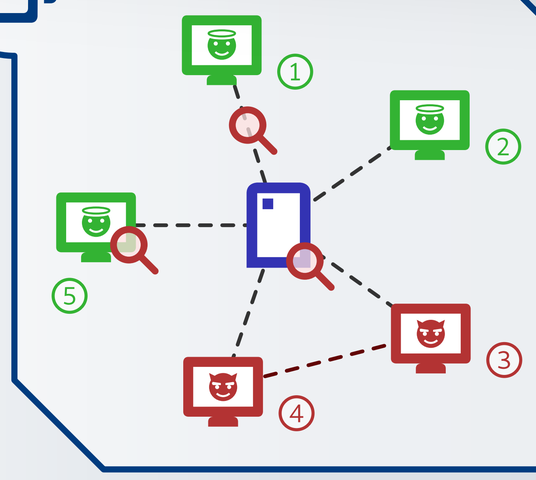
Distributed Momentum for Byzantine-resilient Stochastic Gradient Descent

El-Mahdi El-Mhamdi **Rachid Guerraoui** **Sébastien Rouault**

Byzantine SGD



Let a (non-convex) loss $Q : \mathbb{R}^d \to \mathbb{R}$ with $Q(\theta_t) \triangleq \mathbb{E}_{x \sim \mathcal{D}}[q(\theta_t, x)]$ (same data distribution for each worker)

Each worker i samples $\left(x_t^{(i,1)} \dots x_t^{(i,b)}\right)$ from $\mathcal{D}$, computes $g_t^{(i)} \triangleq \frac{1}{b}\sum_{k=1}^{b} \nabla q\left(\theta_t, x_t^{(i,k)}\right) \approx \nabla Q(\theta_t)$, and sends $g_t^{(i)}$ to the parameter server.

The parameter server **aggregates** the gradients with

$$G_t \triangleq \sum_{u=0}^{t} \mu^{t-u} F\left(g_u^{(1)}, \dots, g_u^{(n)}\right)$$ (classical momentum)

and updates the parameters $\theta_{t+1} = \theta_t - \alpha_t G_t$ with $\theta_0 \in \mathbb{R}^d$, $\alpha_t > 0$

F is called a Gradient Aggregation Rule. For instance, averaging can be used: $F\left(g_t^{(1)}, \dots, g_t^{(n)}\right) = \frac{1}{n}\sum_{i=1}^{n} g_t^{(i)}$

In the presence of Byzantine workers, a more robust aggregation is performed, with a Byzantine-resilient GAR such as:

- Multi-**Krum**
- **Trimmed-Mean**
- **Bulyan** of Multi-Krum
- **Median**
- **Phocas**
- **MeaMed**
- etc...

But there are effective attacks...

State-of-the-art attacks

Byzantine workers may deviate arbitrarily far from the protocol

Let $\varepsilon \in \mathbb{R}_{\geq 0}$ and $a_t \in \mathbb{R}^d$ an "attack" vector, both depending on the attack.

With this family of attack, each Byzantine worker sends the same Byzantine vector $\overline{g_t} + \varepsilon\, a_t$, where $\overline{g_t}$ is an approximation of the real gradient $\nabla Q(\theta_t)$.

Fall of Empires

$\varepsilon = 1.1$

$a_t \triangleq -\overline{g_t}$

Little is Enough

$\varepsilon = 1.5$

$a_t \triangleq -\sigma_t$, where σ_t is the coordinate-wise standard deviation of the honest gradients

Distributed momentum

Instead of $G_t \triangleq \sum_{u=0}^{t} \mu^{t-u} F\left(g_u^{(1)}, \dots, g_u^{(n)}\right)$ (standard formulation)

do $G_t \triangleq F\left(\sum_{u=0}^{t} \mu^{t-u} g_u^{(1)}, \dots, \sum_{u=0}^{t} \mu^{t-u} g_u^{(n)}\right)$ (our formulation)

Experimental assessment

https://github.com/LPD-EPFL/ByzantineMomentum

→ **maximum top-1 cross-accuracy** for all of the 6 Byzantine-resilient GAR named above, each repeated 5 times (= 30 points/box). The dotted blue line is the median of the maximum top-1 cross-accuracy of the 5 runs **without attack**. Each run is seeded for **reproducibility** purpose.

This is only a fraction of the results available in the main paper.

All of these experiments + graphs are reproducible in one command.

home sweet home