-1}^{(\text{coop})}) - r_k^*\right)\right] - \frac{1}{|\mathcal{N}_k^{\leq}|} \frac{\mu L \sum_{l \in \mathcal{N}_k^{\leq}} \sigma_l^2}{2m(2-\mu L)}\right) \\ = &\frac{1}{|\mathcal{N}_k^{\leq}|} \frac{\mu L \sum_{l \in \mathcal{N}_k^{\leq}} \sigma_l^2}{2m(2-\mu L)} + \left(1 - 2m\mu(1 - \frac{\mu L}{2})\right)^{i-t}\left(\mathbb{E}\left[\left(r_k(\theta_{k,t}^{(\text{coop})}) - r_k^*\right)\right] - \frac{1}{|\mathcal{N}_k^{\leq}|} \frac{\mu L \sum_{l \in \mathcal{N}_k^{\leq}} \sigma_l^2}{2m(2-\mu L)}\right). \end{aligned}$$

Given fixed stepsize $\mu \in (0, \frac{2}{L}]$, $\left(1 - 2m\mu(1 - \frac{\mu L}{2})\right) \in [0, 1)$, it is guaranteed that $r_k\left(\theta_{k,i}^{(\text{coop})}\right)$ converges towards r_k^* as $(i - t) \to \infty$. And the expected regret satisfies

$$\lim_{i\to\infty} \sup R_k^{(\text{coop})}(i) = \lim_{i\to\infty} \sup \mathbb{E}\left[r_k(\theta_{k,i}^{(\text{coop})}) - r_k^*\right] = \frac{\mu L}{2m(2-\mu L)} \frac{1}{|\mathcal{N}_k^{\leq}|} \sum_{l\in\mathcal{N}_k^{\leq}} \sigma_l^2. \tag{22}$$

Note that equation (20) gives another upper bound of the expected regret: $\mathbb{E}\left[r_k\left(\theta_{k,i}^{(\text{coop})}\right) - r_k^*\right] \leq \mathbb{E}\left[r_k\left(\hat{\theta}_{k,i}^{(\text{coop})}\right) - r_k^*\right]$. Using this condition, we obtain

$$\lim_{i\to\infty} \mathbb{E}\left[r_k(\theta_{k,i}^{(\text{coop})}) - r_k^*\right] \leq \frac{\mu L \sigma_k^2}{2m(2-\mu L)} = \lim_{i\to\infty} \sup R_k^{(\text{ncop})}(i). \tag{23}$$

Given (20), it holds that

$$\lim_{i\to\infty} \frac{1}{|\mathcal{N}_k^{\leq}|} \sum_{l\in\mathcal{N}_k^{\leq}} \mathbb{E}\left[r_k\left(\hat{\theta}_{l,i}^{(\text{coop})}\right) - r_k^*\right] \leq \lim_{i\to\infty} \mathbb{E}\left[r_k\left(\hat{\theta}_{k,i}^{(\text{coop})}\right) - r_k^*\right].$$

Thus, (22) gives a tighter upper bound of the expected regret than (23). And we conclude

$$\lim_{i\to\infty} \sup R_k^{(\text{coop})}(i) \leq \lim_{i\to\infty} \sup R_k^{(\text{ncop})}(i),$$

which implies that $\sigma_l^2 \leq \sigma_k^2$ for $l \in \mathcal{N}_k^{\leq}$. □

B Simulation Details and Supplementary Results

B.1 Simulation details of Target Localization

The four target locations in $\mathbb{R}^2$ are: $(10.84, 10.76), (20.42, 20.26), (20.51, 10.40), (10.78, 20.30)$. Agents' locations are indicated in Figure 1a. An edge between two agents means they are neighbors. At each iteration, every agent k has a noisy observation (streaming data) of the distance $\boldsymbol{d}_k(i)$ and the unit direction vector $\boldsymbol{u}_{k,i}$ pointing from x_k to its target based on built-in sensors. Let $\theta_k \in \mathbb{R}^2$ denote the estimation of the target location for agent k, then the loss is computed as $\ell_k(\theta_{k,i}; \xi_k^i) = \|\boldsymbol{d}_k(i) - (\theta_k - x_k)^\top \boldsymbol{u}_{k,i}\|^2$, and the agent estimate θ_k using the SGD algorithm as well as the cooperative SGD algorithms with different weight assignment rules. The distance measurement data has noise variance $\sigma_{d,k}^2 \in [0.1, 0.2]$, and the unit direction vector has additive white Guassian noise with diagnonal covariance matrices $R_{u,k} = \sigma_{u,k}^2 I_2$, with $\sigma_{u,k}^2 \in [0.01, 0.1]$ for different k. We tune the step-sizes and forgetting factors from the interval $(0, 1)$ and find the best empirical performance by setting them to be $\mu_k = 0.1$ and $\nu_k = 0.1$ for every normal agent k. φ_{lk}^{-1} and ϕ_{lk}^{-1} are initialized to be zero for all $l \in \mathcal{N}_k$. Byzantine agents are designed to continuously send random values for each dimension from the interval $[15, 16]$ at each iteration.

B.2 Simulation details and supplementary results of Human Action Recognition

We randomly split the data into 75% training and 25% testing for each agent. During training, ten of the thirty agents are randomly selected to have access to much less data (about $\frac{1}{10}^{th}$) than the other agents at each epoch. This is to model the realistic scenario in which some of the agents may have less data samples and they may learn slowly than others. We use mini-batch gradient descent with batch size of 10. We tune the step-sizes and forgetting factors from the interval $(0, 1)$ and find the best empirical performance by setting them to be $\mu_k = 0.01$ and $\nu_k = 0.05$ for every normal agent k. φ_{lk}^{-1} and ϕ_{lk}^{-1} are initialized to be zero for all $l \in \mathcal{N}_k$. Byzantine agents are designed to send a model with very small noisy elements for each dimension from the interval $[0, 0.1]$ at each iteration.

Figure 5 shows the average *testing* loss and classification accuracy of the normal agent when 29 out of 30 agents are Byzantine (the only normal agent has access to the entire training data). Figure 6 and Figure 7 show the mean and range of the average *training* loss and classification accuracy of the normal agents in the case of no attack, with 10 random selected Byzantine agents, and with 29 Byzantine agents. In all the examples, for both training and testing, we observe that the loss-based weight assignment rule (11) outperforms the other rules as well as the non-cooperative case, with respect to the mean and range of the average loss and accuracy, which validates the result indicated by (6). Even in the extreme case in which there is only one normal agent in the network and all of its neighbors are Byzantine, the loss-based weight assignment rule (11) has the same performance as the non-cooperative case, showing its resilience to an arbitrary number of Byzantine agents.

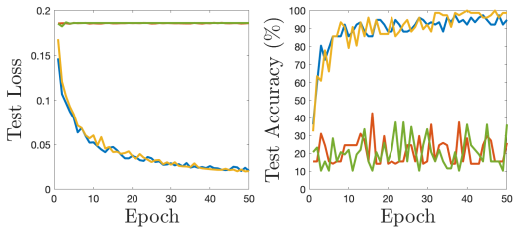
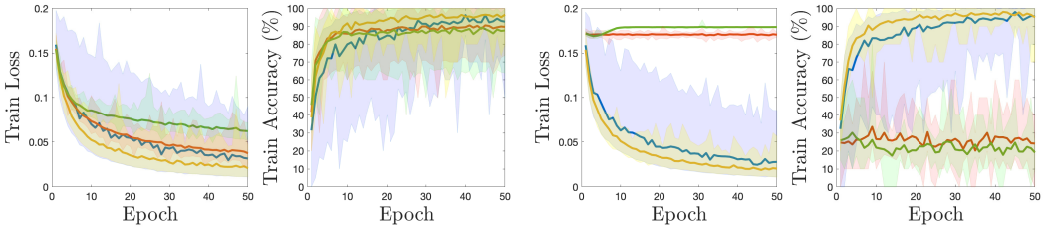


Figure 5: Human Action Recognition: average testing loss and accuracy for normal agents with 29 Byzantine agents.



(a) No attack (b) 10 Byzantine agents

Figure 6: Human Action Recognition: average training loss and accuracy for normal agents.

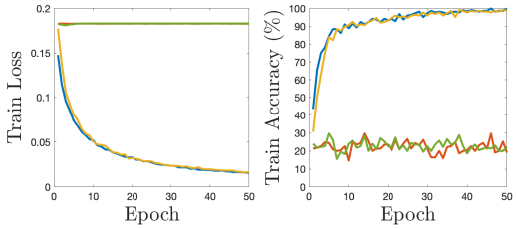


Figure 7: Human Action Recognition: average training loss and accuracy for normal agents with 29 Byzantine agents.

B.3 Simulation details and supplementary results of Digit Classification

The preprocessed examples of the two datasets are given in Figure 8. The details of the CNN architecture is given in Table 1. For each group, we consider that agents have access to uneven sizes of training data. Specifically, for each agent, we randomly feed $200 - 2000$ training data and 400 testing data from the corresponding dataset for each epoch. We use mini-batch gradient descent with batch size of 64. We tune the step-sizes and forgetting factors from the interval $(0, 1)$ and find the best empirical performance by setting them to be $\mu_k = 0.001$ and $\nu_k = 0.05$ for every normal agent. φ_{lk}^{-1} and ϕ_{lk}^{-1} are initialized to be zero for all $l \in \mathcal{N}_k$. Byzantine agents are designed to send a model with very small noisy elements for each dimension from the interval $[0, 0.1]$ at each iteration.

Since the performance of agents in the two groups diverges, we plot the results separately for the two groups. Figure 9 and Figure 10 show the average *testing* loss and classification accuracy of the normal agents in group 1 and group 2, when 8 out of 10 agents (four for each group) are Byzantine (the only normal agent in each group has access to 2000 training data).

Figure 11 and Figure 12 show the mean and range of the average *training* loss and classification accuracy of the normal agents in group 1, in the case of no attack, with 2 Byzantine agents, and with 8 Byzantine agents, which are selected randomly. Figure 13 and Figure 14 show the mean and range of the average *training* loss and classification accuracy of the normal agents in group 2, in the case of no attack, with 2 Byzantine agents, and with 8 Byzantine agents (again selected randomly). In all the examples, for both training and testing, we observe that the loss-based weight assignment rule (11) outperforms the other rules as well as the non-cooperative case, with respect to the mean and range of the average loss and accuracy, thereby validating the result indicated by (6). Even in the extreme case in which there is only one normal agent in each group and all of the other agents

are Byzantine, the loss-based weight assignment rule (11) has the same performance as the non-cooperative case, showing its resilience to an arbitrary number of Byzantine agents.

Comparing the results between groups 1 and 2 reveals that cooperation is most beneficial when there is a substantial divergence in agents' learning performances. Given limited training data, agents in group 1 are able to build refined models. It it is harder for agents receiving less training data in group 2 to achieve a high learning performance as the synthetic digit classification is a more challenging task than the MNIST digit classification. Using the weight assignment rule (11), those agents receiving less data (and therefore, struggling to learn a good model), are able to benefit from the cooperation with the neighbors having learned a refined model. At the same time, agents exhibiting high learning performance will not be negatively affected by such cooperation.

(a) MNIST (b) Synthetic digits

Figure 8: Examples of the digit classification dataset

Table 1: CNN architecture of Digit Classification

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 32, 28, 28]	320
ReLU-2	[-1, 32, 28, 28]	0
MaxPool2d-3	[-1, 32, 14, 14]	0
Conv2d-4	[-1, 64, 14, 14]	18,496
ReLU-5	[-1, 64, 14, 14]	0
MaxPool2d-6	[-1, 64, 7, 7]	0
Conv2d-7	[-1, 64, 7, 7]	36,928
ReLU-8	[-1, 64, 7, 7]	0
MaxPool2d-9	[-1, 64, 3, 3]	0
Linear-10	[-1, 128]	73,856
ReLU-11	[-1, 128]	0
Linear-12	[-1, 10]	1,290

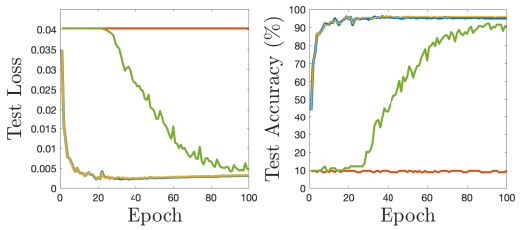


Figure 9: Digit Classification: average testing loss and accuracy for normal agents in group 1, with 8 Byzantine agents (four for each group).

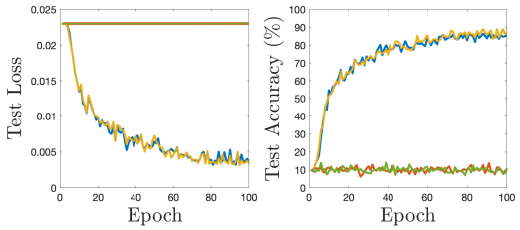
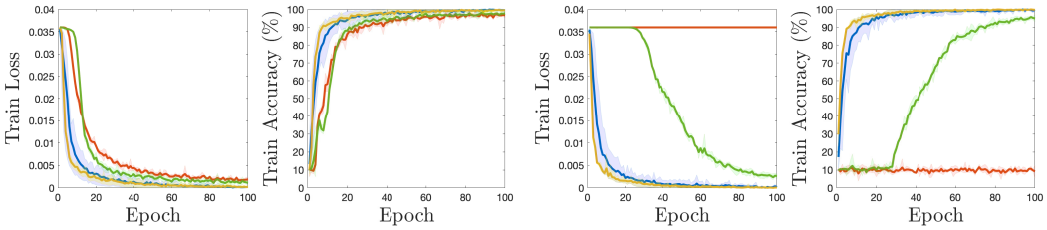


Figure 10: Digit Classification: average testing loss and accuracy for normal agents in group 2, with 8 Byzantine agents (four for each group).



(a) No attack (b) 2 Byzantine agents

Figure 11: Digit Classification: average training loss and accuracy for normal agents in group 1.

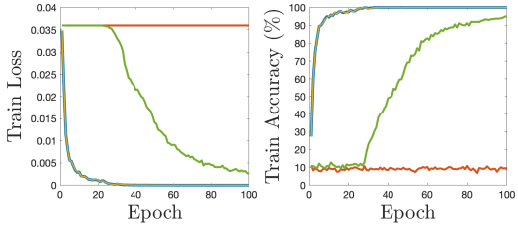
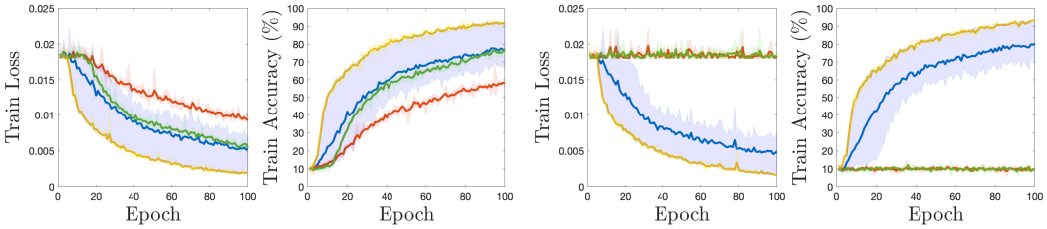


Figure 12: Digit Classification: average training loss and accuracy for normal agents in group 1, with 8 Byzantine agents (four for each group).



(a) No attack (b) 2 Byzantine agents

Figure 13: Digit Classification: average training loss and accuracy for normal agents in group 2.

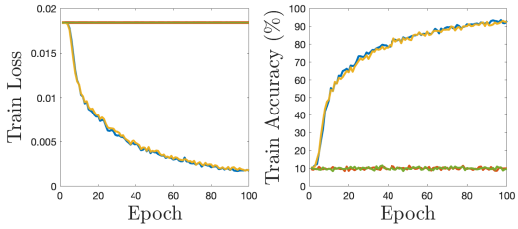


Figure 14: Digit Classification: average training loss and accuracy for normal agents in group 2, with 8 Byzantine agents (four for each group).