2021 American Control Conference (ACC)
New Orleans, USA, May 25-28, 2021

Byzantine-resilient distributed learning under constraints

Dongsheng Ding, Xiaohan Wei, Hao Yu, and Mihailo R. Jovanović

***Abstract*— We consider a class of convex distributed statistical learning problems with inequality constraints in an adversarial scenario. At each iteration, an α-fraction of m machines, which are supposed to compute stochastic gradients of the loss function and send them to a master machine, may act adversarially and send faulty gradients. To guard against defective information sharing, we develop a Byzantine primal-dual algorithm. For $\alpha \in [0, 0.5)$, we prove that after T iterations the algorithm achieves $\tilde{O}(1/T + 1/\sqrt{mT} + \alpha/\sqrt{T})$ statistical error bounds on both the optimality gap and the constraint violation. Our result holds for a class of normed vector spaces and, when specialized to the Euclidean space, it attains the optimal error bound for Byzantine stochastic gradient descent.**

I. Introduction

In this paper, we examine a class of distributed statistical learning problems with inequality constraints in an adversarial scenario. Let $\{f(w; z), z \in \mathcal{Z}\}$ be a collection of closed convex functions whose domains contain the common closed convex set $\mathcal{W} \subset \mathbb{R}^d$, let $\{g_j(w)\}$ be a collection of convex functions on $\mathcal{W}$, and let $\mathcal{D}$ be an unknown distribution over the sample space $\mathcal{Z}$. The objective is to learn a model $w^\star$ by finding the minimizer to the convex program,

$$\begin{aligned} \underset{w \in \mathcal{W}}{\text{minimize}} \quad & F(w) := \mathbb{E}_{z \sim \mathcal{D}}\,[\, f(w; z)\,] \\ \text{subject to} \quad & g_j(w) \leq 0,\ j = 1, \ldots, k. \end{aligned} \tag{1}$$

The formulation (1) includes a broad class of constrained learning problems, e.g., constrained least-squares with $f(w; z) = (y - w^T x)^2$, $z = (x, y)$, and $g_j(w) = A_j w - b_j$. We focus on a distributed computational model with 1 master machine and m worker machines, where the master cannot collect all the data from the distribution $\mathcal{D}$; instead, at each time, the master receives m estimates of the gradient $\nabla F(w)$ from m workers. A popular application is the federated learning [1] where data is spread over a large number of worker machines and the master machine is unable to collect/store all data from mobile devices. In large-scale distributed learning, some machines can fail or even intentionally send malicious information [2]. Thus, studying distributed learning algorithms and their robustness against faulty information is an important and timely topic. In this paper, we consider an adversarial setup where workers behave maliciously by sending arbitrary vectors; they are called Byzantine machines.

Financial support from the National Science Foundation under Awards ECCS-1708906 and ECCS-1809833 is gratefully acknowledged.

D. Ding and M. R. Jovanović are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089. X. Wei is with Facebook Inc., Menlo Park, CA 94025. H. Yu is with Amazon Inc., Seattle, WA 98109. E-mails: dongshed@usc.edu, xiaohanw@usc.edu, eeyuhao@gmail.com, mihailo@usc.edu

Our contribution: In this paper, we propose a new variant of the primal-dual method – Byzantine primal-dual (Byzantine PD) algorithm – for solving problem (1) in an adversarial scenario where an α-fraction of m workers are Byzantine. For $\alpha \in [0, 0.5)$, we prove that after T iterations the algorithm achieves $\tilde{O}(1/T + 1/\sqrt{mT} + \alpha/\sqrt{T})$ statistical error bounds on both the optimality gap and the constraint violation. This result holds for a large class of normed vector spaces and matches the optimal statistical error bound for the problem (1) without constraints in the Euclidean space setting. To the best of our knowledge, our work provides the first study of primal-dual methods for distributed constrained learning problems in the Byzantine adversarial setting.

Related work: Closely related studies on primal-dual methods are references [3]–[9]. For general deterministic convex optimization problems with convex nonlinear constraints, references [4], [5] propose primal-dual algorithms based on drift-plus-penalty [10] and prove $O(1/T)$ convergence rate on both optimality gap and constraint violation. When the objective function and constraints are time-varying, references [3], [6]–[9] propose online primal-dual methods with convergence guarantees regarding regret and constraint violation. However, all of these studies assume access to exact gradients of objective/constraint functions or their samples. It is not the case for practical large-scale distributed learning. To guard against adversarial gradients, this paper generalizes the primal-dual method to Byzantine stochastic optimization in general normed vector spaces.

Our distributed computational model is also relevant to studies of gradient descent [11]–[14] or mirror descent [15] in the Byzantine setting. To mitigate Byzantine machines, the median aggregation has been extensively used. However, apart from reference [15], most other approaches only work in the Euclidean space setting. To add the flexibility to our algorithm, we utilize median aggregation in general normed vector spaces, as done in reference [15]. On the other hand, it is tempting to extend gradient descent methods by adding projection to deal with constraints. However, this is not suitable for our problem since nonlinear constraints can make a projection as hard as solving the original problem. Instead, we employ the primal-dual method to deal with constraints. Thus, our work departs from several unconstrained results, e.g., those reported in references [14], [15], and we focus on the constrained learning problems.

Paper outline: In Section II, we present our main assumptions and describe the algorithm. We conduct our analysis in Section III and present convergence results in Section IV. We conclude the paper in Section V.

978-1-6654-4197-1/$31.00 ©2021 AACC

Notation: The Banach space $(\mathbb{R}^d, \|\cdot\|_p)$, $p \in [1,\infty)$, is 2-smooth if $\rho(s) \leq Cs^2$ where $\rho(s) := \sup_{\|x\|=1,\|y\|=s}\{\frac{1}{2}(\|x+y\|+\|x-y\|)-1\}$ is the smooth modulus and C is a constant. Its dual is $(\mathbb{R}^d, \|\cdot\|_*)$ where the dual norm $\|\cdot\|_*$ is defined as $\|z\|_* = \sup_{\|x\|\leq 1}\langle z, x\rangle$. A function f is L-Lipschitz with respect to a norm $\|\cdot\|$ if $|f(x)-f(y)| \leq L\|x-y\|$. A function is β-smooth with respect to a norm $\|\cdot\|$ if $\|\nabla f(x)-\nabla f(y)\|_* \leq \beta\|x-y\|$. A function is σ-strongly convex with respect to the norm $\|\cdot\|$ if $f(y) \geq f(x)+\langle\nabla f(x), y-x\rangle + \frac{\sigma}{2}\|y-x\|^2$. For a 1-strongly convex function ϕ with respect to the norm $\|\cdot\|$, the Bregman divergence $D(x,y)$ generated by ϕ is $D(x,y) = \phi(x)-\phi(y)-\langle\nabla\phi(y), x-y\rangle$.

II. Assumptions and algorithm

Let $[n] := \{1,\ldots,n\}$ and let us arrange $g_1(w),\ldots,g_k(w)$ in the problem (1) into a vector $g(w) := (g_1(w),\ldots,g_k(w))$ and rewrite constraints as $g(w) \leq 0$.

Assumption 1 (Basic properties):

(i) The domain $\mathcal{W} \in \mathbb{R}^d$ is a compact, convex set with diameter W, and there exist R and G such that $D(x,y) \leq R^2$ for all $x, y \in \mathcal{W}$ and $\|g(w)\| \leq G$ for all $w \in \mathcal{W}$;

(ii) The objective function $F(w)$ and the constraint functions $g_j(w)$, with $j \in [k]$, are convex and smooth on $\mathcal{W}$. Each $g_j(w)$ is also Lipschitz continuous with parameter $L_{j,g}$ for all $j \in [k]$. We denote smoothness parameters of $F(w)$ and $g_j(w)$ as β_F and $\beta_{j,g}$, respectively;

(iii) There exists an optimal solution $w^\star \in \mathcal{W}$ that solves the problem (1).

Assumption 2 (Existence of Lagrange multipliers): There exists a Lagrange multiplier $\lambda^\star := (\lambda_1^\star,\ldots,\lambda_k^\star) \geq 0$ such that $q(\lambda^\star) = \min_{w\in\mathcal{W}}\{F(w)\colon g(w) \leq 0\}$ where $q(\lambda) = \min_{w\in\mathcal{W}}\{F(w)+\langle\lambda, g(w)\rangle\}$ is the dual function.

In our computational model with 1 master and m workers, at iteration t, each worker i receives current iterate w_t, utilizes private data z_t^i to compute the associated gradient, and returns it to the master. There are two possibilities: (i) a non-Byzantine worker returns $\nabla f(w_t; z_t^i)$ with $z_t^i \sim \mathcal{D}$; (ii) a Byzantine worker returns arbitrary vector adversarially. After receiving information from all workers, the master aggregates them for optimization, generates next iterate w_{t+1}, and then broadcasts it. The learning goal for the master is to obtain an approximate solution to the problem (1) after T iterations.

Let $\Omega \subseteq \{1,\ldots,m\}$ be an unknown set of non-Byzantine machines. We denote by $\nabla_t^i := \nabla f(w_t; z_t^i)$ the sampled gradient and $\nabla_t := \nabla F(w_t)$ the true gradient, respectively.

Assumption 3: (i) At each iteration t, there exists $V > 0$ such that $\|\nabla_t^i - \nabla_t\|_* \leq V$ for machine $i \in \Omega$. (ii) There is an α-fraction of Byzantine machines with $\alpha \in [0, 0.5)$.

We only impose Assumption 3 (i) on an unknown set of non-Byzantine machines. This does not simplify the Byzantine issue since we can always take a large V.

Let $w_0 \in \mathbb{R}^d$ be an initial point. Algorithm 1 has two stages. The first stage (lines 3–6) estimates the set Ω_t of non-Byzantine machines by maintaining three estimation sequences for each machine $i \in [k]$, i.e., received gradient ∇_t^i, cumulative gradient B_t^i, and gradient-related value A_t^i,

$$B_t^i := \sum_{\tau=0}^{t}\nabla_\tau^i \quad\text{and}\quad A_t^i := \sum_{\tau=0}^{t}\langle\nabla_\tau^i, w_\tau - w_0\rangle.$$

Let ∇_t^{med} be the median of $\{\nabla_t^1,\ldots,\nabla_t^m\}$, B_t^{med} be the median of $\{B_t^1,\ldots,B_t^m\}$, and A_t^{med} be the median of $\{A_t^1,\ldots,A_t^m\}$. We begin Ω_0 with $[m]$ and update Ω_t by a set of machine i from Ω_{t-1} whose ∇_t^i is $4V$-close to ∇_t^{med}, I_B-close to B_t^{med}, and I_A-close to A_t^{med}. Using Ω_t, we next estimate the population gradient ∇_t by the averaged one (6).

The second stage (line 7) maintains the primal-dual updates in terms of $w_t \in \mathcal{W}$ and $q_{j,t} \in \mathbb{R}$, with $j \in [k]$, using the estimated gradient ξ_t. At iteration t, it updates the primal variable w_{t+1} by a solution to the proximal-type problem with Bregman divergence. The dual variable $q_{j,t+2}$ is updated via a simple $\max$ function. Since the dual update mimics the queueing equation [4], the dual variables are also called virtual queues. In the next section, we discuss some important properties of iterates w_t and $q_{j,t}$, with $j \in [k]$.

III. Preliminaries and basic analysis

Our convergence analysis relies on two gradient-related quantities generated by Algorithm 1,

$$\begin{aligned} E_1 &:= \sum_{t=0}^{T-1}\sum_{i\in\Omega_t}\langle\nabla_t^i - \nabla_t, w_t - w^\star\rangle \\ E_2 &:= \frac{1}{T}\sum_{t=0}^{T-1}\left\|\frac{1}{m}\sum_{i\in\Omega_t}(\nabla_t^i - \nabla_t)\right\|_*^2 \end{aligned} \tag{2}$$

where E_1 determines the bias arising from the stochastic gradient and the adversarial workers, and E_2 is the variance of estimating the true gradient.

By the concentration bound [16] in a 2-smooth Banach space, we can choose proper parameters I_A, I_B, and Δ in Algorithm 1 such that the median aggregation in line 6 allows bounds on E_1 and E_2 in Lemma 1. We refer readers to [15] for details; [14] for a special Euclidean case.

Lemma 1 (Error bounds): [15, Lemmas 8 and 9] Let $(\mathbb{R}^d, \|\cdot\|_*)$ be 2-smooth and let Assumption 3 hold. With probability $1-\delta$, for (2) we have (i) $|E_1| \leq 4WV\Delta\sqrt{2mT} + 16\alpha mWV\Delta\sqrt{2T}$; and (ii) $E_2 \leq 32\alpha^2V^2 + \frac{16V^2\Delta^2}{m}$.

Let $q_t = (q_{1,t},\ldots,q_{k,t})$ be the vector of virtual queues and let $L_t = \frac{1}{2}\|q_t\|^2$ be a quadratic Lyapunov function. We define the Lyapunov drift as $d_t := L_{t+1} - L_t = \frac{1}{2}(\|q_{t+1}\|^2 - \|q_t\|^2)$ for $t \geq 1$. For standard properties of the virtual queues and the Lyapunov drift, we refer readers to Appendix A.

By Assumption 1, the smoothness of g_j with modulus $\beta_{j,g}$ implies that the function $(q_{j,t+1} + g_j(w_t))g_j(w)$ is smooth in w with modulus $(q_{j,t+1} + g_j(w_t))\beta_{g,j}$. By the descent lemma [17, Proposition A.24], we have the following lemma.

Lemma 2: Let Assumption 1 hold. For Algorithm 1 with

$t \geq 0$, we have

$$\begin{aligned}
&\big(q_{t+1} + g(w_t)\big)^T g(w_{t+1}) \\
&\leq \sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big)\big(g_j(w_t) + \langle \nabla g_j(w_t), w_{t+1} - w_t\rangle\big) \\
&\quad + \frac{(q_{t+1} + g(w_t))^T \beta_g}{2} \|w_{t+1} - w_t\|^2
\end{aligned}$$

where $\beta_g := (\beta_{1,g}, \ldots, \beta_{k,g})$.

The following pushback property is useful for analyzing the primal update in line 7 of Algorithm 1.

Lemma 3 (Pushback property): [18, Lemma 1] Let f: $\mathcal{W} \to \mathbb{R}$ be a convex function, $\eta > 0$, and $y \in \mathcal{W}$. If $x^\star = \operatorname{argmin}_{x \in \mathcal{W}} f(x) + \eta D(x, y)$, then $f(x^\star) + \eta D(x^\star, y) \leq f(z) + \eta D(z, y) - \eta D(z, x^\star)$ for any $z \in \mathcal{W}$.

Denote $\iota_t := L_g^2 + \beta_F + (q_{t+1} + g(w_t))^T \beta_g$. We relate the objective function in (1) to the errors (2) via Lemmas 2 and 3 and establish a useful inequality (3) on the drift.

Lemma 4: Let Assumption 1 hold. Suppose $\eta_t > \iota_t$. In Algorithm 1, for $t \geq 0$, we have

$$\begin{aligned}
&d_{t+1} + \frac{1}{m} \sum_{i \in \Omega_t} (F(w_{t+1}) - F(w^\star)) \\
&\leq \frac{1}{m} \sum_{i \in \Omega_t} \big\langle (\nabla_t^i - \nabla_t), w^\star - w_t \big\rangle \\
&\quad + \frac{1}{2(\eta_t - \iota_t)} \left\| \frac{1}{m} \sum_{i \in \Omega_t} (\nabla_t^i - \nabla_t) \right\|_*^2 \\
&\quad + \eta_t \big(D(w^\star, w_t) - D(w^\star, w_{t+1})\big) \\
&\quad + \frac{1}{2} \left(\|g(w_{t+1})\|^2 - \|g(w_t)\|^2\right).
\end{aligned} \tag{3}$$

Proof: See Appendix B. ∎

In Lemma 4, we have established an upper bound on our drift-plus-penalty term in (3). The first two terms in the bound describes bias and variance of gradient estimation that relates to (2). The last two terms account for the regularization and the constraints. Hence, (3) is different from bounds in standard drift-plus-penalty analysis, e.g., references [4], [10], and the constant stepsize rule is no longer valid. We next present an adaptive stepsize rule and establish our convergence results.

IV. Main results

We now provide the convergence analysis of Algorithm 1. We adaptively adjust the parameter η_t using $\iota_t := L_g^2 + \beta_F + (q_{t+1} + g(w_t))^T \beta_g$.

(i) If $\alpha \in [\,1/\sqrt{m},\, 0.5\,)$, we choose

$$\eta_t = \begin{cases} \iota_0 + \sqrt{T}, & t = 0; \\ \max\{\,\eta_{t-1},\, \iota_t + \sqrt{T}\,\}, & t \geq 1. \end{cases} \tag{4}$$

(ii) If $\alpha \in [\,0,\, 1/\sqrt{m}\,)$, we choose

$$\eta_t = \begin{cases} \iota_0 + \sqrt{T/m}, & t = 0; \\ \max\{\,\eta_{t-1},\, \iota_t + \sqrt{T/m}\,\}, & t \geq 1. \end{cases} \tag{5}$$

It is easy to see that η_t is non-decreasing for $t \geq 0$ and $\eta_0 > 0$ from (ii) in Lemma 10.

Lemma 5: Let Assumption 1 hold. Then, for $t \geq 1$,

$$\sum_{\tau=0}^{t-1} \eta_\tau \big(D(w^\star, w_\tau) - D(w^\star, w_{\tau+1})\big) \leq \eta_{t-1} R^2.$$

Proof: See Appendix C. ∎

Algorithm 1 Byzantine Primal-Dual (Byzantine PD)

Initialization: Initial point $w_0 \in \mathcal{W}$, initial virtual queues $q_{j,1} = \max\{0, -g_j(w_0)\}, \forall j \in [k]$, diameters $W, R > 0$, number of iterations T, thresholds $I_A = 4WV\Delta\sqrt{2T}$ and $I_B = 4V\Delta\sqrt{2T}$ where $\Delta := R + 2\sqrt{2\log(8\sqrt{2}\,mT/\delta)}$.

1: $\Omega_0 \leftarrow [m]$;
2: **for all** $t \leftarrow 0$ **to** $T-1$ **do**
3: **for all** $i \leftarrow 1$ **to** m **do**
4: receive $\nabla_t^i \in \mathbb{R}^d$ from worker $i \in [m]$ and update $B_t^i \leftarrow \sum_{\tau=0}^t \nabla_\tau^i$ and $A_t^i \leftarrow \sum_{\tau=0}^t \langle \nabla_\tau^i, w_\tau - w_0\rangle$.
5: **end for**
6: gradient estimation

- $A_t^{\text{med}} \leftarrow \text{median}\{A_t^1, \ldots, A_t^m\}$.
- $B_t^{\text{med}} \leftarrow B_t^i$ where $i \in [m]$ is any machine s.t.
$$|\{j \in [m] : \|B_t^i - B_t^j\|_* \leq I_B\}| > \frac{m}{2}.$$
- $\nabla_t^{\text{med}} \leftarrow \nabla_t^i$ where $i \in [m]$ is any machine s.t.
$$|\{j \in [m] : \|\nabla_t^i - \nabla_t^j\|_* \leq 2V\}| > \frac{m}{2}.$$
- $\Omega_t \leftarrow \Omega_{t-1} \cap \{i \in [m] : |A_t^i - A_t^{\text{med}}| \leq I_A, \|B_t^i - B_t^{\text{med}}\|_* \leq I_B, \text{ and } \|\nabla_t^i - \nabla_t^{\text{med}}\|_* \leq 4V\}$.
- compute the gradient
$$\xi_t = \frac{1}{m} \sum_{i \in \Omega_t} \nabla_t^i. \tag{6}$$

7: primal-dual update

- primal update
$$w_{t+1} \leftarrow \operatorname*{argmin}_{w \in \mathcal{W}} \left\langle \xi_t + \sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big) \nabla g_j(w_t), w \right\rangle + \eta_t\, D(w, w_t).$$
- dual update for all $j \in [k]$
$$q_{j,t+2} \leftarrow \max\big(-g_j(w_{t+1}),\, q_{j,t+1} + g_j(w_{t+1})\big).$$

8: **end for**
9: **Output:** $\bar{w}_T := \frac{1}{T} \sum_{\tau=0}^{T-1} w_{t+1}$

Lemma 6: Let $(\mathbb{R}^d, \|\cdot\|_*)$ be 2-smooth and let Assumptions 1–3 hold. For $1 \leq t \leq T$, with probability at least $1-\delta$, one of the following holds

(i) For $\alpha \in [1/\sqrt{m}, 0.5)$ and η_t given by (4),

$$\|q_{t+1}\| \leq \|\lambda^\star\| + \sqrt{2\eta_{t-1}} R + G + C_1 \tag{7}$$

(ii) For $\alpha \in [0, 1/\sqrt{m})$ and η_t given by (5),

$$\|q_{t+1}\| \;\le\; \|\lambda^\star\| + \sqrt{2\eta_{t-1}}R + G + C_2. \quad (8)$$

Here, $C_1 = \sqrt{C_{1,1}+C_{1,2}}$ and $C_2 = \sqrt{C_{2,1}+C_{2,2}}$ with $C_{1,1} = \frac{8WV\Delta\sqrt{2T}}{\sqrt{m}} + \frac{16V^2\Delta^2\sqrt{T}}{\alpha m}$, $C_{1,2} = 32\alpha(V^2 + \sqrt{2}WV\Delta)\sqrt{T}$, $C_{2,1} = \frac{8WV\Delta\sqrt{2T}+16V^2\Delta^2\sqrt{T}}{\sqrt{m}}$, and $C_{2,2} = 32\alpha(WV\Delta\sqrt{2T} + \alpha V^2\sqrt{mT})$.

Proof: See Appendix D. ∎

Lemma 7: Let $(\mathbb{R}^d, \|\cdot\|_*)$ be 2-smooth and let Assumptions 1–3 hold. For $1 \le t \le T$, with probability at least $1-\delta$, one of the following holds,

(i) For $\alpha \in [1/\sqrt{m}, 0.5)$ and η_t given by (4),

$$\eta_t \;\le\; \eta_1^{\max} := \left(\sqrt{\bar{\eta}_1 + \sqrt{T}} + \sqrt{2}R\|\beta_g\|\right)^2$$

(ii) For $\alpha \in [0, 1/\sqrt{m})$ and η_t given by (5),

$$\eta_t \;\le\; \eta_2^{\max} := \left(\sqrt{\bar{\eta}_2 + \sqrt{T/m}} + \sqrt{2}R\|\beta_g\|\right)^2$$

where $\bar{\eta}_i := L_g^2 + \beta_F + (2G + C_i + \|\lambda^\star\|)\|\beta_g\|$ for $i = 1, 2$.

Proof: See Appendix E. ∎

We are now ready to establish convergence in terms of the optimality gap and the constraint violation.

Theorem 8: Let $(\mathbb{R}^d, \|\cdot\|_*)$ be 2-smooth and let Assumptions 1–3 hold. With probability at least $1-\delta$, if $T \ge m$ one of the following holds,

(i) For $\alpha \in [1/\sqrt{m}, 0.5)$ and η_t given by (4),

$$F(\bar{w}_T) - F(w^\star) \;\le\; O\left(\frac{C_{F,0}}{T} + \frac{C_{F,1}}{\sqrt{mT}} + \alpha\frac{C_{F,2}}{\sqrt{T}}\right);$$

(ii) For $\alpha \in [0, 1/\sqrt{m})$ and η_t given by (5),

$$F(\bar{w}_T) - F(w^\star) \;\le\; O\left(\frac{C_{F,0}}{T} + \frac{C_{F,3}}{\sqrt{mT}} + \alpha\frac{F_{F,4}}{\sqrt{T}}\right)$$

where $\bar{w}_T := \frac{1}{T}\sum_{t=0}^{T-1} w_t$, $C_{F,i}$, $i = 0, 1, \ldots, 4$ are constants that only depend on parameters $\{G, R, W, V, \Delta\}$.

Proof: We show the first case using η_t in (4). We begin with (3) in Lemma 4. Notice that $\eta_t - \iota_t \ge \sqrt{T}$. Thus, we can simplify (3) and show that

$$\begin{aligned}
&\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i\in\Omega_t}(F(w_{t+1}) - F(w^\star)) + \frac{1}{T}\sum_{t=0}^{T-1} d_{t+1} \\
&\le \frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i\in\Omega_t}\langle(\nabla_t^i - \nabla_t), w^\star - w_t\rangle \\
&\quad + \frac{1}{2T\sqrt{T}}\sum_{t=0}^{T-1}\left\|\frac{1}{m}\sum_{i\in\Omega_t}(\nabla_t^i - \nabla_t)\right\|_*^2 \\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1}\eta_t\big(D(w^\star, w_t) - D(w^\star, w_{t+1})\big) \\
&\quad + \frac{1}{2T}\sum_{t=0}^{T-1}\left(\|g(w_{t+1})\|^2 - \|g(w_t)\|^2\right).
\end{aligned}$$

Using E_1 and E_2 in (2) and Lemma 10 (iii) yields

$$\begin{aligned}
&\frac{1}{mT}\sum_{t=0}^{T-1}\sum_{i\in\Omega_t}(F(w_{t+1}) - F(w^\star)) \\
&\le \frac{|E_1|}{mT} + \frac{E_2}{2\sqrt{T}} + \frac{G^2}{2T} \\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1}\eta_t\big(D(w^\star, w_t) - D(w^\star, w_{t+1})\big).
\end{aligned} \quad (9)$$

Lemmas 5 and 7 allow us to bound the right-hand side of (9) via $\frac{|E_1|}{mT} + \frac{E_2}{2\sqrt{T}} + \frac{\eta_1^{\max}R^2}{T}$. Furthermore, substituting the bounds on $|E_1|$ and E_2 in Lemma 1 and the bound on $\eta_1^{\max}$ in Lemma 7 into this term leads to a bound that has the order of $\frac{G^2+R^2(L_g^2+\beta_F+R^2+G+\|\lambda^\star\|)}{T} + \frac{WV\Delta+V^2\Delta^2}{\sqrt{mT}} + \alpha\frac{V+WV\Delta+R^2}{\sqrt{T}} + \frac{R^2(\sqrt{WV\Delta}+V\Delta)}{m^{1/4}T^{3/4}} + \sqrt{\alpha}\frac{R^2(V+\sqrt{WV\Delta})}{T^{3/4}}$. Let $T \ge m$. It is clear that $\frac{1}{m^{1/4}T^{3/4}} \le \frac{1}{\sqrt{mT}}$ and $\frac{\sqrt{\alpha}}{T^{3/4}} \le \frac{\alpha}{\sqrt{T}}$. Thus, we obtain the first bound. We complete the proof by applying the convexity of F to the left-hand side of (9) so that it is lower bounded by $\frac{1}{2}(F(\bar{w}_T) - F(w^\star))$.

Similarly, we can show the second case using η_t in (5) and $\eta_2^{\max}$ in Lemma 7. ∎

Theorem 9: Let $(\mathbb{R}^d, \|\cdot\|_*)$ be 2-smooth and let Assumptions 1–3 hold. With probability at least $1-\delta$, if $T \ge m$ one of the following holds,

(i) For $\alpha \in [1/\sqrt{m}, 0.5)$ and η_t given by (4),

$$g_j(\bar{w}_T) \;\le\; O\left(\frac{C_{g,0}}{T} + \frac{C_{g,1}}{\sqrt{mT}} + \alpha\frac{C_{g,2}}{\sqrt{T}}\right);$$

(ii) For $\alpha \in [0, 1/\sqrt{m})$ and η_t given by (5),

$$g_j(\bar{w}_T) \;\le\; O\left(\frac{C_{g,0}}{T} + \frac{C_{g,3}}{\sqrt{mT}} + \alpha\frac{C_{g,4}}{\sqrt{T}}\right)$$

where $\bar{w}_T := \frac{1}{T}\sum_{t=0}^{T-1} w_{t+1}$, $C_{g,i}$, $i = 0, 1, \ldots, 4$ are constants that only depend on parameters $\{G, R, W, V, \Delta\}$.

Proof: Since g_j, $j \in [k]$ are convex, we can apply the Jensen's inequality and Lemma 11,

$$g_j(\bar{w}_T) \;\le\; \frac{1}{T}\sum_{t=0}^{T-1} g_j(w_{t+1}) \;\le\; \frac{1}{T}\, q_{j,T+1} \;\le\; \frac{1}{T}\,\|q_{T+1}\|.$$

Next, we apply Lemma 7 and discuss two cases.

$$g_j(\bar{w}_T) \;\le\; \begin{cases} \frac{1}{T}\left(\|\lambda^\star\| + \sqrt{2\eta_1^{\max}}R + G + C_1\right), & \text{for (i);} \\ \frac{1}{T}\left(\|\lambda^\star\| + \sqrt{2\eta_2^{\max}}R + G + C_2\right), & \text{for (ii).} \end{cases}$$

Since $x^2 + ax \le (a+1)x^2 + a/4$ for all x and $a \ge 0$, we complete the proof using the expressions for C_1 and C_2 in Lemma 6, $\eta_1^{\max}$ and $\eta_2^{\max}$ in Lemma 7, and $T \ge m$. ∎

Remark 1: When all workers are non-Byzantine, i.e., $\alpha = 0$, the first two terms in the bounds in Theorems 8 and 9 are similar to the rate of mini-batch SGD [19]. The last term determines the effect of Byzantine workers for $\alpha \ne 0$. If the gradients ∇_t^i are unbiased, the dual norm bound of gradients becomes $V = 0$ and the bounds are $O(1/T)$. This matches the optimal rate for the convex stochastic program.

V. Conclusion

We have developed a variant of the primal-dual algorithm for constrained distributed learning problems in the Byzantine setting. We have proved the robustness against Byzantine failures whenever the fraction of Byzantine machines satisfies $\alpha \in [0, 0.5)$. When the objective and constraint functions are convex and smooth the algorithm after T iterations enjoys $\tilde{O}(1/T + 1/\sqrt{mT} + \alpha/\sqrt{T})$ statistical error bounds on both the optimality gap and the constraint violation.

Appendix

A. Properties of Virtual Queues

Lemma 10 (Property of virtual queues): [4, Lemma 3] In line 7 of Algorithm 1, we have

(i) For any $j \in [k]$, $q_{j,t} \geq 0$ for $t \geq 1$;
(ii) For any $j \in [k]$, $q_{j,t} + g_j(w_{t-1}) \geq 0$ for $t \geq 1$;
(iii) $\|q_1\|^2 \leq \|g(w_0)\|^2$ and $\|q_t\|^2 \geq \|g(w_{t-1})\|^2$ for $t \geq 2$.

Lemma 11 (Constraint violation): [4, Lemma 7] Let q_t with $t \geq 1$ be the sequence generated by Algorithm 1. For any $j \in [k]$, we have $q_{j,t+1} \geq \sum_{\tau=0}^{t-1} g_j(w_{\tau+1})$ for all $t \geq 1$.

Lemma 12 (Drift property): [4, Lemma 4] In Algorithm 1, the Lyapunov drift satisfies $d_t \leq q_t^T g(w_t) + \|g(w_t)\|^2$ for all $t \geq 1$.

B. Proof of Lemma 4

Applying Lemma 3 to line 7 of Algorithm 1 with $x^\star = w_{t+1}$, $z = w^\star$, and $y = w_t$ yields

$$\begin{aligned}
&\langle \xi_t, w_{t+1} \rangle + \sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big) \nabla g_j(w_t)^T w_{t+1} \\
&\leq \langle \xi_t, w^\star \rangle + \sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big) \nabla g_j(w_t)^T w^\star \\
&\quad + \eta_t D(w^\star, w_t) - \eta_t D(w^\star, w_{t+1}) - \eta_t D(w_{t+1}, w_t).
\end{aligned}$$

Adding $-\langle \xi_t, w_t \rangle + \sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big)\big(g_j(w_t) - \nabla g_j(w_t)^T w_t\big)$ to both sides of the inequality above and applying the convexity of g_j and the inequality $q_{j,t+1} + g_j(w_t) \geq 0$ from Lemma 10 (ii) lead to

$$\begin{aligned}
&\langle \xi_t, w_{t+1} - w_t \rangle \\
&+ \sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big)\big(g_j(w_t) + \nabla g_j(w_t)^T (w_{t+1} - w_t)\big) \\
&\leq \langle \xi_t, w^\star - w_t \rangle + \sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big) g_j(w^\star) \\
&\quad + \eta_t D(w^\star, w_t) - \eta_t D(w^\star, w_{t+1}) - \eta_t D(w_{t+1}, w_t).
\end{aligned}$$

Moreover, we remove $\sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big) g_j(w^\star)$ without changing the inequality due to feasibility of $w^\star$. Thus,

$$\begin{aligned}
&\langle \xi_t, w_t - w^\star \rangle + \eta_t D(w_{t+1}, w_t) + \\
&\sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big)\big(g_j(w_t) + \nabla g_j(w_t)^T (w_{t+1} - w_t)\big) \\
&\leq \langle \xi_t, w_t - w_{t+1} \rangle + \eta_t \big(D(w^\star, w_t) - D(w^\star, w_{t+1})\big).
\end{aligned} \tag{10}$$

By the convexity of F and the smoothness, i.e., $F(w_t) \geq F(w_{t+1}) - \langle \nabla_t, w_{t+1} - w_t \rangle - \frac{\beta_F}{2}\|w_{t+1} - w_t\|^2$, we can have simplify (10) into

$$\begin{aligned}
&\frac{1}{m} \sum_{i \in \Omega_t} (F(w_{t+1}) - F(w^\star)) \\
&+ \sum_{j=1}^{k} \big(q_{j,t+1} + g_j(w_t)\big)\big(g_j(w_t) + \nabla g_j(w_t)^T (w_{t+1} - w_t)\big) \\
&\leq \frac{1}{m} \sum_{i \in \Omega_t} \langle (\nabla_t^i - \nabla_t), w^\star - w_t \rangle + \frac{\beta_F}{2}\|w_{t+1} - w_t\|^2 \\
&\quad + \left\langle \frac{1}{m} \sum_{i \in \Omega_t} (\nabla_t^i - \nabla_t), w_t - w_{t+1} \right\rangle \\
&\quad + \eta_t \big(D(w^\star, w_t) - D(w^\star, w_{t+1})\big) - \eta_t D(w_{t+1}, w_t).
\end{aligned}$$

We add the inequality in Lemma 2 into the inequality above and use the property of Bregman divergence $D(w_{t+1}, w_t) \geq \frac{1}{2}\|w_{t+1} - w_t\|^2$ first, and then apply $g(w_t)^T g(w_{t+1}) = \frac{1}{2}\big(\|g(w_t)\|^2 + \|g(w_{t+1})\|^2 - \|g(w_t) - g(w_{t+1})\|^2\big)$ and the Lipschitz continuity of g,

$$\begin{aligned}
&\frac{1}{m} \sum_{i \in \Omega_t} (F(w_{t+1}) - F(w^\star)) + q_{t+1}^T g(w_{t+1}) \\
&\leq \frac{1}{m} \sum_{i \in \Omega_t} \big\langle (\nabla_t^i - \nabla_t), w^\star - w_t \big\rangle + \frac{\iota_t - \eta_t}{2} \|w_{t+1} - w_t\|^2 \\
&\quad + \left\langle \frac{1}{m} \sum_{i \in \Omega_t} (\nabla_t^i - \nabla_t), w_t - w_{t+1} \right\rangle \\
&\quad + \eta_t \big(D(w^\star, w_t) - D(w^\star, w_{t+1})\big) \\
&\quad - \frac{1}{2}\big(\|g(w_t)\|^2 + \|g(w_{t+1})\|^2\big)
\end{aligned} \tag{11}$$

where $\iota_t := L_g^2 + \beta_F + (q_{t+1} + g(w_t))^T \beta_g$.

Finally, we apply the drift property in Lemma 12 to (11), the CauchySchwarz inequality, and $bx - ax^2 \leq \frac{b^2}{4a}$, $a, b > 0$,

$$\begin{aligned}
&\left\langle \frac{1}{m} \sum_{i \in \Omega_t} (\nabla_t^i - \nabla_t), w_t - w_{t+1} \right\rangle + \frac{\iota_t - \eta_t}{2} \|w_{t+1} - w_t\|^2 \\
&\leq \left\| \frac{1}{m} \sum_{i \in \Omega_t} (\nabla_t^i - \nabla_t) \right\|_* \cdot \|w_t - w_{t+1}\| \\
&\quad + \frac{\iota_t - \eta_t}{2} \|w_{t+1} - w_t\|^2 \\
&\leq \frac{1}{2} \frac{1}{\eta_t - \iota_t} \left\| \frac{1}{m} \sum_{i \in \Omega_t} (\nabla_t^i - \nabla_t) \right\|_*^2
\end{aligned}$$

where $\eta_t > \iota_t$. Combining this inequality above with (11) yields the desired result.

C. Proof of Lemma 5

We expand $\sum_{\tau=0}^{t-1} \eta_\tau \big(D(w^\star, w_\tau) - D(w^\star, w_{\tau+1})\big)$ into

$$\sum_{\tau=0}^{t-2} (\eta_{\tau+1} - \eta_\tau) D(w^\star, w_{\tau+1}) + \eta_0 D(w^\star, w_0) - \eta_{t-1} D(w^\star, w_t)$$

We complete proof by removing a term $-\eta_{t-1} D(w^\star, w_t) \leq 0$ and applying $D(w^\star, w_\tau) \leq R^2$ and non-decreasing η_τ.

D. Proof of Lemma 6

We show the first case using η_t given in (4). We begin with (3) in Lemma 4. With a slight abuse of notation, we use τ as time index and t as time horizon. According to (4), we have $\eta_\tau - \iota_\tau \geq \alpha\sqrt{T}$. This allows us to simplify (3),

$$\begin{aligned}
&\frac{1}{mt}\sum_{\tau=0}^{t-1}\sum_{i\in\Omega_\tau}(F(w_{\tau+1})-F(w^\star)) + \frac{1}{t}\sum_{\tau=0}^{t-1} d_{\tau+1} \\
\leq\ & \frac{1}{mt}\sum_{\tau=0}^{t-1}\sum_{i\in\Omega_\tau}\langle(\nabla_\tau^i-\nabla_\tau), w^\star - w_\tau\rangle \\
&+ \frac{1}{2\alpha t\sqrt{T}}\sum_{\tau=0}^{t-1}\left\|\frac{1}{m}\sum_{i\in\Omega_\tau}(\nabla_\tau^i-\nabla_\tau)\right\|_*^2 \\
&+ \frac{1}{t}\sum_{\tau=0}^{t-1}\eta_\tau\big(D(w^\star,w_\tau) - D(w^\star,w_{\tau+1})\big) \\
&+ \frac{1}{2t}\sum_{\tau=0}^{t-1}\left(\|g(w_{\tau+1})\|^2 - \|g(w_\tau)\|^2\right).
\end{aligned}$$

Similar to previous notation of E_1 and E_2, for $1 \leq t \leq T$, we introduce $E_1' := \sum_{\tau=0}^{t-1}\sum_{i\in\Omega_\tau}\langle\nabla_\tau^i - \nabla_\tau, w_\tau - w^\star\rangle$ and $E_2' := \frac{1}{t}\sum_{\tau=0}^{t-1}\|\frac{1}{m}\sum_{i\in\Omega_\tau}(\nabla_\tau^i - \nabla_\tau)\|_*^2$. With this notation, we apply $\|q_1\|^2 \leq \|g(w_0)\|^2$ and Lemma 5,

$$\begin{aligned}
&\frac{1}{mt}\sum_{\tau=0}^{t-1}\sum_{i\in\Omega_\tau}(F(w_{\tau+1})-F(w^\star)) \\
&\leq \frac{|E_1'|}{mt} + \frac{E_2'}{2\alpha\sqrt{T}} + \frac{\eta_{t-1}R^2}{t} + \frac{G^2}{2t} - \frac{1}{2t}\|q_{t+1}\|^2.
\end{aligned} \tag{12}$$

Notice that $\frac{1}{mt}\sum_{\tau=0}^{t-1}\sum_{i\in\Omega_\tau}(F(w_{\tau+1}) - F(w^\star)) \geq \frac{1}{2t}\sum_{\tau=0}^{t-1}(F(w_{\tau+1}) - F(w^\star))$. By Assumption 2, we have $F(w_{\tau+1}) + \langle\lambda^\star, g(w_{\tau+1})\rangle \geq F(w^\star)$ and thus,

$$\begin{aligned}
\frac{1}{2t}\sum_{\tau=0}^{t-1}(F(w_{\tau+1})-F(w^\star)) &\geq -\left\langle\lambda^\star, \frac{1}{2t}\sum_{\tau=0}^{t-1} g(w_{\tau+1})\right\rangle \\
&\geq -\frac{1}{2t}\|\lambda^\star\|\cdot\|q_{t+1}\|
\end{aligned}$$

where the second inequality is due to Lemma 11. Combining the inequality above with (12) shows $\|q_{t+1}\|^2 \leq \|\lambda^\star\|\cdot\|q_{t+1}\| + ①$, where $① := \frac{2|E_1'|}{m} + \frac{tE_2'}{\alpha\sqrt{T}} + 2\eta_{t-1}R^2 + G^2$. Solving this quadratic inequality in terms of $\|q_{t+1}\|$ yields

$$\|q_{t+1}\| \leq \tfrac{1}{2}\|\lambda^\star\| + \sqrt{\tfrac{1}{4}\|\lambda^\star\|^2 + ①}. \tag{13}$$

It is easy to verify that the probability bounds on $|E_1|$ and E_2 in Lemma 1 also work for $|E_1'|$ and E_2' for all $1 \leq t \leq T$. Using the inequality $\sqrt{x+y+z} \leq \sqrt{x} + \sqrt{y} + \sqrt{z}$ for $x, y, z \geq 0$, we obtain (7). It is clear from the initial $q_{j,1}$ in Algorithm 1 that (7) holds for $t = 0$. The second case has similar proof as above. We omit it due to the space limit.

E. Proof of Lemma 7

We show the first case using η_t given in (4). We prove it by induction. When $t = 0$, it is easy to verify $\eta_0 \leq \eta_1^{\max}$ by noting $\|q_1 + g(w_0)\| \leq \|q_1\| + \|g(w_0)\| \leq 2G$. Assume $\eta_{t-1} \leq \eta_1^{max}$. We need to show $\eta_t \leq \eta_1^{\max}$. By (4), it is enough to show that $L_g^2 + \beta_F + (q_{t+1} + g(w_t))^T\beta_g + \sqrt{T} \leq \eta_1^{\max}$. Notice $x^2 + y^2 + xy \leq (x+y)^2, \forall x, y \geq 0$.

$$\begin{aligned}
&L_g^2 + \beta_F + (q_{t+1} + g(w_t))^T\beta_g + \sqrt{T} \\
\leq\ & L_g^2 + \beta_F + \|q_{t+1}\|\|\beta_g\| + \|g(w_t)\|\|\beta_g\| + \sqrt{T} \\
\leq\ & \bar{\eta}_1 + \sqrt{T} + \sqrt{2\eta_1^{\max}}R\|\beta_g\| \\
=\ & \bar{\eta}_1 + \sqrt{T} + \sqrt{2}R\|\beta_g\|\sqrt{\bar{\eta}_1 + \sqrt{T}} + (\sqrt{2}R\|\beta_g\|)^2 \\
\leq\ & \eta_1^{\max}
\end{aligned}$$

where we use (7) in the second inequality. By induction, we conclude the proof. The second case has a similar proof.

REFERENCES

[1] J. Konecnỳ, H. Brendan McMahan, F. X. Yu, P. Richtárik, A. Theertha Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[2] D. Peteiro-Barral and B. Guijarro-Berdiñas, "A survey of methods for distributed machine learning," *Progress in Artificial Intelligence*, vol. 2, no. 1, pp. 1–11, 2013.

[3] M. Mahdavi, R. Jin, and T. Yang, "Trading regret for efficiency: online convex optimization with long term constraints," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2503–2528, 2012.

[4] H. Yu and M. J. Neely, "A simple parallel algorithm with an $O(1/t)$ convergence rate for general convex programs," *SIAM J. Optim.*, vol. 27, no. 2, pp. 759–783, 2017.

[5] ——, "A primal-dual parallel method with $O(1/\epsilon)$ convergence for constrained composite convex programs," *arXiv preprint arXiv:1708.00322*, 2017.

[6] H. Yu, M. Neely, and X. Wei, "Online convex optimization with stochastic constraints," in *Advances in Neural Information Processing Systems*, 2017, pp. 1428–1438.

[7] J. Yuan and A. Lamperski, "Online convex optimization for cumulative constraints," in *Advances in Neural Information Processing Systems*, 2018, pp. 6137–6146.

[8] X. Wei, H. Yu, and M. J. Neely, "Online primal-dual mirror descent under stochastic constraints," in *Abstracts of the 2020 SIGMETRICS/Performance Joint International Conference on Measurement and Modeling of Computer Systems*, 2020, pp. 3–4.

[9] H. Yu and M. J. Neely, "A low complexity algorithm with $O(\sqrt{T})$ regret and $O(1)$ constraint violations for online convex optimization with long term constraints," *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 1–24, 2020.

[10] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[11] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 2, p. 44, 2017.

[12] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*, 2018, pp. 5650–5659.

[13] L. Su and J. Xu, "Securing distributed gradient descent in high dimensional statistical learning," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 3, no. 1, 2019.

[14] D. Alistarh, Z. Allen-Zhu, and J. Li, "Byzantine stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2018, pp. 4618–4628.

[15] D. Ding, X. Wei, and M. R. Jovanović, "Distributed robust statistical learning: Byzantine mirror descent," in *Proceedings of the 58th IEEE Conference on Decision and Control*, Nice, France, 2019, pp. 1822–1827.

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

[17] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 2016.

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

[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.