Provably Near-Optimal Federated Ensemble Distillation with Negligible Overhead

Theory-Guided and Efficient Federated Ensemble Distillation Algorithm

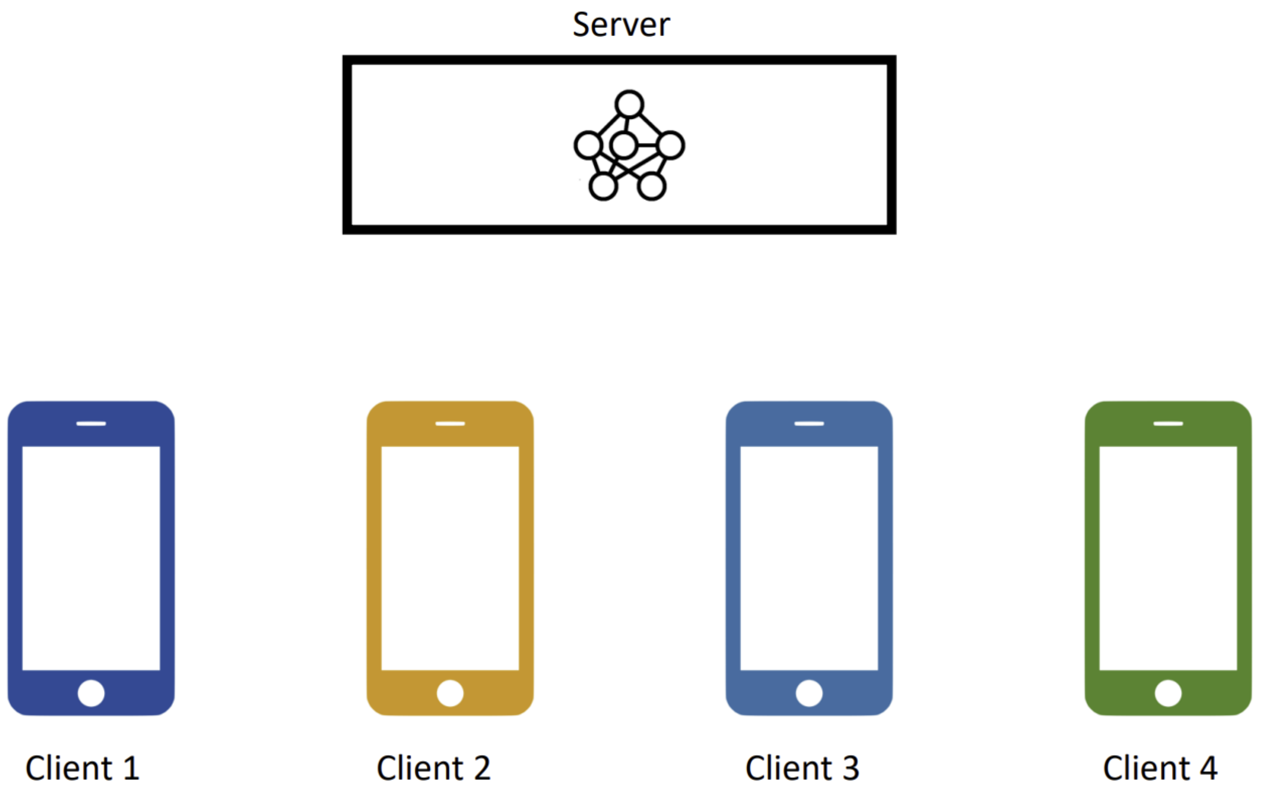
ICML 2025 Poster

KAIST EE InfoLAB

Won-Jun Jang, Hyeon-Seo Park, Si-Hyeon Lee

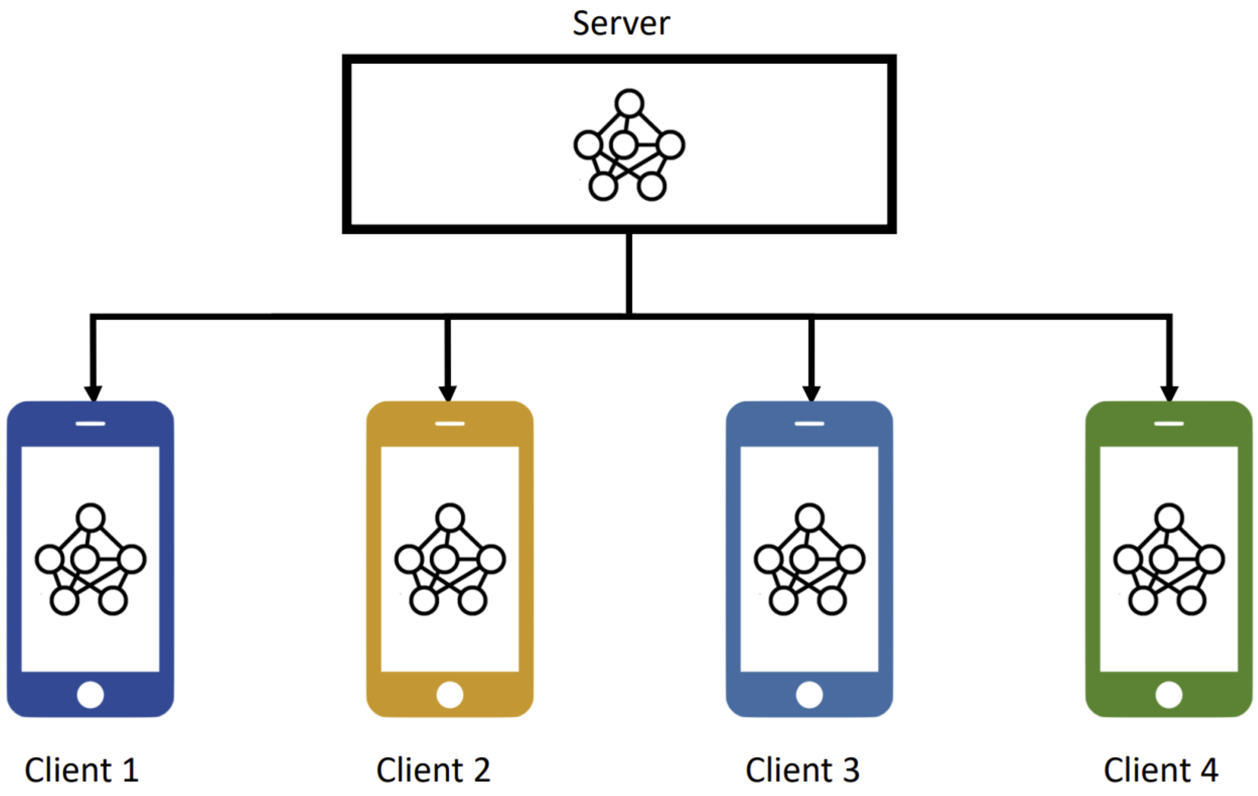
01 Introduction

Federated Learning



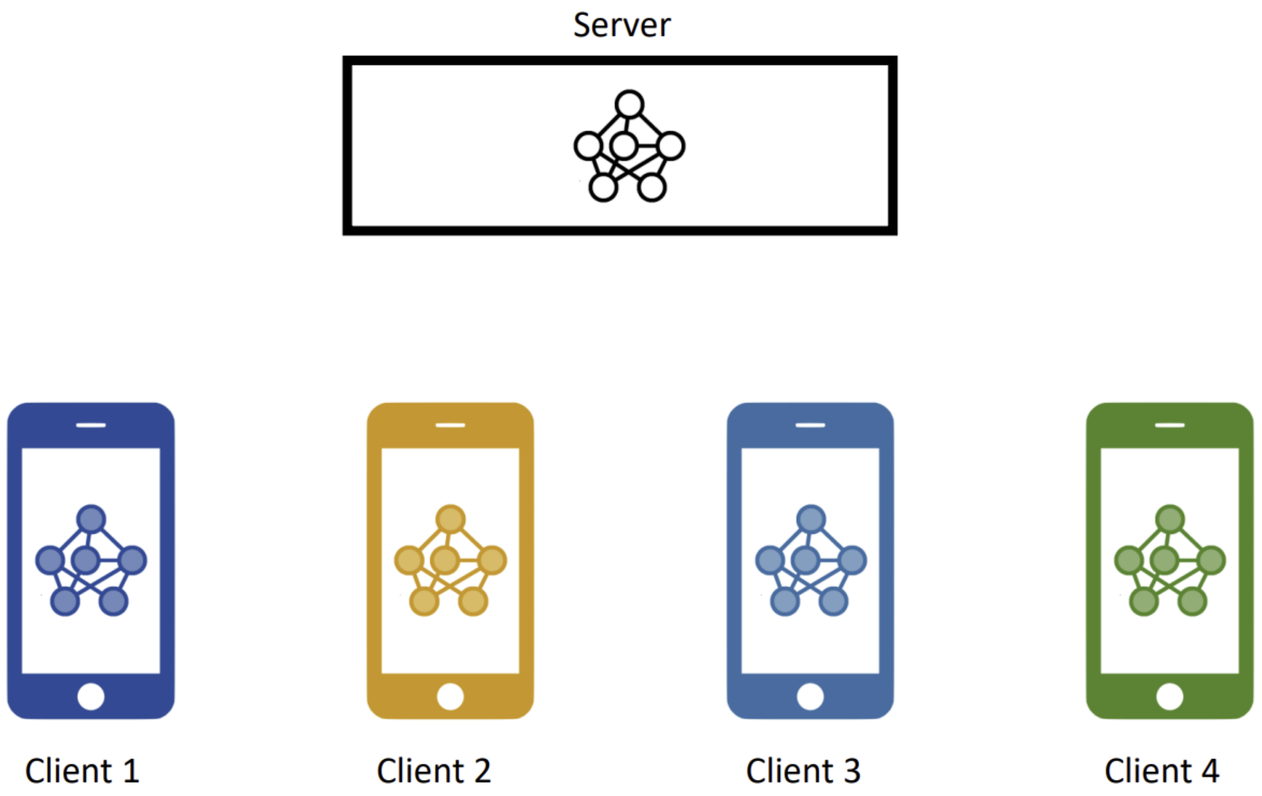
01 Introduction

Federated Learning



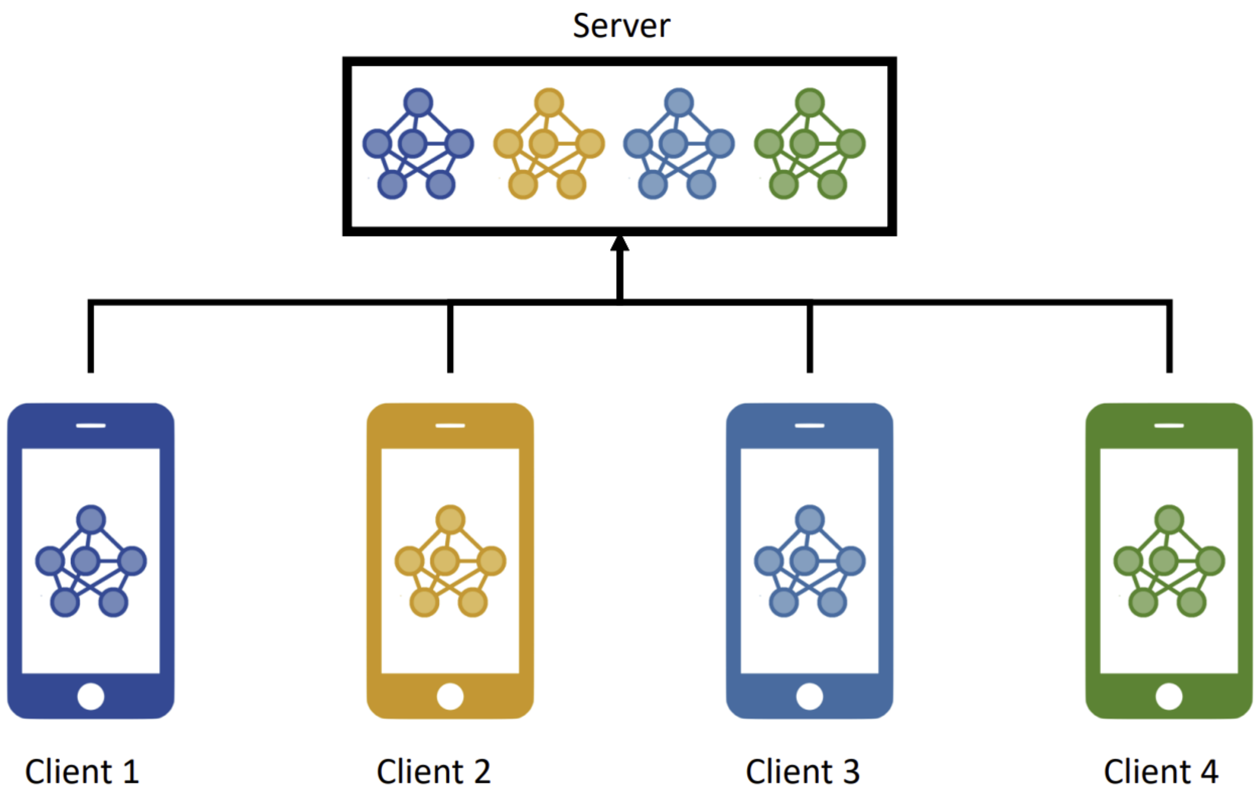
01 Introduction

Federated Learning



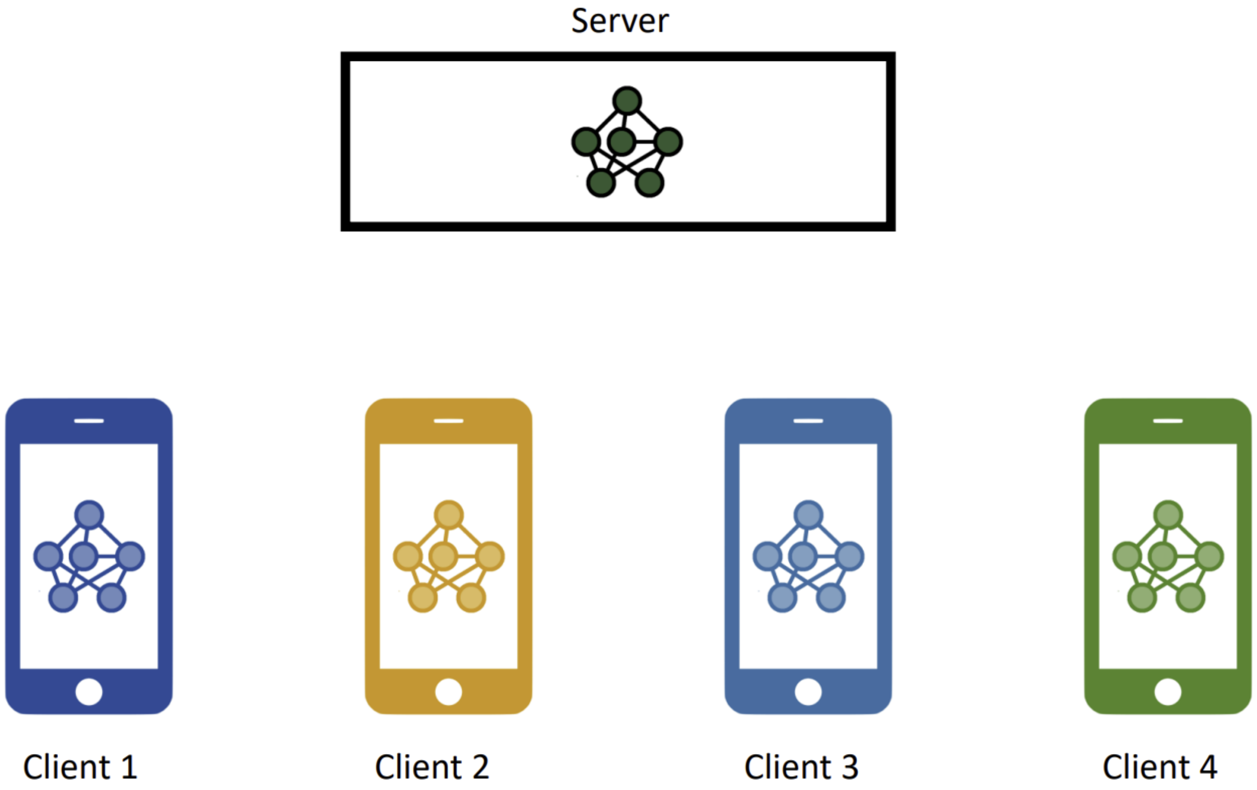
01 Introduction

Federated Learning



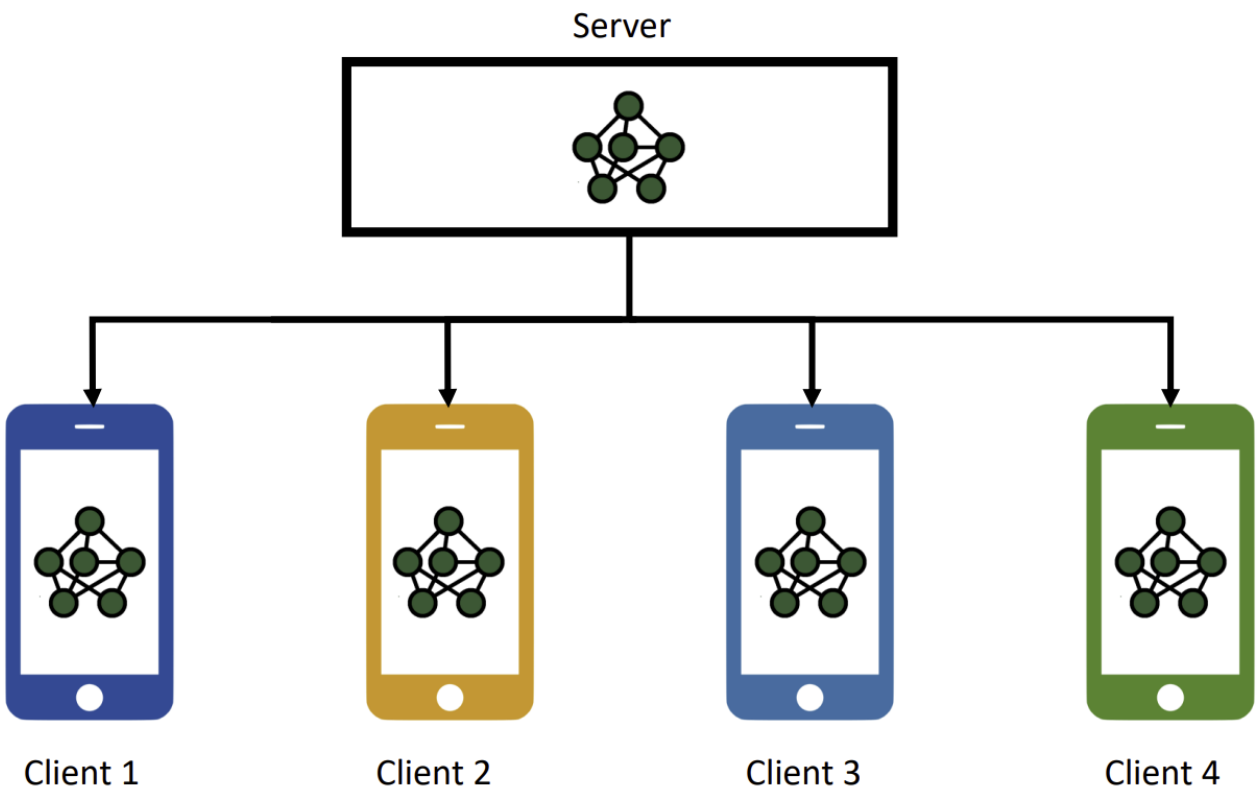
01 Introduction

Federated Learning



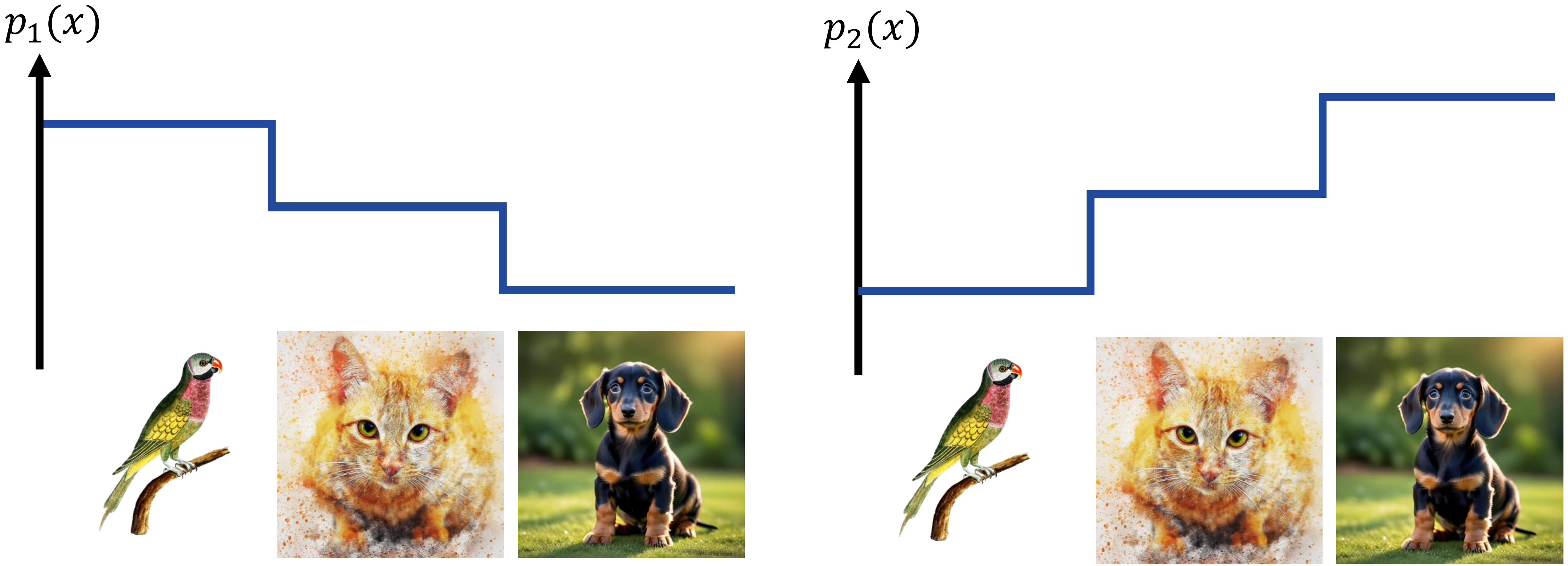
01 Introduction

Federated Learning



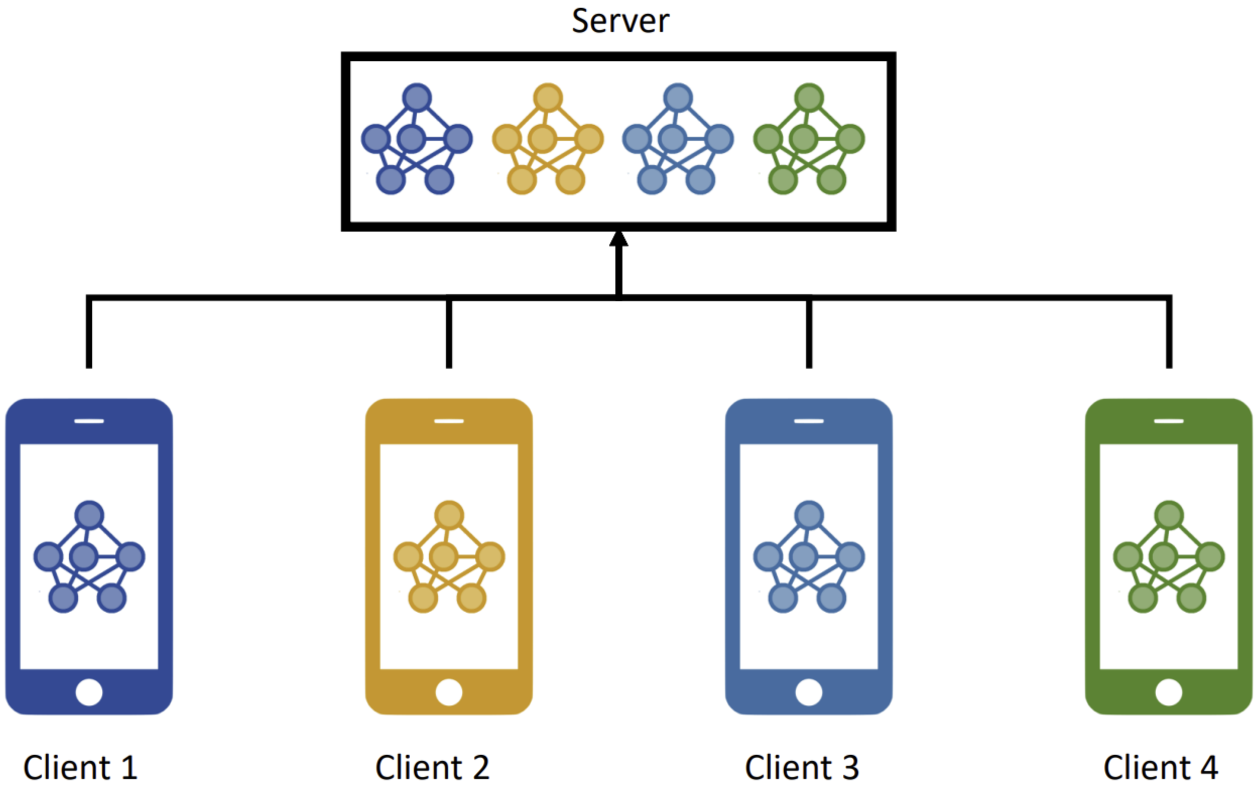
01 Introduction

Problem : Client Data Heterogeneity



01 Introduction

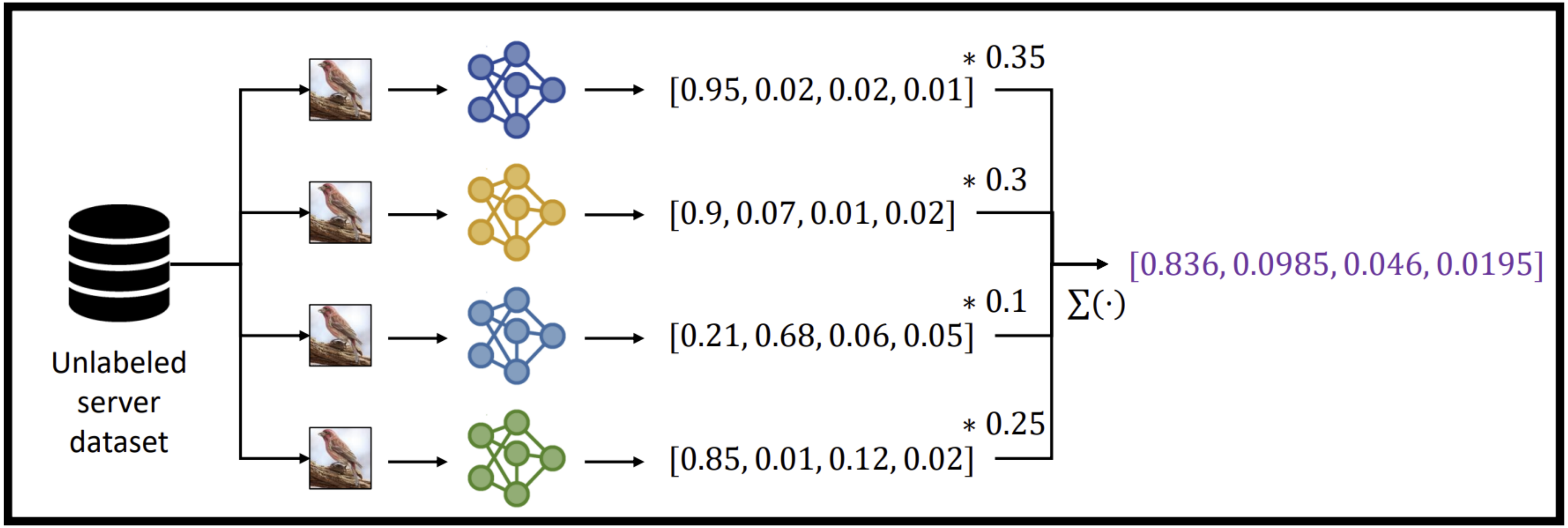
Federated Ensemble Distillation



01 Introduction

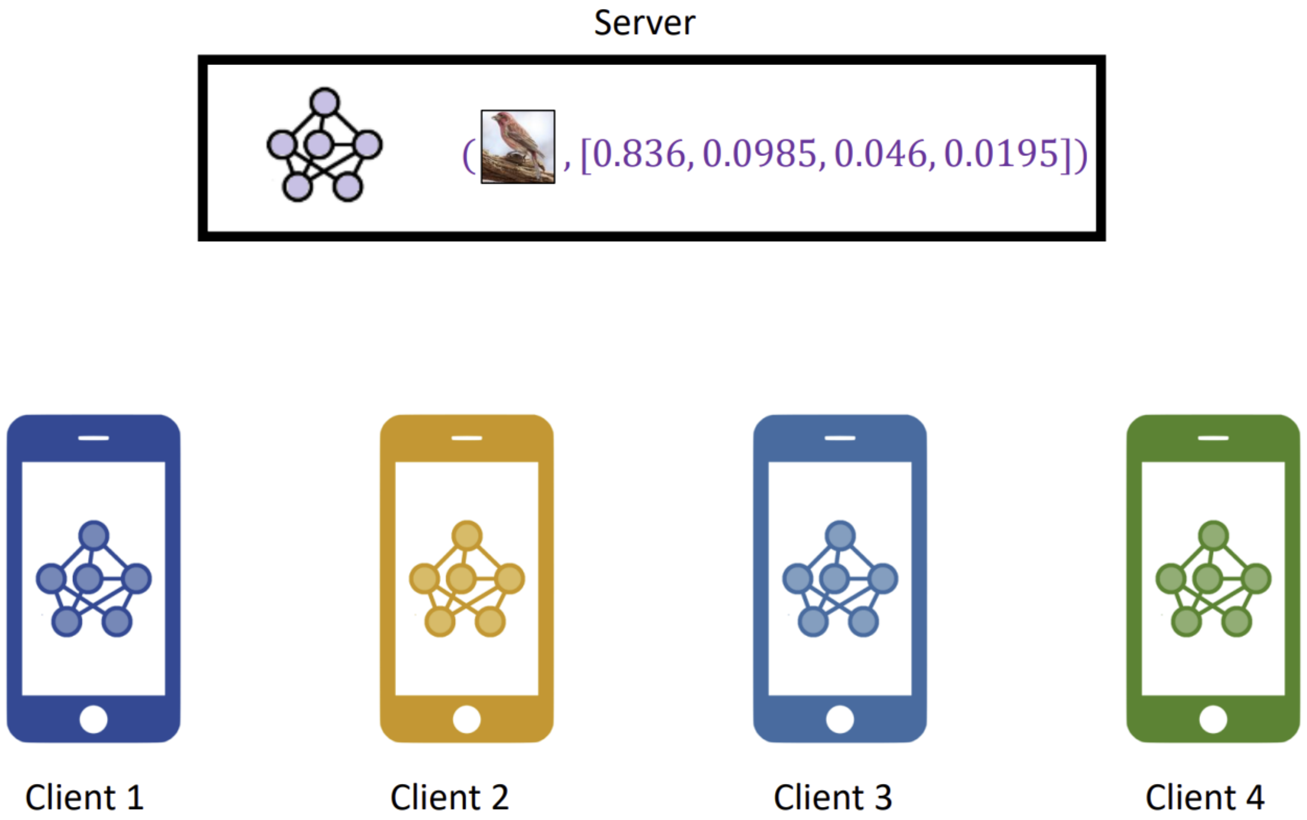
Federated Ensemble Distillation

Server



01 Introduction

Federated Ensemble Distillation



02 Algorithm

Prior Federated Ensemble Algorithms

- Various pseudo-labeling mechanisms are proposed
- Recent weighting mechanisms assign more weight to reliable client
- Our methods provides **the tightest generalization bound for pseudo-label** generated with empirical loss minimizer

Algorithm	Weighting mechanism
FedDF, FedGKD$^+$	Uniform
Fed-ET	$\propto$ variance of output logit
FedHKT, FedDS	$\propto \exp$(entropy of client output softmax)
DaFKD	$\propto$ client discriminator output

02 Algorithm

Theoretical Results

Definition 1. For K clients, the ensemble of their models and weight functions $\{(h_k, w_k)\}_{k=1}^K$ is said to be an optimal model ensemble if the following holds:

$$\mathcal{L}_p\left(\sum_{k=1}^{K} w_k \cdot h_k\right) = \mathbf{E}_p\left[l\left(\sum_{k=1}^{K} w_k(x) \cdot h_k(x), y(x)\right)\right] \leq \min_{h \in \mathcal{H}} \mathcal{L}_p(h) = \mathcal{L}_p(h_p^*). \tag{6}$$

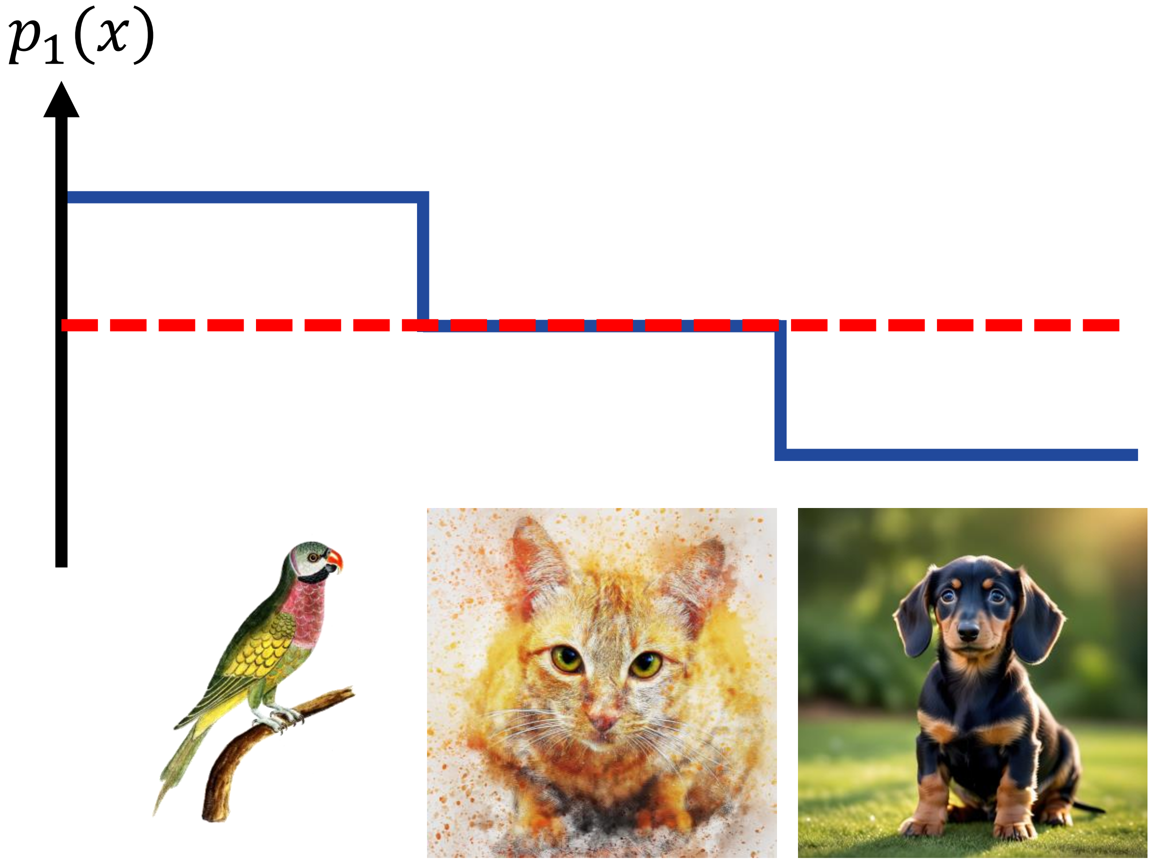
Theorem 3. Let the loss function l be convex. Define the client weight functions $\{w_k^*\}_{k=1}^K$ as follows:

$$w_k^*(x) \triangleq \frac{n_k \cdot p_k(x)}{\sum_{i=1}^K n_i \cdot p_i(x)} = \frac{\pi_k \cdot p_k(x)}{\sum_{i=1}^K \pi_i \cdot p_i(x)}. \tag{10}$$

Then, the ensemble $\{h_{p_k}^*, w_k^*\}_{k=1}^K$ is an optimal model ensemble, i.e., $\mathcal{L}_p\left(\sum_k w_k^* \cdot h_{p_k}^*\right) \leq \mathcal{L}_p(h_p^*)$.

02 Algorithm

Idea



$$p(x) = \frac{1}{2} \cdot \left(p_1(x) + p_2(x)\right)$$

02 Algorithm

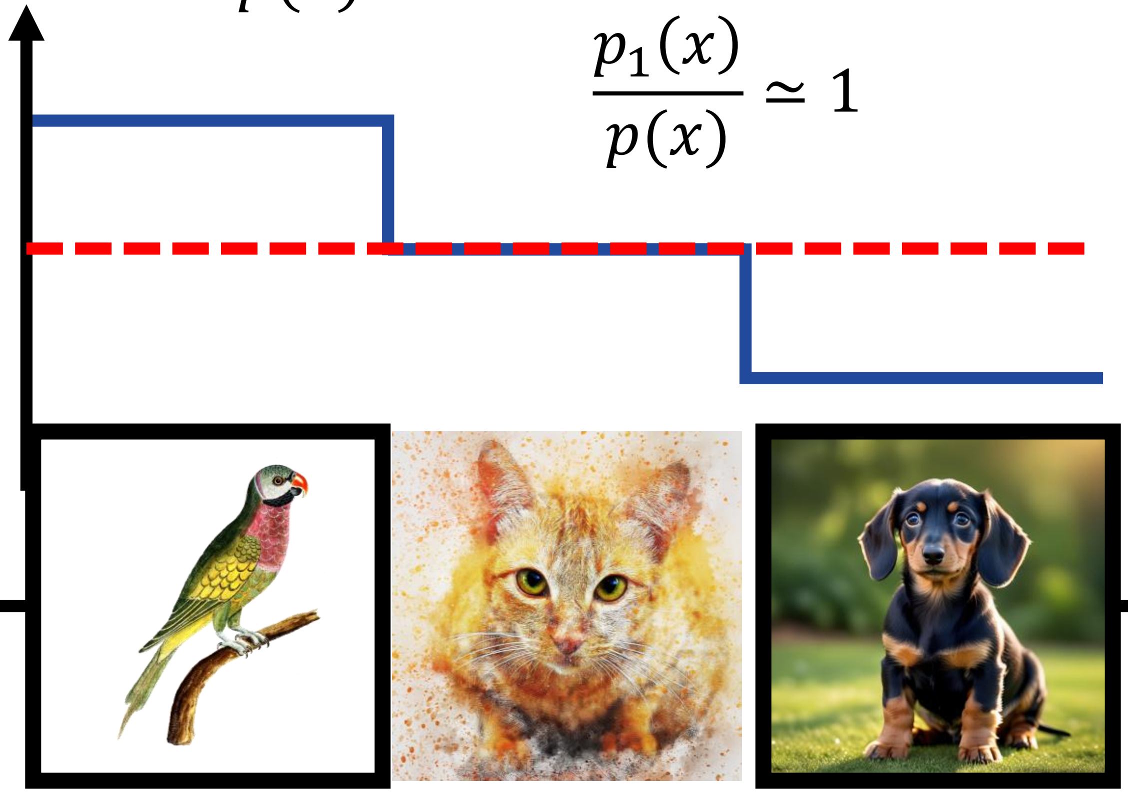
Idea

$p_1(x)$ $\qquad \dfrac{p_1(x)}{p(x)} > 1$

$\dfrac{p_1(x)}{p(x)} \simeq 1$

$p(x) = \dfrac{1}{2} \cdot \left(p_1(x) + p_2(x)\right)$

$\dfrac{p_1(x)}{p(x)} < 1$



Good performance
High accuracy

Bad performance
Low accuracy

02 Algorithm

Theoretical Results

Definition 1. For K clients, the ensemble of their models and weight functions $\{(h_k, w_k)\}_{k=1}^K$ is said to be an optimal model ensemble if the following holds:

$$\mathcal{L}_p\left(\sum_{k=1}^K w_k \cdot h_k\right) = \mathbf{E}_p\left[l\left(\sum_{k=1}^K w_k(x) \cdot h_k(x), y(x)\right)\right] \leq \min_{h \in \mathcal{H}} \mathcal{L}_p(h) = \mathcal{L}_p(h_p^*). \tag{6}$$

Theorem 3. Let the loss function l be convex. Define the client weight functions $\{w_k^*\}_{k=1}^K$ as follows:

$$w_k^*(x) \triangleq \frac{n_k \cdot p_k(x)}{\sum_{i=1}^K n_i \cdot p_i(x)} = \frac{\pi_k \cdot p_k(x)}{\sum_{i=1}^K \pi_i \cdot p_i(x)}. \tag{10}$$

Then, the ensemble $\{h_{p_k}^*, w_k^*\}_{k=1}^K$ is an optimal model ensemble, i.e., $\mathcal{L}_p\left(\sum_k w_k^* \cdot h_{p_k}^*\right) \leq \mathcal{L}_p(h_p^*)$.

02 Algorithm

Theoretical Results

Definition 1. For K clients, the ensemble of their models and weight functions $\{(h_k, w_k)\}_{k=1}^{K}$ is said to be an optimal model ensemble if the following holds:

$$\mathcal{L}_p\left(\sum_{k=1}^{K} w_k \cdot h_k\right) = \mathbf{E}_p\left[l\left(\sum_{k=1}^{K} w_k(x) \cdot h_k(x), y(x)\right)\right] \leq \min_{h \in \mathcal{H}} \mathcal{L}_p(h) = \mathcal{L}_p(h_p^*). \quad (6)$$

Theorem 3. Let the loss function l be convex. Define the client weight functions $\{w_k^*\}_{k=1}^{K}$ as follows:

$$w_k^*(x) \triangleq \frac{n_k \cdot p_k(x)}{\sum_{i=1}^{K} n_i \cdot p_i(x)} = \frac{\pi_k \cdot p_k(x)}{\sum_{i=1}^{K} \pi_i \cdot p_i(x)}. \quad (10)$$

Then, the ensemble $\{h_{p_k}^*, w_k^*\}_{k=1}^{K}$ is an optimal model ensemble, i.e. $\mathcal{L}_p\left(\sum_k w_k^* \cdot h_{p_k}^*\right) \leq \mathcal{L}_p(h_p^*)$.

02 Algorithm

Theoretical Results

Definition 2. (Odds): For $\phi \in (0,1)$, its odds value Φ is defined as $\Phi(\phi) = \frac{\phi}{1-\phi}$.

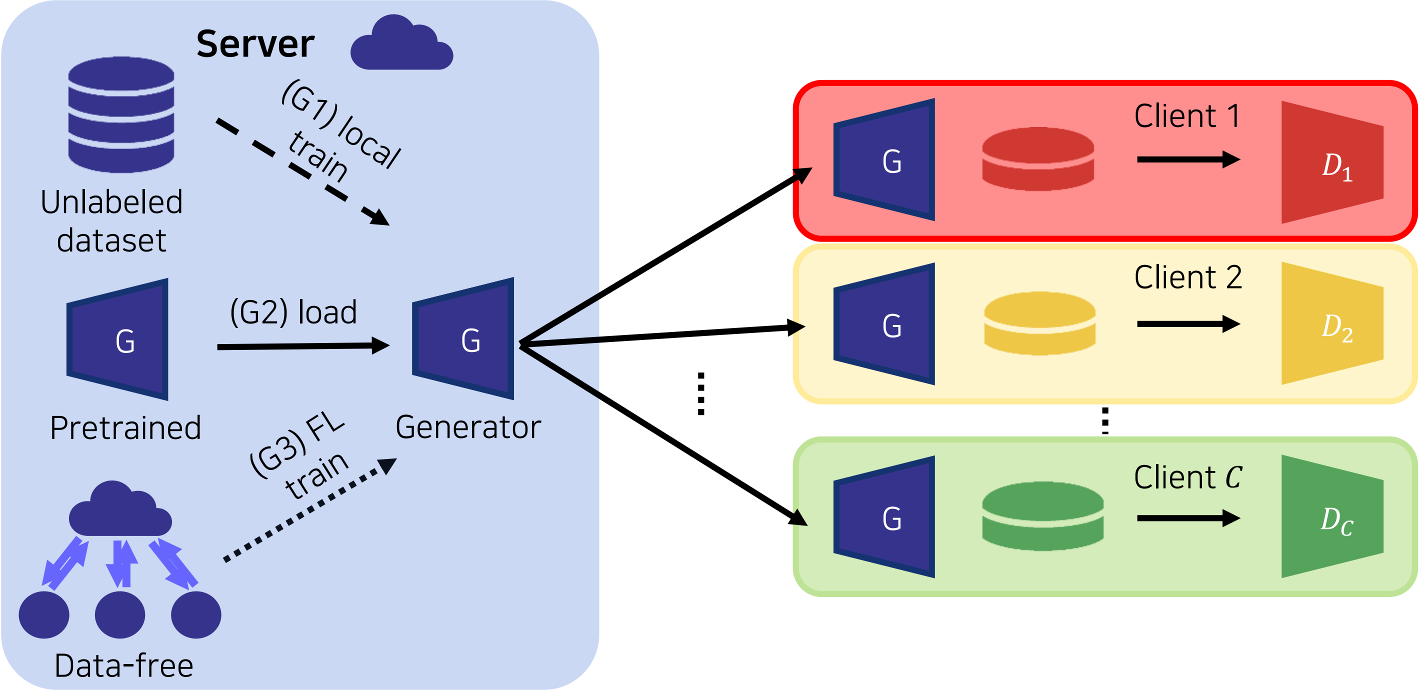
Theorem 4. For a fixed generator G with generating distribution p_g, let D_k be an optimal discriminator for generator G and client k's distribution p_k. Assume that D_k outputs a value over $(0,1)$ using a sigmoid activation function, and let $\Phi_k(x) \triangleq \Phi(D_k(x))$. Then, for $x \in supp(p_g)$, the following holds:

$$\frac{n_k \cdot \Phi_k(x)}{\sum_{i=1}^{K} n_i \cdot \Phi_i(x)} = \frac{\pi_k \cdot p_k(x)}{\sum_{i=1}^{K} \pi_i \cdot p_i(x)} = w_k^*(x). \tag{11}$$

02 Algorithm

FedGO Algorithm

1. Pre-FL : Client discriminators preparation



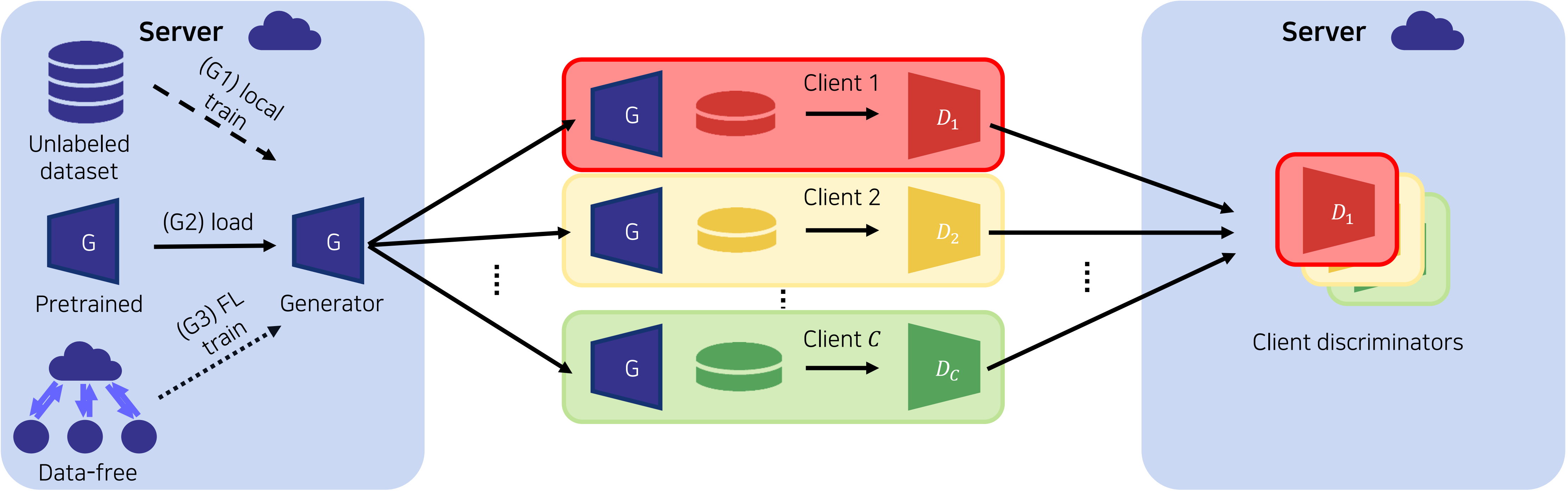
1. Generator preparation (G1 or G2 or G3)

2. Client discriminator training

02 Algorithm

FedGO Algorithm

1. Pre-FL : Client discriminators preparation

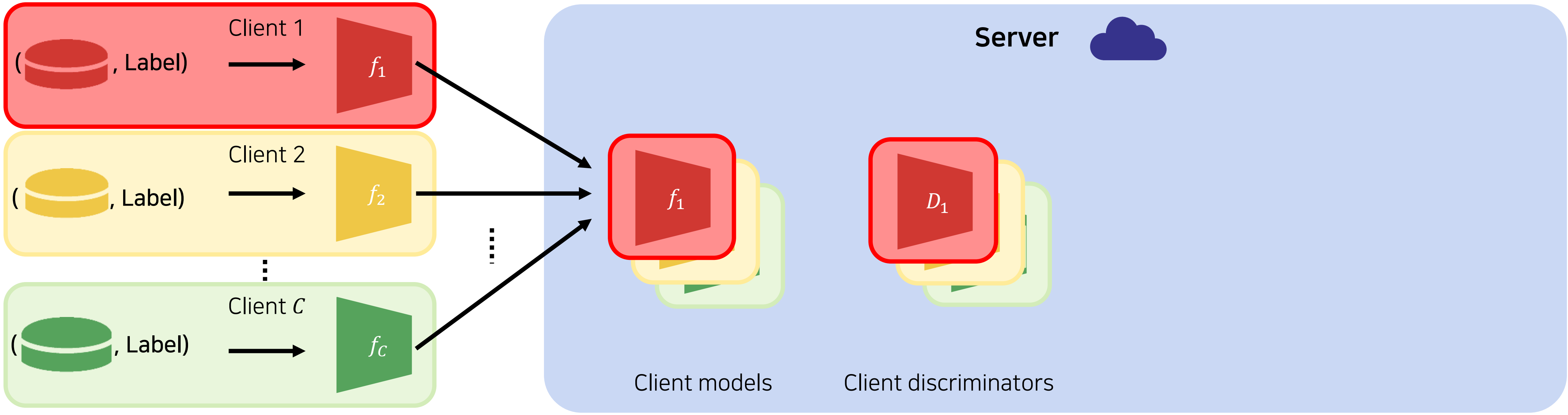


1. Generator preparation (G1 or G2 or G3) 2. Client discriminator training 3. Client discriminators aggregation

02 Algorithm

FedGO Algorithm

2. Main FL : Ensemble distillation along with client discriminators

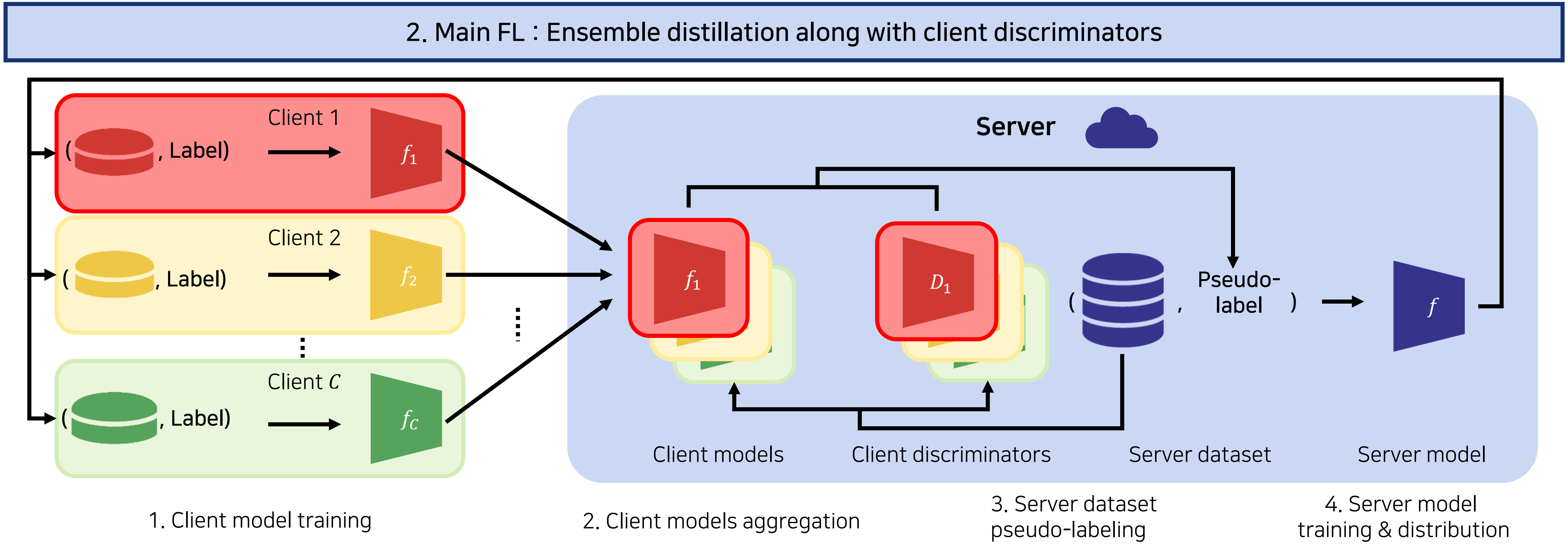


1. Client model training

2. Client models aggregation

02 Algorithm

FedGO Algorithm

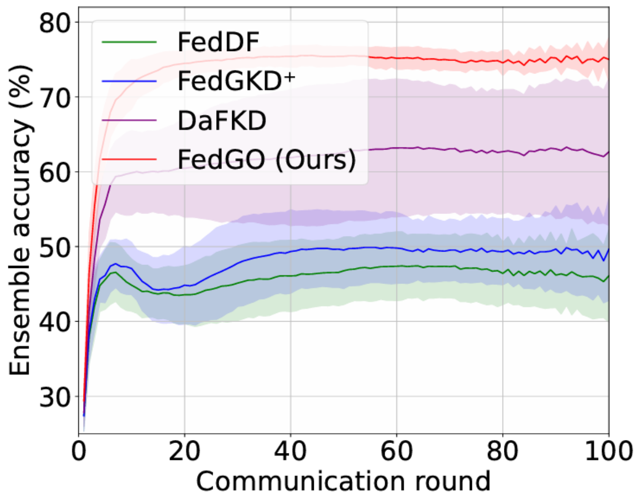


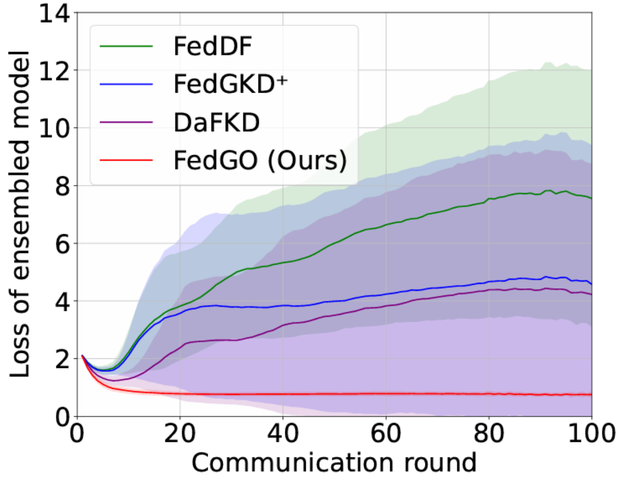
03 Experimental Results

Results

Table 3. Server test accuracy (%) of our FedGO and baselines on three image datasets at the 100-th communication round. A smaller α indicates higher heterogeneity.

	CIFAR-10		CIFAR-100		ImageNet100	
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$
Central Training	85.33±0.25		51.72±0.65		43.20±1.00	
FedAVG	58.65±5.75	46.61±8.54	38.93±0.74	36.66±0.97	29.44±0.41	27.58±0.88
FedProx	64.69±2.15	55.56±9.86	38.21±0.95	34.44±1.26	29.96±0.66	26.99±0.97
SCAFFOLD	61.20±3.98	50.10±10.00	38.15±0.80	36.14±1.06	29.13±0.79	27.08±0.69
FedDisco	56.78±7.22	48.08±8.35	38.81±1.02	36.86±0.88	29.69±0.66	27.54±0.51
FedUV	62.58 ± 4.83	53.80 ± 5.68	38.84 ± 0.79	36.17 ± 1.24	30.09 ± 1.09	27.32 ± 0.65
FedTGP	61.16 ± 6.98	61.51 ± 7.78	39.58 ± 0.10	36.56 ± 0.11	29.21 ± 1.13	26.34 ± 1.02
FedDF	71.56±5.09	59.53±9.88	42.74±1.22	37.18±1.03	33.48±1.00	30.94±1.60
FedGKD$^+$	72.59±4.10	59.96±8.60	43.35±1.14	40.47±1.00	34.10±0.67	31.42±0.93
DaFKD	71.52±5.56	67.51±10.77	44.12±2.25	39.50±0.85	33.34±0.69	31.59±1.46
FedGO (ours)	**79.62**±4.36	**72.35**±9.01	**44.66**±1.27	**41.04**±0.99	**34.20**±0.71	**31.70**±1.55





Paper link

Project link

Thank you.

Provably Near-Optimal Federated Ensemble Distillation with Negligible Overhead