# Secure Distributed Optimization under Gradient Attacks

Shuhua Yu

Carnegie Mellon University

MOPTA 2023

# Acknowledgments

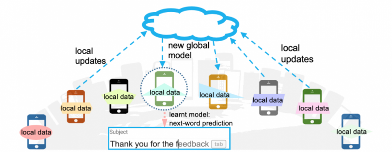
Prof. Soummya Kar
CMU

# Distributed systems

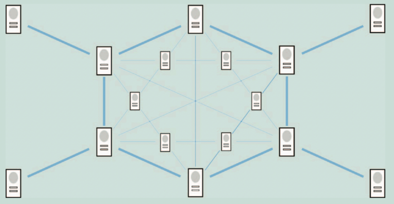
- Cyber-physical systems: power grids, sensor networks.
- Cloud centered devices: smartphones, wearable devices.
- Autonomous vehicle systems: sensors, actuators, multi-vehicle coordination.
- ...
- Goal: secure information processing over distributed systems.

# Two distributed schemes




**Server/client model**

Server coordinates the *global* and *local* information exchange

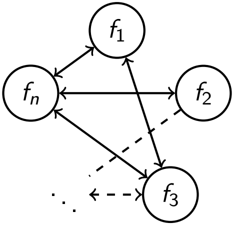


**Decentralized model**

Agents exchange *local* information with direct neighbors over a graph

- Data are distributed over multiple agents due to privacy and scalability.
- We focus on the decentralized model.
  - Flexible: no central server is required.
  - Less communication: communication with neighbors only.

# Decentralized optimization




- Consider a network of agent $i = 1, \ldots, n$.
- Agent $i$ holds a local data distribution $\mathcal{D}_i$, on which we define

$$f_i(\boldsymbol{x}) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i} \ell(\boldsymbol{x}, \xi_i).$$

  for some loss function $\ell$. Examples include: least-squares, logistic-regression, neural networks.
- Agents communicate over a graph to minimize $f(\boldsymbol{x}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{x})$.

# Decentralized SGD

For each agent $i$, at each iteration:

- agent $i$ holds a local decision variable $\boldsymbol{x}_i$;
- agent $i$ computes a stochastic gradient with random noise $\boldsymbol{\xi}_i$,

$$g_i(\boldsymbol{x}_i) = \nabla f_i(\boldsymbol{x}_i) + \boldsymbol{\xi}_i;$$

- agent $i$ employs some weight $w_{ij}$, $w_{ij} > 0$ if agent $j$ is the neighbor of agent $i$;
- decentralized stochastic gradient descent (DSGD), for some stepsize $\alpha$,

$$\boldsymbol{x}_i^+ = \sum_{j=1}^{n} w_{ij}\boldsymbol{x}_i - \alpha g_i(\boldsymbol{x}_i).$$

*Q: $\boldsymbol{\xi}_i$ is typically assumed to be well-behaved, what if it is adversarial?*

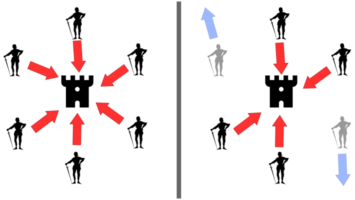
# Gradient attacks

- In DSGD:

$$\boldsymbol{x}_i^+ = \sum_{j=1}^{n} w_{ij}\boldsymbol{x}_i - \alpha \underbrace{\left(\nabla f_i(\boldsymbol{x}_i) + \boldsymbol{\xi}_i\right)}_{\boldsymbol{\xi}_i \text{ can be arbitrarily adversarial}} .$$

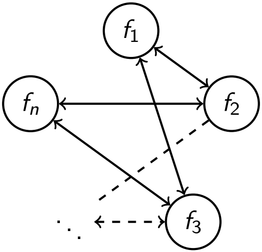
- The local data distribution on some agents can be poisoned. Examples:
  - *Empirical datasets*: replacing labels or specific features of the data.
  - *Streaming data*: sensor measurement corruptions.

# Byzantine fault v.s. gradient attacks



**Byzantine fault**

Byzantine agents themselves *deviate* from the predefined protocols




**Gradient attack**

Data attack only manipulates local functions

*Agents under gradient attacks still follow predefined algorithmic protocol*

# Algorithm development

*Q: How to deal with adversarial noise $\xi_i$?*

Main ideas:

- **Gradient clipping** to control the adversarial noise level on *attacked* agents.
- **Variance reduction (VR)** to approximate the true gradients on *unattacked* agents.

# CLIP-VRG

**Algorithm 0:** CLIP-VRG

**Input:** $\alpha_t, \gamma_t, \eta_t$.

**Initialization:** $\boldsymbol{x}_i^0 = \boldsymbol{x}_j^0, \forall i, j \in [n]$.

**for** $t = 0, \ldots, T-1$ **do**

  **for** *agent* $i \in [n]$ *in parallel* **do**

    Query stochastic gradient oracle that returns $\mathbf{m}_i^t$;

    Update $\mathbf{v}_i^t = \begin{cases} \mathbf{m}_i^t, & t = 0, \\ (1-\eta_{t-1})\mathbf{v}_i^{t-1} + \eta_{t-1}\mathbf{m}_i^t, & t \geq 1, \end{cases}$ (VR) ;

    Compute $k_i^t = \begin{cases} 1, & \|\mathbf{v}_i^t\| \leq \gamma_t, \\ \gamma_t \|\mathbf{v}_i^t\|^{-1}, & \|\mathbf{v}_i^t\| > \gamma_t, \end{cases}$ (Gradient clipping);

    Send $\boldsymbol{x}_i^t - \alpha_t k_i^t \mathbf{v}_i^t$ to all neighbors of agent $i$;

    Update $\boldsymbol{x}_i^{t+1} = \sum_{j=1}^n w_{ij} \left( \boldsymbol{x}_j^t - \alpha_t k_j^t \mathbf{v}_j^t \right)$;

  **end**

**end**

**Output:** $\{\boldsymbol{x}_i^T\}_{i \in [n]}$.

# Problem model

- A subset $[n] \backslash \mathcal{N}$ of agents receives *malicious stochastic gradients*, and we minimize $\sum_{i \in \mathcal{N}} f_i(\boldsymbol{x})$.
- Unattacked $f_i$'s are *convex* and *L-smooth*, $(1/|\mathcal{N}|) \sum_{i \in \mathcal{N}} f_i$ is *$\mu$-strongly convex*. The stochastic gradient $\nabla f_i(\boldsymbol{x}_i) + \boldsymbol{\xi}_i$ on $i \in \mathcal{N}$ satisfies that

$$\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{x}_i] = 0, \ \mathbb{E}\left[\|\boldsymbol{\xi}_i\|^2 \mid \boldsymbol{x}_i\right] \leq \sigma^2 \text{ for some } \sigma > 0.$$

- Unattacked functions $f_i$'s share one common minimizer, but this minimizer is not unique for any $f_i$.
- The fraction of attacked agents $\rho = 1 - |\mathcal{N}|/n < 1/(1 + L/\mu)$.
- The inter-agent communication graph is *undirected* and *connected*.

# CLIP-VRG

- Unattacked agents have access to noisy gradient $\mathbf{m}_i^t = \nabla f_i(\boldsymbol{x}_i^t) + \boldsymbol{\xi}_i^t$, while attacked ones receives arbitrary $\mathbf{m}_i^t$. For $\eta_t = c_\eta(t+\varphi)^{-\tau_\eta}$, compute VR based gradient estimator,

$$\mathbf{v}_i^t = (1-\eta_t)\mathbf{v}_i^{t-1} + \eta_t \mathbf{m}_i^t.$$

- Local clipped updates with decaying clipping threshold $\gamma_t = c_\gamma(t+\varphi)^{-\tau_\gamma}$, stepsize $\alpha_t = c_\alpha(t+\varphi)^{-\tau_\alpha}$,

$$\boldsymbol{x}_i^{t+\frac{1}{2}} = \boldsymbol{x}_i^t - \alpha_t k_i^t \mathbf{v}_i^t, \; k_i^t := \min\left(1, \gamma_t \left\|\mathbf{v}_i^t\right\|^{-1}\right).$$

- Using a *doubly stochastic* and *real symmetric* weight matrix $W$ with $|\lambda_2(W)| \in [0,1)$. Averaging with the iterates with neighbors,

$$\boldsymbol{x}_i^{t+1} = \sum_{j=1}^{n} w_{ij} \boldsymbol{x}_j^{t+\frac{1}{2}}.$$

# Convergence

**Theorem 1 (Yu and Kar 2023)**

*Under aforementioned assumptions, suppose that* $\alpha_t, \gamma_t, \eta_t \in (0,1)$ *are taken as* $\tau_\eta = 2(\tau_\alpha + \tau_\gamma)/3, 2\tau_\gamma < \tau_\alpha < \min(1, 1-\tau_\gamma)$*. Then, for all* $i \in [n]$*, for every* $0 < \tau < \min(\tau_\gamma, (\tau_\alpha - 2\tau_\gamma)/3)$*, we have*

$$\mathbb{P}\left(\lim_{t\to\infty} (t+1)^\tau \left\|\boldsymbol{x}_i^t - \boldsymbol{x}^*\right\| = 0\right) = 1.$$

**Corollary 2**

*We can take* $\tau_\alpha, \tau_\gamma, \tau_\eta$ *in Theorem 1 to achieve that for any* $i \in [n]$*, any* $\epsilon$ *with* $0 < \epsilon < 1/3$*,*

$$\mathbb{P}\Big(\lim_{t\to\infty} (t+1)^{1/3-\epsilon} (f(\boldsymbol{x}_i^t) - f(\boldsymbol{x}^*)) = 0\Big) = 1.$$

# Discussions

- Compared to Byzantine-robust case, we establish an exact convergence in a general topology. (e.g., (Gupta, Doan, and Vaidya 2021); (Wu, Chen, and Ling 2023))
- The assumption $\rho < 1/(1 + L/\mu)$ is tight in that we can find examples where $\rho = 1/(1 + L/\mu)$ leads to the failure of CLIP-VRG.
- Price: The best achievable rate $\mathcal{O}(t^{-1/3})$ is slower than the $\mathcal{O}(t^{-1})$ *almost sure* rate for algorithms designed for non-adversarial scenarios.
- The assumption that all functions share a common minimizer goes beyond the *independent and identically distributed (i.i.d.)* setting.

# Proof sketch

- The local iterate $\boldsymbol{x}_i^t$ converges to the *network average* iterate $\bar{\mathbf{x}}^t = (1/n)\sum_{i\in[n]} \mathbf{x}_i^t$.
- For regular agents $\mathcal{N}$, the recursive estimator $\mathbf{v}_i^t$ for the corresponding true gradient $\nabla f_i(\boldsymbol{x}_i^t)$ is strongly consistent.
- Case 1, if $\bar{\mathbf{x}}^t$ enters some contracting region, we show that $\bar{\mathbf{x}}^t$ would stay in this region and converge to $\boldsymbol{x}^*$ at the same sublinear rate as clipping threshold $\gamma_t$.
- Case 2, if $\bar{\mathbf{x}}^t$ never falls into case 1, then for each iteration $t$, we can lower bound the set of clipping coefficients $\{k_i^t : i \in \mathcal{N}\}$, that leads the $\bar{\mathbf{x}}^t$ sequence to be a time-varying contractive process with a controlled clipping bias.
- Combing the rates of case 1 and case 2.

# Experiments: heterogeneous measurements

Suppose each agent has observations $\mathbf{y}_i^t = \mathbf{H}_i\boldsymbol{\theta}_* + \mathbf{w}_i^t$ where $\mathbf{w}_i^t$ is *white noise*. We formulate an $\ell_2$ loss minimization problem over regular agents $\mathcal{N}$,

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^{625}} \sum_{i\in\mathcal{N}} \mathbb{E}_{\mathbf{w}_i} \left\|\mathbf{H}_i\boldsymbol{x} - \mathbf{y}_i\right\|^2 . \tag{1}$$

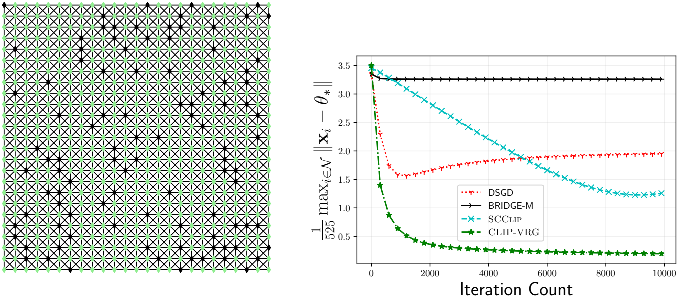


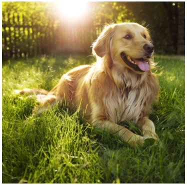

Figure: A 2d-grid network of agents with black agents have arbitrary adversarial measurements. Comparison of the performance of DSGD. BRIDGE-M, SCClip, CLIP-VRG.

# Experiments: collaborative learning

- Given the same datasets $\{\boldsymbol{\theta}_i, y_i\}$ for a binary classification task. Suppose each agent solves the same empirical risk minimization problem to

$$\ell(\boldsymbol{x}, \{\boldsymbol{\theta}_i, y_i\}_{i=1,\dots,n}) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-\boldsymbol{x}^\top \boldsymbol{\theta}_i y_i}\right) + \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2.$$

- For example: classifying cats and dogs.

# Experiments: collaborative learning

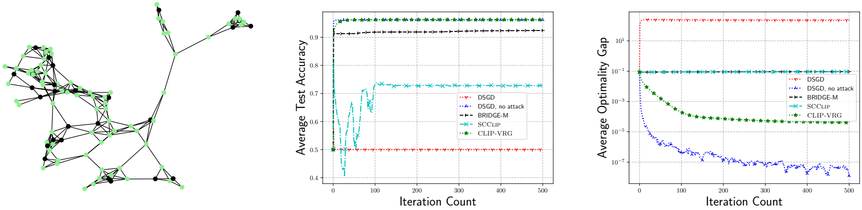



Figure: An undirected random geometric graph of 100 agents with Fashion-MNIST dataset. Performance comparison of DSGD, BRIDGE-M, SCClip, and CLIP-VRG under persistent gradient attacks; and DSGD without attack as baseline.

# Experiments: collaborative learning

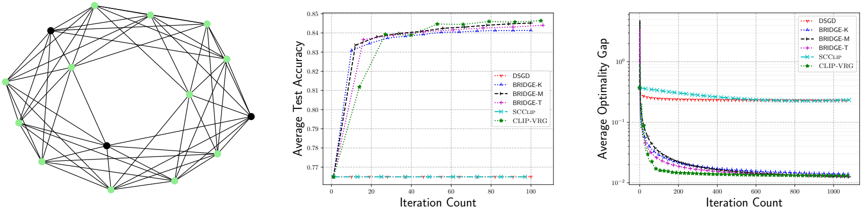



Figure: An connected cycle of 15 agents with a9a dataset. Performance comparison of DSGD, BRIDGE-K, BRIDGE-M, BRIDGE-T, SCCLIP and CLIP-VRG under persistent gradient attacks.

# Future research

- Extend the analysis to more *heterogeneous* case.
- Improve the *convergence rates* in both adversarial ($\rho > 0$) and non-adverarial case ($\rho = 0$).

# Reference

Gupta, Nirupam, Thinh T Doan, and Nitin H Vaidya (2021). “Byzantine fault-tolerance in decentralized optimization under 2f-redundancy”. In: *2021 American Control Conference (ACC)*. IEEE, pp. 3632–3637.

Wu, Zhaoxian, Tianyi Chen, and Qing Ling (2023). “Byzantine-resilient decentralized stochastic optimization with robust aggregation rules”. In: *IEEE Transactions on Signal Processing*.

Yu, Shuhua and Soummya Kar (2023). “Secure Distributed Optimization Under Gradient Attacks”. In: *IEEE Transactions on Signal Processing*.