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

Eduard Gorbunov [1] Samuel Horváth [1] Peter Richtárik [2] Gauthier Gidel [3]

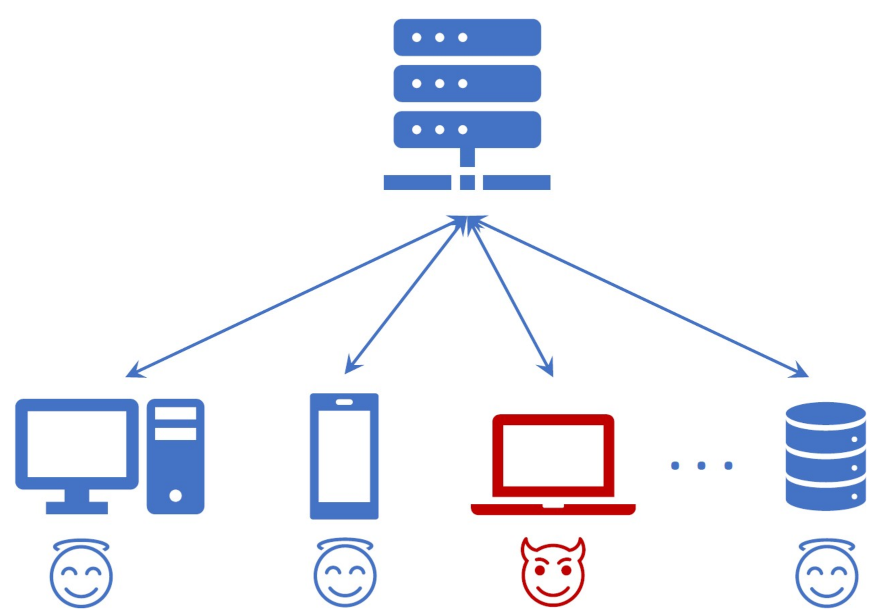
[1]Mohamed bin Zayed University of Artificial Intelligence [2]King Abdullah University of Science and Technology [3]Mila, Université de Montréal

1. Byzantine-Robust Optimization

Distributed optimization problem:

$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\frac{1}{\mathcal{G}}\sum_{i\in\mathcal{G}}f_i(x)\right\},\quad f_i(x)=\frac{1}{m}\sum_{j=1}^{m}f_{i,j}(x)\quad \forall i\in\mathcal{G}$$

- $\mathcal{G}$ is the set of good clients
- $\mathcal{B}$ is the set of *Byzantine workers* – the workers that can arbitrarily deviate from the prescribed protocol (maliciously or not) and are assumed to be omniscient
- $\mathcal{G}\sqcup\mathcal{B}=[n]$ is the set of clients participating in training



Main difficulties in Byzantine-robust optimization:

- When functions are arbitrarily heterogeneous, the problem is impossible to solve
- Fraction of Byzantines $\delta = {}^{B}/_{n}$ should be smaller than $^1/_2$
- Standard approaches based on averaging are vulnerable
- Robust aggregation alone does not ensure robustness [1]

2. Robust Aggregation

Popular aggregation rules:

- $\texttt{Krum}(x_1,\ldots,x_n) := \mathrm{argmin}_{x_i\in\{x_1,\ldots,x_n\}}\sum_{j\in S_i}\|x_j-x_i\|^2$ [7], where $S_i\subseteq\{x_1,\ldots,x_n\}$ are $n-|\mathcal{B}|-2$ closest vectors to x_i
- Robust Fed. Averaging: $\texttt{RFA}(x_1,\ldots,x_n) := \mathrm{argmin}_{x\in\mathbb{R}^d}\sum_{i=1}^n\|x-x_i\|$
- Coordinate-wise Median: $[\texttt{CM}(x_1,\ldots,x_n)]_t := \mathrm{argmin}_{u\in\mathbb{R}}\sum_{i=1}^n|u-[x_i]_t|$

These defenses are vulnerable to Byzantine attacks [8,9] and do not satisfy the following definition.

Definition 1: (δ,c)-Robust Aggregator (modification of the definition from [1])

Assume that $\{x_1,x_2,\ldots,x_n\}$ is such that there exists a subset $\mathcal{G}\subseteq[n]$ of size $|\mathcal{G}|=G\geq(1-\delta)n$ for $\delta<0.5$ and there exists $\sigma\geq 0$ such that $\frac{1}{G(G-1)}\sum_{i,l\in\mathcal{G}}\mathbb{E}[\|x_i-x_l\|^2]\leq\sigma^2$ where the expectation is taken w.r.t. the randomness of $\{x_i\}_{i\in\mathcal{G}}$. We say that the quantity $\widehat{x}$ is (δ,c)-**Robust Aggregator** ((δ,c)-$\texttt{RAgg}$) and write $\widehat{x}=\texttt{RAgg}(x_1,\ldots,x_n)$ for some $c>0$, if the following inequality holds:

$$\mathbb{E}\left[\|\widehat{x}-\overline{x}\|^2\right]\leq c\delta\sigma^2, \tag{1}$$

where $\overline{x}=\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}x_i$. If additionally $\widehat{x}$ is computed without the knowledge of σ^2, we say that $\widehat{x}$ is (δ,c)-**Agnostic Robust Aggregator** ((δ,c)-$\texttt{ARAgg}$) and write $\widehat{x}=\texttt{ARAgg}(x_1,\ldots,x_n)$.

One can robustify $\texttt{Krum}$, $\texttt{RFA}$, and $\texttt{CM}$ using bucketing [1].

Algorithm $\texttt{Bucketing}$: Robust Aggregation using bucketing [1]

1: **Input:** $\{x_1,\ldots,x_n\}$, $s\in\mathbb{N}$ – bucket size, $\texttt{Aggr}$ – aggregation rule
2: Sample random permutation $\pi=(\pi(1),\ldots,\pi(n))$ of $[n]$
3: Compute $y_i=\frac{1}{s}\sum_{k=s(i-1)+1}^{\min\{si,n\}}x_{\pi(k)}$ for $i=1,\ldots,\lceil {}^n/_s\rceil$
4: **Return:** $\widehat{x}=\texttt{Aggr}(y_1,\ldots,y_{\lceil n/s\rceil})$

3. SGD and Variance Reduction

SGD: $x^{k+1}=x^k-\gamma g^k,\quad g^k=\frac{1}{n}\sum_{i=1}^n\nabla f_{i,j_i^k}(x^k)$

✗ Variances of the estimators $\nabla f_{i,j_i^k}(x^k)$ do not go to zero

✗ Byzantines can easily hide in the noise and create a large bias (even if the aggregation is robust)

SAGA [2]: $x^{k+1}=x^k-\gamma g^k,\quad g^k=\frac{1}{n}\sum_{i=1}^n g_i^k,$

$g_i^k=\nabla f_{j_i^k}(x^k)-\nabla f_{i,j_i^k}(w_{i,j_i^k}^k)+\frac{1}{m}\sum_{j=1}^{m}\nabla f_{i,j}(w_{i,j}^k)$

✓ Variances of the estimators g_i^k go to zero

✗ Analysis relies on the unbiasedness: $\mathbb{E}[g_i^k\mid x^k]=\nabla f_i(x^k)$

SARAH/Geom-SARAH/PAGE [3,4,5]:

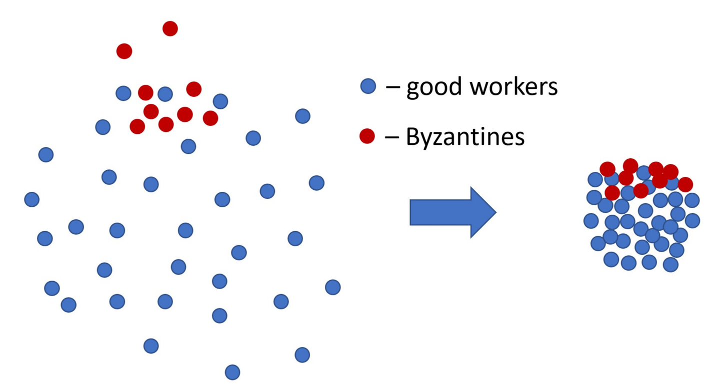
$x^{k+1}=x^k-\gamma g^k,\quad g^k=\frac{1}{n}\sum_{i=1}^n g_i^k,$

$$g_i^k=\begin{cases}\nabla f_i(x^k), & \text{with prob. } p,\\ g_i^{k-1}+\nabla f_{i,j_i^k}(x^k)-\nabla f_{i,j_i^k}(x^{k-1}), & \text{with prob. } 1-p\end{cases}$$

✓ Variances of the estimators g_i^k go to zero

✓ Analysis does not rely on the unbiasedness: $\mathbb{E}[g_i^k\mid x^k]\neq\nabla f_i(x^k)$

How can variance reduction help? *It leaves less space for Byzantines to hide in the noise.*



Main Contributions

◇ **New method: Byz-VR-MARINA.** We make VR-MARINA (VR-method with compression) [6] applicable to Byzantine-robust learning using robust agnostic aggregation [1].

◇ **New SOTA results under more general assumptions.** Under quite general assumptions (no strong assumptions on the compression and second moment of the stochastic gradient; non-uniform sampling is supported), we prove new theoretical convergence results that are tight and outperform known ones when the target accuracy is small enough.

4. Technical Preliminaries

Definition 2: Unbiased Compression

Stochastic mapping $\mathcal{Q}:\mathbb{R}^d\to\mathbb{R}^d$ is called unbiased compressor/compression operator if there exists $\omega\geq 0$ such that for any $x\in\mathbb{R}^d$

$$\mathbb{E}\left[\mathcal{Q}(x)\right]=x,\quad \mathbb{E}\left[\|\mathcal{Q}(x)-x\|^2\right]\leq\omega\|x\|^2. \tag{2}$$

Assumptions

- Smoothness and lower-boundedness: $\forall x,y\in\mathbb{R}^d$ we have $\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\|$ and $f_*=\inf_{x\in\mathbb{R}^d}f(x)>-\infty$
- ζ^2-heterogeneity: $\frac{1}{G}\sum_{i\in\mathcal{G}}\|\nabla f_i(x)-\nabla f(x)\|^2\leq\zeta^2\quad\forall x\in\mathbb{R}^d$
- Global Hessian variance assumption: $\frac{1}{G}\sum_{i\in\mathcal{G}}\|\nabla f_i(x)-\nabla f_i(y)\|^2-\|\nabla f(x)-\nabla f(y)\|^2\leq L_\pm^2\|x-y\|^2$
- Local Hessian variance assumption: $\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\|\widehat{\Delta}_i(x,y)-\Delta_i(x,y)\|^2\leq\frac{\mathcal{L}_\pm^2}{b}\|x-y\|^2$, where $\Delta_i(x,y)=\nabla f_i(x)-\nabla f_i(y)$ and $\widehat{\Delta}_i(x,y)$ is an unbiased mini-batched estimator of $\Delta_i(x,y)$ with batch size b

5. New Method: Byz-VR-MARINA

Algorithm Byz-VR-MARINA: Byzantine-tolerant VR-MARINA

1: **Input:** starting point x^0, stepsize γ, minibatch size b, probability $p\in(0,1]$, number of iterations K, (δ,c)-$\texttt{ARAgg}$
2: **for** $k=0,1,\ldots,K-1$ **do**
3: Get a sample from Bernoulli distribution with parameter p: $c_k\sim\mathrm{Be}(p)$. Broadcast g^k, c_k to all workers
4: **for** $i\in\mathcal{G}$ in parallel **do**
5: $x^{k+1}=x^k-\gamma g^k$
6: Set $g_i^{k+1}=\begin{cases}\nabla f_i(x^{k+1}), & \text{if } c_k=1,\\ g^k+\mathcal{Q}\left(\widehat{\Delta}_i(x^{k+1},x^k)\right), & \text{otherwise,}\end{cases}$ where minibatched estimator $\widehat{\Delta}_i(x^{k+1},x^k)$ of $\nabla f_i(x^{k+1})-\nabla f_i(x^k)$; $\mathcal{Q}(\cdot)$ for $i\in\mathcal{G}$ are computed independently
7: **end for**
8: $g^{k+1}=\texttt{ARAgg}(g_1^{k+1},\ldots,g_n^{k+1})$
9: **end for**

6. Convergence in the Non-Convex Case

Theorem 1

Let the introduced assumptions hold. Assume that $0<\gamma\leq\frac{1}{L+\sqrt{A}}$, where $A=\frac{6(1-p)}{p}\left(\frac{4c\delta}{p}+\frac{1}{2G}\right)\left(\omega L^2+\frac{(1+\omega)\mathcal{L}_\pm^2}{b}\right)+\frac{6(1-p)}{p}\left(\frac{4c\delta(1+\omega)}{p}+\frac{\omega}{2G}\right)L_\pm^2$. Then for all $K\geq 0$ the point $\widehat{x}^K$ chosen uniformly at random from the iterates $x^0,x^1,\ldots,x^K$ produced by Byz-VR-MARINA satisfies

$$\mathbb{E}\left[\|\nabla f(\widehat{x}^K)\|^2\right]\leq\frac{2\Phi_0}{\gamma(K+1)}+\frac{24c\delta\zeta^2}{p}, \tag{3}$$

where $\Phi_0=f(x^0)-f_*+\frac{\gamma}{p}\|g^0-\nabla f(x^0)\|^2$ and $\mathbb{E}[\cdot]$ denotes the full expectation.

- When $\zeta=0$ (homogeneous data) the method converges asymptotically to the exact solution with rate $\mathcal{O}({}^1/_K)$

7. Convergence in PŁ-case

Definition 3: Polyak-Łojasiewicz (PŁ) condition

Function f satisfies Polyak-Łojasiewicz (PŁ) condition with parameter μ if for all $x\in\mathbb{R}^d$ there exists $x^*\in\mathrm{argmin}_{x\in\mathbb{R}^d}f(x)$ such that

$$\|\nabla f(x)\|^2\geq 2\mu\left(f(x)-f(x^*)\right). \tag{4}$$

Theorem 1

Let the introduced assumptions hold and function f satisfies μ-PŁcondition. Assume that $0<\gamma\leq\min\left\{\frac{1}{L+\sqrt{2A}},\frac{p}{4\mu}\right\}$, where $A=\frac{6(1-p)}{p}\left(\frac{4c\delta}{p}+\frac{1}{2G}\right)\left(\omega L^2+\frac{(1+\omega)\mathcal{L}_\pm^2}{b}\right)+\frac{6(1-p)}{p}\left(\frac{4c\delta(1+\omega)}{p}+\frac{\omega}{2G}\right)L_\pm^2$. Then for all $K\geq 0$ the iterates produced by Byz-VR-MARINA satisfy

$$\mathbb{E}\left[f(x^K)-f(x^*)\right]\leq(1-\gamma\mu)^K\Phi_0+\frac{24c\delta\zeta^2}{\mu}, \tag{5}$$

where $\Phi_0=f(x^0)-f_*+\frac{2\gamma}{p}\|g^0-\nabla f(x^0)\|^2$.

- When $\zeta=0$ (homogeneous data) the method converges linearly asymptotically to the exact solution

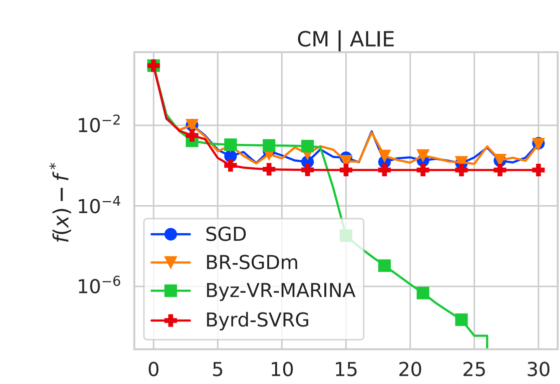
8. Comparison with Prior Work

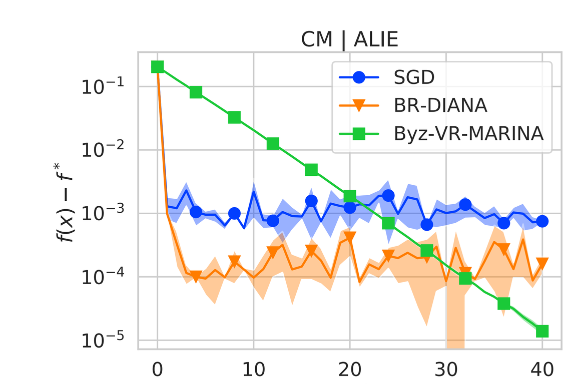
Setup	Method	Complexity (NC)	Complexity (PŁ)
Hom. data, no compr.	BR-SGDm [1]	$\frac{1}{\varepsilon^2}+\frac{\sigma^2(c\delta+1/n)}{b\varepsilon^4}$	✗
	BR-MVR [1]	$\frac{1}{\varepsilon^2}+\frac{\sigma\sqrt{c\delta+1/n}}{\sqrt{b}\varepsilon^3}$	✗
	BTARD-SGD [10]	$\frac{1}{\varepsilon^2}+\frac{n^2\delta\sigma^2}{Cb\varepsilon^2}+\frac{\sigma^2}{nb\varepsilon^4}$	$\frac{1}{\mu}+\frac{\sigma^2}{nb\mu\varepsilon}+\frac{n^2\delta\sigma}{C\sqrt{b\mu\varepsilon}}$
	Byrd-SAGA [11]	✗	$\frac{m^2}{b^2(1-2\delta)\mu^2}$
	Byz-VR-MARINA	$\frac{1+\sqrt{\frac{c\delta m^2}{b^3}+\frac{m}{b^2n}}}{\varepsilon^2}$	$\frac{1+\sqrt{\frac{c\delta m^2}{b^3}+\frac{m}{b^2n}}}{\mu}+\frac{m}{b}$
Het. data, no compr.	BR-SGDm [1]	$\frac{1}{\varepsilon^2}+\frac{\sigma^2(c\delta+1/n)}{b\varepsilon^4}$	✗
	Byrd-SAGA [11]	✗	$\frac{m^2}{b^2(1-2\delta)\mu^2}$
	Byz-VR-MARINA	$\frac{1+\sqrt{\frac{c\delta m^2}{b^2}(1+\frac{1}{b})+\frac{m}{b^2n}}}{\varepsilon^2}$	$\frac{1+\sqrt{\frac{c\delta m^2}{b^2}(1+\frac{1}{b})+\frac{m}{b^2n}}}{\mu}+\frac{m}{b}$
Het. data, compr.	BR-CSGD [12]	✗	$\frac{1}{\mu^2}$
	BR-CSAGA [12]	✗	$\frac{m^2}{b^2\mu^2(1-2\delta)^2}$
	BROADCAST [12]	✗	$\frac{m^2(1+\omega)^{3/2}}{b^2\mu^2(1-2\delta)}$
	Byz-VR-MARINA	$\frac{1+\sqrt{c\delta(1+\omega)(1+\frac{1}{b})}}{p\varepsilon^2}+\frac{\sqrt{(1+\omega)(1+\frac{1}{b})}}{\sqrt{pn}\varepsilon^2}$	$\frac{1+\sqrt{c\delta(1+\omega)(1+\frac{1}{b})}}{p\mu}+\frac{\sqrt{(1+\omega)(1+\frac{1}{b})}}{\sqrt{pn}\mu}+\frac{m}{b}+\omega$

- Dependencies on numerical constants (and logarithms in PŁ setting), smoothness constants, and initial suboptimality are omitted
- $p=\min\{{}^b/_m, {}^1/_{(1+\omega)}\}$ = probability of communication in Byz-VR-MARINA
- Analyses of BR-SGDm, BR-MVR, BTARD-SGD, BR-CSGD, BR-CSAGA rely on uniformly bounded variance assumption
- In the het. case, the methods converge only to the error $\sim\zeta^2$
- The result for BROADCAST is derived for $\omega\leq\frac{\mu^2(1-2\delta)^2}{56L^2(2-2\delta^2)}$

9. Experiments

- We consider a logistic regression model with ℓ_2-regularization and non-convex regularization $\lambda\sum_{i=1}^{d}\frac{x_i^2}{1+x_i^2}$
- We have 4 good workers and 1 Byzantine worker
- A Little is enough (ALIE) attack [8] is considered: the Byzantine workers estimate the mean $\mu_{\mathcal{G}}$ and standard deviation $\sigma_{\mathcal{G}}$ of the good updates, and send $\mu_{\mathcal{G}}-z\sigma_{\mathcal{G}}$, $z>0$
- Byrd-SVRG – a version of Byrd-SAGA with SVRG-estimator instead of SAGA-estimator
- BR-DIANA – a version of BROADCAST with SGD-estimator instead of SAGA-estimator





References

[1] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. ICLR 2022.

[2] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. NeurIPS 2014.

[3] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. ICML 2017.

[4] Samuel Horváth, Lihua Lei, Peter Richtárik, and Michael I. Jordan. Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 2022.

[5] Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. ICML 2021.

[6] Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster non-convex distributed learning with compression. ICML 2021.

[7] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. NeurIPS 2017.

[8] Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. NeurIPS 2019.

[9] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation. UAI 2020.

[10] Eduard Gorbunov, Alexander Borzunov, Michael Diskin, and Max Ryabinin. Secure distributed training at scale. ICML 2021.

[11] Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B Giannakis. Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. IEEE Transactions on Signal Processing, 68:4583–4596, 2020

[12] Heng Zhu and Qing Ling. Broadcast: Reducing both stochastic and compression noise to robustify communication-efficient federated learning. arXiv preprint arXiv:2104.06685, 2021.