Zeno: Distributed Stochastic Gradient Descent with Suspicion-based Fault-tolerance

Cong Xie
University of Illinois at Urbana-Champaign

Oluwasanmi Koyejo
University of Illinois at Urbana-Champaign

Indranil Gupta
University of Illinois at Urbana-Champaign

Main contribution

First approach for Byzantine-tolerant SGD:

- **that tolerates an arbitrary number of Byzantine workers**
- **and provides convergence guarantees for non-convex problems**

Byzantine failures in distributed SGD

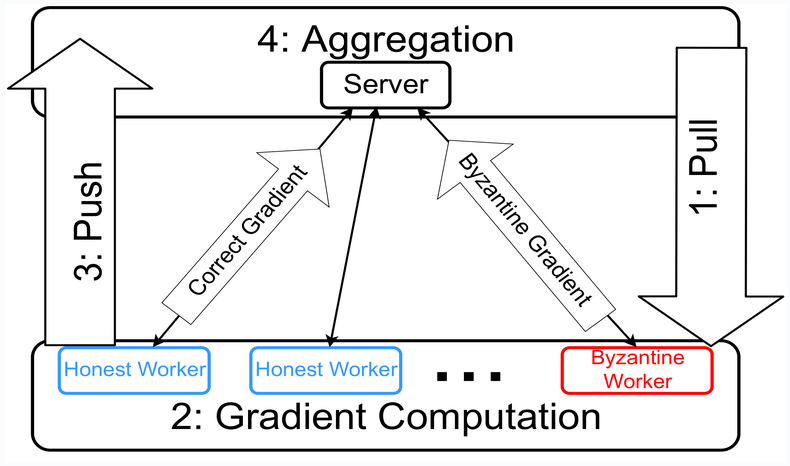
$$\min_{x\in\mathbb{R}^d} F(x), \quad \text{where } F(x) = \mathbb{E}_{z\sim\mathcal{D}}[f(x;z)].$$

- m workers, distributed SGD:

$$x^{t+1} = x^t - \gamma^t \texttt{Aggr}(\{g_i(x^t) : i \in [m]\}),$$

where $\texttt{Aggr}(\cdot)$ is an aggregation rule (e.g., averaging),

$$g_i(x^t) = \begin{cases} \text{Arbitrary} & i\text{th worker is Byzantine,} \\ \nabla F_i(x^t) & \text{otherwise.} \end{cases}$$



Stochastic Descendant Score

Definition 1. *$f_r(x) = \frac{1}{n_r}\sum_{i=1}^{n_r} f(x;z_i)$, $\mathbb{E}[f_r(x)] = F(x)$. For any update (gradient estimator) u:*

$$Score_{\gamma,\rho}(u,x) = f_r(x) - f_r(x-\gamma u) - \rho\|u\|^2.$$

The score is composed of two parts:

- **Estimated descendant of the loss function:** Larger $f_r(x) - f_r(x-\gamma u)$ implies faster convergence.
- **Magnitude of the update:** Smaller $\|u\|^2$ implies smaller change.

Suspicion-based Aggregation

Definition 2. *Sort $\{\tilde{v}_i : i \in [m]\}$ by the stochastic descendant score (Definition 1):*

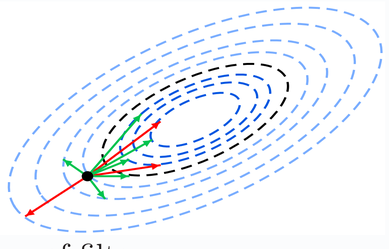
$$Score_{\gamma,\rho}(\tilde{v}_{(1)},x) \geq \ldots \geq Score_{\gamma,\rho}(\tilde{v}_{(m)},x).$$

Aggregate the updates with the $m-b$ highest scores:

$$\texttt{Zeno}_b(\{\tilde{v}_i : i \in [m]\}) = \frac{1}{m-b}\sum_{i=1}^{m-b}\tilde{v}_{(i)}, \quad b > q.$$

Key Intuition

- Contour of loss value
- Black dot: current value
- Arrows: candidates
- Red: Byzantine
- Green: correct
- $b = 3$
- Black dashed circle: boundary of filter
- Inside boundary: harmless updates
- Outside boundary: dropped



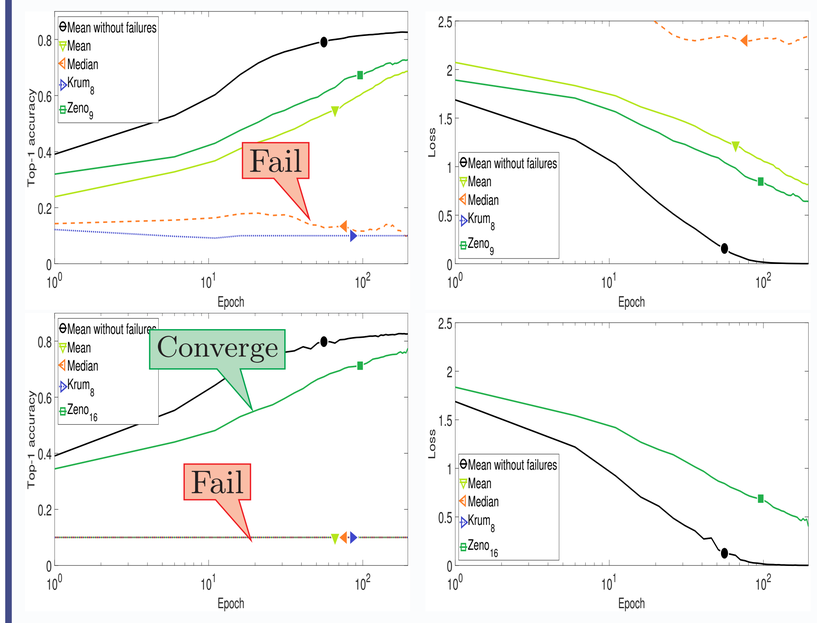
Main result

Convergence to a critical point:

- Larger m and smaller b improve convergence
- L-smooth and μ-weakly convex: $\langle\nabla f(x), y-x\rangle + \frac{\mu}{2}\|y-x\|^2 \leq f(y)-f(x) \leq \langle\nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2$
- Take $\gamma = \frac{1}{L\sqrt{T}}$, $\rho = \frac{\beta\gamma^2}{2}$, $\beta > \max(0,-\mu)$:

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x^t)\|^2}{T} \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{(b-q+1)(m-q)}{(m-b)^2}\right).$$

Experiments



- CIFAR-10 image classification dataset
- Bit-flipping failures: Byzantine workes push negative gradients instead of the true gradients to servers
- Batch size: 100, $n_r = 4$, $\rho = 0.0005$, $\gamma = 0.1$