南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

FedES: Federated Early-Stopping for Hindering Memorizing Heterogeneous Label Noise

Bixiao Zeng[1,2], **Xiaodong Yang**[1], **Yiqiang Chen**[*1,2,3], **Zhiqi Shen**[4], **Hanchao Yu**[5] and **Yingwei Zhang**[1]

[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Peng Cheng Laboratory
[4]Nanyang Technological University
[5]Bureau of Frontier Sciences and Education, Chinese Academy of Sciences
{zengbixiao19b, yangxiaodong, yqchen}@ict.ac.cn, zqshen@ntu.edu.sg, {yuhanchao, zhangyingwei}@ict.ac.cn

IJCAI 2024

Introduction

- Existing **federated noisy label learning (FNLL)** addresses noise heterogeneity by distinguishing noisy clients from clean ones.
- 1) discarding clients
- loss of valuable information / noise residue
- 2) detect noisy clients, employ de-noise strategies(pseudo-labeling, knowledge distillation)
- still treat clients as either noisy or clean

- limited exploration

FedCorr

CVPR 2022

$$\widetilde{\mathcal{D}}_k^n = \mathop{\arg\max}_{\substack{\tilde{\mathcal{D}}\subseteq \mathcal{D}_k^n \\ |\tilde{\mathcal{D}}|=\pi\cdot|\mathcal{D}_k^n|}} L_{CE}(\tilde{\mathcal{D}}; f_G^{(t)});$$

$$\widetilde{\mathcal{D}}_k^{n\,\prime} = \{(x,y)\in \widetilde{\mathcal{D}}_k^n \mid \max(f_G^{(t)}(x)) \ge \theta\};$$

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

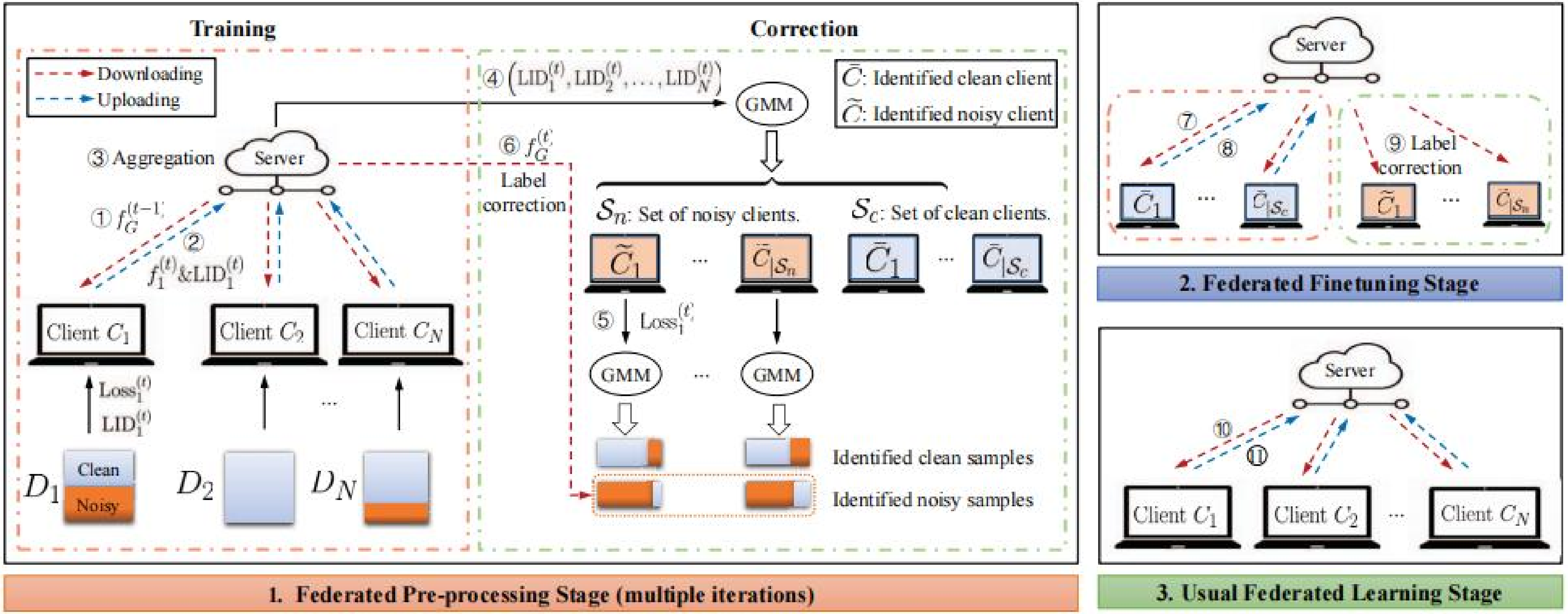


Figure 1. An overview of `FedCorr`, organized into three stages. Algorithm steps are numbered accordingly.

FedNoRo
IJCAI 2023

$$l_i = (l_i^1, l_i^2, ..., l_i^C)^T \in \mathbb{R}^C \qquad l_i^c = \frac{l_i^c - \min_i l_i^c}{\max_i l_i^c - \min_i l_i^c} \qquad d(i) = \min_{j \in S_c} \|w_i - w_j\|_2$$

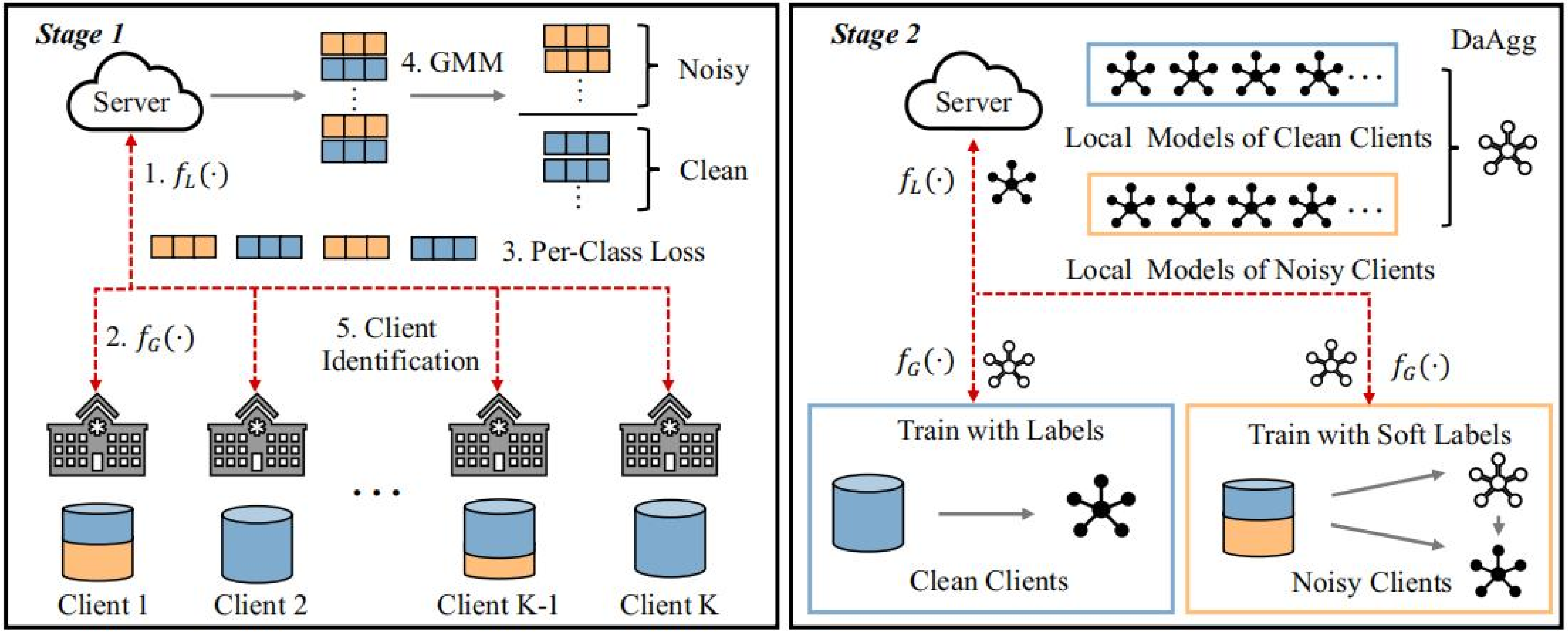


Figure 2: Overview of the proposed two-stage framework FedNoRo.

$$w_g = \sum_{i=1}^{K} \frac{N_i e^{-D(i)}}{\sum_{j=1}^{K} N_j e^{-D(j)}} w_i \qquad y_G = \text{softmax}(\frac{f_G(x)}{T}) \qquad \mathcal{L} = \lambda \mathcal{L}_{KL}(y_p, y_G) + (1-\lambda)\mathcal{L}_{CE}(y_p, \overline{y})$$

FedDiv

AAAI 2024

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

$$\mathbf{p}(\text{“clean”}|x, y; \theta^{(t)}) = P(z = 1|x, y; \theta^{(t)})$$

$$\gamma_{kg}(x, y; \theta_k^{(t)}) = P(z = g|x, y; \theta_k^{(t)}) = \frac{P(\ell(x, y; \theta_k^{(t)})|z = g)P(z = g)}{\sum_{g'=1}^{2} P(\ell(x, y; \theta_k^{(t)})|z = g')P(z = g')}$$

$$\mathcal{D}_k^{\text{clean}} \leftarrow \{(x, y)\,|\,\mathbf{p}(\text{“clean”}|x, y; \theta_k^{(t)}) \geq 0.5, \forall\,(x, y) \in \mathcal{D}_k\}$$

$$\mathcal{D}_k^{\text{noisy}} \leftarrow \{(x,)\,|\,\mathbf{p}(\text{“clean”}|x, y; \theta_k^{(t)}) < 0.5, \forall\,(x, y) \in \mathcal{D}_k\}$$

$$\mathcal{D}_k^{\text{relab}} \leftarrow \{(x, \hat{y})\,|\,\max(\mathbf{p}(x; \theta^{(t)})) \geq \zeta, \forall x \in \mathcal{D}_k^{\text{noisy}}\}$$

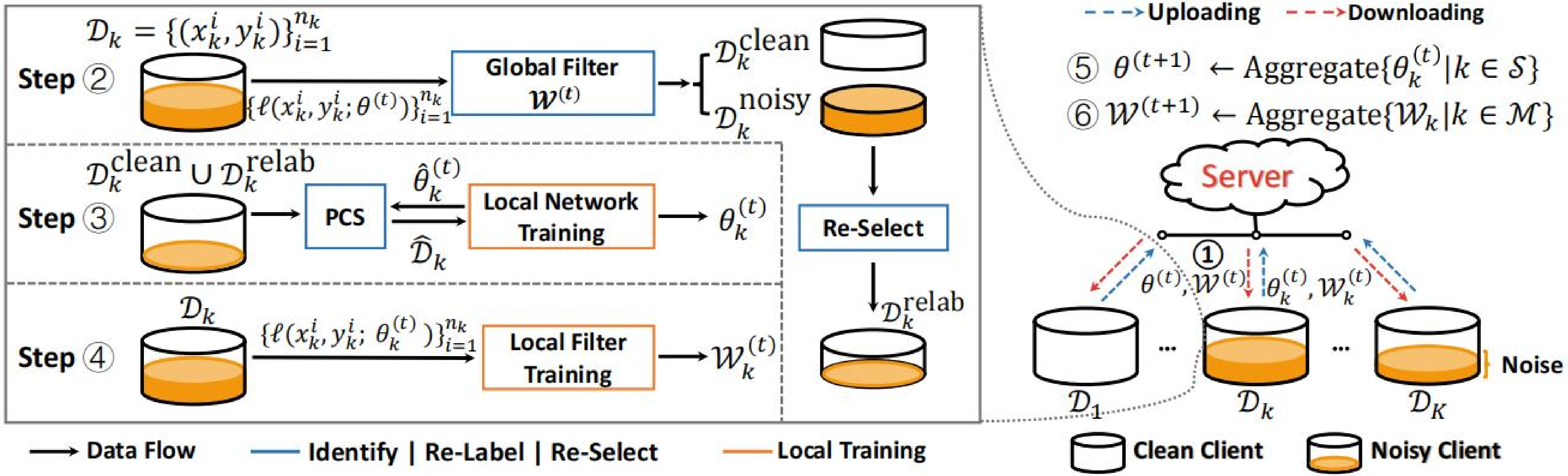


Figure 2: An overview of the training procedure proposed by FedDiv. In this work, the parameters of a local neural model and a local noise filter are simultaneously learned on each client during the local training sessions, while both types of parameters are aggregated on the server.

$$F(x) \leftarrow f(x; \hat{\theta}_k^{(t)}) - \xi \log(\hat{p}_k)$$

$$\hat{p}_k^{(t)} \leftarrow m\hat{p}_k + (1 - m)\frac{1}{n_k} \sum_{x \in \mathcal{D}_k} \mathbf{p}(x; \theta_k^{(t)})$$

$$\mathcal{L}_{final} = \mathcal{L}_{mix} + \eta \mathcal{L}_{reg}$$

$$\tilde{y}(x) = \arg\max F(x)$$

$$\mathcal{D}_k^{\text{resel}} \leftarrow \{(x, y)\,|\,\hat{y}(x) = \tilde{y}(x), \forall (x, y) \in \mathcal{D}_k^{\text{clean}} \cup \mathcal{D}_k^{\text{relab}}\}$$

Introduction

- early-stopping
- explores the dynamic optimization policies during the training of deep neural networks (DNNs)
- memorization effect that **DNNs tend to frst memorize clean labels and then memorize noisy ones**
- Extensive experiments have shown a **positive correlation** between the amount of **clean data and critical parameters, suggesting more clean data need more critical model parameters to memorize them**
- stopping training at a certain time point / on a non-critical segment of DNNs / stopping the training of noise-sensitive layers / **stopping the training of non-critical parameters**
- **these methods all require some prior knowledge(noise rate of training data)**

- In federated learning, noise rates **remain unknown** and exhibit variations among heterogeneous clients

Introduction

1. We present a general noise-robust framework, FedES, to handle noise heterogeneity where clients have varying noise rates instead of a binary noisy-vs-clean problem.
2. We present a general noise-generation approach for modeling federated label noise, incorporating varying noise rates for clients with a continuous spectrum.
3. We estimate each client's noise rate via a signed EMD based on the local and global gradient, without requiring additional information from clients.
4. We demonstrate that FedES outperforms state-of-the-art FL methods on both varying synthetic federated label noise and real-world label noise.

Method——FedES

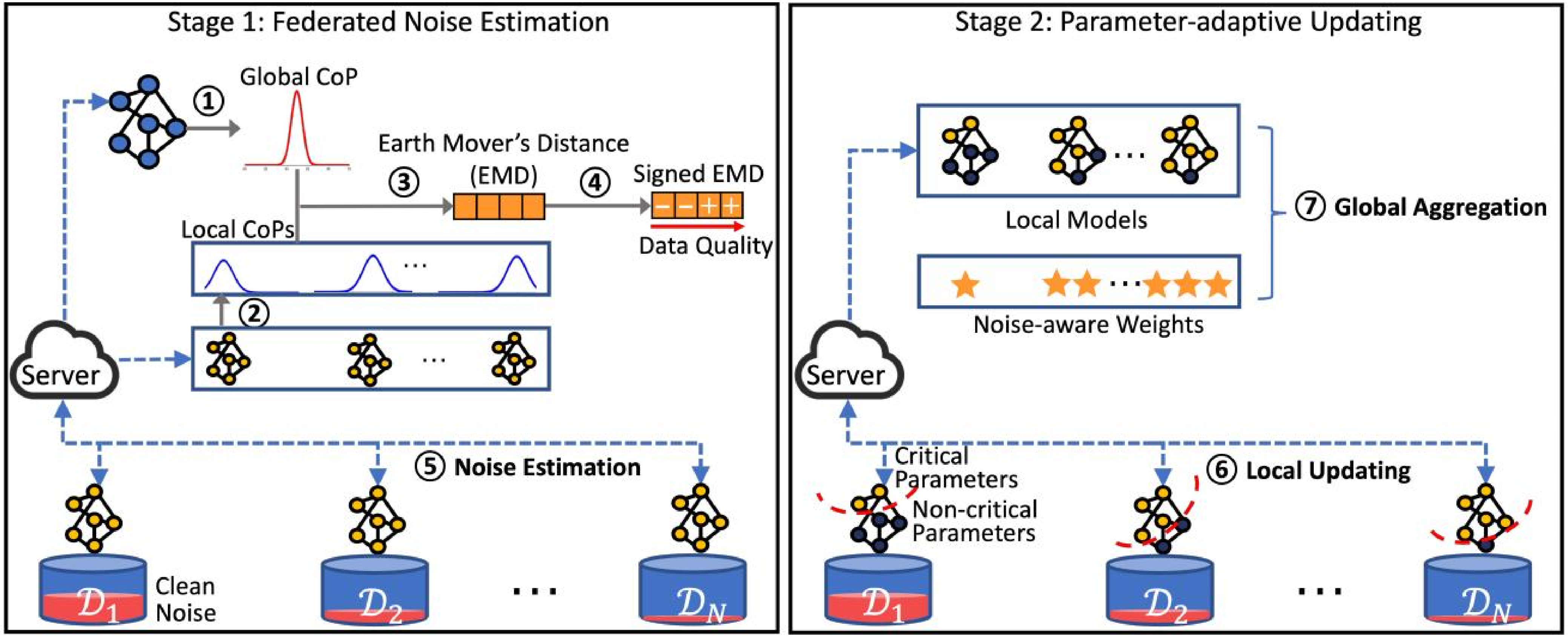
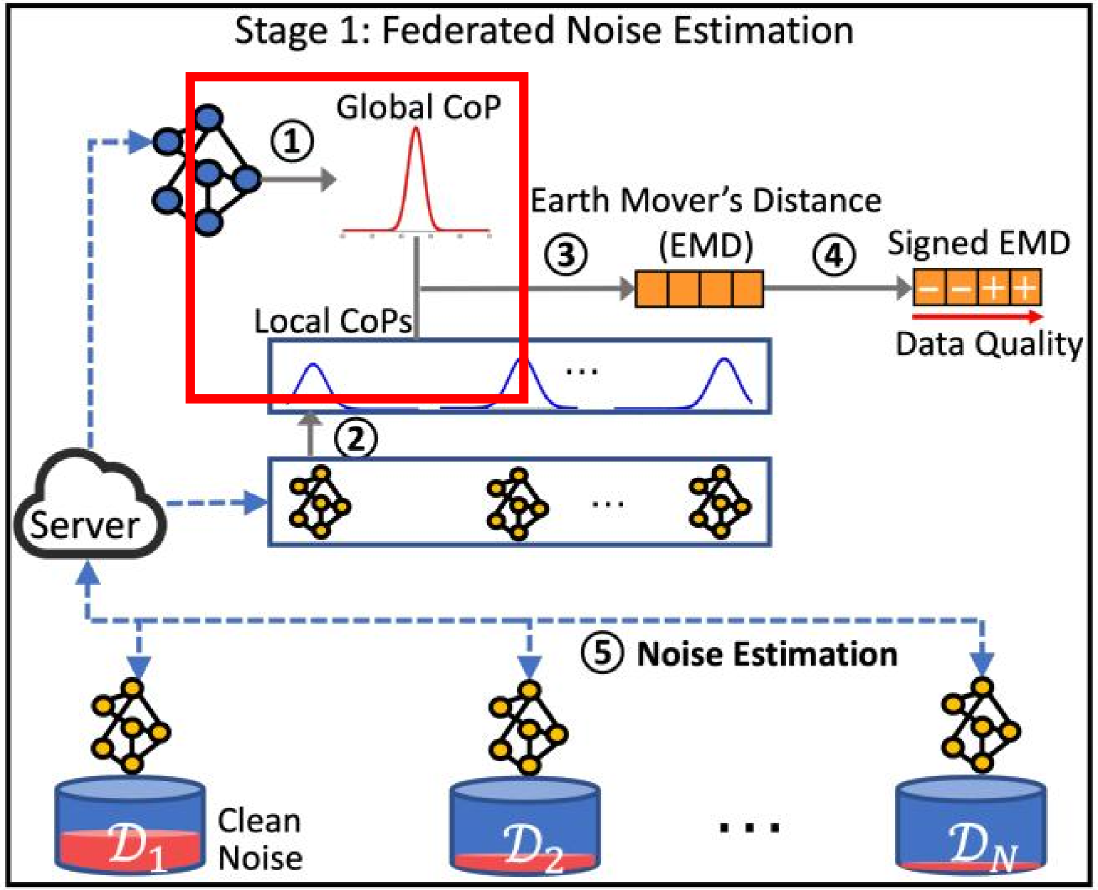


Figure 1: Overview of the proposed two-stage framework FedES.

Method



Criticality of Parameters (CoP)

$$g_i = |\nabla l(\mathrm{w}_i) \times \mathrm{w}_i|, i \in [m]$$

FedAvg Updating

$$\mathcal{W}_n(t+1) = \mathcal{W}(t) - \eta \nabla L_n(\mathcal{W}(t))$$

$$\mathcal{W}(t+1) = \sum_{n}^{N} \frac{|\mathbb{D}_n|}{|\mathbb{D}|} \mathcal{W}_n(t+1)$$

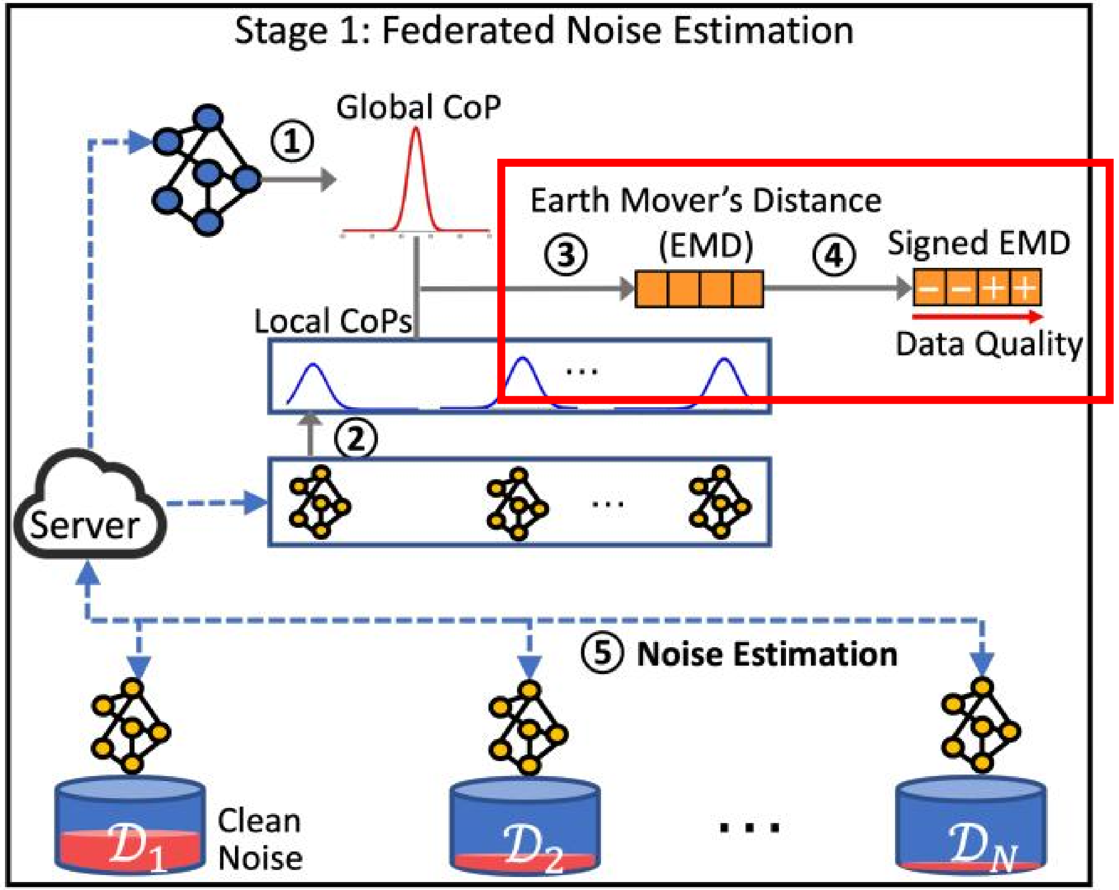
$$= \mathcal{W}(t) - \eta \sum^{N} \frac{|\mathbb{D}_n|}{|\mathbb{D}|} \nabla L_n(\mathcal{W}(t))$$

global CoP

$$\mathbf{g}_s \leftarrow |(\mathcal{W}^{\mathrm{pre}}(s_1+1) - \mathcal{W}^{\mathrm{pre}}(s_1)) * \mathcal{W}^{\mathrm{pre}}(s_1)|$$

local CoP

$$\mathbf{g}_n \leftarrow |(\mathcal{W}_n^{\mathrm{pre}}(s_1+1) - \mathcal{W}^{\mathrm{pre}}(s_1)) * \mathcal{W}^{\mathrm{pre}}(s_1)|$$



Earth Mover's Distance (EMD)

- Higher data quality is associated with a CoP distribution having many large values, and the shape of distributions with varying noise rates differs from the global distribution.
- The distance concerning the CoP of a low-data-quality client may be **the same as** that of a high-data-quality client (a horizontally fipped version of a low-data-quality client).

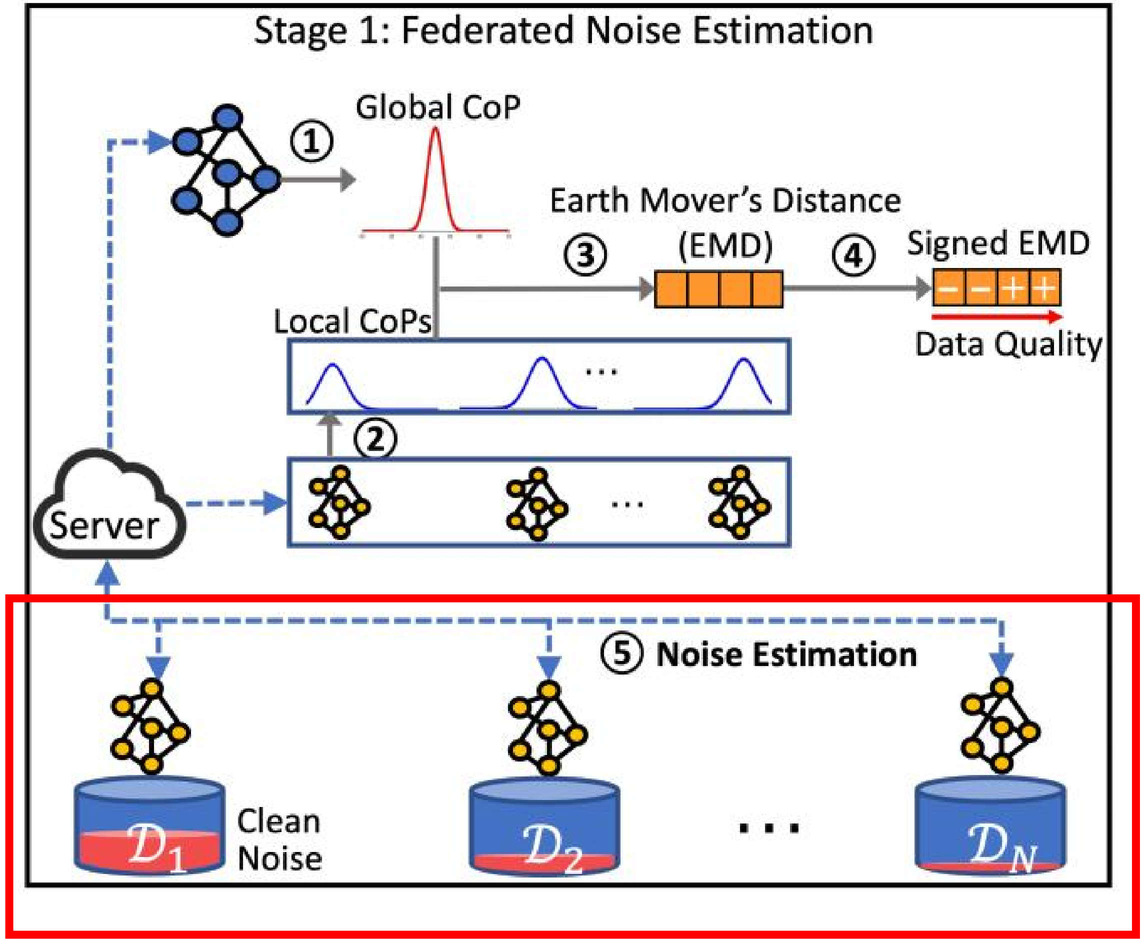
Signed EMD

$$G = [\mathbf{g}_1, ..., \mathbf{g}_N, \mathbf{g}_s] \xrightarrow{\text{GMM}} [\boldsymbol{\mu}_1, ...\boldsymbol{\mu}_N, \boldsymbol{\mu}_{N+1}]$$

$$\begin{aligned} d_n &= \operatorname{sgn}\left(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{N+1}\right) \cdot \operatorname{EMD}\left(\mathbf{g}_n, \mathbf{g}_s\right) \\ &= \operatorname{sgn}\left(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{N+1}\right) \cdot \inf_{\pi \in \Pi(\mathbf{g_n}, \mathbf{g_s})} \mathbb{E}_{(x,y)\sim\pi}[d(x,y)] \end{aligned}$$

$$\operatorname{sgn}(x) = -[x < 0] + [x > 0]$$

Method



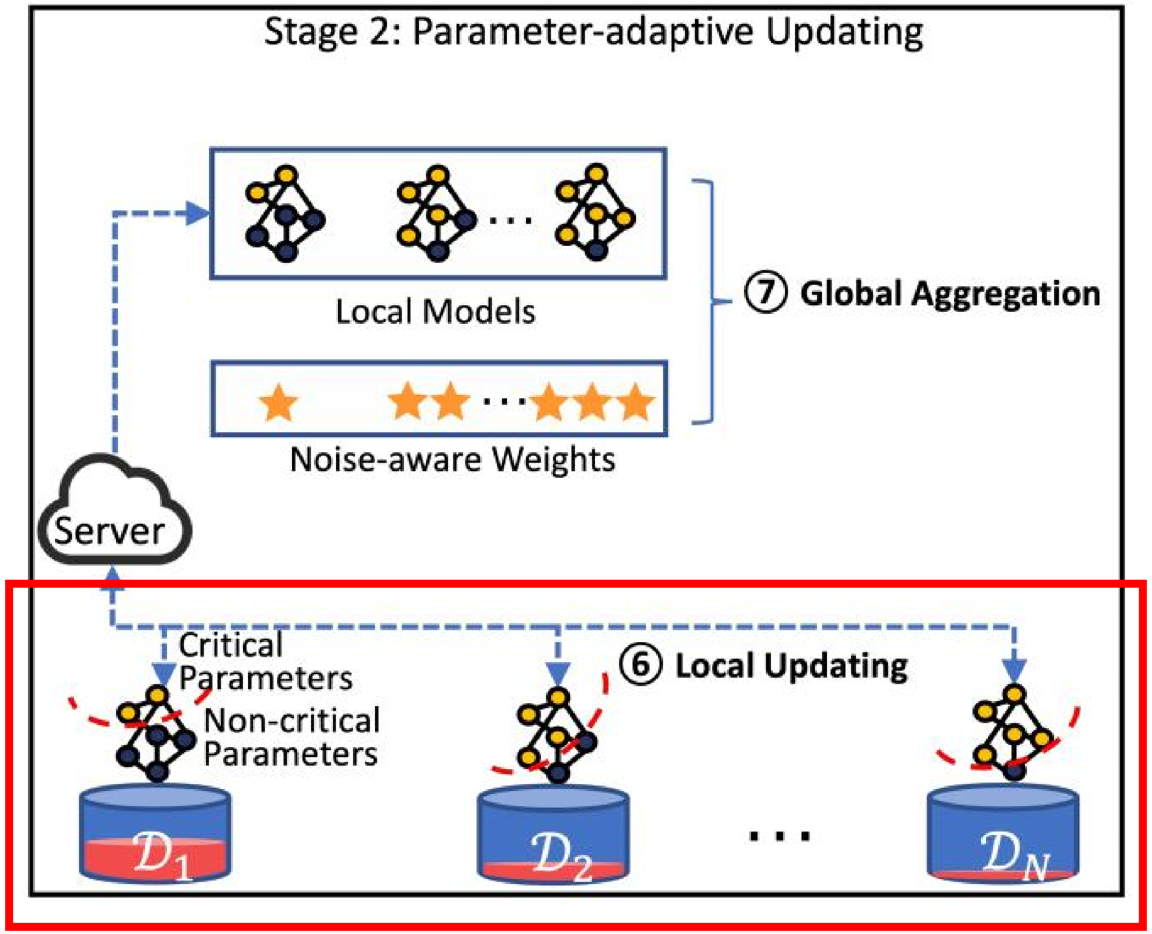
Data quality

$$\tau_n = \frac{|y_n^i \neq y_n^{*i}|}{|\mathbb{D}_n|}, i \in [1, |\mathbb{D}_n|]$$

$$q_n = 1 - \tau_n$$

$$\rho_n = \frac{d_n - \min(\boldsymbol{d})}{\max(\boldsymbol{d}) - \min(\boldsymbol{d})}$$

Method



Parameter-adaptive Updating

$$m_n^c = \rho_n * m$$

$$\mathbf{g}_n^{\downarrow} = \left[g_n^{\downarrow}[1], \ldots, g_n^{\downarrow}[m_n^c], \ldots, g_n^{\downarrow}[m]\right],$$
$$g_n^{\downarrow}[1] \geq \ldots \geq g_n^{\downarrow}[m_n^c] \geq \ldots \geq g_n^{\downarrow}[m]$$

$$\mathcal{M}_n[i] = \begin{cases} 1, & \text{if } g_n^{\downarrow}[1] \geq g[i] \geq g_n^{\downarrow}[m_n^c] \\ 0, & \text{otherwise} \end{cases}$$

Selective gradient decay (SeGD)

$$\mathcal{W}_n(t'+1) \leftarrow \mathcal{W}_n(t') - \eta\rho_n\mathcal{M}_n \odot \nabla L(\mathcal{W}_n(t'))$$

Noise-aware aggregation (NaAgg)

$$\mathcal{W}(t+1) = \sum_{n=1}^{N} \frac{|\mathbb{D}_n|\,\rho_n}{\sum_k |\mathbb{D}_k|\,\rho_k} \mathcal{W}_n(t+1)$$

Experiments

$$\tau_n = \min(\max(\tau, 0), 1), \tau \sim \mathcal{N}(\mu, \sigma)$$

		IID				Non-IID			
		Symmetric		Asymmetric		Symmetric		Asymmetric	
Category	Method	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$
Baseline	FedAvg	78.32±0.36	55.59±0.72	81.62±0.32	50.28±0.03	58.75±0.06	32.56±0.82	63.06±0.66	32.52±0.82
Binary De-noise	S-FedAvg	85.42±0.28	63.72±0.95	88.94±0.62	58.82±0.04	66.55±0.17	41.27±0.53	70.62±0.52	40.72±0.98
	Fair	83.56±0.08	64.35±0.83	87.60±0.22	58.35±0.76	64.04±0.58	41.27±0.18	68.32±0.97	40.64±0.97
	FedNoRo	87.55±0.29	71.00±0.11	83.79±0.14	48.16±0.38	59.36±0.61	35.36±0.27	53.97±0.76	47.62±0.68
General De-noise	Fed-SCE	90.19±0.21	83.00±0.34	84.77±0.10	52.50±0.67	83.66±0.38	65.33±0.56	70.92±0.04	23.63±0.30
	Fed-Mixup	88.72±0.15	74.19±0.69	87.77±0.20	54.61±0.52	70.72±0.48	40.07±0.15	66.71±0.17	31.56±0.83
	Fed-Coteaching	85.38±0.17	73.67±0.20	87.15±0.09	58.20±0.59	76.64±0.73	54.77±0.12	72.25±0.78	22.26±0.71
Ours	**FedES**	**93.09±0.93**	**85.40±0.34**	**90.79±0.91**	**60.34±0.36**	**85.74±0.99**	**68.11±0.48**	**74.54±0.65**	**50.59±0.41**

Table 1: Test Accuracy (%) comparison results on CIFAR-10 datasets under varying synthetic federated label noise

		IID				Non-IID			
		Symmetric		Asymmetric		Symmetric		Asymmetric	
Category	Method	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.3$	$\mu = 0.5$
Baseline	FedAvg	46.22±0.60	30.94±0.87	53.01±0.80	31.66±0.63	42.11±0.36	25.84±0.12	51.72±0.83	33.04±0.01
Binary De-noise	S-FedAvg	53.29±0.00	39.78±0.44	60.33±0.88	40.56±0.53	49.60±0.45	34.22±0.24	58.74±0.06	41.07±0.36
	Fair	51.71±0.19	39.92±0.64	58.54±0.43	39.83±0.05	47.11±0.72	34.44±0.04	56.99±0.79	41.45±0.91
	FedNoRo	59.76±0.38	47.14±0.40	61.13±0.13	33.22±0.75	42.73±0.64	30.40±0.15	50.43±0.06	44.97±0.29
General De-noise	Fed-SCE	57.83±0.51	48.01±0.74	58.05±0.36	33.01±0.43	63.17±0.27	50.20±0.45	57.36±0.11	34.63±0.23
	Fed-Mixup	60.14±0.73	47.05±0.56	62.16±0.59	37.08±0.24	55.86±0.12	40.86±0.18	58.27±0.42	37.57±0.29
	Fed-Coteaching	59.22±0.45	44.27±0.33	58.98±0.50	34.64±0.98	58.45±0.02	42.72±0.43	60.59±0.35	39.03±0.16
Ours	**FedES**	**63.13±0.32**	**50.59±0.68**	**65.11±0.09**	**39.58±0.37**	**65.51±0.75**	**52.96±0.76**	**62.72±0.89**	**47.05±0.11**

Table 2: Test Accuracy (%) comparison results on CIFAR-100 datasets under varying synthetic federated label noise

Experiments

Baseline	Binary De-noise			General De-noise			Ours
FedAvg	S-FedAvg	Fair	FedNoRo	Fed-Mixup	Fed-Coteaching	Fed-SCE	Fed-ES
70.52±0.23	71.33±0.04	71.25±0.50	71.05±0.14	72.61±0.27	71.35±0.23	72.57±0.12	73.03±0.14

Table 3: Test Accuracy (%) comparison results on Clothing1M datasets under real-world label noise

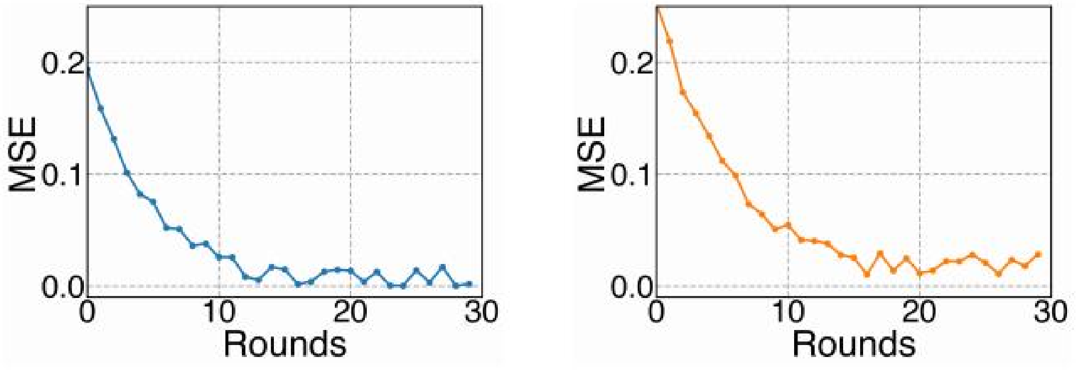
Indicator	Mean	EMD	Sign	CIFAR-10	CIFAR-100
$\hat{q}_n$	✗	✗	✗	0.07	0.13
$P_\phi[n]$	✗	✗	✗	0.05	0.11
ρ_n	✓	✗	✗	0.03	0.09
ρ_n	✗	✓	✗	0.02	0.05
ρ_n	✗	✓	✓	**0.01**	**0.02**

Table 4: MSE comparison results of the first stage ablation study in FedES. Settings: CIFAR-10 dataset ($\mu = 0.5$, noise type: asymmetric, data partition: Non-IID) and CIFAR-100 dataset ($\mu = 0.5$ noise type: asymmetric, data partition: Non-IID)

CS	SeGD	NaAgg	CIFAR-10	CIFAR-100
✗	✗	✗	58.75±0.06	53.01±0.80
✓	✗	✗	67.91±0.15	59.97±0.29
✗	✓	✗	76.15±0.82	62.76±0.64
✗	✗	✓	74.26±0.97	61.17±0.18
✗	✓	✓	**85.74±0.99**	**65.11±0.09**

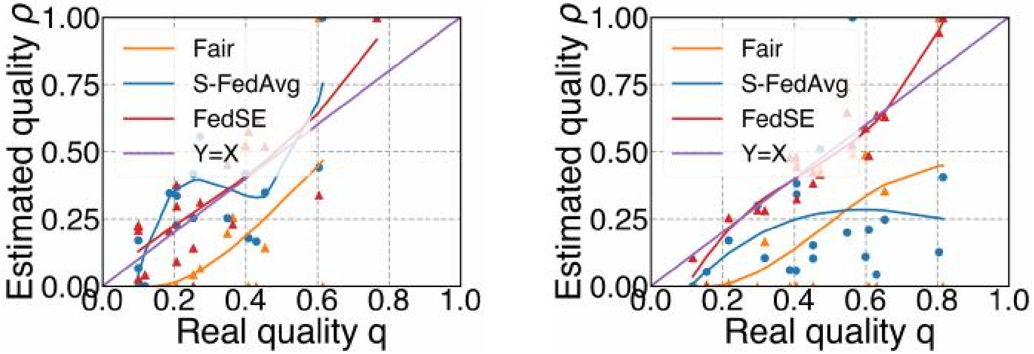
Table 5: Test Accuracy comparison results of the second stage ablation study in FedES. Settings: CIFAR-10 dataset ($\mu = 0.3$, noise type: symmetric, data partition: Non-IID) and CIFAR-100 dataset ($\mu = 0.3$, noise type: asymmetric, data partition: IID)

Experiments



(a) CIFAR-10 (b) CIFAR-100

Figure 3: Ablation study of s_1 for pre-training. Settings: CIFAR-10 dataset ($\mu = 0.5$, noise type: asymmetric, data partition: Non-IID) and CIFAR-100 dataset ($\mu = 0.3$, noise type: asymmetric, data partition: Non-IID)



(a) CIFAR-10 (b) CIFAR-100

Figure 2: Comparison on data quality estimation. Settings: CIFAR-10 dataset ($\mu = 0.5$, noise type: asymmetric, data partition: Non-IID) and CIFAR-100 dataset ($\mu = 0.5$ noise type: asymmetric, data partition: Non-IID)

Thanks