# Robust Federated Learning

## Machine Learning in Untrusted Environments

**Nirupam Gupta**

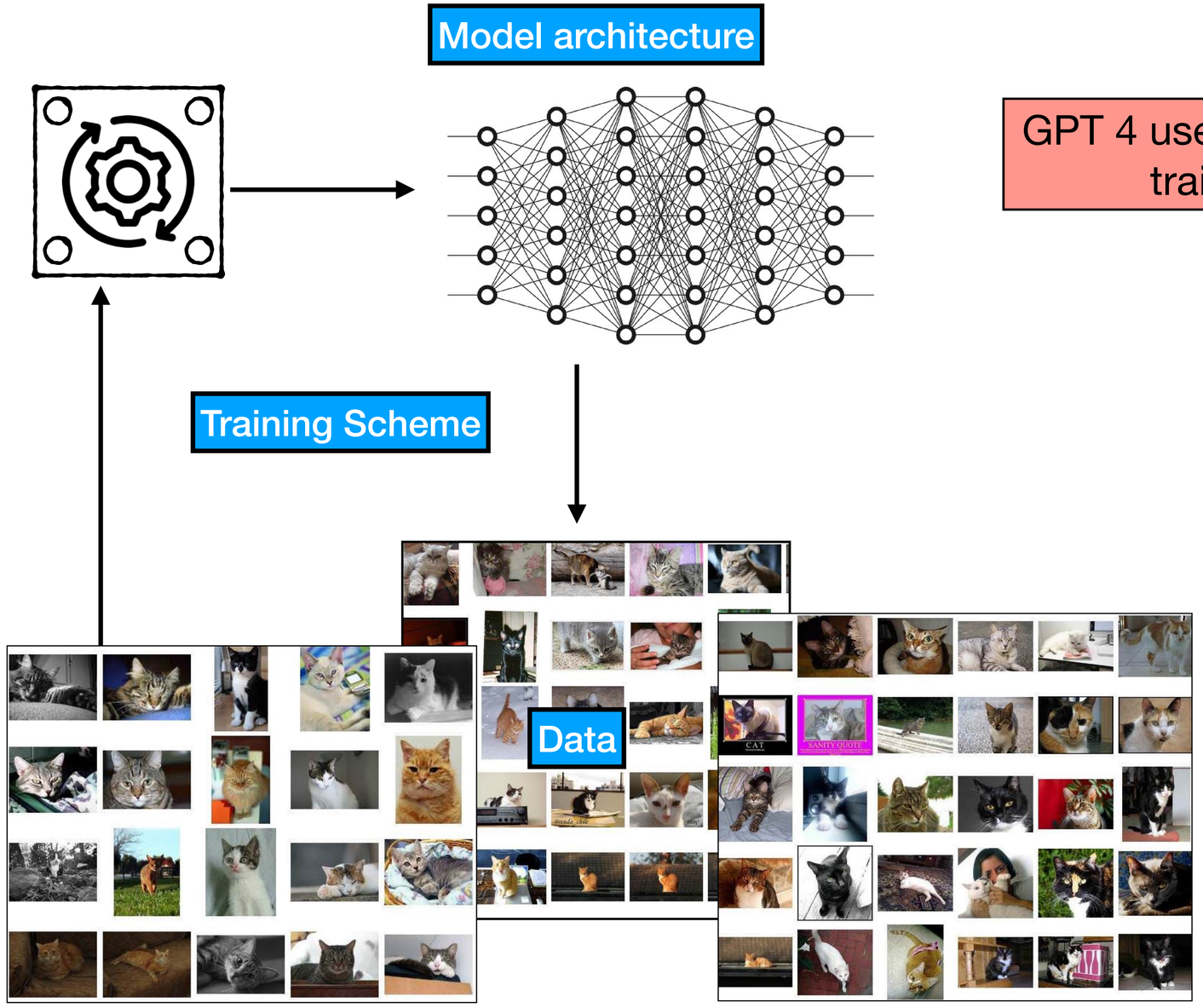
Assistant Professor
Dept. of Computer Science

UNIVERSITY OF COPENHAGEN

MACHINE LEARNING
UNIVERSITY OF COPENHAGEN

Core-AI Spring School 2026, University of Montpellier

# Ever-growing Scale of Machine Learning

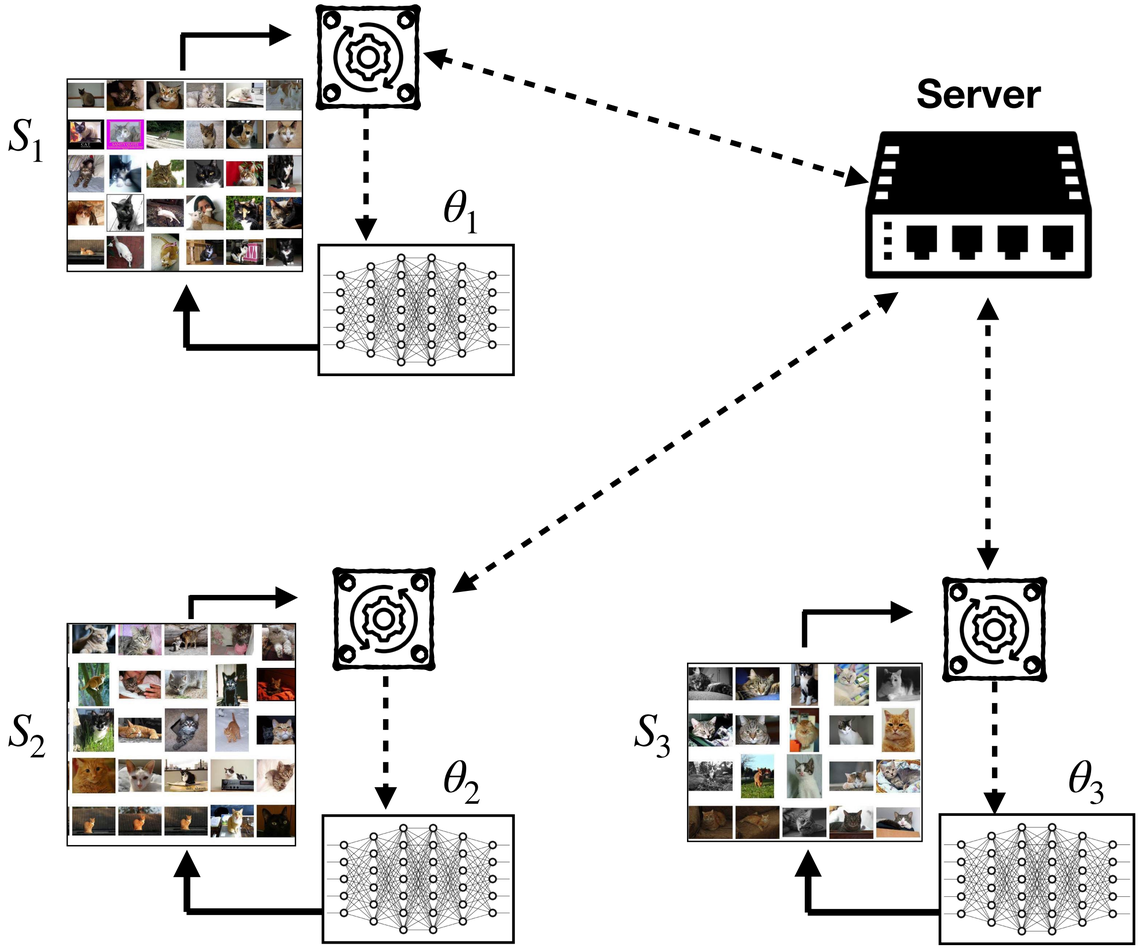



GPT 4 uses NN with **1.7 trillion parameters** trained over **1 PB of text data**

Greater Power

More Data
Larger Model

# Federated Learning: Decentralized Model Training




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

Local loss function

Local samples

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

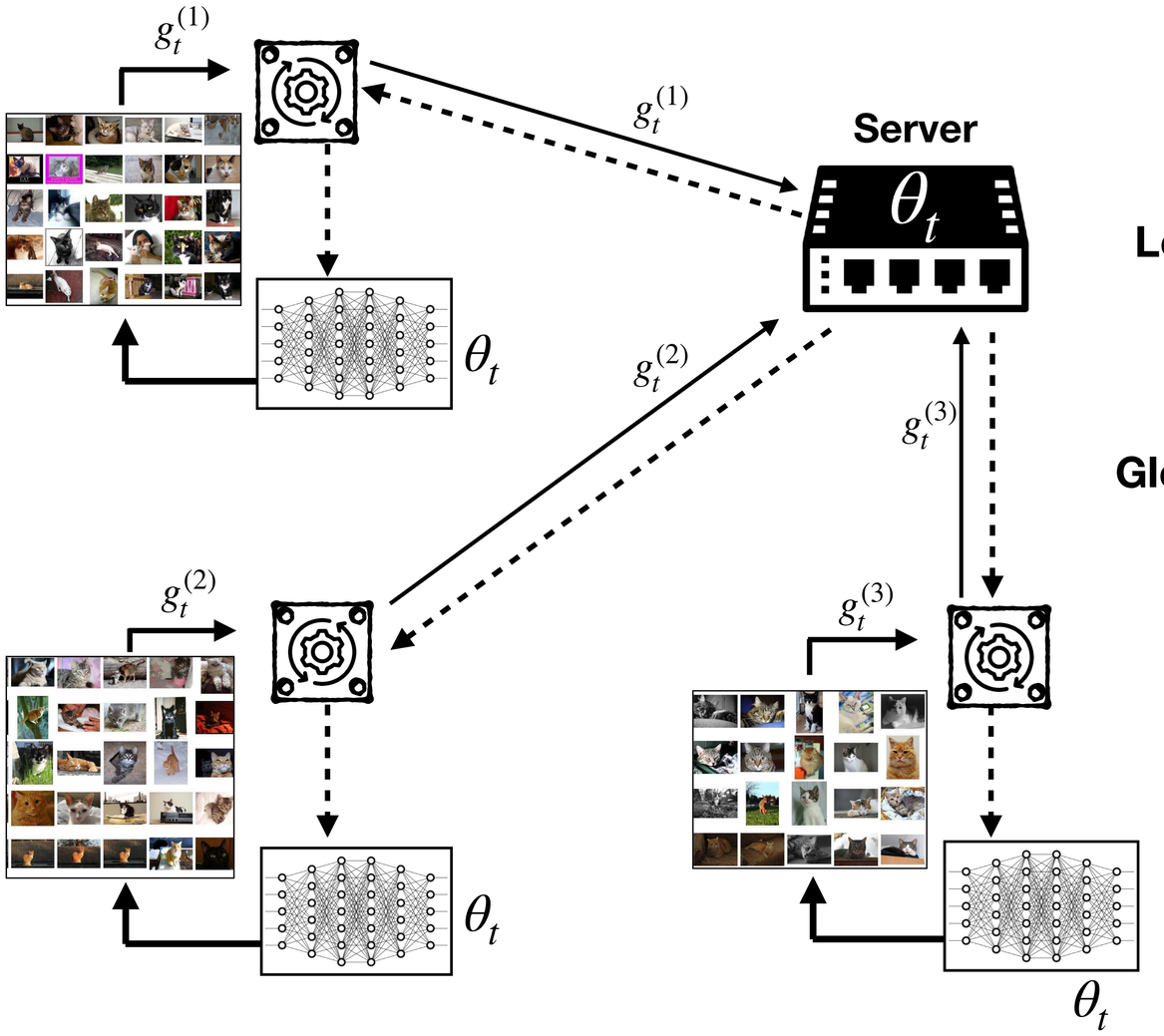
Global loss function

**Benefits**

Scalability
Data ownership control
Data privacy*
Democratization of ML
…

# Distributed Stochastic Gradient Descent (DSGD)

Special Case of Federated Averaging




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

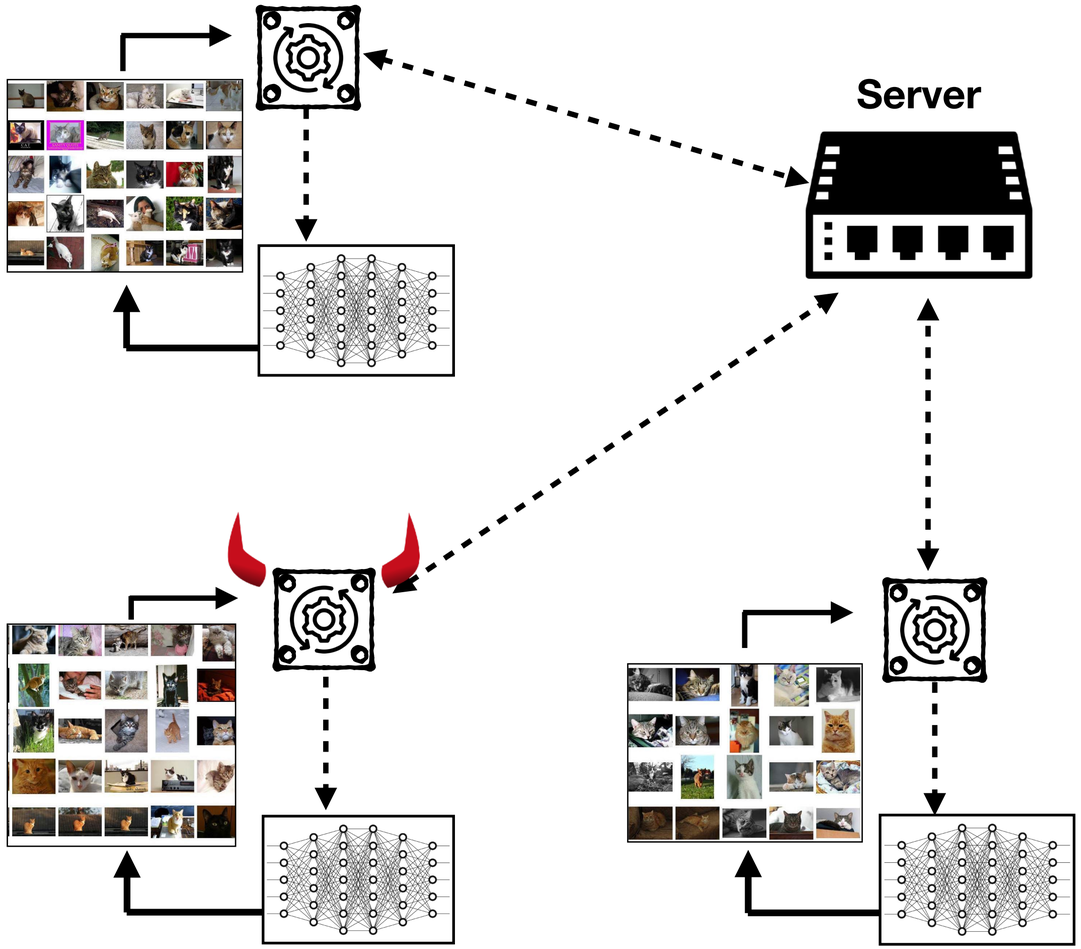
$z_t^{(i)} \sim \mathcal{U}\left(S_i\right)$

**Global Phase:** The server updates the model:

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

$$\theta_t \longrightarrow \arg\min_{\theta \in \Theta} \frac{1}{n} \sum \text{Loss}^{(i)}(\theta)$$

# Vulnerable to Data & Model Poisonings

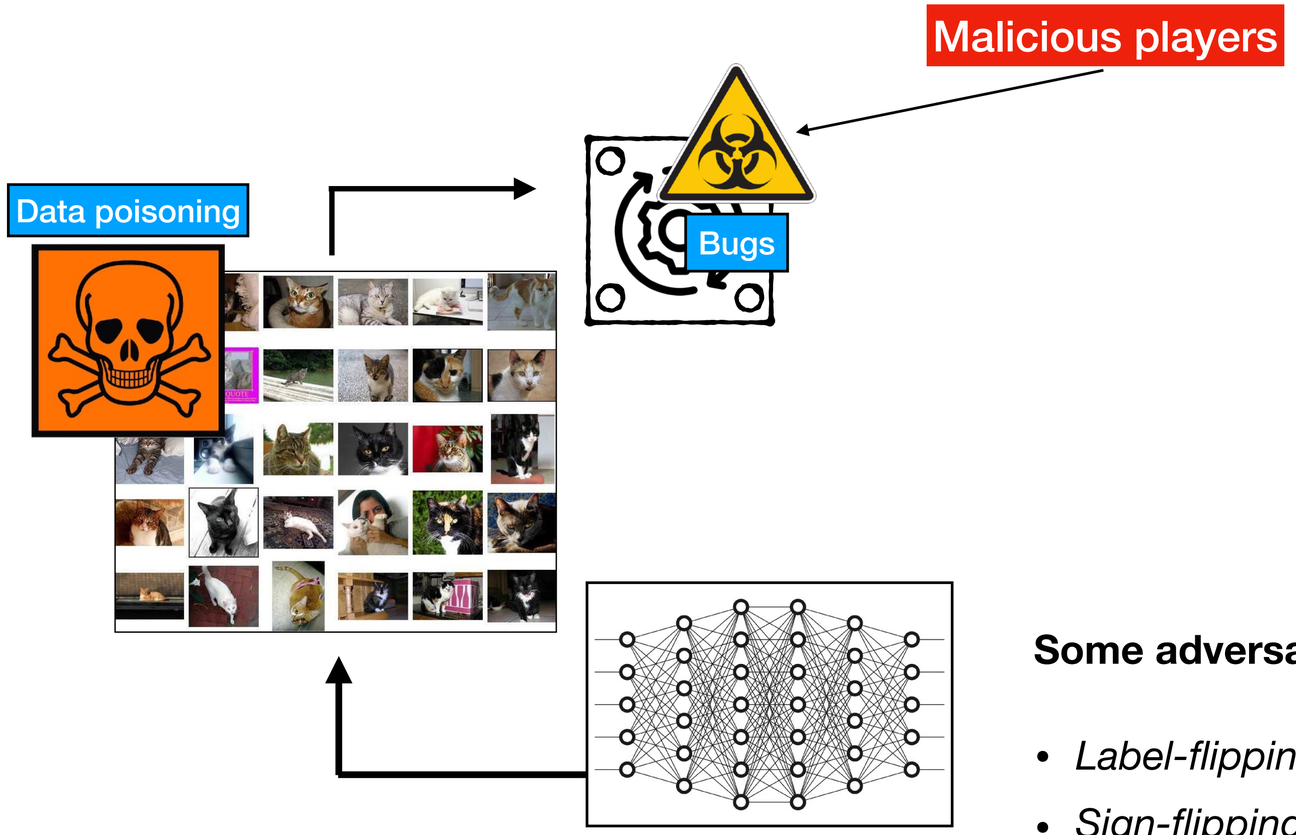


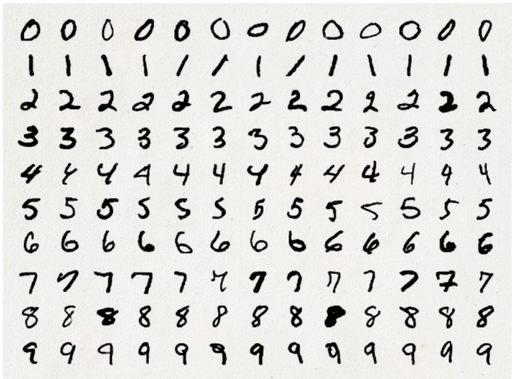

*Not* all data/machines can be trusted

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

Breaks down when some machines “misbehave”

# Reasons Abound for Misbehavior

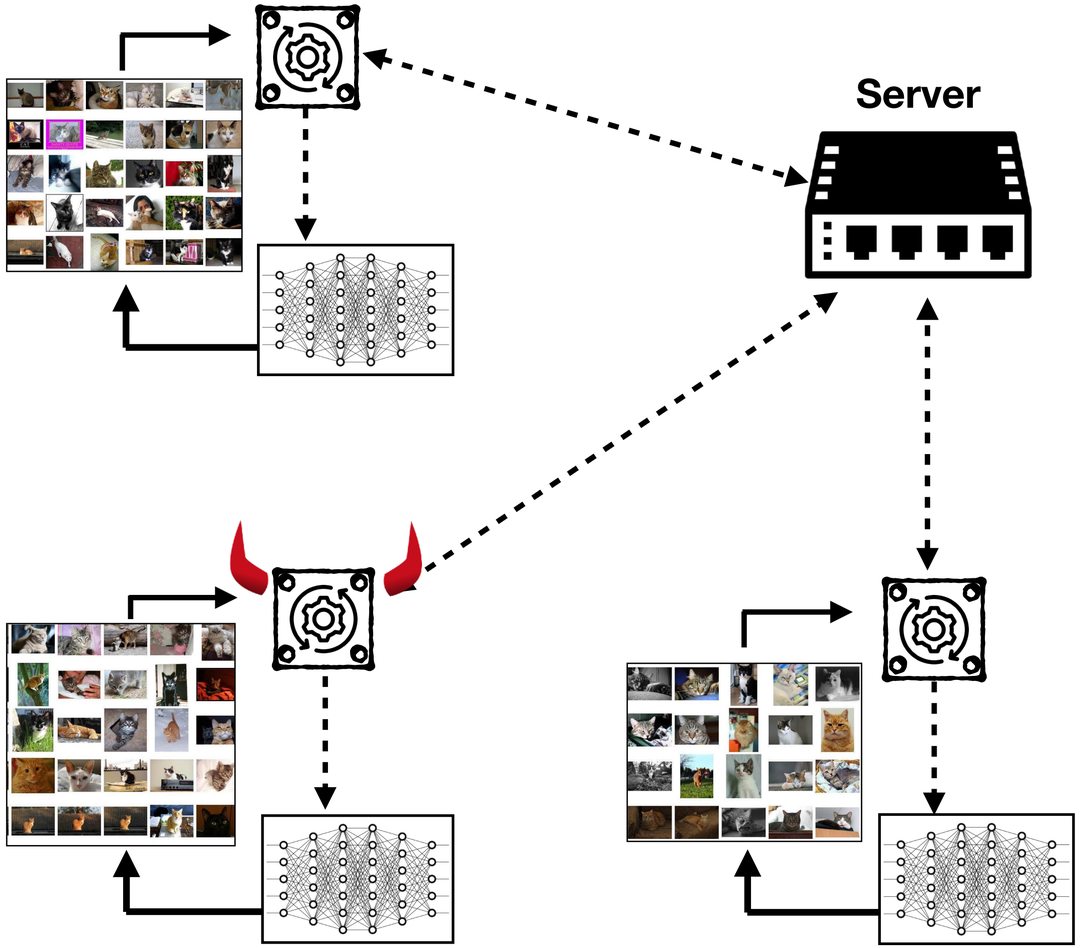


**Some adversarial behaviors:**

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

# Robust Federated/Distributed Learning




$f$ nodes *of unknown identity* are **adversarial**

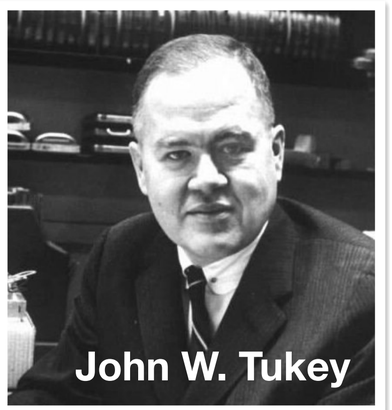
- In minority
- Need not follow the algorithm

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

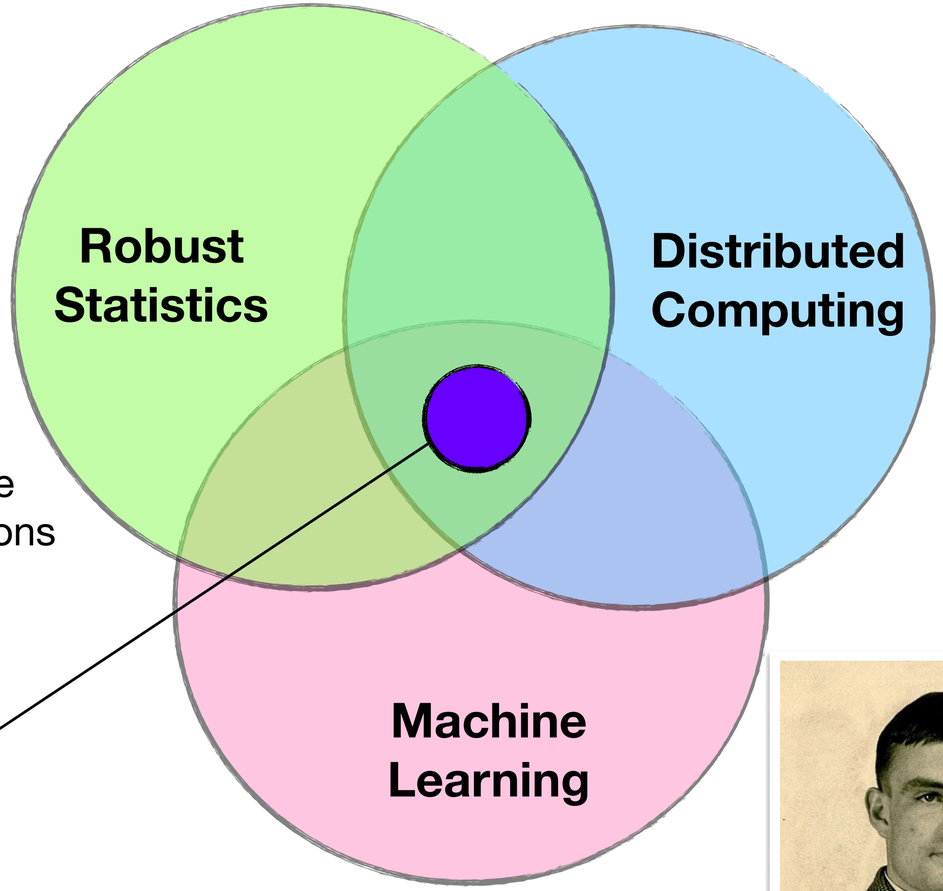
**Robust DL Goal:** $\min_{\theta \in \Theta} \frac{1}{n-f} \sum_{i \in H} \text{Loss}^{(i)}(\theta)$

- Set of *honest* nodes
- Unknown!

# Robust FL: Intersection of Different Worlds



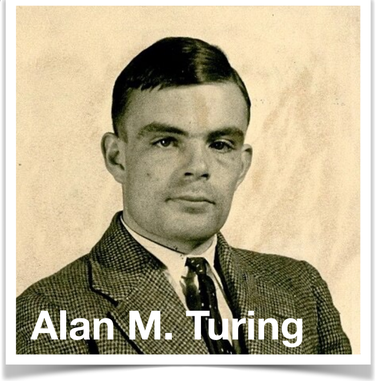

We can estimate true state
despite corrupted observations




Robust Distributed Learning




Algorithms can be distributed
despite untrusted machines



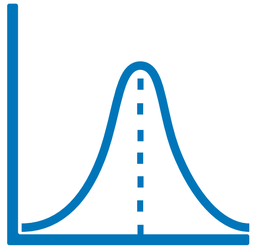

Machines can perform
tasks autonomously

# Three Challenges in Robust FL

State machine replication is not effective

**Local Randomness**

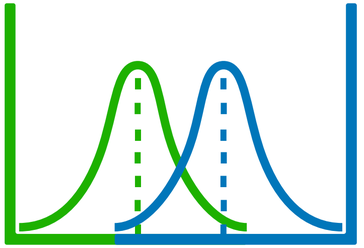
Even honest clients compute noisy updates



Adversarial clients can camouflage as honest ones

**Data Heterogeneity**

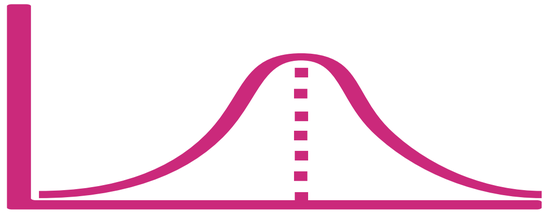
Different clients sample data from different distributions



Local randomness gets amplified

**Data Privacy**

Clients do not share exact information of their data

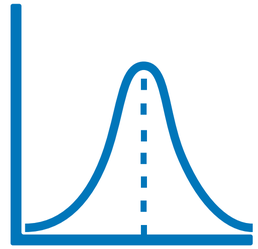


# Challenge I: Tackling Local Randomness

State machine replication is not effective

**Local Randomness**

Even honest clients compute noisy updates



Extreme homogeneity

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

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

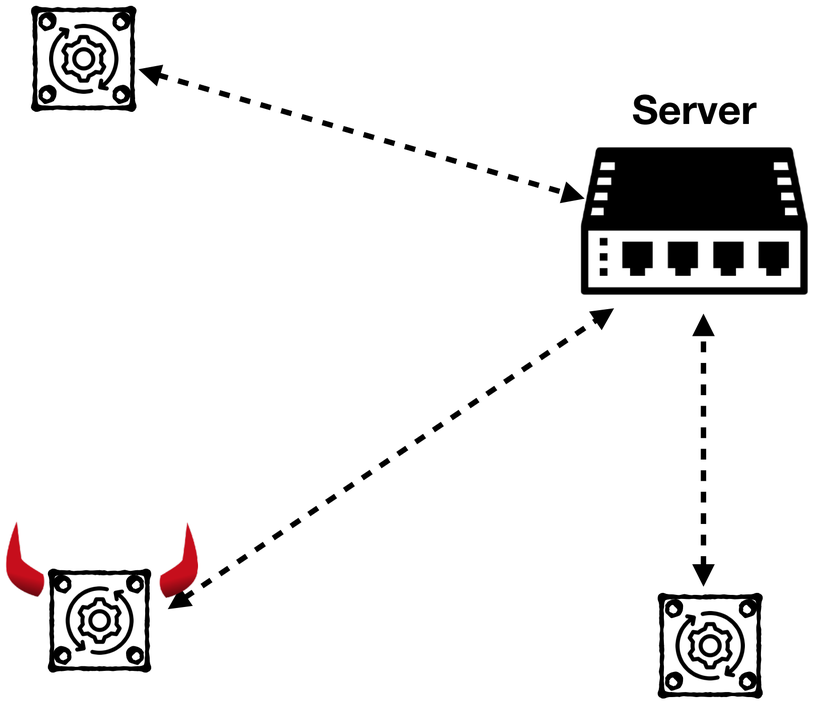
**Can be exploited by adversarial clients**

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

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

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

# Robust FL under Homogeneity




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

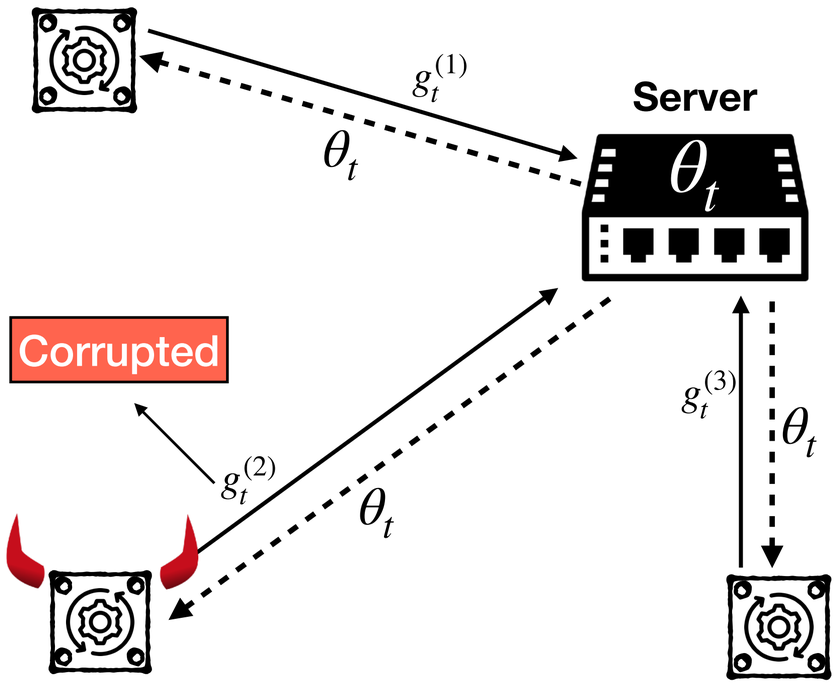
$S_i = S \quad \forall\ i \in H \qquad \text{Loss}^i(\theta) = \text{Loss}(\theta) = \frac{1}{m}\sum_{z \in S} \nabla \text{loss}\,(\theta\ ,\ z)$

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

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 DSGD




**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:

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

Can be arbitrarily manipulated by an adversary

# From Classic DSGD To Robust DSGD

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

"Robust" to some gradients' corruptions

**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)**

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

# Robust Mean Estimation

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

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

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

# $(f, \kappa)$-Robust Averaging: Pathway to Robust Estimation

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

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

Robustness coefficient

Empirical covariance trace

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

$$f < \frac{n}{2}$$

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

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

$$\mathbb{E}\left[\left\|\text{RobAgg}\left(g_t^{(1)}, \ldots, g_t^{(n)}\right) - \nabla \text{Loss}\left(\theta_t\right)\right\|^2\right] \leq 4\kappa\sigma^2$$

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

| Rule | Breakdown | $\kappa$ | When $f \ll n$ |
|---|---|---|---|
| CWMed | $\frac{1}{2}$ | $4\left(1 + \frac{f}{n-2f}\right)^2$ | $\mathcal{O}(1)$ |
| CWTM | $\frac{f+1}{n}$ | $\frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right)$ | $\mathcal{O}\left(\frac{f}{n}\right)$ |

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

# Versatility of $(f, \kappa)$-Robust Averaging

| Aggregation | Breakdown | Robustness Coeff. $\kappa$ | When $f \in \mathcal{O}(1)$ |
|---|---|---|---|
| **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** | $\frac{f+1}{n}$ | $6\left(1+\frac{f}{n-2f}\right)^2$ | $\mathcal{O}(1)$ |

Facilitates robustness analysis that is agnostic to input distribution.

Easy to incorporate with input pre-processing (e.g., kernel operation).

Extends easily to heterogeneous settings.

# Learning Error Rate of Robust DSGD in Homogeneity

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

$(f, \kappa)$-Robust

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

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

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

After $T$ iterations:

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

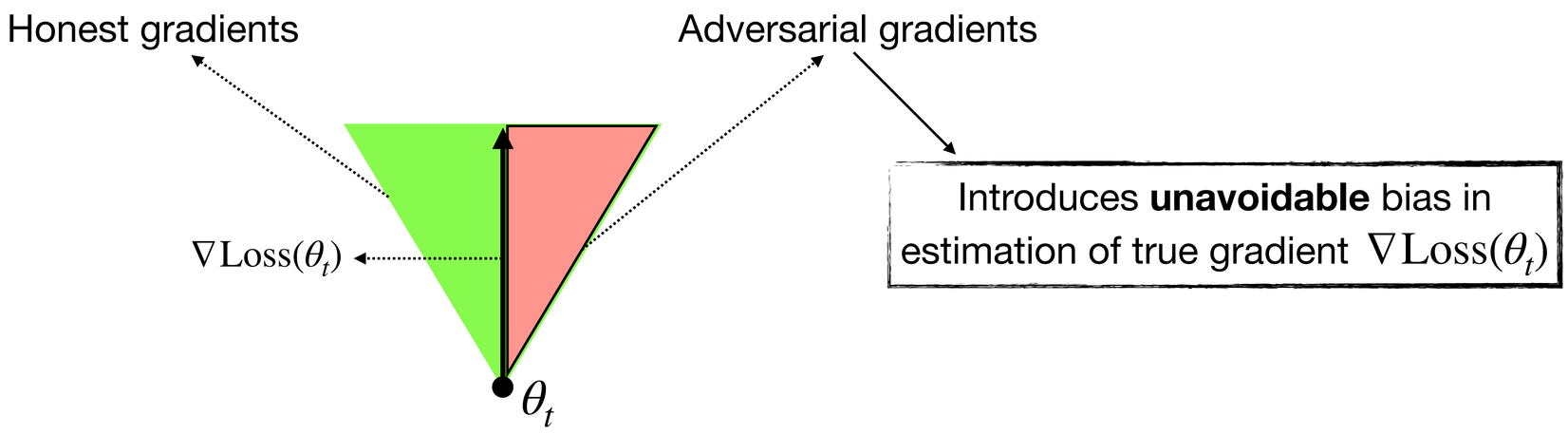
Minimum value

DSGD with $n - f$ honest nodes

Gradient noise

# Can We Do Better?

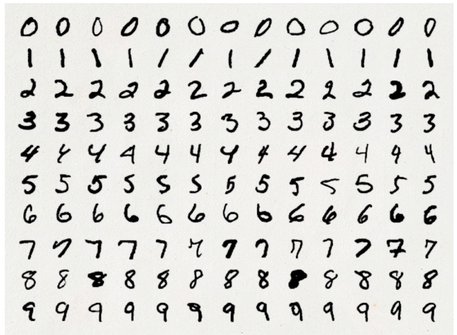
The squared-error in estimating $\nabla \text{Loss}\left(\theta_t\right)$ is in $\Omega\left(\frac{f}{n}\,\sigma^2\right)$ → CWTM is optimal



# Non-Vanishing Dependence on Noise is Problematic

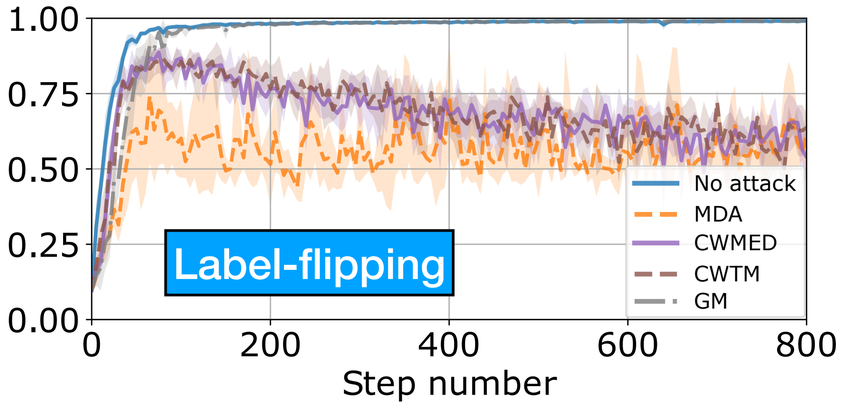
$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \in \mathcal{O}\left(\frac{\sigma^2}{(n-f)T} + \kappa\sigma^2\right)$$

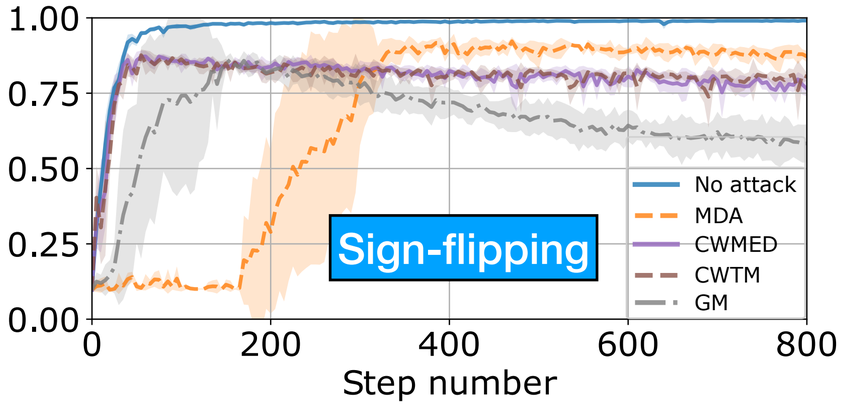
Non-negligible in practice



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**

# Recycle Old Gradients

**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:

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

- $\beta$: Rate of Momentum
- $m_{t-1}^{(i)}$: Past gradients
- $g_t^{(i)}$: Current gradient

**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:** $\beta = 1 - c\,\gamma_t$

After $T$ iterations:

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

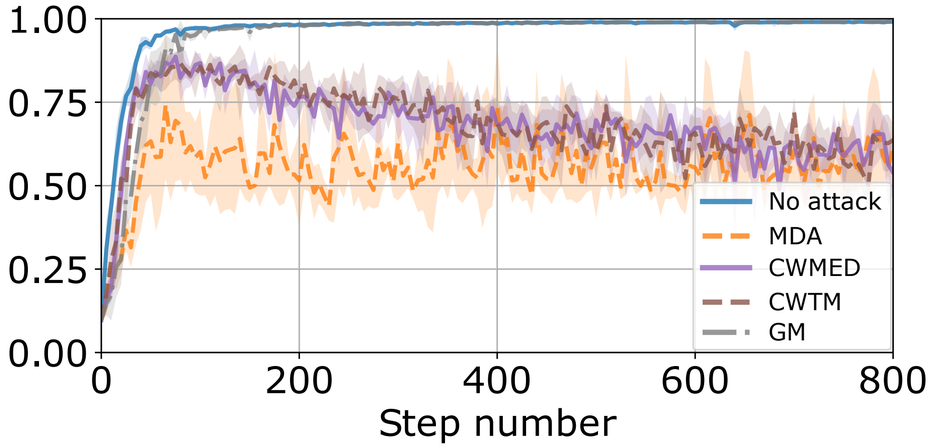
Augments gradient complexity

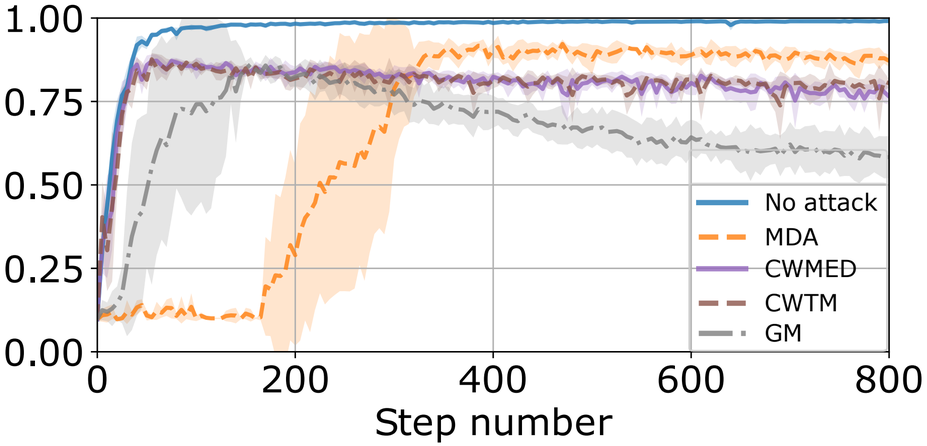
CWTM!

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

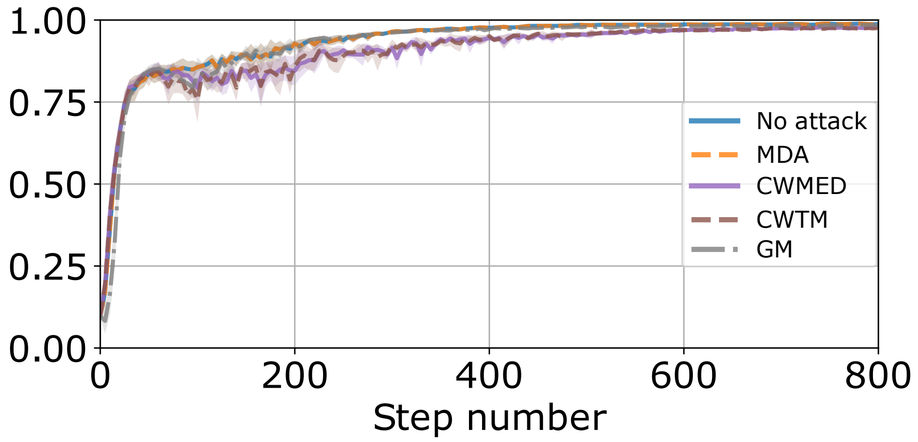
# Empirical Benefits of Momentum

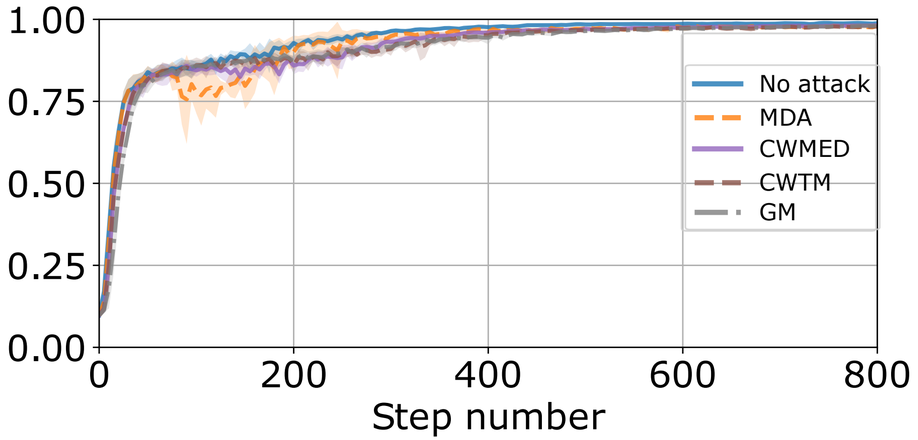
**Without Momentum**




**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

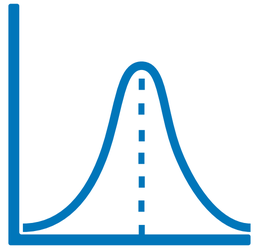
# Take Away: Robustness under Local Randomness

**Robust Averaging**

**Local Momentum**

**Local Randomness**

Even honest machines
compute noisy updates



**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*. 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*. 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*. PMLR, 2023.

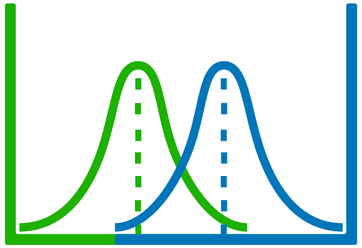
Serverless architecture

# Challenge II: Tackling Heterogeneity

Adversarial clients can camouflage as honest ones

**Data Heterogeneity**

Different 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 $\theta^*$ 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 $\hat{\theta}_S = \arg\min \text{Loss}_S(\theta)$, for any $S \subseteq [n]$

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

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



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

$(f,\kappa)$-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:

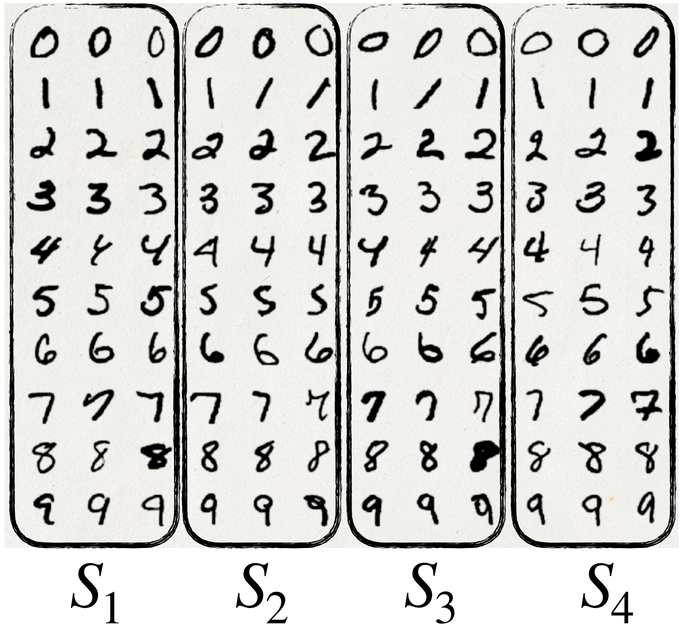
$$\mathbb{E}\left[\text{Loss}\left(\theta_T\right) - \text{Loss}^*\right] \ \in\ \mathcal{O}\left(\left(\frac{1}{n-f}+\kappa\right)\frac{\sigma^2}{T} \ +\ \kappa\zeta\right)$$

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

Unavoidable in general

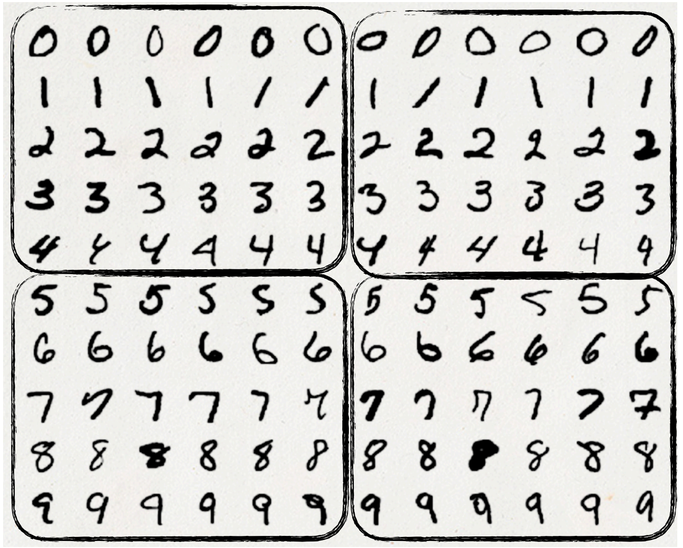
# Different Training Samples $\not\Longleftrightarrow$ Heterogeneity

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



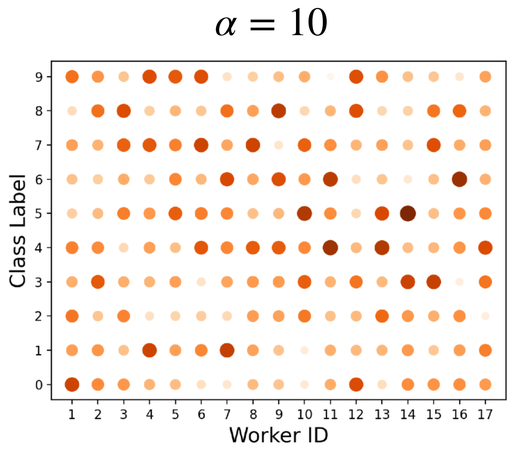
$S_1$ $S_2$ $S_3$ $S_4$

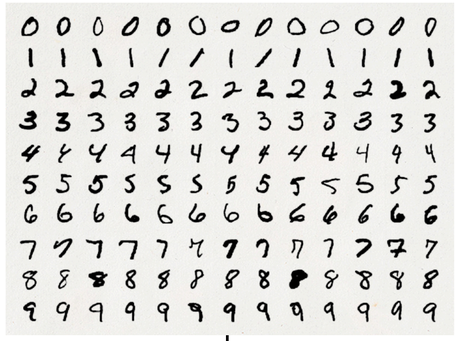
$$\nabla \text{Loss}^i(\theta) \approx \nabla \text{Loss}^j(\theta)$$

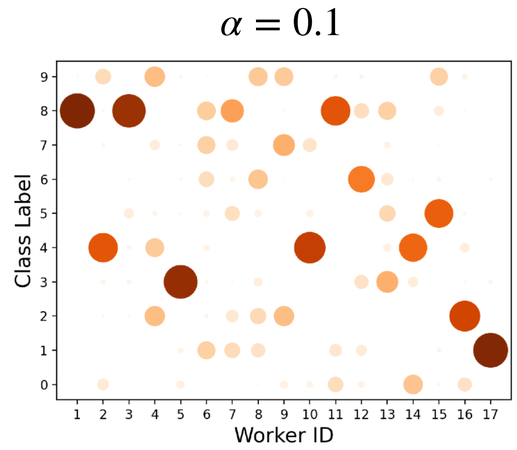


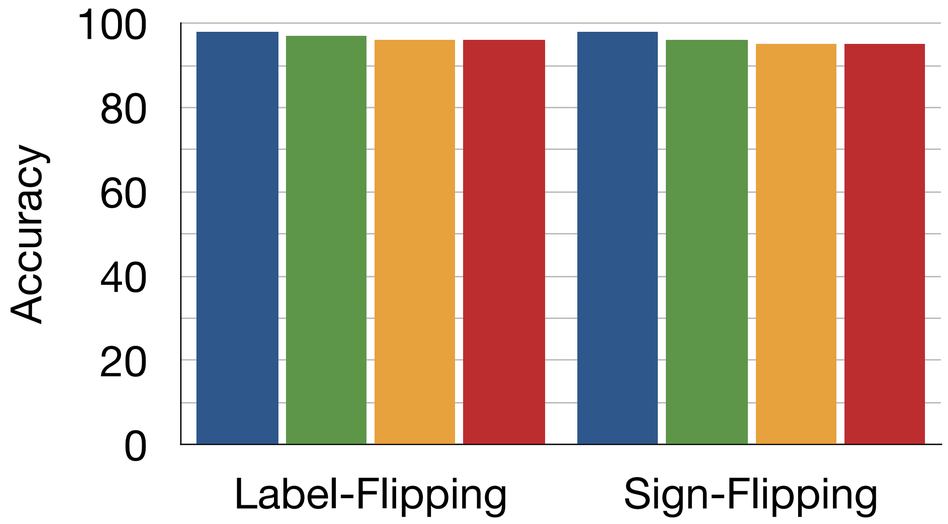
$S_1$ $S_2$ $S_3$ $S_4$

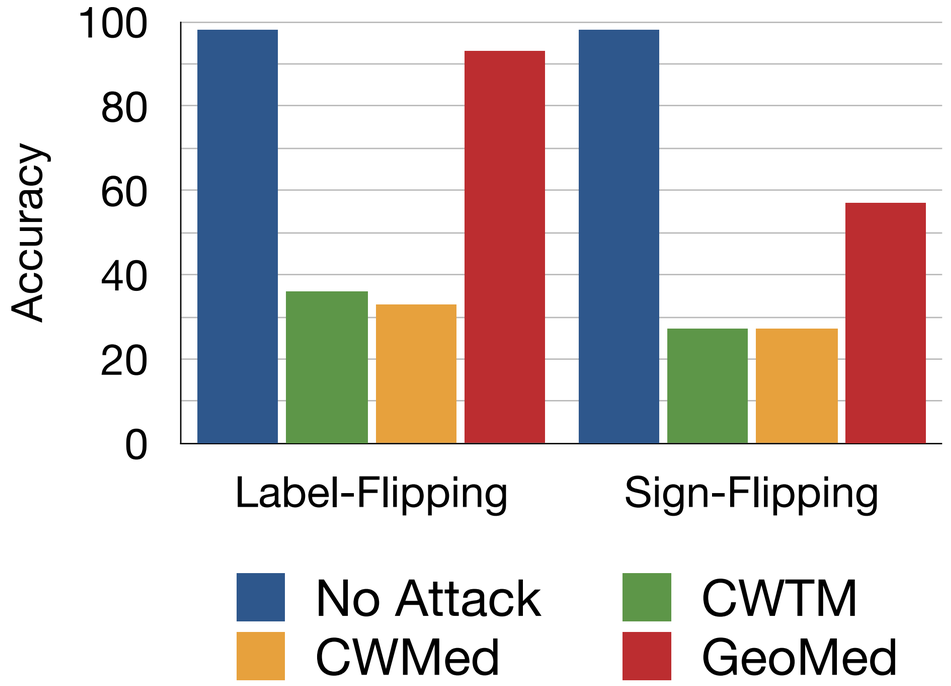
# Numerical Observations under Heterogeneity






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)

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

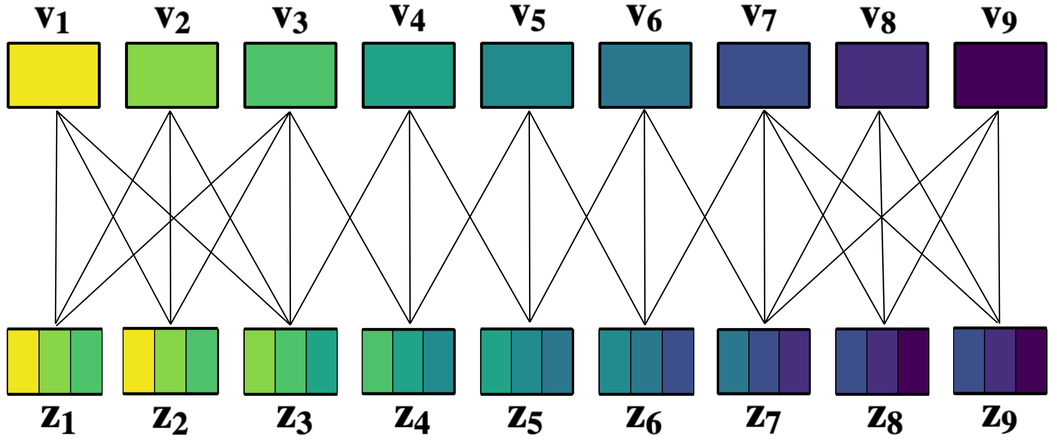
Any median-based aggregation

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

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

Best possible robustness coefficient

Yields optimal learning guarantee

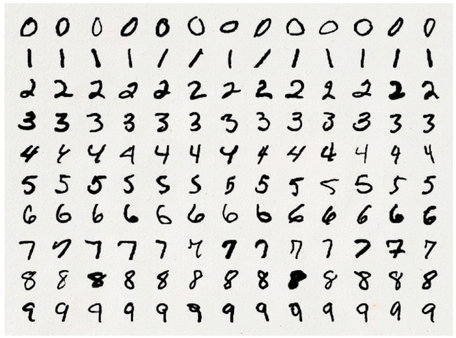


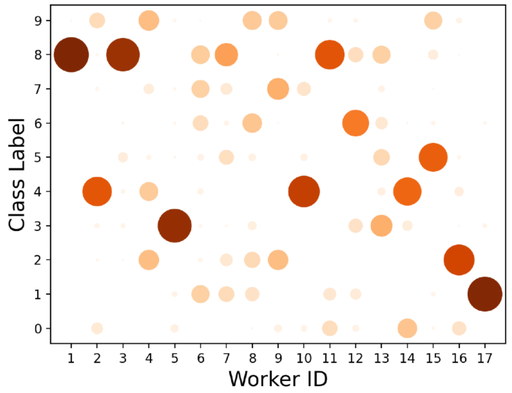

For each $v_i$ choose $n - f$ nearest neighbors in $\{v_1, \ldots, v_n\}$

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

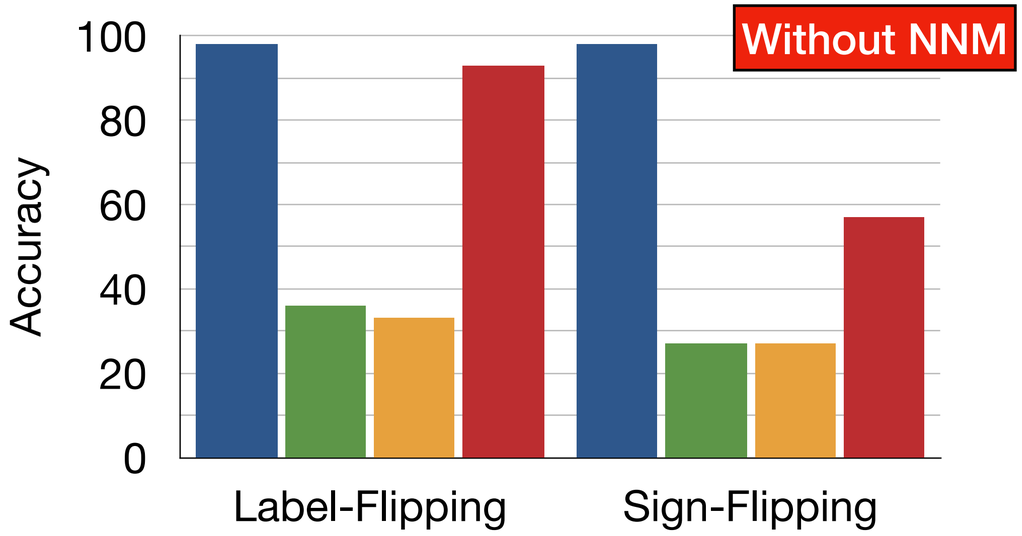
$$\text{Map } v_i \text{ to } z_i := \frac{1}{n-f} \sum_{v \in N_i} v$$

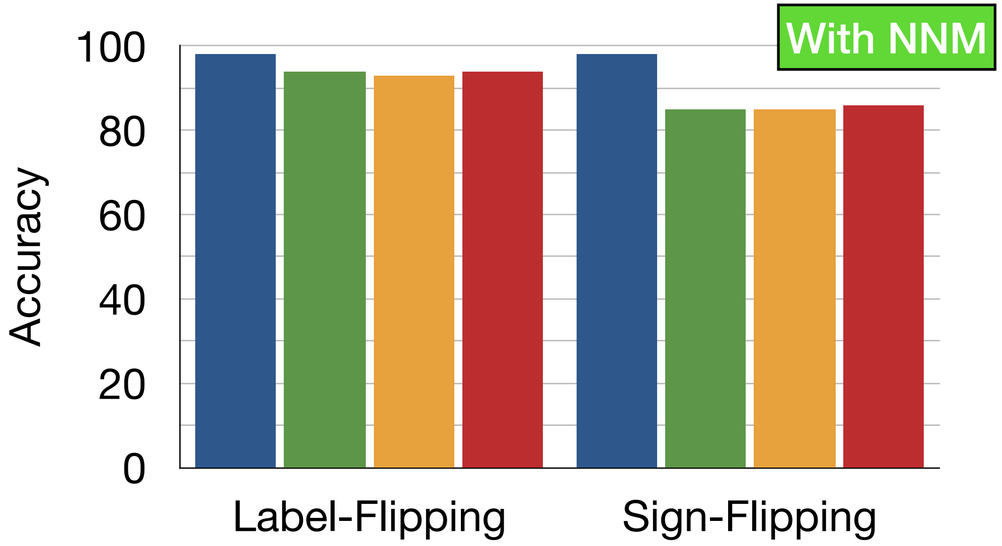
# Practical Benefits of NNM






$\alpha = 0.1$







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 (ARC)

**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 $k = \frac{2f}{n}(n-f) + 1$



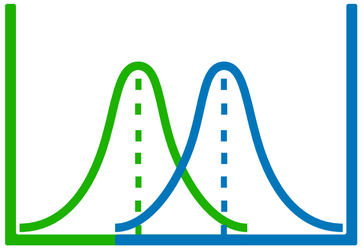
Allouah, Y., Guerraoui, R., Gupta, N., Jellouli, A., Rizk, G., & Stephan, J. Adaptive Gradient Clipping for Robust Federated Learning. In *ICLR, 2025*.

# 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.

Rachid Guerraoui · Nirupam Gupta · Rafael Pinot

**Robust Machine Learning**

Distributed Methods for Safe AI

Today, machine learning algorithms are often distributed across multiple machines to leverage more computing power and more data. However, the use of a distributed framework entails a variety of security threats. In particular, some of the machines may misbehave and jeopardize the learning procedure. This could, for example, result from hardware and software bugs, data poisoning or a malicious player controlling a subset of the machines. This book explains in simple terms what it means for a distributed machine learning scheme to be robust to these threats, and how to build provably robust machine learning algorithms.

Studying the robustness of machine learning algorithms is a necessity given the ubiquity of these algorithms in both the private and public sectors. Accordingly, over the past few years, we have witnessed a rapid growth in the number of articles published on the robustness of distributed machine learning algorithms. We believe it is time to provide a clear foundation to this emerging and dynamic field. By gathering the existing knowledge and democratizing the concept of robustness, the book provides the basis for a new generation of reliable and safe machine learning schemes.

In addition to introducing the problem of robustness in modern machine learning algorithms, the book will equip readers with essential skills for designing distributed learning algorithms with enhanced robustness. Moreover, the book provides a foundation for future research in this area.

ISBN 978-981-97-0687-7



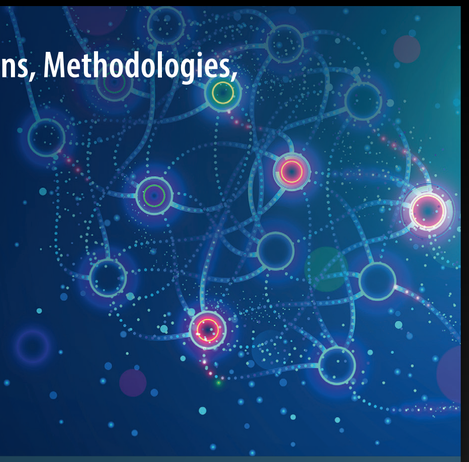
▶ springer.com

Guerraoui · Gupta · Pinot



Robust Machine Learning

Machine Learning: Foundations, Methodologies, and Applications



Rachid Guerraoui
Nirupam Gupta
Rafael Pinot

# Robust Machine Learning

## Distributed Methods for Safe AI



Springer

# Python Library for Robust ML

BYZFL

ByzFL is a Python library for Robust Federated Learning.

**https://byzfl.epfl.ch/**

# Reading List

**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*. 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*. 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*. 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*. 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*. 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* ***(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 distribution of hypothesis?

**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.

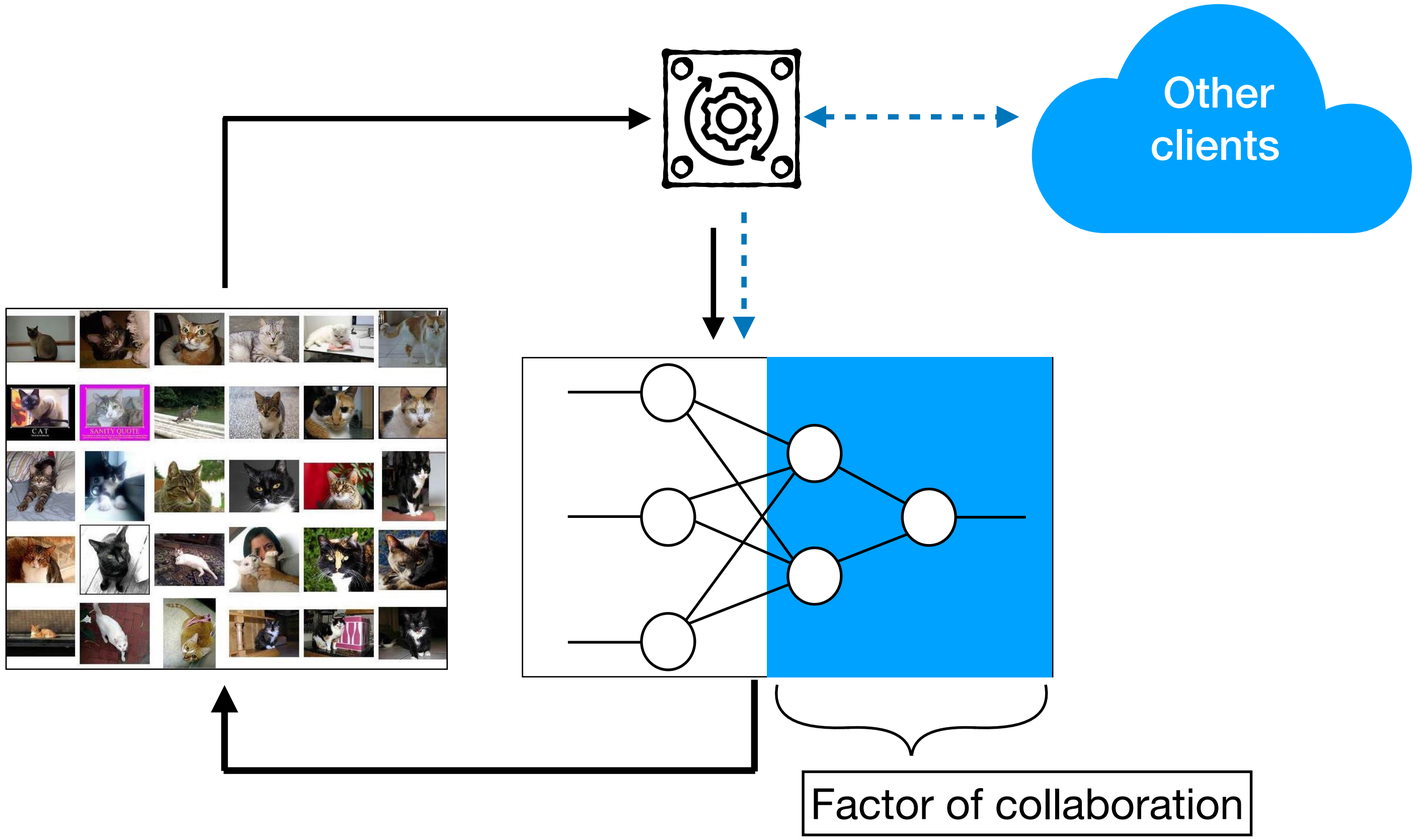
## **Connection to *adversarial examples***

Can Robust SGD defend against evasion attacks?

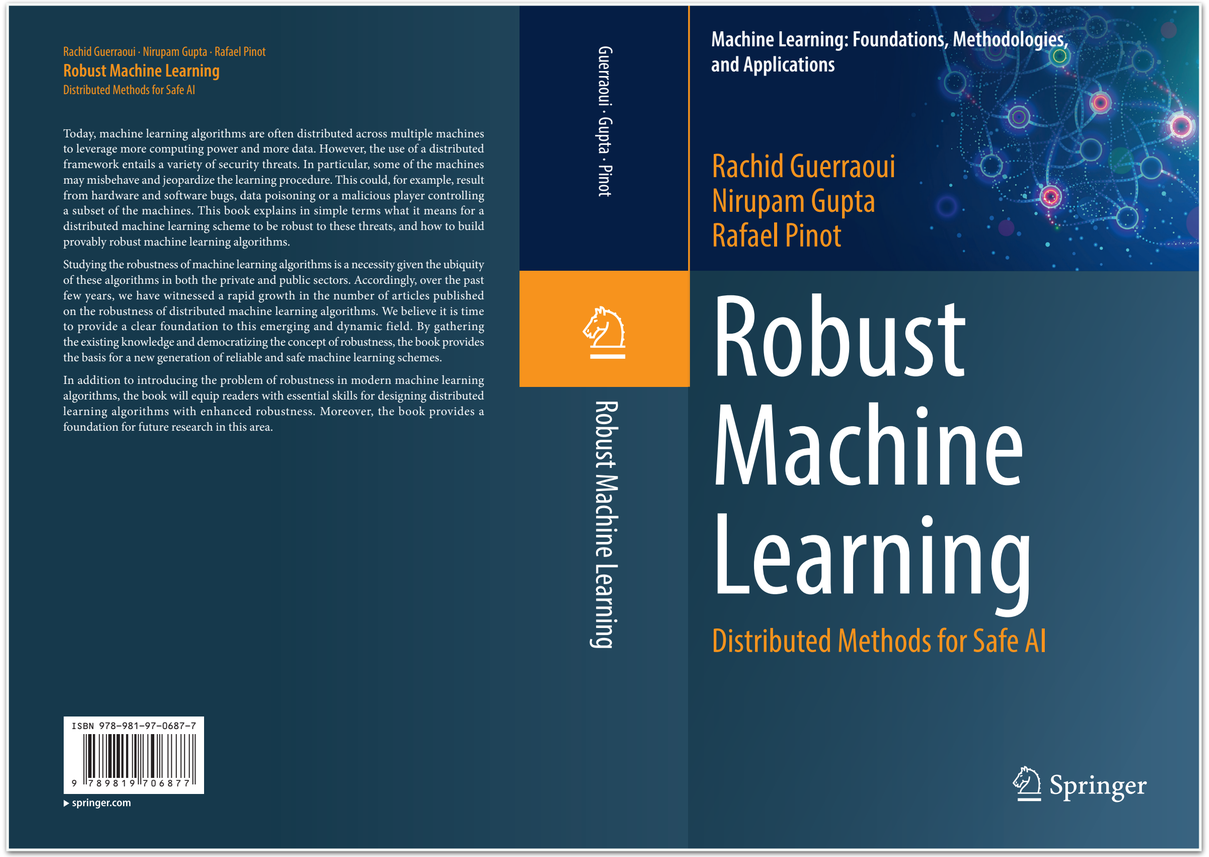
## ***Game-theoretic* approach**

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.




BYZFL

# Thank you!

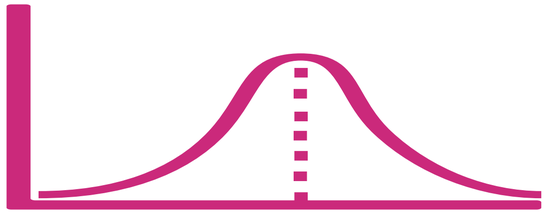
nigu@di.ku.dk

# Challenge III: Robustness with Privacy

Augments local randomness isotropically

**Data Privacy**

Clients do not share exact information on their data

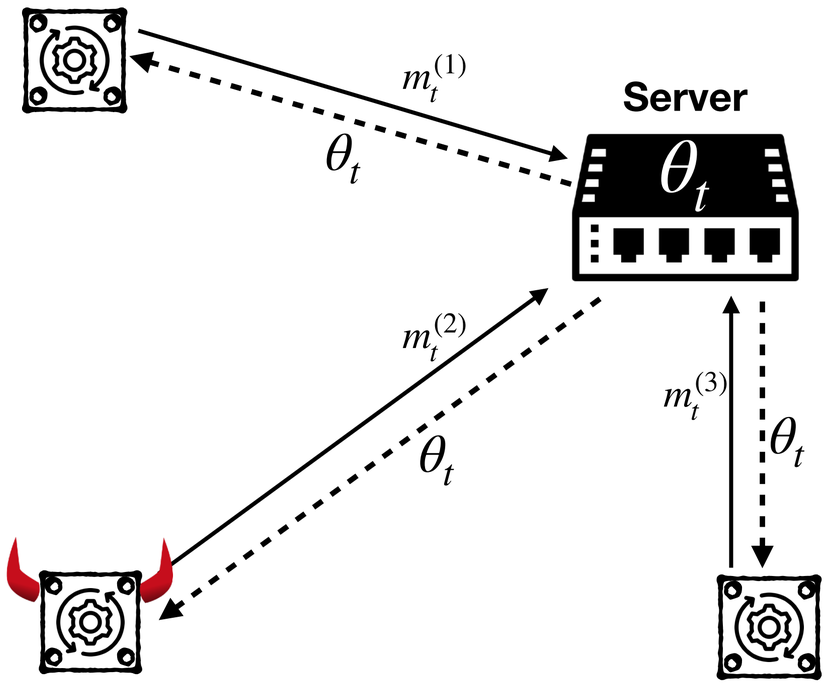


$(\epsilon, \delta)$-Distributed DP

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.

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




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

$\theta_t$ → Membership inference

$\theta_t$ → 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:

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

local stochastic gradient

Additional noise

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)

$(\epsilon, \delta)$-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 adjacent $D,\ D'$

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

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

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 $g_t^{(1)}, \ldots, g_t^{(n)}$ :

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

Inherent gradient noise

**Without distributed momentum** (ignoring $\sigma$):

Overhead due to robustness grows with T

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

# 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

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

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

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

$(f, \kappa)$-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!

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

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

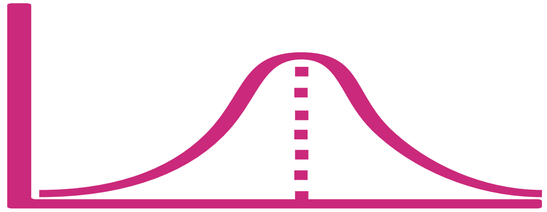
"On the Privacy-Robustness-Utility Trilemma in Distributed Learning". AG**GP**S. ICML 2023

# **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

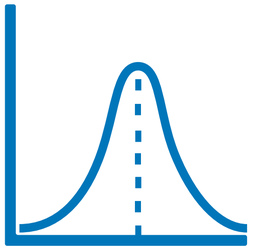
“On the Privacy-Robustness-Utility Trilemma in Distributed Learning”. AG**GP**S**.** ICML 2023

# Take Aways: Three Challenges in Robust DL

State machine replication is not effective

**Local Randomness**

Even honest machines compute noisy updates

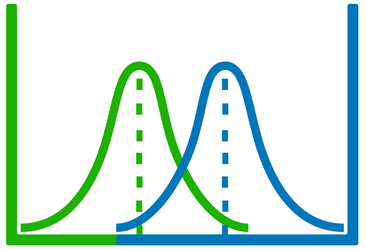


**Robust Averaging**
**Polyak's Momentum**

Adversarial nodes can camouflage as honest nodes

**Data Heterogeneity**

Different nodes sample data from different distributions

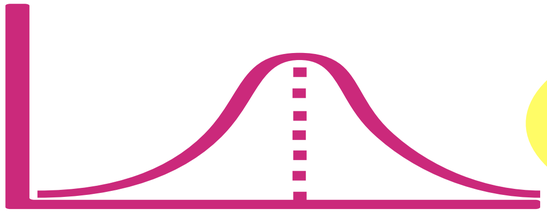


**Pre-aggregation: Nearest Neighbor Mixing (NNM)**

Augments local randomness isotropically

**Data Privacy**

Nodes do not share exact information on their data



**Spectral Robust Averaging**