Distributed Momentum for Byzantine-resilient Stochastic Gradient Descent

El-Mahdi El-Mhamdi  Rachid Guerraoui  Sébastien Rouault

Byzantine SGD
- Parameter Server
- Bidirectional channel
- "Honest" worker
- "Byzantine" worker

State-of-the-art attacks
- Byzantine workers may deviate arbitrarily far from the protocol with this family of attack, each Byzantine worker sends the same Byzantine vector \( \sigma \), where \( \sigma \) is an approximation of the real gradient \( \nabla Q(\theta_i) \).

\[
\begin{align*}
\text{Fall of Empires} & : \quad e = 1.1 \\
\text{Little is Enough} & : \quad e = 1.5 \\
\end{align*}
\]

\( a_i \approx -\sigma_i \), where \( \sigma_i \) is the coordinate-wise standard deviation of the honest gradients.

Distributed momentum
- Instead of
  \[
  G_t \triangleq \sum_{u=0}^{t} \mu^{t-u} F \left( g_u^{(1)}, \ldots, g_u^{(n)} \right)
  \]
  do
  \[
  G_t \triangleq F \left( \sum_{u=0}^{t} \mu^{t-u} g_u^{(1)}, \ldots, \sum_{u=0}^{t} \mu^{t-u} g_u^{(n)} \right)
  \]

Experimental assessment
- \( \mu = 0.7 \) in all experiments
- \( \alpha = 0.95 \) in all experiments
- \( \epsilon = 0.01 \) in all experiments

- 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.

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

\[\text{https://github.com/LPD-EPFL/ByzantineMomentum}\]