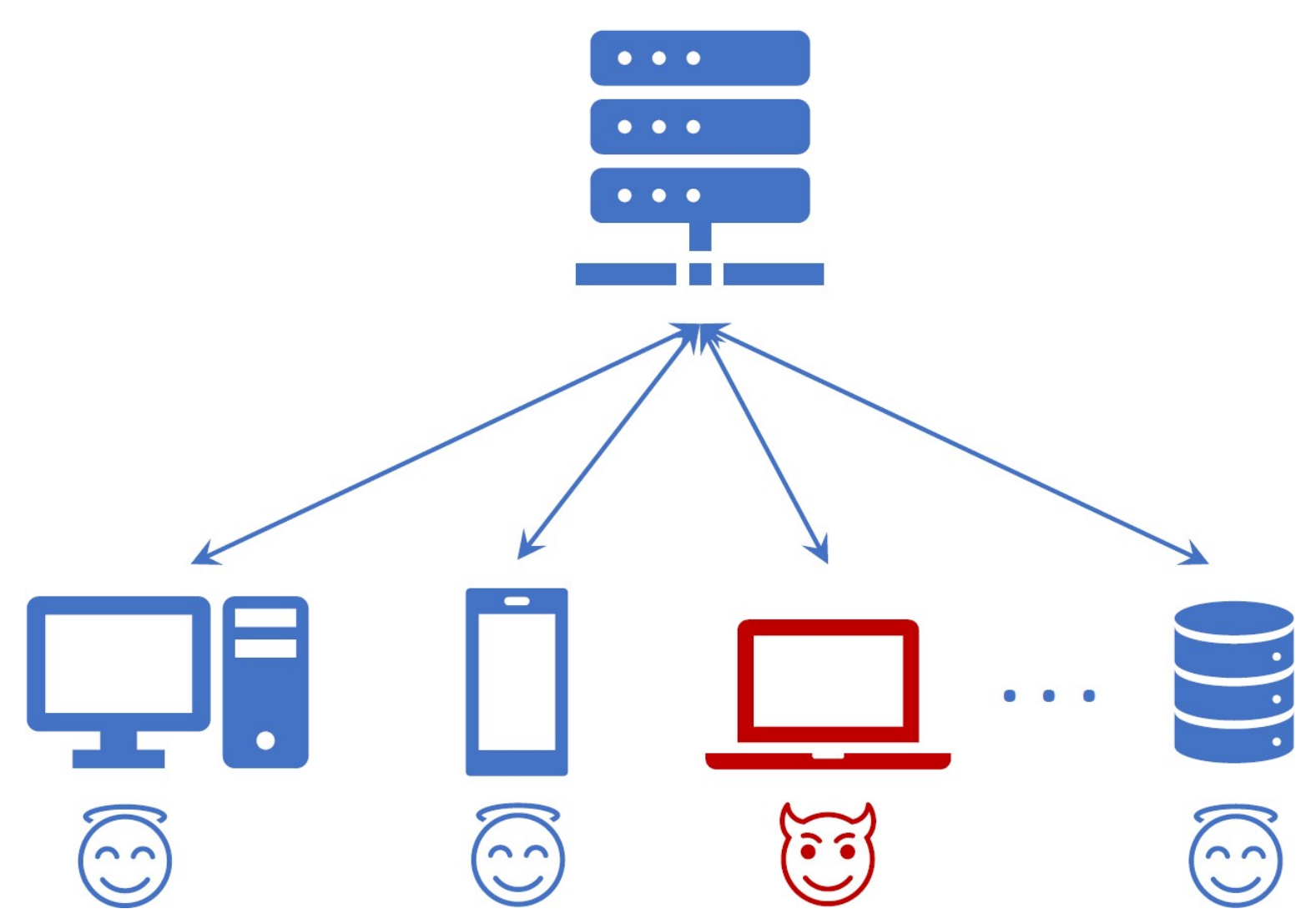


1. Byzantine-Robust Optimization

Distributed optimization problem:

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

- \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:

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

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 $\{x_1, x_2, \dots, x_n\}$ 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 $\frac{1}{G} \sum_{i \in \mathcal{G}} \mathbb{E}[\|x_i - \bar{x}\|^2] \leq \sigma^2$ where the expectation is taken w.r.t. the randomness of $\{x_i\}_{i \in \mathcal{G}}$. We say that the quantity \hat{x} is (δ, c) -**Robust Aggregator** (δ, c) -RAGg and write $\hat{x} = \text{RAGg}(x_1, \dots, x_n)$ for some $c > 0$, if the following inequality holds:

$$\mathbb{E}[\|\hat{x} - \bar{x}\|^2] \leq c\delta\sigma^2, \quad (1)$$

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

One can robustify **Krum**, **RFA**, and **CM** using bucketing [1].

Algorithm Bucketing: Robust Aggregation using bucketing [1]

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

3. SGD and Variance Reduction

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

- \times Variances of the estimators $\nabla f_{i,j_i^k}(x^k)$ do not go to zero
- \times Byzantines can easily hide in the noise and create a large bias (even if the aggregation is robust)

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

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

- \checkmark Variances of the estimators g_i^k go to zero
- \times Analysis relies on the unbiasedness: $\mathbb{E}[g_i^k | x^k] = \nabla f_i(x^k)$

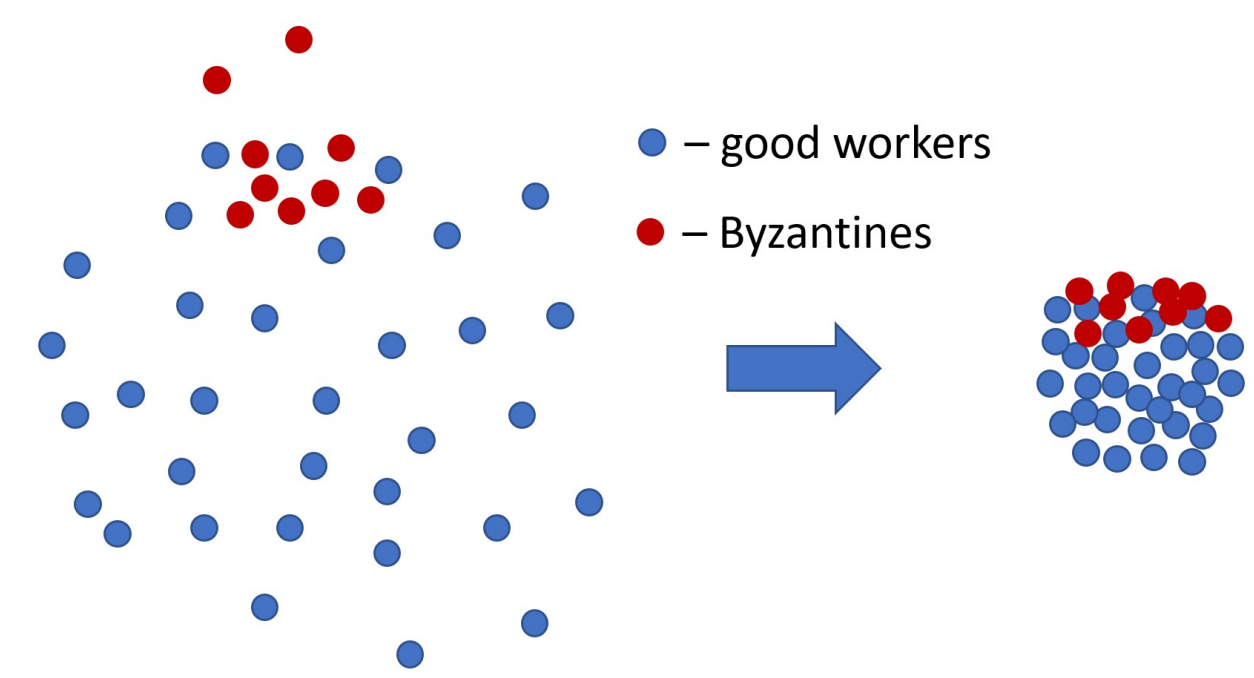
SARAH/Geom-SARAH/PAGE [3,4,5]:

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

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

- \checkmark Variances of the estimators g_i^k go to zero
- \checkmark Analysis does not rely on the unbiasedness: $\mathbb{E}[g_i^k | x^k] \neq \nabla f_i(x^k)$

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

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

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:** $\frac{1}{G} \sum_{i \in \mathcal{G}} \mathbb{E}[\|\hat{\Delta}_i(x, y) - \Delta_i(x, y)\|^2] \leq \frac{L_{\pm}^2}{b} \|x - y\|^2$, where $\Delta_i(x, y) = \nabla f_i(x) - \nabla f_i(y)$ and $\hat{\Delta}_i(x, y)$ 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

- Input:** starting point x^0 , stepsize γ , minibatch size b , probability $p \in (0, 1]$, number of iterations K , (δ, c) -ARAgg
- for** $k = 0, 1, \dots, K - 1$ **do**
- Get a sample from Bernoulli distribution with parameter p : $c_k \sim \text{Be}(p)$. Broadcast g^k , c_k to all workers
- for** $i \in \mathcal{G}$ in parallel **do**
- $x^{k+1} = x^k - \gamma g^k$
- Set $g_i^{k+1} = \begin{cases} \nabla f_i(x^{k+1}), & \text{if } c_k = 1, \\ g_i^k + \mathcal{Q}(\hat{\Delta}_i(x^{k+1}, x^k)), & \text{otherwise,} \end{cases}$ where minibatched estimator $\hat{\Delta}_i(x^{k+1}, x^k)$ of $\nabla f_i(x^{k+1}) - \nabla f_i(x^k)$; $\mathcal{Q}(\cdot)$ for $i \in \mathcal{G}$ are computed independently
- end for**
- $g^{k+1} = \text{ARAgg}(g_1^{k+1}, \dots, g_n^{k+1})$
- 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 $A = \frac{6(1-p)}{p} \left(\frac{4c\delta}{p} + \frac{1}{2G} \right) \left(\omega L^2 + \frac{(1+\omega)L_{\pm}^2}{b} \right) + \frac{6(1-p)}{p} \left(\frac{4c\delta(1+\omega)}{p} + \frac{\omega}{2G} \right) L_{\pm}^2$. Then for all $K \geq 0$ the point \hat{x}^K chosen uniformly at random from the iterates x^0, x^1, \dots, x^K produced by Byz-VR-MARINA satisfies

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

where $\Phi_0 = f(x^0) - f_* + \frac{2\gamma}{p} \|g^0 - \nabla f(x^0)\|^2$ 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 $x^* \in \text{argmin}_{x \in \mathbb{R}^d} f(x)$ such that

$$\|\nabla f(x)\|^2 \geq 2\mu (f(x) - f(x^*)). \quad (4)$$

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 $A = \frac{6(1-p)}{p} \left(\frac{4c\delta}{p} + \frac{1}{2G} \right) \left(\omega L^2 + \frac{(1+\omega)L_{\pm}^2}{b} \right) + \frac{6(1-p)}{p} \left(\frac{4c\delta(1+\omega)}{p} + \frac{\omega}{2G} \right) L_{\pm}^2$. Then for all $K \geq 0$ the iterates produced by Byz-VR-MARINA satisfy

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

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

References

- Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. ICLR 2022.
- Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. NeurIPS 2014.
- Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. ICLR 2017.
- 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.
- Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. ICLR 2021.
- Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster non-convex distributed learning with compression. ICLR 2021.
- Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. NeurIPS 2017.
- Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. NeurIPS 2019.
- Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation. UAI 2020.
- Eduard Gorbunov, Alexander Borzunov, Michael Diskin, and Max Ryabinin. Secure distributed training at scale. ICLR 2021.
- 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.
- Heng Zhu and Qing Ling. Broadcast: Reducing both stochastic and compression noise to robustify communication-efficient federated learning. arXiv preprint arXiv:2104.06685, 2021.

8. Comparison with Prior Work

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

- 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 μ_G and standard deviation σ_G of the good updates, and send $\mu_G - z\sigma_G$, $z > 0$
- 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

