Main difficulties in Byzantine-robust optimization:
• When functions are arbitrarily heterogeneous, the problem is impossible to solve.
• Fracation of Bytinanese δ < ½ should be smaller than ½.
• Standard approaches based on averaging are vulnerable.
• Robust aggregation alone does not ensure robustness.

2. Robust Aggregation

Popular aggregation rules:
Krum: \(f(x_1, \ldots, x_n) = \text{argmin}_{1 \leq i \leq n} \{ f_i(x) \} \) [7], where \( S \subseteq \{1, \ldots, n\} \) and \( |S| = n - |B| \) closest vectors to \( x \).
Robust Fed. Averaging: \( f(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \).
Coordi-nate-Wise Median: \( f(x_1, \ldots, x_n) = \text{median}_{1 \leq i \leq n} \{ f_i(x) \} \).
These defenses are vulnerable to Byzantine attacks [8,9] and do not satisfy the following definition.

4. Technical Preliminaries

1. Byzantine-Robust Optimization

Distributed optimization problem:
\[
\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\} f_i(x) - \frac{1}{m} \sum_{j=1}^{m} f_j(x), \quad \forall i \in G
\]

3. SGD and Variance Reduction

SGD: \( x^{t+1} = x^t - \gamma t \nabla f(x^t) \)

- Variance of the estimates \( \nabla f(x^t) \) do not go to zero.
- Bytanzianese can easily hide in the noise and create a large bias even if the aggregation is robust.

SAGA: \( x^{t+1} = x^t - \gamma t \nabla f(x^t) + \sum_{i=1}^{t} \nabla f_i(x^t) \)

- Variance of the estimates \( g_i^t \) go to zero.

- Analysis relies on the unbiasedness: \( \mathbb{E}[g_i^t] = \nabla f(x^t) \)

How can variance reduction help? \( \nabla f(x) \) is less expensive for Byzantines to hide in the noise.

5. New Method: Byz-VR-MARINA

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

1. Input: starting point \( x^0 \), stepsize \( \gamma \), minibatch size \( \beta \), probability \( p \in (0, 1] \), number of iterations \( K \), \((\delta, c)\)-ARagg
2. for \( k = 1 \) to \( K \)
3. Get a sample from Bernoulli distribution with parameter \( p \) \( c_k \sim \text{Be}(p) \).
4. Broadcast \( g_i^t \) to all workers
5. for \( i \in G \) in parallel do
6. \( x_{ stagnant} \leftarrow x^i \)
7. for \( t = 0 \) to \( B - 1 \)
8. \( \theta_i^t \leftarrow d_{ stagnant} \leftarrow \left( x_{ stagnant} - x^i \right) \)
9. end for

6. Convergence in the Non-Convex Case

Let the introduced assumptions hold. Assume that \( 0 < \gamma \leq \frac{1}{\mu} \frac{1}{\beta} \frac{1}{\sqrt{1 + \kappa}} \kappa \beta \frac{1}{\sqrt{1 + \kappa}} \frac{1}{\sqrt{1 + \kappa}} \frac{1}{\sqrt{1 + \kappa}} k \).

7. Convergence in PL-case

Function satisfies Polyak-Łojasiewicz (PL) condition with parameter \( \mu > 0 \) for all \( x \in \mathbb{R}^d \) there exists \( \varepsilon > 0 \) such that \( \| \nabla f(x) \|^2 \geq 2 \mu \varepsilon (f(x) - f^*)^2 \).

8. Comparison with Prior Work

9. Experiments

- We consider a logistic regression model with \( C \) regularization and non-convex regularization \( \sum_{i=1}^{n} \| x_i \|^2 \).
- We have 4 good workers and 1 Byzantine worker.
- A little is enough (ALIE) attack [8] is considered: The Byzantine workers estimate the mean \( \mu_0 \) and standard deviation \( \sigma_0 \) of the good updates, and send \( \mu_\delta = \mu_0 - \sigma_0 \varepsilon \geq 0 \).
- BR-DIANA: a version of BROADCAST with SGD-estimator instead of SAGA-estimator.

References