

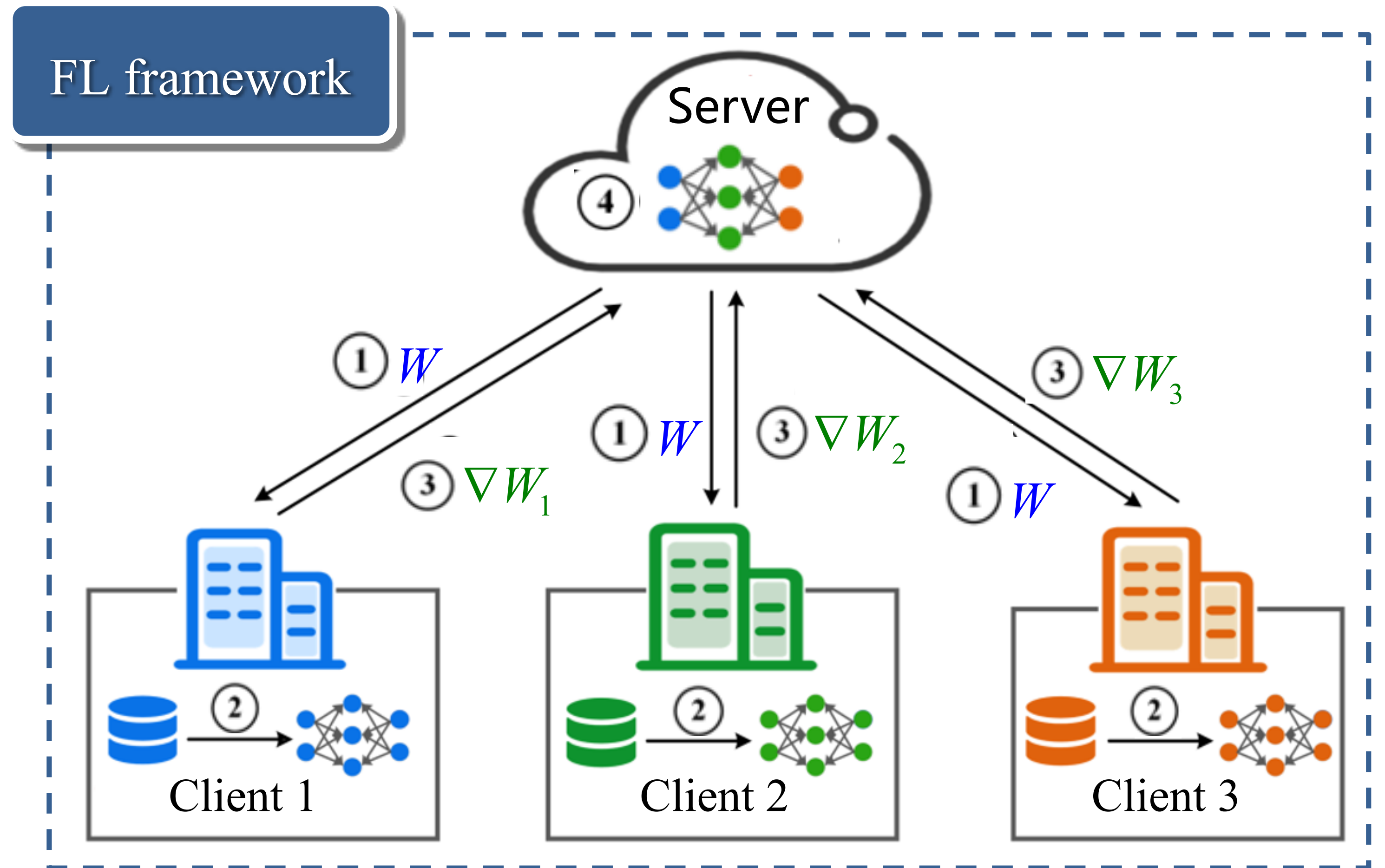
FedDCS: Dynamical Client Selection for Federated Learning in Mobile Scenarios with Label Distribution Shift

Junjie Pang* , Haohua Du* , Zhiyi Liu* , Xiaoya Xu* , and YuanHao Feng†

*Beihang University, †The Hong Kong Polytechnic University

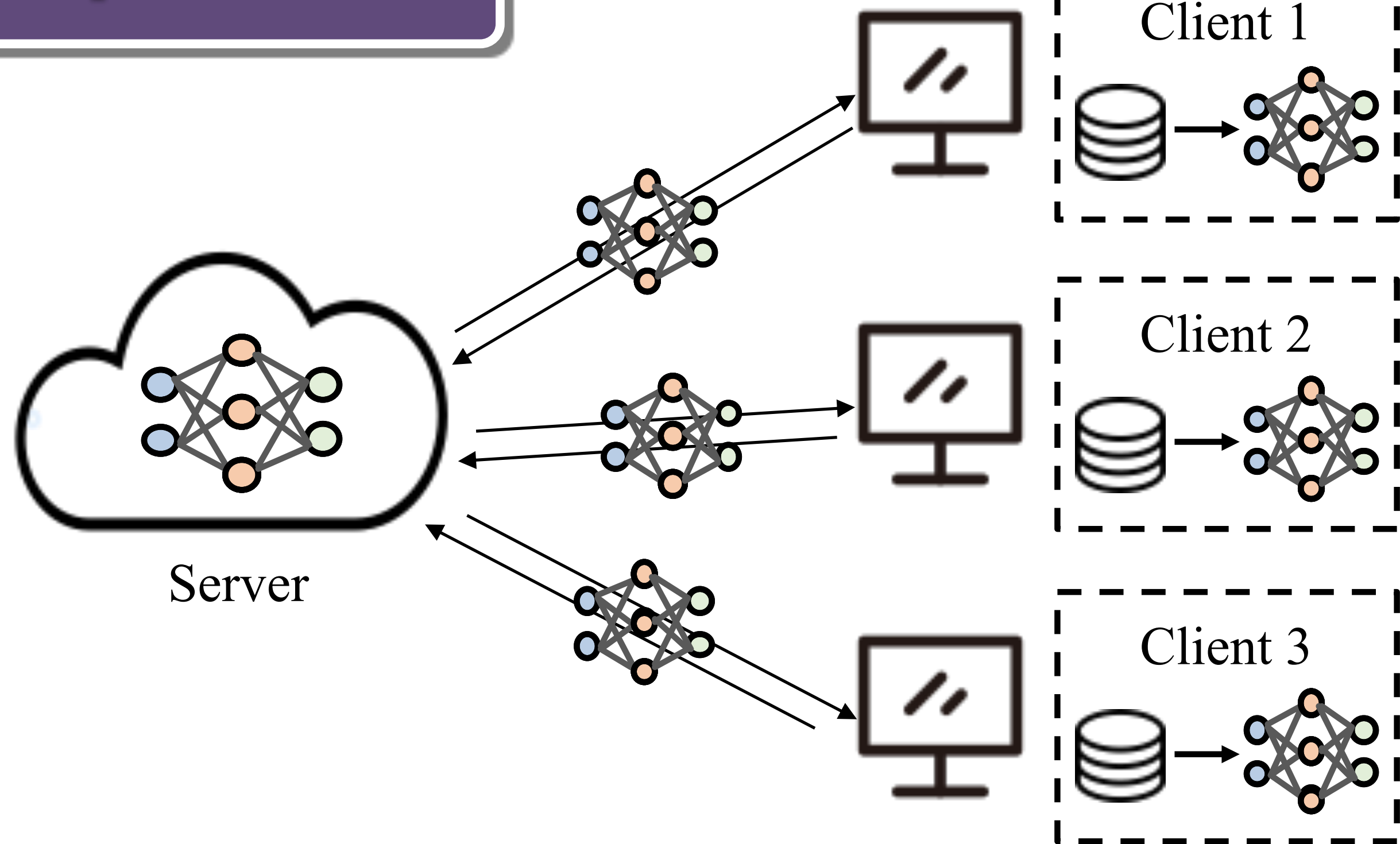
FedDCS: Background and Challenge

Federated learning is a distributed learning paradigm that effectively decouples the dependency of deep learning model training on centralized data storage, thereby protecting the privacy and security of the participants.

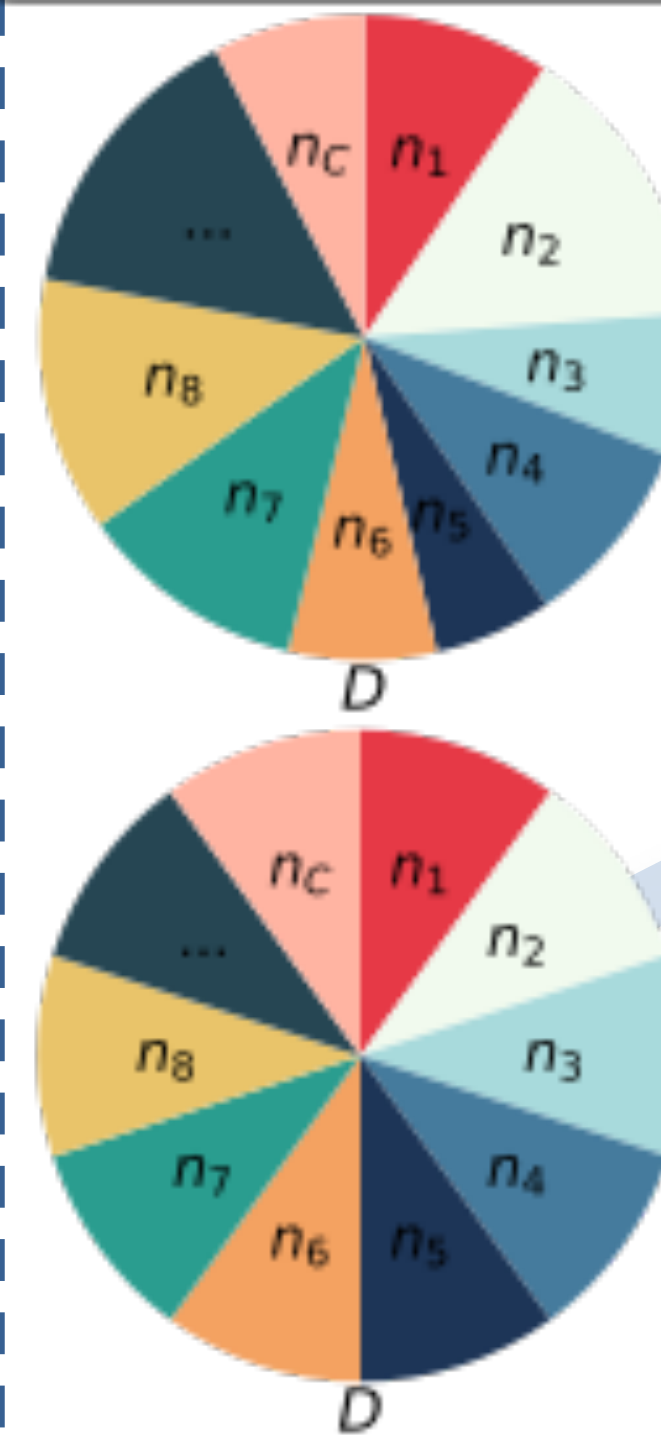


FedDCS: Background and Challenge

Computational Load



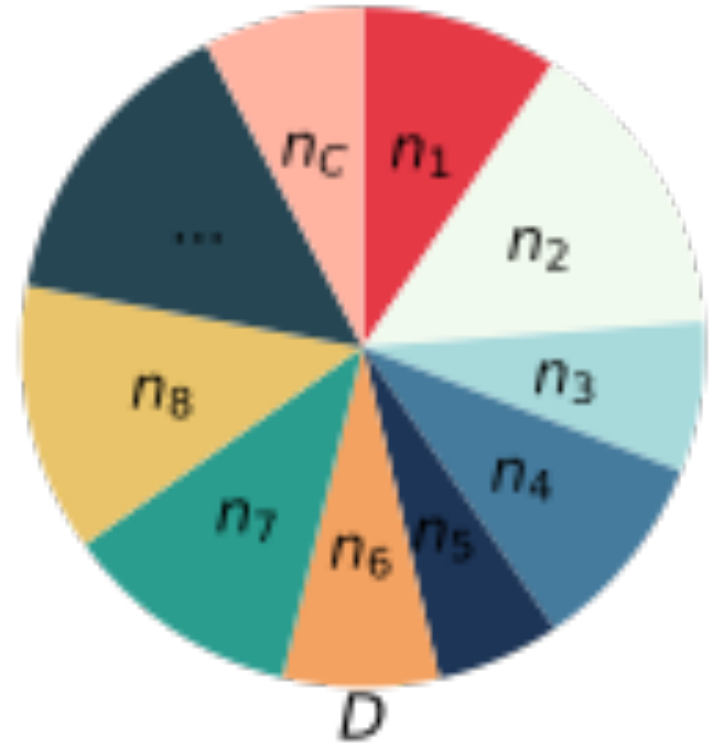
Label Distribution Shift



- $p_i = \frac{n_i}{|D|}$
- Data label distribution:
 $p^k = \{p_1, p_2, p_3, \dots, p_C\}$
- p^* : Optimal data label distribution expected for the FL task
- **Label Distribution Shift**
 p^k is inconsistent with p^* , i.e. p^k deviates from p^* .

FedDCS: Background and Challenge

Label Distribution Shift

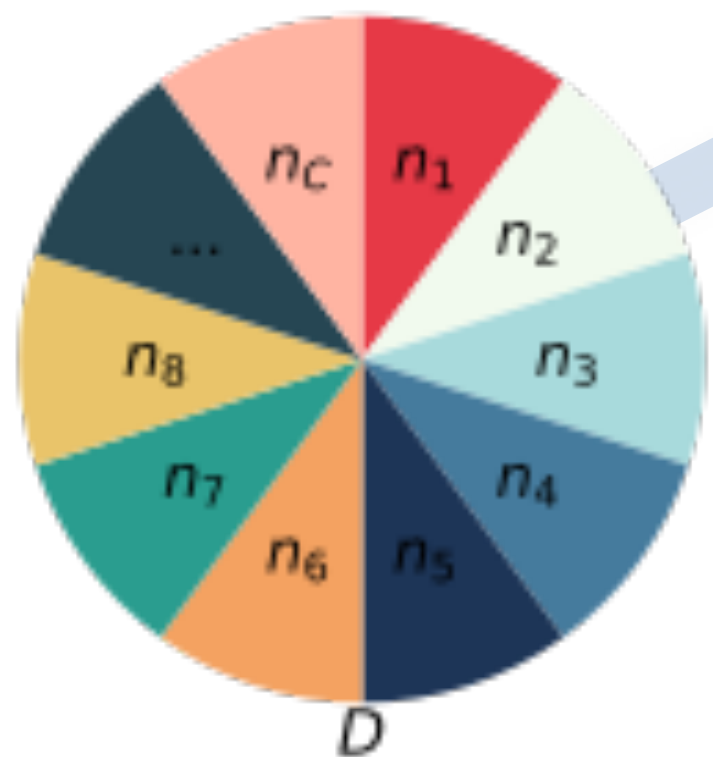


➤ $p_i = \frac{n_i}{|D|}$

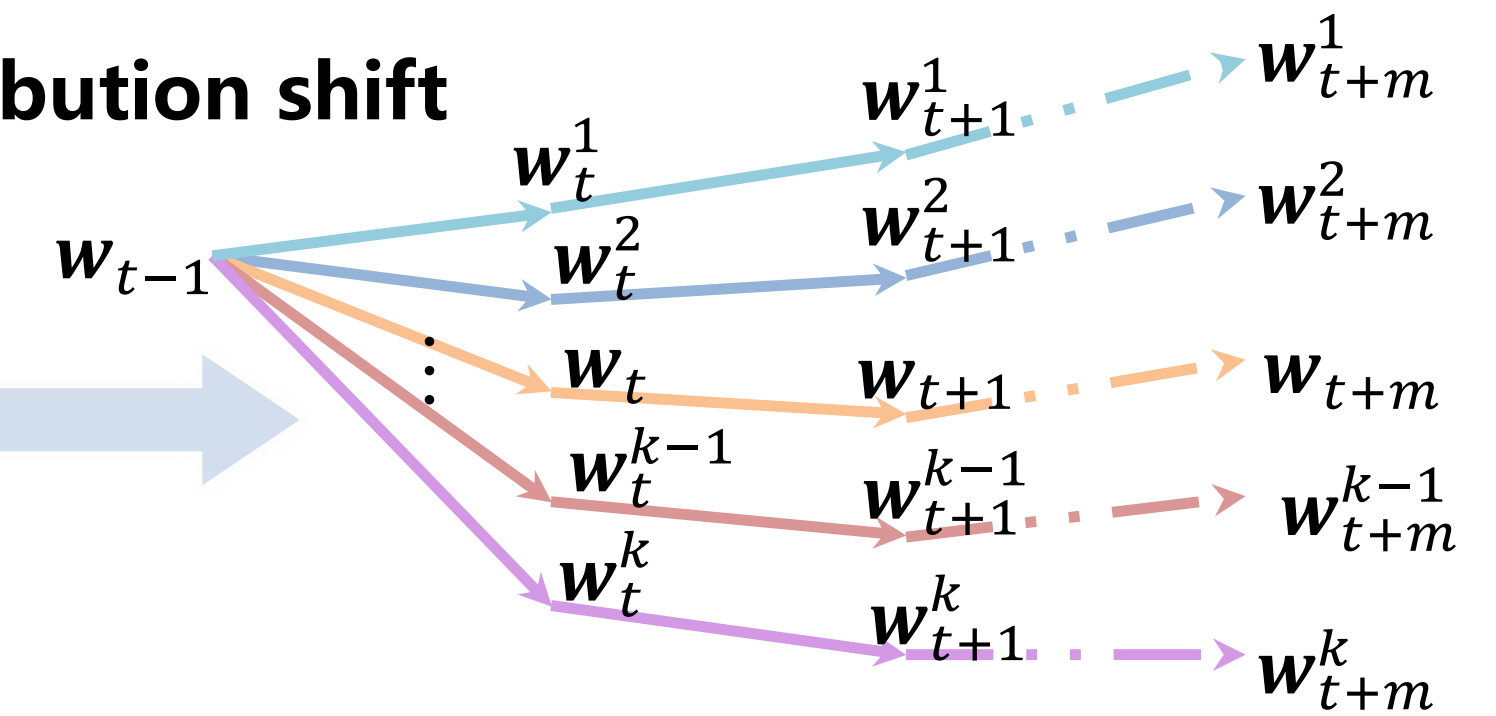
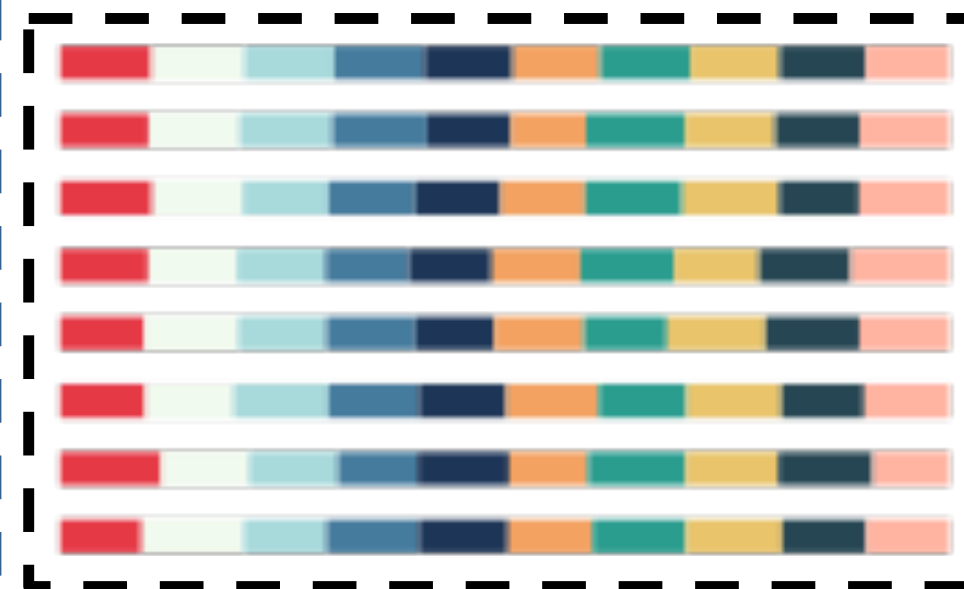
➤ Data label distribution:
 $p^k = \{p_1, p_2, p_3, \dots, p_c\}$

➤ p^* : Optimal data label distribution expected for the FL task

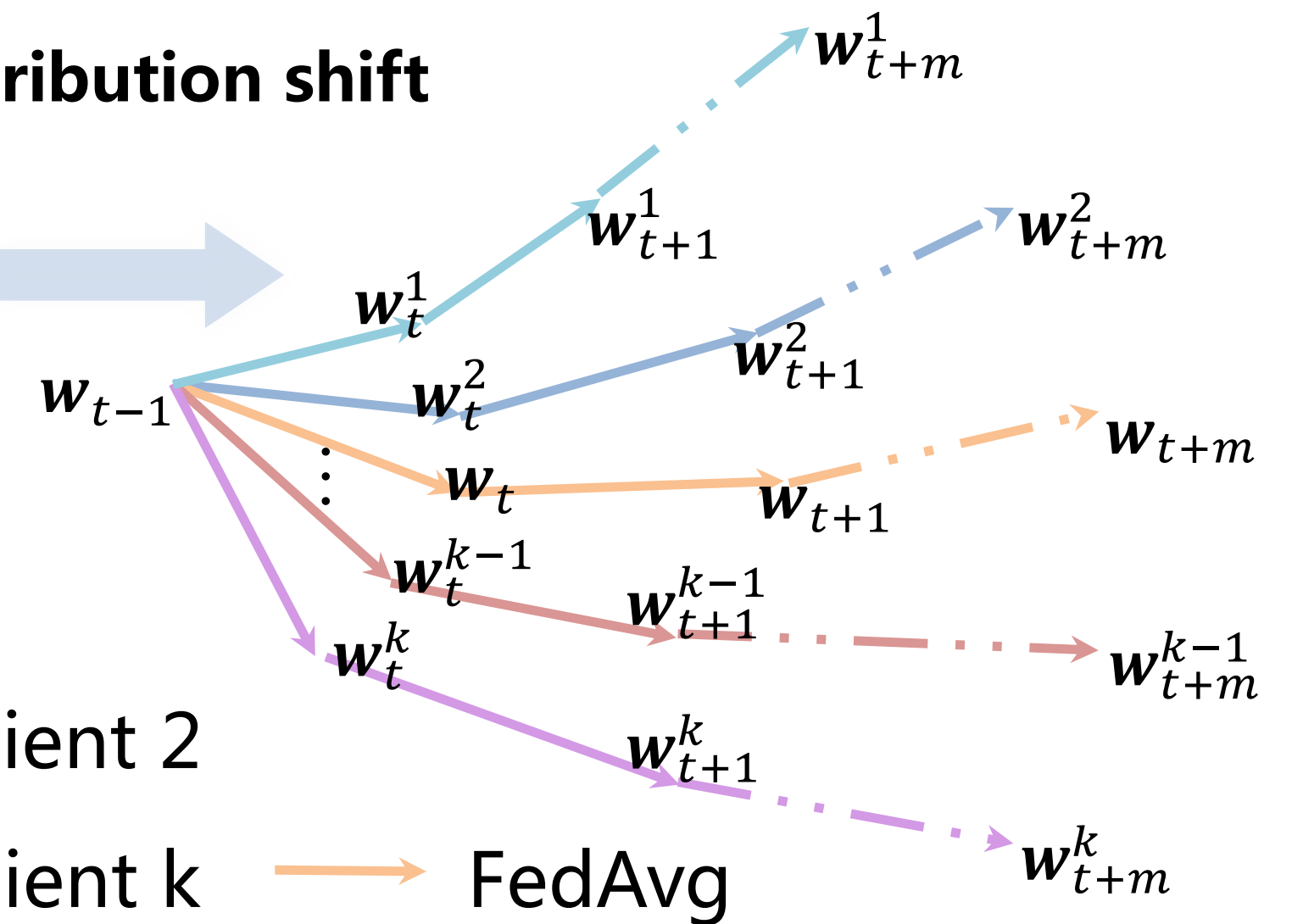
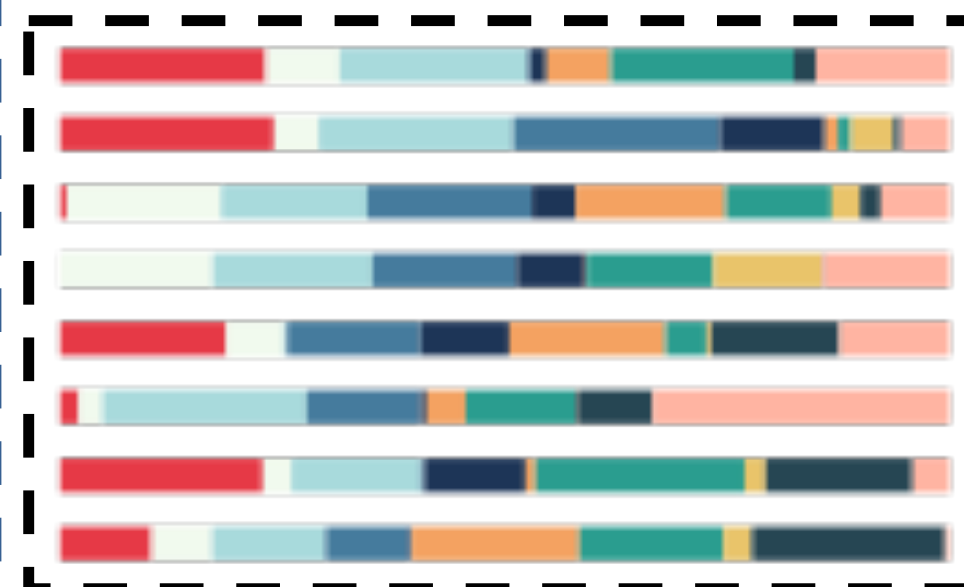
➤ **Label Distribution Shift**



Low degree of label distribution shift



High degree of label distribution shift

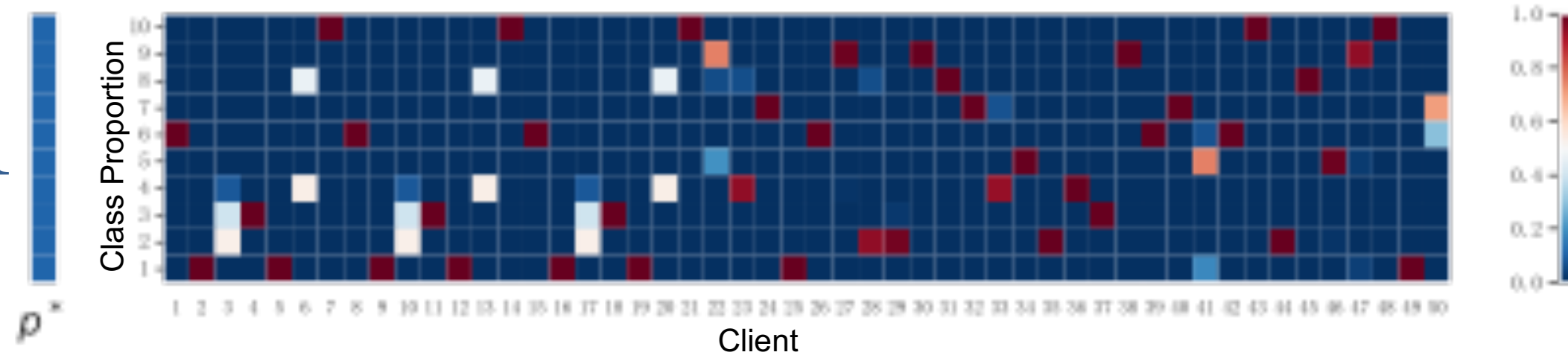


FedDCS: Background and Challenge

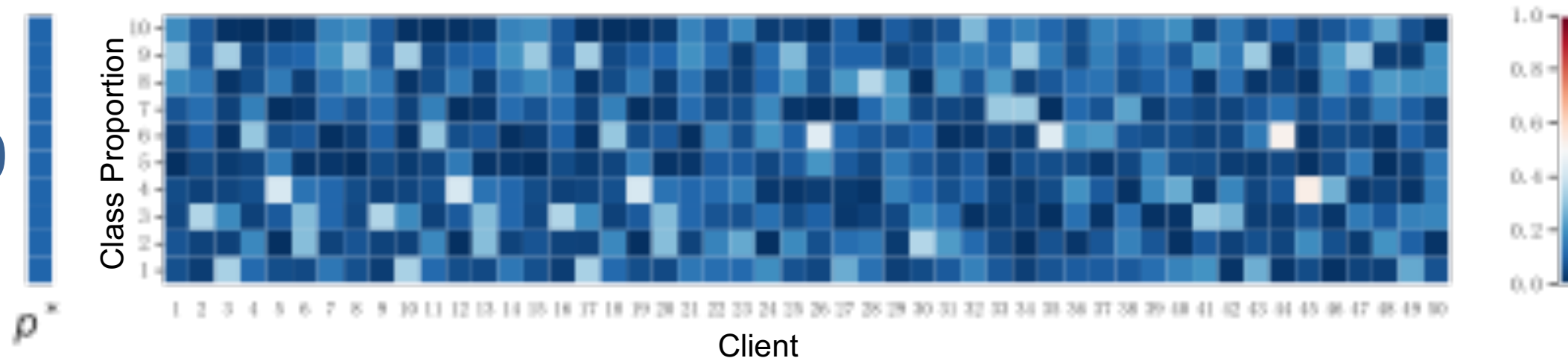
Label Distribution

$$p^k = \text{Dirichlet}(\gamma p^*)$$

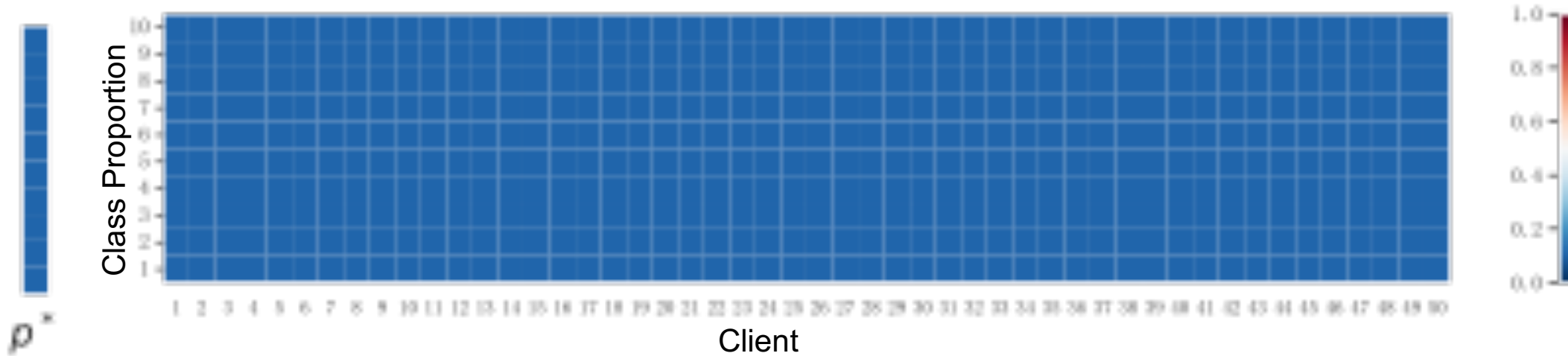
$\gamma = 0.1$



$\gamma = 10$

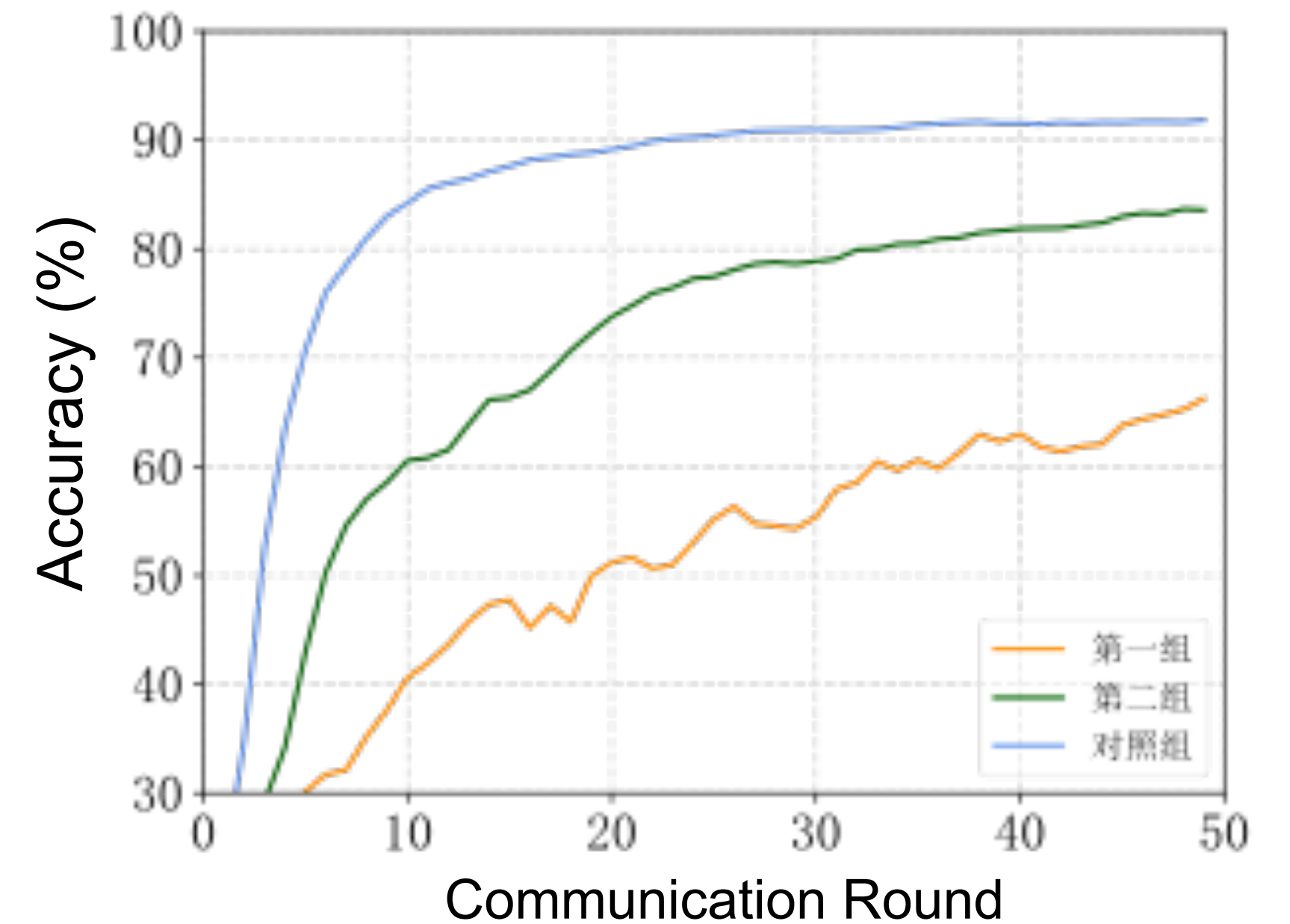


$\gamma = \infty$



Accuracy

The client data label distribution shift significantly reduces the performance of the global model!



FedDCS: Background and Challenge

Recent Advances:

Data Sharing (Tour et al. 2021)

Data Augmentation (Yoon et al. 2021)

Model Correction (Shi et al. 2023)

Knowledge Distillation (Chang et al. 2021)

Client Selection (Wu et al. 2022)

Client Clustering (Xie et al. 2023)

Remain Limitations:

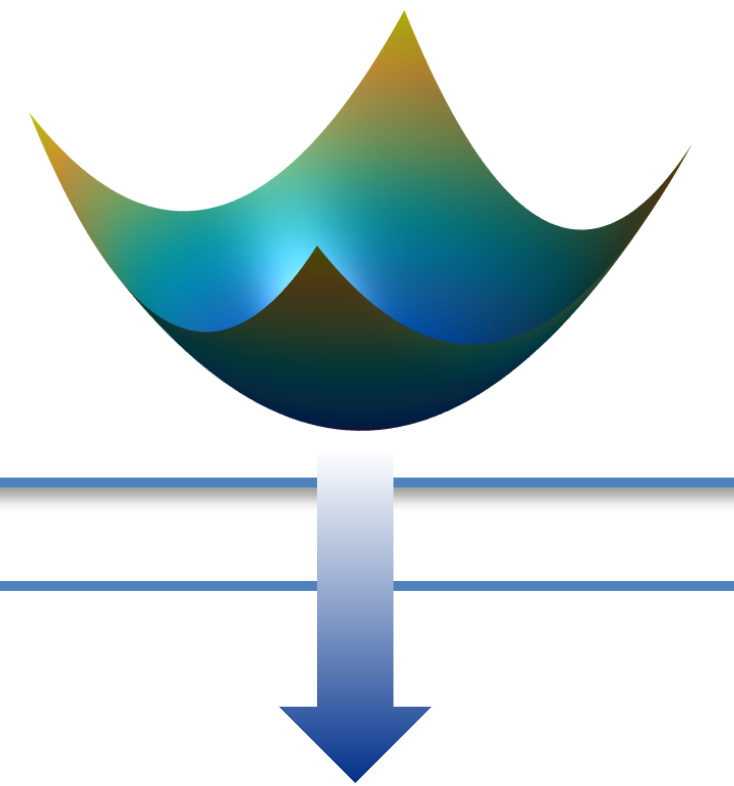
1. Such approaches only focus on the differences in client data distribution without considering the actual target task requirements.
2. Neglecting the dynamic changes in client data.
3. Overlooking the differences in the amount of knowledge contained in the data across different clients.

FedDCS: Design

□ Label Distribution Relevance Estimation

- Assumption1 : The loss function $\mathbb{E}_{x|y}[\log f_i(\mathbf{x}, \mathbf{w})]$ differentiable over the space θ with *Lipschitz* continuous gradient for each class $i \in [C]$. For any $\mathbf{w}, \mathbf{w}' \in \theta$:

$$\|\nabla_{\mathbf{w}} \mathbb{E}_{x|y}[\log f_i(\mathbf{x}, \mathbf{w})] - \nabla_{\mathbf{w}'} \mathbb{E}_{x|y}[\log f_i(\mathbf{x}, \mathbf{w}')]\| \leq L_{x|y=i} \|\mathbf{w} - \mathbf{w}'\|$$



- Earth Mover's Distance (EMD) & model weights

$$\|\mathbf{w}_t^A - \mathbf{w}_t^B\| \leq \left(1 + \eta \sum_{i=1}^C p^B(y=i) L_{x|y=i}\right) \|\mathbf{w}_{t-1}^A - \mathbf{w}_{t-1}^B\| + \eta g_{\max}(\mathbf{w}_{t-1}^A) \sum_{i=1}^C \|p^B(y=i) - p^A(y=i)\|$$

$\mathbf{w}_t^A$ and $\mathbf{w}_t^B$ are the model weights of client A and B, p^A and p^B are corresponding data label distribution, $L_{x|y=i}$ is Lipschitz constant, $g_{\max}(\mathbf{w}) := \max_{i=1}^C \|\nabla_{\mathbf{w}} \mathbb{E}_{x|y}[\log f_i(\mathbf{x}, \mathbf{w})]\|$

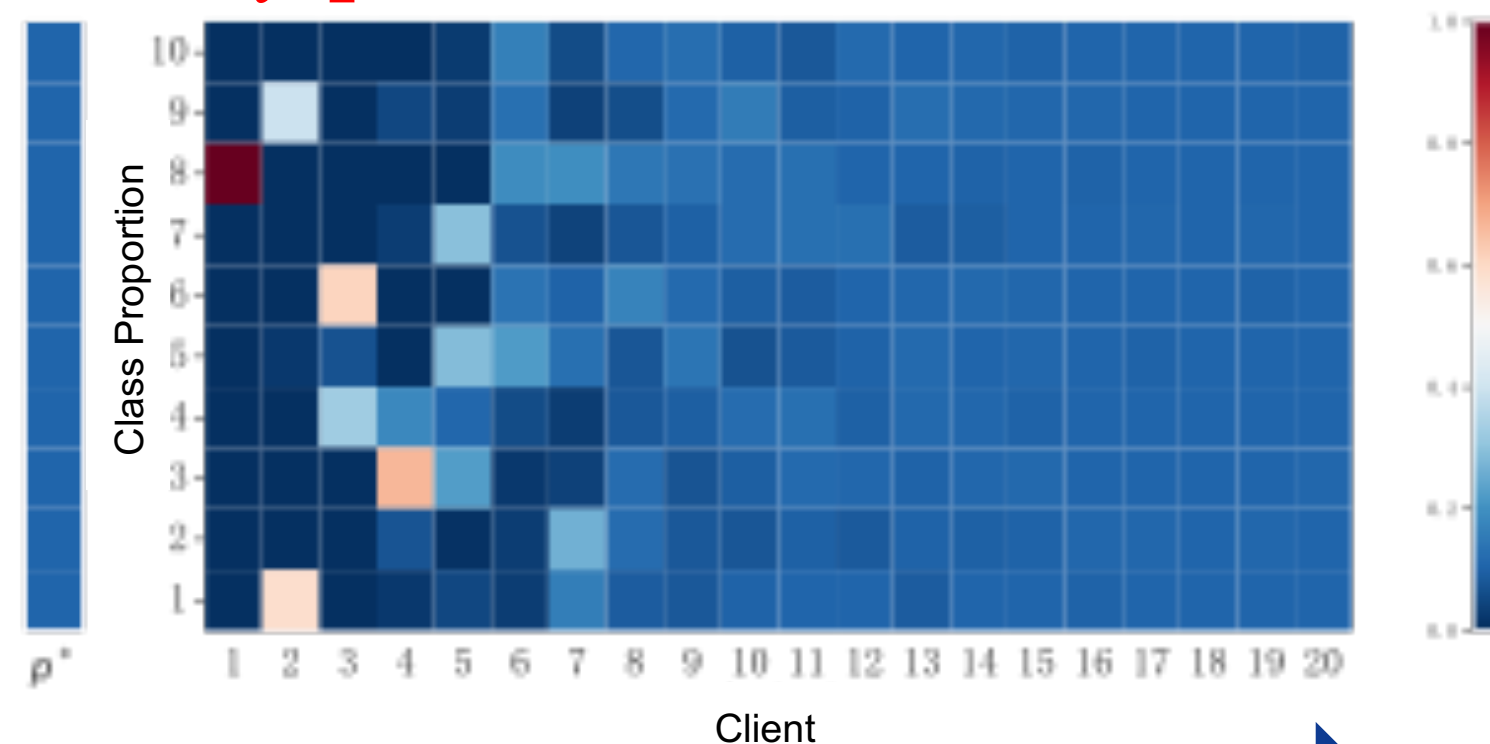
FedDCS: Design

Label Distribution Relevance Estimation

Relevance Quantification

- Quantifying the relevance between data label distribution and the target task:

$$\sum_{i=1}^C \|p^*(y=i) - p_t^k(y=i)\|$$



The relevance between label distribution and the target task strengthens.

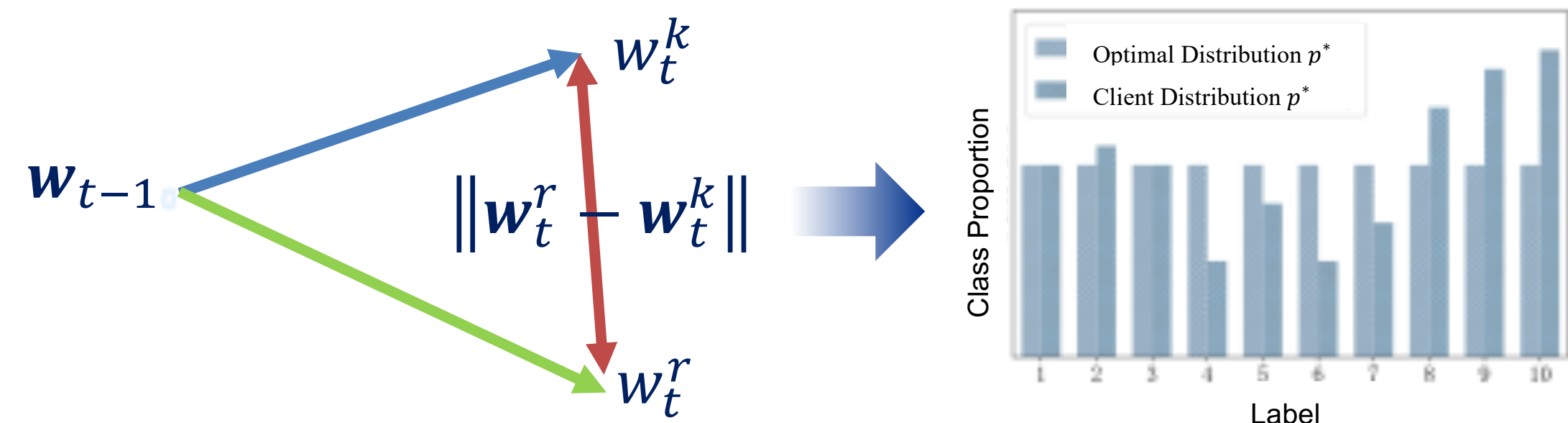
Relevance Estimation

$$D^r \sim p^*$$

- Root model : $\mathbf{w}_t^r = \mathbf{w}_{t-1} - \eta \nabla_{\mathbf{w}} \ell(\mathbf{w}_{t-1}, D^r)$

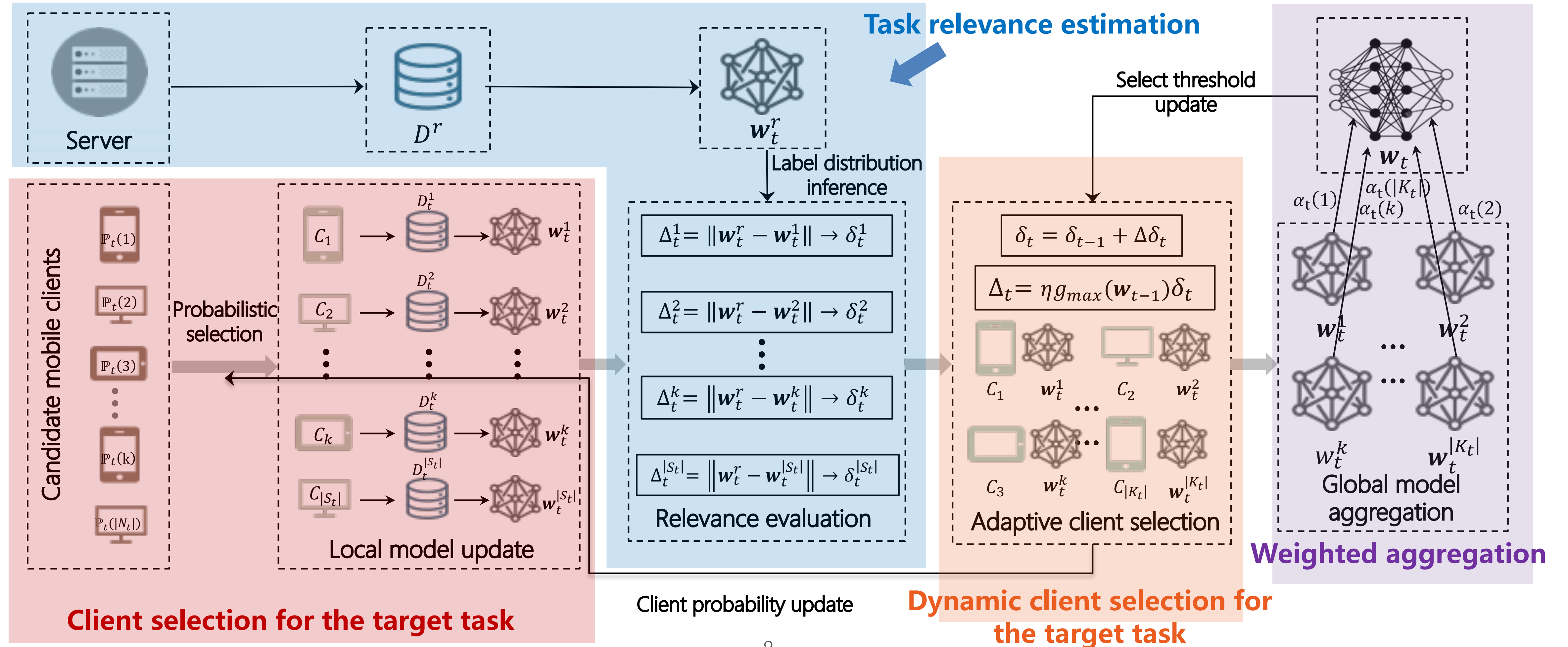
- Relevance Estimation : $\Delta_t^k = \|\mathbf{w}_t^r - \mathbf{w}_t^k\| \leq$

$$\eta g_{\max}(\mathbf{w}_{t-1}) \sum_{i=1}^C \|p^*(y=i) - p_t^k(y=i)\|$$



FedDCS: Design

Framework

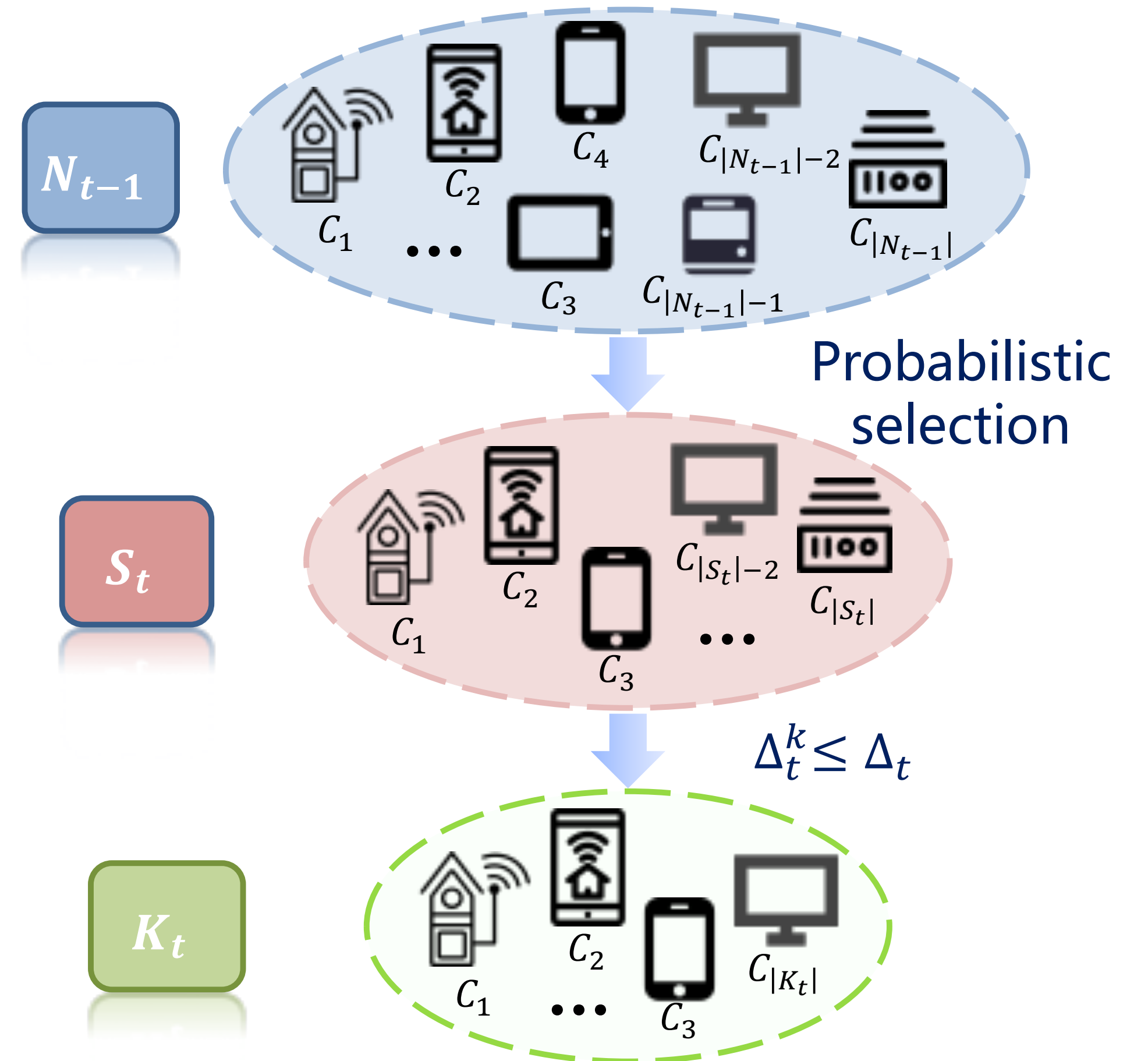


FedDCS: Design

□ Adaptive Client Selection

Two-stage client selection

- 1st stage :
 - Probabilistic select client subset $S_t \subseteq N_{t-1}$ to download the latest global model and update local model.
- 2nd stage :
 - Calculate the threshold of model weights distance : $\Delta_t = \eta \cdot g_{max}(\mathbf{w}_{t-1}) \cdot \delta_t$.
 - Calculate $\ell_2 - norm$: $\Delta_t^k = \|\mathbf{w}_t^r - \mathbf{w}_t^k\|$, select clients $K_t(\Delta_t^k \leq \Delta_t)$ for updating the global model.



FedDCS: Design

□ Client probability update

Update the probabilities of new clients in round t

$$\mathbb{P}_{t+1}(a) = \frac{1}{|N_t|}, a \in A_t$$

Recycle the probabilities of clients left in round t

$$\mathbb{P}_t(Q_t) = \sum_{q \in Q_t} \mathbb{P}_t(q)$$

Update the probabilities of clients in $N_{t-1} \setminus S_t$

$$\Delta \mathbb{P}_{t+1}(k) = \begin{cases} \min \left\{ \frac{\mathbb{P}_t(Q_t) - \mathbb{P}_t(A_t)}{|N_{t-1} \setminus S_t|}, 1 - \mathbb{P}_t(k) \right\}, & \mathbb{P}_t(Q_t) \geq \mathbb{P}_t(A_t) \\ 0 & \mathbb{P}_t(Q_t) < \mathbb{P}_t(A_t) \end{cases}$$

Update the probabilities of clients in $S_t \setminus Q_t$

$$\mathbb{P}_{t+1}(k) = \frac{e^{\Delta_t - \|\mathbf{w}_t^r - \mathbf{w}_t^k\|}}{\sum_{c \in S_t \setminus Q_t} (e^{\Delta_t - \|\mathbf{w}_t^r - \mathbf{w}_t^c\|})} \left(1 - \mathbb{P}_t(A_t) - \sum \mathbb{P}_t(i) \right), k \in S_t \setminus Q_t, i \in N_{t-1} \setminus S_t$$

FedDCS: Design

□ Adaptive adjustment

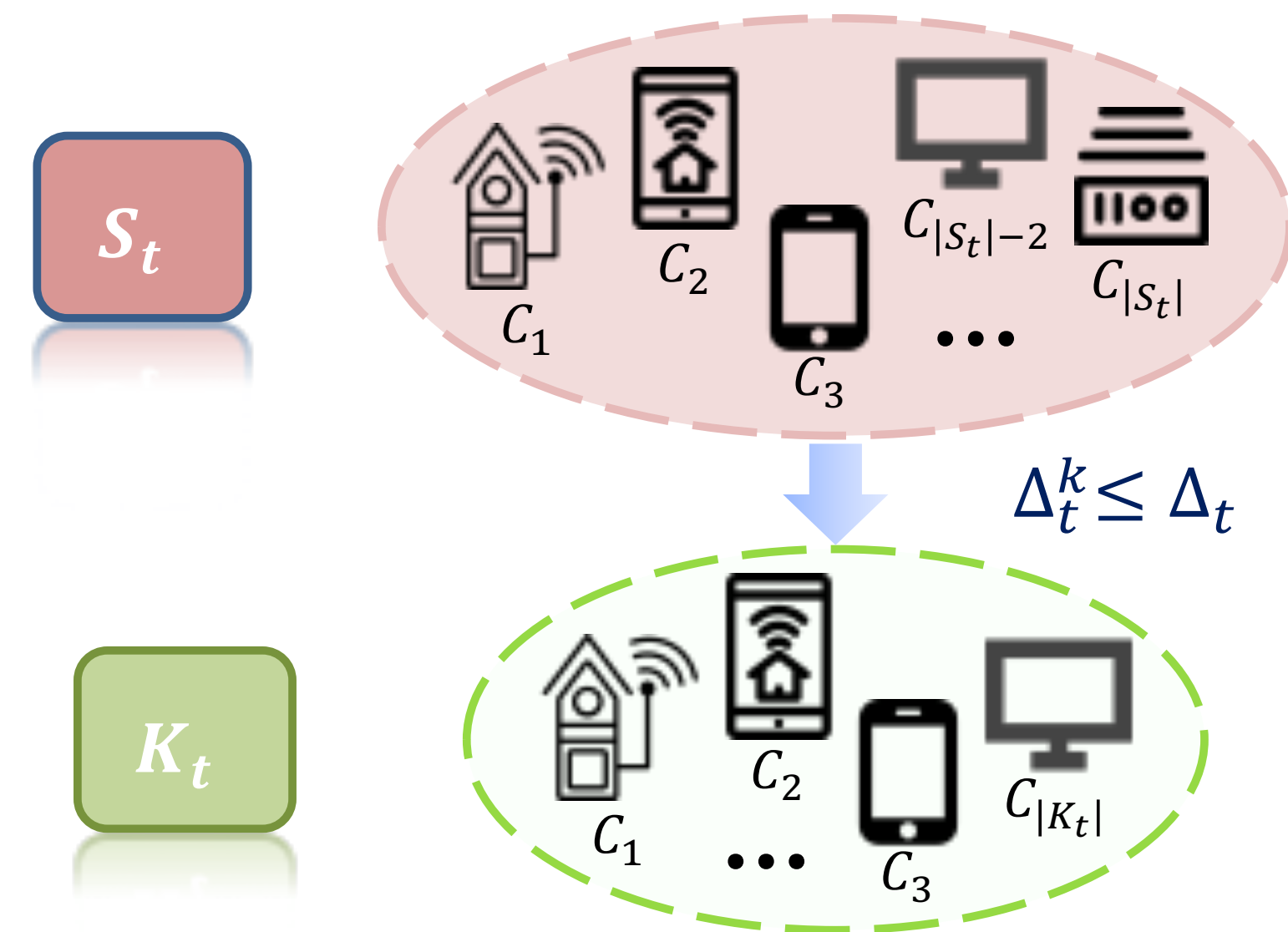
δ -adaptive adjustment

- **Challenge** : δ_t determine the criteria of relevance for client selection, which has a decisive impact on the convergence of FL and the performance of the global model. However, we cannot have any prior knowledge of the optimal value of δ_t .

- **Adaptive search for the optimal δ**

$$\Delta\delta_t = \begin{cases} 0, & \text{if } t = 1 \\ \Delta\delta, & \text{if } t = 2 \\ \Delta\delta_{t-1}, & \text{if } t > 2 \text{ and } \text{loss}_{t-2} > \text{loss}_{t-1} \\ -\Delta\delta_{t-1}, & \text{if } t > 2 \text{ and } \text{loss}_{t-2} \leq \text{loss}_{t-1} \end{cases}$$

$$\delta_t = \delta_{t-1} + \Delta\delta$$



FedDCS: Design

□ Aggregation strategy

Aggregation score

$$s_t(k) = \frac{\mathbf{1} \cdot e^{\Delta_t - \|\mathbf{w}_t^0 - \mathbf{w}_t^k\|}}{\sum_{c \in K_t} (e^{\Delta_t - \|\mathbf{w}_t^0 - \mathbf{w}_t^c\|})} \cdot \frac{n_t^k \cdot 2}{\sum_{c \in K_t} n_t^c}, k \in K_t$$

Normalization

$$\alpha_t^k = \frac{s_t(k)}{\sum_{c \in K_t} s_t(c)}, k \in K_T$$

Update global model

$$\mathbf{w}_t = \sum_{k \in K_t} \alpha_t^k \cdot \mathbf{w}_t^k$$

FedDCS: Experiments

Experimental Settings

➤ Compared Methods

FedPNS(TNSE 2022) , FedAvg(Baseline)

➤ Metrics :

- Accuracy
- Convergence Speed

➤ Client Label Distribution :

$$p^k = \text{Dirichlet}(\gamma p^*)$$

p^* with an uniform distribution :

$$\gamma_1 = 1000 \quad \gamma_2 = 100 \quad \gamma_3 = 0.1$$

p^* with a non-uniform distribution :

$$\gamma_1 = 1000 \quad \gamma_2 = 100 \quad \gamma_3 = 0.1$$

Model	Dataset	η	$ N $	$ S $	n	B	R	E	δ_1	$\Delta\delta$	γ_1	γ_2	γ_3
MLP	MNIST	0.01	50	20	[96, 128]	128	50	2	0.6	0.01	1000	100	0.1
CNN-M	F-MNIST	0.01	50	20	[96, 128]	128	200	2	0.6	0.01	1000	100	0.1
LeNet-5	CIFAR10	0.06	50	20	[192, 256]	128	200	2	0.6	0.01	1000	100	0.1

$|N|$ is the number of initial candidate clients, $|S|$ is the number of clients involved in local model update in each iteration; n is the quantity of training data samples per client; B indicates the data batch size; R represents the number of communications between the server and the clients in FL; E signifies the training epochs for clients in each iteration.

FedDCS: Experiments

Results: Label-balanced p^*

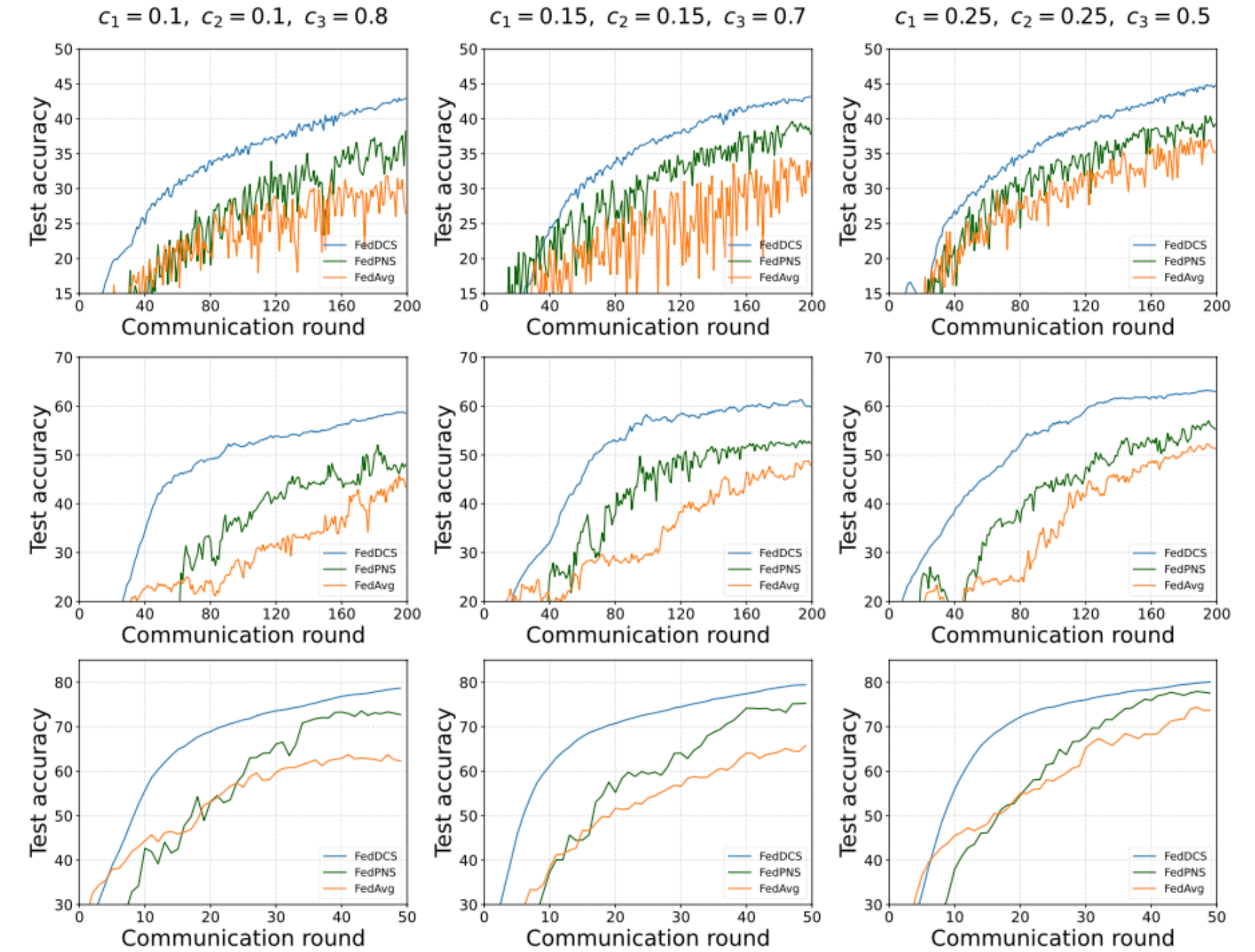
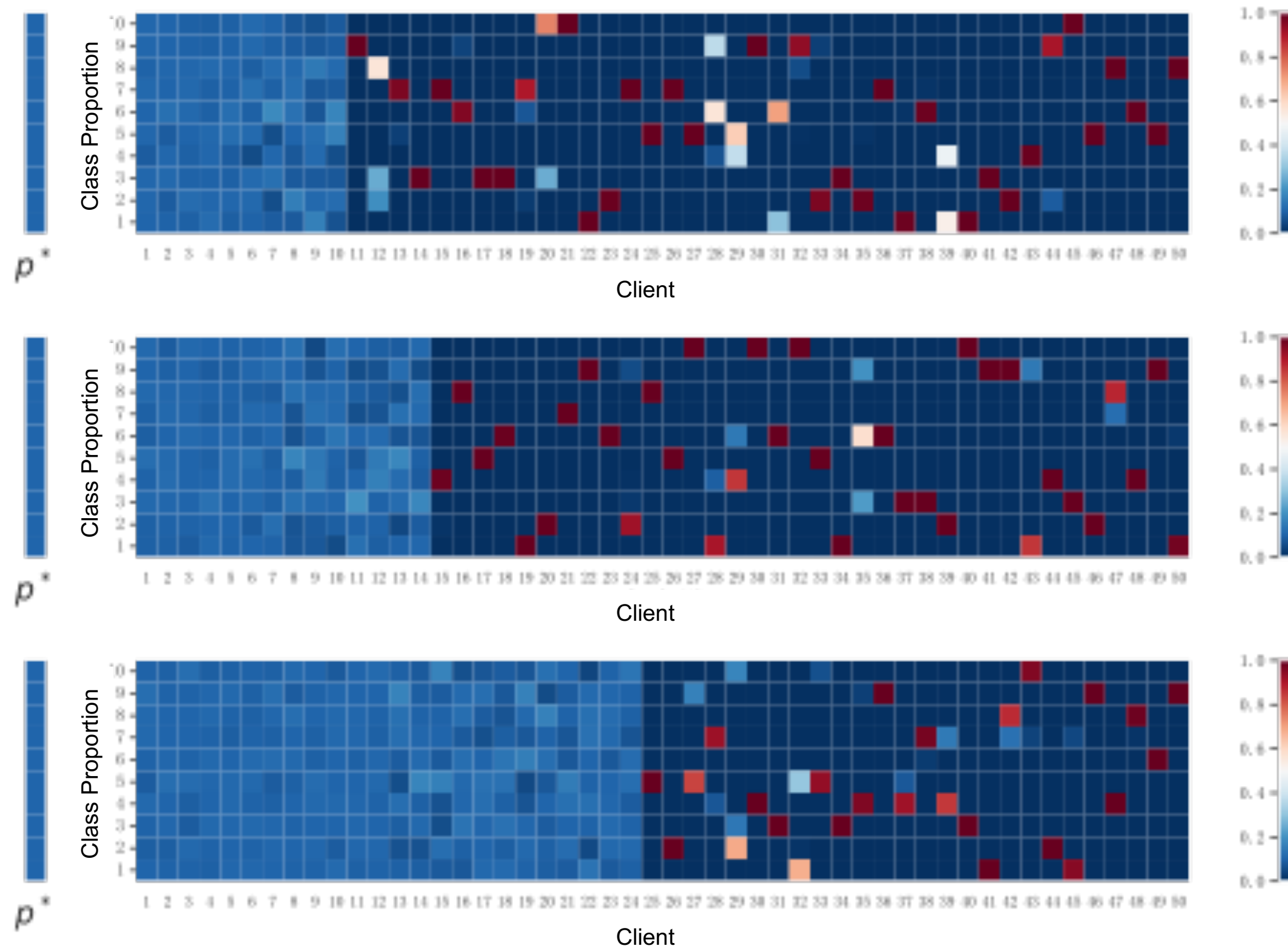
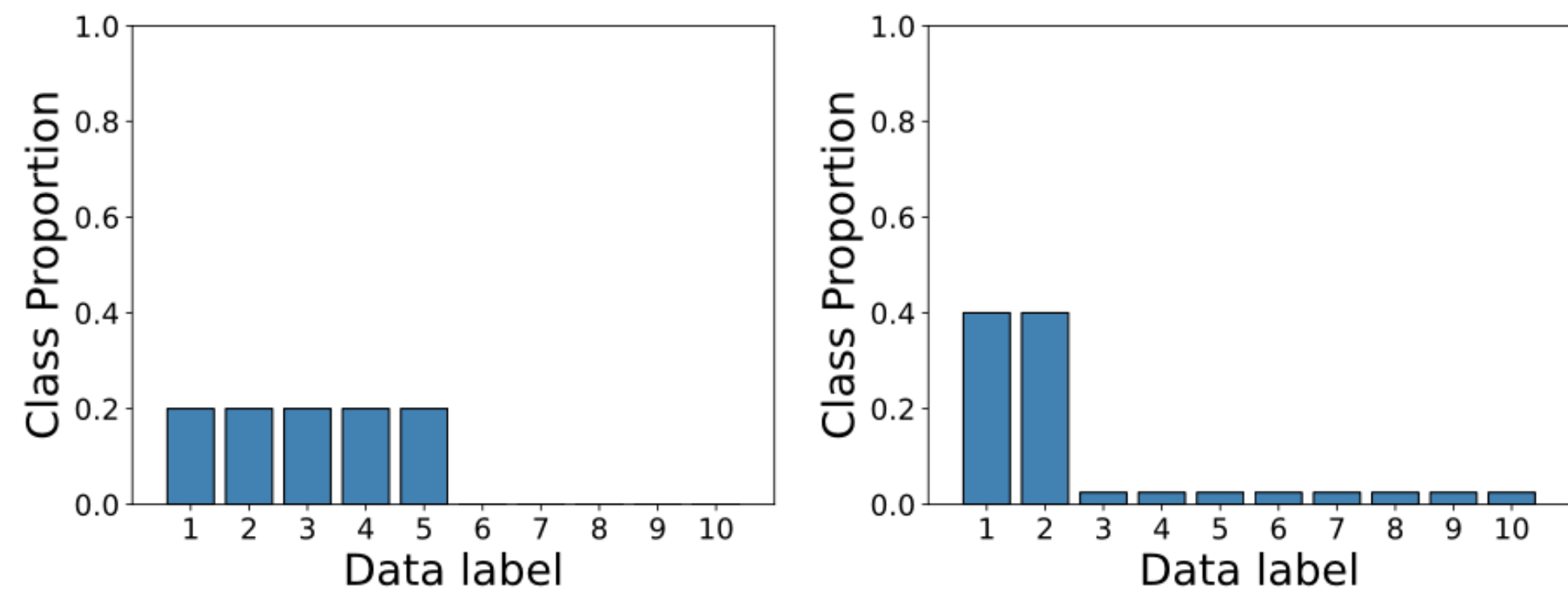


Fig. 3. Performance comparison under label-balanced p^* . (1) Upper: Test accuracy over communication rounds model in CIFAR10. (2) Middle: Test accuracy over communication rounds in F-MNIST. (3) Bottom: Test accuracy over communication rounds in MNIST.

FedDCS: Experiments

Results: Label-unbalanced p^*



(a) Unbalanced label distribution with only five types of labels (b) Unbalanced label distribution emphasizes two types of labels

Fig. 4. Two label unbalanced p^*

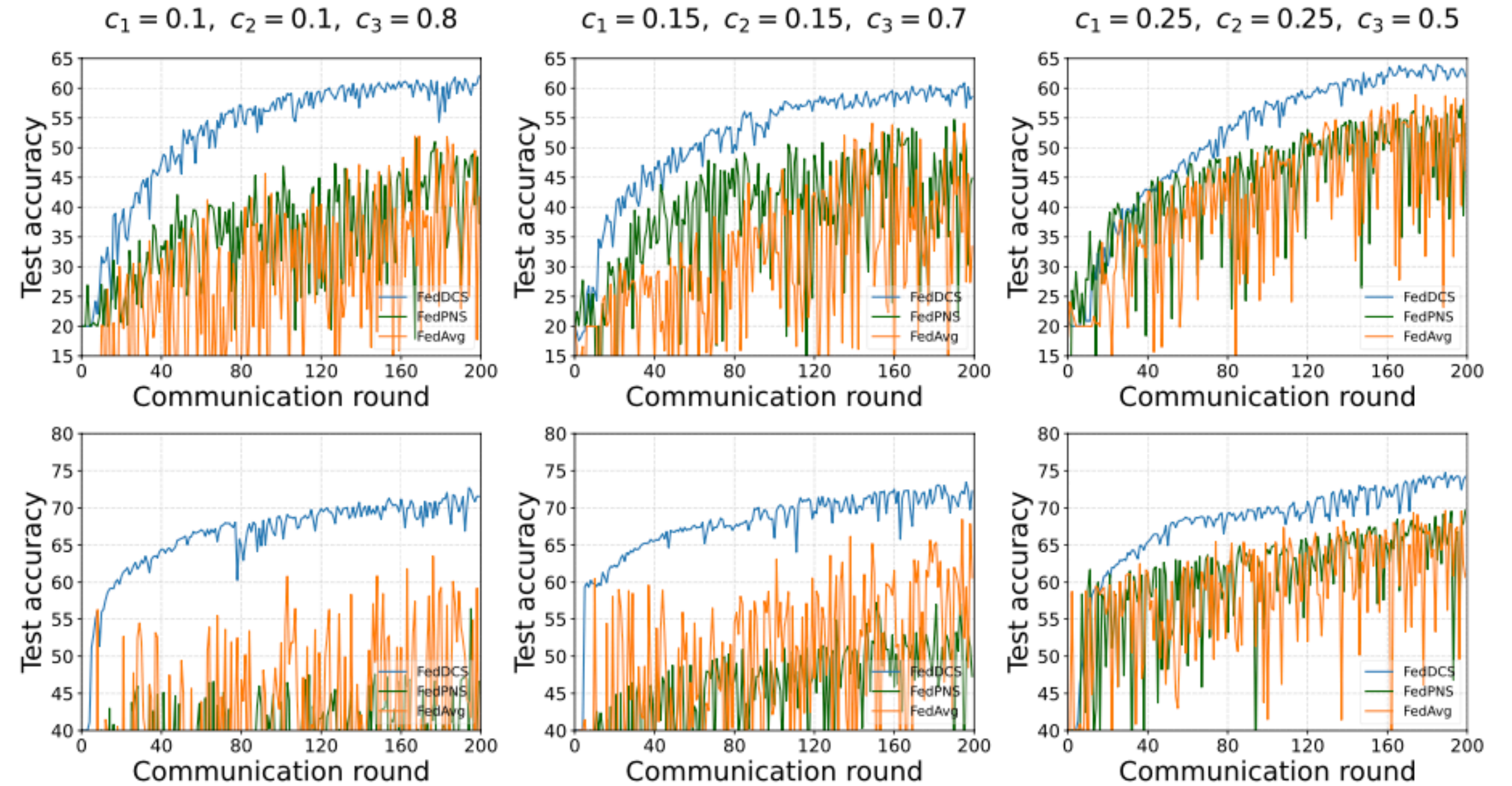


Fig. 5. Performance comparison under two label-unbalanced p^* . (1) Upper: Test accuracy over communication rounds under p^* as show in Fig. 4(a). (2) Bottom: Test accuracy over communication rounds under p^* as show in Fig. 4(b).

FedDCS: Experiments

Impact of threshold δ

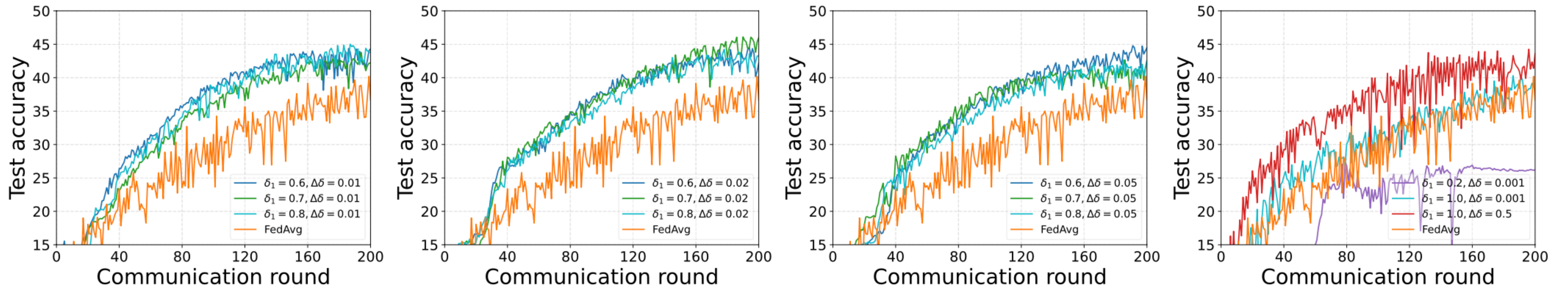


Fig. 6. The effect of different δ_1 and $\Delta\delta$ on the performance of FedDCS. In the first three plots, we fix $\Delta\delta = 0.01, 0.02, 0.05$ respectively and increase δ_1 from 0.6 to 0.8. In the last plot, we study the cases where $\Delta\delta$ is too large or too small.

FedDCS: Contributions and Conclusions



1. Define the label distribution shift tailored to specific FL tasks and propose a dynamic estimation method to capture the relevance of client label distribution to the FL task.
2. Design an adaptive client selection scheme prioritizing clients whose label distribution aligns closely with the FL task, ensuring the convergence of the global model in the desired direction.
3. By conducting extensive experiments, we proved that FedDCS could outperform competing FL methods (FedPNS, FedAvg etc.) in label distribution shift scenarios.

Thank you for your listening!

Q&A