香港浸會大學
HONG KONG BAPTIST UNIVERSITY

Data Resampling for Federated Learning with Non-IID Labels

Zhenheng Tang[1] Zhikai Hu[1] Shaohuai Shi[2] Yiu-ming Cheung[1] Yilun Jin[2] Zhenghang Ren[2] Xiaowen Chu[1]

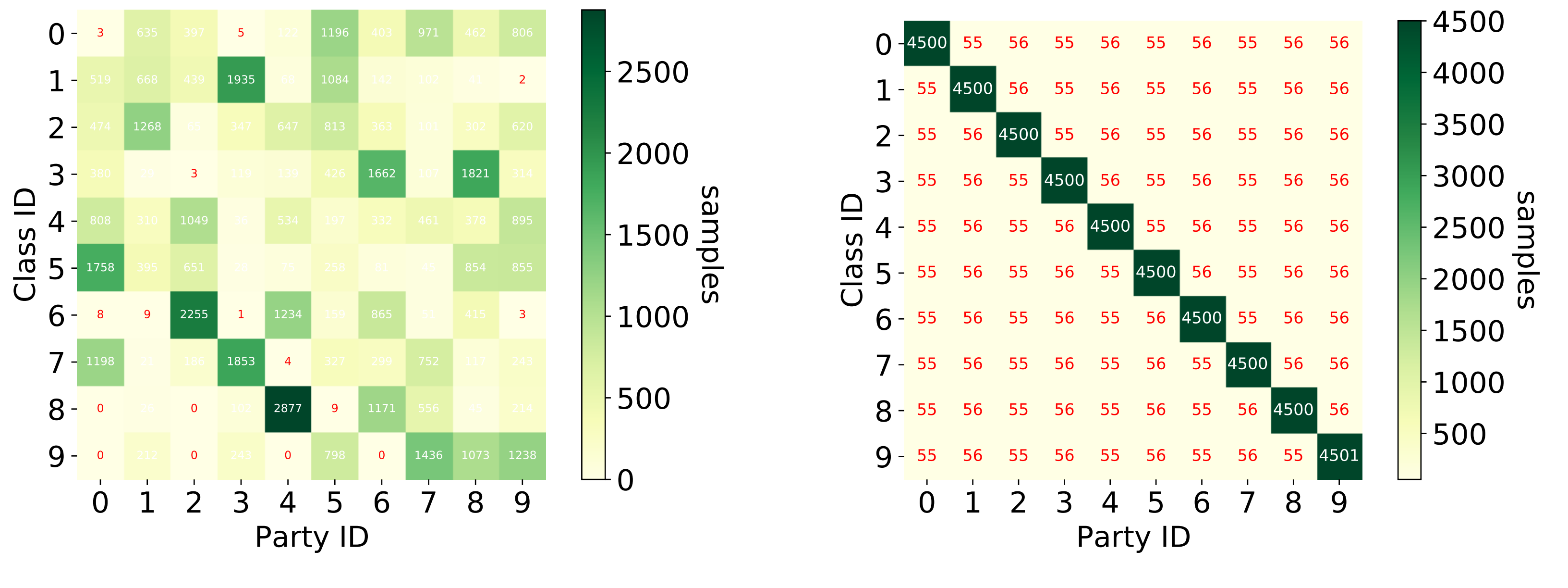
[1]Department of Computer Science, Hong Kong Baptist University
[2]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Introduction

Federated Learning performance suffers from Non-IID data [3][1]. We find that the learning process of one client (without communication) can be seen as the *imbalanced learning*.

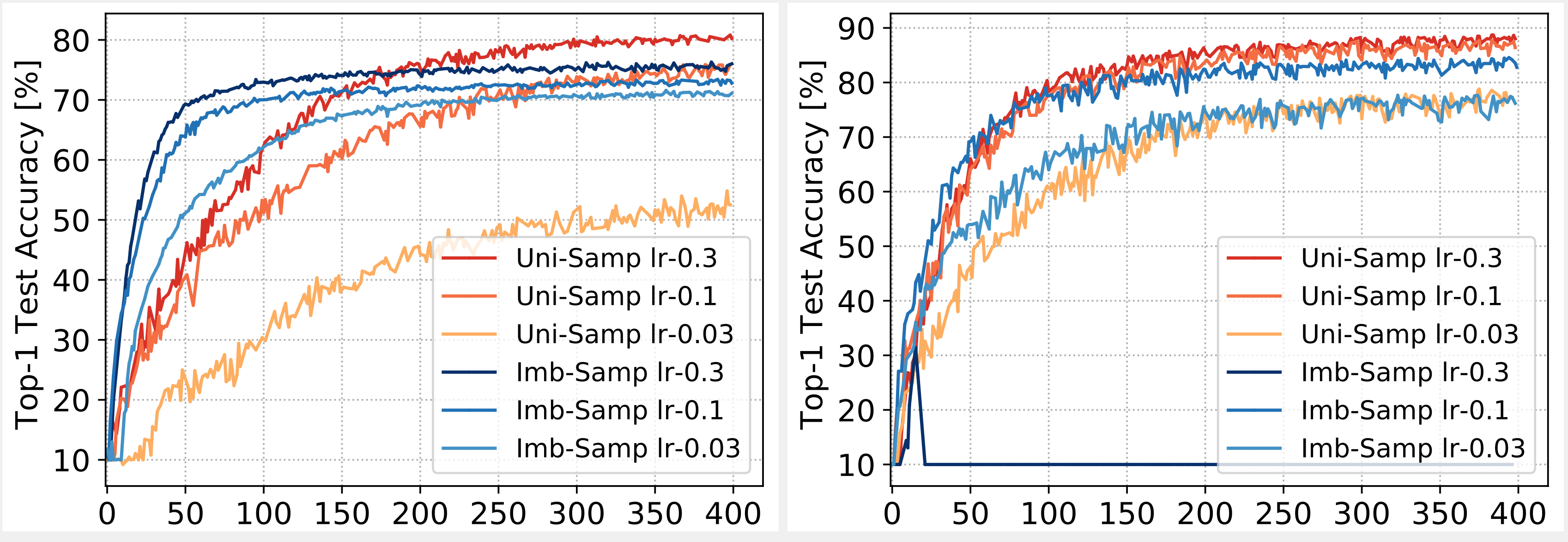
Intuitively, by balancing the sampling probability between labels, the label sampling probability between clients could become similar, thus the data distribution becomes label IID.



(a) LDA[1] partition (b) LLT partition

Figure 1. Visualization of two kinds of Non-IID distribution of CIFAR-10 on 10 clients.

Observation



(a) LLT partition with $\alpha_l = 0.9$ (b) LDA partition with $\alpha_d = 0.5$

Figure 2. FedAvg on CIFAR-10 dataset with 10 clients. Uni-Samp means each Uniform Sampling, and the Imb-Samp means Imbalance Sampling.

Assuming there are K clients with datasets $\mathcal{D}_0, \mathcal{D}_1, \cdots, \mathcal{D}_K$ respectively, and the amount of samples of label $c \in \{1, 2, \cdots, C\}$ in $\mathcal{D}_k$ is $N_{k,c}$, the label of i-th sample in client k is $\boldsymbol{Y}_{(k,i)}$. In our data imbalance sampling scheme, we have:

- Sampling weight of i-th sample in client k: $w_{k,i} = (1-\beta)/(1-\beta^{N_{k,\boldsymbol{Y}_{(k,i)}}})$.
- Sampling probability of i-th sample in client k: $p(k,i) = w_{k,i}/(\sum_i^{N_k} w_{k,i})$.
- Sampling probability ratio of label c_1 and c_2: $q(k,c_1)/q(k,c_2) = N_{k,c_1}p(k,i_1))/(N_{k,c_2}p(k,i_2)) = N_{k,c_1}(1-\beta^{N_{k,c_2}})/(N_{k,c_2}(1-\beta^{N_{k,c_1}}))$.

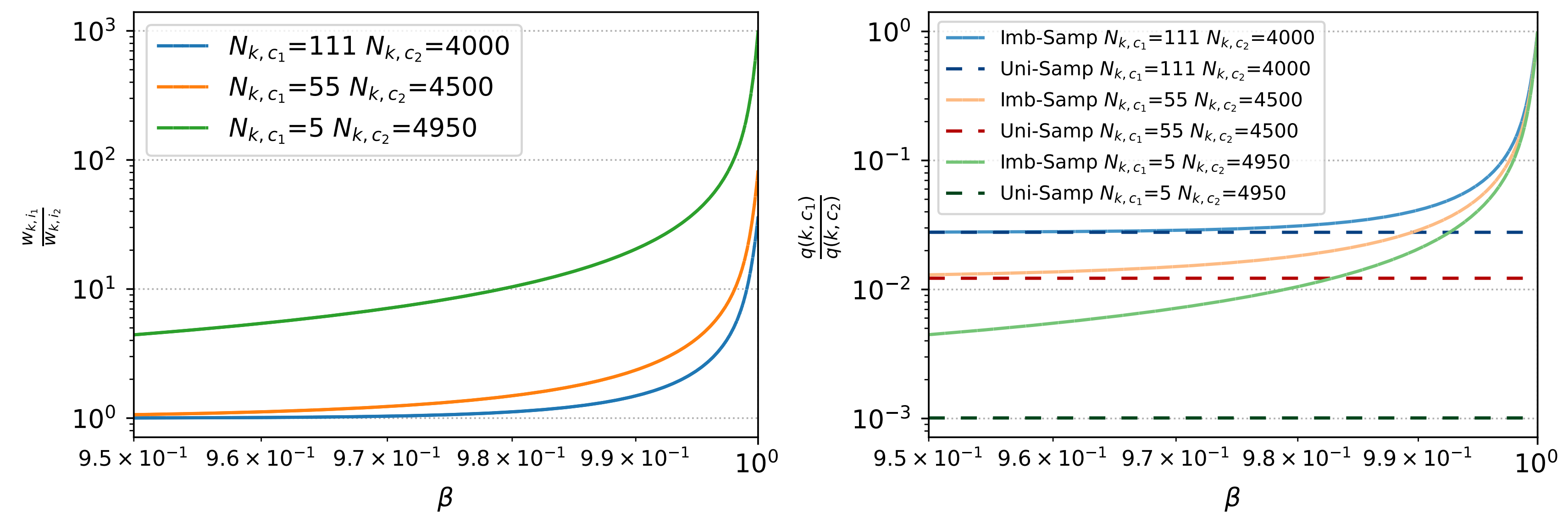
From the results, we found that:

1. By data resampling with high β, the label sampling probabilities $q(k,c)$ are made more similar across clients, which leads to faster convergence.
2. Imbalanced data resampling results in final accuracy decreasing on local dataset.

Approach

Based on the phenomenon and the findings in the preliminary results, we propose to conduct **Imbalanced Weight Decay Sampling (IWDS)**, which decays the resampling degree with the training time, such that

1. The convergence can be accelerated at early stage.
2. The model can learn more information of its own special knowledge well in the late training stage.



(a) Sampling weights ratio of two samples of different labels (b) Sampling probability ratio of two different labels

Figure 3. The relationship between the value of β and how much it can re-balance two classes.

From Fig. 3, we can see higher value of β could make two different label sampling probability become more similar. So, in order to make clients have similar label sampling probability, we firstly use high β at early training stages. And then, to diminish the effect of resampling, we gradually decay it to a value that makes $q(k,c_1)/q(k,c_2)$ close to uniform sampling case.

Therefore, we change the sampling weight of i-th sample in client k into:

$$w_{k,i,t} = (1-\beta_t)/(1-\beta_t^{N_{k,\boldsymbol{Y}_{(k,i)}}}),$$

in which t is the t-th communication round. The β_t is updated during each communication round as

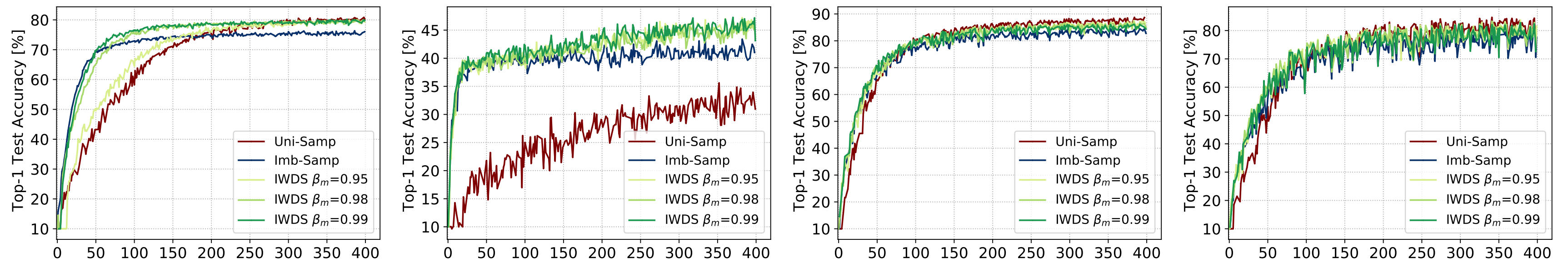
$$\beta_t = \beta_m + (\beta_0 - \beta_m) * \rho^t, \tag{1}$$

in which ρ is the decay rate. This equation makes the β decay from β_0 to β_m with the exponential rate ρ.

Experiment settings

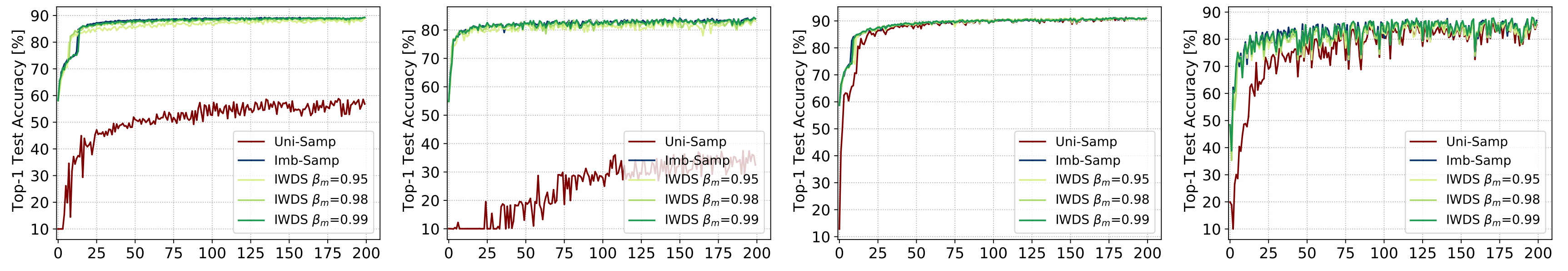
- **FL Algorithms.** We choose the classical Fl algorithm FedAvg [4], the recent FL algorithms FedProx [2] and FedNova [5] to verify the effect of IWDS with different FL algorithms.
- **Datasets and models.** We evaluate our methods on CIFAR-10 with VGG-9, and Fashion-MNIST with a simple CNN used in [4].
- **Datasets partition.** We conduct experiments of FL with 10 clients, and 5 clients will be chosen in every communication round. For both two datasets, 4 different ways of data partition are tested: LDA partition with $\alpha_d = 0.5$ and $\alpha_d = 0.1$, LLT partition with $\alpha_l = 0.9$ and $\alpha_l = 0.99$.

results



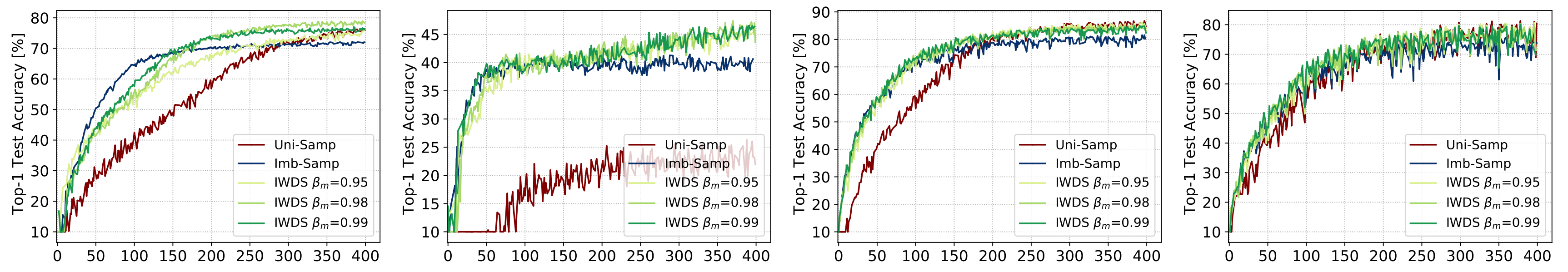
(a) LLT partition with $\alpha_l = 0.9$ (b) LLT partition with $\alpha_l = 0.99$ (c) LDA partition with $\alpha_d = 0.5$ (d) LDA partition $\alpha_d = 0.1$

Figure 4. Test accuracy of FedAvg, using VGG-9 on CIFAR-10.



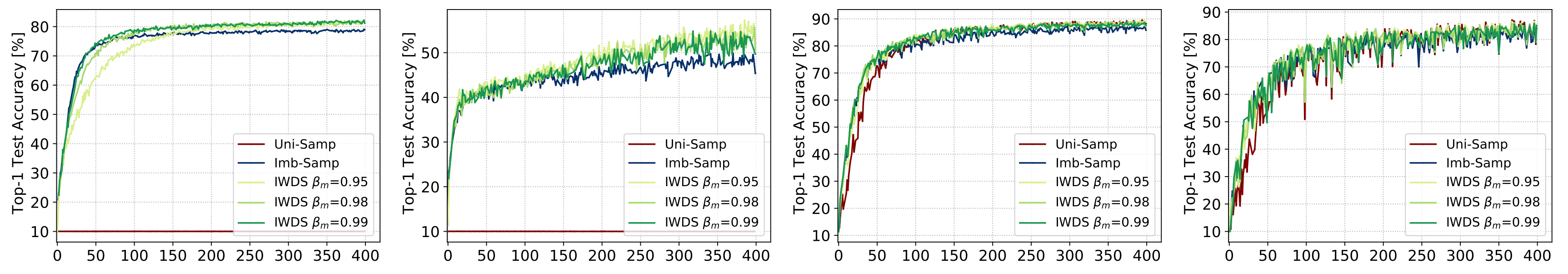
(a) LLT partition with $\alpha_l = 0.9$ (b) LLT partition with $\alpha_l = 0.99$ (c) LDA partition with $\alpha_d = 0.5$ (d) LDA partition $\alpha_d = 0.1$

Figure 5. Test accuracy of FedAvg, using a simple CNN on Fashion-MNIST.



(a) LLT partition with $\alpha_l = 0.9$ (b) LLT partition with $\alpha_l = 0.99$ (c) LDA partition with $\alpha_d = 0.5$ (d) LDA partition with $\alpha_d = 0.1$

Figure 6. Test accuracy of FedProx, using VGG-9 on CIFAR-10.



(a) LLT partition with $\alpha_l = 0.9$ (b) LLT partition with $\alpha_l = 0.99$ (c) LDA partition with $\alpha_d = 0.5$ (d) LDA partition with $\alpha_d = 0.1$

Figure 7. Test accuracy of FedNova with momentum acceleration, using VGG-9 on CIFAR-10.

Experiment results are shown in Fig. (4- 7). From these results, we can see:

- For all experiments with LLT partition, our method IWDS attain the fastest convergence rate and the highest final accuracy.
- When LDA partition, our method IWDS attain the faster convergence rate and the similar final accuracy.

References

- T. Hsu, H. Qi, and M. Brown. Measuring the effects of non-identical data distribution for federated visual classification. *ArXiv*, abs/1909.06335, 2019.
- T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.
- X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020.
- B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.
- J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 7611–7623, 2020.