南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Federated Noisy Label Learning

Reporter: Tong Jin

Noisy Label

Most existing studies assume that training data is **labeled correctly**.

Building a completely clean dataset with high-quality annotation is costly in realistic medical scenarios, as labeling medical data is **time-consuming** and **labor-intensive** requiring expertise.

It would unavoidably introduce noisy labels when hiring **non-professionals to label** or using **automatic labeling techniques**.

Directly learning with such noisy labels

↓

Degrading the generalization performance

Centralized Noisy Label Learning

- **Classification**:

  (1) **Class-conditional noise**: each instance from one class has a fixed probability of being assigned to another.

  (2) **Instance-dependent noise**: a data sample is more likely to be mislabeled due to its content rather than the class label it belongs to.

- **Method**:

  (1) **Loss correction**: aims to correct the loss by estimating the noise transition matrix, adjusting the example labels or weights.

  (2) **Example selection**: separate clean examples from noisy ones based on the small-loss criterion and further consider recognized mislabeled examples as unlabeled ones to perform semi-supervised learning.

Limitations

- **Data Privacy and Inaccessibility**:

  In FL, the data is scattered in clients, and the global data cannot be **centrally accessed**. Noise processing methods that rely on global information cannot be implemented.

- **Non-IID**

  There are **large differences in the distribution** of client data. Traditional methods assume that the data is IID, which leads to the failure of noise detection and correction.

- **Client capability differences**

  The **computing resources** and **storage capacity** of clients are uneven. Complex local noise processing, such as generative adversarial network denoising, may exceed some client load capabilities.

- **Differences in local noise levels**

  The **difference in the proportion or type of noise** between different clients is significant. The global unified noise processing strategy cannot adapt to all clients.

Federated Noisy Label Learning

- **Classification：**
  (1) **Class-conditional noise**

  (2) **Instance-dependent noise**

- **Method：**
  (1) **Loss correction**

  (2) **Example/Client selection**

- **Setting：**

  (1) **Some clients are clean while others are not**

  (2) **Each client has partially noisy data**

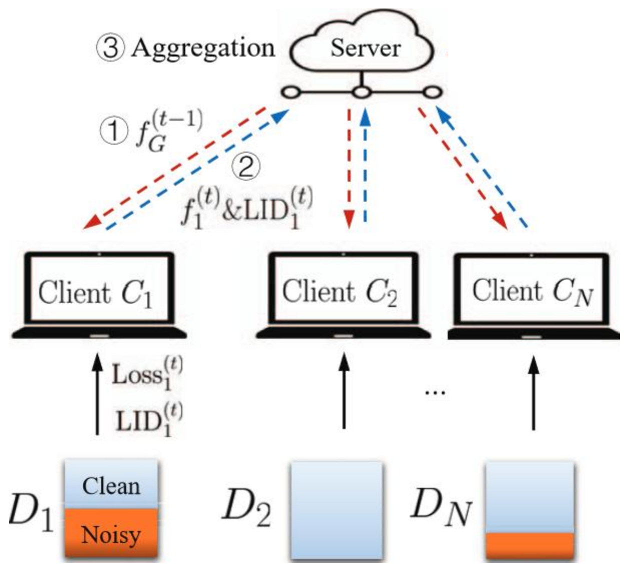
FedNoRo: Towards Noise-Robust Federated Learning by Addressing Class Imbalance and Label Noise Heterogeneity

Nannan Wu[1], **Li Yu**[1], **Xuefeng Jiang**[2], **Kwang-Ting Cheng**[3] and **Zengqiang Yan**[1*]
[1]School of Electronic Information and Communications, Huazhong University of Science and Technology
[2]Institute of Computing Technology, Chinese Academy of Sciences
[3]School of Engineering, Hong Kong University of Science and Technology
{wnn2000, hustlyu}@hust.edu.cn, jiangxuefeng21b@ict.ac.cn, timcheng@ust.hk, z_yan@hust.edu.cn

IJCAI 2023

Challenges

Existing methods for noisy client detection propose to calculate an **average indicator** (e.g. loss) over all samples of each client as its feature and filter out the clients with abnormal features as noisy clients.



Less Effective in Real-World FL

- **Data is highly class-imbalanced from the global perspective.**
- **Data heterogeneous across clients affecting the calculation of indicators.**
- **Label noise is heterogeneous across clients in both strength (different noise rates) and pattern (various forms of label noise).**

Eg. 1

Cancer-specialized hospital A (more malignant cases)
General hospital B (more healthy cases)

A is more likely to produce an abnormal client-wise feature (e.g., large loss values similar to noisy clients) due to class imbalance (i.e., healthy >> malignant).

Eg. 2

Hospital C (more healthy cases)   benign → health
Hospital D (more malignant cases)   benign → malignant

Though both labels are wrong, the loss values (produced by C would be much smaller than D, due to class imbalance (i.e., healthy>>malignant).

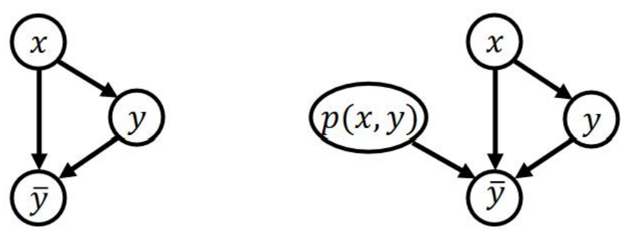
Noise Model

Define the global noise rate ρ as the proportion of noisy clients.

Local noise rate η_i (i.e., the proportion of noisy samples) follows the uniform distribution.

The pattern of noise samples: heterogeneous instance-dependent noise (H-IDN).

Definition 1 (H-IDN). *IDN is heterogeneous if noise transition probability is a function of local data distribution, i.e.,* $\Pr(\overline{Y} = \overline{y} \mid Y = y, X = x) = M_{\overline{y},y,x}(p_i(x,y))$, *where* $M_{\overline{y},y,x}$ *denotes the noise transition matrix of instance* x.



(a) IDN (b) H-IDN

Noise Generation

Algorithm 1 Noise Generation.

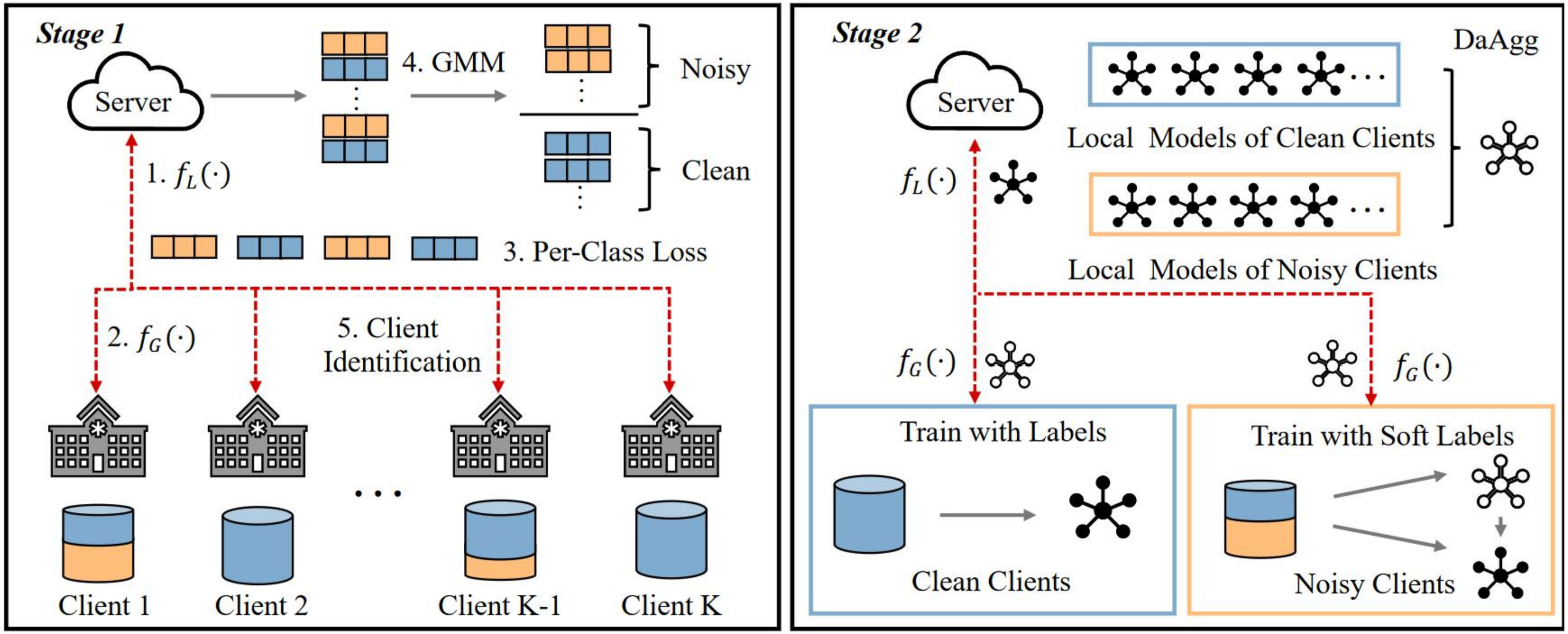
Input: Number of clients K; clean local datasets $\{D_k\}_{k=1}^{K}$; global noise rate ρ; local noise rate distribution parameters η^l, η^u.

```
I = Randomly select ρK elements from [K].
for i in I do
    Initialize a network f_i.
    Train f_i on the local dataset D_i = {(x_j, y_j)}_{j=1}^{N_i}.
    Compute classification probabilities p(Y | x) ∈ [0,1]^{N_i×C} for all samples in D_i.
    Compute the misclassification probability p̃(x) ∈ [0,1]^{N_i} for each sample in D_i (Eq. 1).
    η_i ~ U(η^l, η^u).
    N = Randomly select η_i N_i elements from [N_i] with
    the probability p̃(x)/∑p̃(x).                ▷ Normalization
    for t in N do
        ȳ_t = Randomly select a different label from Y with
        the probability p(Y | x_t).
        Flip y_t to ȳ_t.                          ▷ Add label noise
    end for
end for
```

Output: Local datasets after adding label noise $\{D_k\}_{k=1}^{K}$.

$$\widetilde{p}(x_j) = 1 - p(Y = y_j \mid x_j),$$

Methodology



Overview of the proposed two-stage framework FedNoRo.

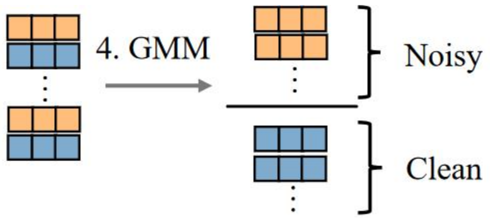
Stage 1: Noisy Client Detection

Train a warmup model for T1 rounds by FedAvg.

The average loss values of all classes on each client i denoted as $l_i = (l_i^1, l_i^2, ..., l_i^C)^T \in \mathbb{R}^C$

Considering the class-missing problem in heterogeneous data, a specific class c may not exist in client i, leading to a missing average loss value, which simply replaced by the minimum value of class c across all clients.

Normalized to [0, 1] : $l_i^c = \dfrac{l_i^c - \min_i l_i^c}{\max_i l_i^c - \min_i l_i^c}.$



Stage 2: Noise-Robust Training

- **Local training phase**

For clean clients, the vanilla **cross-entropy loss** is adopted to train each local model based on clean labels.

For noisy clients, a **knowledge distillation-based** training method is applied. Given any x and its output logit from the global model, a targeted probability distribution is calculated as

$$y_G = \text{softmax}(\frac{f_G(x)}{T}),$$

$$\mathcal{L} = \lambda \mathcal{L}_{KL}(y_p, y_G) + (1-\lambda)\mathcal{L}_{CE}(y_p, \overline{y}), \quad \lambda \text{ grows from 0 to } \lambda_{max}$$

- **Model aggregation phase**

Distance-aware model aggregation is proposed, where a client-wise distance metric is：

$$d(i) = \min_{j \in S_c} \|w_i - w_j\|_2,$$

Normalized to [0, 1] : $D(i) = \frac{d(i)}{\max_j d(j)}.$

Aggregation weight: $w_g = \sum_{i=1}^{K} \frac{N_i e^{-D(i)}}{\sum_{j=1}^{K} N_j e^{-D(j)}} w_i.$

Experimental Results

Performance Results

Category	Methods	ρ	0.0	0.2		0.3		0.4		0.6	
		(η^l, η^u)	$(0.0, 0.0)$	$(0.3, 0.5)$	$(0.5, 0.7)$	$(0.3, 0.5)$	$(0.5, 0.7)$	$(0.3, 0.5)$	$(0.5, 0.7)$	$(0.3, 0.5)$	$(0.5, 0.7)$
FL	FedAvg	Best	69.34	65.81	64.53	62.49	60.82	60.52	58.38	58.46	54.77
		Last	68.92	65.33	63.97	62.10	60.23	60.05	56.79	56.88	50.35
	FedProx	Best	68.16	64.58	63.64	61.81	61.91	60.85	58.50	60.21	57.57
		Last	67.29	63.80	62.65	61.46	61.00	59.99	57.85	59.35	56.63
	FedLA	Best	**73.56**	69.45	69.28	66.84	64.84	66.60	62.39	63.90	58.78
		Last	**73.07**	68.82	68.45	66.20	64.03	63.90	60.51	61.86	54.99
Denoise FL	RoFL	Best	-	42.57	40.42	40.64	39.87	40.35	35.60	39.68	35.35
		Last	-	42.19	40.30	40.95	39.27	40.25	35.39	40.18	35.47
	RHFL	Best	-	57.48	56.91	56.72	55.74	55.26	54.30	54.63	51.00
		Last	-	57.05	56.46	55.75	52.19	54.71	52.08	52.53	49.64
	FedLSR	Best	-	55.99	55.28	54.44	52.27	52.48	48.20	50.89	43.30
		Last	-	55.69	54.77	53.95	51.70	51.92	46.77	49.96	39.81
	FedCorr	Best	-	57.90	56.68	55.02	54.61	53.62	50.82	50.89	47.13
		Last	-	57.68	55.86	54.52	53.62	52.60	49.59	50.24	46.13
Joint	FedCorr+LA	Best	-	64.78	64.29	62.58	62.78	62.19	59.58	57.88	54.53
		Last	-	63.99	63.55	61.90	61.87	61.30	60.06	57.17	53.92
Ours	FedNoRo	Best	-	**70.59**	**70.64**	**70.14**	**69.35**	**70.69**	**69.30**	**67.55**	**63.83**
		Last	-	**70.18**	**69.81**	**69.29**	**68.47**	**70.14**	**68.87**	**67.10**	**63.29**

Table 1: Quantitative BACC (%) comparison results on the ICH dataset under different noise rates. The best results are marked in bold.

Methods	ρ	0.0	0.4	0.6
	(η^l, η^u)	$(0.0, 0.0)$	$(0.5, 0.7)$	$(0.5, 0.7)$
FedAvg	Best	69.09	58.23	54.35
	Last	65.44	61.83	49.88
FedProx	Best	72.20	64.46	58.55
	Last	**69.60**	60.13	49.87
FedLA	Best	**72.55**	66.34	61.20
	Last	68.72	61.18	56.11
RoFL	Best	-	28.45	28.86
	Last	-	27.79	28.29
RHFL	Best	-	46.06	46.67
	Last	-	44.04	45.09
FedLSR	Best	-	30.15	27.24
	Last	-	29.11	26.08
FedCorr	Best	-	42.54	38.40
	Last	-	41.12	37.17
FedCorr+LA	Best	-	60.38	55.40
	Last	-	59.16	54.27
FedNoRo (ours)	Best	-	**68.59**	**66.00**
	Last	-	**64.67**	**60.65**

Table 2: Quantitative BACC (%) results on the ISIC 2019 dataset under different noise rates. The best results are marked in bold.

Ablation Studies

Indicator	LA	*Per-Class*	*Norm.*	Re (%)	Pr (%)	MR (%)
ICH, $\rho = 0.3$, $(\eta^l, \eta^u) = (0.3, 0.5)$						
LID	✗	✗	✗	**100.00**	54.54	0.00
Conf.	✗	✗	✗	16.66	8.33	0.00
Loss	✗	✗	✗	94.58	57.12	0.00
Loss	✓	✗	✗	**100.00**	47.14	0.00
Loss	✗	✓	✗	97.19	90.63	85.77
Loss	✓	✓	✗	99.48	97.78	96.93
Loss	✓	✓	✓	99.70	**98.76**	**98.28**
ICH, $\rho = 0.4$, $(\eta^l, \eta^u) = (0.3, 0.5)$						
LID	✗	✗	✗	87.50	63.63	0.00
Conf.	✗	✗	✗	37.50	25.00	0.00
Loss	✗	✗	✗	78.80	60.57	0.00
Loss	✓	✗	✗	89.41	80.46	0.00
Loss	✗	✓	✗	65.77	**100.00**	37.58
Loss	✓	✓	✗	81.10	**100.00**	47.09
Loss	✓	✓	✓	**90.23**	**100.00**	**88.82**

Table 3: Ablation study of the first stage in FedNoRo.

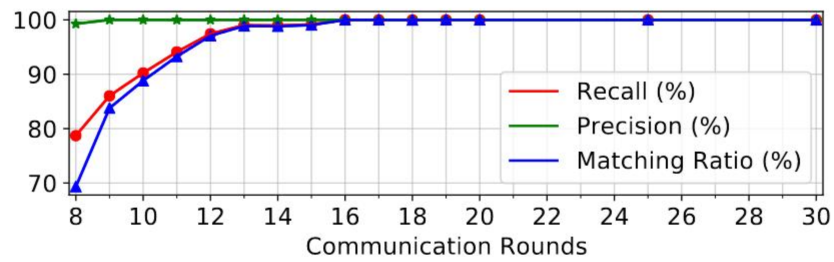


Figure 4: Ablation study of T_1 for warm-up training.

Noisy Clients	De-Noise Strategy		Type	BACC (%)	
	$\mathcal{L}_{KL}$	DaAgg		ICH	ISIC
✓	✗	✗	Best	66.60	61.20
			Last	63.90	56.11
✗	✗	✗	Best	69.67	64.94
			Last	68.07	59.17
✓	✗	✓	Best	69.32	65.72
			Last	68.18	58.60
✓	✓	✗	Best	64.53	60.79
			Last	64.01	43.57
✓	✓	✓	Best	**70.69**	**66.00**
			Last	**70.14**	**60.65**

Table 4: Ablation study of the second stage in FedNoRo. Settings: ρ=0.4 and (η^l, η^u)=(0.3,0.5) for the ICH dataset; ρ=0.6, (η^l, η^u)=(0.5,0.7) for the ISIC dataset.

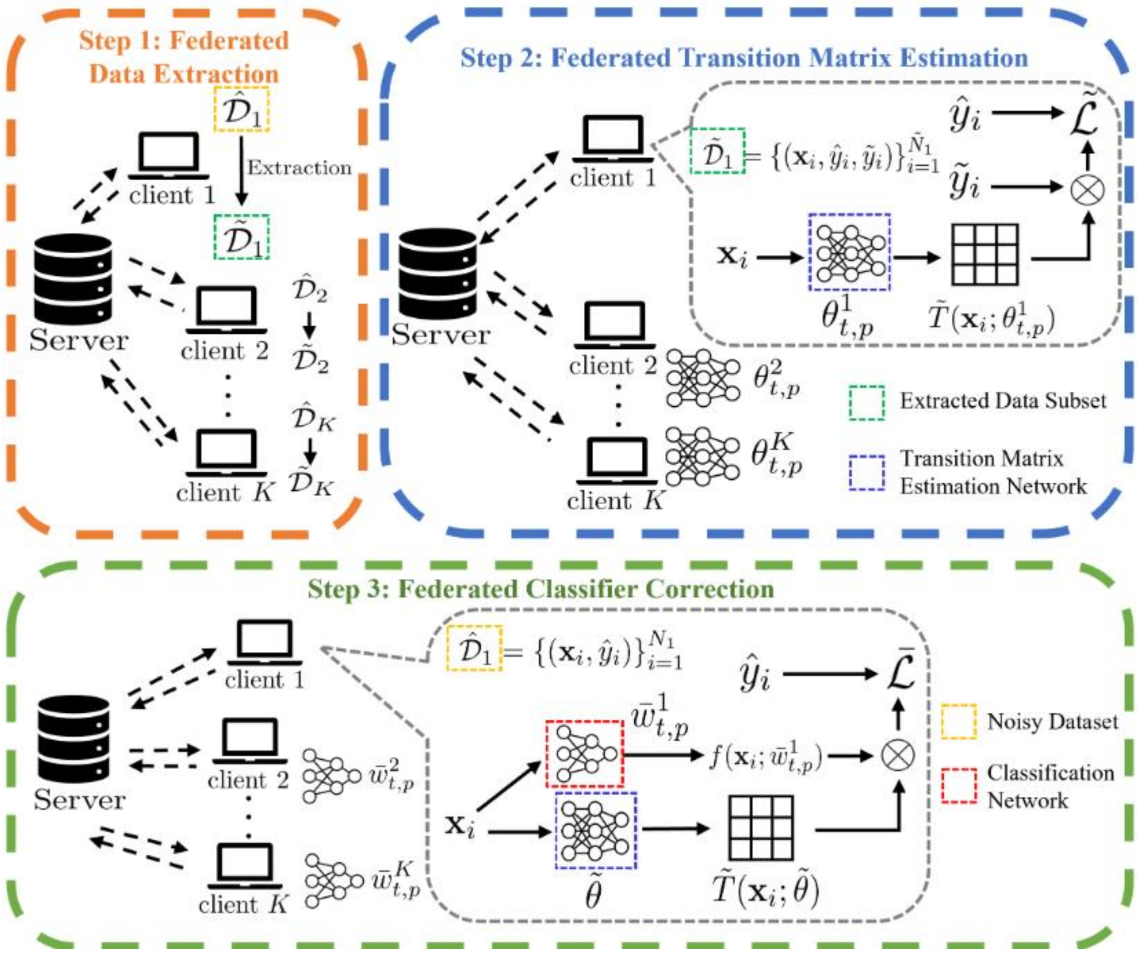
FEDERATED LEARNING WITH INSTANCE-DEPENDENT NOISY LABEL

Lei Wang, Jieming Bian, Jie Xu

Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146, USA

ICASSP 2024

Methodology



An overview of FedBeat

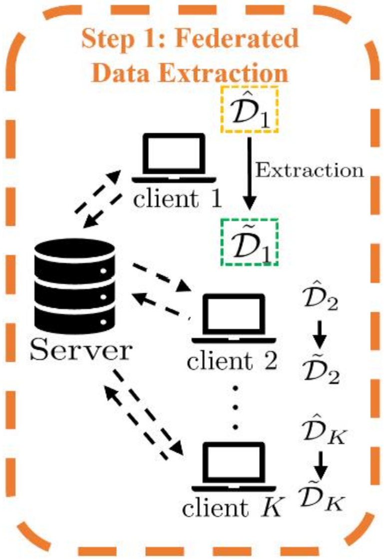
Noise Model

We define $T_{i,j}(x)$ as the i, j-th element of instance-dependent noise transition matrix (IDNTM), representing the transition probability of a sample x with a clean label i transitioning to a noisy label j.

Thus, the noisy class-posterior probability can be inferred by the IDNTM and the clean class-posterior probability as follows:

$$P(\hat{Y} = j | X = \mathbf{x}) = \sum_{i=1}^{C} T_{i,j}(\mathbf{x}) P(Y = i | X = \mathbf{x}).$$

Step1. Federated Data Extraction



Initial training: T1 rounds, and each client performs P1 local steps during each round.

Weak global model: $\bar{w} = \sum_k N_k w^k_{T_1, P_1} / N$

It is returned to the clients to generate pseudo-labels on the training data samples

Bayesian model ensemble:

Denoting the average model $\mu = \bar{w}$, the server then calculates the standard deviation Σ of the local models

$$\Sigma = \operatorname{diag}\left(\sum_k \frac{N_k}{N}\left(w^k_{T_1,P_1} - \mu\right)^2\right).$$

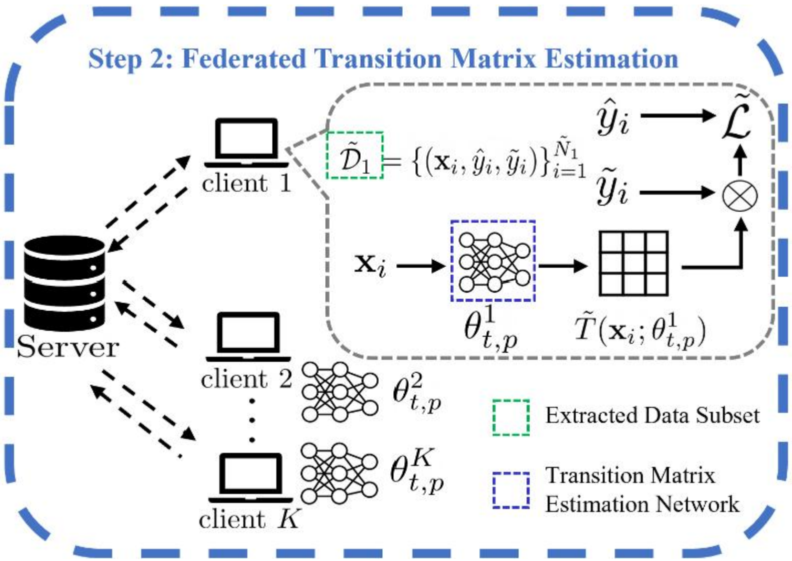
Sample *M* ensemble models $w^{k,(m)} \sim \mathcal{N}(\mu, \Sigma), m = 1, \ldots, M.$

$$\tilde{f}(x_i) = \frac{1}{M}\sum_{m=1}^{M} f\left(\mathbf{x}_i; w^{k,(m)}\right) \qquad \tilde{y}_i = \arg\max_c \tilde{f}_c(\mathbf{x}_i)$$

Extracted dataset: $\tilde{\mathcal{D}}_k = \{(\mathbf{x}_i, \hat{y}_i, \tilde{y}_i); i \in \hat{\mathcal{D}}_k \text{ and } \tilde{f}_{\tilde{y}_i}(\mathbf{x}_i) \geq \tau\}$

Step2. Federated Transition Matrix Estimation

When provided with an input example x, the estimation network θ generates the IDNTM, which contains the probability of transitioning from the clean distribution to the noisy distribution.



T2 rounds training, and each client performs P2 local steps during each round.

$$\theta_{t,0}^k = \theta_t$$

Local training loss function

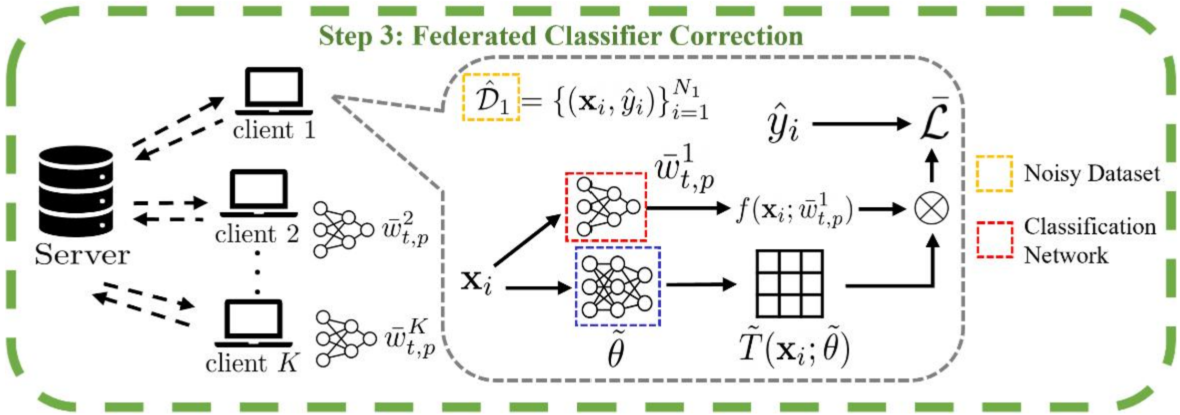
$$\tilde{\mathcal{L}}(\theta_{t,p}^k) = -\frac{1}{\tilde{N}_k}\sum_{i=1}^{\tilde{N}_k}\sum_{c=1}^{C} \hat{\boldsymbol{y}}_{\boldsymbol{i},c} \log\left(\left[\tilde{\boldsymbol{y}}_{\boldsymbol{i}}^{\mathrm{T}} \cdot \tilde{T}\left(\mathbf{x}_i; \theta_{t,p}^k\right)\right]_c^{\mathrm{T}}\right)$$

Global model update

$$\theta_{t+1} = \frac{1}{\tilde{N}}\sum_k \tilde{N}_k \theta_{t,P_2}^k$$

$$P(\hat{Y} = j|X = \mathbf{x}) = \sum_{i=1}^{C} T_{i,j}(\mathbf{x}) P(Y = i|X = \mathbf{x}).$$

Step3. Federated Classifier Correction



Once the estimation network $\tilde{\theta}$ has been trained, it can be employed to correct the weak global classification model $\bar{w}$ derived from step 1.

T3 communication rounds for global model aggregation, with each round consisting of P3 local steps.

Classifier correction loss function:

$$\bar{\mathcal{L}}(\bar{w}_{t,p}^k) = -\frac{1}{N_k} \sum_{i=1}^{N_k} \sum_{c=1}^{C} \hat{\boldsymbol{y}}_{i,c} \log \left(\left[f\left(\mathbf{x}_i; \bar{w}_{t,p}^k\right)^{\mathrm{T}} \tilde{T}\left(\mathbf{x}_i; \tilde{\theta}\right) \right]_c^{\mathrm{T}} \right)$$

$$P(\hat{Y} = j | X = \mathbf{x}) = \sum_{i=1}^{C} T_{i,j}(\mathbf{x}) P(Y = i | X = \mathbf{x}).$$

Performance Results

Table 1. Test accuracy on CIFAR-10 with different IDN rates

	IID		non-IID ($\alpha_{Dir} = 1$)	
	IDN-30%	IDN-50%	IDN-30%	IDN-50%
FedAvg	73.10 ± 0.97	61.99 ± 1.82	61.41 ± 3.26	47.64 ± 1.56
FedProx	71.97 ± 1.16	58.66 ± 0.96	61.57 ± 1.65	47.74 ± 0.95
BLTM-local	45.75 ± 0.55	36.25 ± 0.58	57.64 ± 2.01	49.13 ± 1.27
FedCorr	65.90 ± 1.50	54.41 ± 0.89	62.23 ± 2.34	50.46 ± 2.29
FedBeat(ours)	$\mathbf{81.58 \pm 0.24}$	$\mathbf{74.51 \pm 2.71}$	$\mathbf{72.61 \pm 0.31}$	$\mathbf{58.44 \pm 3.53}$

Table 2. Test accuracy on SVHN with different IDN rates

	IID		non-IID ($\alpha_{Dir} = 1$)	
	IDN-30%	IDN-50%	IDN-30%	IDN-50%
FedAvg	88.35 ± 0.91	74.24 ± 0.24	83.79 ± 0.47	61.86 ± 2.98
FedProx	89.46 ± 0.23	71.81 ± 2.34	84.64 ± 0.21	64.06 ± 1.13
BLTM-local	71.14 ± 1.39	52.26 ± 0.67	70.44 ± 1.34	57.13 ± 2.78
FedCorr	86.18 ± 3.52	70.56 ± 4.63	81.20 ± 1.61	59.37 ± 3.15
FedBeat(ours)	$\mathbf{94.59 \pm 0.25}$	$\mathbf{87.97 \pm 2.90}$	$\mathbf{92.59 \pm 0.40}$	$\mathbf{75.26 \pm 2.89}$

Ablation Studies

Table 3. Test accuracy on SVHN with IDN-30% varying α_{Dir}

	$\alpha_{Dir} = 0.5$	$\alpha_{Dir} = 1$	$\alpha_{Dir} = 5$
FedAvg	81.46 ± 1.66	83.79 ± 0.47	84.98 ± 0.72
FedProx	82.23 ± 1.42	84.64 ± 0.21	84.18 ± 1.01
BLTM-local	73.43 ± 0.93	70.44 ± 1.34	60.08 ± 0.68
FedCorr	76.42 ± 1.88	81.20 ± 1.61	82.61 ± 2.87
FedBeat(ours)	$\mathbf{92.12 \pm 0.49}$	$\mathbf{92.59 \pm 0.40}$	$\mathbf{92.47 \pm 0.31}$

Table 4. Impact of model ensemble

	IDN-30%	IDN-50%
w/o ensemble	92.57% / 1713 / 90.15 ± 0.30	74.24% / **875** / 71.12 ± 0.84
w/ ensemble	**96.01%** / **1736** / $\mathbf{92.59 \pm 0.40}$	**83.75%** / 809 / $\mathbf{75.26 \pm 2.89}$

Table 5. Impact of threshold

	IID		non-IID ($\alpha_{Dir} = 1$)	
	IDN-30%	IDN-50%	IDN-30%	IDN-50%
$\tau = 0.50$	95.49% / 4112	78.81% / 3917	92.28% / 2008	70.05% / 1876
$\tau = 0.65$	97.73% / 3643	90.56% / 1800	96.01% / 1736	83.75% / 809
$\tau = 0.80$	99.21% / 2362	94.77% / 417	97.79% / 1303	85.56% / 250

THANKS