Appendices for the Paper: *pFL-Bench: A Comprehensive Benchmark for Personalized Federated Learning*

We provide more details and experimental results for pFL-Bench in the appendices:

- Sec.A: the details of adopted datasets and models (*e.g.*, tasks, heterogeneous partitions, and model architectures), and the extensions for other datasets and models with pFL-Bench.
- Sec.B: detailed description of methods and metrics in our experiments.
- Sec.C: implementation details including the experimental environments and hyper-parameters.
- Sec.D: more experimental results in terms of generalization (Sec. D.1), fairness (Sec. D.2) and efficiency (Sec.D.3) for all the datasets in Table 1. Besides, to demonstrate the potential and ease of extensibility of the pFL-bench, we also conducted experiments in the heterogeneous device resource scenario based on FedScale [38] (Sec.D.4), as well as experiments incorporating privacy-preserving techniques (Sec.D.5).

A Datasets and Models

Experimental datasets. We present detailed descriptions of the 12 publicly available dataset variants used in pFL-Bench. These datasets are popular in the corresponding fields, and cover a wide range of domains, scales, partition manners, and Non-IID degrees.

- The Federated Extended MNIST (**FEMNIST**) is a widely used FL dataset for 62-class handwritten character recognition [32]. The original FEMNIST dataset contains 3,550 clients and each client corresponds to a character writer from EMNIST [91]. Following [13], we adopt the sub-sampled version in FL-Bench, which contains 200 clients and totally 43,400 images with resolution of 28x28 pixels, and the dataset is randomly split into train/valid/test sets with ratio 3:1:1. [2] within the local data of each client. In pFL-Bench, we use this dataset to vary the client sampling rate in FL processes as shown in Figure 4 in the main body of the paper.
- The **CIFAR10** is a popular dataset for 10-class image classification containing 60,000 colored images with resolution of 32x32 pixels. Follow the heterogeneous partition manners used in [56, 32, 37, 28], we use Dirichlet allocation to split this datasets into 100 clients with different Dirichlet factors as $\alpha = [5, 0.5, 0.1]$ (a smaller α indicates a higher heterogeneous degree). We split the dataset into train/valid/test sets with ratio 4:1:1.
- Corpus of Linguistic Acceptability (**COLA**) is a textual classification datasets from [92, 93], which contains 9,600 English sentences labeled with grammatical correctness. In pFL-Bench, this dataset is partitioned into 50 clients via Dirichlet allocation with $\alpha = 0.4$. We split the dataset into train/valid/test sets with a ratio of about 7:2:1.
- The Stanford Sentiment Treebank (**SST-2**) is a sentiment classification dataset from [92, 94], which contains 68,200 movie reviews sentences labeled with human sentiment. Similar to COLA, in pFL-Bench, this dataset is partitioned into 50 clients with Dirichlet allocation and $\alpha = 0.4$. The train/valid/test sets are with a ratio of about 60:15:1. For **COLA** and **SST-2**, since the test subsets from GLUE [92] are unlabeled (private in the GLUE server), we made new train/val/test partitions different from GLUE versions.
- The **Twitter** dataset is a textual sentiment analysis dataset from [32]. We adopt a subset which contains 13,203 users, the partition manner for this dataset is natural w.r.t. users, and the median number of data samples per user is 7. The train/valid/test sets for each client are with a ratio of about 3:1:1.
- The **Cora** dataset is a citation network that contains 2,708 nodes and 5,429 edges, in which each node indicates a scientific publication classified into one of seven classes [95]. Following FS-G [73], we split it into 5 clients using a community detection algorithm, Louvain [96]. The train/valid/test sets are with ratio about 3:1:1.

[2] For all the adopted datasets, the train/val/test splitting is conducted

- The **Pubmed** dataset contains 19,717 nodes and 44,338 edges. The nodes indicate scientific publications classified into one of three classes [97]. Following FS-G [73], we split it into 5 clients with Louvain community partition. The train/valid/test sets are with a ratio of about 3:1:5.
- The **Citeseer** dataset is a citation network that contains 3,312 nodes and 4,732 edges, in which each node indicates a scientific publication classified into one of six classes [98]. Following FS-G [73], we split it into 5 clients with Louvain community partition. The train/valid/test sets are with a ratio of about 4:1:1.
- The **Movielens1M** contains 1,000,209 ratings from 6,040 users and 3,900 movies [99]. Following the horizontal partition manner used in [100], in pFL-Bench, we split this dataset into 1,000 clients according to *users*. The train/valid/test sets with ratio about 14:3:3.
- The **Movielens10M** contains 10,000,054 ratings from 71,567 users and 10,681 movies [99]. Following the vertical partition manner used in [100], in pFL-Bench, we split this dataset into 1,000 clients according to *items*. The train/valid/test sets are with ratio about 14:3:3.

For all these experimental datasets, we randomly select 20% clients as new clients that do not participate in the FL processes. We summarize some statistics in Table 1 in the main body of the paper. Besides, we illustrate the violin plot of data size per client in Figure 6, the label skew visualization of certain datasets in Figure 7, and clients' pairwise similarity of label distribution in terms of Jensen–Shannon distance in Figure 8. And the smaller the Jensen-Shannon distance, the more similar the compared distributions. We can see that as the degree of heterogeneity increase (the α decreases), the larger the label skew degree and the Jensen-Shannon distances we get. We can perform similar calculations on a variety of FL datasets, and further rank this distances, and in turn select those clients whose distributions are very different but whose models do not perform well for further analysis, understanding and algorithm improvement. Furthermore, all these results show diverse properties across the adopted FL datasets in pFL-Bench, enabling comprehensive comparisons among different methods.

Models To align with previous works [77, 56, 78, 79, 80], we preset a 2-layer CNN for FEMNIST and CIFAR10. Specifically, the model consists of two convolutional layers with 5×5 kernels, max pooling, batch normalization, ReLU activation, and two dense layers. The hidden size is 2,048 and 512 for FEMNIST and CIFAR10 respectively. For the COLA and SST-2 datasets, we preset the pre-trained BERT-Tiny model from [81], which contains 2-layer Transformer encoders with a hidden size of 128. For the Twitter dataset, we preset a LR model with 50d Glove embeddings [3]. For the graph datasets, we preset the graph isomorphism neural network, GIN [82], which contains 2-layer convolutions with batch normalization, the hidden size of 64, and dropout rate of 0.5. For the recommendation datasets, we preset the Matrix Factorization (MF) model [83] with a hidden size of 20 for user and item embeddings.

Remark on the adopted dataset scales and model sizes. It is worth noting that simulation with pFL algorithms on a large client scale is very challenging, due to the fact that we need to maintain distinct (personalized) model object for each client. Let's take the famous benchmark FEMNIST as an example, which has 3,550 users and suppose we adopt the widely-used two-layer CNN network. Although this model only occupies 200MB, maintaining 3,550 such models would consume more than 700GB memory. Different from non-personalized FL algorithms, for which it is feasible to maintain only one model object for all the clients, for pFL algorithms, we may have to switch and cache the personalized models among CPU, GPU and even disks. Due to the large number of methods and datasets included in our benchmark, and the corresponding huge hyper-parameter search space, we used several subsets of the FL datasets to reduce the reproduction and experimental barriers.

Extension. We note that besides the experimental datasets and models introduced above, our code-base is compatible with a large number of datasets from other public popular DataZoos and ModelZoos. We provide the unified dataset, dataloader, and model I/O interfaces with carefully designed modularity, which enables users to easily register and extend the datasets/models with simple and flexible configuration, such as different heterogeneous partition manners, number of clients, new client ratio, model types and model parameter dimensions. Currently, we support datasets from LEAF [32], Torchvision [41], Huggingface datasets [42], FederatedScope (FS) [13]

[3] https://nlp.stanford.edu/data/glove.6B.zip

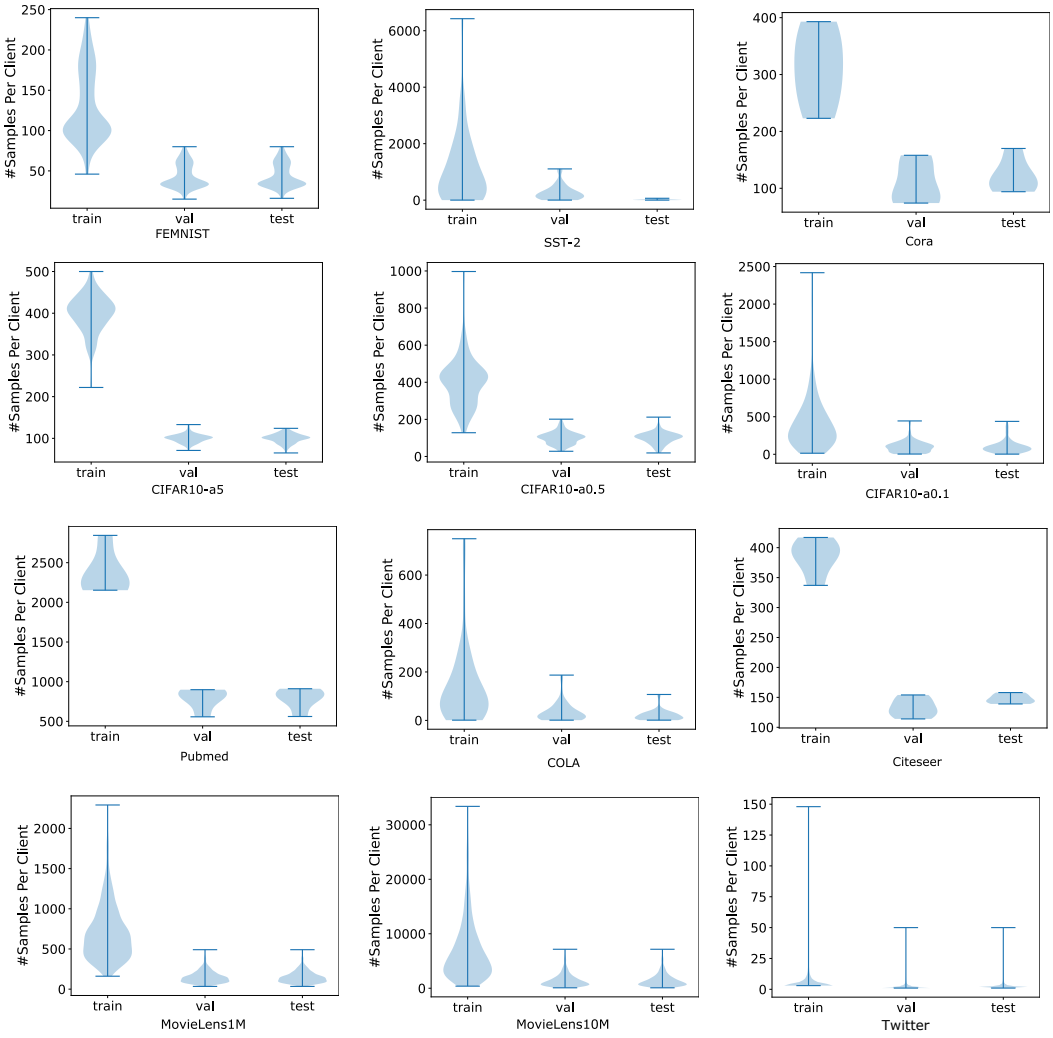

Figure 6: The violin plot of number of samples per client for all the adopted datasets.

and FederatedScope-GNN (FS-G) [73]; and models from Torchvision [41], Huggingface [84], FS [13] and FS-G [73].

B Methods and Metrics

B.1 Methods

We present detailed descriptions of the methods in pFL-Bench, which conveys a range of popular and SOTA methods in three categories including **Non-pFL methods**, **pFL methods** and **Combined variants**.

The following **Non-pFL methods** are considered in pFL-Bench:

- The *Global-Train* method refers to training only a centralized model from all data merged from all clients.
- The *Isolated* method indicates that each client trains its' client-specific model without FL communication. The *Global-Train* and *Isolated* methods provide a good reference to examine the benefits of pFL processes. For these two methods, we omit the un-participated clients.

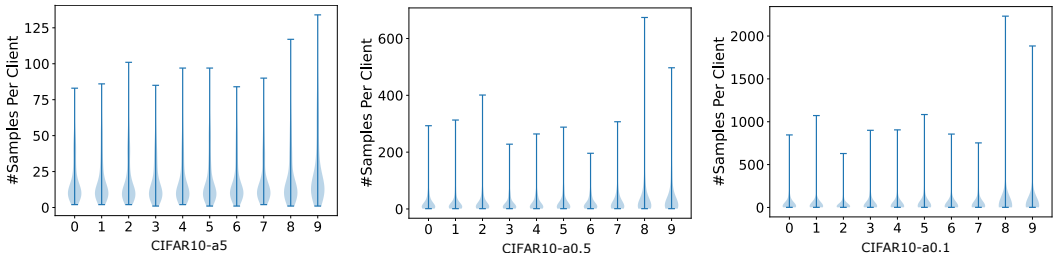

Figure 7: The label skew visualization in terms of number of labels per client for the CIFAR-10 datasets with different Dirichlet allocation factor αs.

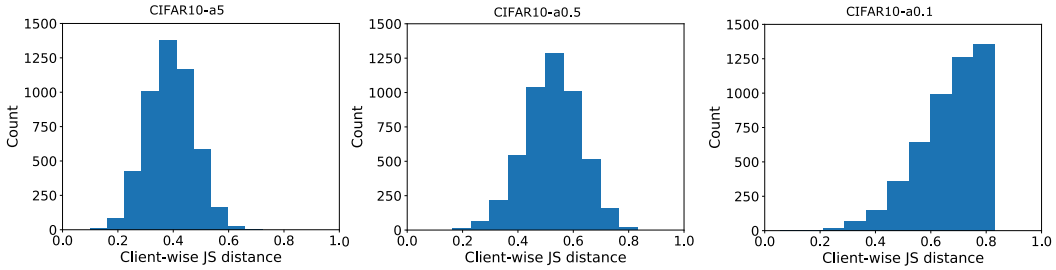

Figure 8: The histogram of clients' pairwise Jensen–Shannon distance in terms of their label distributions. The smaller the Jensen-Shannon distance, the more similar the compared distributions.

- The *FedProx* [85] method leverages proximal term to encourage the updated models at clients not to differ too much from the global model.
- In addition, we include the classical *FedAvg* [1] that average gradients weighted by data size of clients in each FL round.
- The *FedOpt* [43] algorithm is also considered, which generalizes FedAvg by introducing an optimizer for the FL server. We use the SGD as the server optimizer for FedOpt and search its learning rate.

pFL methods. We consider the following representative SOTA methods:

- *FedBN* [86] is a simple yet effective pFL method aiming to handle the feature shift Non-IID challenge. It locally maintains the clients' batch normalization parameters without FL communication and aggregation. In pFL-Bench, we generalize FedBN into the Transformer model by filtering out the layer normalization parameters.
- The *Ditto* [26] is a pFL method aiming to improve the fairness and robustness of FL. For each client, Ditto maintains the local personalized model and global model at the same time. The global model is trained with the same produce in FedAvg and the local model is trained with a personalized regularization according to the global model parameters.
- The *pFedMe* [78] is a meta-learning based method and also regularizes the local models according to the global model parameters. The authors propose to use Moreau envelops based regularization to reduce the complexity caused by Hessian matrix computation, which is required by some meta-learning based pFL methods such as Per-FedAvg [101].
- The pFL-Bench also contains multi-model based pFL methods.
- The *HypCluster* [74] method proposes to split clients into clusters and learns different personalized models for different clusters. The cluster is determined by performance on validation sets. In our experiments, we set the number of clusters as 3 for a fair comparison with FedEM.

  The *FedEM* method [56] assumes the local data distribution is a mixture of multiple underlying distributions. It learns a mixture of multiple local models with Expectation-Maximization algorithm to deal with the data heterogeneity, and can be easily extended to several clustering based and multi-task learning based method pFL methods. In our experiments, we use 3 internal models for FedEM according to the authors' default choice.

Combined variants. It is worth noting that we provide pluggable re-implementations of numerous existing methods in pFL-Bench. This modularity enables users can pick different personalized objects and behaviors to form a new pFL variant. We combine *FedBN*, *FedOpt*, and *Fine-tuning (FT)* with other compatible methods. The *FedBN* combination indicates to make the batch/layer normalization parameters personalized and locally maintained. The *FedOpt* combination indicates introducing the server optimizer into the FL processes. The *Fine-tuning (FT)* combination indicates fine-tuning the local models with a few steps before evaluation within the FL processes.

To facilitate fine-grained ablations and systematic pFL study, we finally compare more than 20 pFL method variants in the experiments. We will continuously include more pFL methods into pFL-Bench.

B.2 Metrics

Here we summarize the monitored metrics in our benchmark and give more detailed description about them.

Generalization. We support server-side and clients-sides monitoring w.r.t. widely used performance metrics such as accuracy, loss, F1, etc. Specifically, in the paper, we denote $\overline{Acc}$ $\overline{Loss}$ be the accuracy Loss average weighted by the number of local data samples, $\widetilde{Acc}$ $\widetilde{Loss}$ be the accuracy Loss of un-participated clients, and Δ be the participation generalization gap.

Fairness. We support different summarizing manners over the evaluated metrics over clients, such as weighted average (e.g., $\overline{Acc}$), uniform average (e.g., $\overline{Acc}'$), the standard deviation (denoted by σ in our paper), and various quantiles such as bottom accuracy $\breve{Acc}$. We report the $\lfloor |\mathcal{C}|/10 \rfloor$-th worst accuracy where $|\mathcal{C}|$ is the number of all evaluated clients. To align with FedEM, the 90th percentile is considered here to omit the particularly noisy results from clients with worse performance with very small data sizes.

System costs. For computational and communication costs, we support to monitor some proxy metrics including FLOPs, communication bytes, and convergence rounds. The FLOPS are counted as the sum of amounts for both training and inference via a per-operator flops counting tool, fvcore/flop_count. [4] The reported communication bytes are counted as the sum of upstream and downstream across all participants until convergence with early stopping. Besides, thanks to the good integration of wandb, our benchmark also supports more runtime metrics including the dynamic utilization of CPU, GPU, memory, disk, etc. [5]

C Implementation

Enviroments. We implement pFL-Bench based on the FS [13] package and PyTorch. The experiments are conducted on a cluster of 8 Tesla V100 and 64 NVIDIA GeForce GTX 1080 Ti GPUs, each machine with 380G memory and Xeon Platinum 8163 2.50GHz CPU containing 96 cores. Our experiments are conducted in the containerized environments with Ubuntu18.04.

We provide versioned *DockerFiles*, the built docker images and experimental datasets in our website with Aliyun storage service. [6] The pFL-Bench and the underlying FS [13] package is continuously developed and maintained by Data Analytics and Intelligence Lab (DAIL) of DAMO Academy. We will actively fix potential issues, track updates and Github release.

Hyper-parameters. For fair comparisons, we first use wandb sweep [7] with the hyper-parameter searching (HPO) algorithm, HyperBand [102], to find the best hyper-parameters for all the methods on all datasets. The validation sets are used and we employ early stopping with a large number of total FL rounds T. We set this hyper-parameter to make almost all methods converge within T

[4] https://github.com/facebookresearch/fvcore/blob/main/docs/flop_count.md

[5] https://docs.wandb.ai/ref/app/features/system-metrics

[6] https://github.com/alibaba/FederatedScope/tree/master/benchmark/pFL-Bench

[7] https://docs.wandb.ai/sweeps

rounds. Specifically, for FEMNIST, CIFAR10, Movielens-1M, Movielens-10M and Twitter, we set $T = 1,000$. For Cola, SST-2, Pubmed, Cora and Citeseer, we set $T = 500$. The batch size is set to be 32 for image datasets, 64 for textual datasets, and 1,024 for the recommendation datasets respectively. For graph datasets, we adopt full batch training. For all methods, we search the local update steps (*i.e.*, the number of local training epochs in each FL round) from $[1, 3, 6]$. For the local SGD learning rate, we search from $[0.05, 0.005, 0.5, 0.01, 0.1, 1, 2]$. For FedOpt, we search the server learning rate from $[0.05, 0.1, 0.5, 1.5]$. For pFedMe, we search the personalized regularization weight from $[0.05, 0.1, 0.2, 0.9]$ and its local meta-learning step from $[1, 3]$. For FedEM, we set its number of mixture models as 3. For Ditto, we search its personalized regularization weight from $[0.05, 0.1, 0.2, 0.5, 0.8]$.

To enable easily reproducible research, we provide standardized and documented scripts including the HPO scripts, the experiments running scripts and searched best configuration files in our code-base (see the link in the above paragraph).

D Additional Experimental Results

D.1 Generalization

The generalization results for FEMNIST, SST-2 and PUBMED are shown in the Table 2 and Figure 3 in the main body of the paper. Here we present the results for other datasets including all the textual datasets (Table 4), all the graph datasets (Table 5), and all the recommendation datasets (Table 6). We note that FedOpt may gain bad results on some datasets when the models contain batch/layer normalization parameters. Besides, for the Twitter and recommendation datasets, we have not compared the FedBN based methods as the used LR and MF models do not contain batch/layer normalization parameters.

D.2 Fairness

The fairness results for FEMNIST, SST-2 and PUBMED are listed in Table 3 in the main body of the paper. Here we present the results for other datasets, including all the textual datasets (Table 7), all the graph datasets (Table 8), and all the recommendation datasets (Table 9).

D.3 Efficiency

The efficiency-accuracy trade-off results for FEMNIST are plotted in Figure 5 in the main body of the paper. Here we present more efficiency-accuracy trade-off results for the experimental datasets, including the FEMNIST datasets with different client sampling rates (Table 11 and Table 12), CIFAR-10 datasets with different α (Table 13 and Table 14), all the textual datasets (Table 15 and Table 16), all the textual datasets (Table 18 and Table 19), and all the recommendation datasets (Table 20 and Table 17).

Besides the reported proxy system metrics such as FLOPs and the number of convergence rounds of FL processes, our benchmark also supports monitoring more runtime metrics. Thanks to the good integration with wandb, we can easily track the usage of system resources in runtime including utilization of CPU, GPU, memory, disk, etc. In Table 21, we report the average and peak process memory usage (in MB) and process running times (in seconds). In general, most pFL algorithms do have higher time and space overheads. We omit to report results for other metrics since we started very many sets of experiments concurrently, taking up as much of the graphics card's memory and maximising CPU/GPU utilisation as possible, these metrics do not differ much from one of our different experiments. However, it is worth noting that these omitted metrics can be used to analyse algorithm bottlenecks in terms of system performance, and to optimise the space-time efficiency in single-experiment scenarios.

D.4 Heterogeneous Device Resources

The proposed pFL-Bench has good extensibility to support experiments in heterogeneous device resource scenarios, where clients have different computational and communication capacities. Specifically, we integrate FedScale [38] into our benchmark with a simulator that executes the behaviors of

Table 4: Accuracy results for both participated clients and un-participated clients on COLA, SST-2 and Twitter datasets. $\overline{Acc}$ indicates the aggregated accuracy weighted by the number of local data samples of participated clients, $\widetilde{Acc}$ indicates the aggregated accuracy of un-participated clients, and Δ indicates the participation generalization gap. **Bold** and underlined indicate the best and second-best results among all compared methods, while red and blue indicate the best and second-best results for original methods without combination "-".

	COLA			SST-2			Twitter		
	$\overline{Acc}$	$\widetilde{Acc}$	Δ	$\overline{Acc}$	$\widetilde{Acc}$	Δ	$\overline{Acc}$	$\widetilde{Acc}$	Δ
Global-Train	69.06	-	-	**80.57**	-	-	55.56	-	-
Isolated	55.96	-	-	60.82	-	-	70.04	-	-
FedAvg	71.85	63.49	-8.36	74.88	80.24	5.36	62.15	61.24	-0.91
FedAvg-FT	68.29	58.66	-9.63	74.14	**83.28**	9.13	70.53	71.17	**0.64**
FedOpt	71.85	59.62	-12.23	72.28	83.06	10.78	62.09	61.64	-0.45
FedOpt-FT	62.59	47.82	-14.77	65.77	80.02	14.25	71.08	71.41	0.33
pFedMe	74.40	67.64	-6.76	71.27	69.34	-1.92	63.45	62.52	-0.94
pFedMe-FT	**78.47**	**76.33**	-2.14	75.61	66.48	-9.13	**84.00**	**71.80**	-12.20
FedBN	71.85	63.49	-8.36	74.88	75.40	0.52	-	-	-
FedBN-FT	66.71	49.87	-16.84	68.81	82.43	13.63	-	-	-
FedBN-FedOPT	71.85	62.48	-9.37	64.70	65.50	0.81	-	-	-
FedBN-FedOPT-FT	67.48	57.59	-9.90	68.65	70.56	1.91	-	-	-
Ditto	55.46	49.90	-5.56	52.03	46.79	-5.24	70.23	49.60	-20.63
Ditto-FT	72.11	52.15	-19.96	56.49	65.50	9.01	69.99	51.32	-18.67
Ditto-FedBN	70.69	49.90	-20.79	56.03	46.79	-9.24	-	-	-
Ditto-FedBN-FT	72.66	53.44	-19.21	53.15	66.49	13.34	-	-	-
Ditto-FedBN-FedOpt	50.25	49.90	-0.35	57.67	46.79	-10.88	-	-	-
Ditto-FedBN-FedOpt-FT	55.01	58.22	**3.21**	52.89	66.49	13.60	-	-	-
FedEM	71.85	63.49	-8.36	75.78	67.67	-8.11	63.44	62.68	-0.75
FedEM-FT	54.90	48.29	-6.61	64.86	81.63	**16.77**	70.97	71.59	0.62
FedEM-FedBN	71.44	63.99	-7.45	75.43	62.81	-12.62	-	-	-
FedEM-FedBN-FT	57.62	58.88	1.26	64.96	81.04	16.08	-	-	-
FedEM-FedBN-FedOPT	71.85	62.82	-9.03	72.25	64.69	-7.56	-	-	-
FedEM-FedBN-FedOPT-FT	57.23	58.88	1.65	62.26	73.87	11.61	-	-	-

clients according to virtual timestamps of their message delivery to the server. The virtual timestamps are updated by the estimated execution time based on clients' computational and communication capacities with the cost model proposed in FedScale. The server employs an over-selection mechanism for clients at each broadcast round and thus some clients' message may be dropped, since the clients have different system capacities and different respond speeds corresponding to real-world mobile devices. [8]

Here we take the Ditto method on FEMNIST dataset as an example and present the results of experiments with heterogeneous device resources in Table 10. Let s to be the clients sampling rate for each FL round, and s_{agg} be the the minimal ratio of received feedback w.r.t. the number of clients for the server to trigger federated aggregation in over-selection mode. For the homo-device case, we set $s = 0.2$ and for the hetero-device case, we set the $s = 0.25$ and $s_{agg} = 0.8$, leading to the same number of clients used for each federated aggregation. From the results, we can see that the hetero-device version has slower convergence speed ($T' = 0$ indicates that the early-stopping is not triggered within the large number of FL rounds $T = 1000$), and it gains worse performance than the homo-device version, especially for the bottom accuracy ($\overline{Acc}'$) and standard deviation of the average accuracy (σ). This shows unfairness among clients due to the fact that some low-resourced clients have too long computation or communication time to make their feedback incorporated into the federated aggregation, calling for more considerations w.r.t. device heterogeneity within pFL algorithms design.

[8] https://github.com/SymbioticLab/FedScale/tree/master/benchmark/dataset/data/device_info

Table 5: Accuracy results for both participated clients and un-participated clients on Pubmed, Cora and Citeseer datasets. $\overline{Acc}$ indicates the aggregated accuracy weighted by the number of local data samples of participated clients, $\widetilde{Acc}$ indicates the aggregated accuracy of un-participated clients, and Δ indicates the participation generalization gap. **Bold** and underlined indicate the best and second-best results among all compared methods, while red and blue indicate the best and second-best results for original methods without combination "-".

	PUBMED			CORA			CITESEER		
	$\overline{Acc}$	$\widetilde{Acc}$	Δ	$\overline{Acc}$	$\widetilde{Acc}$	Δ	$\overline{Acc}$	$\widetilde{Acc}$	Δ
Global-Train	87.01	-	-	**86.10**	-	-	74.03	-	-
Isolated	85.56	-	-	82.48	-	-	69.83	-	-
FedAvg	87.27	72.63	-14.64	81.30	72.14	-9.16	75.58	59.83	-15.74
FedAvg-FT	87.21	79.78	-7.43	82.07	75.84	-6.22	75.63	66.07	-9.57
FedOpt	67.38	53.84	-13.54	70.70	59.58	-11.12	71.59	55.16	-16.43
FedOpt-FT	82.36	64.09	-18.27	82.68	62.17	-20.51	74.34	62.36	-11.99
pFedMe	86.91	71.64	-15.27	83.18	70.13	-13.05	75.30	58.45	-16.85
pFedMe-FT	85.71	77.07	-8.64	82.11	71.48	-10.63	75.35	62.14	-13.20
FedBN	**88.49**	52.53	-35.95	84.13	57.33	-26.80	75.80	51.29	-24.51
FedBN-FT	87.45	**80.36**	**-7.09**	76.20	64.13	-12.07	75.07	64.20	-10.87
FedBN-FedOPT	87.87	42.72	-45.15	84.64	53.27	-31.37	76.20	50.34	-25.86
FedBN-FedOPT-FT	87.54	77.07	-10.47	84.10	68.14	-15.96	**76.70**	62.84	-13.86
Ditto	87.27	2.84	-84.43	83.67	14.53	-69.14	74.79	14.52	-60.27
Ditto-FT	87.47	35.03	-52.44	81.47	72.64	-8.84	76.47	39.27	-37.20
Ditto-FedBN	88.18	2.84	-85.34	81.38	14.53	-66.84	75.35	14.52	-60.83
Ditto-FedBN-FT	87.83	28.52	-59.30	83.25	65.87	-17.38	75.97	36.51	-39.46
Ditto-FedBN-FedOpt	87.81	2.84	-84.97	82.54	14.53	-68.01	75.07	14.52	-60.55
Ditto-FedBN-FedOpt-FT	87.60	18.18	-69.42	82.00	70.75	-11.25	76.42	48.99	-27.43
FedEM	85.64	71.12	-14.52	81.92	72.02	-9.90	75.41	59.60	-15.81
FedEM-FT	85.88	78.08	-7.80	77.32	**79.45**	**2.13**	72.71	66.77	**-5.94**
FedEM-FedBN	88.12	48.64	-39.48	85.07	50.98	-34.09	74.90	41.55	-33.35
FedEM-FedBN-FT	86.38	72.02	-14.35	84.61	76.77	-7.83	75.29	68.38	-6.91
FedEM-FedBN-FedOPT	87.56	42.37	-45.19	84.68	56.45	-28.24	76.08	53.81	-22.27
FedEM-FedBN-FedOPT-FT	87.49	72.39	-15.09	85.02	76.97	-8.05	76.59	**68.62**	-7.96

D.5 Incorporating Differential Privacy

It is interesting and under-explored to investigate the trade-off between personalization and privacy protection, which is important as FL involves the transmitting of local (maybe private) information. We note that FederatedScope supports various privacy-related fundamental components and algorithms, such as Differential Privacy (DP) [103] and privacy attack methods [104] that can be used to examine the privacy-preserving strength. Further research that combines pFL and privacy-preserving techniques will be convenient based on the modularized and extensible design of our benchmark. As a preliminary example, here we demonstrate the combination of the pFL with a Differential Privacy algorithm, the NbAFL [105] that achieves (ϵ, δ)-DP via noise injection and gradient clipping.

In Figure 9, we plot the learning curves of FedAvg and Ditto methods on FEMNIST dataset with various (ϵ, δ)-DP. Generally speaking, for privacy protection, the smaller the protection, the less performance degradation there is. We can see that in the Figure, with larger ϵ and δ, the accuracy ($\overline{Acc}$) is better for both the compared methods, which meets our expectation. Interestingly, Ditto shows significantly better robustness for the dramatically varying privacy protection strengths during the whole learning process than FedAvg. This may be because, in the noise perturbation scenarios, the personalized local model potentially brings up more local optimal points that can be reached for clients. But there is still a gap for the best achievable performance between Ditto and FedAvg, leaving an interesting open question about how to reduce the performance degradation by co-designing personalization and noise injection.

Table 6: Accuracy results for both participated clients and un-participated clients on Movielens-1M and Movielens-10M datasets. $\overline{Loss}$ indicates the loss average weighted by the number of local data samples, $\widetilde{Loss}$ indicates the loss of un-participated clients, and Δ indicates the participation generalization gap. **Bold** and underlined indicate the best and second-best results among all compared methods, while red and blue indicate the best and second-best results for original methods without combination "-".

	Movielens-1M			Movielens-10M		
	$\overline{Loss}$	$\widetilde{Loss}$	Δ	$\overline{Loss}$	$\widetilde{Loss}$	Δ
Global-Train	0.78	-	-	**0.67**	-	-
Isolated	10.35	-	-	11.48	-	-
FedAvg	0.84	14.17	13.33	0.70	13.39	12.68
FedAvg-FT	0.84	9.76	8.92	0.71	**11.07**	10.36
FedAvg-FT-FedOpt	0.85	5.15	4.31	0.73	11.40	10.66
FedOpt	0.83	14.17	13.34	0.71	13.39	12.68
FedOpt-FT	0.83	12.06	11.23	0.74	11.92	11.18
pFedMe	**0.54**	14.18	13.64	13.06	12.73	**-0.33**
pFedMe-FT	0.60	8.20	7.60	0.80	12.59	11.79
Ditto	1.29	14.19	12.89	1.84	13.39	11.55
Ditto-FT	1.35	14.17	12.81	1.69	13.39	11.70
Ditto-FT-FedOpt	1.36	14.15	12.79	2.03	13.39	11.35
FedEM	0.85	14.27	13.43	1.75	13.41	11.65
FedEM-FT	0.85	4.86	4.01	0.87	12.80	11.93
FedEM-FT-FedOpt	0.86	**4.47**	**3.61**	1.43	13.25	11.82

Table 7: Fairness results on COLA, SST-2 and Twitter datasets. $\overline{Acc}'$ indicate the equally-weighted average, σ indicating the standard deviation of the average accuracy, and $\breve{Acc}$ indicating the bottom accuracy. **Bold**, underlined, red and blue indicate the same highlights as used in Table 2.

	COLA			SST-2			Twitter		
	$\overline{Acc}'$	σ	$\breve{Acc}$	$\overline{Acc}'$	σ	$\breve{Acc}$	$\overline{Acc}'$	σ	$\breve{Acc}$
Isolated	56.86	35.34	0.00	59.40	41.29	0.00	67.44	37.86	0.00
FedAvg	51.53	35.96	0.00	76.30	22.02	44.85	56.98	37.97	0.00
FedAvg-FT	58.74	35.21	0.00	75.36	27.67	31.08	68.45	36.19	0.00
FedOpt	57.10	**31.70**	**10.85**	73.78	34.03	17.53	59.95	37.88	0.00
FedOpt-FT	59.77	35.99	0.00	66.17	33.73	15.56	69.20	36.28	0.00
pFedMe	67.58	38.06	0.90	65.08	26.59	27.75	58.18	36.67	0.00
pFedMe-FT	**69.17**	33.63	9.90	74.36	27.02	32.49	**78.82**	**33.28**	**22.22**
FedBN	59.60	35.96	0.00	76.30	22.02	44.85	-	-	-
FedBN-FT	59.69	35.44	0.00	68.50	26.83	29.17	-	-	-
FedBN-FedOPT	59.60	36.03	0.00	65.59	31.07	22.22	-	-	-
FedBN-FedOPT-FT	59.10	35.15	0.00	68.42	28.18	30.71	-	-	-
Ditto	55.14	36.76	0.00	49.94	40.81	0.00	66.90	38.05	0.00
Ditto-FT	63.61	35.02	0.00	54.34	39.26	0.00	66.91	38.08	0.00
Ditto-FedBN	62.68	35.74	0.00	49.44	41.80	0.00	-	-	-
Ditto-FedBN-FT	63.58	34.58	0.00	52.18	39.85	0.00	-	-	-
Ditto-FedBN-FedOpt	52.48	35.89	0.00	55.61	40.43	1.39	-	-	-
Ditto-FedBN-FedOpt-FT	57.13	36.20	0.00	53.16	34.75	9.72	-	-	-
FedEM	51.52	35.96	0.00	**76.53**	23.34	44.44	61.70	37.72	0.00
FedEM-FT	57.80	35.24	0.00	64.29	32.84	12.96	70.19	36.35	0.00
FedEM-FedBN	57.95	34.11	1.52	75.06	**18.48**	**53.33**	-	-	-
FedEM-FedBN-FT	58.74	35.56	1.00	64.33	35.72	8.59	-	-	-
FedEM-FedBN-FedOPT	59.60	35.49	0.00	72.66	27.18	34.17	-	-	-
FedEM-FedBN-FedOPT-FT	56.74	35.61	1.00	58.42	31.21	17.93	-	-	-

Table 8: Fairness results on Pubmed, Cora, and Citeseer datasets. $\overline{Acc}'$ indicate the equally-weighted average, σ indicating the standard deviation of the average accuracy, and $\breve{Acc}$ indicating the bottom accuracy. **Bold**, underlined, red and blue indicate the same highlights as used in Table 2.

	PUBMED			CORA			CITESEER		
	$\overline{Acc}'$	σ	$\breve{Acc}$	$\overline{Acc}'$	σ	$\breve{Acc}$	$\overline{Acc}'$	σ	$\breve{Acc}$
Isolated	84.67	6.26	74.63	81.62	4.67	72.64	69.90	5.76	61.10
FedAvg	86.72	3.93	79.76	81.07	5.06	73.26	75.64	5.03	67.30
FedAvg-FT	86.71	3.86	80.57	81.90	3.06	77.57	75.77	5.00	68.68
FedOpt	66.69	16.69	48.50	70.17	12.56	38.89	71.60	9.20	42.25
FedOpt-FT	81.53	16.69	46.21	82.31	6.35	68.03	74.38	4.13	69.26
pFedMe	86.35	4.43	78.76	82.76	3.40	76.70	75.36	4.79	68.55
pFedMe-FT	85.47	**3.06**	80.95	81.98	**1.63**	79.58	75.40	4.39	70.31
FedBN	**87.97**	3.42	81.77	83.64	3.88	77.53	75.59	**3.80**	71.24
FedBN-FT	87.02	3.47	80.13	76.01	4.32	69.92	75.16	5.04	67.91
FedBN-FedOPT	87.43	4.64	80.81	84.11	3.93	75.44	76.33	4.28	**72.43**
FedBN-FedOPT-FT	87.02	3.94	81.78	83.79	2.39	**80.26**	**76.77**	4.60	71.68
Ditto	86.85	3.98	80.44	83.50	3.54	75.02	75.55	4.23	67.99
Ditto-FT	87.10	3.52	80.46	81.53	3.67	75.63	76.57	4.24	70.31
Ditto-FedBN	87.75	3.70	81.82	81.36	3.34	74.93	75.46	4.83	68.57
Ditto-FedBN-FT	87.43	3.77	81.15	82.21	4.37	72.56	76.05	4.21	69.16
Ditto-FedBN-FedOpt	87.27	3.90	79.14	82.38	2.10	77.69	75.19	4.06	68.60
Ditto-FedBN-FedOpt-FT	87.10	3.79	80.93	81.99	3.67	73.29	76.52	4.43	71.31
FedEM	85.05	4.44	78.51	81.72	3.72	74.95	75.49	4.66	68.50
FedEM-FT	85.54	4.48	79.39	78.43	11.72	56.20	72.88	6.55	63.53
FedEM-FedBN	87.63	4.14	**82.54**	84.45	3.14	76.71	74.29	4.22	69.06
FedEM-FedBN-FT	85.68	4.33	79.44	**84.57**	3.04	80.22	75.40	3.84	70.59
FedEM-FedBN-FedOPT	87.11	4.24	80.32	83.88	4.37	78.93	76.17	4.76	70.54
FedEM-FedBN-FedOPT-FT	87.16	3.66	82.20	84.40	4.43	79.53	76.74	4.53	71.92

Table 9: Fairness results on Movielens-1M and Movielens-10M datasets. $\overline{Loss}'$ indicate the equally-weighted average loss, $\breve{Loss}$ indicates the bottom loss (the largest), and σ indicates the standard deviation of the average loss. **Bold**, underlined, red and blue indicate the same highlights as used in Table 2.

	Movielens-1M			Movielens-10M		
	$\overline{Loss}'$	σ	$\breve{Loss}$	$\overline{Loss}'$	σ	$\breve{Loss}$
Isolated	11.12	2.43	14.34	11.44	1.83	13.75
FedAvg	0.85	0.21	1.13	**0.71**	**0.11**	**0.84**
FedAvg-FT	0.85	0.21	1.12	**0.71**	**0.11**	0.85
FedAvg-FT-FedOpt	0.86	0.22	1.14	0.75	0.12	0.90
FedOpt	0.84	0.21	1.11	0.72	0.12	0.87
FedOpt-FT	0.84	0.21	1.11	0.77	0.13	0.94
pFedMe	**0.55**	**0.11**	**0.69**	12.48	2.44	15.76
pFedMe-FT	0.60	0.12	0.75	0.80	0.13	0.96
Ditto	1.31	0.77	1.69	1.81	0.24	2.12
Ditto-FT	1.35	0.88	1.70	2.30	1.15	4.05
Ditto-FT-FedOpt	1.35	0.79	1.70	1.98	0.27	2.32
FedEM	0.87	0.22	1.15	2.37	1.25	4.09
FedEM-FT	0.87	0.23	1.16	0.98	0.26	1.29
FedEM-FT-FedOpt	0.87	0.22	1.16	1.88	0.94	3.09

Table 10: Comparison between Ditto methods with and without heterogeneous device capabilities on FEMNIST dataset. $\overline{Acc}$ indicates the accuracy average weighted by the number of local data samples, $\widetilde{Acc}$ indicates the accuracy of un-participated clients, and Δ indicates the participation generalization gap. $\overline{Acc}'$ indicate the equally-weighted average, σ indicating the standard deviation of the average accuracy, and $\breve{Acc}$ indicating the bottom accuracy. Efficiency metrics include total FLOPS, communication bytes (Com.) and the convergence round $T' = 0$. The $T' = 0$ indicates the early-stopping is not triggered within the large number of FL rounds $T = 1000$.

	$\overline{Acc}$	$\widetilde{Acc}$	Δ	$\overline{Acc}'$	σ	$\breve{Acc}$	T'	FLOPS	Com.
Ditto, Homo-Device	88.39	2.2	-86.19	87.18	7.52	78.23	849.3G	2.81M	610
Ditto, Hetero-Device	79.76	1.43	-78.33	77.39	11.25	61.76	1.72T	5.72M	0

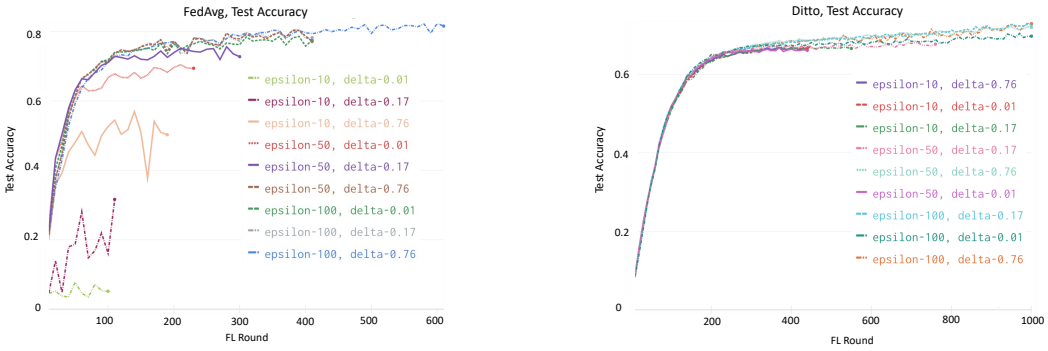

Figure 9: The learning curves of FedAvg and Ditto on FEMNIST dataset with various (ϵ, δ)-DPs.

Table 11: The efficiency-accuracy trade-off results including total FLOPS, communication bytes (Com.), and $\overline{Acc}$ for FEMNIST datasets with different s.

	FEMNIST, $s = 0.2$			FEMNIST, $s = 0.1$			FEMNIST, $s = 0.05$		
	FLOPS	Com.	$\overline{Acc}$	FLOPS	Com.	$\overline{Acc}$	FLOPS	Com.	$\overline{Acc}$
FedAvg	195.38G	1.24M	83.97	159.9G	373.57K	83.84	78.48G	384.0K	83.21
FedAvg-FT	296.8G	1.35M	86.44	155.49G	630.44K	85.45	217.97G	522.82K	84.78
FedOpt	846.22G	1.68M	19.31	272.3G	635.48K	27.29	**29.0G**	**80.81K**	22.53
FedOpt-FT	**77.81G**	**149.97K**	9.69	**54.88G**	**222.47K**	31.02	365.54G	877.15K	29.73
pFedMe	2.24T	1.52M	87.50	516.47G	1012.93K	84.70	216.66G	493.06K	85.23
pFedMe-FT	1.33T	2.19M	88.19	1.22T	953.21K	86.28	585.23G	569.72K	87.37
FedBN	243.31G	1.21M	86.72	120.66G	510.15K	85.10	189.64G	841.57K	84.36
FedBN-FT	208.99G	687.58K	88.51	174.22G	610.31K	87.31	118.16G	407.71K	86.87
FedBN-FedOPT	451.07G	1.78M	88.25	179.28G	758.36K	86.22	109.73G	487.2K	84.04
FedBN-FedOPT-FT	195.3G	696.81K	88.14	217.59G	762.72K	87.91	171.93G	593.18K	87.34
Ditto	2.04T	1.78M	88.39	700.28G	700.96K	88.87	399.96G	533.77K	88.90
Ditto-FT	2.73T	2.26M	85.72	349.06G	600.22K	67.97	433.51G	749.29K	72.08
Ditto-FedBN	1.51T	1.02M	**88.94**	720.27G	623.37K	**89.33**	336.44G	695.85K	67.63
Ditto-FedBN-FT	2.86T	1.89M	86.53	1.42T	1.19M	86.35	410.03G	762.09K	72.84
Ditto-FedBN-FedOpt	2.45T	1.65M	88.73	922.71G	801.91K	89.30	408.13G	493.82K	**88.95**
Ditto-FedBN-FedOpt-FT	2.52T	1.66M	87.02	1.42T	1.19M	87.34	876.33G	993.59K	79.38
FedEM	2.98T	2.0M	84.35	1.99T	1.57M	84.49	1017.32G	1.03M	84.46
FedEM-FT	5.79T	1.67M	86.17	4.12T	1.16M	86.13	3.83T	1.35M	86.11
FedEM-FedBN	10.34T	3.1M	84.37	1.87T	1.26M	84.83	1.23T	1.15M	84.26
FedEM-FedBN-FT	7.55T	2.59M	88.29	4.07T	1.34M	87.49	5.03T	1.29M	86.54
FedEM-FedBN-FedOPT	6.0T	1.79M	82.12	3.89T	2.63M	85.81	1.81T	1.7M	85.16
FedEM-FedBN-FedOPT-FT	13.86T	4.76M	87.54	4.11T	1.36M	87.69	3.62T	1.15M	85.91

Table 12: The convergence results including the convergence round T' and $\overline{Acc}$ for FEMNIST datasets with different s. The $T' = 0$ indicates the early-stopping is not triggered within the large number of FL rounds $T = 1000$.

	FEMNIST, $s = 0.2$		FEMNIST, $s = 0.1$		FEMNIST, $s = 0.05$	
	T'	$\overline{Acc}$	T'	$\overline{Acc}$	T'	$\overline{Acc}$
FedAvg	173.33	82.40	246.67	82.07	350.00	82.31
FedAvg-FT	220.00	85.17	416.67	84.26	476.67	83.45
FedOpt	733.33	19.59	420.00	27.47	73.33	22.45
FedOpt-FT	63.33	10.39	**146.67**	30.88	800.00	30.02
pFedMe	640.00	86.50	670.00	83.69	450.00	84.29
pFedMe-FT	960.00	87.06	630.00	85.05	520.00	86.36
FedBN	273.33	85.38	390.00	83.81	846.67	82.76
FedBN-FT	340.00	87.65	466.67	86.22	410.00	86.01
FedBN-FedOPT	943.33	87.27	580.00	85.09	490.00	82.63
FedBN-FedOPT-FT	360.00	87.13	583.33	87.10	596.67	86.53
Ditto	406.67	87.18	463.33	87.65	486.67	87.80
Ditto-FT	286.67	84.30	396.67	65.86	483.33	70.20
Ditto-FedBN	536.67	**87.82**	476.67	**88.28**	700.00	66.15
Ditto-FedBN-FT	**0.00**	85.16	263.33	84.98	766.67	71.76
Ditto-FedBN-FedOpt	873.33	87.64	613.33	88.20	496.67	**87.84**
Ditto-FedBN-FedOpt-FT	880.00	85.71	263.33	86.00	**0.00**	76.75
FedEM	290.00	82.61	373.33	83.13	346.67	82.91
FedEM-FT	243.33	84.91	276.67	84.77	453.33	84.61
FedEM-FedBN	570.00	82.94	350.00	83.35	430.00	82.98
FedEM-FedBN-FT	476.67	87.09	373.33	86.56	483.33	85.37
FedEM-FedBN-FedOPT	330.00	80.48	730.00	85.01	633.33	83.87
FedEM-FedBN-FedOPT-FT	876.67	86.23	376.67	86.58	430.00	84.98

Table 13: The efficiency-accuracy trade-off results including total FLOPS, communication bytes (Com.), and $\overline{Acc}$ for CIFAR10 datasets with different α.

	CIFAR10, $\alpha = 5$			CIFAR10, $\alpha = 0.5$			CIFAR10, $\alpha = 0.1$		
	FLOPS	Com.	$\overline{Acc}$	FLOPS	Com.	$\overline{Acc}$	FLOPS	Com.	$\overline{Acc}$
FedAvg	**388.27G**	654.43K	73.86	384.14G	646.69K	70.82	305.15G	513.76K	56.21
FedAvg-FT	422.66G	685.52K	73.89	404.17G	654.46K	70.12	259.72G	425.4K	53.99
FedOpt	1.26T	862.74K	54.66	494.38G	833.2K	49.90	651.29G	1.06M	36.41
FedOpt-FT	1.37T	917.13K	57.33	1003.55G	1.59M	51.70	1.36T	895.36K	41.56
pFedMe	3.26T	2.07M	73.92	3.95T	918.26K	70.30	2.51T	594.98K	55.70
pFedMe-FT	3.57T	2.22M	**77.73**	3.57T	825.05K	**72.76**	4.15T	969.52K	**58.65**
FedBN	**388.27G**	540.02K	72.24	375.0G	520.8K	68.32	632.15G	350.44K	56.36
FedBN-FT	394.14G	527.19K	72.55	356.18G	475.9K	67.45	**249.57G**	337.61K	50.52
FedBN-FedOPT	581.98G	809.39K	71.63	581.92G	809.43K	67.13	660.49G	360.45K	57.17
FedBN-FedOPT-FT	638.27G	854.28K	72.32	**288.66G**	**386.11K**	67.72	381.03G	**206.52K**	49.35
Ditto	1.81T	**491.25K**	72.02	1.31T	685.55K	67.96	929.51G	479.8K	45.67
Ditto-FT	4.24T	1.12M	68.59	1.22T	631.15K	58.52	1001.19G	510.88K	37.79
Ditto-FedBN	2.21T	495.12K	71.70	1.87T	418.18K	62.22	940.86G	392.52K	41.59
Ditto-FedBN-FT	5.71T	1.24M	67.90	1.3T	552.87K	57.14	953.17G	392.52K	37.21
Ditto-FedBN-FedOpt	2.33T	520.77K	71.58	1.68T	552.87K	60.78	1.36T	578.53K	47.94
Ditto-FedBN-FedOpt-FT	3.09T	687.53K	68.41	1.01T	431.01K	56.66	1.18T	495.14K	38.73
FedEM	6.78T	1.41M	73.34	7.42T	1.55M	70.56	5.74T	1.21M	53.43
FedEM-FT	12.49T	1.81M	72.89	12.05T	1.74M	65.47	9.75T	1.45M	44.98
FedEM-FedBN	6.88T	1.18M	71.43	8.58T	1.46M	68.42	7.61T	741.57K	55.30
FedEM-FedBN-FT	9.91T	1.17M	70.46	11.44T	1.35M	61.43	9.01T	704.68K	43.48
FedEM-FedBN-FedOPT	11.2T	1.91M	71.43	12.79T	2.18M	67.11	10.63T	1.01M	56.87
FedEM-FedBN-FedOPT-FT	9.45T	1.12M	71.32	8.85T	1.05M	62.03	9.6T	758.5K	41.10

Table 14: The convergence results including the convergence round T' and $\overline{Acc}$ for CIFAR10 datasets with different α. The $T' = 0$ indicates the early-stopping is not triggered within the large number of FL rounds $T = 1000$.

	CIFAR10, $\alpha = 5$		CIFAR10, $\alpha = 0.5$		CIFAR10, $\alpha = 0.1$	
	T'	$\overline{Acc}$	T'	$\overline{Acc}$	T'	$\overline{Acc}$
FedAvg	280.00	73.87	276.67	70.89	213.33	57.71
FedAvg-FT	293.33	73.86	280.00	69.94	173.33	54.47
FedOpt	370.00	54.66	356.67	49.61	466.67	37.30
FedOpt-FT	393.33	57.35	696.67	51.69	383.33	42.01
pFedMe	573.33	73.74	393.33	70.63	243.33	56.46
pFedMe-FT	643.33	**77.82**	353.33	**73.46**	400.00	**59.74**
FedBN	280.00	72.17	270.00	68.28	173.33	57.59
FedBN-FT	273.33	72.50	246.67	67.69	166.67	51.40
FedBN-FedOPT	420.00	71.59	420.00	67.22	186.67	58.19
FedBN-FedOPT-FT	443.33	72.27	200.00	68.07	**106.67**	50.76
Ditto	210.00	71.97	293.33	67.65	196.67	49.77
Ditto-FT	490.00	68.57	270.00	58.63	210.00	38.97
Ditto-FedBN	256.67	71.65	216.67	62.65	203.33	43.07
Ditto-FedBN-FT	660.00	67.78	286.67	57.27	203.33	39.13
Ditto-FedBN-FedOpt	270.00	71.54	286.67	61.12	300.00	48.27
Ditto-FedBN-FedOpt-FT	356.67	68.37	223.33	57.59	256.67	40.67
FedEM	213.33	73.36	233.33	70.56	173.33	54.52
FedEM-FT	273.33	72.84	263.33	65.49	210.00	42.92
FedEM-FedBN	216.67	71.40	270.00	68.64	133.33	57.82
FedEM-FedBN-FT	216.67	70.40	250.00	62.13	126.67	43.97
FedEM-FedBN-FedOPT	353.33	71.39	403.33	67.19	186.67	58.40
FedEM-FedBN-FedOPT-FT	**206.67**	71.27	**193.33**	62.31	136.67	40.70

Table 15: The efficiency-accuracy trade-off results including total FLOPS, communication bytes (Com.), and $\overline{Acc}$ for COLA, SST-2 and Twitter datasets.

	COLA			SST-2			Twitter		
	FLOPS	Com.	$\overline{Acc}$	FLOPS	Com.	$\overline{Acc}$	FLOPS	Com.	$\overline{Acc}$
FedAvg	**22.15G**	310.94K	71.85	954.66G	464.37K	74.88	726.45K	60.32K	62.15
FedAvg-FT	37.23G	251.93K	68.29	1.41T	629.61K	74.14	4.75M	19.26K	70.53
FedOpt	50.71G	299.14K	71.85	808.78G	393.56K	72.28	726.45K	60.32K	62.09
FedOpt-FT	58.3G	393.56K	62.59	707.21G	310.94K	65.77	4.52M	18.35K	71.08
pFedMe	208.22G	1.04M	74.40	4.66T	830.51K	71.27	**603.46K**	29.3K	63.45
pFedMe-FT	240.44G	357.98K	**78.47**	3.76T	629.4K	75.61	12.07M	**15.57K**	**84.00**
FedBN	46.07G	259.58K	71.85	954.66G	464.37K	74.88	-	-	-
FedBN-FT	40.29G	259.58K	66.71	1.06T	476.17K	68.81	-	-	-
FedBN-FedOPT	86.98G	482.11K	71.85	**593.16G**	270.71K	64.70	-	-	-
FedBN-FedOPT-FT	39.08G	**248.45K**	67.48	707.21G	292.96K	68.65	-	-	-
Ditto	133.52G	322.75K	55.46	1.43T	275.53K	52.03	2.64M	38.42K	70.23
Ditto-FT	124.81G	440.77K	72.11	1.85T	334.55K	56.49	4.2M	36.59K	69.99
Ditto-FedBN	96.29G	404.22K	70.69	1.48T	270.71K	56.03	-	-	-
Ditto-FedBN-FT	113.99G	381.97K	72.66	1.58T	**270.7K**	53.15	-	-	-
Ditto-FedBN-FedOpt	163.88G	370.84K	50.25	3.19T	626.86K	57.67	-	-	-
Ditto-FedBN-FedOpt-FT	148.0G	292.96K	55.01	1.58T	537.74K	52.89	-	-	-
FedEM	414.28G	801.7K	71.85	18.46T	1.55M	**75.78**	18.82M	95.64K	63.44
FedEM-FT	1.85T	1.02M	54.90	24.19T	1.32M	64.86	24.19M	35.27K	70.97
FedEM-FedBN	729.78G	753.83K	71.44	16.84T	1.31M	75.43	-	-	-
FedEM-FedBN-FT	1.37T	979.64K	57.62	24.19T	1.24M	64.96	-	-	-
FedEM-FedBN-FedOPT	414.28G	753.83K	71.85	13.34T	1.05M	72.25	-	-	-
FedEM-FedBN-FedOPT-FT	1.49T	1.02M	57.23	22.0T	1.12M	62.26	-	-	-

Table 16: The convergence results including the convergence round T' and $\overline{Acc}$ for COLA, SST-2 and Twitter datasets. The $T' = 0$ indicates the early-stopping is not triggered within a large number of FL rounds, $T = 500$ for COLA and SST-2, and $T = 1000$ for Twitter datasets.

	COLA		SST-2		Twitter	
	T'	$\overline{Acc}$	T'	$\overline{Acc}$	T'	$\overline{Acc}$
FedAvg	43.33	51.53	65.00	76.30	223.33	56.98
FedAvg-FT	**35.00**	58.74	88.33	75.36	70.67	68.45
FedOpt	41.67	57.10	55.00	73.78	220.33	59.95
FedOpt-FT	55.00	59.77	43.33	66.17	66.67	69.20
pFedMe	150.00	67.58	116.67	65.08	106.67	58.18
pFedMe-FT	50.00	**69.17**	88.33	74.36	53.33	**78.82**
FedBN	38.33	59.60	65.00	76.30	-	-
FedBN-FT	38.33	59.69	66.67	68.50	-	-
FedBN-FedOPT	71.67	59.60	40.00	65.59	-	-
FedBN-FedOPT-FT	36.67	59.10	43.33	68.42	-	-
Ditto	45.00	55.14	**38.33**	49.94	140.33	66.90
Ditto-FT	61.67	63.61	46.67	54.34	133.33	66.91
Ditto-FedBN	60.00	62.68	40.00	49.44	-	-
Ditto-FedBN-FT	56.67	63.58	40.00	52.18	-	-
Ditto-FedBN-FedOpt	55.00	52.48	93.33	55.61	-	-
Ditto-FedBN-FedOpt-FT	43.33	57.13	80.00	53.16	-	-
FedEM	38.33	51.52	76.67	**76.53**	163.33	61.70
FedEM-FT	50.00	57.80	65.00	64.29	**40.67**	70.19
FedEM-FedBN	38.33	57.95	68.33	75.06	-	-
FedEM-FedBN-FT	50.00	58.74	65.00	64.33	-	-
FedEM-FedBN-FedOPT	38.33	59.60	55.00	72.66	-	-
FedEM-FedBN-FedOPT-FT	53.33	56.74	58.33	58.42	-	-

Table 17: The convergence results including the convergence round T' and $\overline{Loss}$ for Movielens-1M and Movielens-10M datasets. The $T' = 0$ indicates the early-stopping is not triggered within the large number of FL rounds $T = 1000$.

	Movielens-1M		Movielens-10M	
	T'	$\overline{Loss}$	T'	$\overline{Loss}$
FedAvg	360.33	0.85	470.67	**0.71**
FedAvg-FT	0	0.85	520.33	**0.71**
FedAvg-FT-FedOpt	830.0	0.86	0	0.75
FedOpt	**270.0**	0.84	0	0.72
FedOpt-FT	300.0	0.84	0	0.77
pFedMe	0	**0.55**	**280.33**	12.48
pFedMe-FT	470.0	0.60	840.00	0.80
Ditto	360.67	1.31	910.67	1.81
Ditto-FT	450.33	1.35	0	2.30
Ditto-FT-FedOpt	550.67	1.35	0	1.98
FedEM	700.33	0.87	0	2.37
FedEM-FT	780.00	0.87	0	0.98
FedEM-FT-FedOpt	0	0.87	0	1.88

Table 18: The efficiency-accuracy trade-off results including total FLOPS, communication bytes (Com.), and $\overline{Acc}$ for Pubmed, Cora, and Citeseer datasets.

	PUBMED			CORA			CITESEER		
	FLOPS	Com.	$\overline{Acc}$	FLOPS	Com.	$\overline{Acc}$	FLOPS	Com.	$\overline{Acc}$
FedAvg	40.41G	909.81K	87.27	2.94G	980.35K	81.30	71.06G	2.98M	75.58
FedAvg-FT	31.39G	1.79M	87.21	5.16G	885.92K	82.07	77.54G	2.39M	75.63
FedOpt	163.3G	6.92M	67.38	2.22G	744.07K	70.70	**3.89G**	437.18K	71.59
FedOpt-FT	24.78G	791.28K	82.36	6.27G	1.05M	82.68	25.07G	791.28K	74.34
pFedMe	45.67G	1.0M	86.91	10.02G	1.26M	83.18	193.83G	744.07K	75.30
pFedMe-FT	119.12G	862.1K	85.71	22.76G	2.09M	82.11	64.31G	460.79K	75.35
FedBN	27.53G	875.47K	**88.49**	2.51G	615.38K	84.13	14.01G	441.99K	75.80
FedBN-FT	35.86G	2.04M	87.45	29.13G	4.85M	76.20	28.25G	531.61K	75.07
FedBN-FedOPT	**13.26G**	**632.72K**	87.87	**2.07G**	736.76K	84.64	26.08G	823.45K	76.20
FedBN-FedOPT-FT	45.55G	1.04M	87.54	3.21G	615.38K	84.10	21.97G	771.43K	**76.70**
Ditto	46.9G	909.53K	87.27	7.27G	933.13K	83.67	79.77G	838.49K	74.79
Ditto-FT	43.73G	1.54M	87.47	8.59G	909.31K	81.47	42.34G	744.07K	76.47
Ditto-FedBN	23.72G	752.63K	88.18	8.37G	788.77K	81.38	37.51G	528.68K	75.35
Ditto-FedBN-FT	36.49G	963.57K	87.83	5.46G	**424.65K**	83.25	26.06G	494.0K	75.97
Ditto-FedBN-FedOpt	57.84G	821.99K	87.81	7.45G	702.08K	82.54	20.14G	**407.31K**	75.07
Ditto-FedBN-FedOpt-FT	49.88G	650.06K	87.60	7.92G	615.38K	82.00	35.48G	684.74K	76.42
FedEM	575.33G	3.24M	85.64	61.17G	2.01M	81.92	760.69G	8.36M	75.41
FedEM-FT	647.22G	3.04M	85.88	61.14G	2.7M	77.32	150.55G	2.22M	72.71
FedEM-FedBN	253.94G	2.07M	88.12	18.99G	1.92M	**85.07**	52.64G	1.82M	74.90
FedEM-FedBN-FT	275.77G	3.02M	86.38	34.58G	1.72M	84.61	89.54G	1.47M	75.29
FedEM-FedBN-FedOPT	229.51G	1.87M	87.56	108.73G	2.62M	84.68	419.47G	3.37M	76.08
FedEM-FedBN-FedOPT-FT	233.63G	2.12M	87.49	285.87G	4.92M	85.02	561.46G	3.22M	76.59

Table 19: The convergence results including the convergence round T' and $\overline{Acc}$ for Pubmed, Cora, and Citeseer datasets. The T' indicates the early-stopping is not triggered within the large number of FL rounds $T = 500$.

	PUBMED		CORA		CITESEER	
	T'	$\overline{Acc}$	T'	$\overline{Acc}$	T'	$\overline{Acc}$
FedAvg	63.33	86.72	68.33	81.07	215.00	75.64
FedAvg-FT	128.33	86.71	61.67	81.90	171.67	75.77
FedOpt	**0.00**	66.69	51.67	70.17	**30.00**	71.60
FedOpt-FT	55.00	81.53	75.00	82.31	55.00	74.38
pFedMe	71.67	86.35	90.00	82.76	51.67	75.36
pFedMe-FT	60.00	85.47	150.00	81.98	31.67	75.40
FedBN	83.33	**87.97**	58.33	83.64	41.67	75.59
FedBN-FT	146.67	87.02	183.33	76.01	36.67	75.16
FedBN-FedOPT	60.00	87.43	70.00	84.11	78.33	76.33
FedBN-FedOPT-FT	101.67	87.02	58.33	83.79	73.33	**76.77**
Ditto	63.33	86.85	65.00	83.50	58.33	75.55
Ditto-FT	110.00	87.10	63.33	81.53	51.67	76.57
Ditto-FedBN	71.67	87.75	75.00	81.36	50.00	75.46
Ditto-FedBN-FT	91.67	87.43	**40.00**	82.21	46.67	76.05
Ditto-FedBN-FedOpt	78.33	87.27	66.67	82.38	38.33	75.19
Ditto-FedBN-FedOpt-FT	61.67	87.10	58.33	81.99	65.00	76.52
FedEM	78.33	85.05	48.33	81.72	203.33	75.49
FedEM-FT	73.33	85.54	65.00	78.43	53.33	72.88
FedEM-FedBN	68.33	87.63	63.33	84.45	60.00	74.29
FedEM-FedBN-FT	100.00	85.68	56.67	**84.57**	48.33	75.40
FedEM-FedBN-FedOPT	61.67	87.11	86.67	83.88	111.67	76.17
FedEM-FedBN-FedOPT-FT	70.00	87.16	163.33	84.40	106.67	76.74

Table 20: The efficiency-accuracy trade-off results including total FLOPS, communication bytes (Com.), and $\overline{Loss}$ for Movielens-1M and Movielens-10M datasets.

	Movielens-1M			Movielens-10M		
	FLOPS	Com.	$\overline{Loss}$	FLOPS	Com.	$\overline{Loss}$
FedAvg	343.73M	108.75K	0.84	375.55G	142.58K	**0.70**
FedAvg-FT	995.55M	301.5K	0.84	162.79G	157.72K	0.71
FedAvg-FT-FedOpt	236.68M	250.45K	0.85	21.39G	302.91K	0.73
FedOpt	258.31M	**81.62K**	0.83	**19.75G**	302.91K	0.71
FedOpt-FT	299.3M	90.66K	0.83	52.04G	302.91K	0.74
pFedMe	763.2M	301.51K	**0.54**	131.49G	**24.44K**	13.06
pFedMe-FT	376.37M	142.35K	0.60	93.58G	254.7K	0.80
Ditto	332.94M	108.75K	1.29	72.4G	302.91K	1.84
Ditto-FT	**220.62M**	135.88K	1.35	242.74G	302.91K	1.69
Ditto-FT-FedOpt	269.63M	166.03K	1.36	73.26G	302.91K	2.03
FedEM	2.07G	523.1K	0.85	86.69G	135.15K	1.75
FedEM-FT	4.07G	582.82K	0.85	253.37G	269.78K	0.87
FedEM-FT-FedOpt	5.22G	746.53K	0.86	113.27G	120.19K	1.43

Table 21: Efficiency results in terms of process memory (MB) and running time (seconds). A higher value indicates that more system resources were consumed. The MEM_{avg} and MEM_{peak} indicates average and peak values of process-used memory respectively, and T_{run} indicates the process running time. Similar to above table, **Bold** and underlined indicate the best and second-best results among all compared methods, while red and blue indicate the best and second-best results for original methods without combination "-".

	FEMNIST, $s = 0.2$			SST-2		
	MEM_{avg}	MEM_{peak}	T_{run}	MEM_{avg}	MEM_{peak}	T_{run}
Global-Train	**86.56**	**86.56**	**11.00**	**87.23**	**93.37**	892.00
Isolated	746.85	1351.94	1108.00	118.00	151.07	994.00
FedAvg	10506.71	17707.04	840.00	297.43	399.51	319.00
FedAvg-FT	13400.46	21752.04	1193.00	314.05	371.73	545.00
FedProx	19389.09	20729.26	1415.00	6635.17	7378.15	672.00
FedProx-FT	21209.42	22503.71	1935.00	7804.79	8160.99	1810.00
pFedMe	1880.78	1979.59	7205.00	262.67	366.15	1572.00
pFedMe-FT	18573.50	21332.41	4676.00	282.19	367.81	1080.00
HypCluster	21659.61	22386.94	2803.00	6942.08	7510.02	678.00
HypCluster-FT	22569.13	23588.59	3602.00	7624.09	8232.55	887.00
FedBN	11135.25	15406.92	1011.00	279.89	373.27	332.00
FedBN-FT	17347.31	25649.89	936.00	266.80	331.84	551.00
FedBN-FedOPT	21866.64	35335.47	2881.00	216.22	311.33	**202.00**
FedBN-FedOPT-FT	9827.22	14711.30	845.00	260.59	350.32	283.00
Ditto	1127.92	1268.18	4628.00	149.16	204.95	638.00
Ditto-FT	1110.87	1532.88	8915.00	228.51	292.20	599.00
Ditto-FedBN	1116.05	1227.17	4049.00	196.52	228.59	593.00
Ditto-FedBN-FT	1453.92	1730.06	8528.00	216.48	290.13	627.00
Ditto-FedBN-FedOpt	1176.76	1273.67	5782.00	202.71	237.94	560.00
Ditto-FedBN-FedOpt-FT	1299.42	1564.89	7545.00	275.08	333.58	637.00
FedEM	939.58	1063.35	5070.00	267.37	357.79	1568.00
FedEM-FT	563.77	619.02	7592.00	327.36	374.10	3210.00
FedEM-FedBN	572.14	912.14	12051.00	269.62	337.22	1615.00
FedEM-FedBN-FT	489.82	616.71	9391.00	306.51	367.51	3040.00
FedEM-FedBN-FedOPT	755.61	842.33	8977.00	251.02	343.16	1314.00
FedEM-FedBN-FedOPT-FT	608.29	638.73	19137.00	300.92	353.46	3116.00