Byzantine Fault-Tolerant Parallelized Stochastic Gradient Descent for Linear Regression

Nirupam Gupta and Nitin H. Vaidya

Department of Computer Science,
Georgetown University, Washington, DC 20057, USA
{*first-name*}.{*last-name*}@georgetown.edu

Abstract

This paper addresses the problem of Byzantine fault-tolerance in parallelized stochastic gradient descent (SGD) method for a synchronous master-workers system. The objective is to enable the master to solve for a linear regression problem, using the parallelized SGD method, despite the Byzantine nature of up to (known) constant number of workers [12]. To mitigate the detrimental effects of Byzantine faulty workers, we consider a gradient-filter, named *comparative gradient clipping* (CGC) filter, to robustify the gradient aggregation step of the parallelized SGD method. We show that the parallelized SGD with the CGC filter obtains parameter estimate, for the regression problem, of provably optimal accuracy in presence of up to a certain number of Byzantine faulty workers.

1 Introduction

We consider the problem of Byzantine fault-tolerance in parallelized stochastic gradient descent (SGD) for solving linear regression problem. The system consists of a master, n workers and a data set $\mathcal{Z}$ of m data points $\{Z_i \,|\, i = 1, \ldots, m\}$. Let $m > n$. Each data point Z_i is an ordered pair of d-dimensional row-vector X_i (referred as "independent variable") and scalar Y_i (referred as "dependent variable") that are related linearly in absence of noise. Specifically,

$$Y_i = X_i\, w^* + \xi_i, \quad i = 1, \ldots, m \tag{1}$$

where, $\{\xi_i \,|\, i = 1, \ldots, m\}$ are unknown noise of bounded magnitude, and w^* is the regression parameter that is to be determined.

For each data point $Z_i = (X_i, Y_i)$, for a given estimate w of w^*, we define a squared-error cost function

$$Q_i(w) \triangleq \frac{1}{2}\left(X_i\, w - Y_i\right)^2, \quad i = 1, \ldots, m \tag{2}$$

Following the standard approach of ordinary least squares (OLS) method [2], the parameter w^* can be estimated by minimizing the average squared-error cost over all the data points. Specifically, an estimate of w^* is obtained as

$$\arg\min \frac{1}{m} \sum_{i=1}^{m} Q_i(w) \tag{3}$$

We assume that w^* is unique, and that there exists a compact convex set $\mathcal{W}$ known to the master such that $w^* \in \mathcal{W}$. Note that $\mathcal{W}$ could just be a finite-size box of dimension same as w^*. In practice, such a set $\mathcal{W}$ is known apriori.

A typical (with no faulty workers) parallelized SGD method is an expedited version of SGD [5, 17]. The master starts with an estimate of w^* and iteratively updates it using information received from all the workers. The information sent by each worker is the gradient of the cost function (described above) for a randomly chosen data point in $\mathcal{Z}$ at the current estimate of w^*. The master takes the average of all the gradients received form the workers to update the current estimate [17]. Therefore, effectively, in parallelized SGD the master uses cost gradients of multiple (randomly chosen) data points, unlike SGD wherein only cost gradient of a single (randomly chosen) data point is used, to update its estimates. As the workload of computing cost gradients over a batch of data points is *distributed* amongst multiple workers, parallelized SGD is also sometimes referred to as distributed SGD [3].

1.1 Problem of Byzantine Failures

For the above parallelized SGD framework, we consider the case of Byzantine failures wherein up to f of the n workers are Byzantine faulty [12]. Byzantine faulty workers can send incorrect values for the gradients to the master, and their identity is hidden from the master. Moreover, the faulty workers can even collude. The traditional parallelized SGD – wherein the the master simply takes the average of the gradients reported by all the workers – fails to compute a good estimate of w^* even in presence of a single Byzantine faulty worker [3].

Our objective is to devise an aggregation rule for the master to make the parallelized SGD robust against Byzantine faulty workers. We note that it is impossible to solve this problem if $n \leq 2f$ [3]. Therefore, throughout this paper it is assumed that $n > 2f$.

1.2 Summary of Contribution

We consider a gradient filter, named Comparative Gradient Clipping (CGC), to "robustify" the aforementioned parallelized SGD method against Byzantine faulty workers. The CGC filter was proposed in our prior work [11] for fault-tolerance in distributed linear regression problem. However, in [11] we only considered the deterministic gradient descent method. In this paper we consider the stochastic variant of the gradient descent method which is better suited for regression over large-scale data [5].

In the CGC filter, in each iteration, the gradients with the largest f 2-norms are "clipped" so that their 2-norm equals the norm of the $(f+1)$-th largest gradient (or, equivalently, the $(n-f)$-th smallest gradient). The remaining gradients remain unchanged. The resulting gradients are then averaged to update the current estimate. We refer to the above filter as the *Comparative Gradient Clipping* filter, since the norms of the largest f gradients are clipped to a norm that is "comparable" to the next largest gradient. The formal description of the resultant SGD method with the CGC filter is presented in Section 2.

We have shown, in Section 3, that parallelized SGD with CGC filter obtains a "good" estimate of the regression parameter w^* if the fraction of Byzantine faulty workers are bounded. Moreover, the guaranteed upper bound on the estimation error only grows linearly in f/n. It is interesting

to note that the lower bound on the statistical error rate for distributed learning with Byzantine faulty workers is also linear in f/n [16].

1.3 Related Work

The problem of Byzantine faulty workers in parallelized SGD for machine learning has received considerable attention in recent years [1,3,6–10,14,16]. Unlike most works that rely on gradient filters, [6,9] propose coding schemes that increase the computational workload of the workers to guarantee recovery of correct gradients by the master in presence of Byzantine faulty workers. Whereas, gradient filters, such as ours, do not increase the computational workload of the workers but at the expense of fault-tolerance. In comparison to the gradient filters proposed in [1, 3, 7, 8, 10, 14], the proposed CGC filter is computationally simpler and achieves comparable fault-tolerance. Unlike coordinate-wise trimmed mean and coordinate-wise median filters [16], the fault-tolerance property of the proposed CGC filter does not rely on any assumption on the probability distribution of the data points.

In the past, gradient clipping has been proposed for solving the problem of gradient explosion in training of neural networks [13]. However, the gradient clipping in [13] is threshold based, that is, the gradients are clipped when their norms exceed a constant *threshold*. A similar threshold-based gradient clipping has also been used for improving the differential privacy-accuracy trade-off in distributed stochastic gradient based deep learning [15]. On the other hand, our proposed gradient clipping is *comparative* and does not require any empirical learning of the threshold.

2 Parallelized SGD With Proposed Gradient-Filter

The master starts with an arbitrary estimate $w^0 \in \mathcal{W}$, and updates it iteratively as follows. Let $t \geq 0$ be the iteration index and w^t denote the estimate after t-th iteration.

Steps performed in the t-th iteration are as follows:

S1: The master requests from each worker j the cost gradient for any randomly chosen data point in $\mathcal{Z}$ at the current estimate w^t.

The gradient received by the master from worker j is denoted as g_j^t. If no gradient is received from a particular worker (which must be faulty), then the server assumes a default value for the missing gradient (specifically, $\mathbf{0}$ vector).

S2: Comparative Gradient Clipping (CGC): For a vector v, let $\|v\|$ denote its 2-norm. The master sorts the received gradients as follows,

$$\left\|g_{j_1}^t\right\| \leq \ldots \leq \left\|g_{j_{n-f}}^t\right\| \leq \left\|g_{j_{n-f+1}}^t\right\| \leq \ldots \leq \left\|g_{j_n}^t\right\|$$

Thus, the gradient with the smallest norm, $g_{j_1}^t$, is received from agent j_1, and the gradient with the largest norm, $g_{j_n}^t$, is received from agent j_n.

If $\left\|g_{j_{n-f}}^t\right\| = \mathbf{0}$, then the algorithm terminates and outputs the current value of the estimate

at the master. Otherwise, the master computes "scaled" gradients, $\widehat{g}_k^t$, as

$$\widehat{g}_k^t = \begin{cases} \frac{\left\|g_{j_{n-f}}^t\right\|}{\left\|g_k^t\right\|} g_k^t & , \quad k \in \{j_{n-f+1}, \ldots, j_n\} \\ g_k^t & , \quad k \in \{j_1, \ldots, j_{n-f}\} \end{cases} \tag{4}$$

and updates the estimate,

$$w^{t+1} = \left[w^t - \eta_t \cdot \frac{1}{n} \sum_{k=1}^{n} \widehat{g}_k^t \right]_{\mathcal{W}} \tag{5}$$

where η_t is the step-size, and $[\,\cdot\,]_{\mathcal{W}}$ denotes the Euclidean projection onto $\mathcal{W}$, i.e.

$$[w]_{\mathcal{W}} = \arg\min_{v \in \mathcal{W}} \|w - v\|, \; \forall w \in \mathbb{R}^d$$

3 Fault-Tolerance Property

In this section, we present the fault-tolerance property of the proposed algorithm under the following assumption.

Assumption 1: Assume that matrix $X = [X_1^T, \ldots, X_m^T]^T$ has rank equal to d. This assumption implies that the cost function $\sum_i Q_i(w)$ is strongly convex and that w^* is unique.

To be able to present the results we introduce the following notation.

- For a scalar value s, let $|s|$ denote its absolute value. Then, let $\xi = \max_{i=1}^m |\xi_i|$.
- Let μ denote the maximum value of $\|X_i\|^2$ for $i = 1, \ldots, m$.
- Let $(\cdot)^T$ denote the transpose. Then, $\lambda = \nu/m$, where ν denotes the smallest eigenvalue of matrix $X^T X$. Note that under Assumption 1, $\nu > 0$.
- Let
$$\rho(f) = 1 - \frac{f}{n}\left(1 + \frac{2\mu}{\lambda}\right)$$
- For a probabilistic event $\mathcal{E}$, let $\mathsf{Prob}(\mathcal{E})$ denote its probability.

Theorem 1. *Suppose that Assumption 1 holds. Consider the algorithm described in Section 2, with η_t in (5) satisfying: $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$. If,*

$$\rho(f) > 0 \tag{6}$$

then for,

$$\mathsf{Prob}\left(\lim_{t \to \infty} \left\|w^t - w^*\right\| \leq \left(1 + \frac{2f}{n}\right)\left(\frac{1}{\rho(f)}\right)\left(\frac{\sqrt{\mu}}{\lambda}\right)\xi \right) = 1$$

According to the theorem, the parallelized SGD with the proposed CGC filter converges *almost surely* to a neighborhood of w^*, whose size is directly proportional to ξ (magnitude of noise), if the

fraction of faulty workers f/n is sufficiently small.

If (6) holds then there exists a non-negative constant $\rho_o > 0$ such that

$$\rho(f) \geq \rho_o,$$

Therefore, according to Theorem 1, the estimation accuracy obtained by the CGC filter is (almost surely) less than or equal to

$$\left(1+\frac{2f}{n}\right)\left(\frac{1}{\rho(f)}\right)\left(\frac{\sqrt{\mu}}{\lambda}\right)\xi \leq \left(\frac{\sqrt{\mu}}{\rho_o\,\lambda}\right)\,\xi + \left(\frac{2\sqrt{\mu}}{\rho_o\,\lambda}\right)\,\left(\frac{f}{n}\right)\xi$$

In other words, the estimation accuracy obtained by the CGC filter for sufficiently small f/n is $\mathcal{O}\left(\xi + (f/n)\xi\right)$.

3.1 Proof of Theorem 1

To begin with, let us introduce the following notation.

- Let random variable $\mathcal{F}_t \triangleq \{w^0, \ldots, w^t\}$ denote the history of estimates till $(t+1)$-th iteration.
- Let $\mathfrak{g}^t$ denote $\sum_{k=1}^n \widehat{g}_k^t$. Then, we define

$$\phi_t \triangleq \left\langle w^t - w^*,\, \mathfrak{g}^t \right\rangle$$

- For two random variables V and F, let $\mathbb{E}(V \,|\, F)$ denote the expected value of V given the value of F.

The proof relies on a sufficient criterion for global confinement of a stochastic process with bounded variance [4, Section 5.2]. The criterion is stated as follows:

Lemma 1 (Ref. [4])**.** *Consider the iterative process* (5)*. Suppose that* $\sum_{t=0}^{\infty}\eta_t = \infty$ *and* $\sum_{t=0}^{\infty}\eta_t^2 < \infty$*. If* $\mathbb{E}\left(\left\|\mathfrak{g}^t\right\| \,|\, \mathcal{F}_t\right)$ *is bounded for all* t*, and there exists* $\mathsf{D}^* \in [0,\, \max_{w\in\mathcal{W}}\|w - w^*\|)$ *such that*

$$\mathbb{E}\left(\phi_t \,|\, \mathcal{F}_t\right) > 0 \text{ when } \left\|w^t - w^*\right\| > \mathsf{D}^*$$

then

$$\mathsf{Prob}\left(\lim_{t\to\infty}\left\|w^t - w^*\right\| \leq \mathsf{D}^*\right) = 1$$

Next, we show that the sufficient criterion in Lemma 1 holds under condition (6) and Assumption 1, for

$$\mathsf{D}^* = \left(1+\frac{2f}{n}\right)\,\left(\frac{1}{\rho(f)}\right)\,\left(\frac{\sqrt{\mu}}{\lambda}\right)\,\xi \tag{7}$$

For the proof, we use the following properties of the cost functions.

Some useful properties:
From (1) and (2),

$$\nabla Q_i(w) = X_i^T X_i(w - w^*) - X_i^T \xi_i \tag{8}$$

Using triangle inequality,

$$\|\nabla Q_i(w)\| \leq \left\|X_i^T X_i(w - w^*)\right\| + |\xi_i| \, \|X_i\| = |\langle X_i, \, w - w^* \rangle| \, \|X_i\| + |\xi_i| \, \|X_i\|$$

From Cauchy-Schwartz inequality, $|\langle X_i, \, w - w^* \rangle| \leq \|X_i\| \, \|w - w^*\|$. Therefore,

$$\|\nabla Q_i(w)\| \leq \|X_i\|^2 \, \|w - w^*\| + |\xi_i| \, \|X_i\|$$

Recall, $\xi = \max_i |\xi_i|$ and $\mu = \max_i \|X_i\|^2$. Therefore,

$$\boxed{\|\nabla Q_i(w)\| \leq \mu \, \|w - w^*\| + \sqrt{\mu} \, \xi, \quad \forall w \in \mathcal{W}, \, \forall i} \tag{9}$$

Let $C(w) \triangleq (1/m) \sum_{i=1}^m Q_i(w)$ be the average squared-error cost at estimate w. Then

$$\nabla C(w) = \frac{1}{m} \sum_{i=1}^m \nabla Q_i(w)$$

From (8),

$$\nabla C(w) = \frac{1}{m} \sum_{i=1}^m X_i^T X_i(w - w^*) - \frac{1}{m} \sum_{i=1}^m X_i^T \xi_i$$

As $X^T X = \sum_{i=1}^m X_i^T X_i$, from above we obtain,

$$\nabla C(w) = \frac{1}{m} X^T X(w - w^*) - \frac{1}{m} \sum_{i=1}^m X_i^T \xi_i \tag{10}$$

This implies,

$$\langle w - w^*, \, \nabla C(w) \rangle = \frac{1}{m} (w - w^*)^T X^T X(w - w^*) - \frac{1}{m} \sum_{i=1}^m \left\langle w - w^*, \, X_i^T \xi_i \right\rangle \tag{11}$$

From Cauchy Schwartz inequality,

$$\left\langle w - w^*, \, X_i^T \xi_i \right\rangle \leq \|w - w^*\| \, \left\|X_i^T \xi_i\right\| = |\xi_i| \, \|X_i\| \, \|w - w^*\| \, , \quad \forall i$$

Recall that $\xi = \max_i |\xi_i|$ and $\mu = \max_i \|X_i\|^2$. Therefore,

$$\left\langle w - w^*, \, X_i^T \xi_i \right\rangle \leq \xi \sqrt{\mu} \, \|w - w^*\| \, , \quad \forall i \tag{12}$$

Under Assumption 1, $X^T X$ is a positive definite matrix. Recall that ν denotes the smallest eigenvalue of $X^T X$ and $\lambda = \nu / m$. Therefore,

$$(w - w^*)^T X^T X(w - w^*) \geq \nu \, \|w - w^*\|^2 = m\lambda \, \|w - w^*\|^2 \tag{13}$$

Substituting (12) and (13) implies,

$$\langle w - w^*,\, \nabla C(w)\rangle \geq \lambda \left\|w - w^*\right\|^2 - \xi\sqrt{\mu}\left\|w - w^*\right\|, \quad \forall w \in \mathcal{W} \tag{14}$$

The rest of the proof is divided into 2 steps:

1. In the first step, we show that $\mathbb{E}\left(\left\|g^t\right\| \mid \mathcal{F}_t\right)$ is bounded for all t.
2. In the second step, we show that if condition (6) holds then $\mathbb{E}(\phi_t \mid \mathcal{F}_t) > 0$ when $\left\|w^t - w^*\right\| > \mathsf{D}^*$, for D^* as given in (7).

Let $\mathcal{H}$ and $\mathcal{B}$ denote the set of honest and Byzantine faulty workers, respectively.

Step 1: Note that

$$\left\|\widehat{g_k^t}\right\| \leq \left\|g_{j_{n-f}}^t\right\|, \quad \forall k \in \{1, \ldots, n\},\, t \tag{15}$$

As there are at most f Byzantine faulty workers, for every t there exists $\sigma \in \mathcal{H}$ such that

$$\left\|g_{j_{n-f}}^t\right\| \leq \left\|g_\sigma^t\right\| \tag{16}$$

For an honest worker $j \in \mathcal{H}$, let j_t denotes the index of the data point chosen by j for $(t+1)$-th iteration. Then,

$$g_j^t = \nabla Q_{j_t}(w^t), \quad \forall j \in \mathcal{H}$$

Therefore,

$$\left\|\widehat{g_k^t}\right\| \leq \left\|\nabla Q_{\sigma_t}(w^t)\right\|, \quad \forall k,\, t \tag{17}$$

From (9),

$$\left\|\nabla Q_i(w^t)\right\| \leq \mu\, \left\|w^t - w^*\right\| + \sqrt{\mu}\,\xi, \quad \forall i \in \{1, \ldots, m\} \tag{18}$$

From (17) and (18),

$$\left\|\widehat{g_k^t}\right\| \leq \mu\, \left\|w^t - w^*\right\| + \sqrt{\mu}\,\xi, \quad \forall k,\, t$$

Therefore,

$$\mathbb{E}\left(\left\|\widehat{g_k^t}\right\| \middle|\, \mathcal{F}_t\right) \leq \mu\, \left\|w^t - w^*\right\| + \sqrt{\mu}\,\xi, \quad \forall k,\, t \tag{19}$$

Recall that

$$\mathfrak{g}^t = \sum_{k=1}^{n} \widehat{g_k^t}$$

From triangle inequality,

$$\left\|\mathfrak{g}^t\right\| \leq \sum_{k=1}^{n} \left\|\widehat{g_k^t}\right\|$$

The above implies,

$$\mathbb{E}\left(\left\|\mathfrak{g}^t\right\| \middle|\, \mathcal{F}_t\right) \leq n\left(\mu\, \left\|w^t - w^*\right\| + \sqrt{\mu}\,\xi\right), \quad \forall t \tag{20}$$

As $w^t \in \mathcal{W}$, $\forall t$ and $\mathcal{W}$ is a compact set, there exists

$$\Gamma = \max_{w \in \mathcal{W}} \|w - w^*\| < \infty$$

Thus, $\left\|w^t - w^*\right\| \leq \Gamma$, $\forall t \in \mathbb{Z}_{\geq 0}$. Therefore, from (20) we obtain,

$$\mathbb{E}\left(\left\|\mathfrak{g}^t\right\| \middle| \mathcal{F}_t\right) \leq n\left(\mu\,\Gamma + \sqrt{\mu}\,\xi\right) < \infty, \quad \forall t \tag{21}$$

In other words, $\mathbb{E}\left(\left\|\mathfrak{g}^t\right\| \middle| \mathcal{F}_t\right)$ is bounded for all t.

Step 2: Recall that

$$\phi_t = \left\langle w^t - w^*, \sum_{k=1}^{n} \widehat{g_k^t} \right\rangle, \quad \forall t \in \mathbb{Z}_{\geq 0} \tag{22}$$

For a finite set S, let $|S|$ denote its cardinality. Then, for every iteration t, there exists $\mathcal{H}_1^t \subset \mathcal{H}$ such that $\left|\mathcal{H}_1^t\right| = n - 2f$ and $\mathcal{H}_1^t \subset \{j_1, \ldots, j_{n-f}\}$. Let $[n]$ denote the set $\{1, \ldots, n\}$, then

$$\phi_t = \sum_{k \in \mathcal{H}_1^t} \left\langle w^t - w^*, \widehat{g_k^t} \right\rangle + \sum_{l \in [n] \setminus \mathcal{H}_1^t} \left\langle w^t - w^*, \widehat{g_l^t} \right\rangle$$

As $\widehat{g_k^t} = g_k^t$, $\forall k \in \{j_1, \ldots, j_{n-f}\}$,

$$\phi_t = \sum_{k \in \mathcal{H}_1^t} \left\langle w^t - w^*, g_k^t \right\rangle + \sum_{l \in [n] \setminus \mathcal{H}_1^t} \left\langle w^t - w^*, \widehat{g_l^t} \right\rangle$$

Let $k_t \in \{1, \ldots, m\}$ denote the index of the data point chosen by worker $k \in \mathcal{H}$ for $(t+1)$-th iteration. Then,

$$g_k^t = \nabla Q_{k_t}(w^t), \quad \forall k \in \mathcal{H}$$

Therefore,

$$\phi_t = \sum_{k \in \mathcal{H}_1^t} \left\langle w^t - w^*, \nabla Q_{k_t}(w^t) \right\rangle + \sum_{l \in [n] \setminus \mathcal{H}_1^t} \left\langle w^t - w^*, \widehat{g_l^t} \right\rangle$$

Let $\mathcal{H}_2^t = \mathcal{H} \setminus \mathcal{H}_1^t$. As $\mathcal{B} = [n] \setminus \mathcal{H}$,

$$\phi_t = \sum_{k \in \mathcal{H}_1^t} \left\langle w^t - w^*, \nabla Q_{k_t}(w^t) \right\rangle + \sum_{k \in \mathcal{H}_2^t} \left\langle w^t - w^*, \widehat{g_k^t} \right\rangle + \sum_{l \in \mathcal{B}} \left\langle w^t - w^*, \widehat{g_l^t} \right\rangle$$

Alternately,

$$\begin{aligned}\phi_t = &\sum_{k \in \mathcal{H}_1^t} \left\langle w^t - w^*, \nabla Q_{k_t}(w^t) \right\rangle + \sum_{k \in \mathcal{H}_2^t} \left\langle w^t - w^*, g_k^t \right\rangle - \sum_{k \in \mathcal{H}_2^t} \left\langle w^t - w^*, g_k^t \right\rangle \\ &+ \sum_{k \in \mathcal{H}_2^t} \left\langle w^t - w^*, \widehat{g_k^t} \right\rangle + \sum_{l \in \mathcal{B}} \left\langle w^t - w^*, \widehat{g_l^t} \right\rangle\end{aligned}$$

Similar to $k \in \mathcal{H}_1^t$, substitute $g_k^t = \nabla Q_{k_t}(w^t)$, $\forall k \in \mathcal{H}_2^t$. Therefore,

$$\phi_t = \sum_{k \in \mathcal{H}} \left\langle w^t - w^*, \nabla Q_{k_t}(w^t) \right\rangle + \sum_{k \in \mathcal{H}_2^t} \left\langle w^t - w^*, \widehat{g_k^t} - \nabla Q_{k_t}(w^t) \right\rangle + \sum_{l \in \mathcal{B}} \left\langle w^t - w^*, \widehat{g_l^t} \right\rangle \tag{23}$$

From (8),

$$\left\langle w^t - w^*,\, \nabla Q_{k_t}(w^t) \right\rangle = \left(X_{k_t}(w^t - w^*)\right)^2 - \xi_{k_t}\, X_{k_t}(w^t - w^*), \quad \forall k \in \mathcal{H} \tag{24}$$

Define,

$$\alpha_k^t = \min\left\{1,\, \frac{\left\|g_{j_{n-f}}^t\right\|}{\|g_k^t\|}\right\}, \quad \forall k \in [n] \tag{25}$$

Then,

$$\widehat{g_k^t} = \alpha_k^t\, g_k^t, \quad \forall k \in [n]$$

Thus, for every $k \in \mathcal{H}_2^t$,

$$\left\langle w^t - w^*,\, \widehat{g_k^t} \right\rangle = \alpha_k^t \left\langle w^t - w^*,\, \nabla Q_{k_t}(w^t) \right\rangle$$

Substituting (24) above, we obtain,

$$\left\langle w^t - w^*,\, \widehat{g_k^t} \right\rangle = \alpha_k^t \left(X_{k_t}(w^t - w^*)\right)^2 - \alpha_k^t\, \xi_{k_t}\, X_{k_t}(w^t - w^*)$$

As $\alpha_k^t \geq 0$, from above

$$\left\langle w^t - w^*,\, \widehat{g_k^t} \right\rangle \geq -\alpha_k^t\, \xi_{k_t}\, X_{k_t}(w^t - w^*), \quad \forall k \in \mathcal{H}_2^t \tag{26}$$

From (24) and (26),

$$\left\langle w^t - w^*,\, \widehat{g_k^t} - \nabla Q_{k_t}(w^t) \right\rangle \geq -\left(X_{k_t}(w^t - w^*)\right)^2 + (1 - \alpha_k^t)\, \xi_{k_t}\, X_{k_t}(w^t - w^*), \quad \forall k \in \mathcal{H}_2^t \tag{27}$$

As $0 \leq \alpha_k^t \leq 1$,

$$\left|(1 - \alpha_k^t)\, \xi_{k_t}\, X_{k_t}(w^t - w^*)\right| \leq (1 - \alpha_k^t)\, |\xi_{k_t}|\, \left|X_{k_t}(w^t - w^*)\right| \leq |\xi_{k_t}|\, \left|X_{k_t}(w^t - w^*)\right|$$

Recall that $|\xi_i| \leq \xi$, and from Cauchy-Schwartz inequality

$$\left|X_i(w^t - w^*)\right| \leq \|X_i\| \left\|w^t - w^*\right\| \leq \sqrt{\mu} \left\|w^t - w^*\right\|, \forall i \in \{1, \ldots, m\}.$$

Therefore,

$$\left|(1 - \alpha_k^t)\, \xi_{k_t}\, X_{k_t}(w^t - w^*)\right| \leq \sqrt{\mu}\, \xi\, \left\|w^t - w^*\right\|, \quad \forall k \in \mathcal{H}_2^t$$

Using this in (27) implies that

$$\left\langle w^t - w^*,\, \widehat{g_k^t} - \nabla Q_{k_t}(w^t) \right\rangle \geq -\left(X_{k_t}(w^t - w^*)\right)^2 - \sqrt{\mu}\, \xi\, \left\|w^t - w^*\right\|, \quad \forall k \in \mathcal{H}_2^t$$

Substituting this in (23), we obtain

$$\phi_t \geq \sum_{k \in \mathcal{H}} \left\langle w^t - w^*,\, \nabla Q_{k_t}(w^t) \right\rangle - \sum_{k \in \mathcal{H}_2^t} \left(X_{k_t}(w^t - w^*)\right)^2 - \left|\mathcal{H}_2^t\right| \sqrt{\mu} \xi\, \left\|w^t - w^*\right\| + \sum_{k \in \mathcal{B}} \left\langle w^t - w^*,\, \widehat{g_k^t} \right\rangle$$

As $\|X_i\|^2 \leq \mu$, $\forall i \in \{1, \ldots, m\}$,

$$\phi_t \geq \sum_{k \in \mathcal{H}} \left\langle w^t - w^*,\, \nabla Q_{k_t}(w^t) \right\rangle - \mu\, \left|\mathcal{H}_2^t\right|\, \left\|w^t - w^*\right\|^2 - \left|\mathcal{H}_2^t\right| \sqrt{\mu}\, \xi\, \left\|w^t - w^*\right\| + \sum_{l \in \mathcal{B}} \left\langle w^t - w^*,\, \widehat{g_k^t} \right\rangle \tag{28}$$

From Cauchy-Schwartz inequality,

$$\left\langle w^t - w^*,\, \widehat{g}_l^t \right\rangle \geq - \left\| w^t - w^* \right\| \left\| \widehat{g}_l^t \right\|$$

Substituting this in (28), we obtain,

$$\phi_t \geq \sum_{k \in \mathcal{H}} \left\langle w^t - w^*,\, \nabla Q_{k_t}(w^t) \right\rangle - \mu \left| \mathcal{H}_2^t \right| \left\| w^t - w^* \right\|^2 - \left| \mathcal{H}_2^t \right| \sqrt{\mu}\, \xi \left\| w^t - w^* \right\| - \sum_{l \in \mathcal{B}} \left\| w^t - w^* \right\| \left\| \widehat{g}_l^t \right\|$$

By taking conditional expectation on sides for given $\mathcal{F}_t$, we obtain

$$\begin{aligned} \mathbb{E}\left(\phi_t \,|\, \mathcal{F}_t\right) \geq & \sum_{k \in \mathcal{H}} \left\langle w^t - w^*,\, \mathbb{E}\left(\nabla Q_{k_t}(w^t) \middle| \mathcal{F}_t\right) \right\rangle - \mu \left| \mathcal{H}_2^t \right| \left\| w^t - w^* \right\|^2 - \left| \mathcal{H}_2^t \right| \sqrt{\mu}\, \xi \left\| w^t - w^* \right\| \\ & - \sum_{l \in \mathcal{B}} \left\| w^t - w^* \right\| \mathbb{E}\left(\left\| \widehat{g}_l^t \right\| \middle| \, \mathcal{F}_t \right) \end{aligned}$$

Using (19) above implies,

$$\begin{aligned} \mathbb{E}\left(\phi_t \,|\, \mathcal{F}_t\right) \geq & \sum_{k \in \mathcal{H}} \left\langle w^t - w^*,\, \mathbb{E}\left(\nabla Q_{k_t}(w^t) \middle| \mathcal{F}_t\right) \right\rangle - \mu \left| \mathcal{H}_2^t \right| \left\| w^t - w^* \right\|^2 - \left| \mathcal{H}_2^t \right| \sqrt{\mu}\, \xi \left\| w^t - w^* \right\| \\ & - |\mathcal{B}| \left\| w^t - w^* \right\| \left(\mu \left\| w^t - w^* \right\| + \sqrt{\mu}\, \xi \right) \end{aligned}$$

As $\mathcal{B}$ and $\mathcal{H}_2^t$ are two disjoint sets satisfying $\mathcal{B} \bigcup \mathcal{H}_2^t = [n] \setminus \mathcal{H}_1^t$ and $\left| \mathcal{H}_1^t \right| = n - 2f$, $|\mathcal{B}| + \left| \mathcal{H}_2^t \right| = 2f$. Thus, from above,

$$\mathbb{E}\left(\phi_t \,|\, \mathcal{F}_t\right) \geq \sum_{k \in \mathcal{H}} \left\langle w^t - w^*,\, \mathbb{E}\left(\nabla Q_{k_t}(w^t) \middle| \mathcal{F}_t\right) \right\rangle - 2f\, \mu \left\| w^t - w^* \right\|^2 - 2f\, \sqrt{\mu}\, \xi \left\| w^t - w^* \right\|$$

As $\mathcal{H}$ remains unchanged over t, and k_t is an independent uniform random variable in $\{1, \ldots, m\}$ for every t,

$$\sum_{k \in \mathcal{H}} \mathbb{E}\left(\nabla Q_{k_t}(w^t) \middle| \mathcal{F}_t\right) = |\mathcal{H}|\, \nabla C(w^t)$$

Therefore,

$$\mathbb{E}\left(\phi_t \,|\, \mathcal{F}_t\right) \geq |\mathcal{H}| \left\langle w^t - w^*,\, \nabla C(w^t) \right\rangle - 2f\, \mu \left\| w^t - w^* \right\|^2 - 2f\, \sqrt{\mu}\, \xi \left\| w^t - w^* \right\|$$

Now, using (14) above implies,

$$\mathbb{E}\left(\phi_t \,|\, \mathcal{F}_t\right) \geq |\mathcal{H}|\, \lambda \left\| w^t - w^* \right\|^2 - |\mathcal{H}| \sqrt{\mu}\, \xi \left\| w^t - w^* \right\| - 2f\, \mu \left\| w^t - w^* \right\|^2 - 2f\, \sqrt{\mu}\, \xi \left\| w^t - w^* \right\|$$

Recall that $n - f \leq |\mathcal{H}| \leq n$. Thus,

$$\begin{aligned} \mathbb{E}\left(\phi_t \,|\, \mathcal{F}_t\right) & \geq (n-f)\, \lambda \left\| w^t - w^* \right\|^2 - n \sqrt{\mu}\, \xi \left\| w^t - w^* \right\| - 2f\, \mu \left\| w^t - w^* \right\|^2 - 2f\, \sqrt{\mu}\, \xi \left\| w^t - w^* \right\| \\ & = \left(n\lambda - f\left(\lambda + 2\mu\right)\right) \left\| w^t - w^* \right\|^2 - n\, \sqrt{\mu}\, \xi \left\| w^t - w^* \right\| - 2f\, \sqrt{\mu}\, \xi \left\| w^t - w^* \right\| \\ & = \left(n\lambda - f\left(\lambda + 2\mu\right)\right) \left\| w^t - w^* \right\|^2 - (n + 2f)\, \sqrt{\mu}\, \xi \left\| w^t - w^* \right\| \\ & = \left(n\lambda - f\left(\lambda + 2\mu\right)\right) \left\| w^t - w^* \right\| \left\{ \left\| w^t - w^* \right\| - \left(\frac{(n+2f)\sqrt{\mu}}{n\lambda - f\left(\lambda + 2\mu\right)} \right) \xi \right\} \end{aligned} \tag{29}$$

Inequality (29) implies that if (6) holds, i.e.

$$\rho\left(f\right) = 1 - \frac{f}{n}\left(1 + \frac{2\mu}{\lambda}\right) > 0 \iff n\lambda - f\left(\lambda + 2\mu\right) > 0,$$

then $\mathbb{E}\left(\phi_t \mid \mathcal{F}_t\right) > 0$ when

$$\left\|w^t - w^*\right\| > \frac{(n+2f)\sqrt{\mu}}{n\lambda - f\left(\lambda + 2\mu\right)}\,\xi \tag{30}$$

Note that

$$\frac{(n+2f)\sqrt{\mu}}{n\lambda - f\left(\lambda + 2\mu\right)} = \left(1 + \frac{2f}{n}\right)\left(\frac{1}{\rho(f)}\right)\frac{\sqrt{\mu}}{\lambda}$$

The rest of the proofs follows from Lemma 1.

4 Summary

We present a robust gradient aggregation rule for the parallelized SGD method to mitigate the detrimental effects of Byzantine faulty workers in a synchronous master-workers system. The main component of the proposed robust aggregation rule is a gradient filter, which is referred to as comparative gradient clipping (CGC) filter. We have shown that the resultant parallelized SGD method solves the linear regression problem even in presence of up to a certain number of Byzantine faulty workers. Moreover, we have shows that the obtained bound on the estimation error is linear in f/n.

Acknowledgements

Research reported in this paper was sponsored in part by the Army Research Laboratory under Cooperative Agreement W911NF- 17-2-0196, and by National Science Foundation award 1610543. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the the Army Research Laboratory, National Science Foundation or the U.S. Government.

References

[1] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4618–4628, 2018.

[2] Takeshi Amemiya. *Advanced econometrics*. Harvard University press, 1985.

[3] Peva Blanchard, Rachid Guerraoui, Julien Stainer, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, pages 119–129, 2017.

[4] Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

[5] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[6] Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Draco: Byzantine-resilient distributed training via redundant gradients. In *International Conference on Machine Learning*, pages 903–912, 2018.

[7] Yuan Chen, Soummya Kar, and Jose MF Moura. Resilient distributed estimation through adversary detection. *IEEE Transactions on Signal Processing*, 66(9):2455–2469, 2018.

[8] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *ACM SIGMETRICS Performance Evaluation Review*, 46(1):96–96, 2019.

[9] Deepesh Data, Linqi Song, and Suhas Diggavi. Data encoding for Byzantine-resilient distributed gradient descent. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 863–870. IEEE, 2018.

[10] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

[11] Nirupam Gupta and Nitin H Vaidya. Byzantine fault tolerant distributed linear regression. *arXiv preprint arXiv:1903.08752*, 2019.

[12] Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 4(3):382–401, 1982.

[13] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2, 2012.

[14] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[15] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321. ACM, 2015.

[16] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5636–5645, 2018.

[17] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.