Variance Reduction is an Antidote to Byzantines: Better Rates, Weaker Assumptions and Communication Compression as a Cherry on the Top

Eduard GorbunovSamuel HorváthPeter RichtárikGauthier GidelMBZUAIMBZUAIKAUSTMila & UdeMCanada CIFAR AI Chair

MBZUAI Workshop on Collaborative Learning: From Theory to Practice





October 8, 2022

**PDF** Gorbunov, E., Horváth, S., Richtárik, P., & Gidel, G. (2022). Variance Reduction is an Antidote to Byzantines: Better Rates, Weaker Assumptions and Communication Compression as a Cherry on the Top. *arXiv preprint arXiv:2206.00529*.



Samuel Horváth Assistant professor at MBZUAI



Peter Richtárik Professor at KAUST



Gauthier Gidel Assistant professor at Mila and UdeM

### Outline

1. Byzantine-robust training

2. Robust aggregation

3. Variance reduction and Byzantine-robustness

# **Byzantine-Robust Training**









### Parallel SGD



### Parallel SGD



10

## Parallel SGD



## Parallel SGD Is Fragile



## **The Refined Problem Formulation**



$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{G} \sum_{i \in \mathcal{G}} f_i(x) \right\}$$

#### Good workers form the majority:

- G good workers
- *B* Byzantines (see the page "Byzantine fault" in Wikipedia)
- $\mathcal{G} \sqcup \mathcal{B} = [n], |\mathcal{G}| = G, |\mathcal{B}| = B$
- $B \leq \delta n$ ,  $\delta < 1/2$
- Byzantines are omniscient

## **The Refined Problem Formulation**



$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{G} \sum_{i \in \mathcal{G}} f_i(x) \right\}$$

### Good workers form the majority:

- G good workers
- *B* Byzantines (see the page "Byzantine fault" in Wikipedia)
- $\mathcal{G} \sqcup \mathcal{B} = [n], |\mathcal{G}| = G, |\mathcal{B}| = B$
- $B \leq \delta n$ ,  $\delta < 1/2$
- Byzantines are omniscient

### On the heterogeneity:

- Loss functions on good peers cannot be arbitrary heterogeneous
- In this talk, we will assume that  $\forall i \in \mathcal{G} \rightarrow f_i = f$

## **The Refined Problem Formulation**



**Question:** how to solve such problems?

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{G} \sum_{i \in \mathcal{G}} f_i(x) \right\}$$

#### Good workers form the majority:

- G good workers
- *B* Byzantines (see the page "Byzantine fault" in Wikipedia)
- $\mathcal{G} \sqcup \mathcal{B} = [n], |\mathcal{G}| = G, |\mathcal{B}| = B$
- $B \leq \delta n$ ,  $\delta < 1/2$
- Byzantines are omniscient

### On the heterogeneity:

- Loss functions on good peers cannot be arbitrary heterogeneous
- In this talk, we will assume that  $\forall i \in \mathcal{G} \rightarrow f_i = f$

**Robust Aggregation** 

**Natural idea:** replace the averaging with more robust aggregation rule!

$$x^{k+1} = x^k - \gamma g^k \quad \Longrightarrow \quad x^{k+1} = x^k - \gamma \widehat{g}^k$$
$$g^k = \frac{1}{n} \sum_{i=1}^n g_i^k \quad \Longrightarrow \quad \widehat{g}^k = \operatorname{RAgg}\left(g_1^k, g_2^k, \dots, g_n^k\right)$$

**Question:** how to choose aggregator?

• Geometric median (RFA): Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2019). Robust aggregation for federated learning. arXiv preprint arXiv:1912.13445.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^n \|g - g_i^k\|_2$$

• Geometric median (RFA):

PILLUTIA, K., Kakade, S. M., & Harchaoui, Z. (2019). Robust aggregation for federated learning. arXiv preprint arXiv:1912.13445.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^n \|g - g_i^k\|_2$$

• Coordinate-wise median (CM): Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018, July). Byzantine-robust distributed learning: Towards optimal statistical rates. *In International* 

*Conference on Machine Learning* (pp. 5650-5659). PMLR.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^n \|g - g_i^k\|_1$$

• Geometric median (RFA):

Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2019). Robust aggregation for federated learning. arXiv preprint arXiv:1912.13445.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^n \|g - g_i^k\|_2$$

Coordinate-wise median (CM):
Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018, July). Byzantine-robust distributed learning: Towards optimal statistical rates. *In International Conference on Machine Learning* (pp. 5650-5659). PMLR.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^n \|g - g_i^k\|_1$$

#### • Krum estimator:

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017, December). Machine learning with adversaries: Byzantine tolerant gradient descent. *In Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 118-128).

$$\hat{g}^{k} = \underset{g \in \{g_{1}^{k}, \dots, g_{n}^{k}\}}{\arg\min} \sum_{i \in \mathcal{N}_{n-B-2}(g)} \left\| g - g_{i}^{k} \right\|_{2}^{2}$$

indices of the closest n - B - 2 workers to g

### Simple Example When "Middle-Seekers" Are Good

Let  $d = 1, G = \{1, 2, 3, 4\}, B = \{5, 6\}, g_1^k = 1.5, g_2^k = 2, g_3^k = 2.5, g_4^k = 3$ , and Byzantines are trying to shift the estimator via sending  $g_5^k = g_6^k = 1000$ . In this case,

### Simple Example When "Middle-Seekers" Are Good

Let  $d = 1, G = \{1, 2, 3, 4\}, B = \{5, 6\}, g_1^k = 1.5, g_2^k = 2, g_3^k = 2.5, g_4^k = 3$ , and Byzantines are trying to shift the estimator via sending  $g_5^k = g_6^k = 1000$ . In this case,

- Average of the good workers:  $\bar{g}^k = \frac{1}{4} \sum_{i=1}^4 g_4^k = 2.25$
- Average estimator:  $g^k = \frac{1}{6}\sum_{i=1}^6 g_i^k = 335$
- Median:  $\hat{g}^k$  any number from [2.5, 3]
- Krum estimator:  $\hat{g}^k = 2 \text{ or } 2.5$

## Simple Example When "Middle-Seekers" Are Good

Let  $d = 1, G = \{1, 2, 3, 4\}, B = \{5, 6\}, g_1^k = 1.5, g_2^k = 2, g_3^k = 2.5, g_4^k = 3$ , and Byzantines are trying to shift the estimator via sending  $g_5^k = g_6^k = 1000$ . In this case,

- Average of the good workers:  $\bar{g}^k = \frac{1}{4} \sum_{i=1}^4 g_4^k = 2.25$
- Average estimator:  $g^k = \frac{1}{6}\sum_{i=1}^6 g_i^k = 335$
- Median:  $\hat{g}^k$  any number from [2.5, 3]
- Krum estimator:  $\hat{g}^k = 2 \text{ or } 2.5$

#### "Middle-seekers" can be good for reducing the effect of outliers

### When "Middle-Seekers" Can Be Bad

Karimireddy, S. P., He, L., & Jaggi, M. (2021, July). Learning from history for byzantine robust optimization. *In International Conference on Machine Learning* (pp. 5311-5319). PMLR.



Figure 1: Failure of existing methods on imbalanced MNIST dataset. Only the head classes (class 1 and 2 here) are learnt, and the rest 8 classes are ignored. See Sec. 7.1.



Figure 2: For fat-tailed distributions, median based aggregators ignore the tail. This bias remains even if we have infinite samples.

## A Little Is Enough (ALIE) Attack

Baruch, G., Baruch, M., & Goldberg, Y. (2019). A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32.



Byzantines send the following vectors:

## A Little Is Enough (ALIE) Attack

Baruch, G., Baruch, M., & Goldberg, Y. (2019). A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32.



mean of the good workers

coordinate-wise standard deviation of good workers

- Byzantines choose *z* such that they are close to the "boundary of the cloud"
- Since Byzantines are closer to the mean, "middle-seekers" will treat opposers as outliers

## The Result of ALIE Attack on the Training @ CIFAR10

Baruch, G., Baruch, M., & Goldberg, Y. (2019). A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32.



"No defense" strategy is more robust! Formal definition of robust aggregation is required!

## **Robust Aggregation Formalism**

Karimireddy, S. P., He, L., & Jaggi, M. (2021, July). Learning from history for byzantine robust optimization. *In International Conference on Machine Learning* (pp. 5311-5319). PMLR.

#### Definition of $(\delta, c)$ -robust aggregator

Let  $g_1 \dots, g_n$  be random variables such that there exist a good subset  $\mathcal{G} \subseteq [n]$  of size  $G \ge (1 - \delta)n > n/2$ such that  $\{g_i\}_{i \in \mathcal{G}}$  are independent and for all fixed pairs of good workers  $i, j \in \mathcal{G}$  we have

$$\mathbb{E}\left[\|g_i - g_j\|^2\right] \le \sigma^2.$$

Let  $\bar{g} = \frac{1}{c} \sum_{i \in \mathcal{G}} g_i$ . Then  $\hat{g} = \operatorname{RAgg}(g_1, \dots, g_n)$  is called  $(\delta, c)$ -robust aggregator if for some c > 0

$$\mathbb{E}\left[\|\widehat{g} - \overline{g}\|^2\right] \le c\delta\sigma^2$$

## **Robust Aggregation Formalism**

Karimireddy, S. P., He, L., & Jaggi, M. (2021, July). Learning from history for byzantine robust optimization. *In International Conference on Machine Learning* (pp. 5311-5319). PMLR.

#### Definition of $(\delta, c)$ -robust aggregator

Let  $g_1 \dots, g_n$  be random variables such that there exist a good subset  $\mathcal{G} \subseteq [n]$  of size  $G \ge (1 - \delta)n > n/2$ such that  $\{g_i\}_{i \in \mathcal{G}}$  are independent and for all fixed pairs of good workers  $i, j \in \mathcal{G}$  we have

$$\mathbb{E}\left[\|g_i - g_j\|^2\right] \le \sigma^2.$$

Let  $\bar{g} = \frac{1}{G} \sum_{i \in \mathcal{G}} g_i$ . Then  $\hat{g} = \operatorname{RAgg}(g_1, \dots, g_n)$  is called  $(\delta, c)$ -robust aggregator if for some c > 0

$$\mathbb{E}\left[\|\widehat{g} - \overline{g}\|^2\right] \le c\delta\sigma^2$$

- Medians and Krum estimators do not satisfy this definition
- **Question:** do such aggregators exist?

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

**Bucketing** takes  $\{g_1, \dots, g_n\}$ , positive integer *s*, and aggregator Aggr as an input and returns

$$\widehat{g} = \operatorname{Aggr}(y_1, \dots, y_{\lceil n/s \rceil})$$

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

**Bucketing** takes  $\{g_1, \dots, g_n\}$ , positive integer *s*, and aggregator Aggr as an input and returns

$$\widehat{g} = \operatorname{Aggr}(y_1, \dots, y_{\lceil n/s \rceil})$$

where 
$$y_i = \frac{1}{s} \sum_{k=s(i-1)+1}^{\min\{si,n\}} x_{\pi(k)}$$
 and  $\pi = (\pi(1), \dots, \pi(n))$  is a random permutation of  $[n]$ 

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

**Bucketing** takes  $\{g_1, ..., g_n\}$ , positive integer *s*, and aggregator Aggr as an input and returns

$$\widehat{g} = \operatorname{Aggr}(y_1, \dots, y_{\lceil n/s \rceil})$$

where  $y_i = \frac{1}{s} \sum_{k=s(i-1)+1}^{\min\{si,n\}} x_{\pi(k)}$  and  $\pi = (\pi(1), \dots, \pi(n))$  is a random permutation of [n]

For any  $\delta \leq \delta_{\max}$  and  $s = \lfloor \delta_{\max} / \delta \rfloor$ 

• Krum • Bucketing is  $(\delta, c)$ -robust aggregator with c = O(1) and  $\delta_{\max} < 1/4$ 

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

**Bucketing** takes  $\{g_1, ..., g_n\}$ , positive integer *s*, and aggregator Aggr as an input and returns

$$\widehat{g} = \operatorname{Aggr}(y_1, \dots, y_{\lceil n/s \rceil})$$

where 
$$y_i = \frac{1}{s} \sum_{k=s(i-1)+1}^{\min\{si,n\}} x_{\pi(k)}$$
 and  $\pi = (\pi(1), \dots, \pi(n))$  is a random permutation of  $[n]$ 

For any  $\delta \leq \delta_{\max}$  and  $s = \lfloor \delta_{\max} / \delta \rfloor$ 

- Krum Bucketing is  $(\delta, c)$ -robust aggregator with c = O(1) and  $\delta_{\max} < 1/4$
- RFA Bucketing is  $(\delta, c)$ -robust aggregator with c = O(1) and  $\delta_{max} < 1/2$

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

**Bucketing** takes  $\{g_1, ..., g_n\}$ , positive integer s, and aggregator Aggr as an input and returns

$$\widehat{g} = \operatorname{Aggr}(y_1, \dots, y_{\lceil n/s \rceil})$$

where 
$$y_i = \frac{1}{s} \sum_{k=s(i-1)+1}^{\min\{si,n\}} x_{\pi(k)}$$
 and  $\pi = (\pi(1), \dots, \pi(n))$  is a random permutation of  $[n]$ 

For any  $\delta \leq \delta_{\max}$  and  $s = \lfloor \delta_{\max} / \delta \rfloor$ 

- Krum Bucketing is  $(\delta, c)$ -robust aggregator with c = O(1) and  $\delta_{max} < 1/4$
- RFA  $\circ$  Bucketing is  $(\delta, c)$ -robust aggregator with c = O(1) and  $\delta_{\max} < 1/2$
- CM  $\circ$  Bucketing is  $(\delta, c)$ -robust aggregator with c = O(d) and  $\delta_{\max} < 1/2$

#### Moreover, these estimators are agnostic to $\sigma^2$ !

Variance Reduction and Byzantine-Robustness

## Why Variance Reduction?

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

**Natural idea:** if the variance of good vectors gets smaller, it becomes progressively harder for Byzantines to shift the result of the aggregation from the true mean



- Large variance allows Byzantines to hide in noise and still create large bias
- Hard to detect outliers

- **Small variance** does not allow Byzantines to create large bias easily
- Easy to detect outliers

### Byrd-SAGA: Byzantine-Robust SAGA



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

Finite-sum optimization:

 $\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{m} \sum_{i=1}^m f_j(x) \right\}$ 

*#* of samples in the dataset

loss on *j*-th sample

### Byrd-SAGA: Byzantine-Robust SAGA



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

**Finite-sum optimization:** 

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{m} \sum_{j=1}^m f_j(x) \right\}$$



loss on *j*-th sample

#### **Byrd-SAGA:**

$$x^{k+1} = x^k - \gamma \widehat{g}^k$$

- Good workers compute • SAGA-estimators
- Server uses geometric ٠ median aggregator

$$\begin{split} \widehat{g}^{k} &= \operatorname{RFA}(g_{1}^{k}, \dots, g_{n}^{k}) \\ g_{i}^{k} &= \begin{cases} \nabla f_{j_{i_{k}}}(x^{k}) - \nabla f_{j_{i_{k}}}(\phi_{i,j_{i_{k}}}^{k}) + \frac{1}{m} \sum_{j=1}^{m} \nabla f_{j}(\phi_{i,j}^{k}), & \text{if } i \in \mathcal{G}, \\ *, & \text{if } i \in \mathcal{B} \end{cases} \\ \phi_{i,j}^{k+1} &= \begin{cases} \phi_{i,j}^{k}, & \text{if } j \neq j_{i_{k}}, \\ x^{k}, & \text{if } j = j_{i_{k}} \end{cases} \quad \forall i \in \mathcal{G} \end{cases}$$

## **Complexity of Byrd-SAGA**



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

#### **Assumptions:**

- $\mu$ -strong convexity of f: •
- *L*–smoothness of  $f_1, \ldots, f_m$ : •

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \quad \forall x, y \in \mathbb{R}^d$$
$$\|\nabla f_j(y) - \nabla f_j(x)\| \le L \|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$$

## **Complexity of Byrd-SAGA**



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

#### **Assumptions:**

- $f(y) \ge f(x) + \langle \nabla f(x), y x \rangle + \frac{\mu}{2} \|y x\|^2 \quad \forall x, y \in \mathbb{R}^d$  $\mu$ -strong convexity of f:
- $\|\nabla f_i(y) \nabla f_i(x)\| \le L \|y x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$ *L*–smoothness of  $f_1, \ldots, f_m$ : •

#### **Theorem:**

Let  $\delta < 1/2$  and the above assumptions hold. Then, there exists a choice of the stepsize  $\gamma$  such that the minibatched version of Byrd-SAGA (with batchsize b) produces  $x^k$  satisfying  $\mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] \leq \varepsilon$  after

$$\mathcal{O}\left(\frac{m^2L^2}{b^2(1-2\delta)\mu^2}\log\frac{1}{\varepsilon}\right) \quad \text{iterations}$$

### **Reflecting on the Complexities**

• Complexity of Byrd-SAGA ( $b = 1, \delta > 0$ ):

• Complexity of Byrd-SAGA (b = 1,  $\delta = 0$ ):

• Complexity of SAGA (b = 1,  $\delta = 0$ ):

 $\mathcal{O}\left(\frac{m^2 L^2}{(1-2\delta)\mu^2}\log\frac{1}{\varepsilon}\right)$  $\mathcal{O}\left(\frac{m^2L^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$  $\mathcal{O}\left(\left(m+\frac{L}{\mu}\right)\log\frac{1}{\varepsilon}\right)$ 

### **Reflecting on the Complexities**

 $\mathcal{O}\left(\frac{m^2 L^2}{(1-2\delta)\mu^2}\log\frac{1}{\varepsilon}\right)$ 

 $\mathcal{O}\left(\left(m+\frac{L}{\mu}\right)\log\frac{1}{\varepsilon}\right)$ 

 $\mathcal{O}\left(\frac{m^2L^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$ 

• Complexity of Byrd-SAGA ( $b = 1, \delta > 0$ ):

• Complexity of Byrd-SAGA (b = 1,  $\delta = 0$ ):

• Complexity of SAGA (b = 1,  $\delta = 0$ ):

The reason for such a dramatic deterioration in the complexity of Byrd-SAGA in comparison to SAGA:

$$\mathbb{E}_k[\widehat{g}^k] \neq \nabla f(x^k)$$

#### Analysis of SAGA/SVRG-based methods is very sensitive to unbiasedness!

SARAH/Geom-SARAH/PAGE (1 node case):

$$x^{k+1} = x^k - \gamma g^k$$



Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.



SARAH/Geom-SARAH/PAGE (1 node case):

$$x^{k+1} = x^k - \gamma g^k$$

$$g^{k} = \begin{cases} \nabla f(x^{k}), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b} \sum_{j \in J_{k}} \left( \nabla f_{j}(x^{k}) - \nabla f_{j}(x^{k-1}) \right), & \text{with prob. } 1 - p \end{cases}$$

	P	Dł		
			l	
	Y	1	Ar	loh

Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.



SARAH/Geom-SARAH/PAGE (1 node case):

$$x^{k+1} = x^k - \gamma g^k$$

$$g^{k} = \begin{cases} \nabla f(x^{k}), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b} \sum_{j \in J_{k}} \left( \nabla f_{j}(x^{k}) - \nabla f_{j}(x^{k-1}) \right), & \text{with prob. } 1 - p \end{cases}$$

 $J_k$  – indices in the mini-batch,  $|J_k| = b$ 



Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.



SARAH/Geom-SARAH/PAGE (1 node case):

$$x^{k+1} = x^k - \gamma g^k \qquad \begin{array}{l} p \sim {}^{b/m} - \text{probability of computing the full gradient} \\ g^k = \begin{cases} \nabla f(x^k), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b} \sum\limits_{j \in J_k} \left( \nabla f_j(x^k) - \nabla f_j(x^{k-1}) \right), & \text{with prob. } 1 - p \end{cases}$$

 $J_k$  – indices in the mini-batch,  $|J_k| = b$ 



Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.



SARAH/Geom-SARAH/PAGE (1 node case):

 $J_k$  – indices in the mini-batch,  $|J_k| = b$ 

PDF

Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.



Li, Z., Bao, H., Zhang, X., & Richtárik, P. (2021, July). PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In International Conference on Machine Learning (pp. 6286-6295). PMLR.

Estimator is biased from the beginning!

 $\mathbb{L}_k[g] \neq \mathbf{v} J(\mathcal{L})$ 

### **New Method**: Byz-PAGE

$$x^{k+1} = x^k - \gamma \widehat{g}^k$$
  $\widehat{g}^k = \operatorname{ARAggr}(g_1^k, \dots, g_n^k)$ 

## **New Method**: Byz-PAGE

 $(\delta, c)$ -robust aggregator agnostic to the variance, e.g., Krum/RFA/CM  $\circ$  Bucketing

$$x^{k+1} = x^k - \gamma \widehat{g}^k$$
  $\widehat{g}^k = \operatorname{ARAggr}(g_1^k, \dots, g_n^k)$ 

## New Method: Byz-PAGE

 $(\delta, c)$ -robust aggregator agnostic to the variance, e.g., Krum/RFA/CM  $\circ$  Bucketing

$$x^{k+1} = x^k - \gamma \widehat{g}^k$$
  $\widehat{g}^k = \operatorname{ARAggr}(g_1^k, \dots, g_n^k)$ 

Geom-SARAH/PAGE-estimator

## Complexity of Byz-PAGE (Simplified)

#### Assumptions:

- *f* is lower-bounded:
- *L*-smoothness of  $f_1, \ldots, f_m$ :

$$f_* = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$$
$$\|\nabla f_j(y) - \nabla f_j(x)\| \le L \|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$$

## Complexity of Byz-PAGE (Simplified)

#### **Assumptions:**

• *f* is lower-bounded:

$$f_* = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$$

• L-smoothness of  $f_1, \dots, f_m$ :  $\|\nabla f_j(y) - \nabla f_j(x)\| \le L \|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$ 

#### **Theorem 1:**

Let the above assumptions hold and ARAggr be  $(\delta, c)$ -robust aggregator. Then, there exists a choice of the stepsize  $\gamma$  such that Byz-PAGE produces  $\hat{x}^k$  satisfying  $\mathbb{E}\left[\left\|\nabla f(\hat{x}^k)\right\|^2\right] \leq \varepsilon^2$  after

## Complexity of Byz-PAGE (Simplified)

#### **Assumptions:**

• *f* is lower-bounded:

$$f_* = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$$

• L-smoothness of  $f_1, \dots, f_m$ :  $\|\nabla f_j(y) - \nabla f_j(x)\| \le L \|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$ 

#### **Theorem 1:**

Let the above assumptions hold and ARAggr be  $(\delta, c)$ -robust aggregator. Then, there exists a choice of the stepsize  $\gamma$  such that Byz-PAGE produces  $\hat{x}^k$  satisfying  $\mathbb{E}\left[\left\|\nabla f(\hat{x}^k)\right\|^2\right] \leq \varepsilon^2$  after

$$\mathcal{O}\left(\frac{\left(1+\sqrt{\frac{c\delta m^2}{b^3}+\frac{m}{b^2n}}\right)L\left(f(x^0)-f_*\right)}{\varepsilon^2}\right) \quad \text{iterations}$$

### Remarks on the Results and One Extension

#### Remarks on the results:

- We achieve **new SOTA theoretical results** for Byzantine-robust learning
- When  $\delta = 0$  (no Byzantines), the derived complexity bounds recover the known ones for Geom-SARAH/PAGE
- Therefore, the terms that are not affected by  $\delta$  are **unimprovable**
- **Open question:** are the derived upper bounds optimal?

### Remarks on the Results and One Extension

#### Remarks on the results:

- We achieve **new SOTA theoretical results** for Byzantine-robust learning
- When  $\delta = 0$  (no Byzantines), the derived complexity bounds recover the known ones for Geom-SARAH/PAGE
- Therefore, the terms that are not affected by  $\delta$  are **unimprovable**
- **Open question:** are the derived upper bounds optimal?

#### The extension to the compressed communication case:

• Byz-PAGE:

$$g_i^k = \begin{cases} \nabla f\left(x^k\right), & \text{with prob. } p\\ g^{k-1} + \frac{1}{b} \sum_{j \in J_k} \left(\nabla f_j\left(x^k\right) - \nabla f_j\left(x^{k-1}\right)\right), & \text{with prob. } 1 - p \end{cases}$$

### Remarks on the Results and One Extension

#### Remarks on the results:

- We achieve **new SOTA theoretical results** for Byzantine-robust learning
- When  $\delta = 0$  (no Byzantines), the derived complexity bounds recover the known ones for Geom-SARAH/PAGE
- Therefore, the terms that are not affected by  $\delta$  are **unimprovable**
- **Open question:** are the derived upper bounds optimal?

#### The extension to the compressed communication case:

• Byz-PAGE:

$$g_{i}^{k} = \begin{cases} \nabla f\left(x^{k}\right), & \text{with prob. } p\\ g^{k-1} + \frac{1}{b} \sum_{j \in J_{k}} \left(\nabla f_{j}\left(x^{k}\right) - \nabla f_{j}\left(x^{k-1}\right)\right), & \text{with prob. } 1 - p \end{cases}$$

• Byz-VR-MARINA:

 $g_{i}^{k} = \begin{cases} \nabla f\left(x^{k}\right), & \text{with prob. } p\\ g^{k-1} + \mathcal{Q}\left(\frac{1}{b}\sum_{j\in J_{k}}\left(\nabla f_{j}\left(x^{k}\right) - \nabla f_{j}\left(x^{k-1}\right)\right)\right), & \text{with prob. } 1 - p \end{cases}$ 

Gorbunov, E., Burlachenko, K. P., Li, Z., & Richtárik, P. (2021, July). MARINA: Faster nonconvex distributed learning with compression. In International Conference on Machine Learning (pp. 3788-3798). PMLR.

unbiased compression operator

### **Numerical Results**

- We tested the proposed method on the logistic regression tasks
- In this experiment, we have 4 good workers and 1 Byzantine
- As predicted by the derived results, the proposed method has linear convergence
- Competitors struggle to achieve better loss
- The results are consistent for all tested attacks



# **Concluding Remarks**

## In the Paper We Also Have

- Analysis of the version with compression (Byz-VR-MARINA)
- Analysis under bounded heterogeneity
- Non-uniform sampling of stochastic gradients
- Analysis takes into account data-similarity
- Additional experiments

Thank you!



Link to the paper

