2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Online Learning with Adversaries: A Differential-Inclusion Analysis

Swetha Ganesh, Alexandre Reiffers-Masson, Gugan Thoppe

***Abstract*— We introduce an observation-matrix-based framework for fully asynchronous online Federated Learning (FL) with adversaries. In this work, we demonstrate its effectiveness in estimating the mean of a random vector. Our main result is that the proposed algorithm almost surely converges to the desired mean μ. This makes ours the first asynchronous FL method to have an a.s. convergence guarantee in the presence of adversaries. We derive this convergence using a novel differential-inclusion-based two-timescale analysis. Two other highlights of our proof include (a) the use of a novel Lyapunov function to show that μ is the unique global attractor for our algorithm's limiting dynamics, and (b) the use of martingale and stopping-time theory to show that our algorithm's iterates are almost surely bounded.**

I. Introduction

Federated Learning (FL) [10] is a paradigm for multiple edge/client nodes to collaborate and iteratively solve some global problem with the help of a central server. It has therefore garnered significant interest in machine [19] and reinforcement learning [12]. However, most existing FL methods do not account for failures or adversarial clients, making them ineffective in practice. Also, among those that do, a majority are synchronous [2], [5], [8], [9], [14], [18]: the server waits for inputs from a large number of clients before updating the global estimate. These approaches again are impractical because many edge devices are frequently offline and, when that is not the case, the (inevitable) slow devices decide the overall performance. These issues have put the focus on asynchronous FL, wherein the server updates as soon as one node provides its input. Our work introduces a radically new family of such FL methods.

To our knowledge, there exist only four asynchronous FL methods in the literature: (i.) Kardam [4], (ii.) Zeno++ [15], (iii.) AFLGuard [6], and (iv.) BASGD [17]. The first three use a sophisticated scoring rule for filtering out malicious estimates. However, Kardam's issue is that it drops many correct estimates during attacks. On the other hand, Zeno++ and AFLGuard require a separate validation dataset at the parameter server, which is undesirable from the privacy viewpoint. Finally, BASGD is asynchronous only at the client side: the time between successive server updates is dictated by the stragglers, like in synchronous FL methods. Our proposed approach has none of the above issues.

We only consider the mean estimation FL problem here and discuss our proposed approach only in that context. This problem involves $p \geq 1$ clients (a small but unknown subset of which are malicious), whose joint goal is to estimate the mean μ of a random vector $X \in \mathbb{R}^d$, where $d \leq p$. At its core, our approach considers a tall observation matrix $A \in \mathbb{R}^{p \times d}$ and provides each node i access to the IID samples of the random variable $Y(i) := a_i^T X$, where T denotes transpose and a_i^T is the i-th row of A. We task node i to locally estimate the mean of $Y(i)$. Separately, at every instance $n \geq 0$, the server is tasked to pick a client at random and request for its current local estimate. A honest client is expected to provide its actual estimate; the malicious agent can act arbitrarily (it can even collude with other attackers). The server then is to immediately update its μ-estimate using the gradient of $|a_i^T x - y_n(i)|$, i.e., using one update step of the SGD algorithm that solves $\min_x \|Ax - \mathbb{E}[Y]\|_1$, where $\|\cdot\|_1$ is the ℓ_1 norm. Clearly, the above algorithm is asynchronous since the server updates the μ-estimate immediately upon receiving an input from a client.

The basis for our above approach is as follows. Suppose a matrix A and the vector $b = Ax_*$ are known, but not the vector x_*. Then a natural way to recover x_* is to solve the linear system $Ax = b$. In [7], a variant of this problem is discussed. There, the goal is again to recover x_* but assuming knowledge of only the vector $b' = b + e$ instead of b. The vector e is presumed sparse and represents a one-time malicious attack. Due to e's sparsity, solving for x that minimizes the ℓ_1-norm of $Ax - b'$ is now the natural way to recover x_*. A key result in [7] is that the observation matrix A being robust (see (3) in our work), i.e., has suitable redundancy depending on e's sparsity, is the necessary and sufficient condition for x_* to be the unique solution to this ℓ_1-minimization problem. The vector corresponding to b in our setup is $\mathbb{E}[Y]$. While b' provides the value of b in the non-attacked (but unknown) coordinates, $\mathbb{E}[Y]$ is fully unknown in our case. Thus, our proposed approach above can be seen as a modification of the one in [7] that obtains online estimates of both $\mathbb{E}[Y]$ and μ simultaneously. Note that the malicious agents in [7] attack only once. In contrast, in our case, since the server (unknowingly) will query every malicious node infinitely many times during the algorithm's run, each such node will have infinitely many opportunities to poison the μ-estimation update rule.

Our main contributions can be summarized as follows.

S. Ganesh and G. Thoppe are with the Computer Science and Automation Department, Indian Institute of Science (IISc), Bengaluru 560012, India. SG's research was supported by the Prime Minister's Research Fellowship (PMRF). GT's research was supported in part by DST-SERB's Core Research Grant CRG/2021/00833, in part by DST-SERB's SIRE Fellowship SIR/2022/000444, in part by IISc Startup grants SG/MHRD-19-0054 and SR/MHRD-19-0040, and in part by the Pratiksha Trust Young Investigator award. GT also wishes to thank Michel Benaim for several insightful discussions related to this work during his visit to the University of Neuchatel, Switzerland. Emails: swethaganesh@iisc.ac.in, gthoppe@iisc.ac.in

A. Reiffers-Masson is with the Computer Science Department, IMT Atlantique, 655 Av. du Technopôle, 29280 Plouzané, France. alexandre.reiffers-masson@imt-atlantique.fr

2023 62nd IEEE Conference on Decision and Control (CDC) | 979-8-3503-0124-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CDC49753.2023.10384052

1) **Algorithm**: We propose a novel fully asynchronous FL algorithm and demonstrate its effectiveness for mean estimation. Unlike the sophisticated filtering scheme or the non-corrupt private dataset of Kardam, Zeno++, and AFLGuard, our approach uses a observation matrix to handle adversaries. This matrix choice is not unique; thus, we have a family of algorithms for solving the same problem. Separately, since the gradient of the $\|\cdot\|_1$-error involves a sign function, our algorithm ensures that the impact of an adversarial node in each iteration is limited to a sign change.
2) **Result**: Our main result is that A being robust (as in [7]) is again a necessary and sufficient condition for our algorithm's iterates to converge to μ almost surely. This makes our proposed approach the first asynchronous FL algorithm to have an a.s. convergence guarantee in the presence of adversaries.
3) **Analysis**: Our analysis is novel compared to the existing FL literature. It builds on the Differential Inclusion (DI) and the two-timescale stochastic-approximation theory. The two-timescale part arises because our approach estimates $\mathbb{E}[Y]$ and μ using two stepsize sequences that decay to 0 at different rates. In contrast, we use a DI—a set-valued generalization of an Ordinary Differential Equation (ODE)—to mainly account for the multitude of choices available to an adversarial client in each iteration. DI theory is being used for the first time for analyzing an algorithm in adversarial settings. There are two additional highlights of our proof.
   a) *Lyapunov Function*: Our algorithm is based on the gradient-descent idea for minimizing $\|Ax - \mathbb{E}[Y]\|_1$. Typically, for analyzing such a method, the natural Lyapunov function would have been the objective function. However, in our adversarial setting, we have been unable to verify this claim. We instead prove that $\|x - \mu\|_2^2$ behaves as a Lyapunov function.
   b) *Boundedness of Iterates*: A key step in any ODE/DI based analysis [3] of stochastic algorithms is to show that the algorithm's iterates are stable. In this work, we use a novel martingale and stopping time based approach to show that the algorithm's iterates are almost surely bounded.

II. Setup, Algorithm, and Main Result

We describe here the statistical problem we study, our proposed algorithm to solve it, and our main result that describes the limiting behavior of this algorithm.

Setup: $X \in \mathbb{R}^d$ is a random variable with finite mean and finite covariance matrix entries. There are p agents to collect statistics about X, but an unknown subset M, with $|M| \leq m$, are malicious or adversarial. Specifically, the i-th agent has access to samples of the random variable $Y(i) := a_i^T X$, where $a_i \in \mathbb{R}^d$ is a known deterministic vector. At time $n \geq 1$, a central server picks index i_n uniformly at random from $\{1, \ldots, p\}$ and queries agent i_n for an independent sample of $Y(i_n)$. Agent i_n returns an actual sample if it is non-adversarial, and an arbitrary real number otherwise (the value can change on each query and can depend on the history[1]). In either case, $Y_n(i_n)$ denotes the obtained sample.

Goal: Develop an online algorithm to estimate $\mu := \mathbb{E}[X]$ using the sequence $(Y_n(i_n))$.

Algorithm: Our approach is based on the gradient descent idea for minimizing $\|Ax - \mathbb{E}[Y]\|_1$. Starting from an arbitrary $x_0 \in \mathbb{R}^d$ and $y_0 \in \mathbb{R}^p$, our proposed algorithm to learn μ at the central server is, for $n \geq 0$,

$$\begin{aligned} x_{n+1} &= x_n + \alpha_n a_{i_{n+1}} [\text{sign}(y_n(i_{n+1}) - a_{i_{n+1}}^T x_n)] \\ y_{n+1} &= y_n + \beta_n [Y_{n+1}(i_{n+1}) - y_n(i_{n+1})] u_{i_{n+1}}, \end{aligned} \tag{1}$$

where u_i is i-th column of the $p \times p$-identity matrix and, for any $r \in \mathbb{R}$,

$$\text{sign}(r) = \begin{cases} -1 & \text{if } r < 0, \\ 0 & \text{if } r = 0, \\ 1 & \text{if } r > 0. \end{cases} \tag{2}$$

In (1), the variables indexed by n are known at time n, while the ones by $n+1$ are not. Note that the coordinates of y_n corresponding to malicious nodes are directly fed into x_n's update rule.

Assumptions: Apart from the conditions on $X, (i_n)$, and $Y_n(i_n)$ stated in the setup, we presume that the matrix A and stepsize sequences (α_n) and (β_n) satisfy the following.

1) **Observation matrix**: The matrix A is tall ($p > d$), has full column rank, and satisfies
$$\sum_{i \in K^c} |a_i^T x| > \sum_{i \in K} |a_i^T x| \tag{3}$$
for all $x \in \mathbb{R}^d \setminus \{0\}$ and $K \subseteq \{1, \ldots, p\}$ with $|K| = m$.
2) **Stepsize**: (α_n) and (β_n) are monotonically decreasing positive reals such that $\max\{\alpha_0, \beta_0\} \leq 1$, $\sum_{n \geq 0} \alpha_n = \sum_{n \geq 0} \beta_n = \infty$, $\lim_{n \to \infty} \alpha_n / \beta_n = \lim_{n \to \infty} \beta_n = 0$, and $\max\{\sum_{n \geq 0} \alpha_n^2, \sum_{n \geq 0} \beta_n^2, \sum_{n \geq 0} \alpha_n \gamma_n\} < \infty$, where $\gamma_n = \sqrt{\beta_n \ln(\sum_{k=0}^n \beta_k)}$. An example is $\alpha_n = n^{-\alpha}$, $\alpha \in (2/3, 1]$, and $\beta_n = n^{-\beta}$, $\beta \in (1/2, 1] \cap (2(1-\alpha), \alpha)$.

Our main result is stated below and is derived using a DI-based set-valued analysis. As we discuss in Section II-A, such an analysis is natural for (1) due to its sub-gradient nature and, importantly, the presence of adversaries. Let $h : \mathbb{R}^d \to 2^{\mathbb{R}^d}$ (the power set of $\mathbb{R}^d$) be given by

$$h(x) = \left\{ \frac{1}{p} \sum_{i=1}^p a_i \lambda_i : (\lambda_1, \ldots, \lambda_p) \in \Lambda(x) \right\}, \tag{4}$$

where $\Lambda(x)$ includes all $(\lambda_1, \ldots, \lambda_p)$ for which

$$\lambda_i \in \begin{cases} \{\text{sign}(\mathbb{E}[Y(i)] - a_i^T x)\}, & i \in M^c \text{ and } a_i^T x \neq \mathbb{E}[Y(i)], \\ [-1, +1], & \text{otherwise.} \end{cases}$$

Theorem 1. *The following statements hold.*

1) μ is the unique Globally Asymptotically Stable Equilibrium (GASE) for the DI

$$\dot{x}(t) \in h(x(t)). \tag{5}$$

[1]Such adversaries are commonly referred to as omniscient.

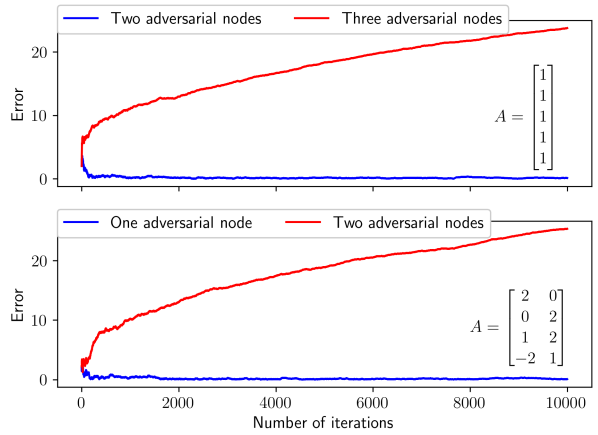

Fig. 1. Error incurred by Algorithm 1 ($\|x_n - \mu\|$) against the number of iterations (n). Within each subplot, the same measurement model is used with the only difference being the number of adversaries. The first subplot concerns the geometric median problem with $p = 5$, while the second considers a generic matrix A (see Section II-A).

2) *There exists some constant* $\Lambda > 0$ *such that*

$$\limsup_{n\to\infty} \frac{\|y_n - \mathbb{E}[Y]\|_{M^c}}{\gamma_n} \leq \Lambda \qquad a.s.,$$

where $\|y\|_{M^c} = \sqrt{\sum_{i\in M^c} y^2(i)}$.

3) $x_n \to \mu$ *a.s.*

The DI in (5) corresponds to the update rule of x_n in (1) with $y_n(i) \equiv \mathbb{E}[Y(i)]$ for $i \in M^c$, and the sign expression replaced with an arbitrary value in $[-1, +1]$, otherwise. Our first result states that every solution of this DI will converge to μ, irrespective of the sign choices made at the adversarial nodes (in a continuous time sense). The second statement provides the asymptotic rate at which $|y_n(i) - \mathbb{E}[Y(i)]| \to 0$, $i \in M^c$, on every sample path. While this result assumes that the stepsizes are square-summable, it can be extended to cover the case of even non-square summable stepsizes; see [13] for details. Our third and final result states that the actual (x_n) iterates in (1) also behave like the solutions of (5) and almost surely converge to μ. However, because the sign function is not continuous, this is not a simple consequence of the first two statements. Instead, we have to rely on a more complex two-timescale DI analysis, and a separate boundedness result for (x_n) based on the theory of martingales and stopping times.

A. Motivation for a DI-based Analysis

In this subsection, we give a simple example on why our algorithm will converge to μ even in the presence of adversarial measurements. We use a simplified set-up to illustrate the necessity of the DI analysis.

Let A be a vector of all ones. This implies that $\mathbb{E}Y(i) = \mu \in \mathbb{R}$, for all i. Our problem setup then reduces to computing $x \in \mathbb{R}$ that minimises $\sum_{i=1}^{p} |x - \mathbb{E}Y(i)|$. The solution to this minimisation problem is called the geometric median [14]. Consider Algorithm (1) in the deterministic setting, where all agents i are given $\mathbb{E}Y(i)$, instead of having to estimate it. Then, $y_n(i)$ will be μ, for $i \in M^c$, and any arbitrary value otherwise. It can be seen that the synchronous version of update (1) can be written as

$$x_{n+1} = x_n + \alpha_n \Bigg[\underbrace{\sum_{i\in M^c} \lambda_n(i)}_{\text{Unperturbed subgradient}} + \sum_{i\in M} \underbrace{\lambda_n(i)}_{\text{Adversarial noise}} \Bigg],$$

where $\lambda_n(i) = \text{sign}(\mu - x_n)$, if $i \in M^c$ and $x_n \neq \mu$, and an arbitrary value in $[-1, 1]$, otherwise. Clearly, $\lambda_n(i)$ is the subgradient of $-|x - \mathbb{E}Y(i)|$ when $i \in M^c$ and the perturbed subgradient given by the adversary, otherwise. The above update rule cannot be analysed using traditional ODE based approaches. Firstly, the update can now take a set of values at each x_n. This is because $\lambda_n(i)$, $i \in M$, can take any value in $[-1, 1]$, regardless of x_n. Moreover, $\text{sign}(\mu - x)$ is discontinuous at $x = \mu$, while ODE approaches require that this function be Lipschitz continuous. Thus, the differential inclusion approach is preferred since it is capable of handling discontinuities and capturing the evolution of a set-valued map. The associated DI for the above update is given by:

$$\dot{x}(t) \in \left\{ |M^c| \text{sign}(\mu - x) + \sum_{i\in M} v_i : v_i \in [-1, 1] \right\},$$

when $x \neq \mu$ and

$$\dot{x}(t) \in \left\{ \sum_{i=1}^{p} v_i : v_i \in [-1, 1] \right\},$$

when $x = \mu$. The DI is modified at $x = \mu$ to make it continuous in a set-valued sense.

Note that if $|M^c| > |M|$ (equivalent to (3)), it follows that $\lim_{t\to+\infty} x(t) = \mu$. The intuition is as follows: if $\mu \neq x$, the sign of $|M^c|\text{sign}(\mu - x) + \sum_{i\in M} v_i$ will be always the same as the $\text{sign}(\mu - x)$ and therefore the drift of the DI is controlled by the $\text{sign}(\mu - x)$ and not by the adversaries. The performance of our algorithm for this problem with $p = 5$ is shown in Figure 1. Here, condition (3) holds if $|M| = 2$, but not when $|M| = 3$. Consequently, our algorithm converges in the presence of two adversaries but diverges in the presence of three adversaries.

More generally, condition (3) is necessary and sufficient for our algorithm to converge. We emphasize that this condition is necessary even in the absence of noise and thus cannot be relaxed. A less obvious case where condition (3) holds is given in the bottom subplot of Figure 1. For the matrix A in this example, the condition holds for $|M| = 1$, but not when $|M| = 2$.

III. Proof of Theorem 1

We first discuss our proof strategy and then provide the details. Since $y_n(i)$'s estimate for $i \in M^c$ is not influenced by the Y samples of other nodes, one would intuitively expect $\|y_n - \mathbb{E}[Y]\|_{M^c} \to 0$. Hence, (5) is the natural object for studying (x_n)'s behaviors. However, because the sign function is discontinuous, (x_n)'s evolution cannot be viewed as a simple perturbation of (5)'s solutions as in [3,

pg. 17]. Instead, we rely on a two-timescale DI analysis [16]. Henceforth, $\|\cdot\|$ will denote the Euclidean norm.

A. Informal Outline of Two-timescale Analysis

Our algorithm (1) is of a two-timescale nature because $\alpha_n/\beta_n \to 0$. Thus, the changes in the x_n values eventually appear negligible compared to that of y_n, which, in turn, implies (x_n) and (y_n)'s behaviors can be studied in a decoupled fashion. Loosely, our analysis proceeds via the following prescribed steps from [16].

1) (y_n)'s analysis: We set $x_n \equiv x$ for some arbitrary x, and look at $y_n(i)$'s evolution for $i \in M^c$; we ignore what happens at the adversarial nodes. In our case, $y_n(i)$'s evolution is not influenced by the value of x in any way. Further, its limiting ODE can be guessed to be $\dot{z}(t) = \frac{1}{p}(\mathbb{E}[Y(i)] - z(t))$. Since this scalar ODE is linear and has $\mathbb{E}[Y(i)]$ as its unique GASE, it follows from a standard single-timescale stochastic approximation analysis [3, Chapters 2 and 3] that $|y_n(i) - \mathbb{E}[Y(i)]| \to 0$.
2) (x_n)'s analysis: From (x_n)'s perspective, (y_n) would appear to have converged to its limit point. Accordingly, in x_n's update rule, we now set $y_n(i) = \mathbb{E}[Y(i)]$, for $i \in M^c$, and allow for arbitrary values for adversarial i's. This leads to the set-valued DI dynamics (5). In the rest of this section, we formally prove that μ is its only attractor (Section III-B), the original (x_n) sequence in (1) is almost surely bounded (Section III-C), and it almost surely converges to μ (Section III-D).

B. Analysis of the DI in (5)

We first check that (5) is a well-defined DI. Recall that, for an (autonomous) ODE to be well-defined, one sufficient condition is that its driving function be Lipschitz continuous. In particular, this guarantees the existence and uniqueness of a solution for any initial point. Similarly, a DI is well-defined when its set-valued driving function h is Marchaud, i.e., Lipschitz continuous in a set-valued sense (defined below). In general, solutions of a DI from a given starting point are not unique, but the above condition ensures existence.

For $x \in \mathbb{R}^d$, let $Z(x) := M \cup \{i : a_i^T(x-\mu) = 0\}$.

Lemma 1. *The function h defined in (4) is Marchaud, i.e.,*

1) $h(x)$ is convex and compact for all $x \in \mathbb{R}^d$;
2) $\exists K_h > 0$ such that, for all $x \in \mathbb{R}^d$, $\sup_{y\in h(x)} \|y\| \le K_h(1+\|x\|)$; and
3) h is upper semicontinuous or, equivalently, $\{(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^d : \theta \in h(x)\}$ is closed.

Hence, the DI in (5) is well-defined.

Proof: The first two conditions are easy. For h's upper semi-continuity, it suffices to check if (x_n) and (θ_n) are such that $x_n \to x$, $\theta_n \in h(x_n)\ \forall n$, and $\theta_n \to \theta$, then $\theta \in h(x)$.

For $i \in Z(x)^c$, $a_i^T(x-\mu)$ is either > 0 or < 0. This fact along with $x_n \to x$ then implies $\exists n_0 \ge 0$ such that, for $n \ge n_0$, we have $\text{sign}(a_i^T(x_n-\mu)) = \text{sign}(a_i^T(x-\mu))$ for all $i \in Z(x)^c$ and, hence, $Z(x)^c \subseteq Z(x_n)^c$. Thus, $h(x_n) \subseteq h(x)$ for all $n \ge n_0$, which implies $(\theta_n)_{n\ge n_0} \subseteq h(x)$. The desired result now follows since $h(x)$ is compact. □

We now show that μ is (5)'s unique GASE.

Proof of Statement 1, Theorem 1: It suffices to show that $V(x) = \frac{1}{2}\|x-\mu\|^2$ is a Lyapunov function [1] for the DI in (5) with respect to $\{\mu\}$. Clearly, $V(x) = 0$ if and only if $x = \mu$. Further, for any $x \ne \mu$ and $\theta \equiv \frac{1}{p}\sum_{i=1}^p a_i\lambda_i \in h(x)$,

$$\begin{aligned}
&\nabla V(x)^T\theta \\
&= \frac{1}{p}\sum_{i=1}^{p} \lambda_i a_i^T(x-\mu) \\
&= \frac{1}{p}\left[-\sum_{i\in M^c}|a_i^T(x-\mu)| + \sum_{i\in M}\lambda_i a_i^T(x-\mu)\right] \qquad (6)\\
&\le \frac{1}{p}\left[-\sum_{i\in M^c}|a_i^T(x-\mu)| + \sum_{i\in M}|a_i^T(x-\mu)|\right] \qquad (7)\\
&< 0, \qquad (8)
\end{aligned}$$

where (6) holds since $\lambda_i a_i^T(x-\mu) = -|a_i^T(x-\mu)|$ for $i \in M^c$, (7) is true because $r \le |r|$ for any $r \in \mathbb{R}$ and $|\lambda_i| \le 1$, while (8) follows from (3) since $|M| \le m$.

The claim now follows from [1, Proposition 3.25]. □

C. Almost Sure Boundedness of (x_n)

We use martingale and stopping time theory to show that (x_n) obtained using (1) is almost surely bounded.

Our proof needs a few intermediate results. In relation to (x_n) and (y_n) in (1), define the following. For $n \ge 0$, let

$$b_n = \frac{1}{p}\sum_{i\in M^c} a_i[\text{sign}(y_n(i) - a_i^T x_n) - \text{sign}(\mathbb{E}[Y](i) - a_i^T x_n)], \quad (9)$$

$$g(x_n, y_n) = \frac{1}{p}\sum_{i\in M^c} a_i\text{sign}(\mathbb{E}[Y](i) - a_i^T x_n) + \frac{1}{p}\sum_{i\in M} a_i\text{sign}(y_n(i) - a_i^T x_n),$$

and

$$M_{n+1} = a_{i_{n+1}}^T[\text{sign}(y_n(i_{n+1}) - a_{i_{n+1}}^T x_n] - g(x_n, y_n) - b_n. \quad (10)$$

In the above terms, the update rule in (1) can be written as

$$x_{n+1} = x_n + \alpha_n[g(x_n, y_n) + b_n + M_{n+1}]. \quad (11)$$

Note that $g(x_n, y_n) \in h(x_n)$. Therefore, one can view $g(x_n, y_n)$ as the update direction that is prescribed by (5), b_n as a perturbation that arises since, for $i \in M^c$, $y_n(i) \ne \mathbb{E}[Y(i)]$ a.s. for any finite n, and M_{n+1} as the noise.

Lemma 2. *The following statements are true.*

1) For $x \in \mathbb{R}^d$, let $\phi(x) = \frac{1}{p}\sum_{i\in M^c}|a_i^T x| - \frac{1}{p}\sum_{i\in M}|a_i^T x|$. Then there exists $\eta > 0$ such that $\phi(x) \ge \eta\|x\|\ \ \forall x$.
2) $|(x_n-\mu)^T b_n| \le \frac{2\sqrt{|M^c|}}{p}\|y_n - \mathbb{E}[Y]\|_{M^c}$.
3) $(x-\mu)^T\theta \le -\eta\|x-\mu\|$ for any $\theta \in h(x)$.

4) Let $C_M := \sup_{1\le i\le p}\|a_i\|$. Then, for any $n \ge 0$,

$$\|x_{n+1}-\mu\|^2 \le \|x_0-\mu\|^2 + \sum_{k=0}^{n}\alpha_k(x_k-\mu)^T M_{k+1} + \frac{2}{p}\sum_{k=0}^{n}\alpha_k\|y_k-\mathbb{E}[Y]\|_{M^c} + C_M^2\sum_{k=0}^{n}\alpha_k^2.$$

Proof: The first statement is trivially true for $x = 0$. Hence, suppose $x \neq 0$. It suffices to show that $\exists\,\eta > 0$ such that $\phi(x) \ge \eta$ for any x with unit norm. However, this holds since (a) ϕ is continuous and $\{x \in \mathbb{R}^d : \|x\| = 1\}$ is a compact set: thus, ϕ attains its minimum; and (b) $\phi(x) > 0$ for any $x \neq \mu$ on account of (3).

For the second statement, note that

$$|\text{sign}(r_1-r_0) - \text{sign}(r_2-r_0)| \le 2\delta_{|r_1-r_2|\ge|r_0-r_2|}$$

for any $r_0, r_1, r_2 \in \mathbb{R}$, where δ denotes the indicator function. Combining this with the fact that $\mathbb{E}[Y(i)] = a_i^T\mu$, for $i \in M^c$, gives

$$\begin{aligned}
&|(x_n-\mu)^T b_n| \\
&\le \frac{2}{p}\sum_{i\in M^c}|a_i^T(x_n-\mu)|\delta_{|y_n(i)-\mathbb{E}[Y(i)]|\ge|a_i^Tx_n-a_i^T\mu|} \\
&\le \frac{2}{p}\sum_{i\in M^c}|y_n(i)-\mathbb{E}[Y(i)]|\delta_{|y_n(i)-\mathbb{E}[Y(i)]|\ge|a_i^Tx_n-a_i^T\mu|} \\
&\le \frac{2}{p}\sum_{i\in M^c}|y_n(i)-\mathbb{E}Y(i)| \\
&\le \frac{2\sqrt{|M^c|}}{p}\|y_n-\mathbb{E}[Y]\|_{M^c},
\end{aligned}$$

as desired.

We now discuss the third statement. Let $\theta \in h(x)$ be arbitrary. Then,

$$\begin{aligned}
&(x-\mu)^T\theta \\
&\le \frac{1}{p}\left[-\sum_{i\in M^c}|a_i^T(x-\mu)| + \sum_{i\in M}|a_i^T(x-\mu)|\right] \\
&\le -\phi(x-\mu),
\end{aligned}$$

where the first relation follows as in (7), and the second relation holds from ϕ's definition. The claim now follows from our first statement above.

Finally, we derive the fourth statement. From (11),

$$\|x_{n+1}-\mu\|^2 = \|x_n-\mu\|^2 + \alpha_n^2\|g(x_n,y_n)+b_n+M_{n+1}\|^2 + 2\alpha_n(x_n-\mu)^T[g(x_n,y_n)+b_n+M_{n+1}].$$

Statement 3 along with the fact that $g(x_n,y_n) \in h(x_n)$ shows $(x_n-\mu)^T g(x_n,y_n) \le -\eta\|x_n-\mu\|$, while Statement 2 gives the bound on $(x_n-\mu)^T b_n$. Separately, $\|g(x_n,y_n)+b_n+M_{n+1}\| = \|a_{i_{n+1}}\| \le C_M$. It now follows that

$$\|x_{n+1}-\mu\|^2 \le \|x_n-\mu\|^2 - \alpha_n\eta\|x_n-\mu\| + \frac{2\sqrt{|M^c|}\alpha_n}{p}\|y_n-\mathbb{E}[Y]\|_{M^c} + \alpha_n(x_n-\mu)^T M_{n+1} + C_M^2\alpha_n^2.$$

The desired claim is now easy to see. □

Presuming Statement 2 in Theorem 1 holds, we are now ready to show that (x_n) is bounded almost surely,

Proposition 1. $\sup_{n\ge0}\|x_n\| < \infty$ *a.s.*

Proof: Let (γ_n) be as in Theorem 1. Fix an arbitrary integer $r \ge 1$, and let $C_r := \frac{2r\sqrt{|M^c|}}{p}\sum_{k=0}^{\infty}\alpha_k\gamma_k + C_M^2\sum_{k=0}^{\infty}\alpha_k^2 < \infty$, and $T(r)$ be the stopping time $\inf\left\{n \ge 0 : \frac{1}{\gamma_n}\|y_n-\mathbb{E}[Y]\|_{M^c} > r\right\}$. Next, for $n \ge 0$, let

$$S_n = \|x_0-\mu\|^2 + 2\sum_{k=0}^{n-1}\alpha_k(x_k-\mu)^T M_{k+1} + C_r.$$

Clearly, (S_n) and, hence, $(S_n^r) \equiv (S_{n\wedge T(r)})$ is a martingale.

Let $(x_n^r) \equiv (x_{n\wedge T(r)})$. Then Statement 4 of Lemma 2 shows $\|x_n^r-\mu\|^2 \le S_n^r\ \forall\, n \ge 0$. This implies (S_n^r) is a non-negative martingale and, hence, converges almost surely. Therefore, (x_n^r) is bounded almost surely.

Finally, note that

$$\begin{aligned}
E &:= \left\{\sup_{n\ge0}\|x_n\| = \infty\right\} \\
&\qquad \cap \left[\bigcup_{r=1}^{\infty}\left\{\sup_{n\ge0}\frac{\|y_n-\mathbb{E}[Y]\|_{M^c}}{\gamma_n} \le r\right\}\right] \\
&= \bigcup_{r=1}^{\infty}\left\{\sup_{n\ge0}\|x_n^r\| = \infty, \sup_{n\ge0}\frac{\|y_n-\mathbb{E}[Y]\|_{M^c}}{\gamma_n} \le r\right\} \qquad (12) \\
&\subseteq \bigcup_{r=1}^{\infty}\left\{\sup_{n\ge0}\|x_n^r\| = \infty\right\},
\end{aligned}$$

where (12) follows from the fact that $\sup_{n\ge0}\frac{\|y_n-\mathbb{E}[Y]\|_{M^c}}{\gamma_n} \le r$ implies $x_n = x_n^r$ for all n. Since (x_n^r) is almost surely bounded for any $r \ge 1$, we get $\mathbb{P}(E) = 0$. From Statement 2 in Theorem 1, we also have that

$$\mathbb{P}\left(\bigcup_{r=1}^{\infty}\left\{\sup_{n\ge0}\frac{\|y_n-\mathbb{E}[Y]\|_{M^c}}{\gamma_n} \le r\right\}\right) = 1.$$

The desired claim now follows since, for any events E_1 and E_2, $\mathbb{P}(E_1) = 1$ and $\mathbb{P}(E_2^c \cap E_1) = 0$ imply $\mathbb{P}(E_2) = 1$. □

D. Rest of the Proof

In this section, we discuss the proofs of Statements 2 and 3 of Theorem 1.

Statement 2 follows from [11, Theorem 1], which provides a law of iterated logarithm type result for generic stochastic approximation algorithms. That work assumes that the iterates almost surely converge, but this can be shown using the results in [3, Chapters 2 and 3], as discussed in Section III-A.

To prove Statement 3, we rely on [16, Theorem 4], which looks at convergence of generic two-timescale algorithms with set-valued limiting dynamics. Specifically, this latter result assumes (x_n)'s limiting DI has a global attractor (see A10 there), and states that, if ten other conditions (labelled A1 - A9 and A11 there) hold, then x_n converges

to this global attractor a.s. These ten conditions concern the behaviors of x_n and y_n's driving functions, stepsizes, and noise. Below we provide a brief commentary on why these assumptions hold for (1). The reader should note that the role of x_n and y_n is flipped in [16]: the changes in y_n eventually appear negligible compared to that of x_n. The analysis there also accounts for Markov noise, but it can ignored using the approach suggested in Remark 3 there. Finally, for all of (y_n)'s analysis below, we ignore the evolution at adversarial nodes: instead, we account for them directly in the definition of the DI in (5).

Assumptions A1 and A2 of [16] hold when the limiting DIs associated with x_n and y_n are Marchaud. For (1), this can be established like in the proof of our Lemma 1. Assumptions A3 and A4 concern Markov noise and, hence, trivially hold true in our case. Assumption A5 is on stepsizes and it holds in our case because we also assume those conditions. Assumption A8 there holds if the (x_n) and (y_n) iterates are bounded almost surely. Proposition 1 here proves it for (x_n), while, for (y_n), it follows easily from [3, Chapter 3, Theorem 7] due to its linear nature. Assumptions A6 and A7 hold if the contributions of the additive noise terms are eventually negligible. This can be established as in [3, Chapter 2, (2.19)], which holds in our case because our iterates are bounded a.s. and the noise growth rate condition of (2.13) trivially holds in our context. Assumptions A9 and A11 hold, if for each fixed x, the limiting DI for (y_n) has a unique GASE. As discussed in Section III-A, in our case, the dynamics of (y_n) is not influenced by the value of x and $\{\mathbb{E}[Y(i)] : i \in M^c\}$ is the global attractor for any x. Finally, Assumption A10 requires that (x_n)'s limiting DI has a unique global attractor. We established this in Statement 1 of our Theorem 1.

IV. Discussions and Future Directions

In this work, we developed a fully-asynchronous algorithm for mean estimation in the presence of adversaries. Thereafter, we developed a novel DI-based two-timescale analysis to rigorously show its a.s. convergence.

We now discuss some simple extensions of our work, where we can relax certain assumptions.

Non-zero kernel: The condition (3) fails for all matrices A with a non-zero kernel. Thus, Theorem 1 cannot be used for fat matrices or tall matrices with non full rank. However, we can obtain a similar result by relaxing condition (3) to hold only for points outside the kernel of A. Note that in this case, there are several $x \in \mathbb{R}^d$ such that $Ax = \mathbb{E}Y$. Under this modified assumption, it can be shown that the DI always converges to one such point. To see this, the function $\frac{1}{2}\|x-\mu\|_2^2$, with μ as solution of $Ax = \mathbb{E}Y$, would remain a Lyapunov function in this case. Applying a variant of LaSalle's invariance theorem would then give us that the DI converges to an invariant subset of $\{x : Ax = \mathbb{E}Y\}$.

Perturbed samples: Suppose that, instead of being provided samples of $Y(i) = a_i^T X$, we only have access to samples of form $Y(i) = a_i^T X + b(i)$, where $b(i)$ is some random or deterministic perturbation. The only condition imposed on $b(i)$ is that its magnitude remains bounded by some constant B for each i. We can extend the result in Theorem 1 to this setting using similar arguments as discussed in the previous case. However, the Lyapunov function would need to be redefined and may have discontinuous derivatives.

References

[1] Benaïm, M., Hofbauer, J., Sorin, S.: Stochastic approximations and differential inclusions. SIAM Journal on Control and Optimization **44**(1), 328–348 (2005)

[2] Bernstein, J., Zhao, J., Azizzadenesheli, K., Anandkumar, A.: signsgd with majority vote is communication efficient and fault tolerant. In: International Conference on Learning Representations (2018)

[3] Borkar, V.S.: Stochastic approximation: a dynamical systems viewpoint, vol. 48. Springer (2009)

[4] Damaskinos, G., Guerraoui, R., Patra, R., Taziki, M., et al.: Asynchronous byzantine machine learning (the case of sgd). In: International Conference on Machine Learning. pp. 1145–1154. PMLR (2018)

[5] Data, D., Diggavi, S.: Byzantine-resilient high-dimensional sgd with local iterations on heterogeneous data. In: International Conference on Machine Learning. pp. 2478–2488. PMLR (2021)

[6] Fang, M., Liu, J., Gong, N.Z., Bentley, E.S.: Aflguard: Byzantine-robust asynchronous federated learning. In: Proceedings of the 38th Annual Computer Security Applications Conference. pp. 632–646 (2022)

[7] Fawzi, H., Tabuada, P., Diggavi, S.: Secure state-estimation for dynamical systems under active adversaries. In: 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton). pp. 337–344. IEEE (2011)

[8] Gorbunov, E., Horváth, S., Richtárik, P., Gidel, G.: Variance reduction is an antidote to byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top. In: The Eleventh International Conference on Learning Representations (2022)

[9] Jin, R., Huang, Y., He, X., Dai, H., Wu, T.: Stochastic-sign sgd for federated learning with theoretical guarantees. arXiv preprint arXiv:2002.10940 (2020)

[10] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)

[11] Pelletier, M.: On the almost sure asymptotic behaviour of stochastic algorithms. Stochastic processes and their applications **78**(2), 217–244 (1998)

[12] Qi, J., Zhou, Q., Lei, L., Zheng, K.: Federated reinforcement learning: Techniques, applications, and open challenges. arXiv preprint arXiv:2108.11887 (2021)

[13] Thoppe, G.C., Kumar, B.: A law of iterated logarithm for multi-agent reinforcement learning. Advances in Neural Information Processing Systems **34**, 17927–17938 (2021)

[14] Wu, Z., Ling, Q., Chen, T., Giannakis, G.B.: Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. IEEE Transactions on Signal Processing **68**, 4583–4596 (2020)

[15] Xie, C., Koyejo, S., Gupta, I.: Zeno++: Robust fully asynchronous sgd. In: International Conference on Machine Learning. pp. 10495–10503. PMLR (2020)

[16] Yaji, V.G., Bhatnagar, S.: Stochastic recursive inclusions in two timescales with nonadditive iterate-dependent markov noise. Mathematics of Operations Research **45**(4), 1405–1444 (2020)

[17] Yang, Y.R., Li, W.J.: Basgd: Buffered asynchronous sgd for byzantine learning. In: International Conference on Machine Learning. pp. 11751–11761. PMLR (2021)

[18] Yin, D., Chen, Y., Kannan, R., Bartlett, P.: Byzantine-robust distributed learning: Towards optimal statistical rates. In: International Conference on Machine Learning. pp. 5650–5659. PMLR (2018)

[19] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y.: A survey on federated learning. Knowledge-Based Systems **216**, 106775 (2021)