Robust Machine Learning

Towards Scalable Learning from Untrusted Data

Nirupam Gupta

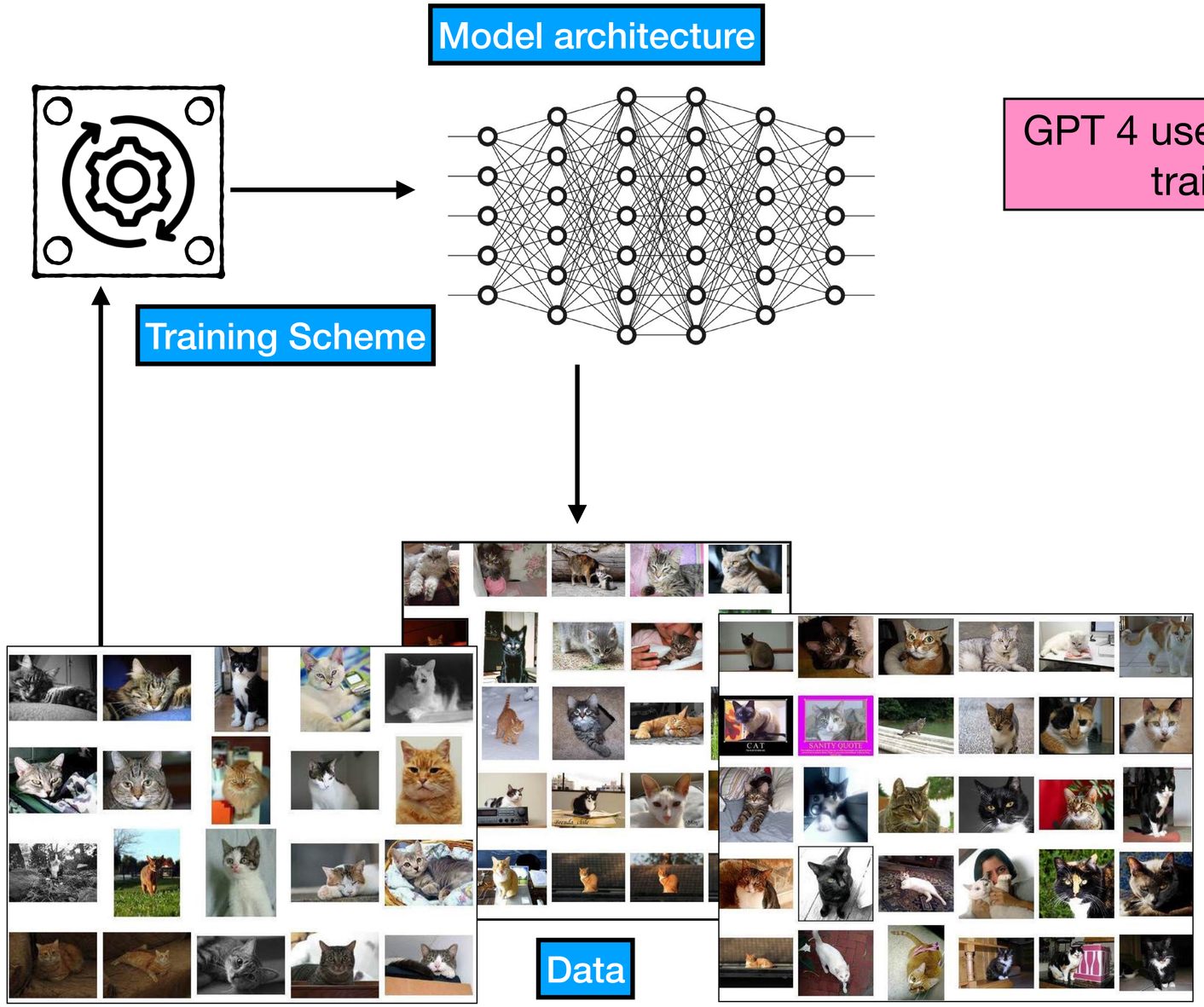
Assistant Professor
Department of Computer Science

UNIVERSITY OF COPENHAGEN

MACHINE LEARNING
UNIVERSITY OF COPENHAGEN

Learning Theory Summer School (LTSS), Copenhagen 2026

Ever-growing Scale of Machine Learning

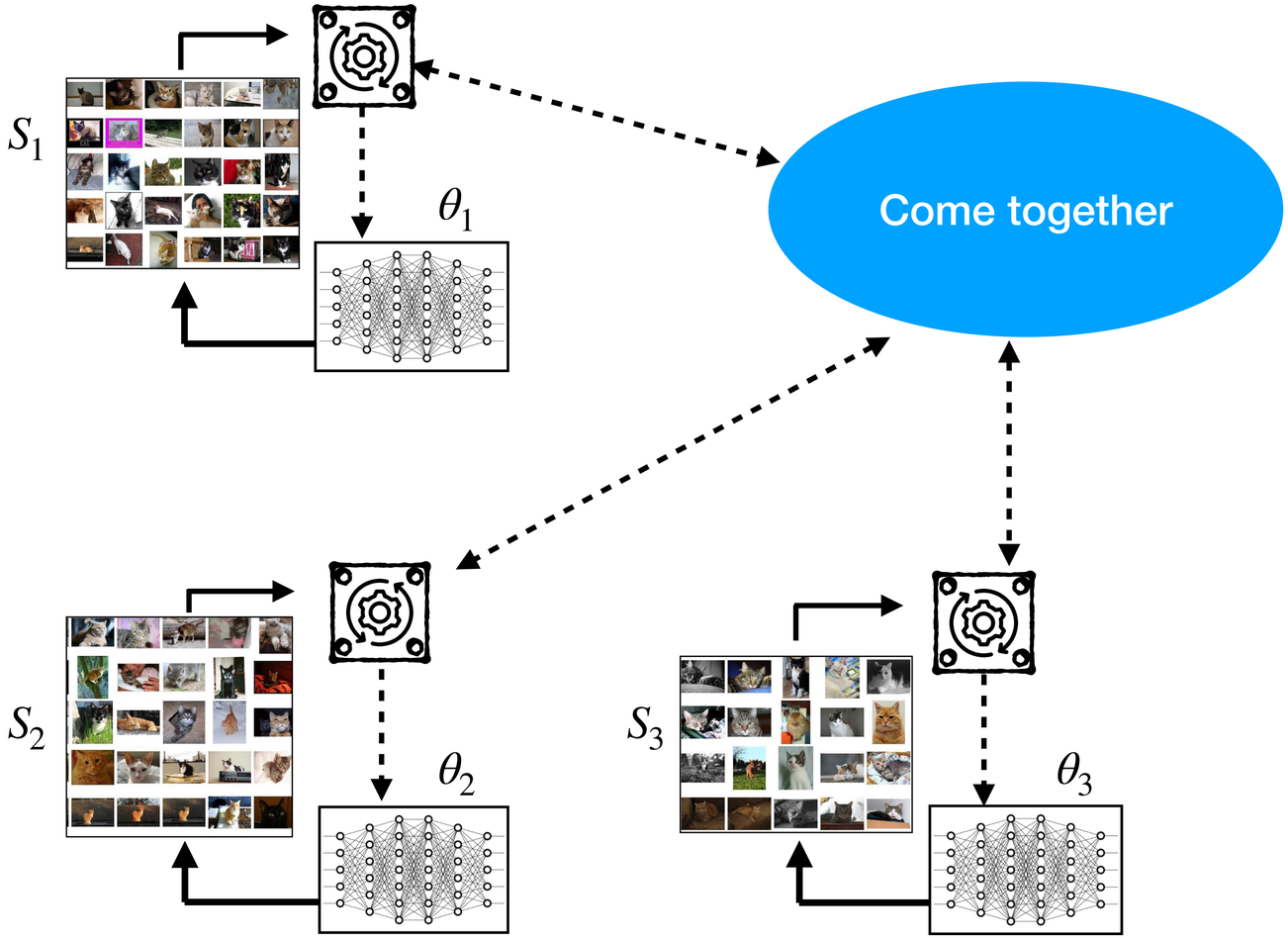


GPT 4 uses NN with **1.7 trillion parameters** trained over **1 PB of text data**

Greater Power

More Data
Larger Model

Distributed Learning: Model Training on Decentralized Data



$$\text{Loss}^{(i)}(\theta) := \frac{1}{m} \sum_{z \in S_i} \text{loss}(\theta, z)$$

Local loss function

Local samples

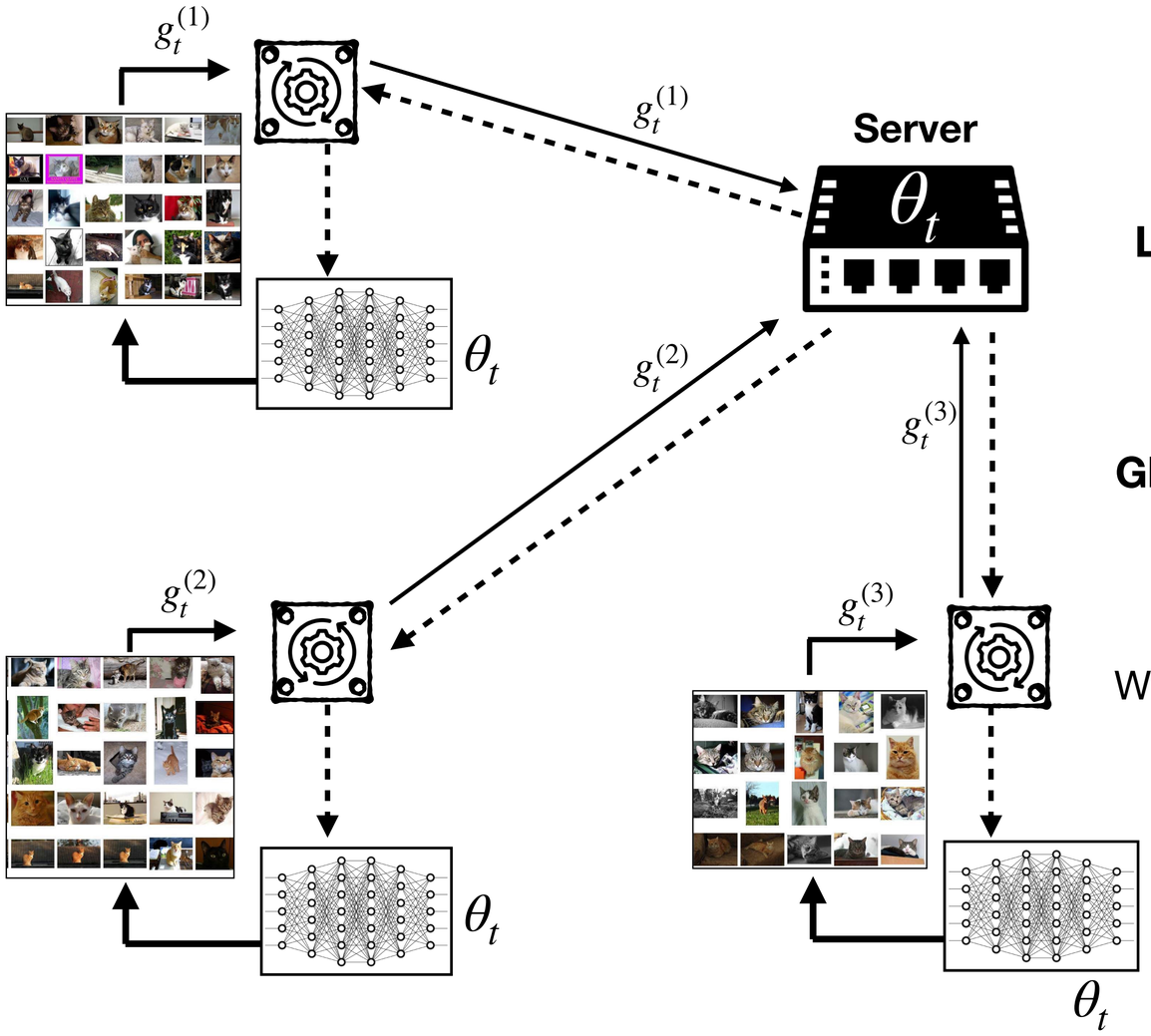
Goal: $\min_{\theta \in \Theta} \text{Loss}(\theta) \triangleq \frac{1}{n} \sum \text{Loss}^{(i)}(\theta)$

Benefits

Scalability
Data ownership control
Data privacy*
Democratization of ML
…

Distributed Stochastic Gradient Descent (D-SGD)

Special Case of Federated Averaging



Local Phase: Each *node* i computes $g_t^{(i)} := \nabla \text{loss}\left(\theta_t\ ,\ z_t^{(i)}\right)$

$z_t^{(i)} \sim U\left(S_i\right)$

Global Phase: The server updates the model:

$$\theta_{t+1} = \theta_t - \gamma_t\ \text{Avg}\left(g_t^{(1)}\ ,\ \cdots\ ,\ g_t^{(n)}\right)$$

When nodes hold identical data, this reduces to mini-batch GD.

$$\theta_t \longrightarrow \underset{\theta \in \Theta}{\arg\min}\ \text{Loss}\left(\theta\right)$$

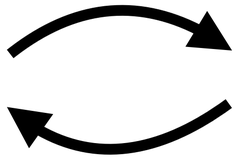
Learning Error Rate of D-SGD

Global Phase

Each *node* i computes $g_t^{(i)} := \nabla \text{loss}\left(\theta_t\ ,\ z_t^{(i)}\right)$

$z_t^{(i)} \sim U\left(S_i\right)$

Local Phase



The server updates the model:

$$\theta_{t+1} = \theta_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} g_t^{(i)}$$

Assumption I: Loss function is Lipschitz smooth, i.e., $\exists\ \lambda$ s.t. $\left\| \nabla \text{Loss}(\theta) - \nabla \text{Loss}(\theta') \right\| \leq \lambda \left\| \theta - \theta' \right\|$

$$\text{Loss}(\theta) - \text{Loss}(\theta') \leq \langle \nabla \text{Loss}(\theta'), \theta - \theta' \rangle + \frac{\lambda}{2} \left\| \theta - \theta' \right\|^2$$

Assumption II: Local gradients have bounded noise, $\exists\ \sigma$ s.t. $\left\| g_t^{(i)} - \mathbb{E}\left[g_t^{(i)}\right] \right\|^2 \leq \sigma^2$ where $\mathbb{E}\left[g_t^{(i)}\right] = \nabla \text{Loss}^{(i)}\left(\theta_t\right)$

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \left\| \nabla \text{Loss}\left(\theta_t\right) \right\|^2\right] \leq \mathcal{O}\left(\sqrt{\frac{\sigma^2}{nT}}\right)$$

Learning Error Rate of D-SGD (quick proof)

$$\mathbb{E}_t\left[\text{Loss}\left(\theta_{t+1}\right) - \text{Loss}\left(\theta_t\right)\right] \leq -\gamma_t\left(1 - \frac{\lambda\gamma_t}{2}\right)\left\|\nabla\text{Loss}\left(\theta_t\right)\right\|^2 + \gamma_t^2\frac{\lambda\sigma^2}{2n}$$

If $\gamma_t = \gamma \leq \frac{1}{\lambda}$ then $\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\text{Loss}\left(\theta_t\right)\right\|^2\right] \leq \frac{2}{\gamma}\mathbb{E}\left[\text{Loss}\left(\theta_1\right) - \text{Loss}\left(\theta_{T+1}\right)\right] + \gamma\frac{\lambda\sigma^2}{n}T$

Since $\text{Loss}\left(\theta_{T+1}\right) - \text{Loss}^*$ where $\text{Loss}^* = \min\text{Loss}(\theta)$, substituting $\gamma = 1/\left(\lambda\sqrt{T}\right)$,

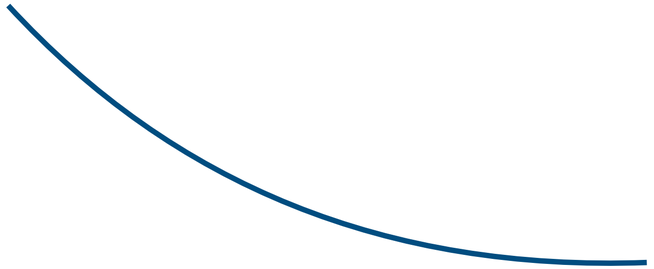
$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\text{Loss}\left(\theta_t\right)\right\|^2\right] \leq \left(2\lambda\left(\mathbb{E}\left[\text{Loss}\left(\theta_1\right)\right] - \text{Loss}^*\right) + \frac{\sigma^2}{n}\right)\frac{1}{\sqrt{T}}$$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla\text{Loss}\left(\theta_t\right)\right\|^2\right] \leq \mathcal{O}\left(\frac{1}{n\sqrt{T}}\right)$$

Learning Error Rate of D-SGD (cont'd)

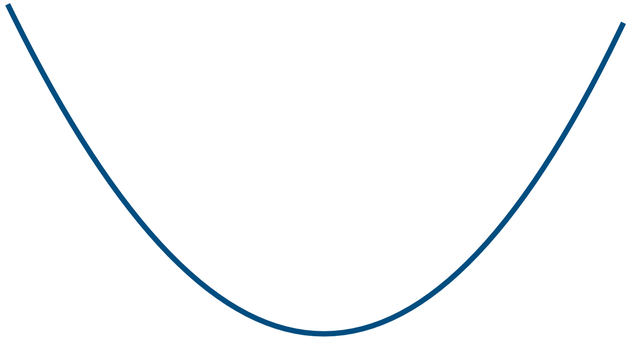
Convexity: $\text{Loss}(\theta') \geq \text{Loss}(\theta) + \langle \nabla \text{Loss}(\theta), \theta' - \theta \rangle$

$$\mathbb{E}\left[\text{Loss}\left(\theta_{T+1}\right) - \text{Loss}^*\right] \leq \mathcal{O}\left(\sqrt{\frac{\sigma^2}{nT}}\right)$$

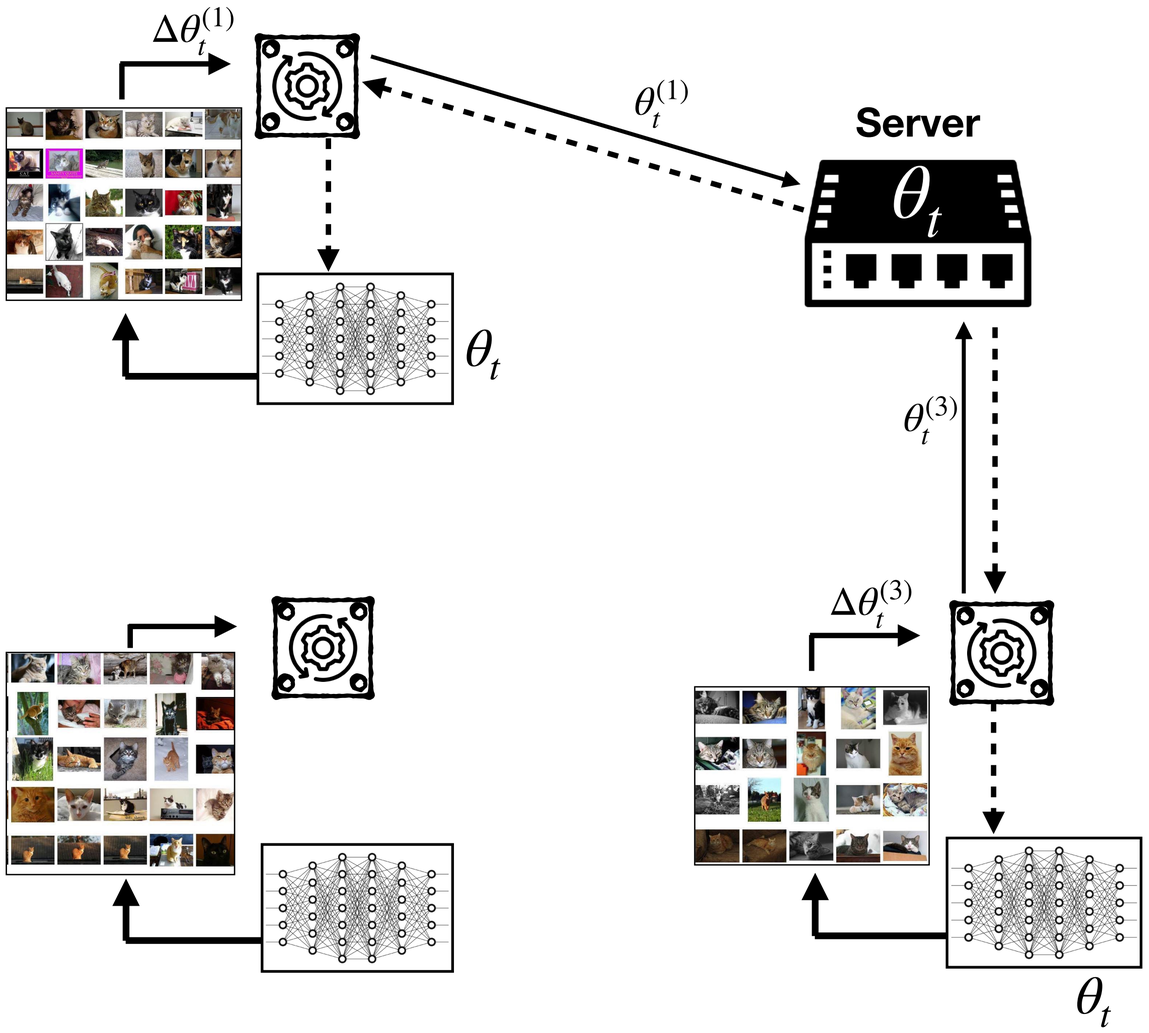


Strong convexity: $\text{Loss}(\theta') \geq \text{Loss}(\theta) + \langle \nabla \text{Loss}(\theta), \theta' - \theta \rangle + \mu \left\| \theta' - \theta \right\|^2$

$$\mathbb{E}\left[\text{Loss}\left(\theta_{T+1}\right) - \text{Loss}^*\right] \leq \mathcal{O}\left(\frac{\sigma^2}{nT}\right)$$



Federated Averaging: Local SGD



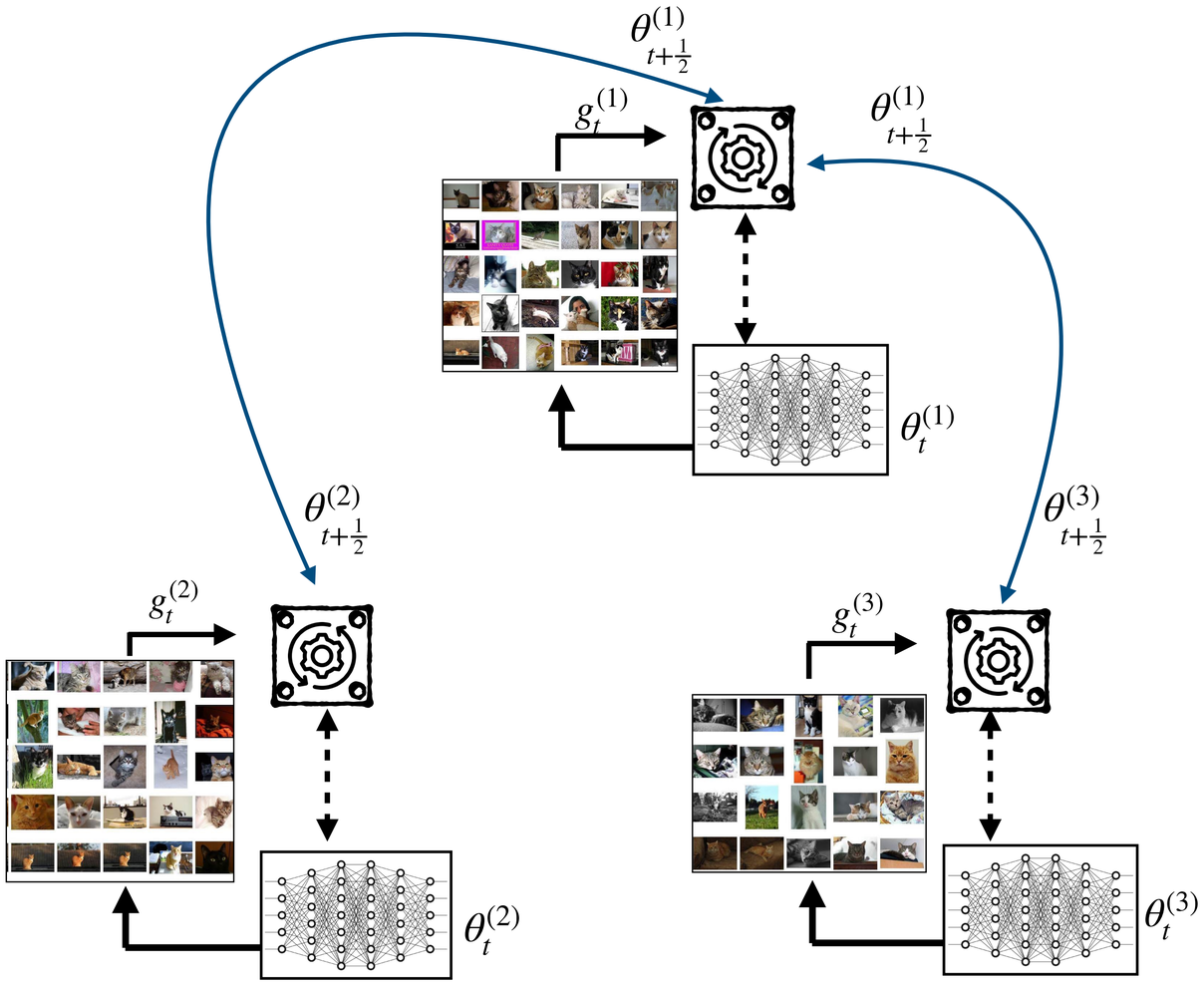
Partial participation: Server may only ask a few available clients to compute the updates.

Local Phase: Each selected client i computes an updated model $\theta_t^{(i)}$, upon running several local SGD steps.

Global Phase: The server updates the model:

$$\theta_{t+1} = \theta_t - \gamma_t \frac{1}{|S_t|} \sum_{i \in S_t} \left(\theta_t - \theta_t^{(i)}\right)$$

Distributed SGD in Peer-to-Peer Architecture (Serverless)



$$\min_{\theta^{(i)},\forall i} \quad \frac{1}{n}\sum \text{Loss}^{(i)}\left(\theta^{(i)}\right)$$

$$\text{Subject to}: \quad \theta^{(i)} = \theta^{(j)} \ , \ \forall i, j$$

Phase I: Each *node* i updates the local model:

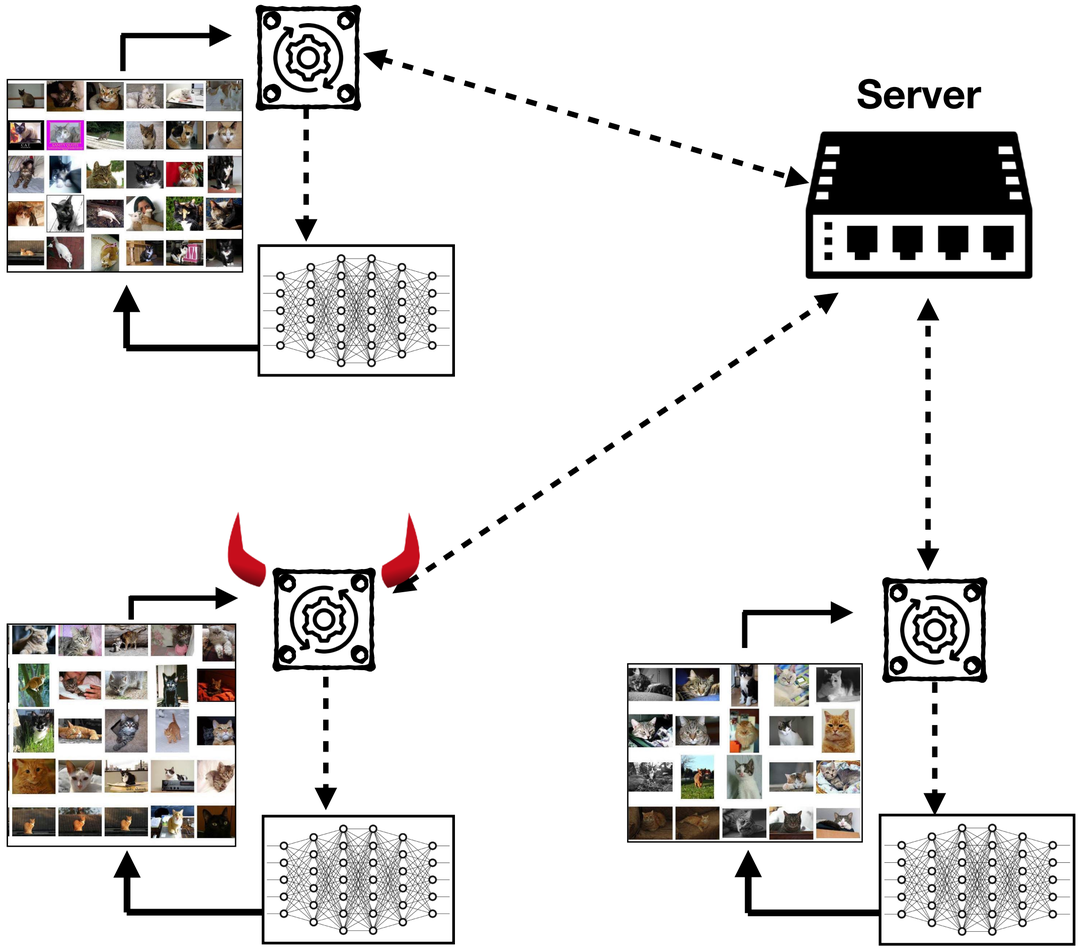
$$\theta^{(i)}_{t+\frac{1}{2}} = \theta^{(i)}_t - \gamma_t \ g^{(i)}_t$$

Phase II: Nodes gossip:

Gossip rate

$$\theta^{(i)}_{t+1} = \theta^{(i)}_{t+\frac{1}{2}} - \alpha_t \sum_{j\in\mathcal{N}_t^{(i)}} \left(\theta^{(i)}_{t+\frac{1}{2}} - \theta^{(j)}_{t+\frac{1}{2}}\right)$$

Vulnerability to Data & Model Poisonings

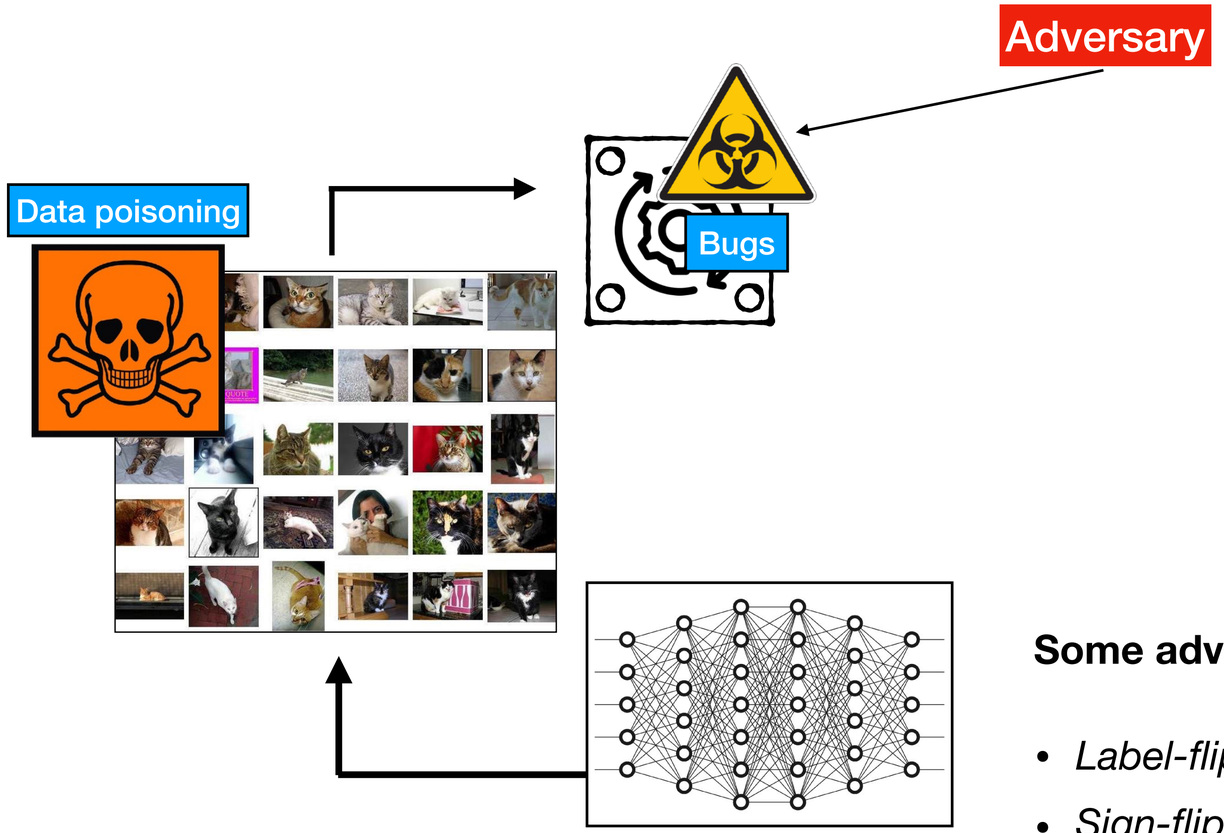


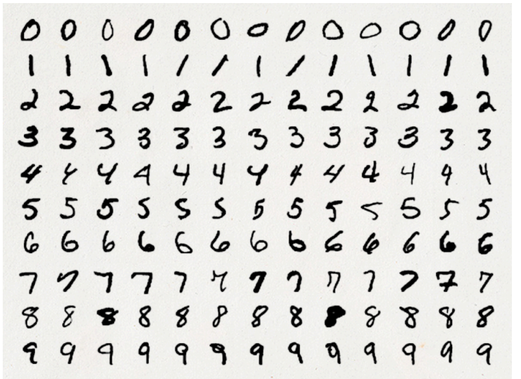
Not all data/machines can be certified (trusted)

$$\theta_{t+1} = \theta_t - \gamma_t \, \mathsf{Avg}\left(g_t^{(1)}, \ \ldots \ , g_t^{(n)}\right)$$

Breaks down when some machines “misbehave”

Common sources of poisoning



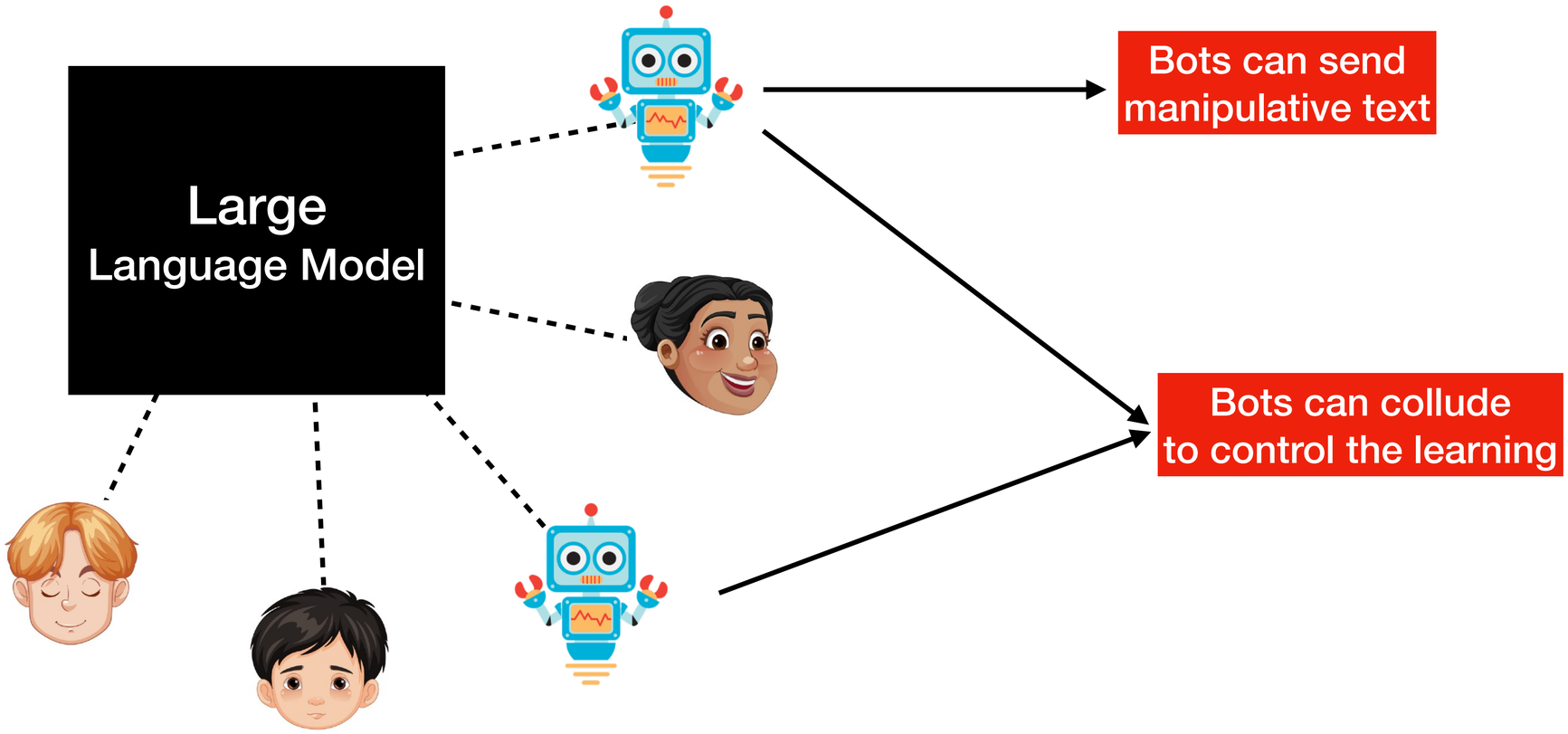


Some adversarial behaviors:

- *Label-flipping* ***(data poisoning)****:* swap the labels, e.g., $(4\ ,\ 9) \rightarrow (9\ ,\ 4)$
- *Sign-flipping* ***(model poisoning)****:* toggle gradient's sign $g_t^{(i)} \rightarrow -\ g_t^{(i)}$

Data Poisoning is Common in the Age of Bots

Some fake news is created not for humans, but for misleading language models



Bots are Spreading Faster & Getting Smarter

Communications of the ACM, 2016

818717

Today's social bots are sophisticated and sometimes menacing. Indeed, their presence can endanger online ecosystems as well as our society.

BY EMILIO FERRARA, ONUR VAROL, CLAYTON DAVIS, FILIPPO MENCZER, AND ALESSANDRO FLAMMINI

The Rise of Social Bots

"exhibit human-like behavior"

CNBC, 2017

(~ 300 million total)

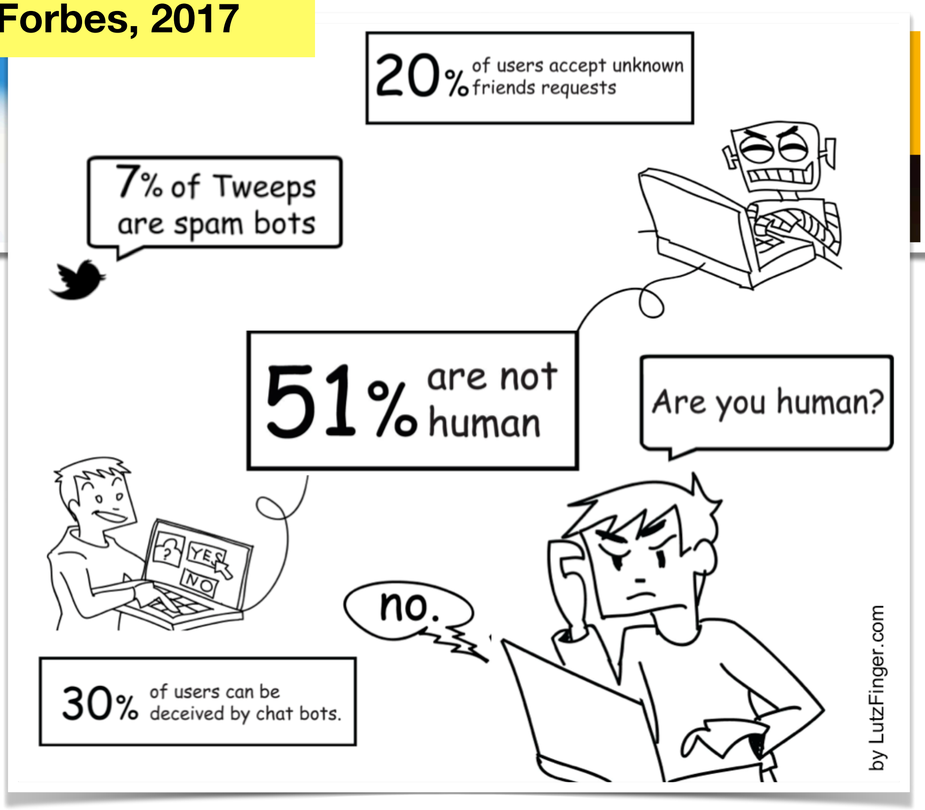
As many as 48 million Twitter accounts aren't people, says study

PUBLISHED FRI, MAR 10 2017•1:09 PM EST | UPDATED FRI, MAR 10 2017•7:56 PM EST

Michael Newberg
@MIKENEWBERG

SHARE

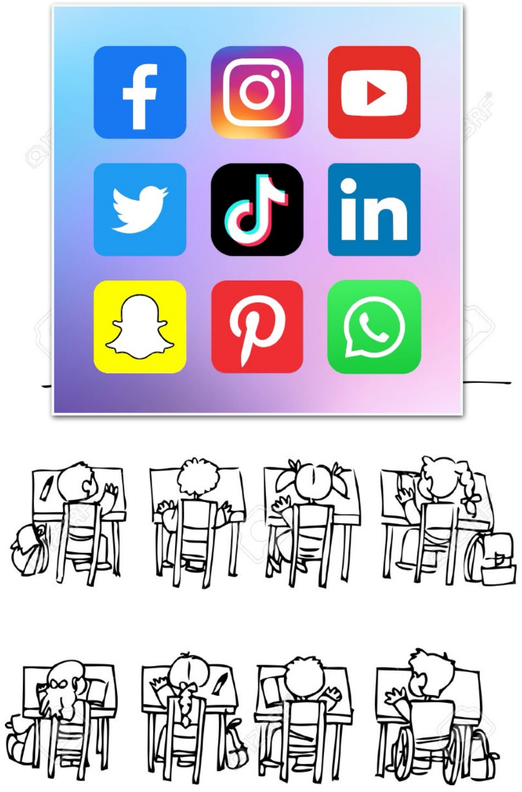
Forbes, 2017



Can we Learn Without Trusting the Source?

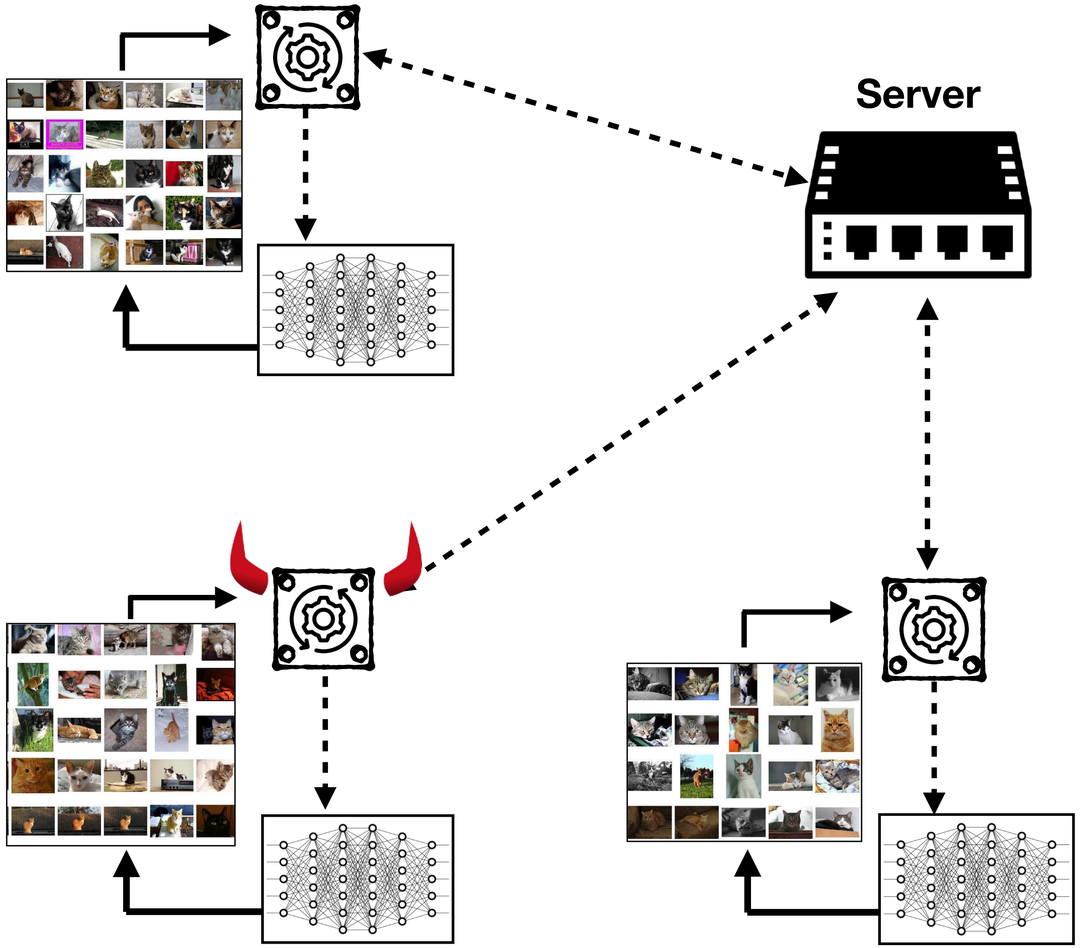
Copyright: Carla Francesca Castagno

Vs



Copyright: Carla Francesca Castagno

(Byzantine-)Robust Machine Learning



f nodes (silos) *of unknown identity* are **adversarial**

$< n/2$

Need not follow the algorithm

$$\text{Loss}^{(i)}(\theta) := \frac{1}{m} \sum_{z \in S_i} \text{loss}(\theta, z)$$

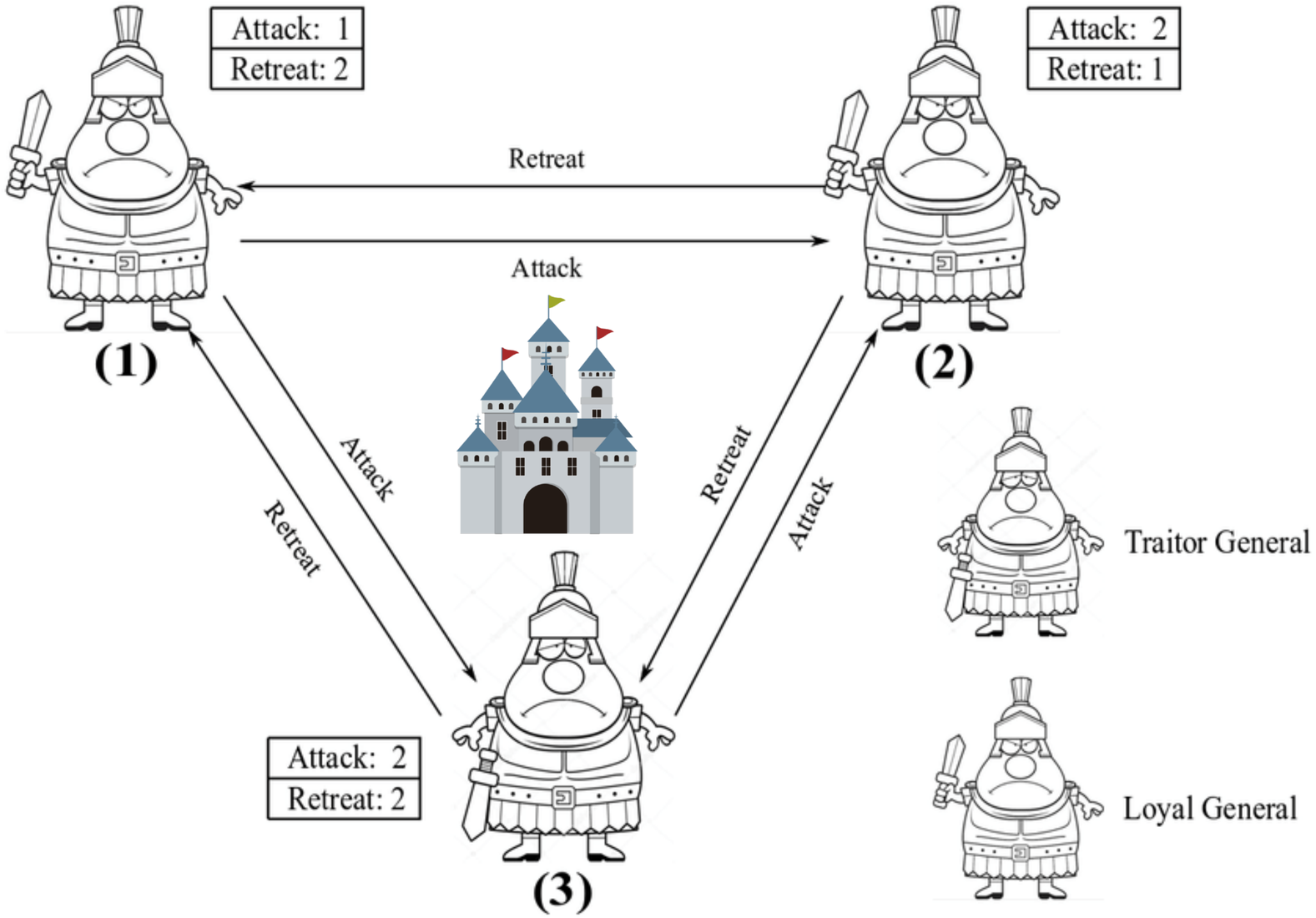
Robust DL Goal: $$\min_{\theta \in \Theta} \frac{1}{n-f} \sum_{i \in H} \text{Loss}^{(i)}(\theta)$$

Set of *honest* nodes

Unknown!

Byzantine Generals' Problem

- Leslie Lamport (*Turing Award - 2013*)

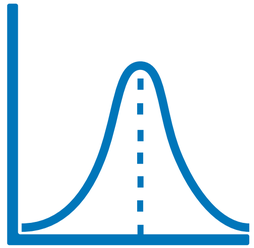


Three Challenges in Robust FL

State machine replication is not efficient

Local Randomness

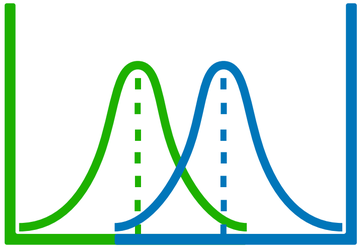
Honest nodes/clients *independently* compute noisy updates



Adversarial nodes can camouflage as honest ones

Data Heterogeneity

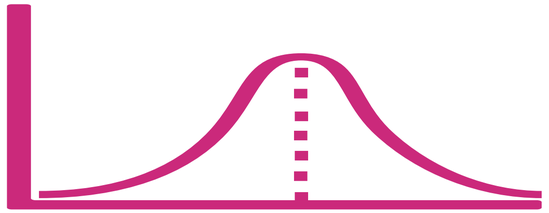
Different nodes sample data from different distributions



Local randomness gets amplified

Data Privacy

Nodes do not share exact information of their data

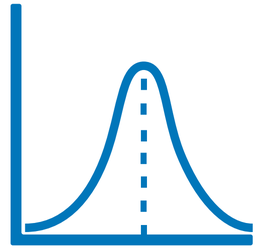


Challenge I: Tackling Local Randomness

State machine replication is not efficient

Local Randomness

Honest nodes independently compute noisy updates



Mini-batch GD

$$S_i = S \quad \forall\ i \in H$$

$$\text{Loss}^i(\theta) = \text{Loss}(\theta) = \frac{1}{m} \sum_{z \in S} \nabla \text{loss}\,(\theta\ ,\ z)$$

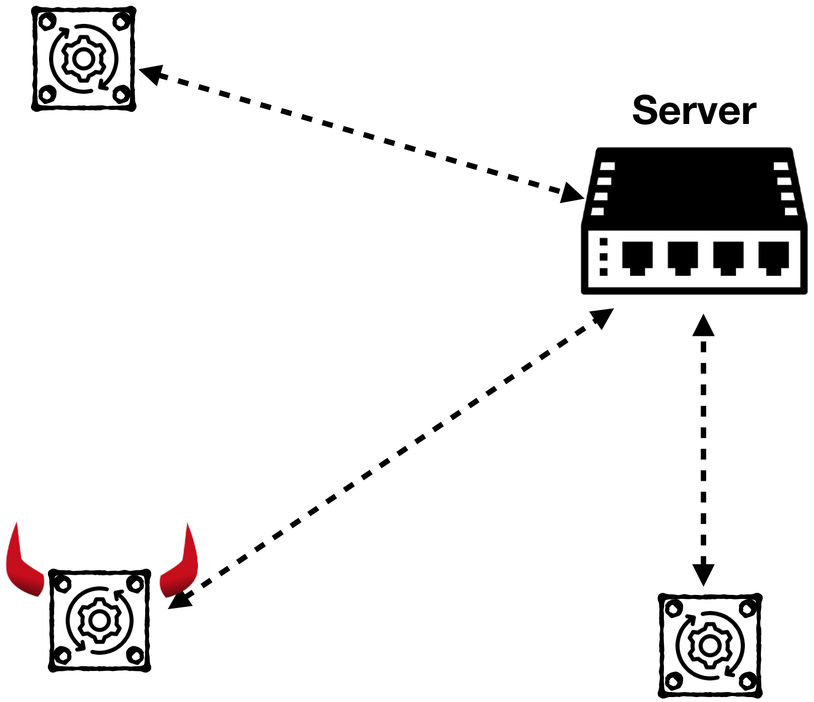
Can be exploited by adversarial nodes

$$g_t^{(i)} := \nabla \text{loss}\left(\theta_t\ ,\ z_t^{(i)}\right)$$

where $z_t^{(i)} \sim \mathcal{U}\,(S)$

Noisy estimate → $\nabla \text{Loss}(\theta_t)$

Robust Mini-batch Gradient Descent



$f < \frac{n}{2}$ nodes are adversarial

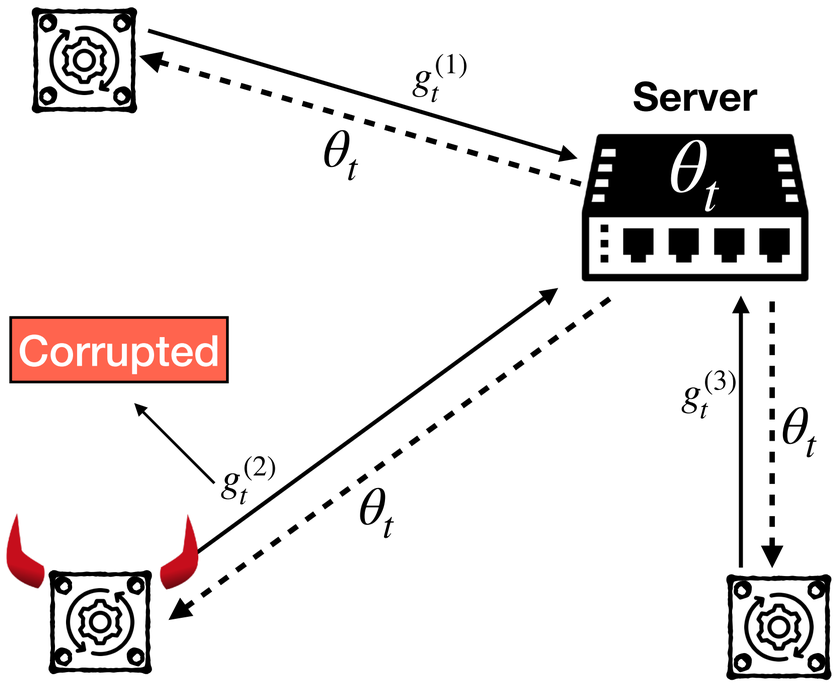
$S_i = S \quad \forall\ i \in H \qquad \text{Loss}^i(\theta) = \text{Loss}(\theta) = \frac{1}{m} \sum_{z \in S} \text{loss}\,(\theta\ ,\ z)$

Minimize $\quad \text{Loss}(\theta) := \frac{1}{m} \sum_{z \in S} \text{loss}\,(\theta\ ,\ z)$

Problem reduces to estimating $\nabla \text{Loss}(\theta)$

Diakonikolas, Ilias, et al. "Being robust (in high dimensions) can be practical." *International Conference on Machine Learning*. PMLR, 2017.

Shortcoming of Classic Mini-batch Gradient Descent



Local Phase: Each *node* i computes $g_t^{(i)} := \nabla \text{loss}\left(\theta_t\ ,\ z_t^{(i)}\right)$ where $z_t^{(i)} \sim \mathcal{U}\,(S)$

Global Phase: The server updates the model:

$$\theta_{t+1} = \theta_t - \gamma_t\ \text{Avg}\left(g_t^{(1)}\ ,\ \cdots\ ,\ g_t^{(n)}\right)$$

Can be arbitrarily corrupted by an adversary

From Classic to Robust Mini-Batch GD

$$\text{Avg}\left(g_t^{(1)}, \ldots, g_t^{(n)}\right) \quad \Rightarrow \quad \text{RobAgg}\left(g_t^{(1)}, \ldots, g_t^{(n)}\right)$$

Robust to some corrupted gradients

Examples: $d = 2,\ n = 5$ and $f = 1$. Let $\left(g_t^{(1)}, \ldots, g_t^{(5)}\right) = \begin{pmatrix} 2 & 3 & 2 & 1 & 12 \\ 5 & 6 & 4 & 5 & 0 \end{pmatrix}$

Coordinate-wise Median **(CWMed)**

$$\text{CWMed}\begin{pmatrix} 2 & 3 & 2 & 1 & 12 \\ 5 & 6 & 4 & 5 & 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$$

Averaging

$$\text{Avg}\begin{pmatrix} 2 & 3 & 2 & 1 & 12 \\ 5 & 6 & 4 & 5 & 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$$

Coordinate-wise Trimmed Mean **(CWTM)**

$$\text{CWTM}\begin{pmatrix} 2 & 3 & 2 & \cancel{1} & \cancel{12} \\ 5 & \cancel{6} & 4 & 5 & \cancel{0} \end{pmatrix} = \begin{pmatrix} 2.3 \\ 4.6 \end{pmatrix}$$

Robust Mini-batch GD — Reducible to Robust Mean

$$\text{RobAgg}\left(g_t^{(1)}, \ldots, g_t^{(n)}\right); \quad g_t^{(i)} := \nabla \text{loss}\left(\theta_t, z\right), \quad z \sim \mathcal{U}(S)$$

Estimate the true mean

$n - f$ honest gradients → In majority

$$\text{Minimize} \quad \mathbb{E}\left[\left\|\text{RobAgg}\left(g_t^{(1)}, \ldots, g_t^{(n)}\right) - \nabla \text{Loss}\left(\theta_t\right)\right\|^2\right]$$

$$\nabla \text{Loss}(\theta_t) = \mathop{\mathbb{E}}_{z \sim \mathcal{U}(S)}\left[\nabla \text{loss}\left(\theta_t, z\right)\right]$$

Mean of honest gradients

$$\text{Bounded “variance”:} \quad \mathbb{E}_{z \sim \mathcal{U}(S)}\left[\left\|\nabla \text{loss}\left(\theta_t, z\right) - \nabla \text{Loss}(\theta_t)\right\|^2\right] \leq \sigma^2$$

(f, κ)-Robust Averaging: A Solution to Robust Estimation

Aggregation $F : (v_1, \ldots, v_n) \mapsto v$ $\quad \forall G \subseteq [n]$ of size $n - f$

$$\|v - \bar{v}_G\|^2 \leq \kappa \left(\frac{1}{n-f} \sum_{i \in G} \|v_i - \bar{v}_G\|^2 \right)$$

Robustness coefficient

$$\bar{v}_G = \frac{1}{n-f} \sum_{i \in G} v_i$$

Empirical “variance”

Sanity check: If $n - f$ vectors are identical the output is that vector.
If honest vectors are separated by distance δ, the error is in $\mathcal{O}\left(\kappa\ \delta^2\right)$

When RobAgg $\left(g_t^{(1)}\ ,\ \ldots\ ,\ g_t^{(n)}\right)$ is (f, κ)-robust, we have

$$\mathbb{E}\left[\left\|\text{RobAgg}\left(g_t^{(1)}\ ,\ \ldots\ ,\ g_t^{(n)}\right)\ -\ \nabla \text{Loss}\left(\theta_t\right)\right\|^2\right] \leq \mathcal{O}\left(\left(\kappa + \frac{1}{n}\right)\ \sigma^2\right)$$

(f, κ)-Robust Averaging (cont'd)

Aggregation $F : (v_1, \ldots, v_n) \mapsto v$ $\xrightarrow{\forall G \subseteq [n] \text{ of size } n-f}$ $\|v - \bar{v}_G\|^2 \leq \kappa \left(\frac{1}{n-f} \sum_{i \in G} \|v_i - \bar{v}_G\|^2 \right)$

$$\mathbb{E}\left[\left\| \text{RobAgg}\left(g_t^{(1)}, \ldots, g_t^{(n)}\right) - \frac{1}{n-f} \sum_{i \in H} g_t^{(i)} \right\|^2 \right] \leq 4\kappa\sigma^2$$

$$\mathbb{E}\left[\left\| \text{CWTM}\left(g_t^{(1)}, \ldots, g_t^{(n)}\right) - \frac{1}{n-f} \sum_{i \in H} g_t^{(i)} \right\|^2 \right] \in \mathcal{O}\left(\frac{f}{n}\sigma^2\right)$$

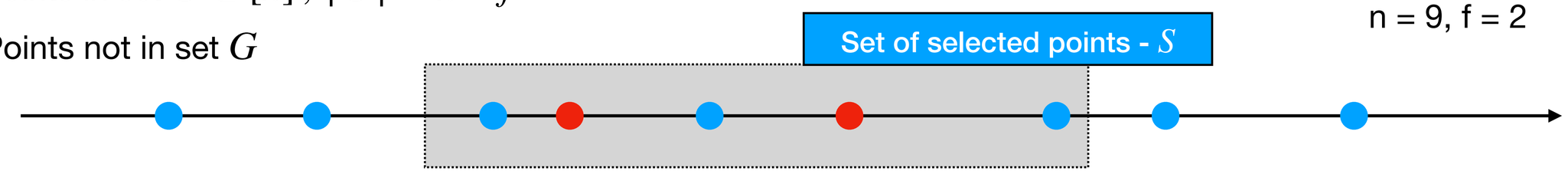
$$f < \frac{n}{2} \qquad \kappa \in \Omega\left(\frac{f}{n}\right)$$

Agg.	κ	When $f \ll n$
CWMed	$4\left(1 + \frac{f}{n-2f}\right)^2$	$\mathcal{O}(1)$
CWTM	$\frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right)$	$\mathcal{O}\left(\frac{f}{n}\right)$

Allouah, Youssef, et al. "Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity." *Int. Conf. on Artificial Intelligence and Statistics*. PMLR, 2023.

(f, κ)-Robustness of Trimmed Mean

- Points in set $G \subset [n]$, $|G| = n - f$
- Points not in set G



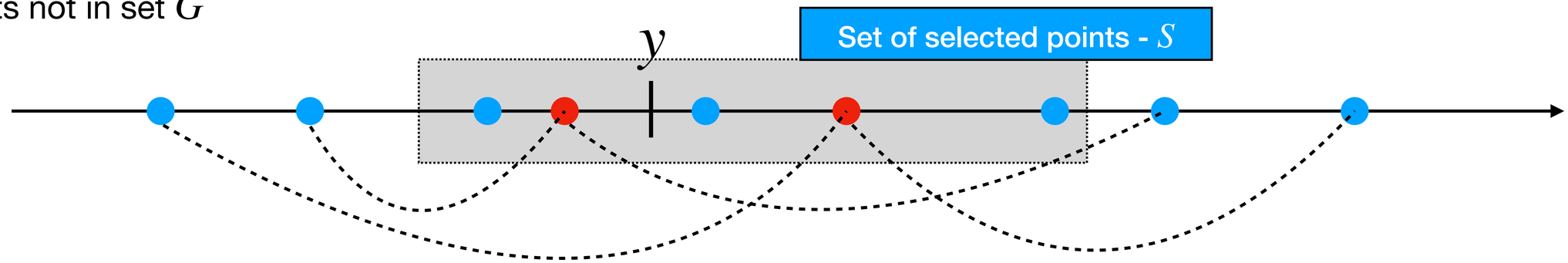
Let $v_1 \leq \ldots \leq v_n$, then $\hat{v} = \mathrm{TM}(v_1, \ldots, v_n) = \dfrac{1}{n-2f} \sum_{i=f}^{n-f-1} v_i$

$$\left(\hat{v} - \bar{v}_G\right)^2 = \frac{1}{(n-2f)^2} \left(\sum_{i \in S} v_i - \bar{v}_G \right)^2 = \frac{1}{(n-2f)^2} \left(\sum_{i \in S \setminus G} v_i - \bar{v}_G + \sum_{i \in S \cap G} v_i - \bar{v}_G \right)^2$$

$$\left(\hat{v} - \bar{v}_G\right)^2 = \frac{1}{(n-2f)^2} \left(\sum_{i \in S \setminus G} v_i - \bar{v}_G - \sum_{i \in G \setminus S} v_i - \bar{v}_G \right)^2 \leq \frac{|S \setminus G| + |G \setminus S|}{(n-2f)^2} \left(\sum_{i \in S \setminus G} \left(v_i - \bar{v}_G\right)^2 + \sum_{i \in G \setminus S} \left(v_i - \bar{v}_G\right)^2 \right)$$

(f, κ)-Robustness of Trimmed Mean (Cont'd)

- Points in set $G \subset [n]$, $|G| = n - f$
- Points not in set G



$$\left(\hat{v} - \bar{v}_G\right)^2 \leq \underbrace{\frac{|S\backslash G| + |G\backslash S|}{(n-2f)^2}}_{\leq 3f} \left(\sum_{i \in S\backslash G} \left(v_i - \bar{v}_G\right)^2 + \sum_{i \in G\backslash S} \left(v_i - \bar{v}_G\right)^2 \right)$$

$$\forall j \in S\backslash G\ ,\ \exists\, i_j, \in G\backslash S\ , \quad \textbf{such that} \quad (v_j - y)^2 \leq (v_{i_j} - y)^2$$

$$\left(\hat{v} - \bar{v}_G\right)^2 \leq \frac{3f}{(n-2f)^2} \left(\sum_{i \in G\backslash S} \left(v_i - \bar{v}_G\right)^2 + \sum_{i \in G\backslash S} \left(v_i - \bar{v}_G\right)^2 \right) \leq \frac{6f(n-f)}{(n-2f)^2} \left(\frac{1}{n-f} \sum_{i \in G} \left(v_i - \bar{v}_G\right)^2 \right)$$

κ

(f, κ)-Robustness of Geometric Median

$$\hat{v} = \mathrm{GM}(v_1, \ldots, v_n) = \min_{v} \sum_{i=1}^{n} \|v_i - v\|$$

For all $i \in G$, use: $\|\hat{v} - \bar{v}_G\| \leq \|\hat{v} - v_i\| + \|v_i - \bar{v}_G\|$

For all $i \in [n] \backslash G$, use: $\|\hat{v} - \bar{v}_G\| \geq \|\bar{v}_G - v_i\| - \|v_i - \hat{v}\|$

Niceness of (f, κ)-Robust Averaging

Aggregation	Breakdown	Robustness Coeff. κ	When $f \ll n$
CWMed	$\frac{1}{2}$	$4\left(1+\frac{f}{n-2f}\right)^2$	$\mathcal{O}(1)$
CWTM	$\frac{f+1}{n}$	$\frac{6f}{n-2f}\left(1+\frac{f}{n-2f}\right)$	$\mathcal{O}\left(\frac{f}{n}\right)$
GeoMed	$\frac{1}{2}$	$4\left(1+\frac{f}{n-2f}\right)^2$	$\mathcal{O}(1)$
Krum	$\frac{f+1}{n}$	$6\left(1+\frac{f}{n-2f}\right)^2$	$\mathcal{O}(1)$

Robustness analysis is agnostic to input distribution.

Incorporate input pre-processing (e.g., clipping) easily.

Extends easily to heterogeneous settings.

Allouah, Youssef, et al. "Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity." *Int. Conf. on Artificial Intelligence and Statistics*. PMLR, 2023.

Learning Error Rate of Robust Mini-batch GD

$$\theta_{t+1} = \theta_t - \gamma_t \, \text{RobAgg}\left(g_t^{(1)}, \ \dots \ , g_t^{(n)}\right)$$

(f, κ)-Robust

$\text{Loss}(\theta)$ → Lipschitz smooth

$\text{Loss}(\theta)$ → Strongly convex

$$\mathbb{E}_{z \sim \mathcal{U}(S)}\left[\|\nabla \text{loss}(\theta, z) - \nabla \text{Loss}(\theta)\|^2\right] \leq \sigma^2$$

Minimum value

After T iterations:

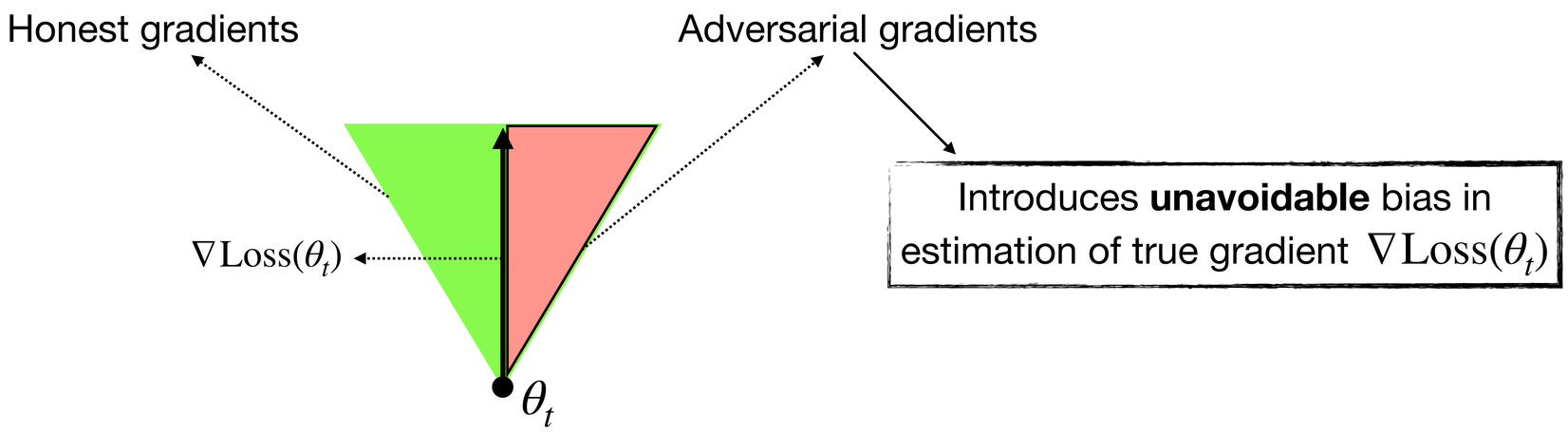
$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \in \mathcal{O}\left(\frac{\sigma^2}{(n-f)T} + \kappa\sigma^2\right)$$

DSGD with $n - f$ honest nodes

Gradient noise

Can We Do Better?

The squared-error in estimating $\nabla \text{Loss}\left(\theta_t\right)$ is in $\Omega\left(\frac{f}{n}\,\sigma^2\right)$ → Trimmed-Mean is optimal



Introduces **unavoidable** bias in estimation of true gradient $\nabla \text{Loss}(\theta_t)$

Why Can't We Do Better?

n points sampled from $P' = \left(1 - \frac{f}{n}\right) P + \frac{f}{n} Q$, where $P = 0$ w.p. 1, and $Q = \sigma\sqrt{\frac{n}{f}}$ w.p. 1.

An algorithm $\mathscr{A}$ cannot tell if the honest distribution is either P or P',

Huber's contamination model

Both p and p' are valid honest distributions: $\sigma_p^2 = 0$ and $\sigma_{p'}^2 = \sigma^2\left(1 - \frac{f}{n}\right)$.

We have $\mu_p = 0$ and $\mu_{p'} = \sigma\sqrt{\frac{f}{n}}$. Thus, $|\mu_p - \mu_{p'}|^2 = \sigma^2 \frac{f}{n}$.

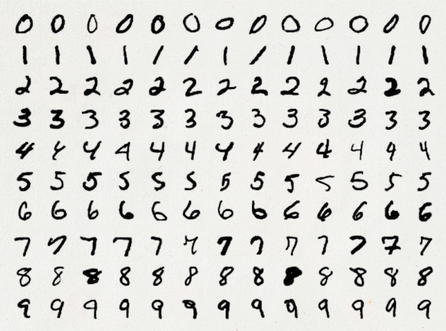
Therefore, $\mathscr{A}$ incurs an error overhead in estimating the true mean $\in \Omega\left(\sigma^2 \frac{f}{n}\right)$.

Overall error $\in \Omega\left(\sigma^2\left(\frac{1}{n} + \frac{f}{n}\right)\right)$.

Non-Vanishing Robustness Error is Problematic in Practice

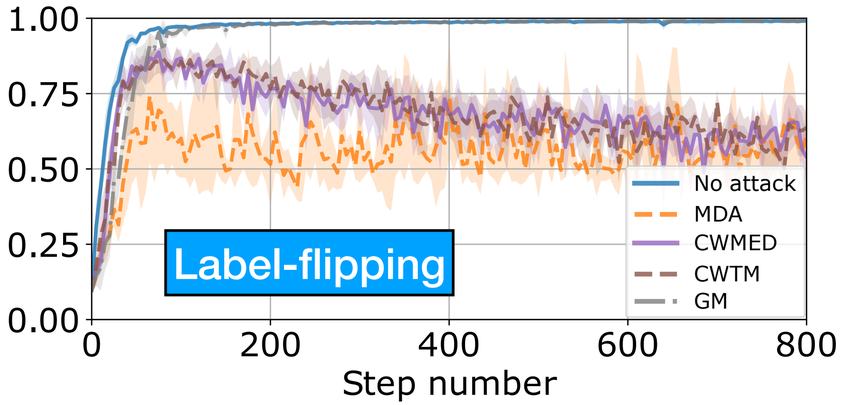
$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \in \mathcal{O}\left(\frac{\sigma^2}{(n-f)T} + \kappa\sigma^2\right)$$

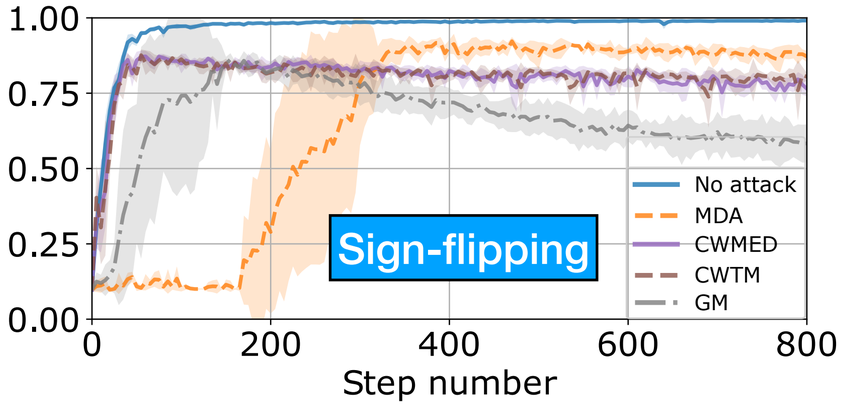
Non-negligible in practice



Training **CNN** using $n = 15$ nodes where $f = 5$ nodes are adversarial

MNIST dataset is equally divided amongst the honest nodes





Mitigating the Impact of Noise

$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \in \mathcal{O}\left(\frac{\sigma^2}{(n-f)T} + \kappa\sigma^2\right)$$

Reduce the noise

Few Ideas -

Increase local batch-sizes → Imposes high computation cost

Stochastic variance reduction → Requires computing full-batch gradients w.p. > 0

Incorporate Local Gradient Momentum

Assumption: Identity of corrupted nodes is fixed throughout the training procedure

Local Phase: Each honest *node* i computes $g_t^{(i)} := \nabla \text{loss}\left(\theta_t\ ,\ z_t^{(i)}\right)$ and computes (Polyak's) momentum:

$$m_t^{(i)} = \beta\, m_{t-1}^{(i)} \ +\ (1-\beta)\, g_t^{(i)}$$

Past gradients

Global Phase: $\theta_{t+1} = \theta_t - \gamma_t\ \text{RobAvg}\left(m_t^{(1)}\ ,\ \ldots\ ,\ m_t^{(n)}\right)$

Introduces bias $\approx \beta$

Honest variance reduces by factor $(1-\beta)$

Controlled momentum rate: $\beta = 1 - c\,\gamma_t$

Byzantine Machine Learning Made Easy by Resilient Averaging of Momentums
FGGPS. ICML, 2022.

$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \ \in\ \mathcal{O}\left(\left(\frac{1}{n-f} + \kappa\right)\frac{\sigma^2}{T}\right)$$

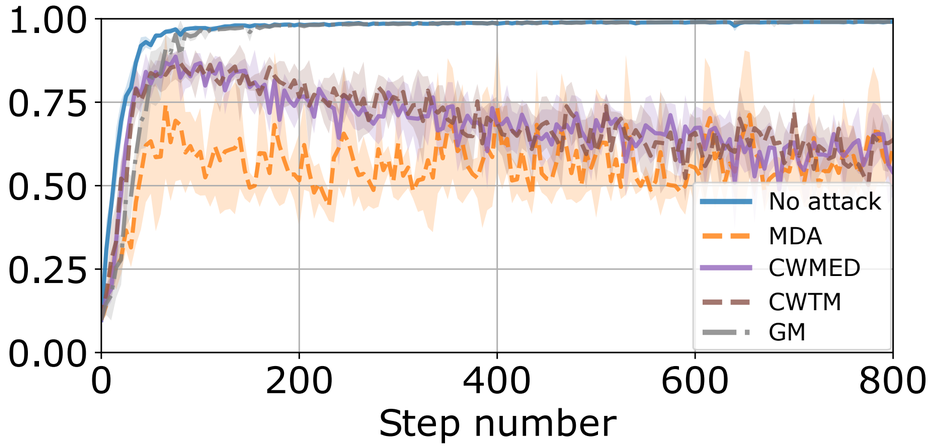
Augments gradient complexity

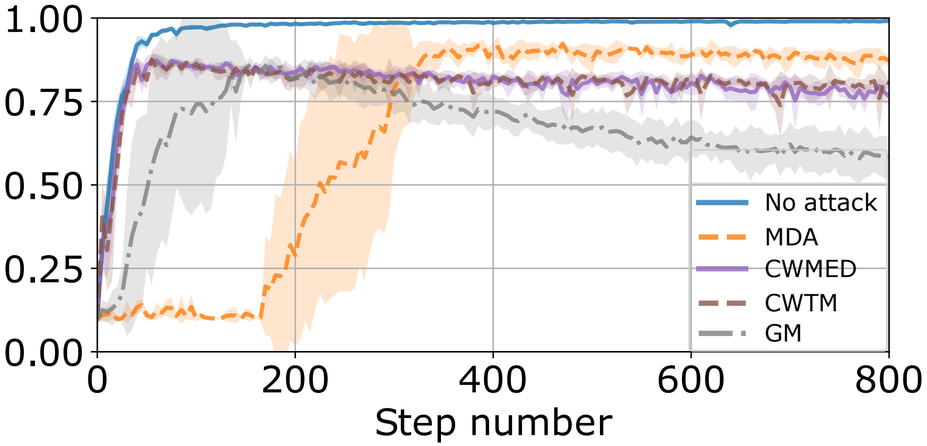
Trimmed-mean

Optimal when $\kappa \in \mathcal{O}\left(f/n\right)$

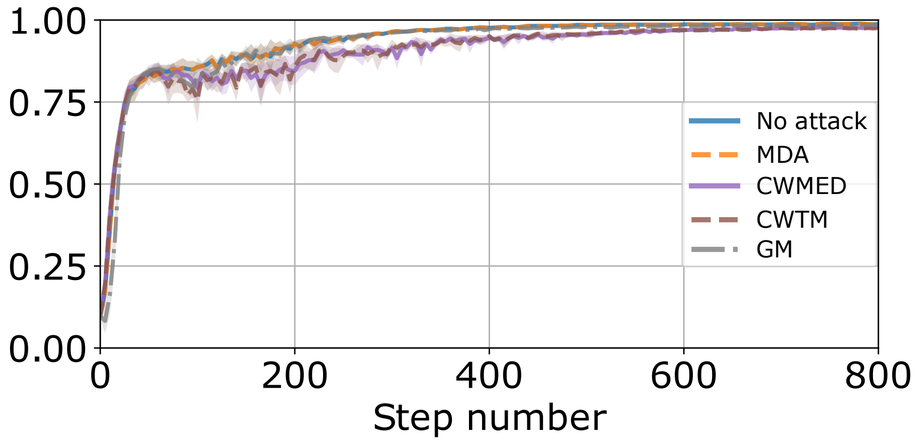
Empirical Benefits of Momentum

Without Momentum

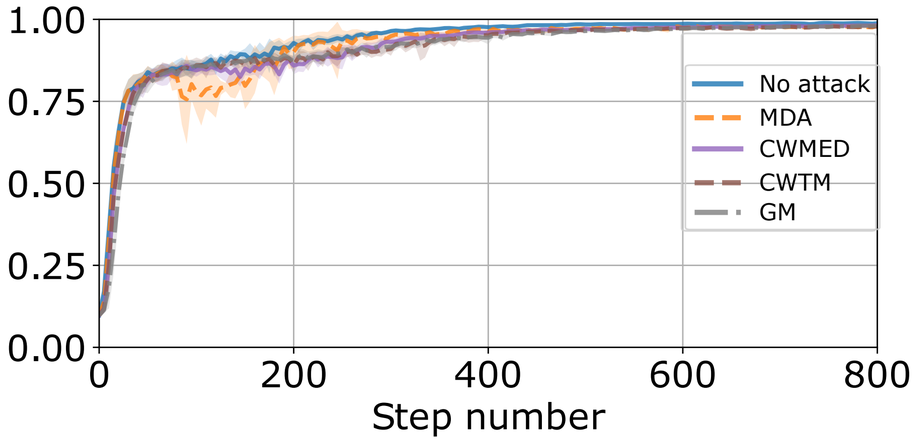




With Momentum $\beta = 0.9$



Label-flipping



Sign-flipping

Training **CNN** on $n = 15$ nodes where $f = 5$ nodes are adversarial.
MNIST dataset is equally divided amongst the honest nodes

Data Poisoning $\not\leftarrow$ Model (gradient) Poisoning

Data poisoning

$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \in \Omega\left(\frac{1+f}{n} \cdot \frac{1}{T}\right)$$

Gradient poisoning

$$\in \mathcal{O}\left(\frac{1+f}{n} \cdot \frac{K}{T}\right)$$

Condition number of $\text{Loss}(\theta)$

Adapting to the learning procedure **does not help** an adversary

On The Relevance of Byzantine Robust Optimization Against Data Poisoning
S. Farhadkhani, R. Guerraoui, N. Gupta, and R. Pinot. JMLR, 2026.

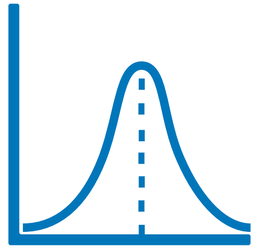
Take Away: Robustness under Local Randomness

Robust Averaging

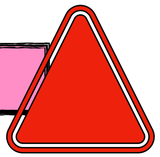
Local Momentum

Local Randomness

Even honest machines compute noisy updates



Adapting to the learning procedure **does not help** an adversary



Byzantine Machine Learning Made Easy by Resilient Averaging of Momentums
S. Farhadkhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. In *International Conference on Machine Learning (ICML),* 2022.

Fixing by Mixing: A Recipe for Optimal Byzantine ML under Heterogeneity
Y. Allouah, S. Farhadkhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Robust Collaborative Learning with Linear Gradient Overhead
S. Farhadkhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. In *ICML*, 2023.

Serverless architecture

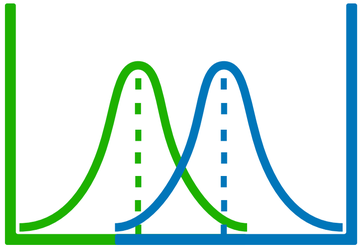
Byzantine Failures Hurt Generalization More Than Data Poisoning
T. Boudou, B. Le Bars, N. Gupta, and A. Bellet. In *International Conference on Machine Learning (ICML),* 2026.

Challenge II: Tackling Heterogeneity

Adversarial nodes can camouflage as honest ones

Data Heterogeneity

Different nodes/clients sample data from different distributions



$S_i \neq S_j$ for any $i, j \in H$ → **Adversarial clients can exploit heterogeneity to inflict more damage**

Impossibility for Robust FL under Heterogeneity

Compute parameters θ^* minimizing $\text{Loss}_H(\theta) := \frac{1}{n-f} \sum_{i \in H} \text{Loss}^i(\theta)$

This goal is generally impossible to achieve, unless the heterogeneity is bounded

Unless $(2f, \varepsilon) - \text{Redundancy}$

Let $\hat{\theta}_S = \arg\min \text{Loss}_S(\theta)$

We guarantee $\hat{\theta}$ s.t. $\|\hat{\theta} - \theta^*\|^2 \leq \varepsilon$ *only if* $\forall\ S' \subseteq S \subseteq [n]$,

with $|S'| = n - 2f$, $|S| = n - f$, we have $\|\hat{\theta}_{S'} - \hat{\theta}_S\|^2 \leq \varepsilon$.

L. Shuo, N. Gupta, and N. H. Vaidya. "Approximate byzantine fault-tolerance in distributed optimization." *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing (PODC),* 2021.

Robust DSGD (with Momentum) under Heterogeneity

Local Phase: Each honest *node* i computes $g_t^{(i)} := \nabla \text{loss}\left(\theta_t\ ,\ z_t^{(i)}\right)$ where $z_t^{(i)} \sim \mathcal{U}\left(S_i\right)$ and updates local Polyak's momentum:

$$m_t^{(i)} = \beta_t\, m_{t-1}^{(i)} \ +\ (1-\beta_t)\, g_t^{(i)}$$

Global Phase: The coordinator updates the model: $\theta_{t+1} = \theta_t - \gamma_t\ \text{RobAvg}\left(m_t^{(1)}\ ,\ \ldots\ ,\ m_t^{(n)}\right)$

(f,κ)-robust averaging

Heterogeneity bound

$$\frac{1}{n-f}\sum_{i\in H}\left\|\nabla \text{Loss}^{(i)}(\theta) - \nabla \text{Loss}(\theta)\right\|^2 \leq \zeta$$

After T iterations:

$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \ \in\ \mathcal{O}\left(\left(\frac{1}{n-f}+\kappa\right)\frac{\sigma^2}{T} \ +\ \kappa\,\zeta\right)$$

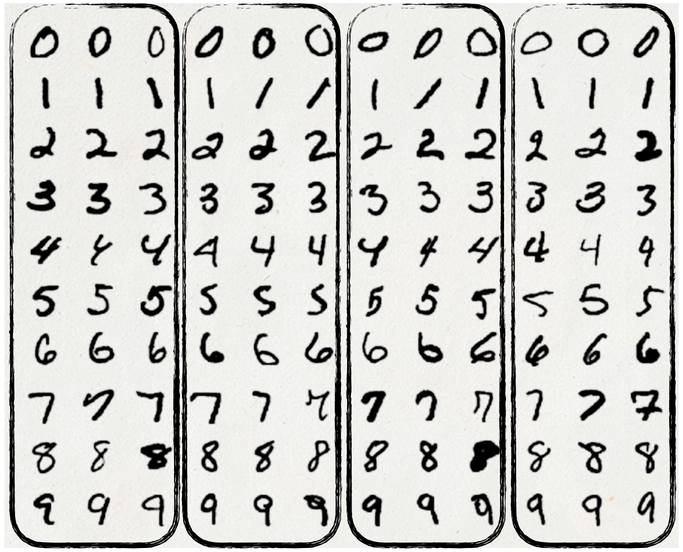
Optimal when $\kappa \in \mathcal{O}\left(f/n\right)$

Unavoidable in general

Different Training Samples $\not\Longleftrightarrow$ Heterogeneity

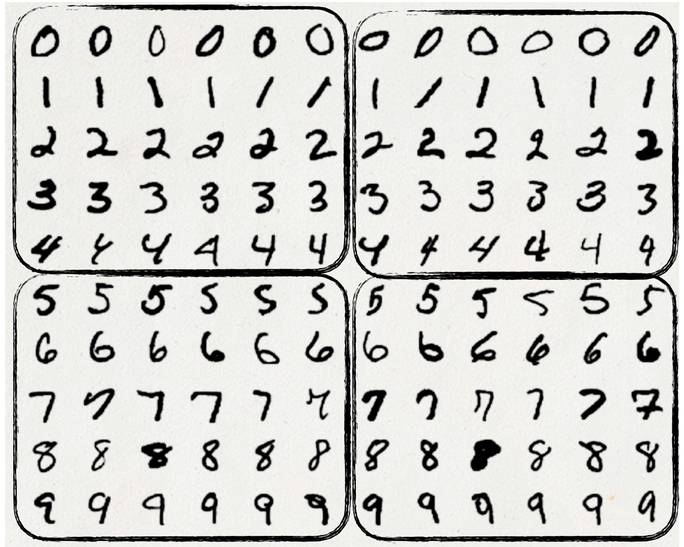
Label-wise heterogeneity

$$\frac{1}{n-f}\sum_{i\in H}\|\nabla \text{Loss}^{(i)}(\theta)-\nabla \text{Loss}(\theta)\|^2 \ll \frac{1}{n-f}\sum_{i\in H}\|\nabla \text{Loss}^{(i)}(\theta)-\nabla \text{Loss}(\theta)\|^2$$



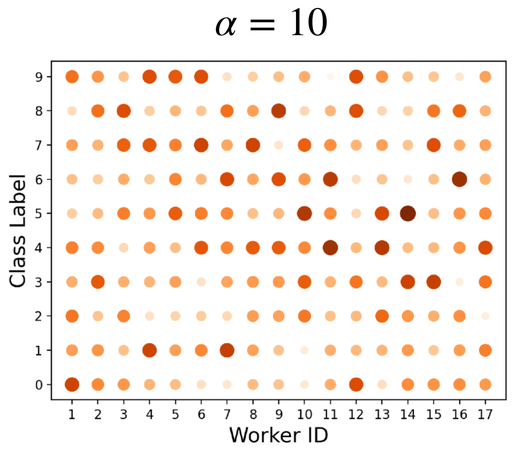
S_1 S_2 S_3 S_4

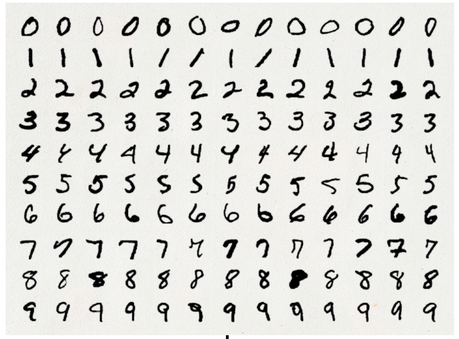
$$\nabla \text{Loss}^{i}(\theta) \approx \nabla \text{Loss}^{j}(\theta)$$

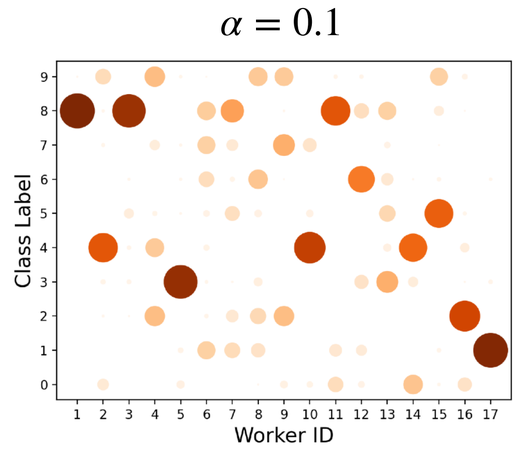


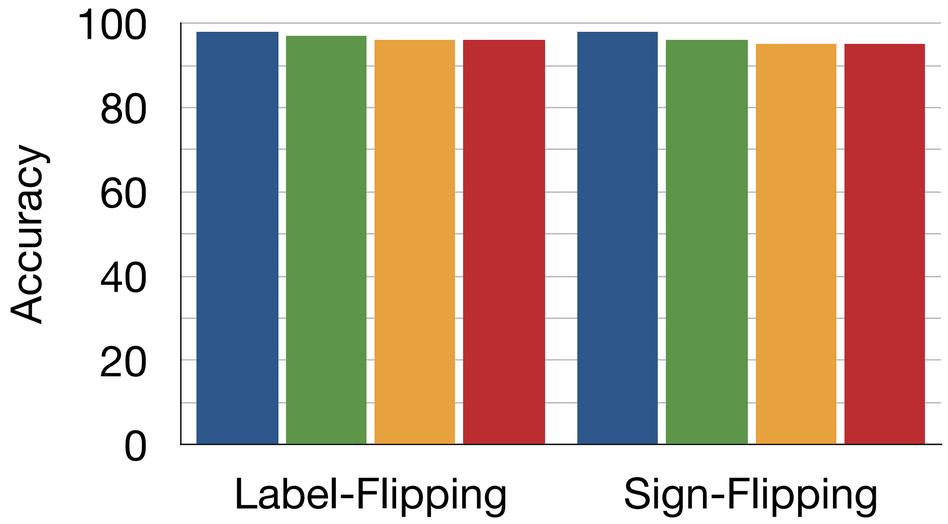
S_1 S_2 S_3 S_4

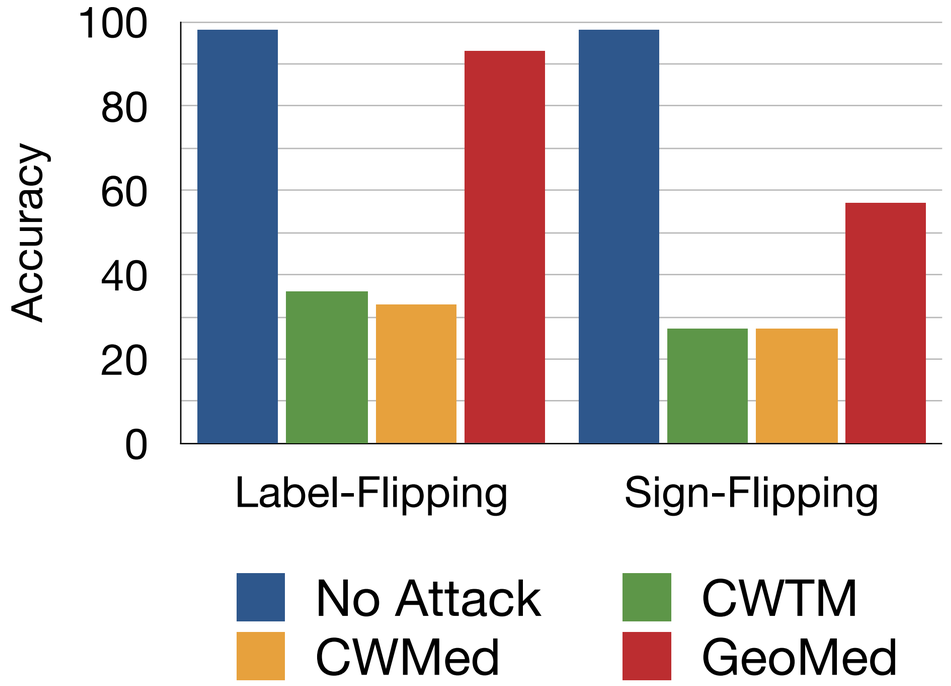
Numerical Observations under Heterogeneity











Training **CNN** on $n = 17$ nodes where $f = 4$ nodes are adversarial. MNIST dataset is divided amongst nodes using Dirichlet distribution with parameters $\alpha = 10$ and $\alpha = 0.1$

Fixing by Mixing: Reducing Effective Heterogeneity

Pre-Aggregation - Nearest Neighbor Mixing (NNM)

$$\text{RobAvg}\left(\text{NNM}\left(m_t^{(1)}, \ldots, m_t^{(n)}\right)\right)$$

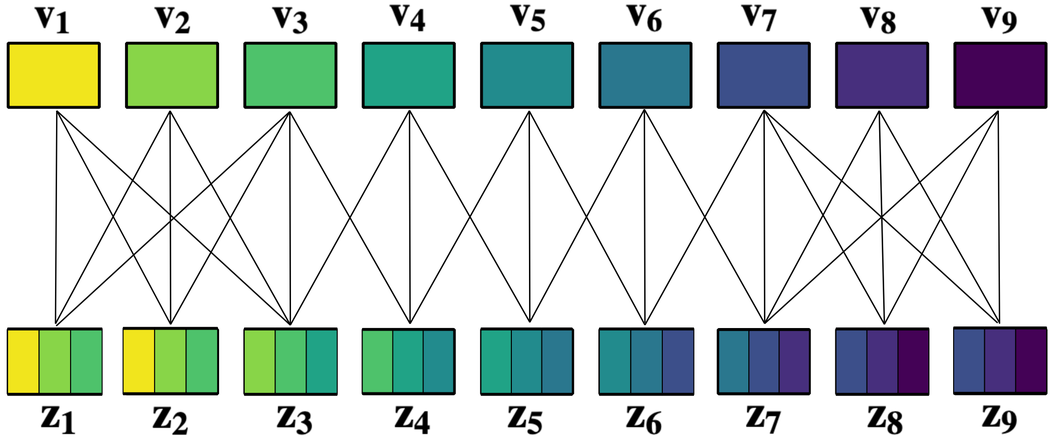
Any median-based aggregation

If F is (f, κ)-robust with $\kappa \in \mathcal{O}(1)$ then

$F \circ \text{NNM}$ is (f, κ)-robust with $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

Best possible robustness coefficient

Yields optimal learning guarantee

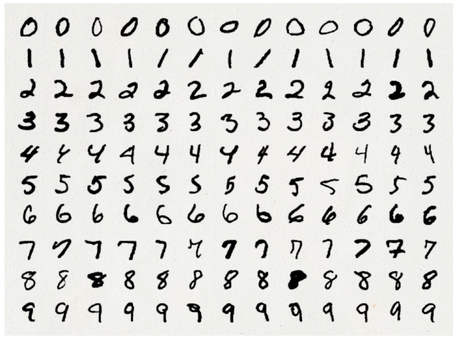


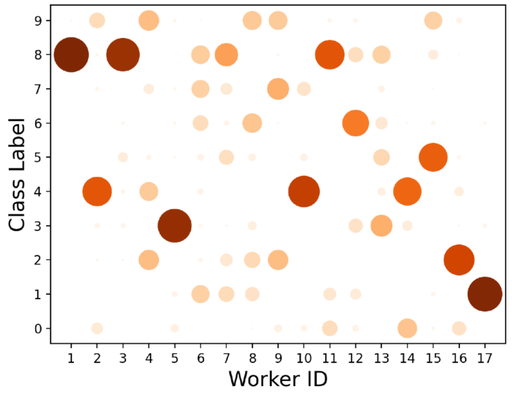
For each v_i choose $n - f$ nearest neighbors in $\{v_1, \ldots, v_n\}$

Let N_i be the set of $n - f$ vectors nearest to v_i

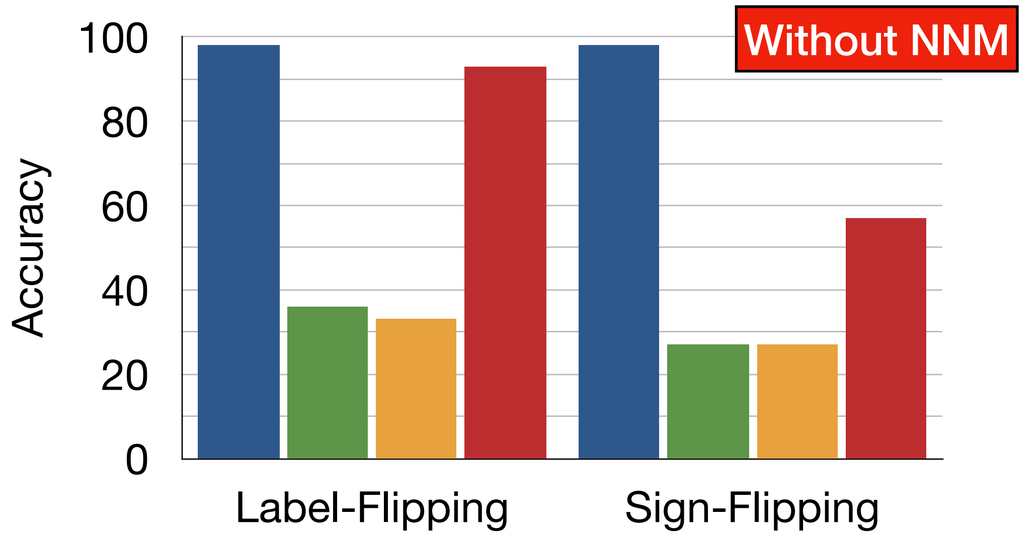
$$\text{Map } v_i \text{ to } z_i := \frac{1}{n-f} \sum_{v \in N_i} v$$

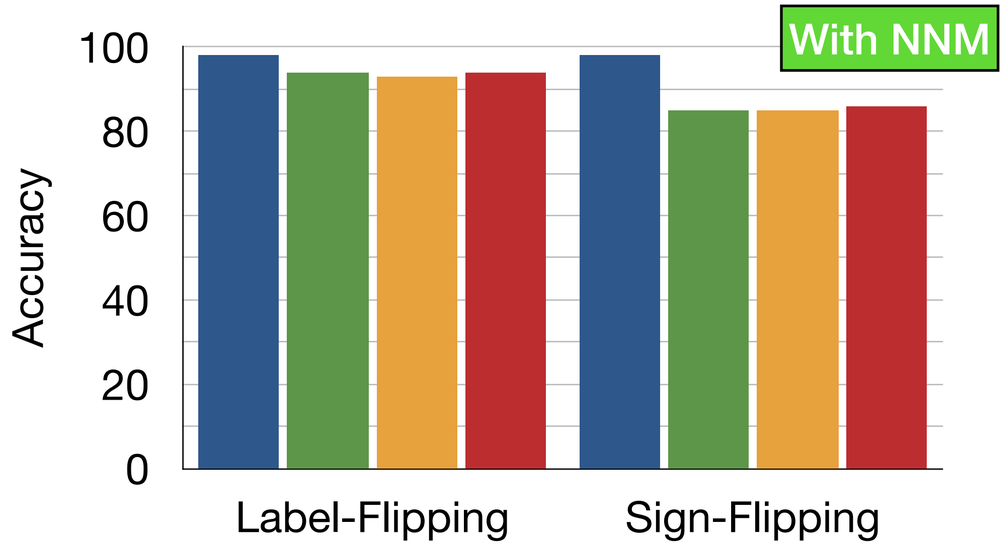
Practical Benefits of Nearest Neighbor Mixing (NNM)





$\alpha = 0.1$





Training **CNN** on $n = 17$ nodes where $f = 4$ nodes are adversarial. MNIST dataset is divided amongst nodes using Dirichlet distribution with parameters $\alpha = 0.1$

(G, B)-Dissimilarity & Adaptive Robust Clipping

Refined heterogeneity model

(G, B)—Gradient Dissimilarity

$$\frac{1}{n-f}\sum_{i\in H}\|\nabla \text{Loss}^{(i)}(\theta) - \nabla \text{Loss}_H(\theta)\|^2 \leq G + B\,\|\nabla \text{Loss}_H(\theta)\|^2$$

$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \;\in\; \Omega\left(\frac{fG}{n-(2+B)f}\right)$$

This lower bound can be remedied under “good” model initialization

ARC

Sort the inputs: $\|v_1\| \leq \ldots \leq \|v_n\|$ and clip them with clipping threshold $C = \|v_k\|$ where $k = \frac{2f}{n}(n-f) + 1$

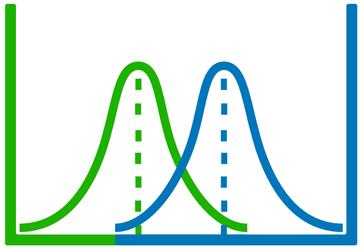
Allouah, Y., Guerraoui, R., Gupta, N., Jellouli, A., Rizk, G., & Stephan, J. Adaptive Gradient Clipping for Robust Federated Learning. In *ICLR, 2025*.

Take Away: Robustness under Heterogeneity

Nearest Neighbor Mixing & Adaptive Robust Clipping

Data Heterogeneity

Different clients sample data from different distributions



Fixing by Mixing: A Recipe for Optimal Byzantine ML under Heterogeneity
Y. Allouah, S. Farhadkhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. *In AISTATS, 2023.*

Robust distributed learning: Tight error bounds and breakdown point under data heterogeneity
Y. Allouah, R. Guerraoui, N. Gupta, R. Pinot, and G. Rizk. *In NeurIPS, 2023.*

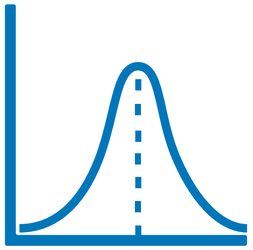
Adaptive Gradient Clipping for Robust Federated Learning
Y. Allouah, R. Guerraoui, N. Gupta, A. Jellouli, G. RIzk, and J. Stephan. *In ICLR, 2025.*

Take Aways: Two Challenges in Robust ML

State machine replication is not efficient

Local Randomness

Even honest machines compute noisy updates

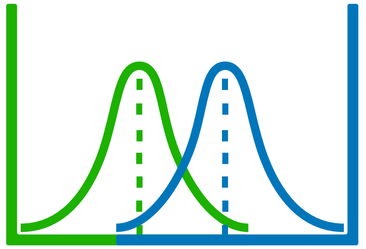


Robust Averaging
Local Momentum

Adversarial nodes can camouflage as honest nodes

Data Heterogeneity

Different nodes sample data from different distributions

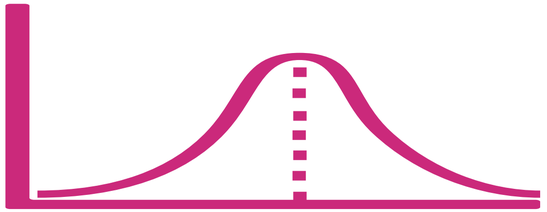


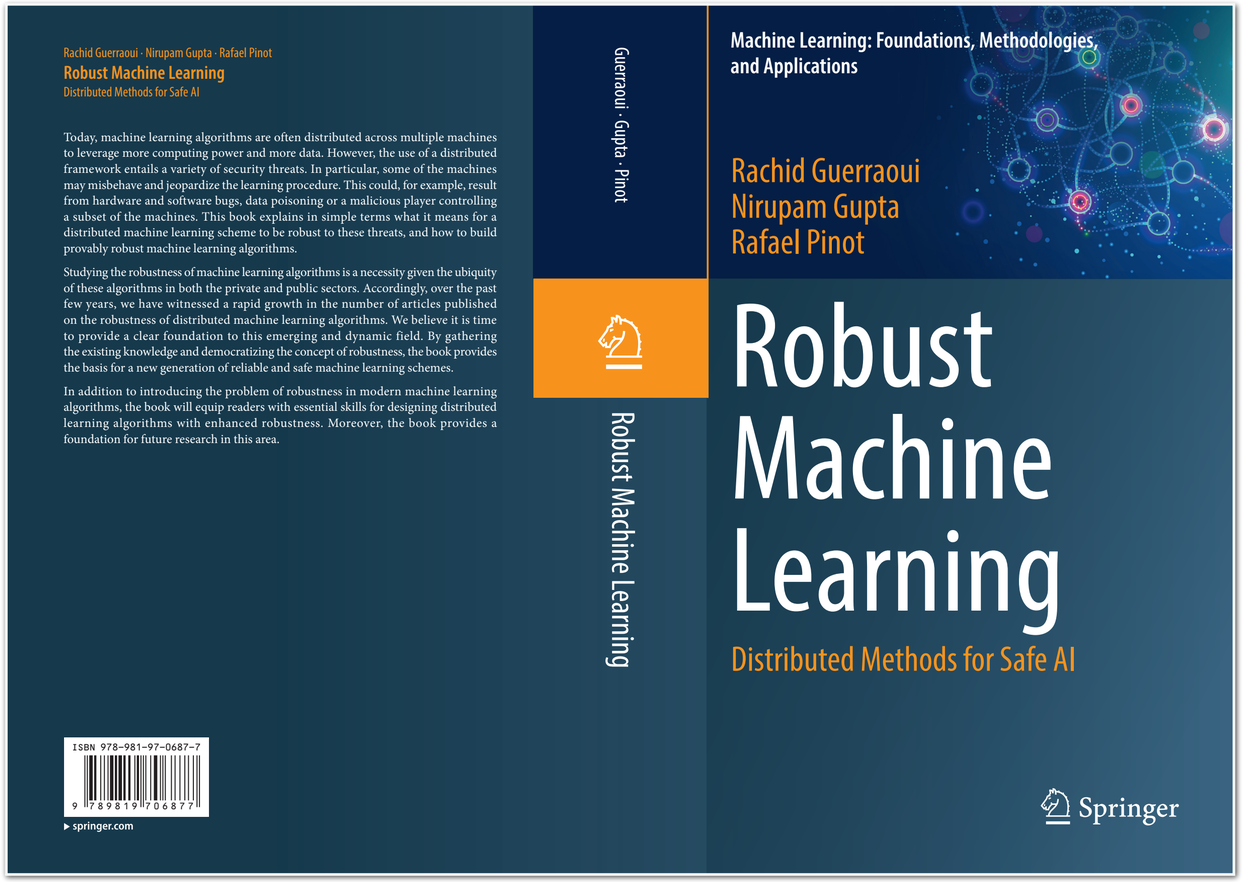
Nearest Neighbor Mixing &
Adaptive Robust Clipping

Augments local randomness isotropically

Data Privacy

Nodes do not share exact information on their data





Reading material:

1. Chapter 4 - Robust averaging
2. Chapter 5 - Nearest neighbor mixing
3. Chapter 6 - Local gradient momentum

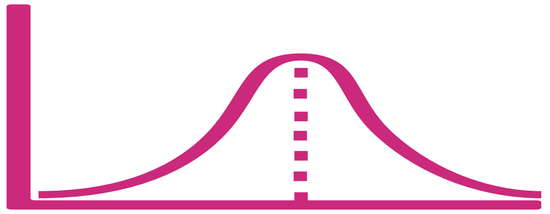
nigu@di.ku.dk

Challenge III: Robustness with Privacy

Augments local randomness isotropically

Data Privacy

Clients do not share exact information on their data

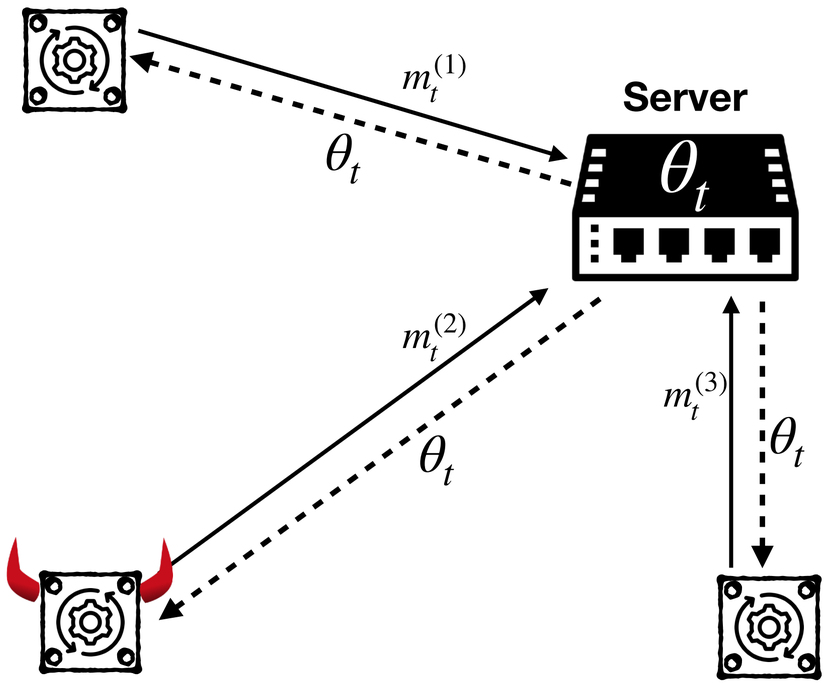


(ϵ, δ)-Distributed DP

Communication between any node i and server is (ϵ, δ)-DP w.r.t. the data held by node i

"Is Interaction Necessary Distributed Private Learning?" A. Smith et al. IEEE S&P 2017.

Robust-DSGD (Similar to DSGD) Leaks Privacy of Data



"Membership Inference Attacks Against Machine Learning Models." Shokri et al. 2016

θ_t → Membership inference

θ_t → Model inversion

"Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures." Fredrikson et al. CCS 2015

$\left(\theta_t , m_t^{(i)}\right)_{i,t}$ → Data reconstruction

"Deep Leakage from Gradients." Zhu et al. NeurIPS 2019

"On the Privacy-Robustness-Utility Trilemma in Distributed Learning". AG**GP**S**.** ICML 2023

Quick Recap of Differentially Private DSGD

Local Phase: Each *honest* node i computes noisy local gradient:

$$g_t^i := \text{Clip}\left(\nabla \text{loss}(\theta_t\ ,\ z)\ ,\ C\right) + \eta_t$$

with $\eta_t \sim \mathcal{N}\left(0,\ \sigma_{\text{DP}}^2\ I_d\right)$, where $\text{Clip}(v,\ C) = \min\left\{1,\ \frac{C}{\|v\|}\right\}\ v$

Gaussian

Clipping

Global Phase: The server averages the gradients $g_t^{(1)}, \ldots, g_t^{(n)}$:

$$\widehat{g}_t := \text{Avg}\left(g_t^{(1)}, \ldots, g_t^{(n)}\right)\ ,$$

updates the current model: $\theta_{t+1} = \theta_t - \gamma_t\ \widehat{g}_t$.

Distributed Differential Privacy (DP)

(ϵ, δ)-Distributed DP

Communication between any node i and server is $(\epsilon,\ \delta)$-DP w.r.t. the data held by node i

"Is Interaction Necessary Distributed Private Learning?" A. Smith et al. IEEE S&P 2017.

Algorithm $\mathscr{A}$ is $(\epsilon,\ \delta)$-DP if for $D \sim D'$

$$\Pr\left(\mathscr{A}(D)\right) \leq e^{\epsilon} \Pr\left(\mathscr{A}(D')\right) + \delta$$

"Our Data, Ourselves: Privacy via Distributed Noise Generation" C. Dwork et al. Eurocrypt 2006.

Privacy by DP-DSGD

Running T iterations of DP-DSGD

By RDP composition and subsampling amplification theorems, we get

"Rényi Differential Privacy." *Mironov, Ilya.* IEEE CSF,2017.

$$\text{If } \sigma_{\text{DP}} \approx C \max \left\{ 1 \ , \ \frac{\sqrt{T \log(1/\delta)}}{m \epsilon} \right\}$$

Grows with T !!

then DP-DSGD ensures (ϵ, δ)-Distributed DP

Learning Error Rate of DP-DSGD

At the completion of T iterations: $\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \in \mathcal{O}\left(\frac{\sigma^2 + d\sigma_{\text{DP}}^2}{nT}\right)$

For providing $(\epsilon,\ \delta)$-distributed DP: $\sigma_{\text{DP}} = kC \max\left\{1\ ,\ \frac{\sqrt{T\log(1/\delta)}}{m\,\epsilon}\right\}$

Some constant

$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \in \mathcal{O}\left(\frac{\sigma^2}{nT} + \frac{d\log(1/\delta)}{nm^2\epsilon^2}\right)$

Privacy-Accuracy trade-off

What About Robust DP-DSGD?

Global Phase: The server robustly aggregates averages the gradients $g_t^{(1)}, \ldots, g_t^{(n)}$:

$$\widehat{g}_t := \mathsf{RobAvg}\left(g_t^{(1)}, \ldots, g_t^{(n)}\right)$$

Without distributed momentum:

Overhead due to robustness grows with T

$$\mathbb{E}\left[\mathrm{Loss}\left(\theta_T\right) - \mathrm{Loss}^*\right] \in \mathcal{O}\left(\frac{d\log(1/\delta)}{nm^2\epsilon^2} + \kappa T\frac{d\log(1/\delta)}{m^2\epsilon^2} + \kappa\zeta\right)$$

Robust Averaging with Distributed Momentum

$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \in \mathcal{O}\left(\frac{d\log(1/\delta)}{m^2\epsilon^2}\left(\frac{1}{n} + \kappa\right) + \kappa\zeta\right)$$

At best $\kappa \in \mathcal{O}\left(f/n\right)$

Lower Bound:

No dimension

$$\text{Loss}\left(\theta_T\right) - \text{Loss}^* \in \Omega\left(\frac{d\log(1/\delta)}{nm^2\epsilon^2} + \frac{f}{n}\cdot\frac{\log(1/\delta)}{m^2\epsilon^2} + \frac{f}{n}\zeta\right)$$

Dimension Can Be Big!

ML Model	Common Applications	Number of Parameters (d)
Resnet-50	Vision, regression	23×10^6
Resnet-18	Vision, regression	18×10^6
GPT-2	NLP, vision	1.5×10^9
GPT-3	NLP, vision	175×10^9

$(f, \kappa)-$Spectral Robust Averaging

Aggregation $F : (v_1, \ldots, v_n) \mapsto \hat{v}$ for all $G \subseteq [n]$ with $|G| = n - f$,

(f, κ)-Robust averaging

$$\frac{1}{n-f} \sum_{i \in G} \|v_i - \bar{v}_G\|^2$$

$$\|\hat{v} - \bar{v}_G\|^2 \leq \kappa \, \lambda_{\max} \left(\frac{1}{n-f} \sum_{i \in G} \left(v_i - \bar{v}_S\right) \left(v_i - \bar{v}_G\right)^\top \right)$$

Spectral norm

No dimension!

$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \in \mathcal{O}\left(\frac{d \log(1/\delta)}{nm^2\epsilon^2} + \kappa \cdot \frac{\log(1/\delta)}{m^2\epsilon^2} + \kappa\zeta \right)$$

Matches LB if

$$\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$$

SMEA: Optimal Spectral Robust Averaging

Smallest Maximum Eigenvalue Averaging

$$G^* \in \underset{G \subseteq [n]\,,\, |G| = n-f}{\arg\min} \lambda_{\max} \left(\frac{1}{n-f} \sum_{i \in G} \left(v_i - \bar{v}_G\right) \left(v_i - \bar{v}_G\right)^\top \right)$$

$$F\left(v_1, \ldots, v_n\right) := \bar{v}_{G^*}$$

SMEA is (f, κ)-Spectral Robust with $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

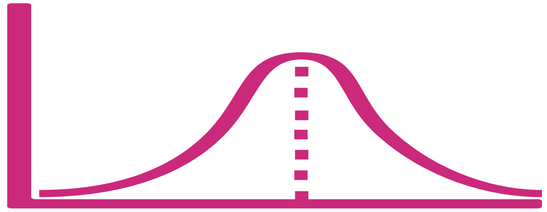
"On the Privacy-Robustness-Utility Trilemma in Distributed Learning". AG**GP**S. ICML 2023

Take Away III: Robustness with Privacy

Augments local randomness isotropically

Data Privacy

Nodes do not share exact information on their data



Spectral Averaging
Local Momentum

“Differential Privacy and Byzantine Resilience in SGD: Do They Add Up?”. G**GP**RS**.** PODC 2022

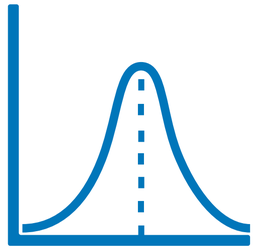
“On the Privacy-Robustness-Utility Trilemma in Distributed Learning”. AG**GP**S**.** ICML 2023

Take Aways: Three Challenges in Robust DL

State machine replication is not efficient

Local Randomness

Even honest machines compute noisy updates

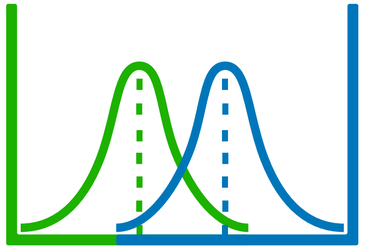


Robust Averaging
Local Momentum

Adversarial nodes can camouflage as honest nodes

Data Heterogeneity

Different nodes sample data from different distributions

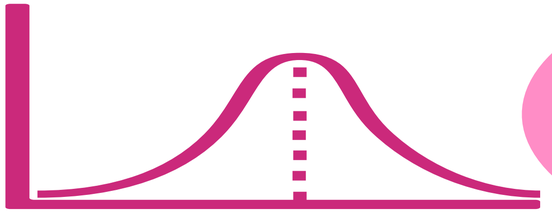


Nearest Neighbor Mixing &
Adaptive Robust Clipping

Augments local randomness isotropically

Data Privacy

Nodes do not share exact information on their data



Spectral Averaging
Local Momentum

Bibliography

Fault-Tolerance in Distributed Optimization: The Case of Redundancy
N. Gupta and N. H. Vaidya. In *Principles of Distributed Computing*. ACM, 2020.

Differential Privacy and Byzantine Resilience in SGD: Do They Add Up?
R. Guerraoui, N. Gupta, R. Pinot, S. Rouault and J. Stephan. In *Principles of Distributed Computing (PODC)*. ACM, 2021.

Presented in the talk

Approximate Byzantine Fault-Tolerance in Distributed Optimization
S. Liu, N. Gupta and N. H. Vaidya. In *Principles of Distributed Computing*. ACM, 2021.

Byzantine Machine Learning Made Easy by Resilient Averaging of Momentums
S. Farhadkhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. In *International Conference on Machine Learning (ICML)*. PMLR, 2022.

Fixing by Mixing: A Recipe for Optimal Byzantine ML under Heterogeneity
Y. Allouah, S. Farhadkhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2023.

Robust Collaborative Learning with Linear Gradient Overhead
S. Farhadkhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

On the Privacy-Robustness-Utility Trilemma in Distributed Learning
Y. Allouah, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

Byzantine Machine Learning: A Primer
R. Guerraoui, N. Gupta and R. Pinot. *ACM Computing Surveys*, 2023.

Robust Distributed Learning: Tight Error Bounds and Breakdown Point under Data Heterogeneity
Y. Allouah, R. Guerraoui, N. Gupta, R. Pinot, and G.Rizk. In *Advances in Neural Information Processing Systems (NeurIPS)* ***(Spotlight)***, 2023.

Robust Machine Learning: Distributed Methods for Safe AI
R. Guerraoui, N. Gupta and R. Pinot. *Springer Nature Publishing Company,* 2023. **[BOOK]**

Open Problems

***Generalizability* of Robust DSGD**

Influence of robustness on accuracy for unseen points?

Boudou, Thomas, Le Bars, Batiste, Gupta, N., & Bellet, Aurélien. (2025).
Generalization under Byzantine & Poisoning Attacks: Tight Stability Bounds in Robust Distributed Learning. *arXiv e-prints*, arXiv-2506.

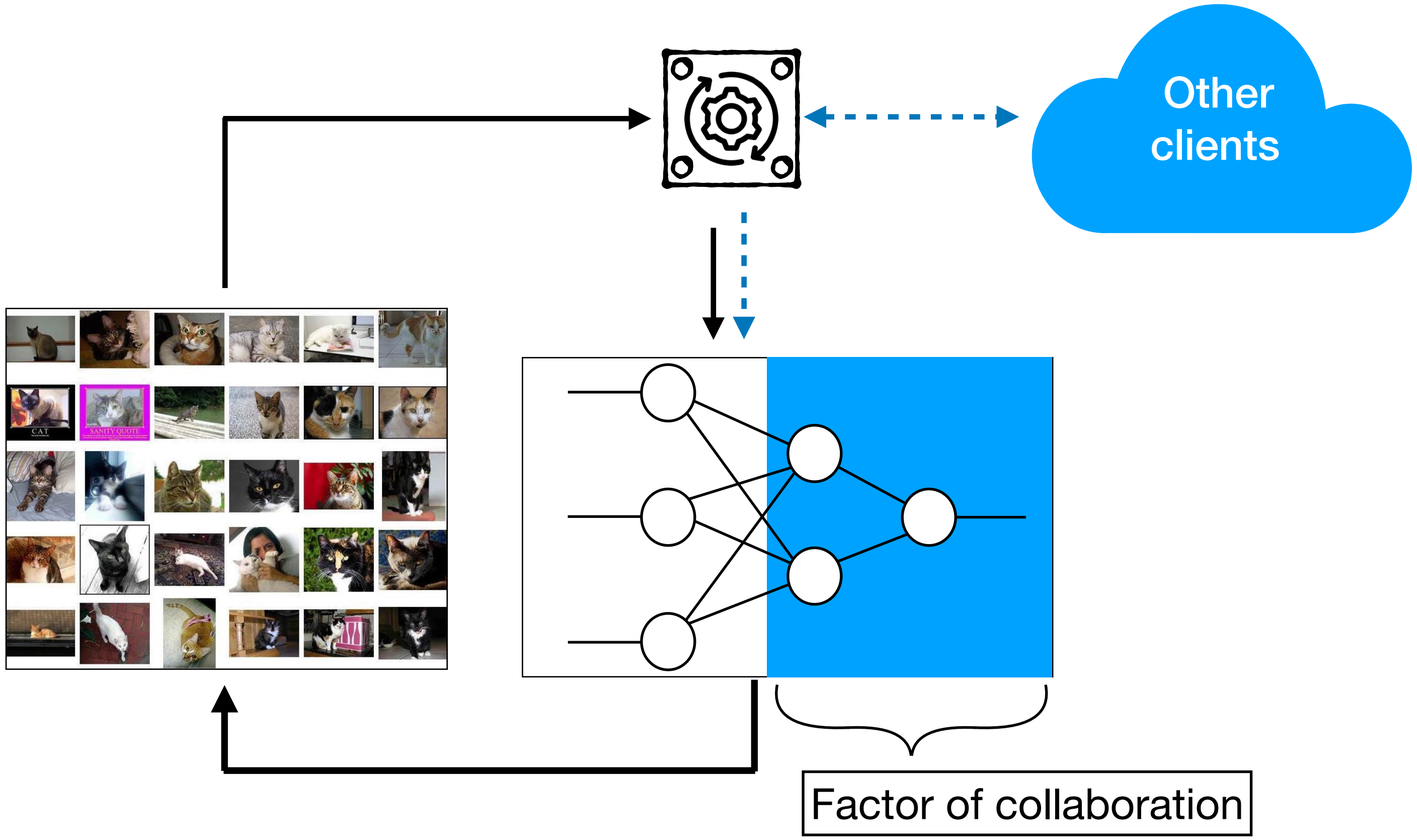
Connection to *adversarial examples*

Can Robust SGD defend against evasion attacks?

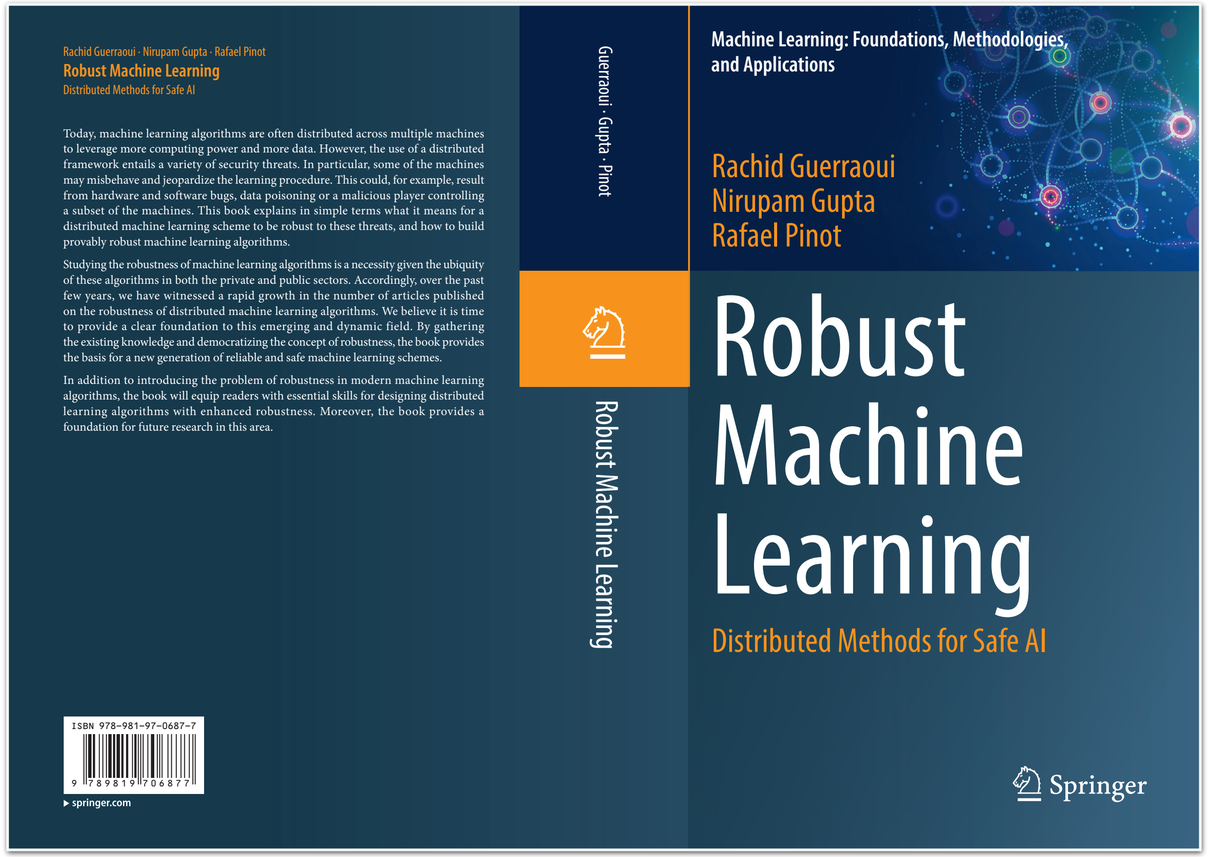
***Game-theoretic* approach**

Tight adversary-aware robustness guarantees

Future Avenue: Robustness with Personalization



Allouah, Youssef, **Abdellah El Mrini**, Rachid Guerraoui, N. Gupta, and Rafael Pinot.
"Fine-tuning personalization in federated learning to mitigate adversarial clients." *NeurIPS,* 2024.



Thank you!

nigu@di.ku.dk

BYZFL