Byzantine-resilient distributed learning under constraints

Dongsheng Ding

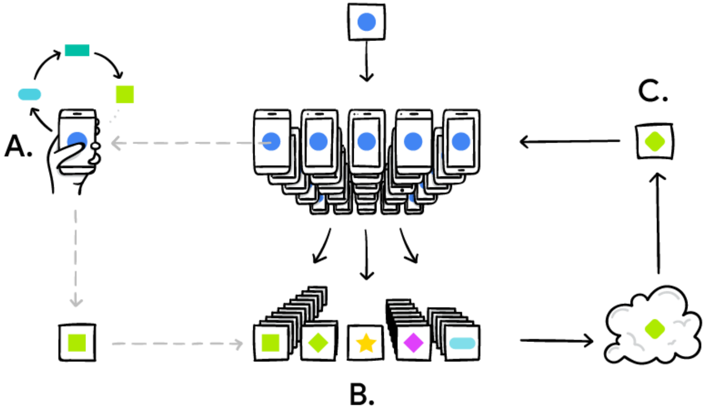
a joint work with

Xiaohan Wei, Hao Yu, Mihailo Jovanović

2021 American Control Conference

Motivating application

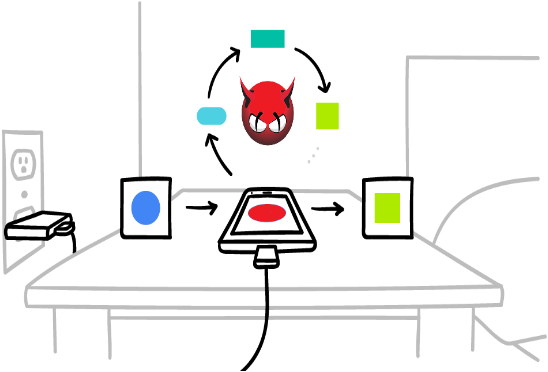
- FEDERATED LEARNING



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

Google AI, Blog ’17

Byzantine fault



- FAULT SOURCES
  - ⋆ Machine failures
  - ⋆ Communication errors
  - ⋆ Malicious users

Byzantine failure model

- STOCHASTIC LEARNING PROBLEM WITH CONSTRAINTS

$$\begin{aligned} \underset{w \in \mathcal{W}}{\text{minimize}} \quad & F(w) \ := \ \mathbb{E}_{z \sim \mathcal{D}}\left(f(w; z)\right) \\ \text{subject to} \quad & g_j(w) \ \leq \ 0, \ j \ = \ 1, \ldots, k. \end{aligned}$$

Byzantine failure model

- STOCHASTIC LEARNING PROBLEM WITH CONSTRAINTS

$$\begin{aligned} \underset{w \in \mathcal{W}}{\text{minimize}} \quad & F(w) \; := \; \mathbb{E}_{z \sim \mathcal{D}} \left(f(w; z) \right) \\ \text{subject to} \quad & g_j(w) \; \leq \; 0, \; j \; = \; 1, \ldots, k. \end{aligned}$$

1 master and m workers

- $\star$ g_j – deterministic constraints
- $\star$ $z_t^i \sim \mathcal{D}$, $i \in \{1, \ldots, m\}$

Byzantine failure model

- STOCHASTIC LEARNING PROBLEM WITH CONSTRAINTS

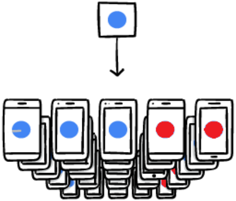
$$\begin{aligned} \underset{w \in \mathcal{W}}{\text{minimize}} \quad & F(w) \;:=\; \mathbb{E}_{z \sim \mathcal{D}}\left(f(w; z)\right) \\ \text{subject to} \quad & g_j(w) \;\leq\; 0, \;\; j \;=\; 1, \ldots, k. \end{aligned}$$

1 master and m workers

- $\star$ g_j – deterministic constraints
- $\star$ $z_t^i \sim \mathcal{D}$, $i \in \{1, \ldots, m\}$

m gradients at time t

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



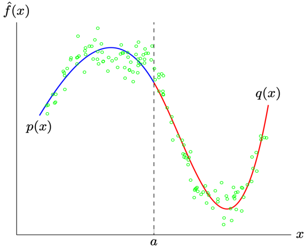
α fraction

Example

- CONSTRAINED LINEAR REGRESSION

$$
\begin{aligned}
f(w; z) &= \frac{1}{2}\,(x^T w - y)^2 \\
g_j(w) &= A_j w - b_j, \; j = 1, \ldots, k.
\end{aligned}
$$

- $\star$ $z := (x, y)$ – data from distribution $\mathcal{D}$
- $\star$ $g_j(w) \leq 0$ – constraints, e.g., smooth spline fitting



Boyd, Vandenberghe, Applied Linear Algebra ’18

Adding Byzantine resilience to primal-dual methods

Identification of "good" workers

- MEDIAN AGGREGATION
  - robust to outliers

sequence	mean	median
$1,\ 2,\ 3,\ 6,\ 7,\ 8,\ 10$	5.3	6
$10^{-3},\ 2,\ 3,\ 6,\ 7,\ 8,\ 10^{3}$	145.6	6

  - breakdown point 50%

Bregman divergence

$$D(x, y) \;:=\; \phi(x) \;-\; \phi(y) \;-\; \nabla\phi(y)^T(x \;-\; y)$$

⋆ ϕ – differentiable, 1-strongly convex w.r.t. $\|\cdot\|$

- EXAMPLES

⋆ $\phi(x) \;=\; \frac{1}{2}\,\|x\|_2^2$ strongly convex w.r.t. $\|\cdot\|_2$

$$D(x, y) \;=\; \frac{1}{2}\,\|x \;-\; y\|_2^2$$

⋆ $\phi(x) \;=\; \sum_i x(i) \log x(i)$ strongly convex w.r.t. $\|\cdot\|_1$

$$D(x, y) \;=\; \sum_i x(i) \log \frac{x(i)}{y(i)}$$ KL divergence

Bubeck, Found. Trends Mach. Learn. ’15

Byzantine primal-dual method

- BYZANTINE PRIMAL MIRROR DESCENT

$$w_{t+1} := \operatorname*{argmin}_{w \in \mathcal{W}} \left\langle \xi_t + \sum_{j=1}^{k} (q_{j,t+1} + g_j(w_t)) \nabla g_j(w_t), w - w_t \right\rangle + \eta_t D(w, w_t)$$

Byzantine primal-dual method

- BYZANTINE PRIMAL MIRROR DESCENT

$$w_{t+1} := \operatorname*{argmin}_{w \in \mathcal{W}} \left\langle \xi_t + \sum_{j=1}^k (q_{j,t+1} + g_j(w_t)) \nabla g_j(w_t), w - w_t \right\rangle + \eta_t \, D(w, w_t)$$

  - ⋆ ξ_t – stochastic estimate of the gradient $\nabla F(w_t)$

$$\xi_t = \frac{1}{m} \sum_{i \in \Omega_t} \nabla_t^i, \quad \Omega_t \text{ – set of "good" workers}$$

Byzantine primal-dual method

- BYZANTINE PRIMAL MIRROR DESCENT

$$w_{t+1} \;:=\; \operatorname*{argmin}_{w \in \mathcal{W}} \Big\langle \xi_t + \textstyle\sum_{j=1}^{k} (q_{j,t+1} + g_j(w_t)) \nabla g_j(w_t), w - w_t \Big\rangle + \eta_t\, D(w, w_t)$$

  - $\star$ ξ_t – stochastic estimate of the gradient $\nabla F(w_t)$

$$\xi_t = \frac{1}{m} \sum_{i \in \Omega_t} \nabla_t^i, \quad \Omega_t - \text{set of "good" workers}$$

- DUAL UPDATE

$$q_{j,t+2} = \max\left(-g_j(w_{t+1}),\; q_{j,t+1} + g_j(w_{t+1}) \right)$$

- IMPORTANT QUANTITIES
  - $\star$ $\nabla_t^1, \ldots, \nabla_t^m$ – gradients (normal or Byzantine)
  - $\star$ $A_t^1, \ldots, A_t^m$ – gradient related values

$$A_t^i \;:=\; \sum_{\tau=1}^{t} \left\langle \nabla_\tau^i, w_\tau - w_1 \right\rangle$$

  - $\star$ $B_t^1, \ldots, B_t^m$ – accumulated gradients

$$B_t^i \;:=\; \sum_{\tau=1}^{t} \nabla_\tau^i$$

used to update the set of "good" workers

Ding, Wei, Jovanović. CDC ’19
Alistarh, Allen-Zhu, Li, NeurIPS ’18

- UPDATE THE SET OF 'GOOD' WORKERS

$\Omega_t \quad \leftarrow \quad i \in \Omega_{t-1}$ satisfies

$$\begin{cases} |A_t^i - A_{\text{med}}| \;\leq\; I_A \\ \|B_t^i - B_{\text{med}}\|_* \;\leq\; I_B \\ \|\nabla_t^i - \nabla_{\text{med}}\|_* \;\leq\; 4C \end{cases}$$

$$\Omega_0 = \{1, \ldots, m\}$$

$$\|\nabla_t^i - \nabla_t\|_* \leq C$$

$$I_A = 4WC\Delta\sqrt{2T}$$

$$I_B = 4C\Delta\sqrt{2T}$$

$$\Delta = \Theta\left(\sqrt{\log \frac{mT}{\delta}}\right)$$

Ding, Wei, Jovanović. CDC '19

Convergence result

- Optimality gap

$$F(\bar{w}) - F(w^\star) \le \tilde{O}\left(\frac{1}{T} + \frac{1}{\sqrt{mT}} + \frac{\alpha}{\sqrt{T}}\right) \quad \text{w.h.p.}$$

- Constraint violation

$$g_j(\bar{w}) \le \tilde{O}\left(\frac{1}{T} + \frac{1}{\sqrt{mT}} + \frac{\alpha}{\sqrt{T}}\right) \quad \text{w.h.p.}$$

$\alpha \in [0, 0.5)$

$\bar{w} = \frac{1}{T}\sum_{t=0}^{T-1} w_t$

F, g – convex and smooth

Convergence result

- OPTIMALITY GAP

$$F(\bar{w}) - F(w^\star) \;\le\; \tilde{O}\left(\frac{1}{T} + \frac{1}{\sqrt{mT}} + \frac{\alpha}{\sqrt{T}}\right) \quad \text{w.h.p.}$$

- CONSTRAINT VIOLATION

$$g_j(\bar{w}) \;\le\; \tilde{O}\left(\frac{1}{T} + \frac{1}{\sqrt{mT}} + \frac{\alpha}{\sqrt{T}}\right) \quad \text{w.h.p.}$$

$\alpha \in [\,0,\, 0.5\,)$

$\bar{w} = \frac{1}{T}\sum_{t=0}^{T-1} w_t$

F, g – convex and smooth

$$\text{matches the rate of} \begin{cases} \text{Byzantine MD \& SGD}, & \alpha \neq 0 \\ \text{batch SGD}, & \alpha = 0 \end{cases}$$

Summary

- Results
  - ⋆ Byzantine primal-dual method
  - ⋆ Optimal rate

- Ongoing effort
  - ⋆ Nonsmooth optimization problems
  - ⋆ Byzantine constraints

Thank you !