FLGuard : Byzantine-robust Federated Learning via Contrastive Models

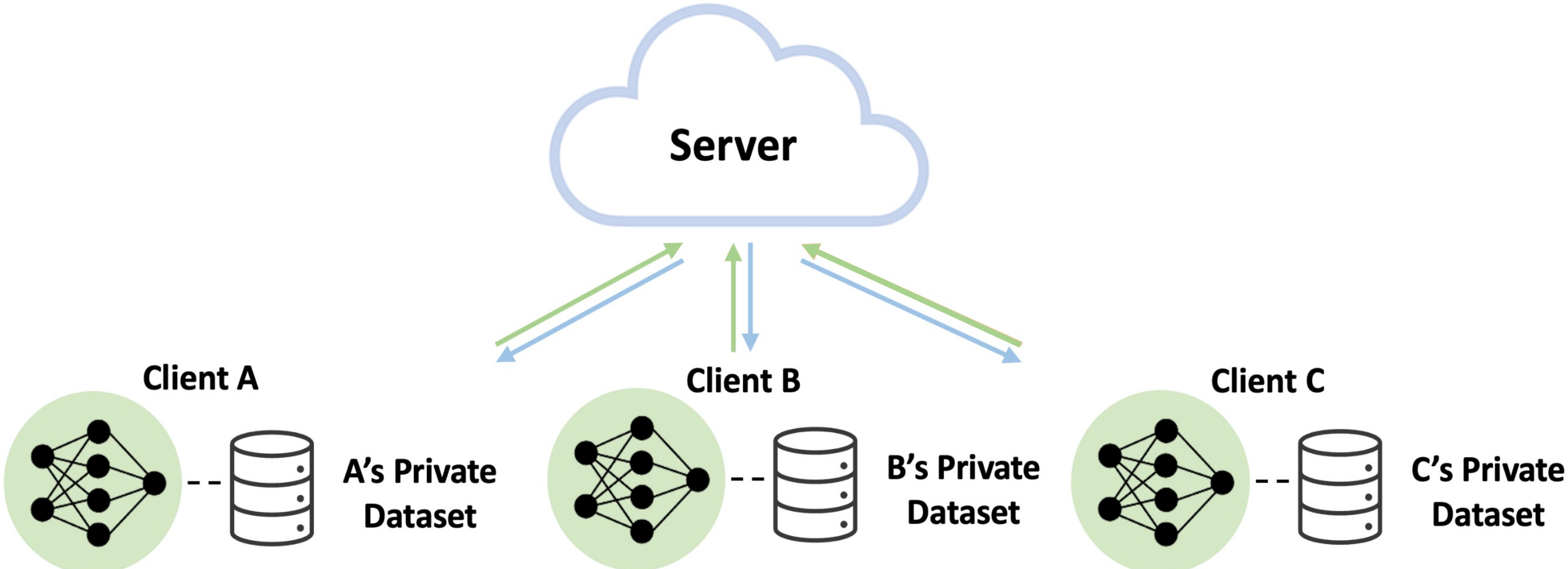
Younghan Lee[1], Yungi Cho[1], Woorim Han[1], Ho Bae[2] and Yunheung Paek[1]

[1] Seoul National University

[2] Ewha Woman's University

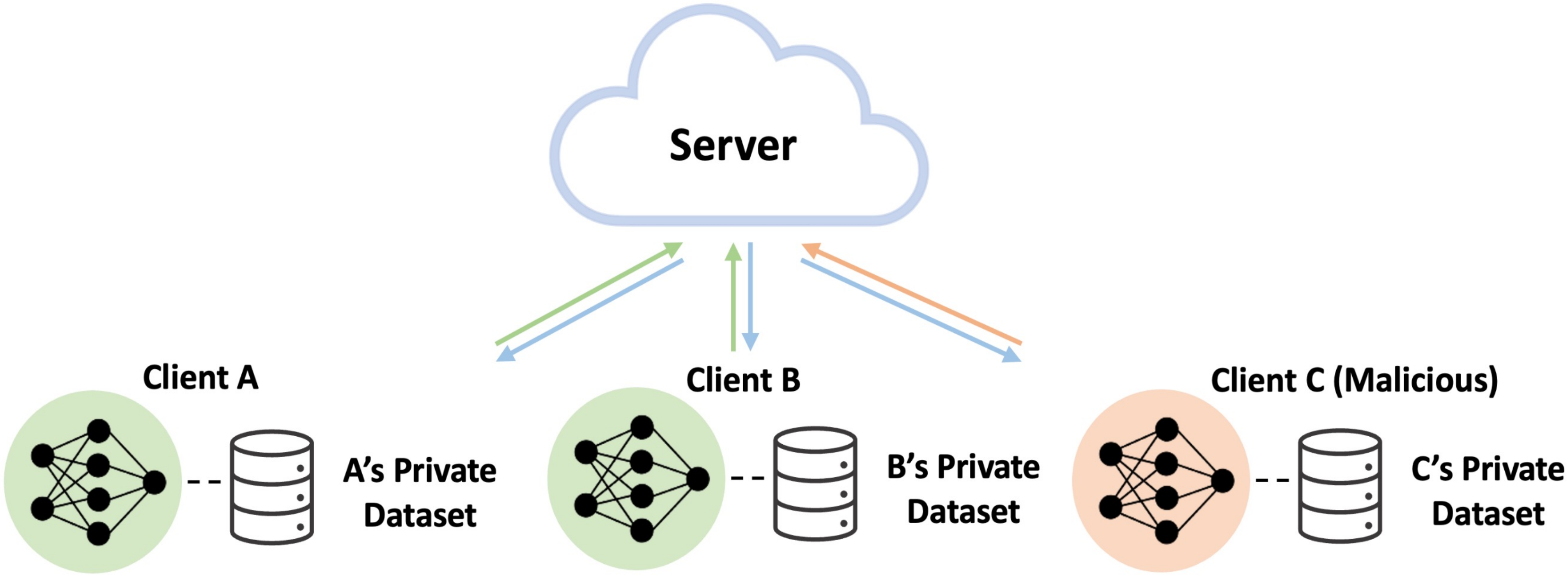
Table of Contents

FL – How it works



	Step 1	Step 2	Step 3
Server	Initializes the global model		Updates the global model
Client	Downloads the global model	Performs local updates & uploads to the server	Awaits the server

FL – Poisoning attacks



- Malicious clients attempts to degrade the performance of AI model
  - Model poisoning attack & Data poisoning attack
- Threat to integrity and availability of AI model

FL – Poisoning attacks

Type	Adversaries' Capability	Adversaries' Knowledge	
		Local Updates of Benign Clients	Server's AGR Algorithm
Type-1 (T1)	Model Poisoning	✓	✓
Type-2 (T2)	Model Poisoning	✗	✓
Type-3 (T3)	Model Poisoning	✓	✗
Type-4 (T4)	Model Poisoning	✗	✗
Type-5 (T5)	Data Poisoning	✗	✗

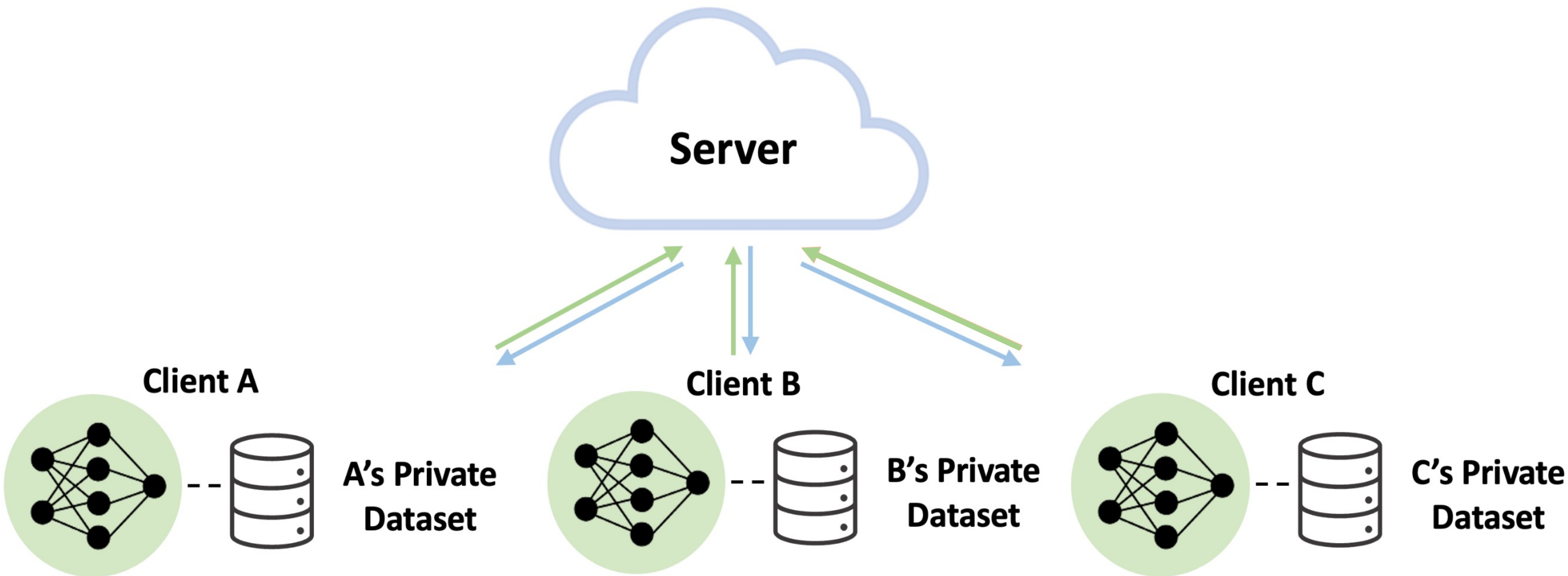
- Adversaries' objective is indiscriminate
  - aims to misclassify any data samples
- Type-1 represents the strongest adversaries

Byzantine-robust FL

- Preserve the performance of AI model
  - Fidelity – Not sacrifice accuracy when no adversaries are present
  - Robustness –Persist the accuracy when adversaries are present
  - Efficiency – Not cause an overhead that will delay the training

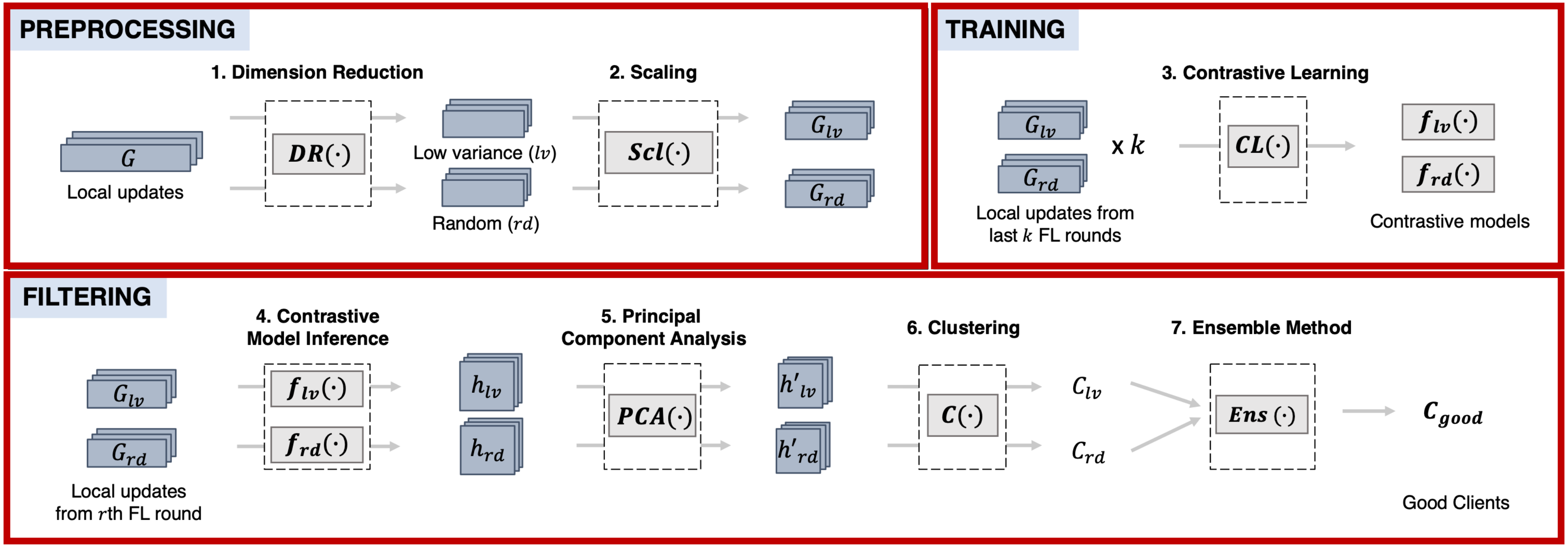
- Current Limitation
  - Requires additional information about FL
    - Number of malicious clients present in FL (statistical info)
    - Auxiliary dataset
  - Not effective under non-IID settings

Our Method – Overview

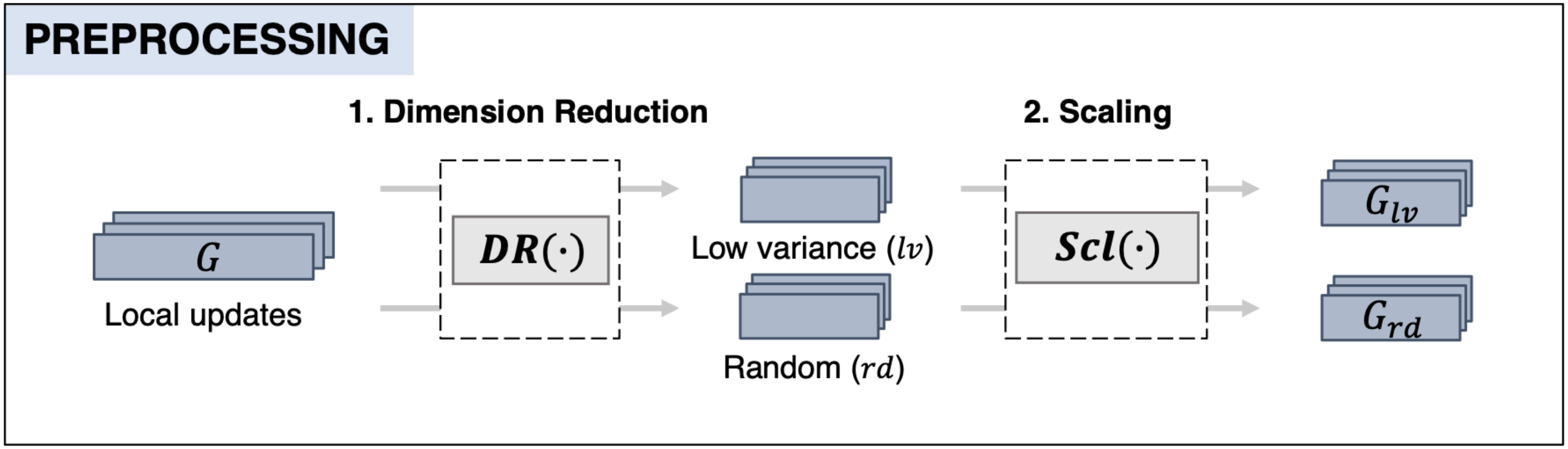


	Step 1	Step 2	Step 3
Server	Initializes the global model		Updates the global model
Client	Downloads the global model	Performs local updates & uploads to the server	Awaits the server

Our Method – In details

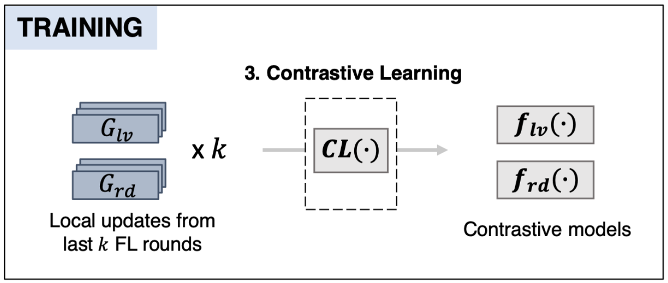


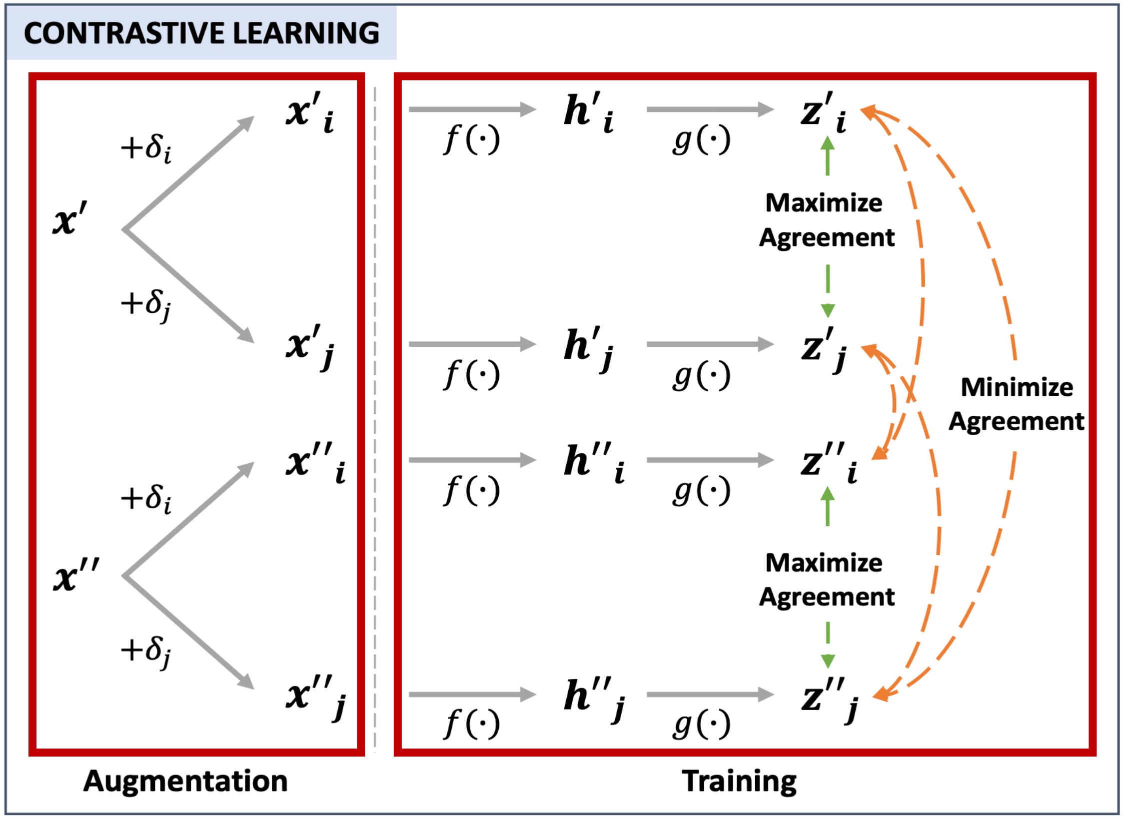
Our Method – Preprocessing



- Two different dimension reduction techniques
  - Different attacks requires different techniques to defend
- MaxAbs scaler to keep the sign of original local updates
  - Improves the accuracy by 20% compared to MinMax scaler

Our Method – Contrastive Learning

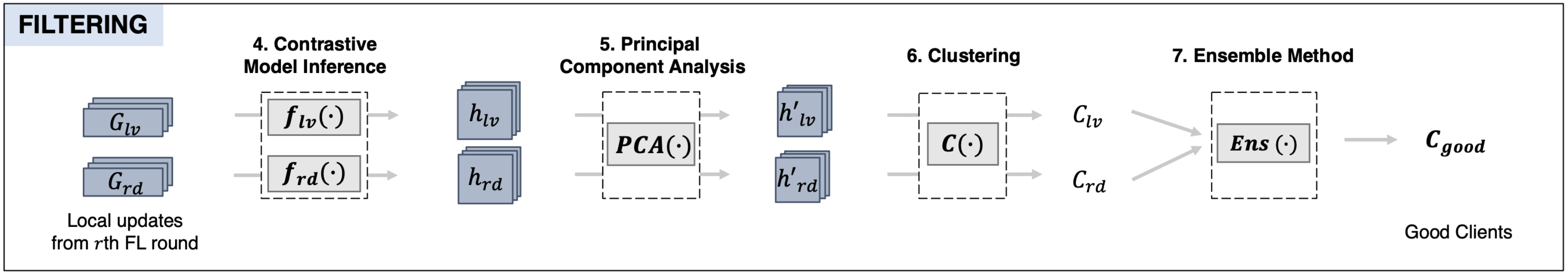




- Gaussian noise
- $l_2 similarity$

$$l(z_{i,j}) = -\log \frac{\exp(sim(z_{i,j})/\tau)}{\sum_{k=1}^{2B} \mathbb{1}_{[k\neq i]} \exp(sim(z_{i,k})/\tau)}$$

Our Method – Filtering



- Projects the local updates to representation: h_{lv} and h_{rd}
- PCA to reduce the dimension down for clustering
- Agglomerative hierarchical clustering to form two clusters
- We consider bigger cluster as the benign and ensemble: $C_{lv} \cap C_{rd}$

Experiments – Setup

	Details	MNIST	CIFAR-10	FEMNIST
R	# FL Rounds	300	2,000	1,500
N	# Total Clients	100	50	3,400
M	# Malicious Clients	20	10	680
P	# Participating Clients	N		60
b	Batch Size	250	32	250
η	Global Learning Rate	0.001	0.01	0.001
Opt	Optimizer	Adam	SGD	Adam
$Arch$	Global Model Architecture	FCN	ResNet-14	ConvNet

Experiments – Poisoning attacks

- Goal is to create malicious local updates
- $g_m = g_b + \gamma p,\ where\ p = perturbation\ vector$
- $Inverse\ unit\ vector = -\left({}^{g_b}/_{\|g_b\|_2}\right),\ Inverse\ sign = -\operatorname{sgn}(f_{avg}\ (g_b))$
- Finding an optimal γ is the main challenge

- Threat model Type 1 & 2
  - Static optimization approach (USENIX Security 20)
  - Dynamic optimization approach (NDSS 21)
- Threat model Type 3 & 4
  - Little Is Enough (NIPS 19)
  - Min-Sum and Min-Max (NDSS 21)
  - Sign Flip (NIPS 19)
- Threat model Type 5
  - Static Label Flip (NDSS 21)
  - Dynamic Label Flip (S&P 22)

Experiments – Fidelity

Dataset (Distr.)	AGR	No Attack
MNIST-0.1 (IID)	TrMean (ICML'18)	96.98
	MKrum (NIPS'17)	96.37
	Bulyan (ICML'18)	95.92
	DnC (NDSS'21)	97.06
	FLTrust (NDSS'21)	95.96
	SignGuard (ICDCS'22)	97.20
	FLGuard (Ours)	**97.24**
CIFAR10 (IID)	TrMean (ICML'18)	71.92
	MKrum (NIPS'17)	71.41
	Bulyan (ICML'18)	56.62
	DnC (NDSS'21)	72.44
	FLTrust (NDSS'21)	70.74
	SignGuard (ICDCS'22)	70.64
	FLGuard (Ours)	**72.73**

Dataset (Distr.)	AGR	No Attack
MNIST-0.5 (Non-IID)	TrMean (ICML'18)	95.96
	MKrum (NIPS'17)	96.19
	Bulyan (ICML'18)	94.38
	DnC (NDSS'21)	96.57
	FLTrust (NDSS'21)	95.47
	SignGuard (ICDCS'22)	**96.94**
	FLGuard (Ours)	96.79
FEMNIST (Non-IID)	TrMean (ICML'18)	80.62
	MKrum (NIPS'17)	83.69
	Bulyan (ICML'18)	69.89
	DnC (NDSS'21)	83.87
	FLTrust (NDSS'21)	81.83
	SignGuard (ICDCS'22)	83.56
	FLGuard (Ours)	**84.74**

- FedAvg without attacks (baseline)
- MNIST-0.1 : 97.24%, CIFAR-10 : 73.54%, MNIST-0.5 : 97.16%, FEMNIST : 84.11%

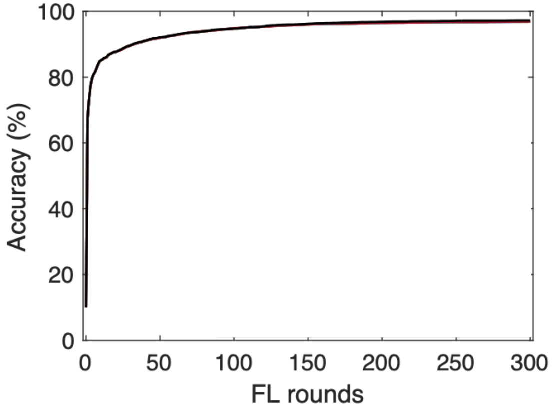
Experiments – Robustness

Dataset (Distr.)	AGR	Type-1						Type-2					
		STAT-OPT		DYN-OPT		Adaptive		STAT-OPT		DYN-OPT		Adaptive	
		Krum	TM	Krum	TM	DnC	FLG	Krum	TM	Krum	TM	DnC	FLG
CIFAR10 (IID)	TrMean (ICML'18)	58.16	46.53	63.82	52.11	69.32	71.83	58.73	44.48	58.89	53.33	59.88	71.39
	MKrum (NIPS'17)	31.82	45.11	41.94	71.49	51.58	69.56	41.66	41.92	49.49	71.06	68.06	69.58
	Bulyan (ICML'18)	31.84	40.38	35.86	49.39	41.96	67.51	33.12	38.35	36.47	48.15	50.89	64.89
	DnC (NDSS'21)	72.73	71.55	64.94	72.65	47.28	70.88	72.22	72.85	70.76	71.96	72.08	70.84
	FLTrust (NDSS'21)	71.00	56.98	65.02	70.80	67.45	71.47	70.27	53.04	71.29	70.54	71.06	68.99
	SignGuard (ICDCS'22)	66.72	72.52	68.63	69.66	70.21	69.70	60.88	72.71	68.71	72.38	70.86	70.68
	FLGuard (Ours)	**73.44**	**72.71**	**73.86**	**73.60**	**72.48**	**72.02**	**73.19**	**72.99**	**73.15**	**73.44**	**73.30**	**72.36**
FEMNIST (Non-IID)	TrMean (ICML'18)	56.78	76.10	62.52	63.69	49.44	80.06	73.98	76.73	74.28	75.68	78.84	81.11
	MKrum (NIPS'17)	5.05	78.60	5.02	83.71	4.87	81.00	5.67	78.13	4.85	83.74	8.81	81.34
	Bulyan (ICML'18)	53.41	72.77	53.13	66.74	53.47	75.95	48.89	74.98	53.53	67.20	56.50	79.63
	DnC (NDSS'21)	6.97	82.76	4.85	84.03	4.87	80.71	5.26	81.06	4.98	83.63	26.79	80.65
	FLTrust (NDSS'21)	4.60	83.30	35.79	4.58	52.83	80.07	4.68	83.61	36.88	4.46	5.09	79.98
	SignGuard (ICDCS'22)	80.37	**84.19**	10.12	83.58	8.87	**82.15**	8.75	83.96	8.77	83.40	77.64	81.73
	FLGuard (Ours)	**84.14**	83.80	**84.30**	**84.19**	**83.22**	81.86	**83.12**	**84.02**	**82.11**	**83.94**	**81.51**	**83.44**

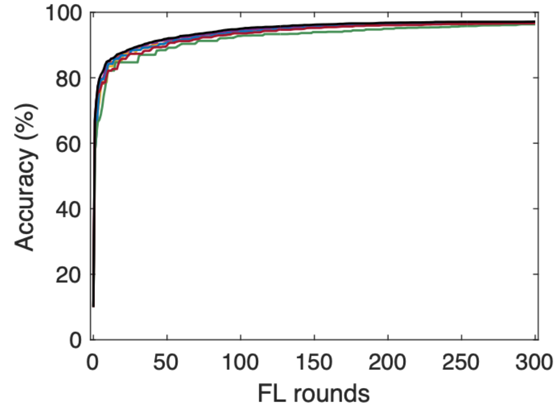
Experiments – Robustness

Dataset (Distr.)	AGR	Type-3						Type-4						
		LIE		Min-Max		Min-Sum		LIE		Min-Max		Min-Sum		SF
		uv	*sgn*	*uv*	*sgn*	*uv*	*sgn*	*uv*	*sgn*	*uv*	*sgn*	*uv*	*sgn*	
CIFAR10 (IID)	TrMean (ICML'18)	72.65	71.61	57.65	60.65	69.36	69.07	73.42	**75.28**	64.85	59.33	68.34	64.67	57.73
	MKrum (NIPS'17)	71.31	71.55	65.32	44.18	65.69	56.39	70.64	70.35	59.21	43.32	64.02	46.61	62.62
	Bulyan (ICML'18)	73.42	53.86	38.80	37.58	45.21	41.07	64.96	54.44	39.31	38.70	40.28	42.67	48.54
	DnC (NDSS'21)	72.24	71.23	71.61	71.29	70.94	55.05	72.24	73.09	71.83	71.57	71.49	60.13	**73.54**
	FLTrust (NDSS'21)	72.24	68.22	71.29	70.54	72.24	57.57	71.00	66.11	70.27	71.45	70.45	57.35	69.83
	SignGuard (ICDCS'22)	72.50	70.68	60.11	69.99	70.05	69.87	72.81	71.96	58.85	68.43	69.95	68.51	53.06
	FLGuard (Ours)	**73.60**	**73.13**	**72.95**	**73.50**	**72.38**	**72.50**	**74.25**	72.85	**72.87**	**72.97**	**72.06**	**72.87**	72.06
FEMNIST (Non-IID)	TrMean (ICML'18)	83.12	83.95	72.09	57.64	81.26	64.11	82.21	83.17	73.62	71.05	80.49	72.24	79.73
	MKrum (NIPS'17)	83.86	72.30	80.18	4.85	82.90	9.33	83.68	**83.80**	78.13	4.87	82.79	11.25	78.23
	Bulyan (ICML'18)	82.60	71.41	58.29	61.21	72.34	34.50	80.86	72.28	57.43	60.39	73.09	45.37	68.70
	DnC (NDSS'21)	83.93	83.59	83.34	44.11	83.36	5.69	83.84	83.63	80.93	5.25	83.51	5.42	81.44
	FLTrust (NDSS'21)	**84.92**	81.97	4.64	59.68	76.16	58.83	83.66	4.85	6.64	6.17	5.63	6.40	14.27
	SignGuard (ICDCS'22)	83.90	83.56	80.09	8.73	83.58	76.80	83.79	83.74	80.10	8.13	83.27	10.80	78.43
	FLGuard (Ours)	84.32	**84.04**	**84.07**	**84.39**	**84.19**	**82.62**	**83.90**	82.41	**83.47**	**82.63**	**84.08**	**83.53**	**83.79**

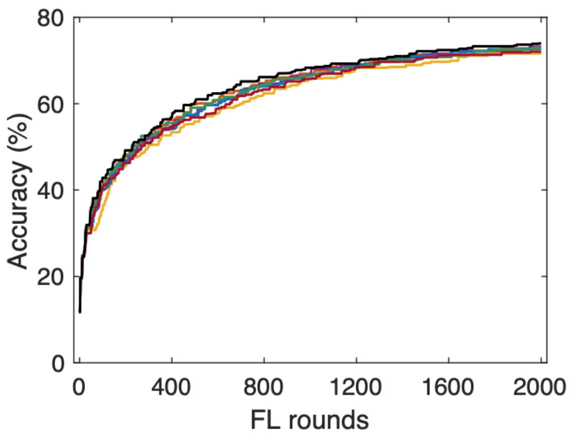
Experiments – Efficiency



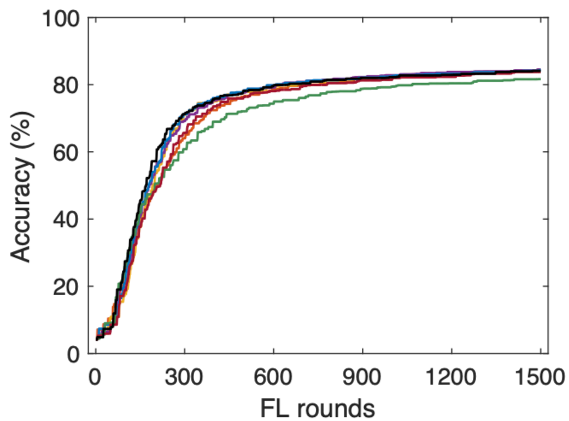
(a) MNIST-0.1



(c) MNIST-0.5



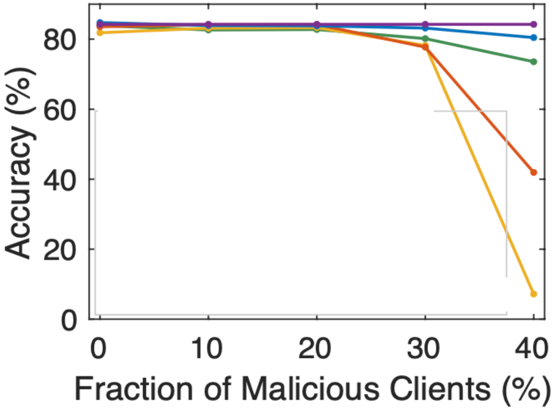
(b) CIFAR-10



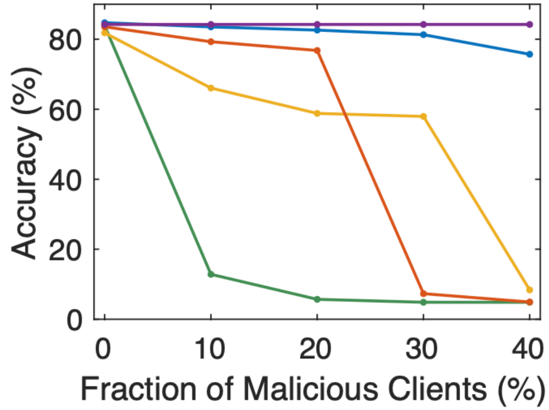
(d) FEMNIST

- No extra FL rounds
- 22.1s to train contrastive models
- 78ms for filtering

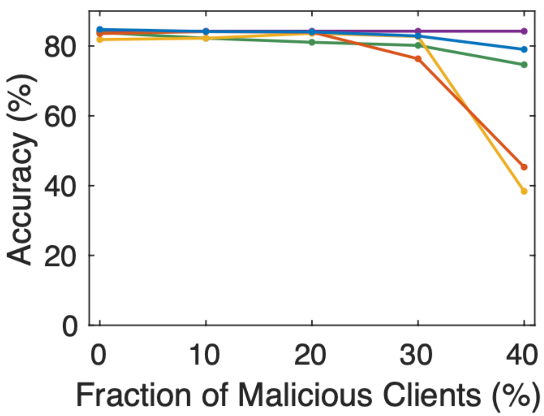
Experiments – Malicious Client %



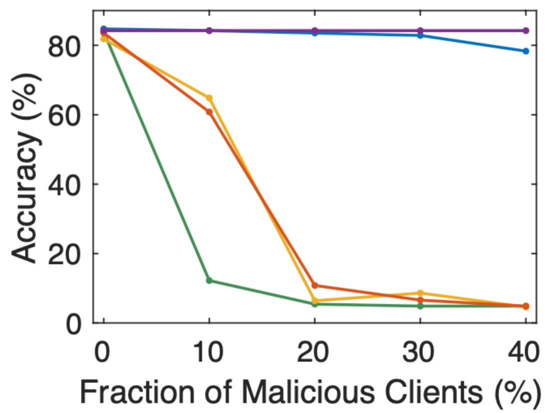
(a) T1 (STAT-OPT)



(c) T3 (Min-Sum)



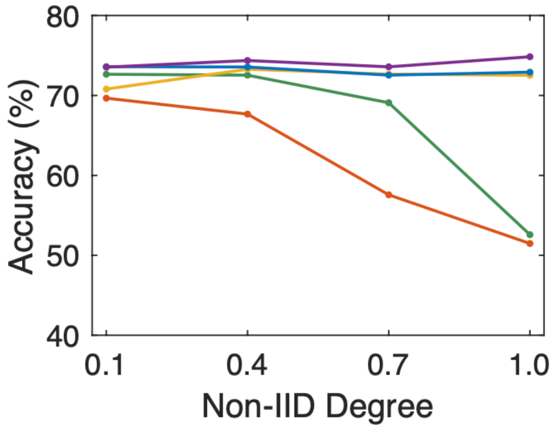
(b) T2 (STAT-OPT)



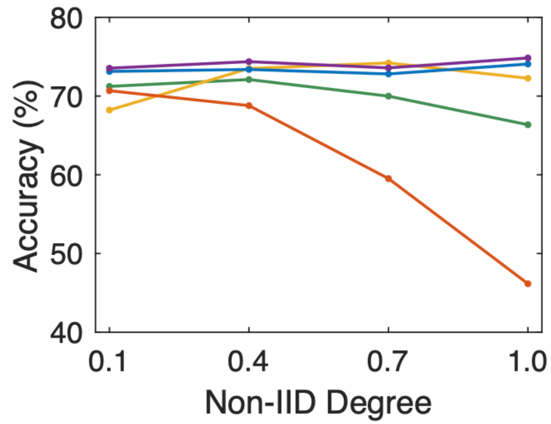
(d) T4 (Min-Sum)

- FEMNIST dataset

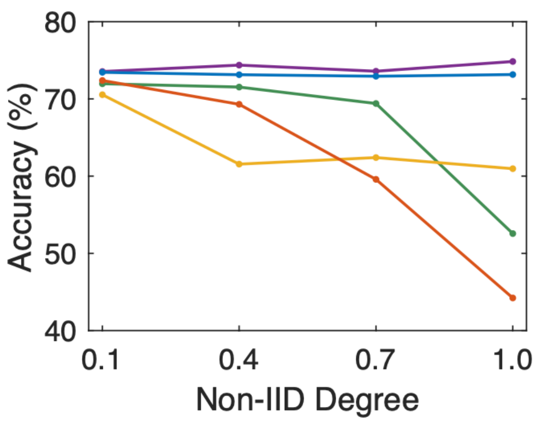
Experiments – Non-IID Degree



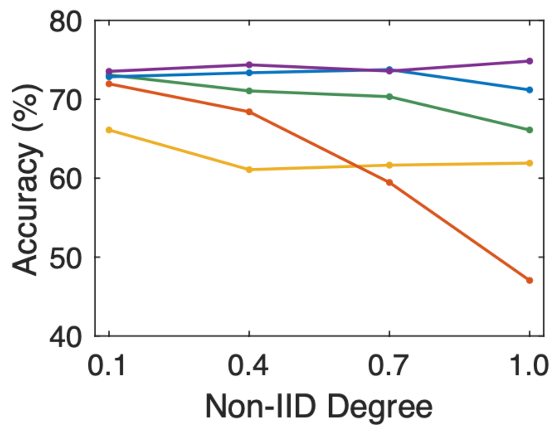
(e) T1 (DYN-OPT)



(g) T3 (LIE)



(f) T2 (DYN-OPT)



(h) T4 (LIE)

- CIFAR10 dataset

Conclusion

- Byzantine-robust FL by employing **contrastive learning**
- FLGuard operates **without prior knowledge** regarding FL
  - No information about the number of malicious client (statistical info)
  - No auxiliary dataset
- FLGuard is robust in **both IID and non-IID dataset settings**
  - No catastrophic failure in non-IID dataset settings
- FLGuard is robust under **an extreme adversarial settings**
  - High percentage of malicious client present
  - Extremely non-IID settings

Thank you for your attention !

Q&A ?