# FedSlice: Secure and Communication-efficient Federated Learning by Model Slicing

## A. Experimental Details on Non-IID Datasets

### A.1. Dirichlet Data Distributions.

In this section, we will use the Dirichlet distribution to simulate the scenario of non-independent and identically distributed (non-IID) data in federated learning on edge devices. As there are various manifestations of non-IID data, although we present in Appendix B the experimental results where the volume of datasets in each edge device follows a normal distribution, the distribution of dataset labels owned by each edge device in this scenario is actually uniform. Therefore, we adopt the popular Dirichlet distribution as another experiment for non-IID datasets, that is, we conduct some experiments on the situation where the volume of same-category labels in each edge device follows a Dirichlet distribution.

$$q \sim \mathrm{Dir}(\alpha p) \tag{1}$$

The size of the labeled categories in the dataset is assumed to be M. These datasets include Fashion-MNIST (Xiao et al., 2017), CIFAR10 (Krizhevsky et al., 2009) and AG News (Zhang et al., 2015) used in previous experiments, the number of training samples on each edge device is obtained by sampling these source datasets with a probability distribution as the parameter vector $q$ ($q_i \geq 0, i \in [1, M]$). In order to generate non-IID data samples, sampling can be performed according to the probability of the Dirichlet distribution corresponding to equation, where the probability vector p represents the prior distribution of M categories of labels, and the parameter $\alpha$ controls the heterogeneity of the sample data produced by the edge devices. The distribution of sampled data according to the probability distribution $q$ in equation (1) changes with the value of the parameter $\alpha$. When $\alpha \to \infty$, the data distribution on each edge device produced is consistent with the prior distribution of the original datasets. In another extreme case, when $\alpha \to 0$, the sample data of each edge device will only be randomly sampled from one category of samples. Simply put, the larger the parameter $\alpha$, the closer the distribution is to a uniform distribution; the smaller the parameter $\alpha$, the more concentrated the distribution is.

We will conduct some experiments under $\alpha = 1$ and $\alpha = 100$, i.e., the volume of data samples for each label in the training set on each edge device follows $bR_k \sim Dir(\alpha p)$, specifically, $bR_k \sim Dir(p)$ or $bR_k \sim Dir(100p)$, where the vector p is an all-ones vector of size M. In our algorithm, the weight of each edge device is set to $|bR_k|$, that is, the size of the dataset owned by each edge device.

### A.2. Datasets.

**Fashion-MNIST** is a dataset comprising of 28×28 grayscale images of 70,000 fashion products from 10 categories, with 7,000 images per category (Xiao et al., 2017). We divide these 70,000 images into a training set and a test set, where the training set has 60,000 images, and the test set has 10,000 images. The training set for each edge device is randomly sampled from these 60,000 images. The remaining 10,000 images will be allocated to the global server for evaluating the accuracy of the model in each global epoch. Since the Fashion-MNIST dataset is relatively simple, we used the most classical CNN model containing 32,340 parameters.

**CIFAR10** dataset (Canadian Institute for Advanced Research, 10 classes) is a subset of the Tiny Images dataset and consists of 60,000 32x32 color images (Krizhevsky et al., 2009). Similar to the setup of the Fashion-MNIST dataset, we divide the CIFAR10 dataset into a training set containing 50,000 images and a test set of 10,000 images. The training set for each edge device is randomly sampled from these 50,000 images. The remaining 10,000 images will be allocated to the global server for evaluating the accuracy of the model in each epoch of the global competition.

**AG News (AG's News Corpus)** is a subdataset of AG's corpus of news articles constructed by assembling titles and description fields of articles from the 4 largest classes (Zhang et al., 2015). The AG News contains 30,000 training and 1,900 test samples per class. Similar to the setup of the Fashion-MNIST dataset and the CIFAR10 dataset, we slice the AG

*Table 1.* Top-1 accuracy on the Fashion-MNIST, CIFAR10, and AG News datasets, where CR denotes the communication compression rate, the total number of edge devices is $K$, while the number of edge devices involved in training for each experiment is $k$, and $bR_k \sim Dir(100 * p)$ or $bR_k \sim Dir(100 * p)$ denotes the different size distribution of the training set over these edge devices.

| Algorithms | $K$ | $k$ | $CR$ | Acc. of Fashion-MNIST | | Acc. of CIFAR10 | | Acc. of AG News | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $bR_k \sim Dir(p)$ | $bR_k \sim Dir(100p)$ | $bR_k \sim Dir(p)$ | $bR_k \sim Dir(100p)$ | $bR_k \sim Dir(p)$ | $bR_k \sim Dir(100p)$ |
| FDQM | 256 | 128 | 1x | 74.14% | 77.21% | 68.43% | 70.50% | 86.03% | 87.14% |
| FDQM | 256 | 256 | 1x | 76.75% | 79.54% | 69.96% | 73.77% | 86.13% | 88.53% |
| FDQM | 2048 | 1024 | 1x | 82.40% | 84.92% | 72.12% | 76.02% | 86.91% | 89.91% |
| FDQM | 2048 | 2048 | 1x | **83.69**% | **88.00**% | **76.91**% | **77.46**% | 87.63% | 89.40% |
| PerFedAvg | 256 | 128 | 1x | 72.25% | 79.57% | 65.46% | 69.92% | 87.61% | 87.55% |
| PerFedAvg | 256 | 256 | 1x | 74.59% | 79.41% | 66.12% | 70.87% | 87.18% | 87.91% |
| PerFedAvg | 2048 | 1024 | 1x | 76.47% | 80.23% | 71.09% | 72.45% | **88.20**% | 89.66% |
| PerFedAvg | 2048 | 2048 | 1x | 79.35% | 82.05% | 74.91% | 76.10% | 87.43% | **90.58**% |
| CSER | 256 | 128 | 128x | 56.21% | 65.23% | 50.42% | 56.80% | 71.14% | 73.25% |
| CSER | 256 | 256 | 256x | 62.45% | 68.80% | 51.26% | 57.73% | 71.26% | 73.53% |
| CSER | 2048 | 1024 | 1024x | 65.21% | 70.38% | 54.65% | 57.96% | 72.21% | 73.94% |
| CSER | 2048 | 2048 | 2048x | 67.89% | 74.14% | 54.31% | 59.42% | 72.45% | 75.43% |
| TopK-EF21 | 256 | 128 | 128x | 60.22% | 66.50% | 52.29% | 56.55% | 70.56% | 71.05% |
| TopK-EF21 | 256 | 256 | 256x | 62.57% | 66.45% | 52.02% | 57.30% | 72.65% | 72.31% |
| TopK-EF21 | 2048 | 1024 | 1024x | 65.73% | 68.92% | 55.31% | 59.80% | 71.48% | 75.94% |
| TopK-EF21 | 2048 | 2048 | 2048x | 66.42% | 67.65% | 56.65% | 61.27% | 72.40% | 76.39% |
| FedSlice | 256 | 128 | 128x | 69.57% | 75.62% | 60.46% | 64.12% | 83.60% | 85.14% |
| FedSlice | 256 | 256 | 256x | 72.39% | 76.42% | 64.11% | 65.07% | 85.28% | 86.77% |
| FedSlice | 2048 | 1024 | 1024x | 75.43% | 78.55% | 69.33% | 71.59% | 86.21% | 88.40% |
| FedSlice | 2048 | 2048 | 2048x | 79.78% | 83.75% | 70.61% | 72.67% | 87.70% | 88.05% |

News dataset into a training set containing 120,000 description fields and a test set of 7,600 description fields. The training set for each edge device is randomly sampled from these 120,000 description fields. The remaining 7,600 description fields will be allocated to the global server for evaluating the accuracy of the model in each epoch of the global competition.

### A.3. Models.

The neural network models utilized for the experiments in this section are consistent with those used in the main text and Appendix B. Specifically, we employed the Convolutional Neural Network (CNN) model, the ResNet-18 (Szegedy et al., 2017), and Encoder-Decoder (Cho et al., 2014) model (LSTM as an encoder and GRU as a decoder (Fu et al., 2016) with an attention mechanism (Hu, 2020)).

### A.4. Experimental Details.

We evaluated these algorithms using the test set accuracy for each training epoch on various neural network models. We compared the top-1 accuracy of our algorithm with the CSER (Xie et al., 2020), TopK-EF21 (Richtárik et al., 2021), PerFedAvg (Fallah et al., 2020), and FDQM (Guo et al., 2022) algorithms under the same Dirichlet data distribution. Our algorithm, along with CSER and TopK-EF21, still compresses communication, a capability that PerFedAvg and FDQM algorithms do not possess. Since our algorithm's strength is communication compression, it may not have an advantage under the condition of non-independent and identically distributed data.

For experimental validation, we utilized the CNN model on the MNIST dataset, where the global epoch was set to 150. We employed the ResNet-18 model for training on the CIFAR10 dataset over 500 epochs. For the AG News dataset, we set a global training epoch of 250. These configurations were chosen to best showcase the performance and capabilities of our algorithm across diverse datasets and model architectures.

To simulate a situation where not all edge devices participate in computations, we adhere to the assumptions outlined in the main text. That is, we assume the total number of edge devices is $K$, while the number of edge devices involved in training for each experiment is $k$. For each global epoch, the edge devices participating in the training are randomly selected from all available edge devices.

### A.5. Experimental Results.

Our experimental results are presented in Table 1. Under the condition of non-independent and identically distributed (non-IID) data, our algorithm clearly outperforms the CSER and TopK-EF21 algorithms in terms of accuracy. When the number of edge devices is relatively large ($k$=2048), our algorithm also maintains a relatively good accuracy rate compared to algorithms PerFedAvg and FDQM. While there are some experiments where the accuracy rate is not as good as these two algorithms, our algorithm has the significant advantage of reducing the amount of communication, a capability that algorithms PerFedAvg and FDQM do not possess. Similarly, for the CSER and TopK-EF21 algorithms, our algorithm achieves better accuracy under the same communication compression rate when the dataset has a Dirichlet distribution.

Additionally, in scenarios where not all edge devices are involved in the training (i.e., $K$!=$k$), our algorithm is capable of maintaining better accuracy and stability. This demonstrates the robustness of our algorithm in handling diverse and dynamic edge computing environments, further enhancing its applicability in real-world federated learning scenarios.

Overall, our algorithms are able to strike a balance between the communication compression rate and stability in non-IID scenarios. Non-IID scenarios include not only those where the data volume of edge devices follows a normal distribution (as presented in Appendix B), but also those where the labels of edge devices follow a Dirichlet distribution (as discussed in this section).

While the current iteration of the FedSlice algorithm has certain limitations, we remain committed to ongoing research and improvement. This includes exploring metrics of the weight factor indicators within the FedSlice algorithm. Potential avenues for exploration include incorporating the weight factor into the loss function, or calculate heterogeneity metrics for edge datasets with privacy guarantees. These approaches could potentially enhance the performance and applicability of our algorithm.

## References


Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Fallah, A., Mokhtari, A., and Ozdaglar, A. E. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Neural Information Processing Systems*, 2020. URL https://api.semanticscholar.org/CorpusID:227276412.

Fu, R., Zhang, Z., and Li, L. Using lstm and gru neural network methods for traffic flow prediction. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 324–328. IEEE, 2016.

Guo, Y., Tang, Q., Wang, Y., and Ren, J. Fdqm: Four-dimensional quantitative measure for statistical heterogeneity in federated learning. *2022 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 956–961, 2022. URL https://api.semanticscholar.org/CorpusID:252469462.

Hu, D. An introductory survey on attention mechanisms in nlp problems. In *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2*, pp. 432–448. Springer, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Richtárik, P., Sokolov, I., and Fatkhullin, I. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xie, C., Zheng, S., Koyejo, S., Gupta, I., Li, M., and Lin, H. Cser: Communication-efficient sgd with error reset. *Advances in Neural Information Processing Systems*, 33:12593–12603, 2020.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.