2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Distributed robust statistical learning: Byzantine mirror descent

Dongsheng Ding, Xiaohan Wei, and Mihailo R. Jovanović

***Abstract*— We consider the distributed statistical learning problem in a high-dimensional adversarial scenario. At each iteration, m worker machines compute stochastic gradients and send them to a master machine. However, an α-fraction of m worker machines, called Byzantine machines, may act adversarially and send faulty gradients. To guard against faulty information sharing, we develop a distributed robust learning algorithm based on mirror descent. This algorithm is provably robust against Byzantine machines whenever $\alpha \in [0, 1/2)$. For smooth convex functions, we show that running the proposed algorithm for T iterations achieves a statistical error bound $\tilde{O}(1/\sqrt{mT} + \alpha/\sqrt{T})$. This result holds for a large class of normed spaces and it matches the known statistical error bound for Byzantine stochastic gradient in the Euclidean space setting. A key feature of the algorithm is that the dimension dependence of the bound scales with the dual norm of the gradient; in particular, for probability simplex, we show that it depends logarithmically on the problem dimension d. Such a weak dependence is desirable in high-dimensional statistical learning and it has been known to hold for the classical mirror descent but it appears to be new for the Byzantine gradient scenario.**

I. Introduction

Modern statistical learning usually requires algorithms to learn a prediction model in a high-dimensional space from massive amounts of data. Since the training data are often spread across a large number of local worker machines, efficient algorithms should make local machines collaboratively learn a shared model, while maintaining distributed computation and storage. This is one of the main topics of Federated Learning [1]–[3]. However, such a distributed learning brings new challenges not faced by classical centralized learning including communication/storage failure and adversarial attacks from malicious worker machines. Thus, robustness against arbitrary unpredictable corruptions becomes crucial to improving the learning efficiency, a topic that has been investigated in a series of recent papers [4]–[7].

A well-known model that accounts for abnormal worker machines is the Byzantine failure model [8]. In this model, faulty Byzantine machines can send arbitrary messages to the master machine. We assume that these Byzantine machines have a complete knowledge of the system and learning algorithms and that they can collude with each other. One recent surging research interest is to investigate the robustness of different optimization and learning algorithms against Byzantine failures with provable statistical or computational guarantees. Some recent works along this direction include batch gradient descent [4], [5], [7], stochastic gradient descent (SGD) [1], [6], [9], and alternating direction method of multipliers (ADMM) [10].

Financial support from the National Science Foundation under Award ECCS-1809833 and the Air Force Office of Scientific Research under Award FA9550-16-1-0009 is gratefully acknowledged.

D. Ding, X. Wei, and M. R. Jovanović are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089. E-mails: dongshed@usc.edu, xiaohanw@usc.edu, mihailo@usc.edu

A. System model

We assume that the training data z are sampled from some unknown distribution $\mathcal{D}$ on the sample space $\mathcal{Z}$. Let $f(w; z)$ be a corresponding loss function where $w \in \mathcal{W} \subseteq \mathbb{R}^d$ and $\mathcal{W}$ is the parameter space. The goal of the statistical learning is to learn a model $w^\star$ defined as the minimizer of the population loss function $F(w) := \mathbb{E}_{z\sim\mathcal{D}}\,[f(w; z)]$,

$$w^\star \;=\; \operatorname*{argmin}_{w\in\mathcal{W}} \; F(w). \tag{1}$$

The computational model consists of one master machine and m worker machines. At iteration t, the ith worker machine receives a data point z_t^i which is sampled independently from the distribution $\mathcal{D}$. The empirical risk function for the population loss $F(w)$ at iteration t becomes $F_t(w) = \frac{1}{m}\sum_{i=1}^m f(w; z_t^i)$. Among m worker machines, we assume that an α-fraction of them are Byzantine, meaning that they can send arbitrary messages to the master machine synchronously. At iteration t, each worker machine receives the current iterate w_t, utilizes local data point z_t^i to compute the associated gradient and returns it to the master machine. There are two possibilities: (1) a non-Byzantine machine returns $\nabla f(w_t, z_t^i)$ where $z_t^i \sim \mathcal{D}$; (2) a Byzantine machine adversarially returns arbitrary vector. After receiving information from all worker machines, the master machine aggregates them for the optimization routine and generates the next iterate w_{t+1}, and then broadcasts w_{t+1} to all worker machines. After T iterations, the master machine should obtain an approximate solution to the optimal solution in (1).

In this paper, we study such distributed computational models that can be used for general distributed statistical learning problems. We propose a new learning algorithm that includes mirror descent algorithm [13] as the optimization routine and study its convergence properties. Benefited from mirror descent, we are enabled to adapt our algorithm to the problem geometry so that its convergence has a weak dependence on the problem dimension d.

B. Related work

Among various works on distributed statistical learning under Byzantine failures [1], [4]–[7], [9], the closely-related work is [6]. The authors present a robust variant of SGD where the gradient is calculated through a median aggregation of gradients from worker machines on the fly. Using the property of median and bounded good gradients, the authors achieve a convergence rate $O(V\,(1/\sqrt{mT}+\alpha/\sqrt{T}))$,

where V is the Euclidean norm bound of the stochastic gradient which usually scales with $\sqrt{d}$. A more challenging scenario shown in [7] is that the good gradients obtained by earnest worker machines are unbounded heavy-tailed. Robust aggregation methods such as adaptive truncated mean of sampled gradients or entry-wise median-of-mean are used in [7] to design robust batch gradient algorithms. The achieved statistical rate is $O(\text{poly}(d)\,(1/\sqrt{mT}+\alpha/\sqrt{T}))$, which only works well in low-dimensional scenarios due to the polynomial factor $\text{poly}(d)$.

In the earlier results of [4], [5], other types of aggregation methods such as the geometric median-of-mean are used to estimate the population gradient in robust batch gradient algorithms. However, compared to the aforementioned work, the dependence of their statistical rates on α is suboptimal, e.g., $O(\text{poly}(d)\,\sqrt{\alpha/T})$. In [14], [15], different heuristic mechanisms are proposed for defending against an arbitrary number of Byzantine machines in batch gradient methods.

C. Our contributions

We build on mirror descent [13] and introduce a simple distributed robust learning algorithm against Byzantine machines for $\alpha \in [0, 1/2)$. This method enables us to adapt the algorithm to the problem geometry so that the error bound has a mild dimension dependence. In any normed space $(\mathbb{R}^d, \|\cdot\|)$, when the function is convex and smooth, running the proposed algorithm for T steps achieves the statistical error bound $\tilde{O}(V\left(1/\sqrt{mT}+\alpha/\sqrt{T}\right))$ where $\tilde{O}(\cdot)$ omits some logarithm factors and V is the dual norm bound of gradients. Particularly, when the space is a probability simplex with ℓ_1-norm, the dual norm bound is dimension-free. This is a desirable feature in high-dimensional statistical learning for it implies that the statistical error bound scales logarithmically to the dimension d.

D. Organization of the paper

In Section II, we prepare some preliminaries. In Section III, we introduce the Byzantine mirror descent algorithm. We show the convergence analysis in Section IV. We discuss the probability simplex case in Section V and we close the paper in Section VI.

II. Preliminaries

Let $\|\cdot\|$ be any norm in $\mathbb{R}^d$. The dual norm $\|\cdot\|_*$ is defined as $\|g\|_* = \sup_{\|x\|\le 1} g^T x$. Let $\mathcal{W} \in \mathbb{R}^d$ be a compact, convex set with diameter W.

Definition 1: Let $f : \mathcal{W} \to \mathbb{R}$ be a differentiable function.

(1) f is σ-strongly convex w.r.t. $\|\cdot\|$, if $f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle + \frac{\sigma}{2}\|y-x\|^2$, $\forall x, y \in \mathcal{W}$;
(2) f is L-smooth w.r.t. $\|\cdot\|$, if $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x-y\|$, $\forall x, y \in \mathcal{W}$;
(3) f is G-Lipschitz (continuous) w.r.t. $\|\cdot\|$, if $\|\nabla f(x)\|_* \le G$, $\forall x \in \mathcal{W}$.

Definition 2: Let $\mathcal{X} \subset \mathbb{R}^d$ be an open convex set such that $\mathcal{W} \subset \bar{\mathcal{X}}$ and $\mathcal{X} \cap \mathcal{W} \neq \emptyset$.

(1) $\Phi : \mathcal{X} \to \mathbb{R}$ is a mirror map, if it satisfies: (i) Φ is 1-strongly convex w.r.t. $\|\cdot\|$, and differentiable; (ii) $\nabla\Phi(\mathcal{X}) = \mathbb{R}^d$; (iii) $\lim_{x\to\partial\mathcal{X}} \|\nabla\Phi(x)\| = +\infty$;
(2) The Bregman divergence associated with Φ is

$$D_\Phi(x, y) \;=\; \Phi(x) \;-\; \Phi(y) \;-\; \nabla\Phi(y)^T(x-y).$$

Examples of such mirror maps and the associated Bregman divergences are given as follows. The simplest example is when $\mathcal{X} = \mathbb{R}^d$ and $\Phi(x) = \frac{1}{2}\|x\|_2^2$. The function Φ is a mirror map that is 1-strongly convex w.r.t. $\|\cdot\|_2$, and the associated Bregman divergence is $D_\Phi(x,y) = \frac{1}{2}\|x-y\|_2^2$; A more interesting example is when $\mathcal{X} = \{x \in \mathbb{R}^d : x(i) > 0\}$ and $\Phi(x) = \sum_{i=1}^d x(i)\log x(i)$ (i.e., negative entropy). By Pinsker's inequality, the mirror map Φ is 1-strongly convex w.r.t. $\|\cdot\|_1$ on the probability simplex $\{x \in \mathbb{R}^d : x(i) \ge 0, \sum_{i=1}^d x(i) = 1\}$, and the associated Bregman divergence is given by $D_\Phi(x,y) = \sum_{i=1}^d x(i)\log\frac{x(i)}{y(i)}$, also known as Kullback-Leibler divergence.

Definition 3: The Banach space $(\mathbb{R}^d, \|\cdot\|)$ is 2-smooth, if the modulus of smoothness

$$\rho(\tau) \;=\; \sup_{\|x\|=1, \|y\|=\tau} \left\{\frac{1}{2}(\|x+y\| + \|x-y\|) - 1\right\}$$

behaves as $\rho(\tau) \le C\tau^2$ for some constant C.

Remark 1: This definition is unintuitive from the first sight, however, one can show that for any norm $\|\cdot\|_p, 1 < p < \infty$, the corresponding Banach space $(\mathbb{R}^d, \|\cdot\|_p)$ is 2-smooth. Furthermore, when $(\mathbb{R}^d, \|\cdot\|_*)$ is a 2-smooth Banach space, there always exists a $\Phi(\cdot)$ which is 1-strongly convex w.r.t. the primal norm $\|\cdot\|$ (see [16]). Such a condition is crucial since our algorithm is of proximal gradient type and it ensures the existence of a strongly convex regularizer. But when $\|\cdot\|_* = \|\cdot\|_\infty$, $(\mathbb{R}^d, \|\cdot\|_\infty)$ is not 2-smooth. We will address this additional case separately in Section V.

For succinctness, we denote $D = D_\Phi$. We take a variant of the mirror descent scheme [13] based on a mirror map Φ as follows. The algorithm maintains iterates $\{w_t, \xi_t\}_{t=1}^\infty$ within $\mathcal{W}\times\mathbb{R}^d$. At iteration t, it computes the gradient $\xi_t = \nabla F(w_t)$ and performs the updates,

$$w_{t+1} \;=\; \operatorname*{argmin}_{w\in\mathcal{W}} \; \eta\,\langle \xi_t, w - w_t\rangle \;+\; D(w, w_t) \qquad (2)$$

where $w_1 \in \mathcal{X}$ and $\eta > 0$ is the constant stepsize. We state an important property of iterates $\{w_t, \xi_t\}_{t=1}^\infty$ in Lemma 1; see [17, Lemma 14] for a general version.

Lemma 1: Let Φ be a mirror map. Let $\{w_t, \xi_t\}_{t=1}^\infty$ be generated by (2). Then, we have $\langle \xi_t, w_t - w\rangle \le \langle \xi_t, w_t - w_{t+1}\rangle - D(w_{t+1}, w_t)/\eta + (D(w, w_t) - D(w, w_{t+1}))/\eta$.

In the Byzantine setup, since Byzantine machines can send wrong gradients, it is difficult to obtain an unbiased estimate of the population gradient. The following concentration result will be useful for us to deal with this issue.

Lemma 2: [18, Theorem 3] Let $(\mathbb{R}^d, \|\cdot\|)$ be a 2-smooth Banach space. Assume $X_1, \cdots, X_T$ be a martingale difference sequence in $\mathbb{R}^d$ with $\|X_t\| \le M$. For any $\delta > 0$, it holds

$$P\left(\left\|\sum_{k=1}^T X_k\right\| \;\ge\; 2\sqrt{2T}\left(R + 2\sqrt{2\log\tfrac{\sqrt{2}}{\delta}}\right)M\right) \;\le\; \delta$$

where $R^2 := \sup_{x,y\in\mathcal{W}} D_\Phi(x,y)$.

When gradient estimates are not far away from the population gradient, the mirror descent algorithm is expected to return an approximate solution to the optimization problem (1) after T iterations. Let $w_1 \in \mathcal{W}$ be an initial point. Let α-fraction of m worker machines be Byzantine where $\alpha \in [0, 1/2)$. Let $\Omega \subseteq [m]$ be an unknown set of good or non-Byzantine machines. At iteration t, all worker machines receive the same iterate w_t from the master machine. Then they utilize their own sampled data points z_t^i to compute local gradients $\nabla_t^i := \nabla f(w_t, z_t^i)$ and send them back, if they satisfy Assumption 1.

Assumption 1: Let $\nabla_t := \nabla F(w_t)$. At each iteration t, there exists $V > 0$ such that worker machine $i \in \Omega$ satisfies

$$\|\nabla_t^i - \nabla_t\|_* \;\le\; V.$$

The rational behind Assumption 1 is that the gradients of all good machines behave mildly, not far away from the population gradient ∇_t which is unknown. Although Byzantine machines may also satisfy this assumption, their effect on the solution quality is negligible as we will show in the next section.

III. Byzantine mirror descent

We propose a robust variant of mirror descent to the Byzantine setting. We describe this algorithm in Algorithm 1 as Byzantine mirror descent. Let $[T] = \{1, 2, \cdots, T\}$ be a set of iteration indices and $[m] = \{1, 2, \cdots, m\}$ be a set of worker machines. At each iteration $t \in [T]$, the algorithm maintains iterate $(w_t, \xi_t) \in \mathcal{W} \times \mathbb{R}^d$ and a set of "good" worker machine in $\Omega_t \subset [m]$. The gradient estimate ξ_t is computed using the averaged gradients in Ω_t,

$$\xi_t \;=\; \frac{1}{m}\sum_{i\in\Omega_t} \nabla_t^i \tag{3}$$

where the set Ω_t is maintained by tracking two running sequences for each worker machine $i \in [m]$,

$$A_t^i \;=\; \sum_{k=1}^{t} \langle \nabla_k^i, w_k - w_1\rangle, \;\; B_t^i \;=\; \sum_{k=1}^{t} \nabla_k^i.$$

By the martingale concentration, for each good worker machine i, B_T^i will concentrate at the population sum $B_T^\star = \sum_{k=1}^T \nabla_k$ with a maximum deviation of $\sqrt{T}$. If some worker machines are too far away from the mean, we can mark them as Byzantine machines. However, some Byzantine machines may hide themselves with small deviation in terms of B_t^i. To identify such Byzantine machines, we further consider the martingale concentration for A_t^i. If A_t^i is too far away from the population sum $A_t^\star = \sum_{k=1}^t \langle \nabla_k, w_k - w_1\rangle$, we mark them as Byzantine machines. By removing these Byzantine machines, we put the rest in the set Ω_t as an estimated set of good machines at iteration t.

We start Ω_1 with $[m]$. The set Ω_t contains all machines $i \in \Omega_{t-1}$, whose A_t^i, B_t^i and ∇_t^i are close to their medians $A_t^{\text{med}}, B_t^{\text{med}}$, and ∇_t^{med} with thresholds I_A, I_B and $4V$ respectively. We will show that if we choose these thresholds appropriately, then Ω_t contains all good machines, i.e., $\Omega \subset \Omega_t$. Thus, (3) works as an estimation of the population gradient at iteration t in terms of the boundedness of errors,

$$\begin{aligned} E_1 &= \sum_{t=1}^{T}\sum_{i\in\Omega_t} \langle \nabla_t^i - \nabla_t, w_t - w^*\rangle \\ E_2 &= \frac{1}{T}\sum_{t=1}^{T} \left\| \frac{1}{m}\sum_{i\in\Omega_t} (\nabla_t^i - \nabla_t) \right\|_* \end{aligned} \tag{4}$$

where E_1 is the accumulative optimality bias, and E_2 measures the variance of estimating the population gradient.

Algorithm 1 Byzantine mirror descent

Input: Learning rate $\eta > 0$, starting point $w_1 \in \mathcal{X}$, diameters $W, R > 0$, number of iterations T, constant thresholds $I_A = 4WV\Delta\sqrt{2T}$ and $I_B = 4V\Delta\sqrt{2T}$, and $\Delta = R + 2\sqrt{2\log\frac{8\sqrt{2}mT}{\delta}}$.

$\Omega_1 \leftarrow [m]$;
for all $t \leftarrow 1$ **to** T **do**
  for all $i \leftarrow 1$ **to** m **do**
    receive $\nabla_t^i \in \mathbb{R}^d$ from worker machine $i \in [m]$
    $$A_t^i \;\leftarrow\; \sum_{k=1}^{t}\langle \nabla_k^i, w_k - w_1\rangle, \;\; B_t^i \;\leftarrow\; \sum_{k=1}^{t} \nabla_k^i$$
  end for
  $A_t^{\text{med}} := \text{median}(A_t^1, \cdots, A_t^m)$
  $B_t^{\text{med}} \leftarrow B_t^i$ where $i \in [m]$ is any machine s.t.
  $$|\{j \in [m] : \|B_t^i - B_t^j\|_* \le I_B\}| \;>\; \frac{m}{2}$$
  $\nabla_t^{\text{med}} \leftarrow \nabla_t^i$ where $i \in [m]$ is any machine s.t.
  $$|\{j \in [m] : \|\nabla_t^i - \nabla_t^j\|_* \le 2V\}| \;>\; \frac{m}{2}$$
  $\Omega_t \leftarrow \{i \in \Omega_{t-1} : |A_t^i - A_t^{\text{med}}| \le I_A \wedge \|B_t^i - B_t^{\text{med}}\|_* \le I_B \wedge \|\nabla_t^i - \nabla_t^{\text{med}}\|_* \le 4V\}$
  $$\xi_t = \frac{1}{m}\sum_{i\in\Omega_t} \nabla_t^i$$
  $w_{t+1} = \text{argmin}_{w\in\mathcal{W}} \;\; D(w, w_t) + \eta\,\langle \xi_k, w - w_t\rangle$
end for

IV. Convergence analysis

In this section, we show upper bounds for errors E_1 and E_2, and provide statistical error bound analysis when the objective function is convex and smooth.

A. Three events

Denote $\Delta := R + 2\sqrt{2\log\frac{8\sqrt{2}mT}{\delta}}$. We use Lemma 2 to identify probability bounds for martingale sequences A_t^i, B_t^i and ∇_t^i over iteration t. We describe them as follows.

Proposition 3: Let

$$A_t^\star \;=\; \sum_{k=1}^{t} \langle \nabla_k, w_k - w_1\rangle,\; A_t^{\text{med}} \;=\; \text{median}(A_t^1, \cdots, A_t^m).$$

With probability at least $1-\delta/4$, we have

(1) for all $i \in \Omega$ and $t \in [T]$, $|A_t^i - A_t^\star| \le 2WV\Delta\sqrt{2t}$;
(2) for all $i \in \Omega$ and $t \in [T]$, $|A_t^i - A_t^{\text{med}}| \le 4WV\Delta\sqrt{2t}$ and $|A_t^\star - A_t^{\text{med}}| \le 2WV\Delta\sqrt{2t}$;
(3) $|\sum_{i\in\Omega}(A_t^i - A_T^\star)| \le 2WV\Delta\sqrt{2Tm}$.

We denote this event by Event_A.

Proof: See Appendix A. ∎

Proposition 4: Let $B_t^\star = \sum_{k=1}^t \nabla_k$, and $B_t^{\text{med}} = B_t^i$ where i is any machine in $[m]$ such that at least half of machines $j \in [m]$ satisfies $\|B_t^j - B_t^i\|_* \le 4V\Delta\sqrt{2t}$.

With probability at least $1-\delta/4$, we have

(1) for all $i \in \Omega$ and $t \in [T]$, $\|B_t^i - B_t^\star\|_* \le 2V\Delta\sqrt{2t}$;
(2) for all $t \in [T]$, each $i \in \Omega$ is a valid choice for $B_t^{\text{med}} = B_t^i$;
(3) for all $i \in \Omega$ and $t \in [T]$, $\|B_t^i - B_t^{\text{med}}\|_* \le 4V\Delta\sqrt{2t}$ and $\|B_t^\star - B_t^{\text{med}}\|_* \le 6V\Delta\sqrt{2t}$;
(4) $\|\sum_{i\in\Omega}(B_t^i - B_T^\star)\|_* \le 2V\Delta\sqrt{2Tm}$.

We denote this event by Event_B.

Proof: See Appendix B. ∎

Proposition 5: With probability at least $1-\delta/4$, we have

$$\left\|\frac{1}{m}\sum_{i\in\Omega}(\nabla_t^i - \nabla_t)\right\|_*^2 \le \frac{8V^2\Delta^2}{m}$$

for all $t \in [T]$. We denote this event by Event_C.

Proof: See Appendix C. ∎

B. Key upper bounds

Denote $I_A := 4WV\Delta\sqrt{2T}$ and $I_B := 4V\Delta\sqrt{2T}$. We compute $\nabla_t^{\text{med}} = \nabla_t^i$ where i is any machine in $[m]$ such that at least half of machines $j \in [m]$ satisfies $\|\nabla_t^i - \nabla_t^j\|_* \le 2V$.

Proposition 6: For all $t \in [T]$, each $i \in \Omega$ is a valid choice for $\nabla_t^{\text{med}} = \nabla_t^i$, and $\|\nabla_t^{\text{med}} - \nabla_t\|_* \le 3V$.

Proof: See Appendix D. ∎

Proposition 7: If Event_A and Event_B hold, then $\Omega \subseteq \Omega_t$ for $t \in [T]$.

Proof: See Appendix E. ∎

We show upper bounds for E_1 and E_2 in Lemmas 8 and 9.

Lemma 8: If Event_A and Event_B hold, then

$$|E_1| \le 4WV\Delta\sqrt{2Tm} + 16\alpha mWV\Delta\sqrt{2T}.$$

Proof: See Appendix F. ∎

Lemma 9: If $\mathsf{Event}_A, \mathsf{Event}_B$ and Event_C hold, we have

$$E_2 \le \frac{16V^2\Delta^2}{m} + 32\alpha^2V^2.$$

Proof: See Appendix G. ∎

C. Statistical error bound

Theorem 10: Let $(\mathbb{R}^d, \|\cdot\|_*)$ be a 2-smooth Banach space. Let the objective function $F(w)$ be G-Lipschitz and L-smooth, and Assumption 1 hold. Suppose $\eta \le \frac{1}{2L}$, $I_A = 4WV\Delta\sqrt{2T}$, and $I_B = 4V\Delta\sqrt{2T}$. Then, with probability at least $1-\delta$, we have

$$F(\bar{w}) - F(w^*) \le \frac{2R^2}{\eta T} + \frac{8WV\Delta(\sqrt{2mT} + 4\alpha m\sqrt{2T})}{mT} + \eta\left(\frac{32V^2\Delta^2}{m} + 64\alpha^2V^2\right)$$

where $\bar{w} := \frac{1}{T}\sum_{t=1}^T w_{t+1}$.

Proof: First we have the decomposition $\sum_{t=1}^T\langle\xi_t, w_t - w^\star\rangle = \frac{E_1}{m} + \frac{1}{m}\sum_{t=1}^T\sum_{i\in\Omega_t}\langle\nabla_t, w_t - w^*\rangle$. By the convexity and the L-smoothness of F, we know $\langle\nabla_t, w_t - w^*\rangle \ge F(w_t) - F(w^\star) \ge F(w_{t+1}) - \langle\nabla_t, w_{t+1} - w_t\rangle - \frac{L}{2}\|w_{t+1} - w_t\|^2$. Thus, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\langle\xi_t, w_t - w^\star\rangle \\
&\ge \frac{E_1}{mT} + \frac{1}{mT}\sum_{t=1}^T\sum_{i\in\Omega_t}(F(w_{t+1}) - F(w^*)) \\
&\quad - \frac{1}{mT}\sum_{t=1}^T\sum_{i\in\Omega_t}(\langle\nabla_t, w_{t+1} - w_t\rangle + \frac{L}{2}\|w_{t+1} - w_t\|^2).
\end{aligned} \tag{5}$$

On the other hand, we substitute $w = w^\star$ and $\xi_t = \frac{1}{m}\sum_{i\in\Omega_t}\nabla_t^i$ into the inequality in Lemma 1. The time average of this inequality over t becomes

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\langle\xi_t, w_t - w^*\rangle \\
&\le \frac{1}{T}\sum_{t=1}^T\left\langle\frac{1}{m}\sum_{i\in\Omega_t}\nabla_t^i, w_t - w_{t+1}\right\rangle \\
&\quad - \frac{1}{\eta T}\sum_{t=1}^T D(w_{t+1}, w_t) + \frac{D(w^\star, w_1) - D(w^\star, w_{T+1})}{\eta T}.
\end{aligned}$$

Due to the 1-strongly convexity of Φ in the primal norm $\|\cdot\|$, we have $D(w_{t+1}, w_t) \ge \|w_{t+1} - w_t\|^2/2$. Combing the above inequality with (5) yields

$$\begin{aligned}
&\frac{1}{mT}\sum_{t=1}^T\sum_{i\in\Omega_t}(F(w_{t+1}) - F(w^*)) \\
&\le \frac{1}{T}\sum_{t=1}^T\left\langle\frac{1}{m}\sum_{i\in\Omega_t}(\nabla_t - \nabla_t^i), w_{t+1} - w_t\right\rangle \\
&\quad - \frac{1}{T}\sum_{t=1}^T\left(\frac{1}{2\eta} - \frac{L}{2}\right)\|w_t - w_{t+1}\|^2 + \frac{D(w^\star, w_1)}{\eta T} - \frac{E_1}{mT} \\
&\le \frac{1}{T}\sum_{t=1}^T\left\|\frac{1}{m}\sum_{i\in\Omega_t}(\nabla_t - \nabla_t^i)\right\|_*\|w_{t+1} - w_t\| \\
&\quad - \frac{1}{4\eta T}\sum_{t=1}^T\|w_t - w_{t+1}\|^2 + \frac{R^2}{\eta T} - \frac{E_1}{mT} \\
&\le \frac{R^2}{\eta T} - \frac{E_1}{mT} + \eta E_2
\end{aligned}$$

where we use $\frac{1}{2\eta} - \frac{L}{2} \ge \frac{1}{4\eta}$ and $D(w^\star, w_1) \le R^2$ for the second inequality; the third inequality is due to $2\|a\|_*\|b\| \le \|a\|_*^2 + \|b\|^2$.

Finally, we collect the upper bounds on $|E_1|$ and E_2 and use the convexity of F to complete the proof.

$$\begin{aligned}\frac{1}{mT}\sum_{t=1}^{T}\sum_{i\in\Omega_t}(F(w_{t+1})-F(w^*)) &\geq \frac{1}{2T}\sum_{t=1}^{T}(F(w_{t+1})-F(w^*)) \\ &\geq \frac{1}{2}(F(\bar{w})-F(w^*)).\end{aligned}$$

■

Remark 2: When the gradients ∇_t^i are exactly the population gradient ∇_t, the dual norm bound of gradients becomes $V=0$. The first term $O(R^2/T)$ matches the standard rate for the mirror descent [13] when the objective function is convex and smooth. The error bound in Theorem 10 can be tuned to be optimal by choosing η carefully. There are two cases. (i) when $\alpha m \geq \sqrt{m}$, we choose $\eta = \min(1/(\alpha V\sqrt{T}), 1/(2L))$; and (ii) when $\alpha m < \sqrt{m}$, we choose $\eta = \min(\sqrt{m/T}/V, 1/(2L))$. Thus, we have

$$F(\bar{w})-F(w^*) \;\leq\; \tilde{O}\left(V\left(\frac{R^2}{T}+\frac{1}{\sqrt{mT}}+\frac{\alpha}{\sqrt{T}}\right)\right)$$

When there are no Byzantine worker machines, i.e. $\alpha=0$, the first two terms are similar to the rate for the classical mini-batch SGD [19]. The last term accounts for Byzantine worker machines if $\alpha\neq 0$.

V. Probability simplex case

When the dual norm $\|\cdot\|_*=\|\cdot\|_\infty$, the associated Banach space $(\mathbb{R}^d,\|\cdot\|_\infty)$ is not 2-smooth and Lemma 2 does not apply. Even though, we can still use the proposed Byzantine mirror descent, but with some changes in the algorithm setting.

We set the mirror map Φ be the negative entropy, and the constraint set $\mathcal{W}$ be the probability simplex. Thus, we replace the type of norm in all previous bounds for gradients by ∞-norm. First, the upper bound in Assumption 1 becomes entry-wise bounded. Second, not Lemma 2, we use the Azuma's inequality [20, Theorem 17] to establish similar concentration results in Section IV-A and Section IV-B. We omit them here due to the limit of space. We just mention some changes in the algorithm. Since Φ is 1-strongly convex w.r.t. $\|\cdot\|_1$ on $\mathcal{W}$, the diameter of $\mathcal{W}$ is $W=1$. We choose $I_A=I_B=4V\Delta\sqrt{T}$ and $\Delta=\sqrt{\log\frac{16mT}{\delta}}$. Upper bounds for E_1 and E_2 are given as follows, which are similar to [6, Lemma 3.6] and [6, Lemma 3.7].

$$\begin{aligned}|E_1| &\leq 4V\Delta\sqrt{Tm}+16\alpha mV\Delta\sqrt{T} \\ E_2 &\leq \frac{4V^2\Delta^2}{m}+32\alpha^2V^2.\end{aligned} \tag{6}$$

Theorem 11: Let $(\mathbb{R}^d,\|\cdot\|_*)$ be equipped with the norm $\|\cdot\|_*=\|\cdot\|_\infty$. Let the objective function $F(w)$ be G-Lipschitz and L-smooth, and Assumption 1 hold. Let the mirror map $\Phi(w)$ be the negative entropy. Suppose $\eta\leq\frac{1}{2L}$ and $I_A=I_B=4V\Delta\sqrt{T}$. Then, with probability at least $1-\delta$, we have

$$F(\bar{w})-F(w^*)\leq\frac{2(\log d)^2}{\eta T}+\frac{8V\Delta(\sqrt{mT}+4\alpha m\sqrt{T})}{mT}+\eta\left(\frac{4V^2\Delta^2}{m}+32\alpha^2V^2\right)$$

where $\bar{w}=\frac{1}{T}\sum_{t=1}^{T}w_{t+1}$ and d is the problem dimension.

Proof: The proof is similar to proving Theorem 10. We use the error bounds in (6) and $R^2=\log d$. ■

Remark 3: If we choose η optimally as in Remark 2, then

$$F(\bar{w})-F(w^*) \;\leq\; \tilde{O}\left(\frac{\log d}{T}+\frac{1}{\sqrt{mT}}+\frac{\alpha}{\sqrt{T}}\right)$$

The upper bounds G and V in Definition 1 and Assumption 1 become entry-wise constants when $\|\cdot\|_*=\|\cdot\|_\infty$. The only left dimension factor is $\log d$. Therefore, the above error bound is almost dimension-free, meaning that it scales as $O(\log d)$.

VI. Conclusion

In this paper, we propose a new variant of the mirror descent for robust distributed learning problems in Byzantine setting. We show the robustness of this algorithm to Byzantine failures whenever the fraction of Byzantine machines $\alpha\in[0,1/2)$. In T iterations, we show the statistical error bound $O(1/\sqrt{mT}+\alpha/\sqrt{T})$ when the objective function is convex and smooth. In the probability simplex case, the statistical error bound enjoys the almost dimension-free property, i.e., it scales as $O(\log d)$, where d is the problem dimension.

Appendix

A. Proof of Proposition 3

(1) Note that $\mathbb{E}[\langle\nabla_k^i-\nabla_k, w_k-w_1\rangle]=0$ and $|\langle\nabla_k^i-\nabla_k, w_k-w_1\rangle|\leq\|\nabla_k^i-\nabla_k\|_*\cdot\|w_k-w_1\|\leq VW$. Therefore, we can use Lemma 2 by choosing $X_k=\langle\nabla_k^i-\nabla_k, x_k-x_1\rangle$. First we can obtain that $|A_t^i-A_t^\star|\leq 2WV\Delta\sqrt{2t}$ holds with probability at least $1-\frac{\delta}{8mT}$. Then we take an union bound over $i\in\Omega$ and $t\in[T]$.

(2) Since $|\Omega|>m/2$, a special case of (1) is that $|A_t^\star-A_t^{\text{med}}|\leq 2WV\Delta\sqrt{2t}$. By the triangle inequality, we have $|A_t^i-A_t^{\text{med}}|\leq 4WV\Delta\sqrt{2t}$.

(3) If we apply Lemma 2 on $\{X_1,X_2,\cdots,X_{T|\Omega|}\}=\{\langle\nabla_k^i-\nabla_k, w_k-w_1\rangle\}_{k\in[T],i\in\Omega}$, we can have a similar bound as (1).

B. Proof of Proposition 4

(1) Note that $\mathbb{E}[\nabla_i^k]=\nabla_k$ and $\|\nabla_i^k-\nabla_k\|_*\leq V$. Similarly, we apply Lemma 2 with $X_k=\nabla_i^k-\nabla_k$ and then take a union bound over $i\in\Omega$ and $t\in[T]$.

(2) From (1), it is clear that every $i,j\in\Omega$ have $\|B_t^i-B_j^t\|_*\leq 4V\Delta\sqrt{2t}$. Therefore, each $i\in\Omega$ is a candidate for B_t^{med}.

(3) By the definition of B_t^{med} and the triangle inequality, this is a consequence of (1).

(4) If we apply Lemma 2 with $\{X_1, X_2, \cdots, X_{T|\Omega|}\} = \{\nabla_i^k - \nabla_k\}_{k\in[T], i\in\Omega}$, we can have a similar bound as in (1).

C. Proof of Proposition 5

We apply Lemma 2 with $X_t = \nabla_t^i - \nabla_t$ for all $i \in \Omega$ first and then take an union bound over $t \in [T]$.

D. Proof of Proposition 6

According to Assumption 1, we have $\|\nabla_t^i - \nabla_t^j\|_* \leq 2V$ for $i, j \in \Omega$. Due to $\alpha \in [0, 1/2)$, every $i \in \Omega$ is a candidate for $\nabla_t^{\text{med}} = \nabla_t^i$. For the second part, we show it by contradiction. If $\|\nabla_t^{\text{med}} - \nabla_t\|_* > 3V$, then $\|\nabla_t^{\text{med}} - \nabla_t^i\|_* \geq \|\nabla_t^{\text{med}} - \nabla_t\|_* - \|\nabla_t^i - \nabla_t\|_* > 2V$ for $i \in \Omega$. This contradicts the definition of ∇_t^{med} due to $\alpha < 1/2$.

E. Proof of Proposition 7

By Proposition 3(2), Proposition 4(3), and Proposition 6 that, we know $|A_t^i - A_t^{\text{med}}| \leq I_A, \|B_t^i - B_t^{\text{med}}\|_* \leq I_B$, and $\|\nabla_t^{\text{med}} - \nabla_i^t\|_* \leq 4V$ for any $i \in \Omega$. Therefore, no elements from Ω will be removed from Ω_t for $t \in [T]$.

F. Proof of Proposition 8

Let $T_i \in \{0, 1, \cdots, T\}$ be the maximum iteration index so that $i \in \Omega_{T_i}$. The E_1 can be decomposed into two sums,

$$\begin{aligned}
&\sum_{t\in[T]} \sum_{i\in\Omega_t} \langle \nabla_t^i - \nabla_t, w_t - w^* \rangle \\
&= \sum_{i\in\Omega} (A_T^i - A_T^\star + \langle B_T^i - B_T^\star, w_1 - w^* \rangle) \\
&\quad + \sum_{i\notin\Omega} (A_{T_i}^i - A_{T_i}^\star + \langle B_{T_i}^i - B_{T_i}^\star, w_1 - w^* \rangle).
\end{aligned} \tag{7}$$

For the first sum, by Proposition 3(1) and Proposition 4(4), we have $|\sum_{i\in\Omega}(A_T^i - A_T^\star)| \leq 2WV\Delta\sqrt{2Tm}$ and $|\sum_{i\in\Omega}\langle B_t^i - B_T^\star, w_1 - w^* \rangle| \leq 2WV\Delta\sqrt{2Tm}$. For the second sum over $i \notin \Omega$, by the definition of I_A and T_i, we have $|A_{T_i}^i - A_{T_i}^{\text{med}}| \leq 4WV\Delta\sqrt{2T_i} \leq I_A$. Combining this with Proposition 3(2) yields

$$|A_{T_i}^i - A_{T_i}^\star| \leq |A_{T_i}^i - A_{T_i}^{\text{med}}| + |A_{T_i}^{\text{med}} - A_{T_i}^\star| \leq 6WV\Delta\sqrt{2T}.$$

Similarly, we have $\|B_{T_i}^i - B_{T_i}^{\text{med}}\|_* \leq 4V\Delta\sqrt{2T_i} \leq I_B$. Combining this with Proposition 4(3) yields

$$\begin{aligned}
\|B_{T_i}^i - B_{T_i}^\star\|_* &\leq \|B_{T_i}^i - B_{T_i}^{\text{med}}\|_* + \|B_{T_i}^{\text{med}} - B_{T_i}^\star\|_* \\
&\leq 10V\Delta\sqrt{2T}.
\end{aligned}$$

Finally, we collect these bounds for (7) and use the fact $|[m]\backslash\Omega| = \alpha m$ to complete the proof.

G. Proof of Proposition 9

For each $t \in [T]$, we have

$$\begin{aligned}
&\left\| \frac{1}{m} \sum_{i\in\Omega_t} (\nabla_t^i - \nabla_t) \right\|_*^2 \\
&\leq 2\left\| \frac{1}{m} \sum_{i\in\Omega} (\nabla_t^i - \nabla_t) \right\|_*^2 + 2\left\| \frac{1}{m} \sum_{i\in\Omega_t\backslash\Omega} (\nabla_t^i - \nabla_t) \right\|_*^2 \\
&\leq \frac{16V^2\Delta^2}{m} + 32\alpha^2V^2
\end{aligned}$$

where the first inequality follows $\|a+b\|_*^2 \leq 2\|a\|_*^2 + 2\|b\|_*^2$, and the second uses Proposition 5 and Proposition 6 by noting that $|\Omega_t\backslash\Omega| \leq \alpha m$.

REFERENCES

[1] B. McMahan and D. Ramage, "Federated learning: Collaborative machine learning without centralized training data," pp. 5650–5659, 2017. [Online]. Available: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[2] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[3] J. Konečnỳ, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *arXiv preprint arXiv:1511.03575*, 2015.

[4] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, p. 44, 2017.

[5] L. Su and J. Xu, "Securing distributed gradient descent in high dimensional statistical learning," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 3, no. 1, pp. 12:1–12:41, 2019.

[6] D. Alistarh, Z. Allen-Zhu, and J. Li, "Byzantine stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2018.

[7] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1206–1212.

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

[9] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, 2017, pp. 119–129.

[10] Q. Li, B. Kailkhura, R. Goldhahn, P. Ray, and P. K. Varshney, "Robust decentralized learning using ADMM with unreliable agents," *arXiv preprint arXiv:1710.05241*, 2017.

[11] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Math. Program. A*, vol. 120, no. 1, pp. 221–259, 2009.

[12] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *J. Mach. Learn. Res.*, vol. 11, no. Oct, pp. 2543–2596, 2010.

[13] S. Bubeck, "Convex optimization: algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[14] C. Xie, S. Koyejo, and I. Gupta, "Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6893–6901.

[15] X. Cao and L. Lai, "Distributed gradient descent algorithm robust to an arbitrary number of Byzantine attackers," 2018.

[16] J. Borwein, A. Guirao, P. Hájek, and J. Vanderwerff, "Uniformly convex functions on banach spaces," *Proceedings of the American Mathematical Society*, vol. 137, no. 3, pp. 1081–1091, 2009.

[17] X. Wei, H. Yu, and M. J. Neely, "Online primal-dual mirror descent under stochastic constraints," *arXiv preprint arXiv:1908.00305*, 2019.

[18] A. Rakhlin and K. Sridharan, "On equivalence of martingale tail bounds and deterministic regret inequalities," in *Proceedings of the 2017 Conference on Learning Theory*, vol. 65, 2017, pp. 1704–1722.

[19] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction using mini-batches," *J. Mach. Learn. Res.*, vol. 13, no. Jan, pp. 165–202, 2012.

[20] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: a survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, 2006.