Re-thinking Federated Active Learning based on Inter-class Diversity

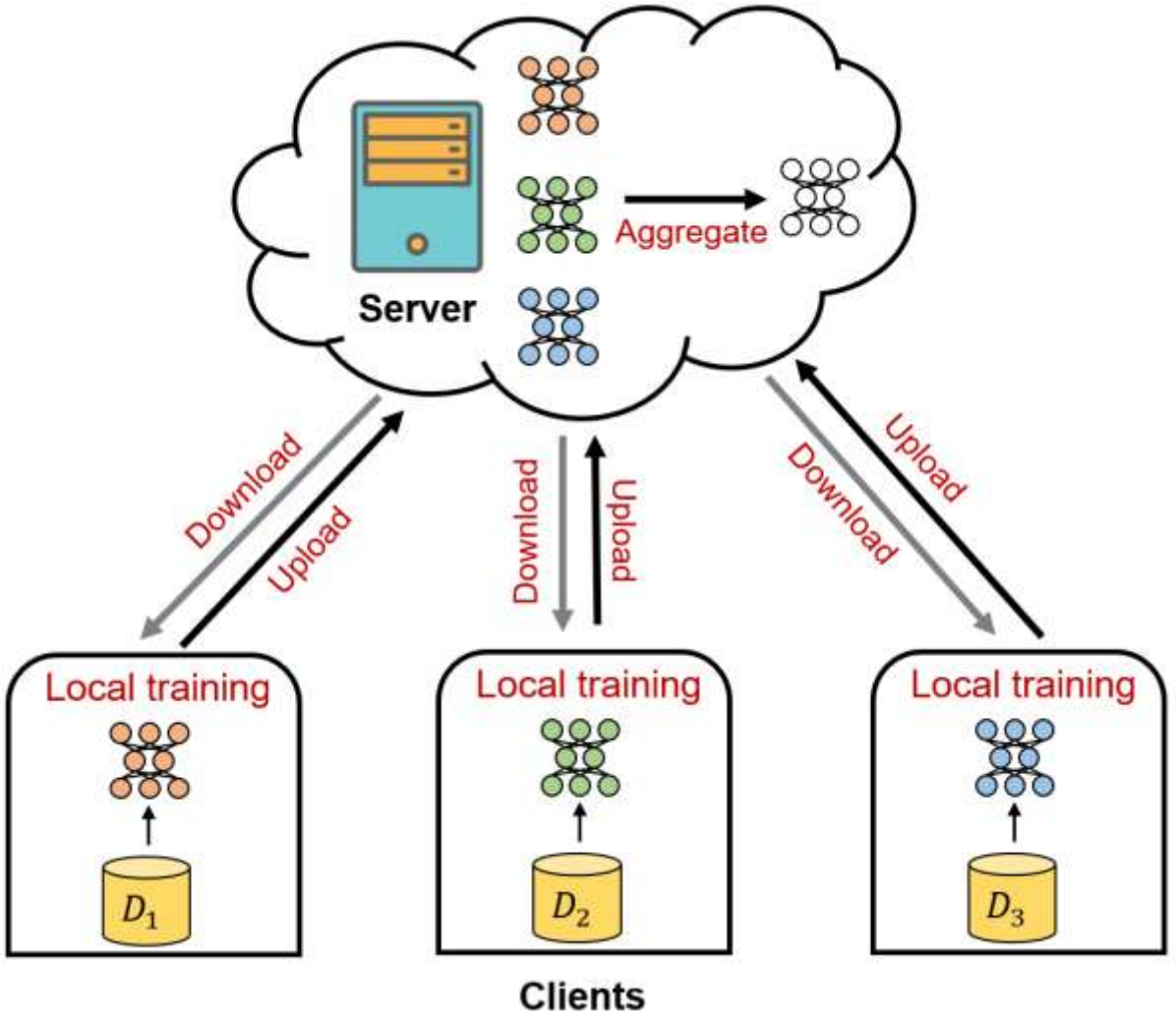
SangMook Kim[1*] Sangmin Bae[1*] Hwanjun Song[2†] Se-Young Yun[1†]

[1]KAIST AI [2]NAVER AI LAB

{sangmook.kim, bsmn0223, yunseyoung}@kaist.ac.kr ghkswns91@gmail.com

CVPR 2023

Background: Federated Learning



Goal: enable collaborative modeling while ensuring data privacy

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:

initialize w_0

for each round $t = 1, 2, \dots$ **do**

$m \leftarrow \max(C \cdot K, 1)$

$S_t \leftarrow$ (random set of m clients)

for each client $k \in S_t$ **in parallel do**

$w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$

$m_t \leftarrow \sum_{k \in S_t} n_k$

$w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$ *// Erratum*[4]

ClientUpdate(k, w): *// Run on client k*

$\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)

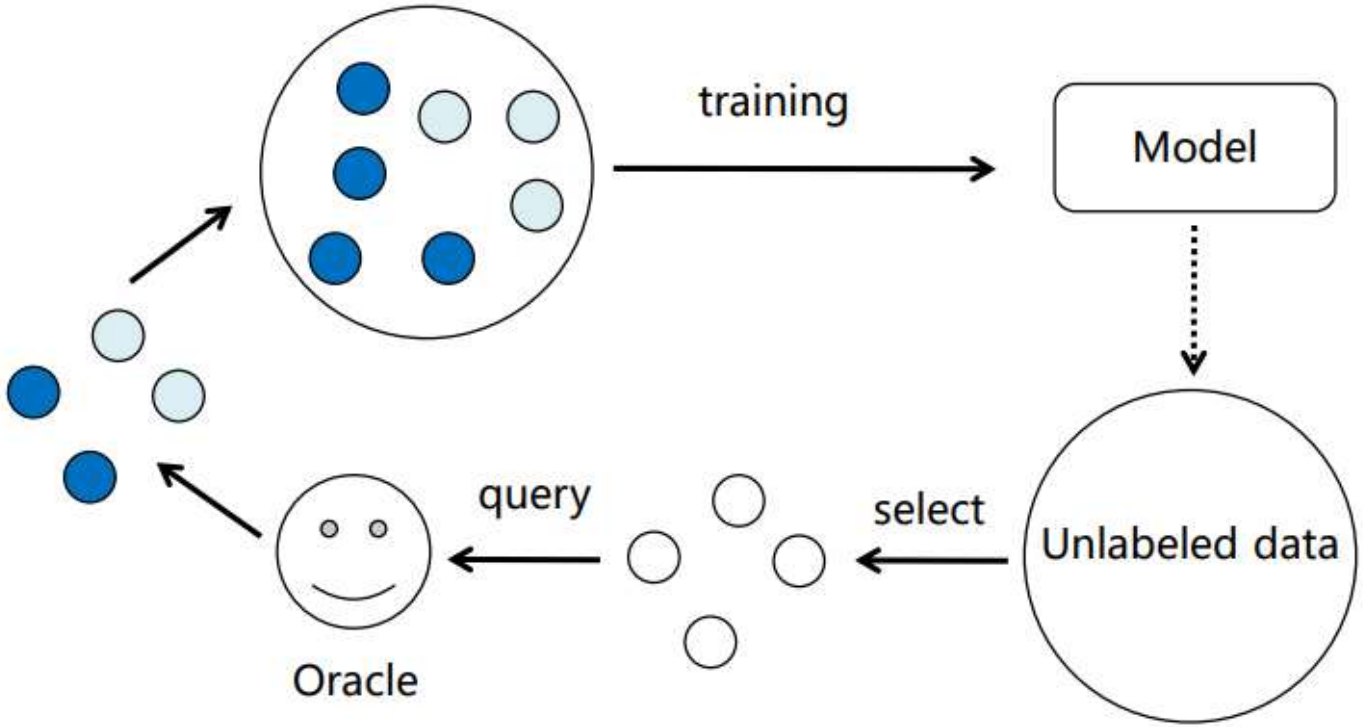
for each local epoch i from 1 to E **do**

for batch $b \in \mathcal{B}$ **do**

$w \leftarrow w - \eta \nabla \ell(w; b)$

return w to server

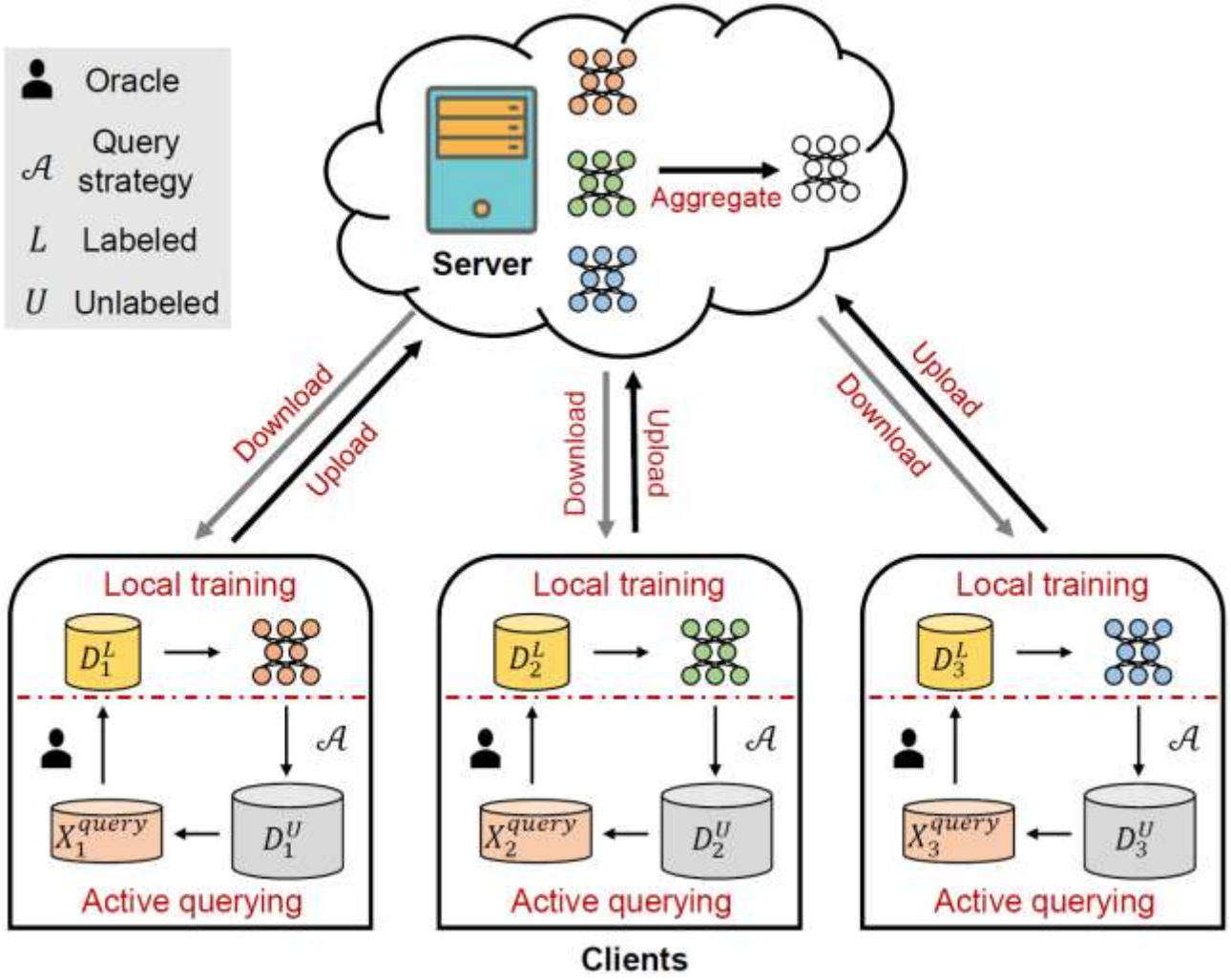
Background: Active Learning



- Uncertainty-based sampling
  - Least-confidence
  - Margin
  - Entropy
- Diversity-based sampling
  - CoreSet
- Hybird sampling
  - BADGE

Goal: query less for more.

Motivation



- Data heterogeneity:
  - The active selection algorithm has to ensure inter-class diversity from both local and global perspectives.
- Two available query-selecting models:
  - A global model, which is globally optimized through the FL pipeline.
  - A local-only model, which can be separately trained only for each client.

α: local heterogeneity level
ρ: global imbalance ratio

Observation

CIFAR-10 $\alpha \in \{0.1, 1.0, \infty\}$ $\rho \in \{1, 5, 10, 20\}$ Entropy Four seeds {1, 2, 3, 4}

$$a_r = \{a_{r,1}, \ldots, a_{r,4}\}$$

Definition 1. [36] *Let a_r^i and a_r^j be the set of accuracies for two different FAL strategies i and j. Then, t-score at AL round r is formulated as:*

$$t_r^{ij} = \frac{\sqrt{4}\mu_r^{ij}}{\sigma_r^{ij}}, \quad \text{where } \mu_r^{ij} = \frac{1}{4}\sum_{l=1}^{4}\left(a_{r,l}^i - a_{r,l}^j\right)$$
$$\text{and } \sigma_r^{ij} = \sqrt{\frac{1}{3}\sum_{l=1}^{4}\left(\left(a_{r,l}^i - a_{r,l}^j\right) - \mu_r^{ij}\right)}. \tag{5}$$

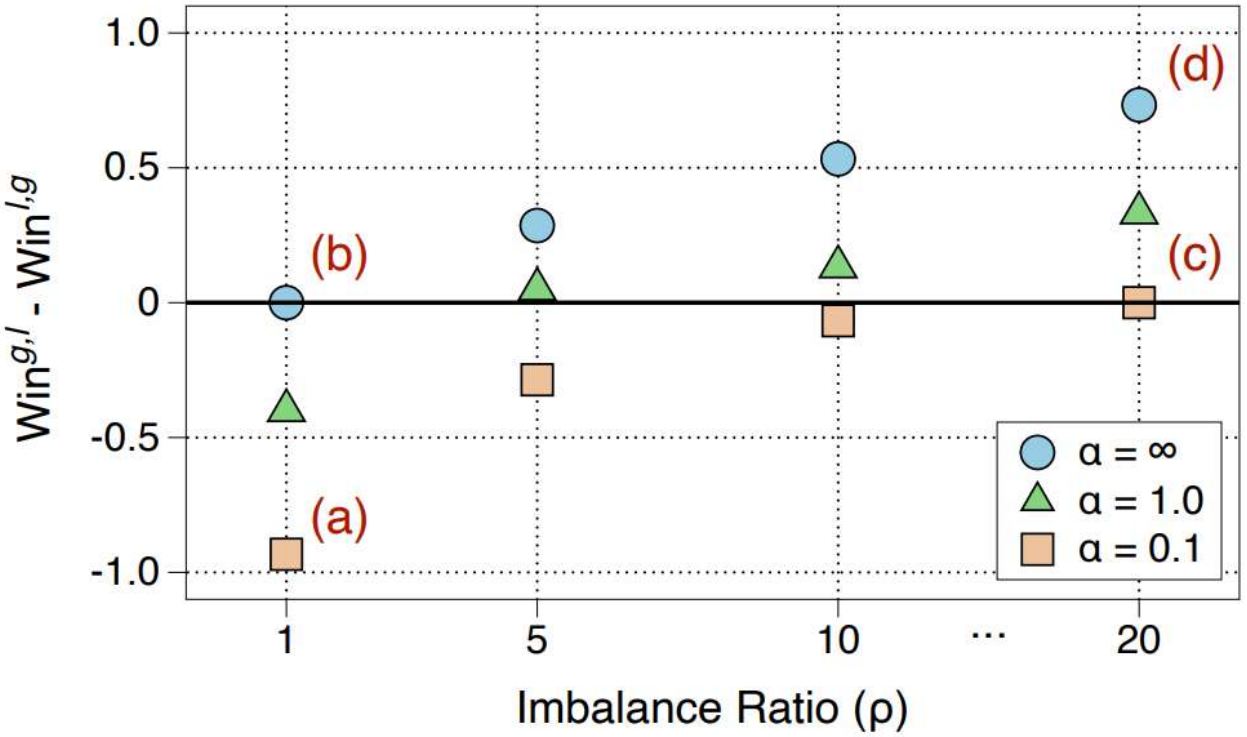
Here, the strategy i is considered to beat the strategy j if $t_r^{ij} > 2.776$. Therefore, the *winning rate* for all AL rounds is formulated as follows:

$$\text{win}^{ij} = \sum_{r=1}^{R}\frac{1}{R}\mathbb{1}_{t_r^{ij} > 2.776}. \tag{6}$$

The value of winning rate becomes 1 if the strategy i beats the strategy j over all AL rounds.

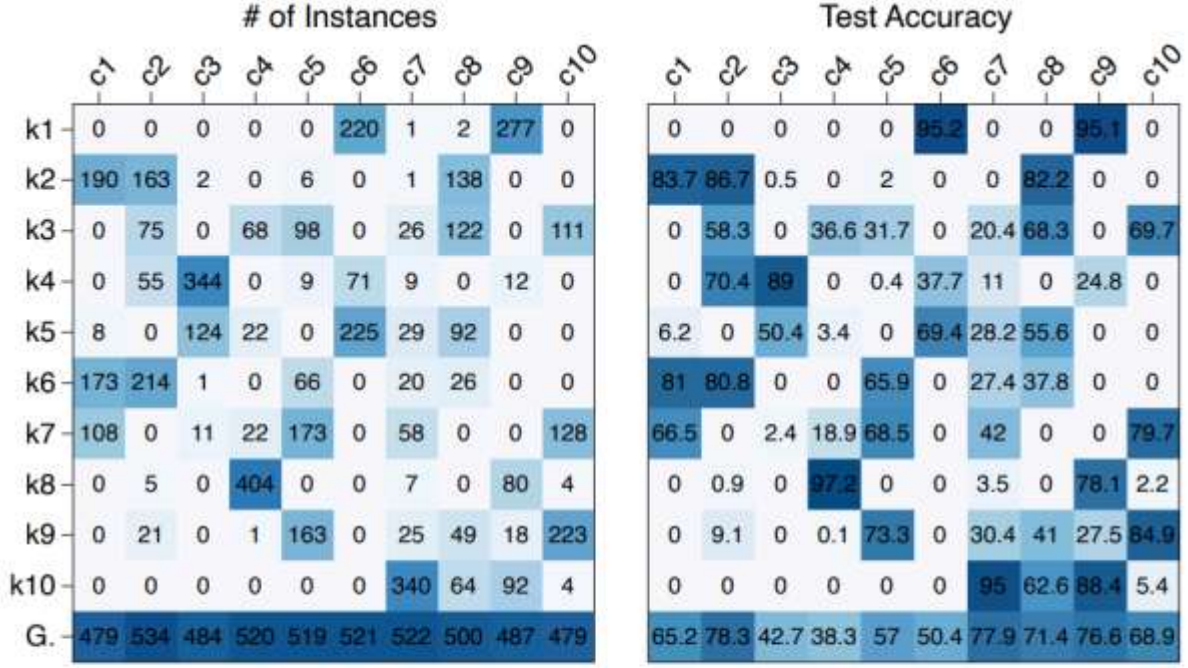
Observation

1. The superiority of local-only and global query-selecting models varies according to the degree of local heterogeneity and global imbalance ratio.

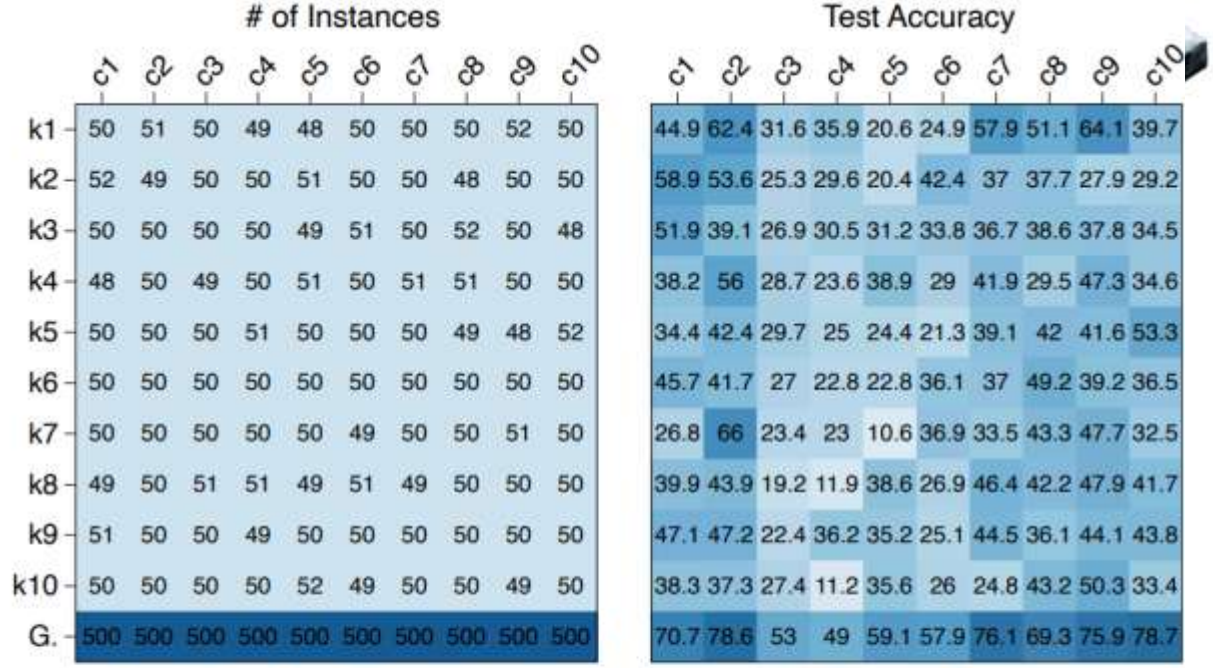


2. As local heterogeneity increases (α ↓), a local-only query selector is preferred due to the increased significance of local inter-class diversity.

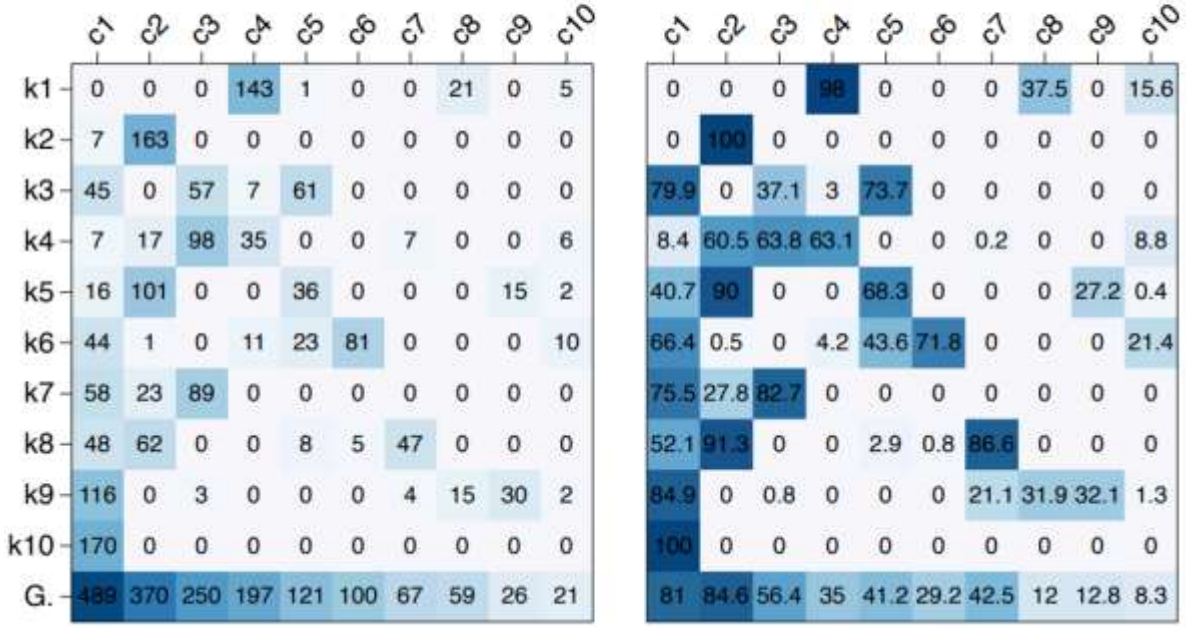
3. As the degree of global class imbalance increases (ρ ↑), it is more advantageous to exploit a global model that alleviates the global class imbalance.



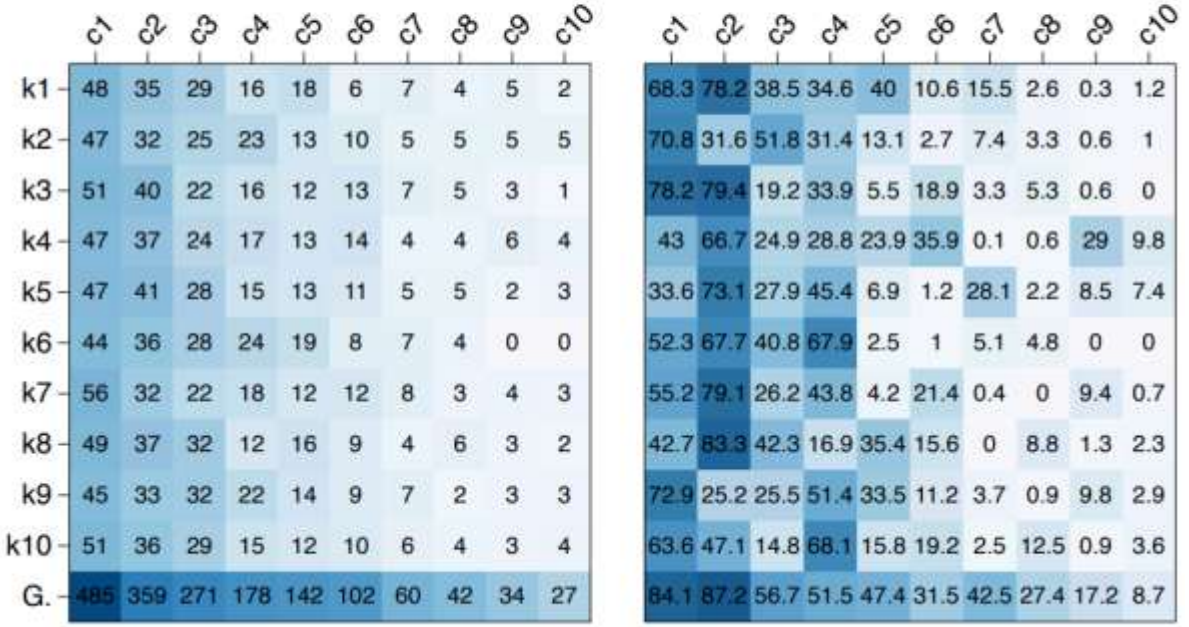
(a) Low global imbalance ($\rho = 1$) and high heterogeneity ($\alpha = 0.1$).



(b) Low global imbalance ($\rho = 1$) and low heterogeneity ($\alpha = \infty$).

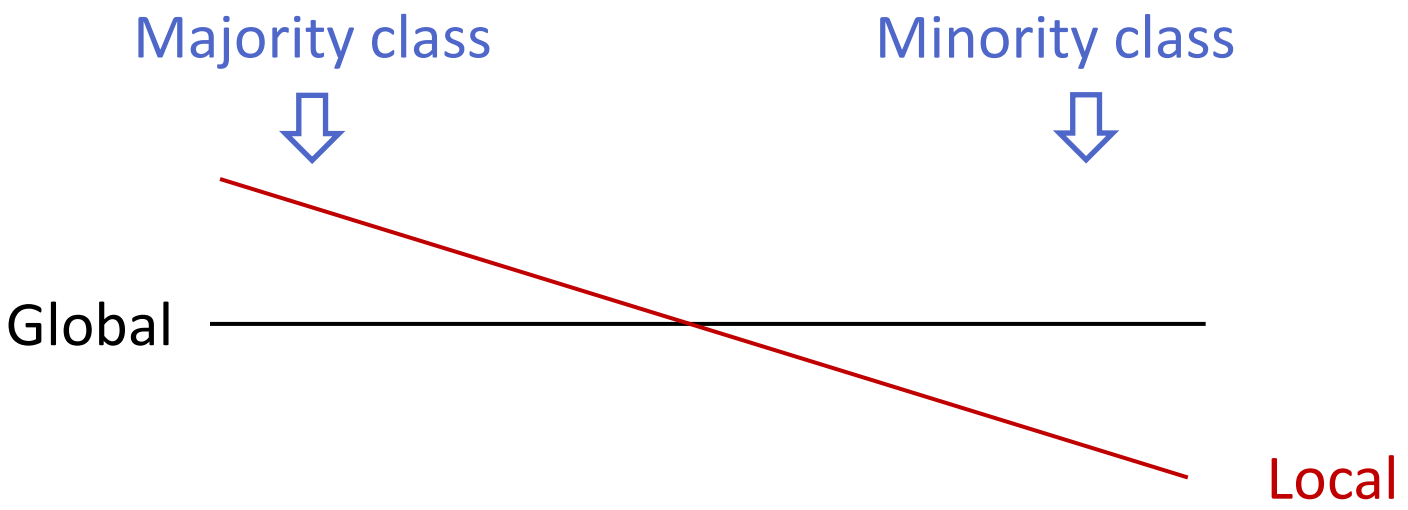


(c) High global imbalance ($\rho = 20$) and high heterogeneity ($\alpha = 0.1$).



(d) High global imbalance ($\rho = 20$) and low heterogeneity ($\alpha = \infty$).

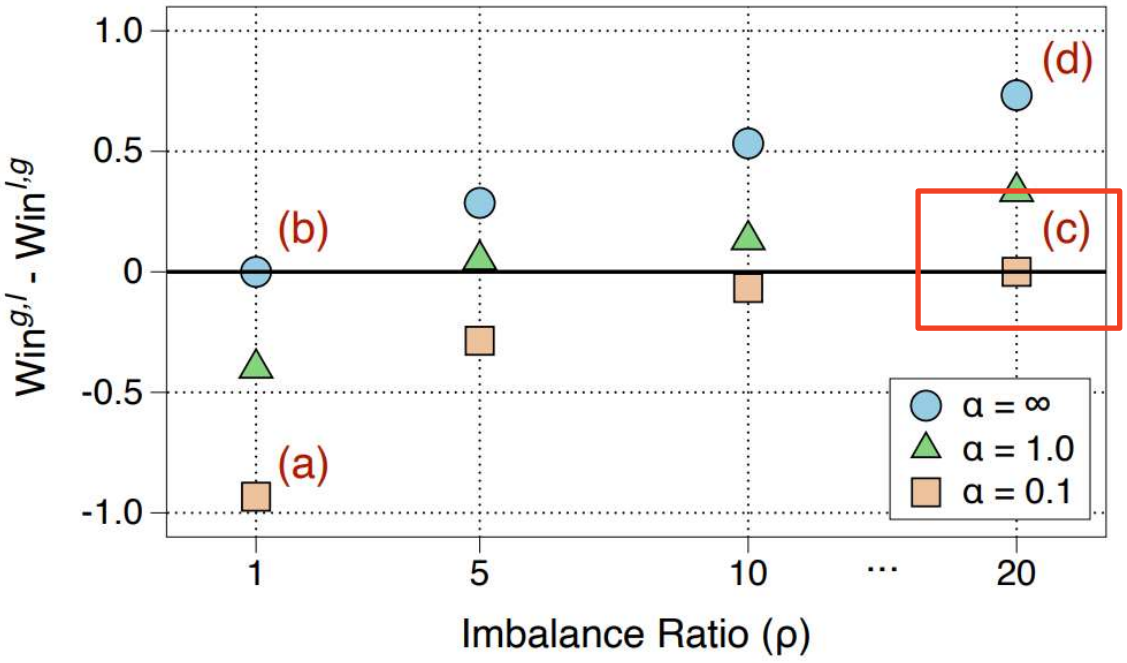
Observation



Observation

- As local heterogeneity increases (α ↓), a local-only query selector is preferred due to the increased significance of local inter-class diversity.

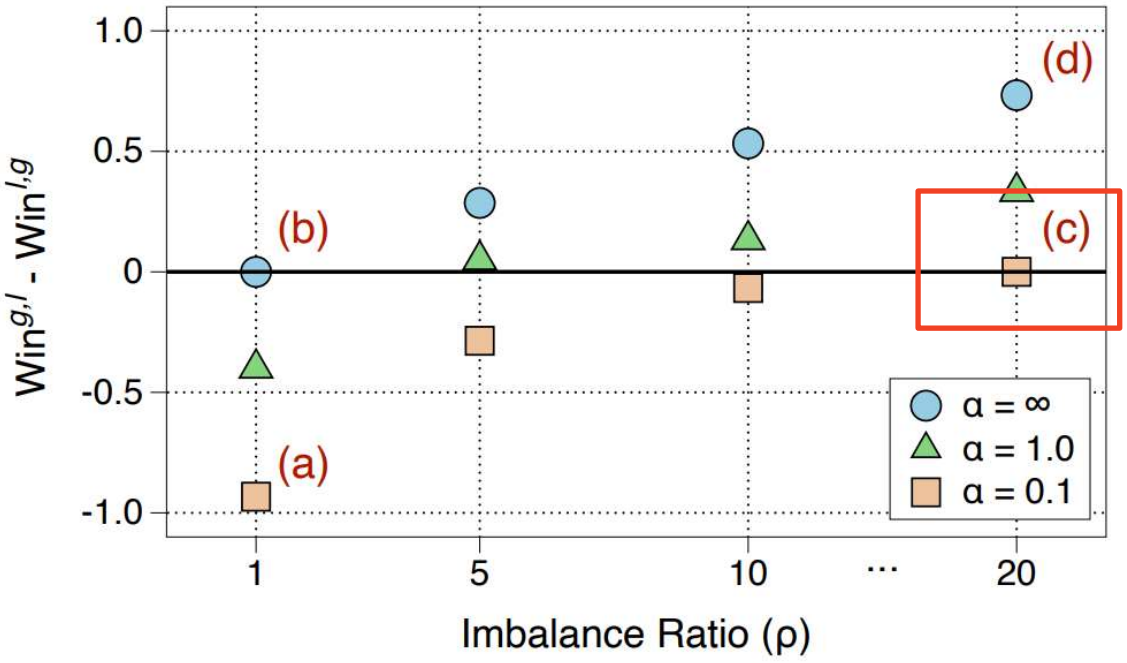
Case	Model	Obs. 2: Local EMD (↓)			
		10%	20%	30%	40%
(a)	G	0.632	0.638	0.641	0.643
	L	0.632	0.597	0.592	0.595
(b)	G	0.049	0.077	0.070	0.084
	L	0.049	0.042	0.054	0.059
(c)	G	0.692	0.680	0.676	0.674
	L	0.692	0.641	0.633	0.636
(d)	G	0.371	0.298	0.284	0.274
	L	0.371	0.313	0.293	0.290



Observation

- As the degree of global class imbalance increases (ρ ↑), it is more advantageous to exploit a global model that alleviates the global class imbalance.

Case	Model	Obs. 3: Global EMD (↓)			
		10%	20%	30%	40%
(a)	G	0.019	0.064	0.086	0.095
	L	0.019	0.050	0.050	0.046
(b)	G	0.014	0.070	0.066	0.063
	L	0.014	0.025	0.044	0.053
(c)	G	0.377	0.300	0.294	0.294
	L	0.377	0.334	0.326	0.321
(d)	G	0.368	0.294	0.282	0.272
	L	0.368	0.309	0.287	0.288



LoGo

Algorithm 1 FAL framework with LoGo algorithm

Input: initialized parameter Θ; unlabeled data U^1; sampling strategy $\mathcal{A}$; labeling budget B; clients number K; AL round R;
Output: trained parameter Θ^{R*}

 # Alternating AL and FL Procedure

for $k = 1, \ldots, K$ **do**
    Randomly sample $L_k^1 = \{x_1, \ldots, x_B\}$ from U_k^1, and $U_k^2 = U_k^1 \setminus L_k^1$
    Get the labeled set D_k^1 from the oracles
end for
$\Theta^{1*} = \texttt{FedAvg}(\Theta, D^1, K)$

for $r = 2, \ldots, R$ **do**
    for $k = 1, \ldots, K$ **do**
        $D_k^r, U_k^{r+1} = \texttt{LoGo}(\Theta^{(r-1)*}, D_k^{r-1}, U_k^r)$
    end for
    $\Theta^{r*} = \texttt{FedAvg}(\Theta, D^r, K)$
end for

LoGo

Function `LoGo`:

 # Macro Step
Train a local-only model $\Theta_{k*}^{(r-1)}$ from the scratch only using D_k^{r-1}
For each $x \in U_k^r$, calculate the gradient embedding $g_{\hat{y}}^x$ by Eq. (7)
Cluster U_k^r into B clusters$(\mathcal{C}_1, ..., \mathcal{C}_B)$ by Eq. (8)

 # Micro Step
$L_k^r = \emptyset$
for $\mathcal{C}_i = \mathcal{C}_1, \ldots, \mathcal{C}_B$ **do**
    $L_k^r = L_k^r \cup \{\mathcal{A}(\mathcal{C}_i, \Theta^{(r-1)*}, 1)\}$
    $D_k^r = D_k^{r-1} \cup D_k^r$ and $U_k^{r+1} = U_k^r \setminus L_k^r$
end for
return D_k^r, U_k^{r+1}

$$g_c^x = -\frac{\partial}{\partial W_c}\ell_{CE}(x, \hat{y}; \Theta_{k*}^r) = z \cdot (\mathbb{1}_{[\hat{y}=c]} - p_c),$$
$$\hat{y} = \arg\max_{c \in [C]} p_c$$

Function `FedAvg`:

for $FL\ round$ **do**
    Distribute Θ to the all client
    for $k = 1, \ldots, K$ **do**
        Train Θ_k on D_k^r by minimizing $\mathbb{E}_{D_k^r}[\ell(x, y; \Theta_k)]$
    end for
    $\Theta = (\sum_k \Theta_k)/K$
end for
return Θ

Experiments

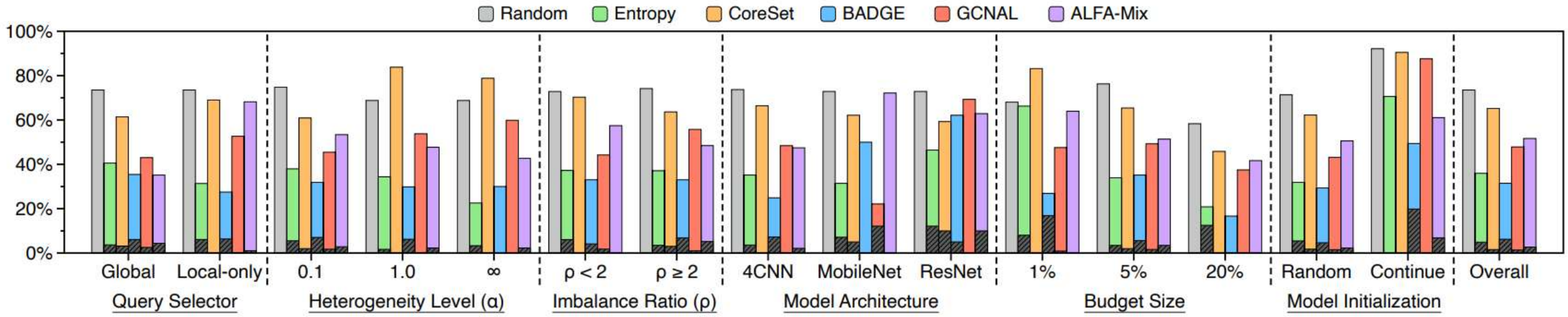
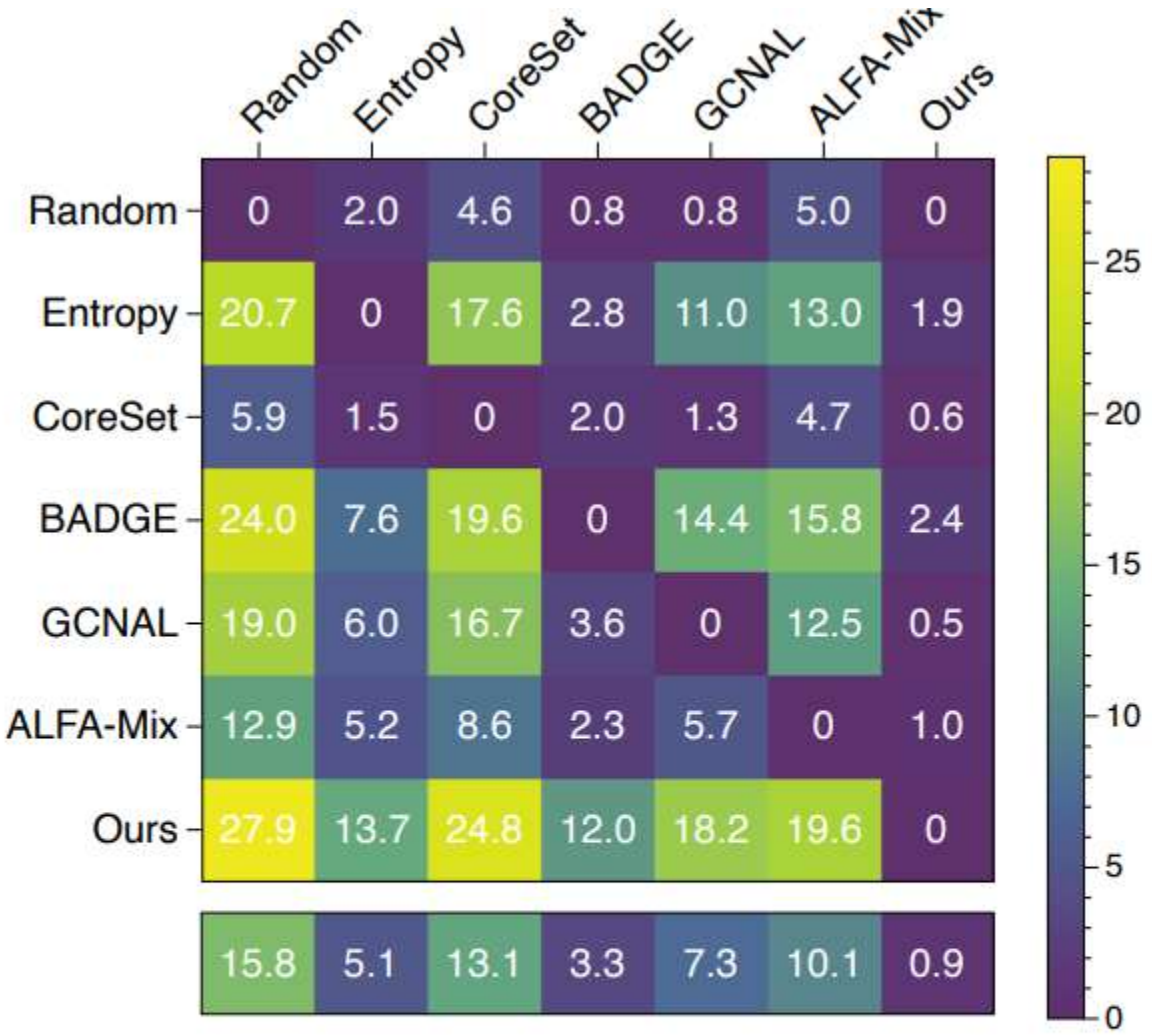


Figure 4. Winning percentage across six categories. We also added defeat percentage, the black hatched bar that represents the percentage at which LoGo has been defeated by each baseline. Among total experiments, only statistically reliable values (t-score > 2.776) are considered. Thus, the lower value of the colored bar and the higher value of the black bar indicate a more comparable baseline.

$$\text{win}^{ij} = \sum_{r=1}^{R} \frac{1}{R} \mathbb{1}_{t_r^{ij} > 2.776}$$

Experiments



$$\text{win}^{ij} = \sum_{r=1}^{R} \frac{1}{R} \mathbb{1}_{t_r^{ij} > 2.776}$$

Figure 5. Pairwise penalty matrix over 38 experimental settings. The value P_{ij} indicates the number of times that the i-th strategy outperforms the j-th strategy (*i.e.*, sum of win^{ij} in Eq. (6) over 38 settings). The last row is the average number of times the j-th strategy is defeated by the rest strategies; the lower, the better.

Experiments

Method	Model	CIFAR-10				SVHN				PathMNIST				DermaMNIST			
		20%	40%	60%	80%	20%	30%	40%	50%	20%	30%	40%	50%	20%	30%	40%	50%
Random	-	64.19	69.07	71.63	72.81	80.90	83.07	84.22	84.77	68.41	72.70	73.76	75.49	71.70	72.57	72.66	72.86
Entropy [41]	*G*	64.02	69.12	71.87	73.33	82.08	84.61	85.88	86.31	71.54	74.39	75.91	76.65	72.49	72.63	73.02	73.20
	L	66.29	71.45	73.51	74.02	82.09	84.58	85.69	86.18	**76.52**	78.29	78.71	79.10	71.38	72.04	72.22	72.65
Coreset [37]	*G*	64.66	69.43	71.75	73.1	80.94	82.74	83.81	84.46	74.84	76.24	76.85	76.80	72.02	72.16	72.34	72.74
	L	64.06	68.79	71.49	73.28	80.94	82.92	83.78	84.48	72.53	76.06	76.28	76.86	71.13	71.48	72.15	72.38
BADGE [6]	*G*	65.12	69.57	72.11	73.53	82.81	84.82	85.89	86.2	72.21	74.38	75.53	76.97	72.59	73.09	73.23	73.45
	L	66.32	71.28	73.41	74.28	82.69	84.67	85.61	86.1	76.48	78.51	78.42	78.68	71.35	72.13	72.25	72.99
GCNAL [8]	*G*	65.40	70.05	72.41	73.42	82.05	84.07	85.09	85.61	75.51	77.79	78.13	78.81	72.01	72.60	73.07	73.17
	L	65.62	70.18	72.36	73.42	81.92	83.58	84.55	85.10	74.85	76.46	77.18	77.45	71.95	72.91	72.91	73.29
ALFA-Mix [34]	*G*	65.45	69.87	72.24	73.29	83.02	84.99	**86.05**	86.33	73.34	74.83	76.31	77.43	72.39	73.14	73.27	73.10
	L	64.14	68.79	71.03	72.6	81.08	82.55	83.62	84.33	71.10	75.01	75.81	76.70	71.51	72.18	72.94	73.28
LoGo (ours)	*G*, *L*	**66.50**	**71.70**	**73.80**	**74.49**	**83.46**	**85.31**	86.02	**86.38**	76.32	**78.72**	**79.51**	**79.58**	**72.61**	**73.18**	**73.33**	**73.77**

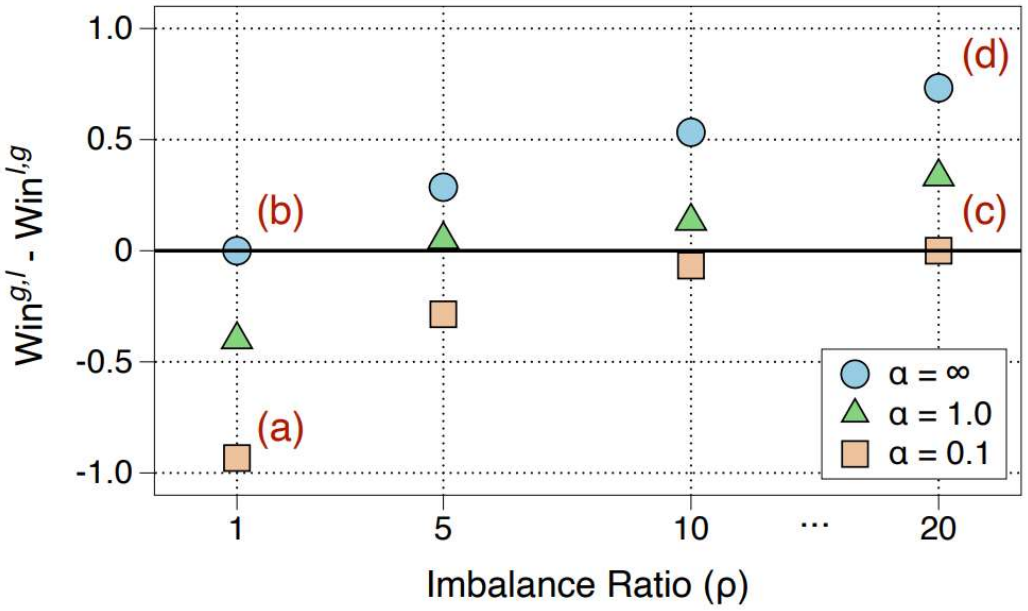
Table 3. Comparison of test accuracy on four benchmarks with $\alpha = 0.1$. We reported the results with four random seeds. The baselines, except for Random sampling, are combined with two query selector models, G and L that stands for a global or local-only model, respectively. **Bold** and underline mean Top-1 and Top-2, respectively.

Experiments

Method	Strategy	CIFAR-10				SVHN		
		20%	40%	60%	80%	20%	30%	40%
Ens. Logit	+Entropy	64.53	70.36	73.02	74.28	81.81	84.64	85.87
	+BADGE	65.55	70.31	72.83	73.97	82.77	84.76	85.90
Ens. Rank	+Entropy	65.90	70.92	73.34	74.20	82.15	84.38	85.64
	+BADGE	66.21	70.98	73.15	74.01	83.02	85.05	85.86
Fine-tuning	+Entropy	65.10	70.75	73.21	74.23	82.53	85.05	86.01
	+BADGE	65.82	70.95	72.94	74.12	82.59	84.89	85.82
LoGo (ours)	-	**66.50**	**71.70**	**73.80**	**74.49**	**83.46**	**85.31**	**86.02**

Table 4. Comparison of test accuracy on two benchmarks (α=0.1) with baselines using both global and local information.

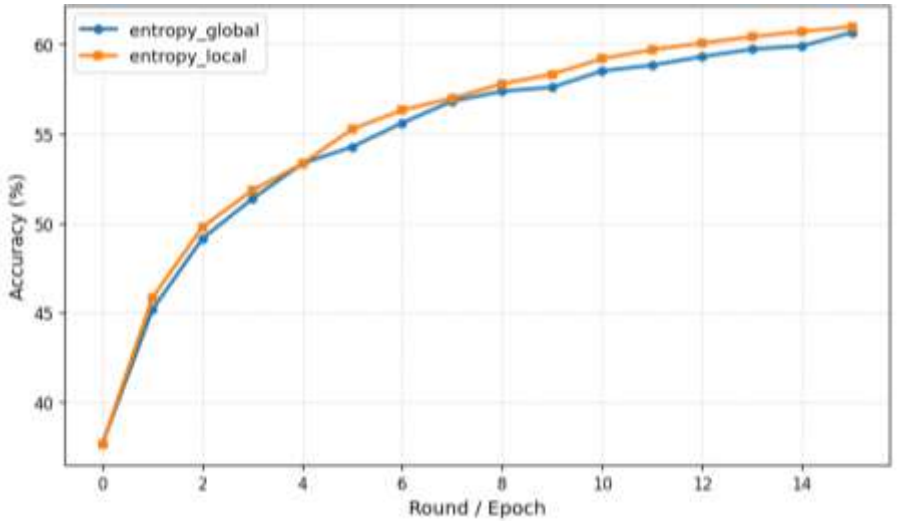
科研进展



Here, the strategy i is considered to beat the strategy j if $t_r^{ij} > 2.776$. Therefore, the *winning rate* for all AL rounds is formulated as follows:

$$\text{win}^{ij} = \sum_{r=1}^{R} \frac{1}{R} \mathbb{1}_{t_r^{ij} > 2.776}. \tag{6}$$

The value of winning rate becomes 1 if the strategy i beats the strategy j over all AL rounds.



Seed1 (80%)

Global:

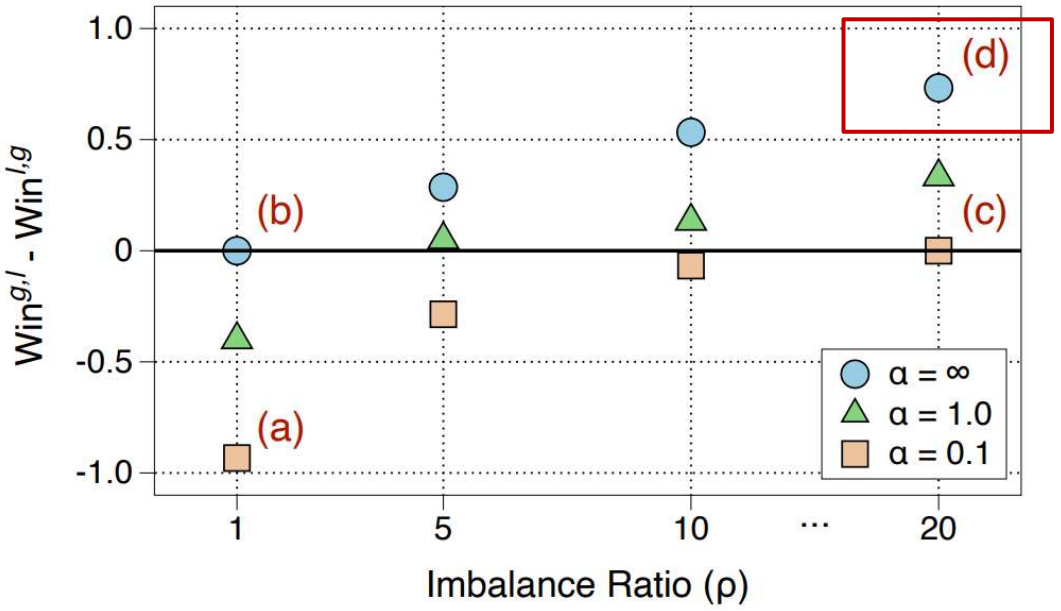
{0: 3041, 1: 2277, 2: 2008, 3: 1402, 4: 1004, 5: 816, 6: 549, 7: 325, 8: 300, 9: 178} 7326 3771 803

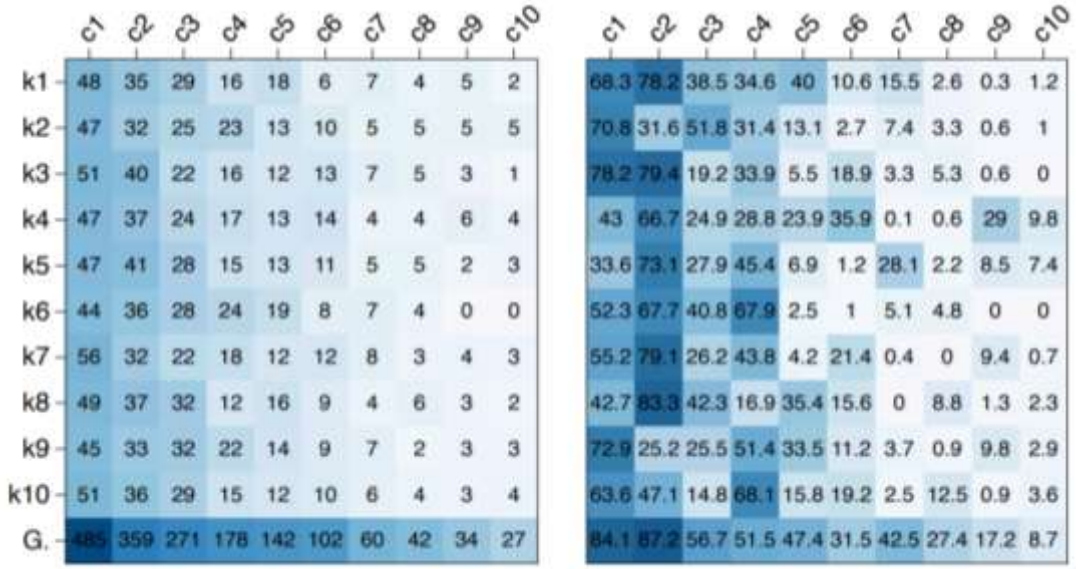
Local:

{0: 3410, 1: 2387, 2: 1705, 3: 1329, 4: 1009, 5: 694, 6: 498, 7: 381, 8: 292, 9: 195} 7502 3530 868

Seed1 (50%): 3870 2386 544 vs 4178 1989 633

科研进展





(d) High global imbalance ($\rho = 20$) and low heterogeneity ($\alpha = \infty$).

Seed1 (80%)
Global:
{0: 3180, 1: 1942, 2: 1941, 3: 1516, 4: 1090, 5: 798, 6: 552, 7: 373, 8: 296, 9: 212}
Local:
{0: 2880, 1: 2018, 2: 2076, 3: 1531, 4: 1147, 5: 805, 6: 567, 7: 422, 8: 266, 9: 188}

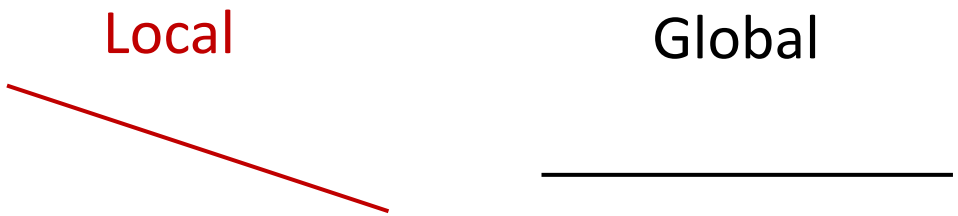
Seed1 (50%)
Global:
{0: 1464, 1: 889, 2: 1163, 3: 980, 4: 740, 5: 555, 6: 357, 7: 263, 8: 223, 9: 166}
Local:
{0: 1479, 1: 1024, 2: 1180, 3: 959, 4: 745, 5: 452, 6: 403, 7: 327, 8: 113, 9: 118}

Seed1 (25%)
Global: {0: 498, 1: 310, 2: 441, 3: 405, 4: 319, 5: 202, 6: 140, 7: 93, 8: 79, 9: 63}
Local: {0: 480, 1: 437, 2: 399, 3: 354, 4: 272, 5: 157, 6: 185, 7: 166, 8: 49, 9: 51}

科研进展

只有在**全局不平衡率**高且**本地与全局数据分布**高度重合时，用全局模型采样更好

- ✓ 如何衡量全局不平衡率？
- ✓ 如何衡量本地与全局数据分布差异？

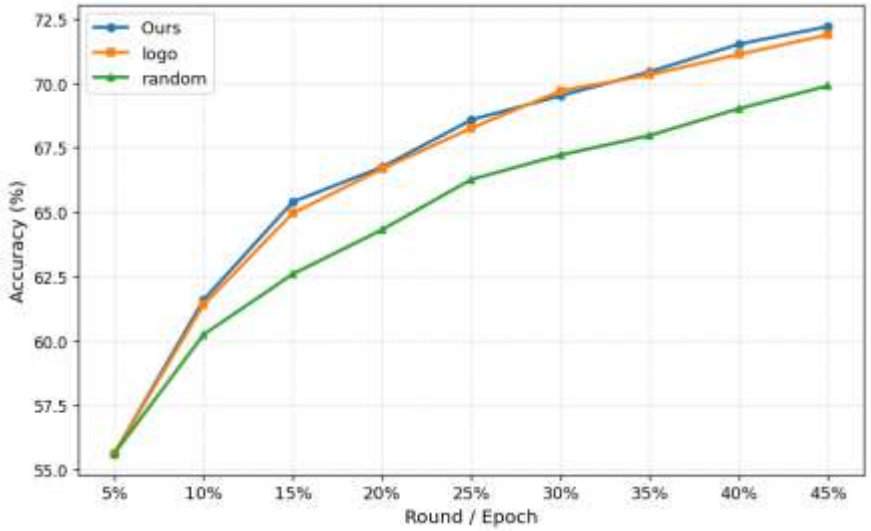


LoGo:	1123.74	1197.12	900.33	954.07
尝试1：	1120.86	1196.92	896.55	962.88
按类别：	1118.44	1198.26	907.04	970.88
+多样性：	1124.82	1199.98	910.67	978.81

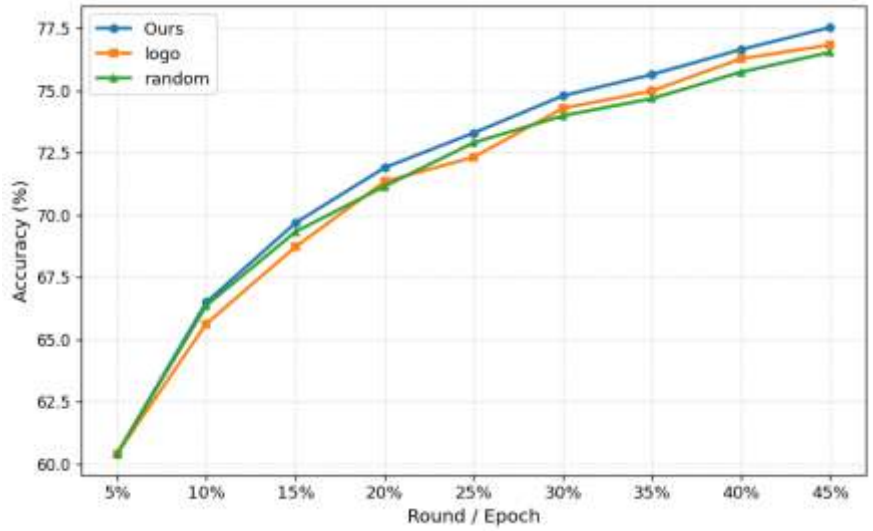
```
global_mean_probs = global_mean_probs.numpy()
local_mean_probs = local_mean_probs.numpy()

idx = np.fromiter(count_label.keys(), dtype=int)
global_mean_probs = global_mean_probs[idx]
local_mean_probs = local_mean_probs[idx]

distance = (np.abs(global_mean_probs - local_mean_probs)/np.maximum(global_mean_probs, local_mean_probs)).mean()

global_ratio = global_mean_probs.min() / global_mean_probs.max()
```

科研进展

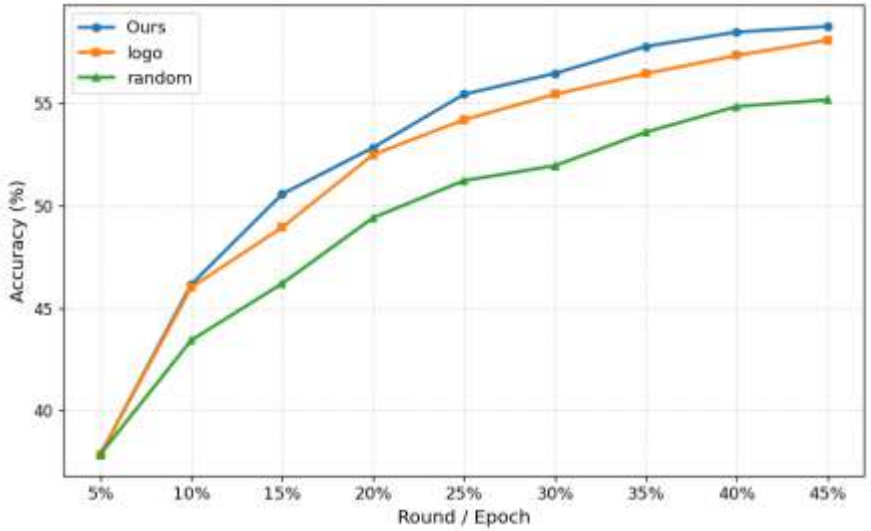
$\rho = 1$，$\alpha = 0.1$



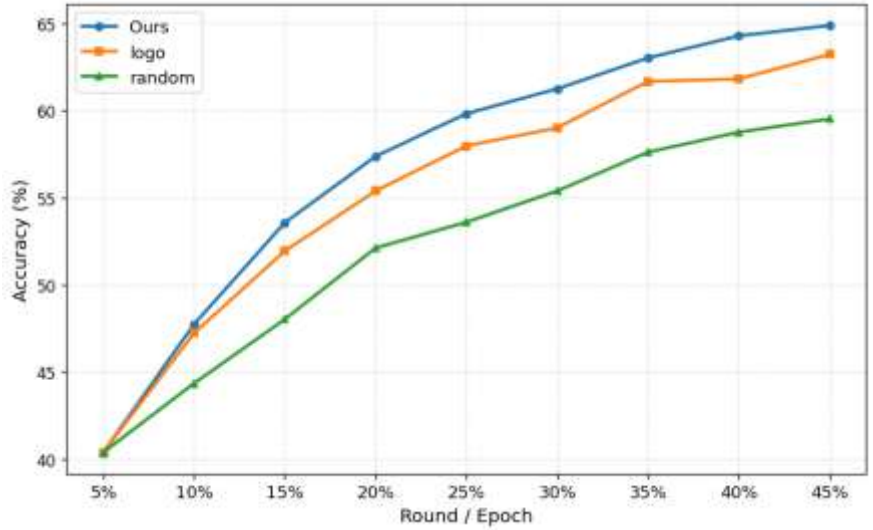
$\rho = 1$，$\alpha = 100$



$\rho = 20$，$\alpha = 0.1$



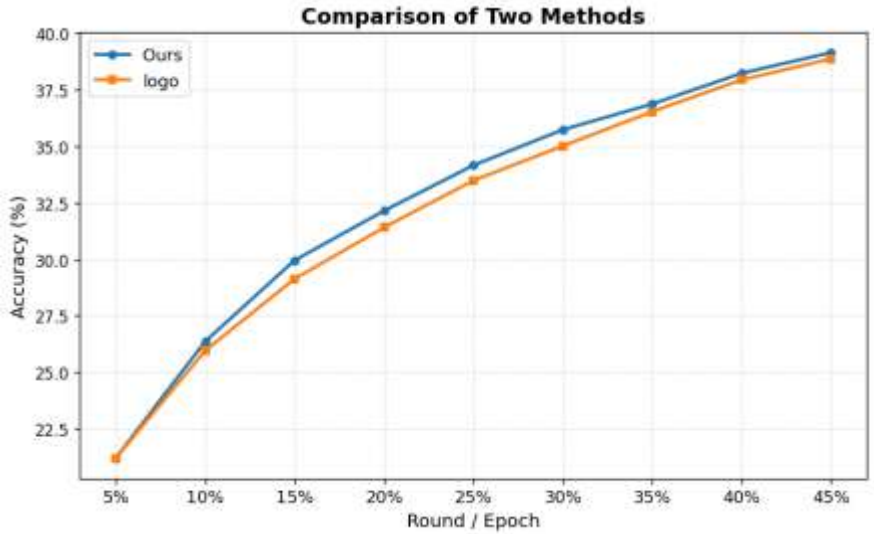
$\rho = 20$，$\alpha = 100$



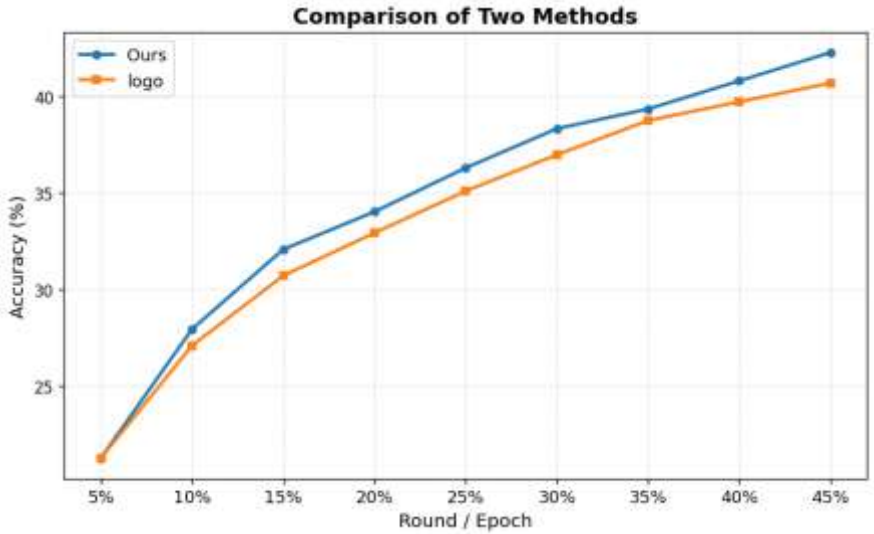
CIFAR-10

科研进展

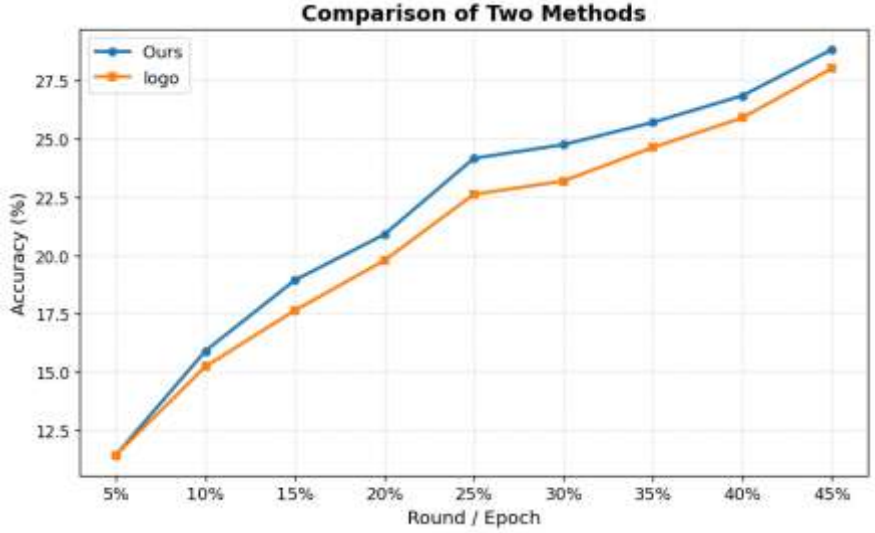
$\rho = 1$，$\alpha = 0.1$



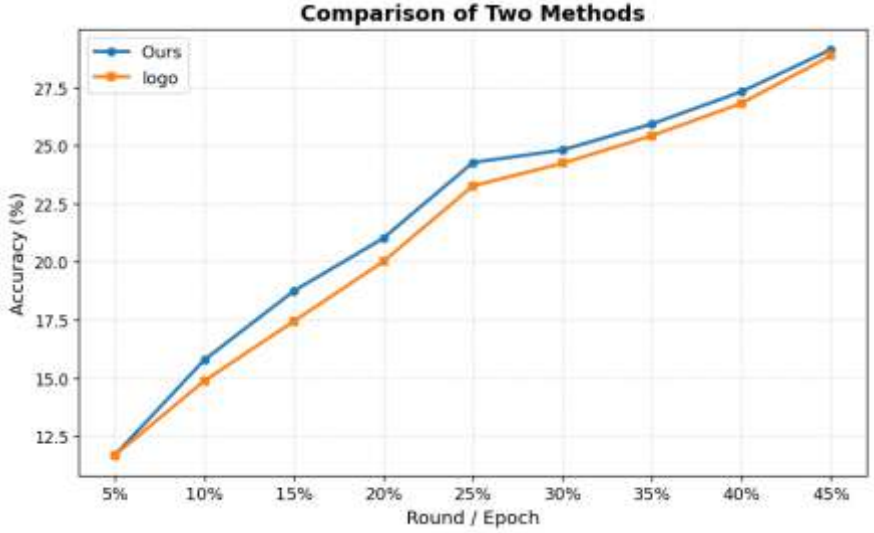
$\rho = 1$，$\alpha = 100$



$\rho = 20$，$\alpha = 0.1$



$\rho = 20$，$\alpha = 100$



CIFAR-100

THANKS