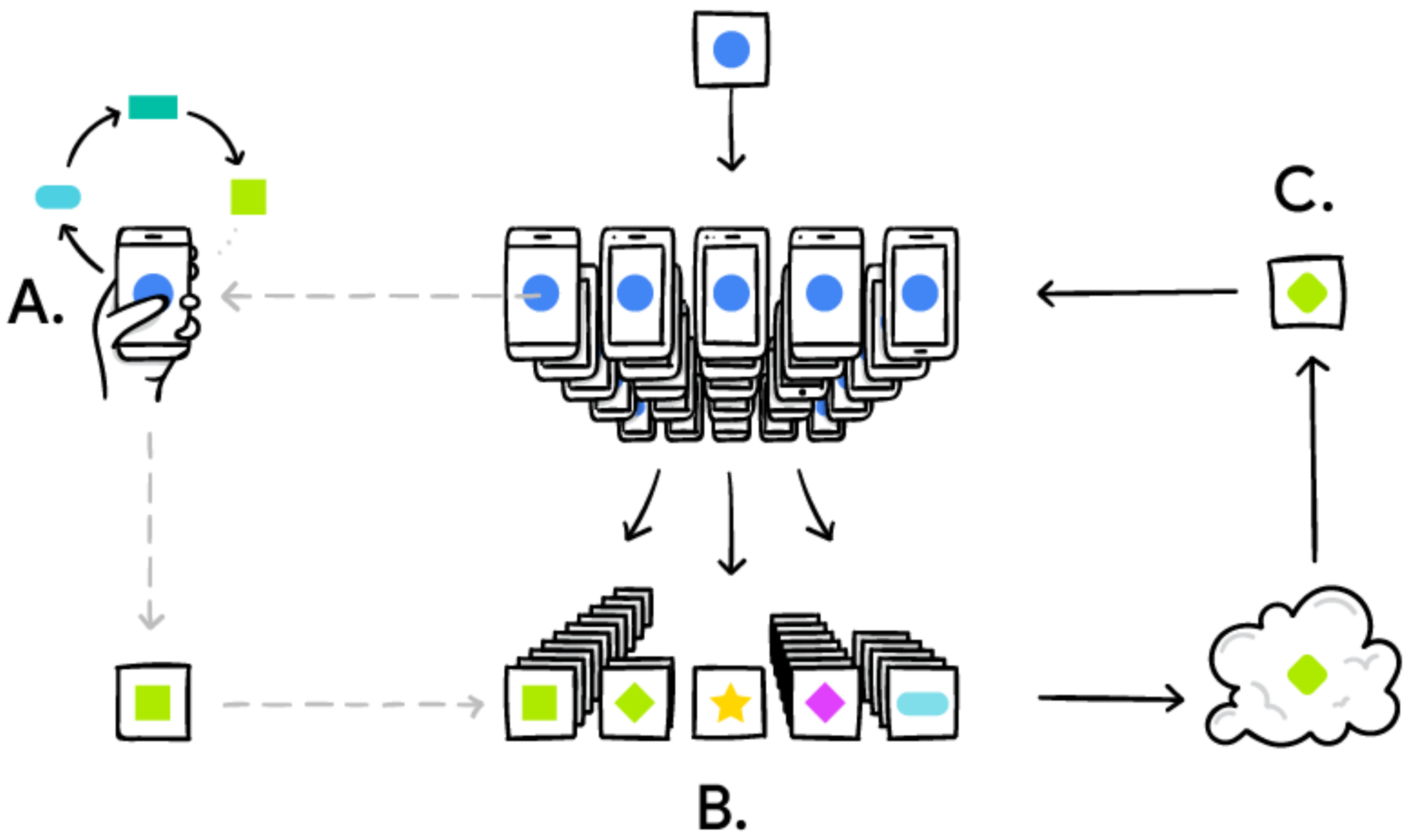
Distributed robust statistical learning: Byzantine mirror descent

DONGSHENG DING, XIAOHAN WEI, MIHAILO R. JOVANOVIĆ

USC Viterbi
School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

MOTIVATION

Federated learning [1]

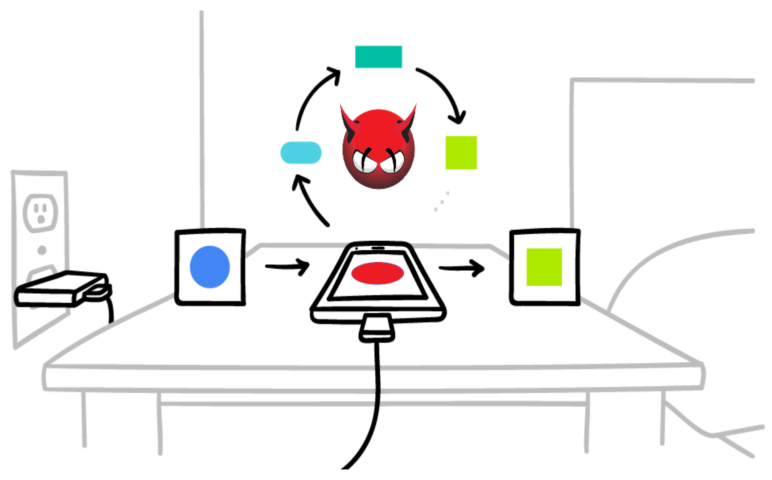


A. Worker machine B. Master machine C. Shared model

Byzantine failure model

A small, but unknown fraction of machines are assumed to behave arbitrarily.

- Comm.\Comp. failure
- Malicious user



Challenges

- Robustness
- Comm.\Comp. complexity
- Dimension scalability

PROBLEM FORMULATION

Objective

$$\underset{w}{\text{minimize}} \quad F(w) \; := \; \mathbb{E}_{z\sim\mathcal{D}}\,[f(w;z)]$$
$$\text{subject to} \quad w \;\in\; \mathcal{W}$$

- $\mathcal{D}$ – unknown distribution
- $\mathcal{W}$ – $\{w \in \mathbb{R}^d : \|w - w_1\| \le W\}$

Byzantine stochastic gradient descent [2]

- m – total number of machines
- $\alpha \in [0, 0.5)$ – fraction of machines that are Byzantine
- αm – total number of Byzantine machines
- T – Total number of iterations

$$w_t$$

Random sample $z_t^i \sim \mathcal{D}$

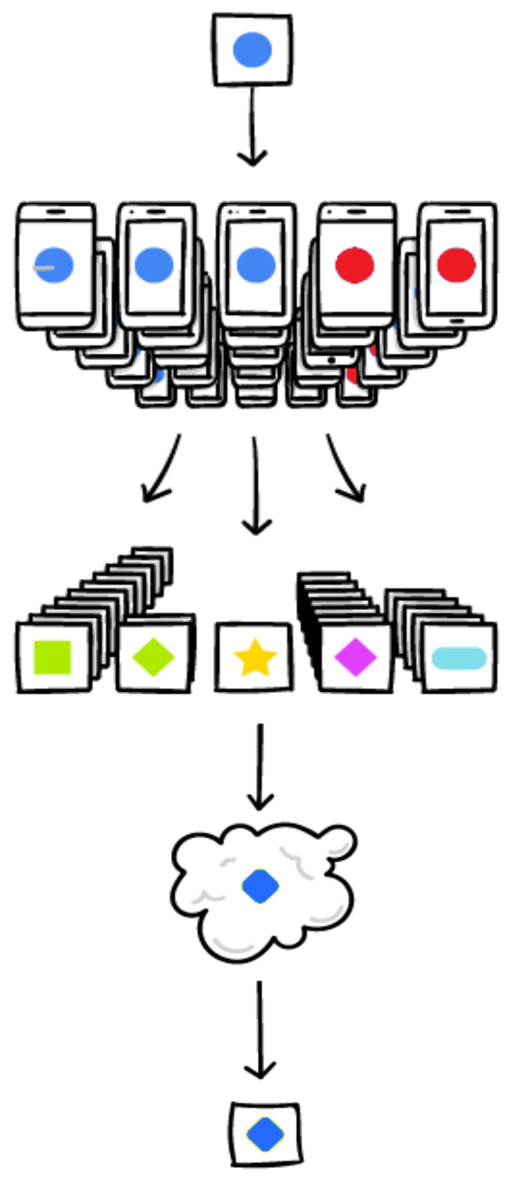
$$f(w_t; z_t^1), \ldots, f(w_t; z_t^i), \ldots, f(w_t; z_t^m)$$

$$\nabla_t^1, \ldots, \nabla_t^i, \ldots, \nabla_t^m$$

$$\text{"}\nabla_t^1, \ldots, \nabla_t^i, \ldots, \nabla_t^m\text{"}$$

$$\mathsf{good}_t \;\leftarrow\; \|\nabla_t^i - \nabla_t\|_* \;\le\; V$$

$$w_{t+1} \;\leftarrow\; \mathsf{GD}\left(w_t, \; \tfrac{1}{m}\textstyle\sum_{i\in \mathsf{good}_t} \nabla_t^i\right)$$



BYZANTINE MIRROR DESCENT

- **Master machine:** send w_t to all machines
- **Worker machine** $i \in [m]$:
  - compute local gradient:

$$\nabla_t^i \;=\; \begin{cases} \nabla f(w_t; z_t^i) & \text{normal machine} \\ \text{"}\nabla f(w_t; z_t^i)\text{"} & \text{Byzantine machine} \end{cases}$$

  - measure reliability of gradient:

$$A_i \;\leftarrow\; \sum_{k=1}^{t} \langle \nabla_k^i, w_k - w_1 \rangle, \quad B_i \;\leftarrow\; \sum_{k=1}^{t} \nabla_k^i$$

  - send A_i, B_i and ∇_t^i to master machine
- **Master machine:** aggregate gradients
  - identify good candidates: good_t

$$A_{\text{med}} \;=\; \text{median}\{A_1, \cdots, A_m\}$$
$$B_{\text{med}} \;\leftarrow\; B_i \text{ satisfies } \big|\{j \in [m] : \|B_i - B_j\|_* \le I_B\}\big| > \frac{m}{2}$$
$$\nabla_{\text{med}} \;\leftarrow\; \nabla_t^i \text{ satisfies } \left|\left\{j \in [m] : \|\nabla_t^i - \nabla_t^j\|_* \le 2V\right\}\right| > \frac{m}{2}$$
$$\mathsf{close}_t = \left\{i : |A_i - A_{\text{med}}| \le I_A, \|B_i - B_{\text{med}}\|_* \le I_B, \|\nabla_t^i - \nabla_{\text{med}}\|_* \le 4V\right\}$$
$$\mathsf{good}_t \leftarrow \mathsf{good}_{t-1} \cap \mathsf{close}_t$$

$$\xi_t \;=\; \frac{1}{m} \sum_{i\in \mathsf{good}_t} \nabla_t^i$$

  - mirror descent

$$w_{t+1} \;=\; \arg\min_{w\in\mathcal{W}} \left\{ D(w, w_t) \;+\; \eta \,\langle \xi_t, w - w_t \rangle \right\}$$

CONVERGENCE RESULT

Setup

- $\Phi(x) = \sum_1^d w(i) \log w(i)$ – mirror map
- $\mathcal{W} = \{w : \mathbb{1}^T w = 1, w \ge 0\}$ – probability simplex
- $\|\nabla_t^i - \nabla f(w_t)\|_\infty \le V$ – bounded gradient
- $I_A = I_B = 4V\Delta\sqrt{T}, \Delta = \sqrt{\log(\frac{16mT}{\delta})}$

Error bound [3]

Suppose F is G-Lipschitz and L-smooth.
If $\eta \le \frac{1}{2L}$, then, with probability $1-\delta$,

$$F(\bar{w}) - F(w^\star) \;\le\; \frac{2\log(d)^2}{\eta T} + \frac{8V\Delta\left(\sqrt{mT} + 4\alpha m\sqrt{T}\right)}{mT} + \eta\left(\frac{4V^2\Delta^2}{m} + 32\alpha^2 V^2\right)$$

Moreover, if we choose η optimally, then,

$$F(\bar{w}) - F(w^\star) \;\le\; O\left(\underbrace{\frac{\log d}{T} + \frac{1}{\sqrt{mT}}}_{\text{mini-batch SGD}} + \underbrace{\frac{\alpha}{\sqrt{T}}}_{\text{Byzantine}}\right)$$

REFERENCES

[1] B. McMahan, D. Ramage, "Federated learning: collaborative machine learning without centralized training data", *Google AI Blog*, 2017.

[2] D. Alistarh, Z. Allen-Zhu, J. Li, "Byzantine stochastic gradient descent", *NeurIPS*, 2018.

[3] D. Ding, X. Wei, M. R. Jovanović, "Distributed robust statistical learning: Byzantine mirror descent", *CDC*, 2019. To appear.