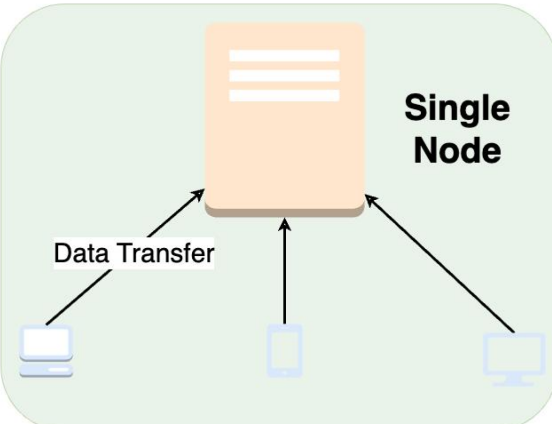
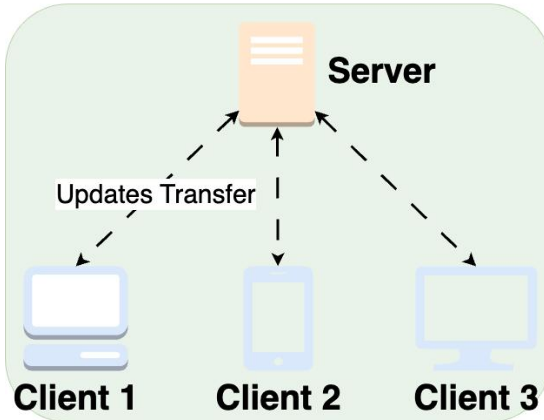
Distributional and Byzantine Robust Decentralized Federated Learning

Kaichuang Zhang, The University of Texas Rio Grande Valley
Ping Xu, The University of Texas Rio Grande Valley
Zhi Tian, George Mason University

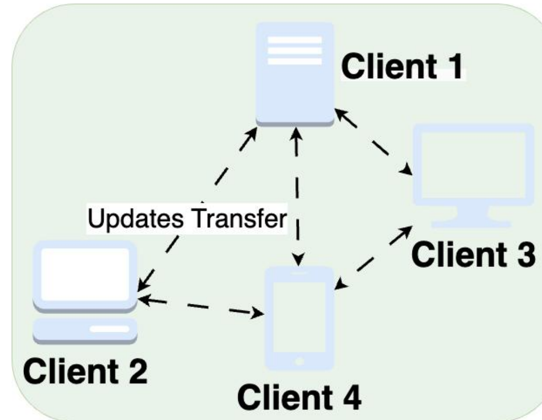
Decentralized Federated Learning



Traditional Machine Learning



Federated Learning



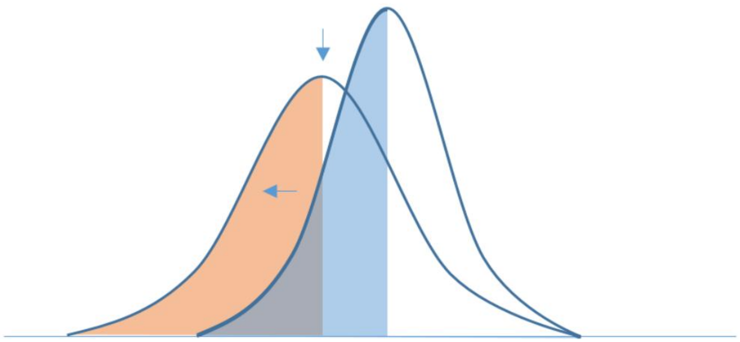
Decentralized Federated Learning

- **Traditional Machine Learning**
  - Collect datasets into a single node.
  - Data privacy issue.
- **Federated Learning (FL) [1]**
  - Without data sharing.
  - Server vulnerable.
- **Decentralized Federated Learning (DFL)**
  - Without a server.
  - More Robust.

Still have some Robustness issues in DFL:
Distributional Shift and **Byzantine Attack.**

Challenge 1: Distributional Shift in DFL

- **Distributional shift:** a mismatch between the distributions of train data and test data.



- Traditional Machine Learning under **Empirical Risk Minimization (ERM)** assumes that train data and test data share the same distribution, but usually fails in practice.

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{x_n \sim \mathcal{D}_n} \, f_n(w; x_n)$$

- **Prior work to resolve distributional shifts**
  - Adaptive Regularization [2]
  - Invariant Risk Minimization [3]
  - Distributionally Robust Optimization (DRO) [4]

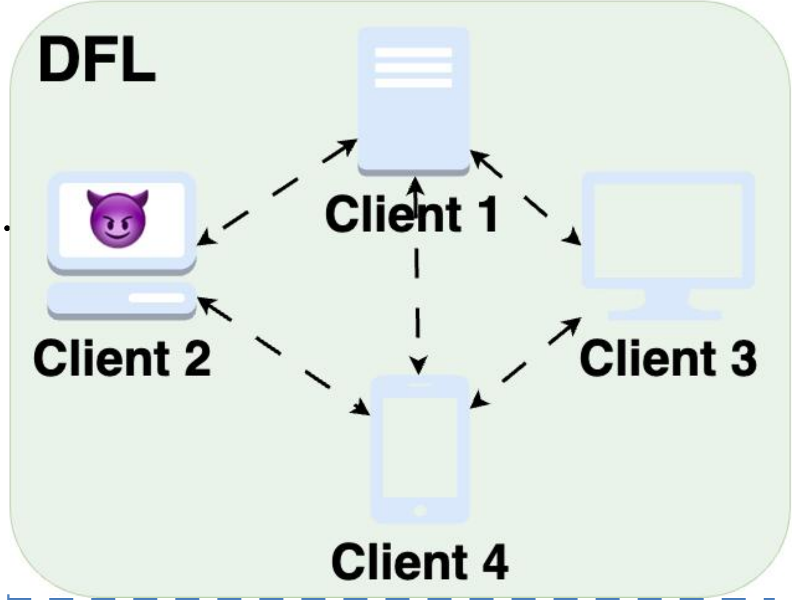
$$\min_{w \in \mathbb{R}^d} \sup_{\mathcal{Q} \in \Omega} \mathbb{E}_{x \sim \mathcal{Q}} \, f(w; x).$$

    - Construct the ambiguity set Ω based on probability distance, Wasserstein distance, etc.

Challenge 2: Byzantine Attack in DFL

- ***Byzantine attacks*** refer to dishonest clients who send arbitrary malicious information to intentionally disrupt the entire system.

- **Byzantine Robust Aggregation Algorithms**
  - Statistics
    - Median [1], Trimmed Mean [1], etc.
  - Anomaly detection
    - pre-trained autoencode [2]
  - Performance evaluation
    - Requires an evaluation dataset.
    - Straightforward and efficient.



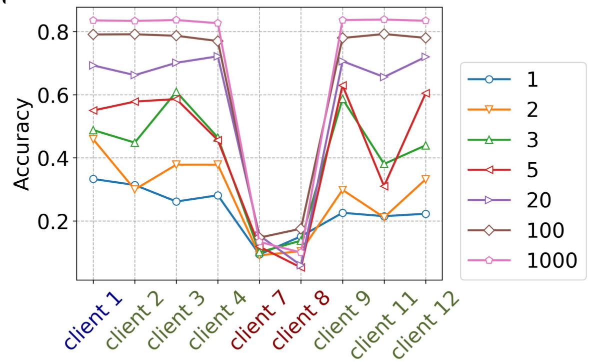
- **Our contributions**
  - A Byzantine-robust aggregation algorithm designed to eliminate the negative impact of Byzantine attackers.
  - The **first** framework that achieves Byzantine robustness and distributional robustness simultaneously.

Byzantine Robust: LPE-TSR

Local Performance Evaluation with Temperature-Scaled Softmax Reweighting (LPE-TSR)

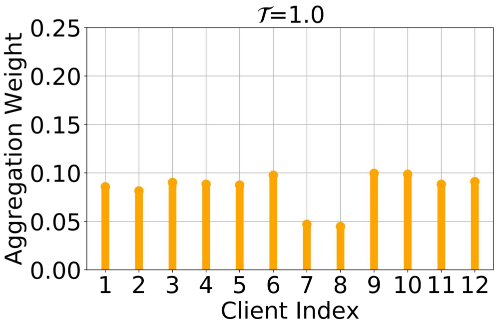
Local Performance Evaluation

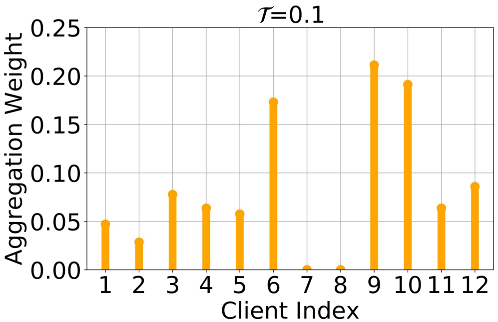
- Determine the updates are benign or malicious using an evaluation dataset.



Temperature-Scaled Softmax Reweighting

- Assign larger weights to benign updates and smaller weights to malicious updates.
- Accelerate convergence. $\mathrm{Softmax}_T(z_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$





LPE-TSR summary:

Step 1 Build evaluation dataset; → **Step 2** Perform evaluation; → **Step 3** Filter out malicious updates; → **Step 4** Aggregation;

DB-Robust DSGD

Distributional and Byzantine Robust Decentralized Stochastic Gradient Descent

- **Distributed Wasserstein DRO for handling distributional shifts.**
  - Wasserstein distance to build ambiguity sets: $\mathcal{Q}_n : W_c(\mathcal{Q}_n, \mathcal{D}_n) \leq \rho_n$
  - Optimization problem:
    - ERM: $$\min_{w \in \mathbb{R}^d} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{x_n \sim \mathcal{D}_n} f_n(w; x_n)$$
    - DRO: $$\min_{w \in \mathbb{R}^d} \frac{1}{|\mathcal{B}|} \sum_{n=1}^{|\mathcal{B}|} \sup_{\mathcal{Q}_n : W_c(\mathcal{Q}_n, \mathcal{D}_n) \leq \rho_n} \mathbb{E}_{x_n \sim \mathcal{Q}_n} f_n(w; x_n)$$

- **The Byzantine Robust Aggregation Algorithm is a plugin in DB-Robust DSGD framework .**

DB-Robust DSGD

Traditional DFL

$$\min_{w \in \mathbb{R}^d} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{x_n \sim \mathcal{D}_n} f_n(w; x_n)$$

- Local SGD training;

$$w_n^{t+\frac{1}{2}} = w_n^t - \eta^t \nabla f_n(w_n^t; x_n^t), x_n^t \sim \mathcal{D}_n$$

- Communication;

$$W_n = \{w_n^{t+\frac{1}{2}}\} \cup \{w_m^{t+\frac{1}{2}} | m \in \mathcal{N}_n\}$$

- Aggregation;

$$w_n^{t+1} = \sum_{w_i \in W_n} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} w_i^{t+\frac{1}{2}}$$

Distributional and Byzantine robust DFL

$$\min_{w \in \mathbb{R}^d} \frac{1}{|\mathcal{B}|} \sum_{n=1}^{|\mathcal{B}|} \sup_{\mathcal{Q}_n : W_c(\mathcal{Q}_n, \mathcal{D}_n) \leq \rho_n} \mathbb{E}_{x_n \sim \mathcal{Q}_n} f_n(w; x_n)$$

- Local SGD training;

$$w_n^{t+\frac{1}{2}} = w_n^t - \eta^t \nabla f_n(w_n^t; x_n^t), x_n^t \sim \mathcal{Q}_n$$

$$w_n^{t+\frac{1}{2}} = \star$$

- Communication;

$$W_n = \{w_n^{t+\frac{1}{2}}\} \cup \{w_m^{t+\frac{1}{2}} | m \in \mathcal{N}_n\}$$

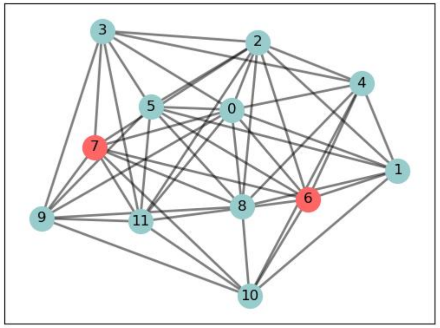
- **Robust** Aggregation;

$$w_n^{t+1} = \mathbf{BRAgg}(w_i^{t+\frac{1}{2}} | w_i^{t+\frac{1}{2}} \in W_n);$$

where N is the number of clients in DFL system; w is local the weights of local model; D_n is the original data distribution of client n; D is the sum of D_n; f is loss function; and η is the learning rate; BRAgg() is Byzantine Robust Aggregation Rules.

Experimental Setups

- **Network topology.** A random undirected graph containing $|\mathcal{B}|$ benign nodes and $|\mathcal{M}|$ malicious nodes, characterized by a connection probability of ρ .
- **Datasets.**
  - Fashion MNIST: An image dataset of 10 categories, containing 60,000 (train) and 10,000 (test) samples.
  - Spambase: 4,601 email samples and 57 features.



|B| = 10, |M| = 2, p=0.7.

- **Distributional shifts.**
- **Byzantine attacks.**

$$\begin{cases} L_1 \text{ shift: } \|z - x\| \leq q \\ L_2 \text{ shift: } \|z - x\|_2 \leq q \end{cases}$$

  - Gaussian Attack (GA) , Sign-Flipping Attack (S-F),
  - A Little Is Enough Attack (ALIE), Same-Value Attack (SA)
- **Evaluation metric.** We report the average test accuracy (**Acc.**) of all local models on the test dataset.

Experimental Result

Evaluate Byzantine robustness of LPE-TSR

Benchmark

- **No-AT:** no attacker, theoretical upper bound;
- Median / Trimmed Mean and Krum.

Result

- **LPE-TSR** is better than **No-AT** in some scenarios;
- **LPE-TSR** is better than Median / Trimmed Mean and Krum.

ATTACK	DATA	No-AT	Median	Trimmed Mean	Krum	LPE-TSR(ours)
GA	IID	82.32	82.21	82.31	82.31	82.41
	Non-IID	82.27	82.01	82.15	81.83	82.20
S-F	IID	82.32	79.71	79.20	82.31	82.40
	Non-IID	82.27	74.62	73.20	81.83	82.20
ALIE	IID	82.32	82.19	82.33	82.15	82.37
	Non-IID	82.27	81.10	81.85	79.88	81.12
SA	IID	82.32	79.79	79.22	82.19	82.40
	Non-IID	82.27	74.82	73.29	10.00	82.20

Experimental Result

Evaluate robustness of DB-Robust DSGD

Benchmark

- Empirical Risk Minimization (ERM).

Result

- DB-Robust (*) is better than ERM when just Byzantine Attack or Distributional Shift is exist;
- DB-Robust (*) outperforms ERM under the same scenarios when Byzantine Attack and Distributional Shift is exist at the same time;

	ERM	DB-Robust(Median)	DB-Robust(TM)	DB-Robust(Krum)	DB-Robust(LPE-TSR)
No-Shifts & No-AT	93.48	92.1	93.02	92.83	92.89
No-Shifts & GA	54.35	91.92	92.44	93.94	92.60
L1 Shifts & No-AT	90.92	90.60	91.20	92.21	91.20
L2 Shifts & No-AT	68.03	71.54	71.28	68.70	71.69
L1 Shifts & GA	54.95	90.61	90.89	91.10	90.59
L2 Shifts & GA	52.90	71.39	71.62	71.24	71.50

Conclusion & Future Work

Conclusion

- ***Local Performance Evaluation with Temperature-Scaled Softmax Reweighting (LPE-TSR)*** efficiently and effectively mitigates the negative impact of Byzantine clients in DFL systems.
- ***DB-Robust DSGD*** addresses distributional shifts and Byzantine attacks simultaneously. Distributed Wasserstein DRO is used to mitigate distributional shifts, while Byzantine-robust aggregation algorithms counter Byzantine attacks.
- Experimental results demonstrate that the proposed algorithms achieve superior accuracy and robustness compared to benchmark methods.

Future work

- ***Theoretical analysis*** of DB-Robust DSGD and LPE-TSR.
- Other methods to address distribution shifts.
- ***Theoretical analysis*** of why LPE-TSR outperforms No-AT.

References

1. McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Artificial intelligence and statistics. PMLR, 2017: 1273-1282.

2. Hansen L K, Rasmussen C E, Svarer C, et al. Adaptive regularization[C]//Proceedings of IEEE Workshop on Neural Networks for Signal Processing. IEEE, 1994: 78-87.

3. Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization[J]. arXiv preprint arXiv:1907.02893, 2019.

4. Lin F, Fang X, Gao Z. Distributionally robust optimization: A review on theory and applications[J]. Numerical Algebra, Control and Optimization, 2022, 12(1): 159-212.

5. Shi J, Wan W, Hu S, et al. Challenges and approaches for mitigating byzantine attacks in federated learning[C]//2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2022: 139-146.

6. Yin D, Chen Y, Kannan R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates[C]//International conference on machine learning. Pmlr, 2018: 5650-5659.

7. Li S, Cheng Y, Liu Y, et al. Abnormal client behavior detection in federated learning[J]. arXiv preprint arXiv:1910.09933, 2019.