A Joint Training-Calibration Framework for Test-Time Personalization with Label Distribution Shift in Federated Learning

Jian Xu [1] **Shao-Lun Huang** [1]

Abstract

The data heterogeneity has been a challenging issue in federated learning in both training and inference stages, which motivates a variety of approaches to learn either personalized models for participating clients or test-time adaptations for unseen clients. One such approach is employing a shared feature representation and a customized classifier head for each client. However, previous works either do not utilize the global head with rich knowledge or assume the new clients have enough labeled data, which significantly limit their broader practicality. In this work, we propose a lightweight framework to tackle the label shift issue in model deployment by test priors estimation and model prediction calibration. We emphasize the importance of training a balanced global model in FL and the general effectiveness of prior estimation approaches. Numerical evaluation results on benchmark datasets with various label distribution shift cases demonstrate the superiority of our proposed framework.

1. Introduction

Personalized Federated Learning (PFL) (Tan et al., 2022; Kairouz et al., 2021) has attracted increasing attentions to tackle the data heterogeneity issue that becomes one of the key bottlenecks of federated learning (McMahan et al., 2017; Yang et al., 2019; Li et al., 2020). The heterogeneous data across clients are usually caused by either feature distribution shift or label distribution shift, as different devices generate/collect data separately and may have specific preferences, making it hard to learn a single global model that works well at all clients (Zhao et al., 2018; Zhu et al., 2021; Li et al., 2022). Different from the traditional FL, personalized approaches aim at learning a customized model for each client that has better performance than the global model. Such settings can be motivated by many realistic FL applications, where clients (e.g., hospitals and corporations) may wish to satisfy client-specific tasks.

While advanced algorithms have been developed for PFL, they also have some limitations as they only generate personalized models during the training stage and rely on enough labeled local data for unseen/new clients during the inference stage (Collins et al., 2021; Shamsian et al., 2021). Moreover, the resulting personalized models naturally lose the robustness to label distribution shift, since the personalized model may only fit the local labeled class distribution while the global model obtained by FL could generalize to all classes (Jiang & Lin, 2023). Besides, it is possible that some devices do not support the model re-training after deployment due to hardware/software constraints or model intellectual property protection (Sun et al., 2021).

Adapting the learned model to new clients with low cost after FL is both challenging and important for real-world applications. To this end, we adapt the trained global model to new clients by only calibrating the classifier prediction without global knowledge forgetting. We focus on the label distribution shift scenarios, where the target classification categories could vary for new clients. To perform classifier calibration, both the a-priori label distributions of training and test sets are essentially required (Alexandari et al., 2020; Tian et al., 2020). The key challenge is how to evaluate the label distribution information in the FL systems with the privacy constraints and only unlabeled test data for new clients, or even online emerging data points. For this purpose, we apply the balanced-softmax (Ren et al., 2020) to debias the global model training and utilize the historic prediction information to obtain reliable test-time prior estimations.

Our contributions. We propose a simple yet effective PFL framework with flexible prediction Calibration (`FedCal`) to improve the overall performance of client-specific tasks. The proposed framework does not rely on any extra labeled data and also not modify the model parameters, only making use of some prediction statistics to dynamically adjust the class priors during the inference, incuring minimal compu-

[1]Tsinghua-Berkeley Shenzhen Institute, Shenzhen International Graduate School, Tsinghua University. Correspondence to: Jian Xu <xujian20@mails.tsinghua.edu.cn>, Shao-Lun Huang <shaolun.huang@sz.tsinghua.edu.cn>.

Workshop of Federated Learning and Analytics in Practice, colocated with 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA.

tational costs. Evaluation results on benchmark datasets with label shift verified the effectiveness in achieving higher performance than the plain global model prediction on test sets with unknown label shifts.

2. Related Work

Most personalized FL methods achieve the model adaptation during the training process for the participating clients (Deng et al., 2020; Dinh et al., 2020; Li et al., 2021b; Zhang et al., 2021b) or by local fine-tuning after global training based on local labeled data (Yu et al., 2020; Cheng et al., 2021). In particular, model decoupling of feature extractor and classifier head is widely studied (Arivazhagan et al., 2019; Liang et al., 2020; Collins et al., 2021). Combing global and local classifiers after FL is investigated in (Chen & Chao, 2022) and (Marfoq et al., 2022). A mixture of multiple global models with theoretical interpretation and new clients generalization is also proposed (Marfoq et al., 2021; Wu et al., 2023). Besides, a server-side hypernetwork is employed for generating customized local models (Shamsian et al., 2021; Amosy et al., 2022). However, most methods cannot customize the model to new clients without labeled data. There is also a line of work addressing the test-time label shift issues in the centralized settings (Azizzadenesheli et al., 2019; Lipton et al., 2018; Alexandari et al., 2020; Garg et al., 2020; Zhang et al., 2021a; Ma et al., 2022), ***however the training set is not directly accessible in FL***. Only few works tackle the test-time adaptation to new distributions in the context of FL (Jiang & Lin, 2023; Amosy et al., 2022).

3. Preliminary and Motivation

3.1. Problem Setup

We consider a setup with m clients and a central server, where each client i is equipped with its own data distribution $P_{XY}^{(i)}$ on $\mathcal{X} \times \mathcal{Y}$, and $P_{XY}^{(i)}$ and $P_{XY}^{(j)}$ could be different for any pair of client i and j. We also assume that the server does not have any prior knowledge about the data distribution of the participating clients, which is usually the case in FL. Let $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ denote the loss function given local model $\boldsymbol{w}_i$ and data point ξ_i sampled from $P_{XY}^{(i)}$, e.g., cross-entropy loss, then the underlying optimization goal of PFL can be formalized as

$$\min_{\mathbf{w}} \left\{ F(\mathbf{w}) := \frac{1}{m} \sum_{i=1}^{m} \mathcal{E}_{\xi_i} \left[\ell(\boldsymbol{w}_i; \xi_i) + \mathcal{R}(\boldsymbol{w}_i, \boldsymbol{w}_g) \right] \right\}, \quad (1)$$

where $\mathbf{w} = (\boldsymbol{w}_g, \boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_m)$ denotes the collection of global model and all local models. $\mathcal{R}$ is a regularization between local and global models. The global model $\boldsymbol{w}_g$ can be further used for new clients. The feature embedding function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$ is a learnable network parameterized by $\boldsymbol{\theta}_f$ and d is the dimension of feature embedding. The linear classifier head $\boldsymbol{g}$ parameterized by $\boldsymbol{\phi}_g$ is responsible for making predictions. For example, LG-FedAvg (Liang et al., 2020) keeps $\boldsymbol{\theta}_f$ local and only shares the $\boldsymbol{\phi}_g$, while FedRep (Collins et al., 2021) shares $\boldsymbol{\theta}_f$ and keeps $\boldsymbol{\phi}_g$ local. However, some clients may only participate the FL system after the training phase and do not have labeled data to train the missing model part. In such cases, how to generate a customized model is of significantly importance.

3.2. Local Prediction Calibration

Unlike previous works that maintain the local trained classifier, we focus on generalize to new clients without labeled data or new test sets for old clients during the deployment stage. Therefore, a global classifier head that is capable to discriminate all possible categories is essentially required. Rather than combining the global and local heads (Chen & Chao, 2022; Jiang & Lin, 2023), we argue that simply calibrating the classifier prediction can offer significantly higher performance than many personalization approaches.

Probabilistic interpretation of model predictions. As a common interpretation, training a deep network by minimizing the cross-entropy actually tries to approximate the true conditional distributions:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} -\frac{1}{n} \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \boldsymbol{w}), \quad (2)$$

where the loss is computed over samples $\mathbf{x}_i$ with known labels y_i, $\boldsymbol{w}^*$ are parameters of the trained model. Let us assume that a model is well-trained to provide an good estimate of posterior probabilities of classes $c_1, \ldots, c_K \in [K]$ given a test point $\mathbf{x}_i \in \mathcal{X}$:

$$f(c_k|\mathbf{x}_i, \boldsymbol{w}^*) \approx p_s(c_k|\mathbf{x}_i). \quad (3)$$

When the prior class probabilities $p_t(c_k)$ in the test[1] set differ from the training set, the posterior $p_t(c_k|\mathbf{x}_i)$ changes as well. As we focus on label shift, we assume that the class-wise conditional distribution $p(\mathbf{x}_i|c_k)$ remain (almost) unchanged for the unknown test set, which describe the statistical properties of observations belonging to class c_k:

$$p(\mathbf{x}_i|c_k) = \frac{p_s(c_k|\mathbf{x}_i)p(\mathbf{x}_i)}{p_s(c_k)} = \frac{p_t(c_k|\mathbf{x}_i)p(\mathbf{x}_i)}{p_t(c_k)}. \quad (4)$$

Then the prediction calibration can be formulated as

$$p_t(c_k|\mathbf{x}_i) := \frac{p_t(c_k)}{p_s(c_k)} \cdot p_s(c_k|\mathbf{x}_i). \quad (5)$$

The training priors $p_s(c_k)$ can be empirically quantified as the proportion of samples labeled as c_k in the training set. The test-time priors $p_t(c_k)$ are, however, often unknown for new clients, especially for the online test deployment where data will not be stored due to limited storage capability. In FL, the training class priors are also difficult to know as local

[1]We use index s and t to denote the training (source) and test (target) distributions, respectively.

statistics may not be shared or the global model may not fit the overall priors perfectly due to partial client participation and non-linear change after global aggregation. However, if the global model is approximately unbiased, then the challenge of training priors estimation could be avoided by directly setting them as the uniform distributions.

3.3. Learning an Unbiased Global Model

While significant efforts have been devoted for better model training, most methods will introduce extra communication and computation costs (Karimireddy et al., 2020; Li et al., 2021a; Acar et al., 2021). Previous works have applied the balanced-softmax (BSM) for local training (Chen & Chao, 2022; Zhang et al., 2022), which is promising for resulting more balanced global model. In this work, we also follow this strategy due to its simplicity and utility. We leave the better global model training strategy as future works.

4. Test-Time Prior Estimation

In this work, we provide three methods for test-time prior estimation, which can be selected or combined flexibly.

4.1. Estimation by Pseudo Labeling

By assuming the model has a good fit to the training data distribution, we can estimate the prior distribution by counting the pseudo labels generated by the model prediction.

$$p_t(c_k) := \frac{1}{n}\sum_{i=1}^{n}\mathcal{I}\{c_k = \arg\max_c f(c|\mathbf{x}_i, \boldsymbol{w}^*)\}, \quad (6)$$

where $\mathcal{I}\{\cdot\}$ is the indication function and advanced sample selections could also be applied to obtain a better estimation.

4.2. Estimation by Maximum Likelihood

A theoretically sound approach to estimate the unknown test-time label distributions is maximizing the likelihood of the test observations (Saerens et al., 2002; du Plessis & Sugiyama, 2014; Alexandari et al., 2020):

$$\begin{aligned} L(\mathbf{x}_1, ...\mathbf{x}_n) &= \prod_{i=1}^{n} p_t(\mathbf{x}_i) = \prod_{i=1}^{n}\left[\sum_{k=1}^{K} p_t(\mathbf{x}_i, c_k)\right] \\ &= \prod_{i=1}^{n}\left[\sum_{k=1}^{K} p(\mathbf{x}_i|c_k)p_t(c_k)\right] \end{aligned} \quad (7)$$

To solve this problem, a simple EM algorithm can be derived with the following steps:

$$\textbf{E-Step:} \quad p_t^{(\tau)}(c_k|\mathbf{x}_i) = \frac{p_s(c_k|\mathbf{x}_i)p_t^{(\tau)}(c_k)}{\sum_{k=1}^{K} p_s(c_k|\mathbf{x}_i)p_t^{(\tau)}(c_k)} \quad (8)$$

$$\textbf{M-Step:} \quad p_t^{(\tau+1)}(c_k) = \frac{1}{n}\sum_{i=1}^{n} p_t^{(\tau)}(c_k|\mathbf{x}_i) \quad (9)$$

where τ is the iteration step index, Eq. (8) is the Expectation-step, Eq. (9) is the Maximization-step, and $p_t^0(c_k)$ could be initialized with a uniform distribution.

4.3. Estimation by Feature Matching

If the global class-wise prototypes $\mathbf{C} = \{\mathbf{c}_k\}_{k=1}^K$ (i.e., the mean feature representation for each class) are available by aggregating local prototypes, then it is possible to estimate the priors $\mathbf{p} = \{p_k\}_{k=1}^K$ by the following mean-feature matching (FM) optimization objective (we justify the rationality of this approach in Appendix A):

$$\begin{aligned} \mathbf{p} &= \arg\min_{\mathbf{p}} \frac{1}{2}\|\mathbf{C}^T\mathbf{p} - \mathbf{f}\|^2 \\ &= \arg\min_{\mathbf{p}} \frac{1}{2}\mathbf{p}^T\mathbf{C}\mathbf{C}^T\mathbf{p} - (\mathbf{C}\mathbf{f})^T\mathbf{p}, \\ &\text{s.t. } \sum_{k=1}^{K} p_k = 1;\ \forall k: p_k \geq 0. \end{aligned} \quad (10)$$

where $\mathbf{f}$ is the mean test feature representation of test observations. Note that this objective is a quadratic optimization problem and can be solved by the cvxpy toolbox efficiently.

4.4. Online Adaptive Calibration

Moreover, the test set might emerge continually and should be predicted online, instead of testing after collecting all the test data points (Wu et al., 2021; Bai et al., 2022). For the online test scenarios, the test data may not be stored due to limited storage capacity. Instead, we can record the original predicted posteriors for each test point as $\{\hat{\mathbf{y}}_i\}_{i=1}^n$, which are memory efficient and can be accumulated within the maximum buffer size to derive an increasingly better estimation of label distribution in the online settings. When the global prototypes are available, we can also record the point-wise low-dimentional feature representation.

4.5. Special Consideration

In many FL studies, the local test set only has partial categories but uniformly distributed (pathological case) (McMahan et al., 2017; Collins et al., 2021), we claim that with such a assumption, the estimated priors could be further rectified by simply setting the elements with 0/1. To this end, clustering- or threshold-based methods could be employed. Note that in the absence of such prior knowledge, it might be too radical to conduct such an estimate rectification.

5. Evaluation

Experimental Setup. We conduct experiments on four popular datasets with simple CNN models, including MNIST, Fashion-MNIST, CIFAR-10 and CINIC-10. ***We focus on the pathological cases*** and compare our three variants (PL-, EM-, FM-based) with popular PFL baselines as listed in Ta-

Table 1. The comparison of label-shift test accuracy (%) on different datasets. We apply full participation for FL system with 20 clients.

Method	MNIST		Fashion-MNIST		CIFAR-10		CINIC-10	
	Local Test	OoC Test	Local Test	OoC Test	Local Test	OoC Test	Local Test	OoC Test
Local-only	93.73	46.47	88.97	43.62	84.17	23.10	75.43	20.79
FedAvg	95.25	95.25	86.34	86.34	68.43	68.43	52.73	52.73
FedPer	96.07	77.83	**91.56**	67.02	88.35	26.41	81.21	28.46
FedRep	95.39	71.29	90.05	54.20	84.83	24.32	78.63	22.20
Ditto	**97.04**	90.55	90.80	71.46	**88.63**	36.64	81.21	27.11
Fed-RoD	96.43	58.25	91.28	78.15	87.60	50.19	81.32	36.38
kNN-Per	96.39	66.65	91.59	55.72	87.77	31.98	81.05	26.42
FedTHE	96.41	71.37	91.25	82.73	87.03	59.63	81.46	43.47
pFedHN	94.91	90.69	88.53	81.67	87.07	31.82	79.40	24.65
ODPFL-HN	94.07	94.06	86.32	86.17	64.82	64.80	50.63	50.47
FedCal-PL	96.68	**96.55**	91.41	**91.41**	88.48	87.92	**82.16**	69.11
FedCal-EM	\	**96.55**	\	90.35	\	**88.48**	\	**74.04**
FedCal-FM	\	95.62	\	**91.41**	\	88.07	\	70.25

ble 1. For new clients without labeled data, we only compare with FedAvg. More details are provided in Appendix B.

Performance Comparison. We first consider a setup with 20 clients and full participation. Similar to (Jiang & Lin, 2023), both in-client and ***out-of-client (OoC)*** tests are conducted to evaluate the generalization performance. From the results in Table 1, we can find that prediction calibration not only can achieve competitive performance with advanced PFL methods when given local true priors, but also dominates the OoC test with unknown label shifts, which demonstrates the effectiveness and benefit of calibration.

Generalize to New Clients. We vary the number of classes per-client C to simulate different levels of shift. The results in Fig. 1 clearly show that FedCal achieves better accuracy than the plain global model under all levels of label shift. We also consider the test data points emerge online and could have non-stationary prior transitions. We split the test process into multiple time slots (20 samples in each slot) and repeat prior estimation after each time slot by using information from the latest 5 slots. We further consider a step change of test prior with another set of randomly selected C classes to assess the algorithm resilience. The accuracy curves are plotted in Fig. 2, where our methods still outperform the FedAvg. However, we also find the accuracy fluctuates a lot, which means both the online calibration strategy and the inter-class fairness of the global model should be further improved.

Table 2. Ablation Studies on CIFAR-10 with two FL settings. Components: 1) Calibration, 2) rectification, 3) Balanced-SoftMax.

FL Setting	1)	1) 2)	1) 3)	1) 2) 3)
20 clients, 100%	83.13	87.86	84.69	**88.48**
100 clients, 10%	77.46	80.61	83.79	**86.78**

Effect of Design Components. To show the necessity of a balanced global model and the benefit of prior rectification, here we add a more challenging case with 100 clients and 10% participating ratio in each round on the CIFAR-10. Clients may not get the final global model but only the intermediate one, and a new unknown test set is randomly assigned to each client for OoC test. As reported in Table 2, without a balanced global model, the prior estimation is actually unreliable and the calibrated prediction accuracy is very low. In the full participating case, the impact is largely mitigated. In contrast, the combination of BSM and prior rectification can achieve the best results in both cases.

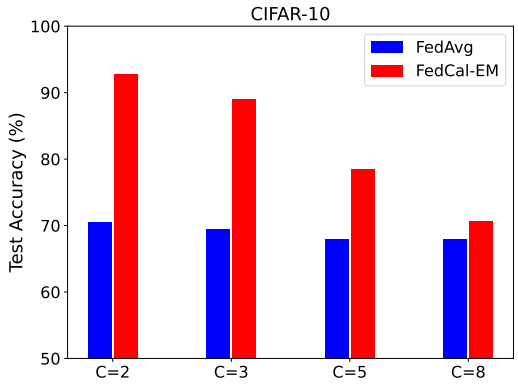


Figure 1. Static test on CIFAR-10 with various C.

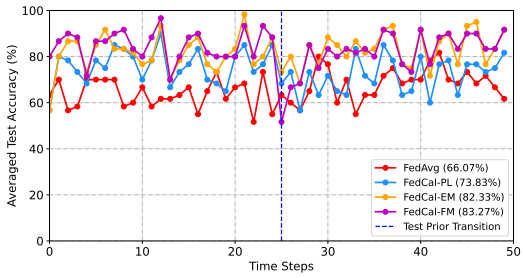


Figure 2. Comparison of online test on CIFAR-10 dataset.

6. Conclusion and Future Work

In this work, we introduce several calibration methods for building customized prediction in new clients, providing empirical justification for their utilities in label shift settings. Future work includes investigating the test-time adaptation with feature-level distribution shift, and developing a more robust global model against various test data shifts.

Acknowledgements

The research is supported in part by the Shenzhen Science and Technology Program under Grant KQTD20170810150821146, National Key R&D Program of China under Grant 2021YFA0715202.

References

Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., and Saligrama, V. Federated learning based on dynamic regularization. In *9th International Conference on Learning Representations, ICLR*, 2021.

Alexandari, A., Kundaje, A., and Shrikumar, A. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020.

Amosy, O., Eyal, G., and Chechik, G. On-demand unlabeled personalized federated learning, 2022.

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *7th International Conference on Learning Representations, ICLR*, 2019.

Bai, Y., Zhang, Y., Zhao, P., Sugiyama, M., and Zhou, Z. Adapting to online label shift with provable guarantees. In *NeurIPS*, 2022.

Chen, H. and Chao, W. On bridging generic and personalized federated learning for image classification. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

Cheng, G., Chadha, K., and Duchi, J. Fine-tuning in federated learning: a simple but tough-to-beat baseline, 2021.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Dinh, C. T., Tran, N. H., and Nguyen, T. D. Personalized federated learning with moreau envelopes. In *Conference on Neural Information Processing Systems, NeurIPS*, 2020.

du Plessis, M. C. and Sugiyama, M. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. C. A unified view of label shift estimation. In *Annual Conference on Neural Information Processing Systems 2020, NeurIPS*, 2020.

He, C., Li, S., So, J., Zhang, M., Wang, H., Wang, X., Vepakomma, P., Singh, A., Qiu, H., Shen, L., Zhao, P., Kang, Y., Liu, Y., Raskar, R., Yang, Q., Annavaram, M., and Avestimehr, S. Fedml: A research library and benchmark for federated machine learning. *CoRR*, abs/2007.13518, 2020.

Jiang, L. and Lin, T. Test-time robust personalization for federated learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., and et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

Li, Q., He, B., and Song, D. Model-contrastive federated learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2021a.

Li, Q., Diao, Y., Chen, Q., and He, B. Federated learning on non-iid data silos: An experimental study. In *38th IEEE International Conference on Data Engineering, ICDE*, 2022.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.*, 37(3):50–60, 2020.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021b.

Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., Salakhutdinov, R., and Morency, L.-P. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Lipton, Z. C., Wang, Y., and Smola, A. J. Detecting and correcting for label shift with black box predictors. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018.

Ma, W., Chen, C., Zheng, S., Qin, J., Zhang, H., and Dou, Q. Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *Medical Image Computing and Computer Assisted Intervention - MICCAI*, 2022.

Marfoq, O., Neglia, G., Bellet, A., Kameni, L., and Vidal, R. Federated multi-task learning under a mixture of distributions. In *Advances in Neural Information Processing Systems, NeurIPS*, 2021.

Marfoq, O., Neglia, G., Vidal, R., and Kameni, L. Personalized federated learning through local memorization. In *International Conference on Machine Learning, ICML*, 2022.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., and Li, H. Balanced meta-softmax for long-tailed visual recognition. In *Annual Conference on Neural Information Processing Systems 2020, NeurIPS*, 2020.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Comput.*, 14(1):21–41, 2002.

Shamsian, A., Navon, A., Fetaya, E., and Chechik, G. Personalized federated learning using hypernetworks. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

Sun, Z., Sun, R., Lu, L., and Mislove, A. Mind your weight(s): A large-scale study on insufficient machine learning model protection in mobile apps. In *30th USENIX Security Symposium, USENIX Security*, 2021.

Tan, A. Z., Yu, H., Cui, L., and Yang, Q. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Tian, J., Liu, Y., Glaser, N., Hsu, Y., and Kira, Z. Posterior re-calibration for imbalanced datasets. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020.

Wu, R., Guo, C., Su, Y., and Weinberger, K. Q. Online adaptation to label distribution shift. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS*, 2021.

Wu, Y., Zhang, S., Yu, W., Liu, Y., Gu, Q., Zhou, D., Chen, H., and Cheng, W. Personalized federated learning under mixture of distributions, 2023.

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

Xu, B., Wang, N., Chen, T., and Li, M. Empirical evaluation of rectified activations in convolutional network, 2015.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2), 2019. ISSN 2157–6904.

Yu, T., Bagdasaryan, E., and Shmatikov, V. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Zhang, J., Menon, A. K., Veit, A., Bhojanapalli, S., Kumar, S., and Sra, S. Coping with label shift via distributionally robust optimisation. In *9th International Conference on Learning Representations, ICLR*, 2021a.

Zhang, J., Li, Z., Li, B., Xu, J., Wu, S., Ding, S., and Wu, C. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning, ICML*, 2022.

Zhang, M., Sapra, K., Fidler, S., Yeung, S., and Alvarez, J. M. Personalized federated learning with first order model optimization. In *9th International Conference on Learning Representations, ICLR*. OpenReview.net, 2021b.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv:1806.00582*, 2018.

Zhu, H., Xu, J., Liu, S., and Jin, Y. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

Appendix

A. Rationality of Feature Matching

Suppose that there are $n_{t,k}$ test samples for class $k \in [K]$ and the empirical test priors based on true labels are $\mathbf{p}_t = \{p_{t,k}\}_{k=1}^K$, with $n_t = \sum_{k=1}^K n_{t,k}$ and $p_{t,k} = n_{t,k}/n_t$. Then, the mean feature representation $\mathbf{f}$ is given by

$$\mathbf{f} = \frac{1}{n_t}\sum_{i=1}^{n_t} \boldsymbol{f}(x_i) = \sum_{k=1}^{K} \frac{n_{t,k}}{n_t} \frac{\sum_{i=1}^{n} \mathbb{I}\{y_i = k\}\boldsymbol{f}(x_i)}{n_k} = \sum_{k=1}^{K} p_{t,k}\boldsymbol{f}_k, \tag{11}$$

where $\boldsymbol{f}_k$ is the class-wise prototype of test set. Denote $\mathbf{F} = \{\boldsymbol{f}_k\}_{k=1}^K$ as the collection of test prototypes and $\Delta\mathbf{p} = \mathbf{p} - \mathbf{p}_t$ as the estimate error, then the optimization problem in (10) can be rewritten as

$$\begin{aligned}
\Delta\mathbf{p}^* &= \arg\min_{\Delta\mathbf{p}} \frac{1}{2}\|\mathbf{C}^T(\Delta\mathbf{p} + \mathbf{p}_t) - \mathbf{F}^T\mathbf{p}_t\|^2 \\
&= \arg\min_{\Delta\mathbf{p}} \frac{1}{2}\|\mathbf{C}^T\Delta\mathbf{p} + (\mathbf{C} - \mathbf{F})^T\mathbf{p}_t\|^2 \\
&= \arg\min_{\Delta\mathbf{p}} \frac{1}{2}\Delta\mathbf{p}^T\mathbf{C}\mathbf{C}^T\Delta\mathbf{p} + \mathbf{p}_t{}^T(\mathbf{C} - \mathbf{F})\mathbf{C}^T\Delta\mathbf{p}, \\
&\text{s.t.} \sum_{k=1}^{K} \Delta p_k = 0.
\end{aligned} \tag{12}$$

Notice that the optimization problem in (12) is still a quadratic programming. Suppose that the feature prototype matrix $\mathbf{C}$ is non-singular, then the closed-form solution can be derived as

$$\Delta\mathbf{p}^* = \mathbf{1}^T[\mathbf{C}\mathbf{C}^T]^{-1}\mathbf{C}(\mathbf{C} - \mathbf{F})^T\mathbf{p}_t \frac{[\mathbf{C}\mathbf{C}^T]^{-1}\mathbf{1}}{\mathbf{1}^T[\mathbf{C}\mathbf{C}^T]^{-1}\mathbf{1}} - [\mathbf{C}\mathbf{C}^T]^{-1}\mathbf{C}(\mathbf{C} - \mathbf{F})^T\mathbf{p}_t, \tag{13}$$

where $\mathbf{1}$ is a K-dimensional vector with all elements equal to 1. Therefore, the squared error of test priors estimate is determined by the deviation between training and test prototype matrix $\|(\mathbf{C} - \mathbf{F})^T\mathbf{p}_t\|^2$, which means we have

$$\|\Delta\mathbf{p}^*\|^2 = \mathcal{O}\left(\|(\mathbf{C} - \mathbf{F})^T\mathbf{p}_t\|^2\right). \tag{14}$$

Now, we further assume that the feature representation of each class follows a Gaussian distribution $G(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$ with $\boldsymbol{\sigma}^2 = \max\{\boldsymbol{\sigma}_k\}_{k=1}^K$, and the per-class training prototype is averaged over n_s samples. For the random training and test sample generation process in expectation we have

$$\mathbb{E}\left[\boldsymbol{f}_k\right] = \boldsymbol{\mu}_k, \quad \mathbb{E}\|\boldsymbol{f}_k - \boldsymbol{\mu}_k\|^2 \leq \boldsymbol{\sigma}/n_{t,k}, \quad \forall k \in [K] \tag{15}$$

$$\mathbb{E}\left[\boldsymbol{c}_k\right] = \boldsymbol{\mu}_k, \quad \mathbb{E}\|\boldsymbol{c}_k - \boldsymbol{\mu}_k\|^2 \leq \boldsymbol{\sigma}/n_s, \quad \forall k \in [K] \tag{16}$$

Then, take the expectation over sampling process again, we have the following error bound

$$\begin{aligned}
\mathbb{E}\|(\mathbf{C} - \mathbf{F})^T\mathbf{p}_t\|^2 &= \mathbb{E}\sum_{k=1}^{K}\left\|\frac{n_k}{n}(\boldsymbol{c}_k - \boldsymbol{f}_k)\right\|^2 \\
&= \mathbb{E}\sum_{k=1}^{K}\left\|\frac{n_k}{n}(\boldsymbol{c}_k - \boldsymbol{\mu}_k + \boldsymbol{\mu}_k - \boldsymbol{f}_k)\right\|^2 \\
&= \mathbb{E}\sum_{k=1}^{K}\left[\frac{n_k^2}{n^2}\left(\|\boldsymbol{c}_k - \boldsymbol{\mu}_k\|^2 + 2(\boldsymbol{c}_k - \boldsymbol{\mu}_k)^T(\boldsymbol{\mu}_k - \boldsymbol{f}_k) + \|\boldsymbol{\mu}_k - \boldsymbol{f}_k\|^2\right)\right] \\
&\leq \mathbb{E}\sum_{k=1}^{K}\left[\frac{n_{t,k}^2}{n_t^2}\left(\frac{\boldsymbol{\sigma}^2}{n_s} + \frac{\boldsymbol{\sigma}^2}{n_{t,k}}\right)\right] \\
&= \mathbb{E}\sum_{k=1}^{K}\left[\frac{n_{t,k}^2\boldsymbol{\sigma}^2}{n_t^2 n_s} + \frac{n_{t,k}\boldsymbol{\sigma}^2}{n_t^2}\right] \\
&\leq (\frac{1}{n_s} + \frac{1}{n_t})\boldsymbol{\sigma}^2.
\end{aligned} \tag{17}$$

Therefore, when the size of training data in FL and the size of unlabeled test set are sufficiently large, the squared error of prior estimate could be small enough for a reliable model prediction calibration. It is worth noting that both discriminative feature prototypes and accurate prior estimate depend on a well-trained global model, which means the the first step towards generalizing to unseen clients is selecting enough devices in FL training stage and obtaining a sufficiently good global model. Otherwise, even the prior estimate is accurate, the resulted prediction performance is still limited, which indeed bridges the intrinsic connection between the global model and personalization.

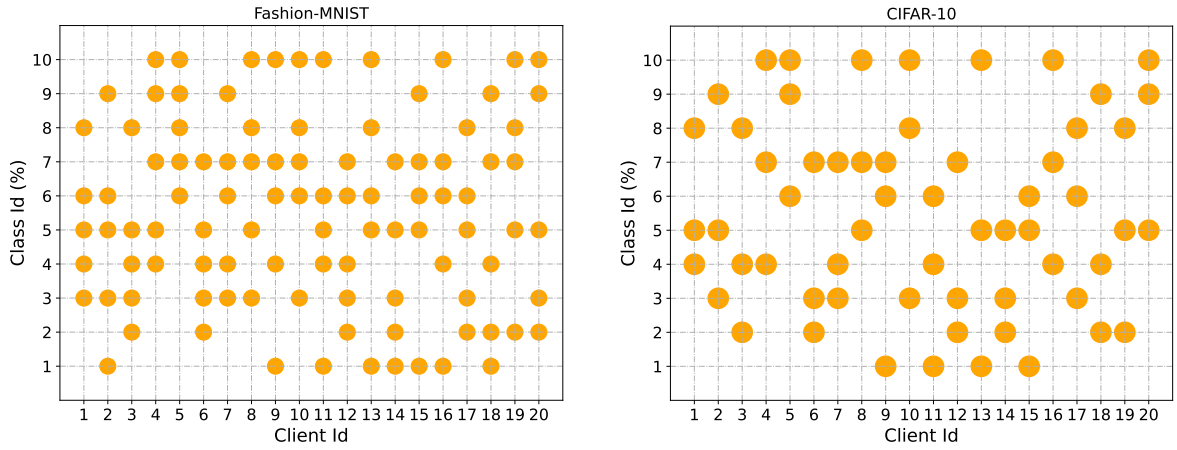
B. Details of Experimental Setup

B.1. Datasets and Models

We consider image classification tasks and evaluate our method on four popular datasets: (1) MNIST is a 10-class digit classification dataset; (2) Fashion-MNIST with 10 categories of clothes (Xiao et al., 2017); (3) CIFAR-10 with 10 categories of color images (Krizhevsky & Hinton, 2009); and (4) CINIC-10 (He et al., 2020), which is more diverse than CIFAR-10 as it is constructed from two different sources: ImageNet and CIFAR-10. We construct two different CNN models for MNIST/Fashion-MNIST and CIFAR-10/CINIC-10, respectively. The first CNN model is constructed by two convolution layers with 16 and 32 channels respectively, each followed by a max pooling layer, and two fully-connected layers with 128 and 10 units before the softmax output. We use the LeakyReLU (Xu et al., 2015) activation function. The second CNN model is similar to the first one but has one more convolution layer with 64 channels.

B.2. Data Partitioning

Similar to (McMahan et al., 2017; Collins et al., 2021), we make all clients have the same data size and each of the clients has C classes which are randomly selected from the whole class set. We evenly divide the local training samples over available classes. Specifically, for MNIST and Fashion-MNIST, C=5 categories are randomly selected for each client as these two learning tasks are relatively easy and result in high accuracy if C is two small, while CIFAR-10 and CINIC-10 are relatively hard to learn and 3 categories are chosen for each client. For all experiments, each client will be assigned with 1000 training samples and 500 test samples, where the local test set has the same label distribution with local training set by default. For new/unseen clients after federated training stage, we change the random seed and generate new 10 clients with C changing over $\{2, 3, 5, 8\}$. For online test settings, we generate three different test label sets and select any two of which as a pair that arise sequentially with a transition time point. We repeat the online test with three distinct pairs in total and report the averaged results in the test accuracy curves. Take the setup on Fashion-MNIST and CIFAR-10 with 20 clients as instance, the label distributions across all clients are presented in the following figures.



(a) Fashion-MNIST (b) CIFAR-10

Figure 3. Data label distributions of 20 clients with random samples partitioning from a limited number of classes.

B.3. Implementation Details

Training Settings. We employ the min-batch SGD as the local optimizer for all approaches. The learning rate η of local training is set to 0.01 for MNIST/Fashion-MNIST, and 0.02 for CIFAR-10/CINIC-10. The weight decay is set to 5e-4 and the momentum is set to 0.5. The batch size is fixed to $B = 50$ for all datasets. The number of local training epochs is set to $E = 5$ for all federated learning approaches unless explicitly specified. And the number of global communication rounds is set to 200 for all datasets, where all FL approaches have little or no accuracy gain with more communications. For all methods, we report the average test accuracy of all local models for performance evaluation.

Compared Methods. We compare our proposed FedCal with the following approaches: a baseline approach named Local-only, where each client only trains model on its own local data; parameter decoupling based methods, including FedPer (Arivazhagan et al., 2019) and FedRep (Collins et al., 2021) that learn personal classifier on top of a shared feature extractor, FedRoD (Chen & Chao, 2022) and FedTHE (Jiang & Lin, 2023) that combine global and local prediction heads. multi-task learning based methods Ditto (Li et al., 2021b); pFedHN (Shamsian et al., 2021) that employs a server-side hypernetwork to generate the personalized model for each client. ODPFL-HN (Amosy et al., 2022) trains an encoder with the same structure of the local model to generate client embeddings based on unlabeled data, which are fed into a hypernetwork to generate customized models. It is worth noting that a common drawback of pFedHN and ODPFL-HN is that they cannot support parallel client local training in each round and thus inefficient. Moreover, they involve the client representation transmission, which has potential privacy risk. We do not compare the methods in centralized settings as they usually need access the (offline) training set, which is not practical in FL settings.

Hyper-parameters. For all FedCal variants, we use a simple threshold-based method to rectify the estimated priors. More precisely, we consider there are at least two classes in the test set and denote y_s the second largest element in the prior vector, then all the elements that are smaller than $\alpha \cdot y_s$ are set to zero, where α is a tunable coefficient and is set to 0.5 by default. For kNN-Per, we tune the λ over $\{0.5, 0.6, 0.7, 0.8\}$ and select 0.6. The hyper-parameters in FedTHE are adopted from the paper. For pFedHN and ODPFL-HN, we set both the server learning rate and the client learning rate to 0.01.

Training priors and class-wise prototypes. In our proposed methods, to perform the test-time calibration, the training priors are needed in all calibration-based variants and the class-wise prototypes are specially required in the FM-based variant. However, how to obtain them is a non-trivial issue as those information are not naturally available for the FL service provider due to privacy protection needs. In our work, as mentioned before, we try to learn an unbiased global model such that it has the similar performance as the model trained on a class-balanced training set, to avoid the estimation of training priors. While advanced client selection methods and other approaches have been developed in the literature, we choose the balanced-softmax based local training strategy to achieve this goal with minimal costs of computation and communication. For the class-wise prototypes, one way is to collect and aggregate the local prototypes, which might be practical or might not be feasible as the local label distribution information would be exposed to the server. Another way is to maintain a prototype dictionary in the server and iteratively update the dictionary by distributing to local clients and aggregating locally updated ones. Moreover, other privacy-preserving techniques could also be integrated to enhance the system security.

C. Additional Experimental Results

Table 3. Comparison between FedCal with and without rectification in pathological cases

Method	MNIST		Fashion-MNIST		CIFAR-10		CINIC-10	
	w/o rect.	w/ rect.	w/o rect.	w/ rect.	w/o rect.	w/ rect.	w/o rect.	w/ rect.
FedCal-PL	96.87	96.55 (-0.32)	88.71	91.41 (+2.70)	80.05	87.92 (+7.87)	62.92	69.11 (+6.09)
FedCal-EM	96.97	96.55 (-0.42)	89.25	90.35 (+1.10)	84.69	88.48 (+3.79)	68.18	74.04 (+5.86)
FedCal-FM	96.65	95.62 (-1.03)	90.32	91.42 (+1.10)	85.29	88.07 (+2.78)	69.84	70.25 (+0.41)

C.1. More Discussions on Prior Rectification

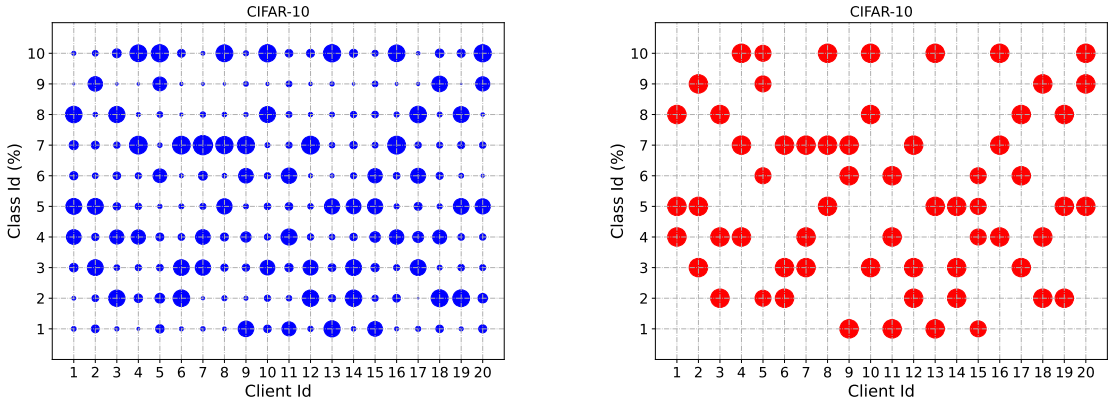
As mentioned in the main text, when we have the prior knowledge about that the test set is uniformly distributed for existing categories, we can utilize the rectification to generate more precise test priors. Here, we compare the performance of FedCal with and without the rectification in the pathological settings. As demonstrated in Table 3, in most cases, the rectification can promote the test accuracy except the MNIST dataset, where the classification task is relatively easy and the rectification

will instead slightly reduce the accuracy. It also can be found that as the difficulty of task increases, the effect of rectification becomes more significant for pseudo-label and maximum likelihood based estimation, while the feature-matching based estimation is more insensitive to the rectification and outperforms the other two variants when no rectification is conducted.

However, this operation is kind of risky as imbalanced test sets generally exist in many real-world applications, e.g., the disease diagnosis and anomaly detection, where the positive and negative samples could be highly-imbalanced in both training and test stages. Therefore, the prior rectification is only recommended when we have prior knowledge of the balanced test scenarios. Moreover, the results in Table 3 also indicate that when the classification tasks are relatively easy or the global model performs well, the marginal gain of rectification is rather limited. In such cases, the rectification operation is not suggested as the original prior estimate can already lead to a significant improvement over the global model.

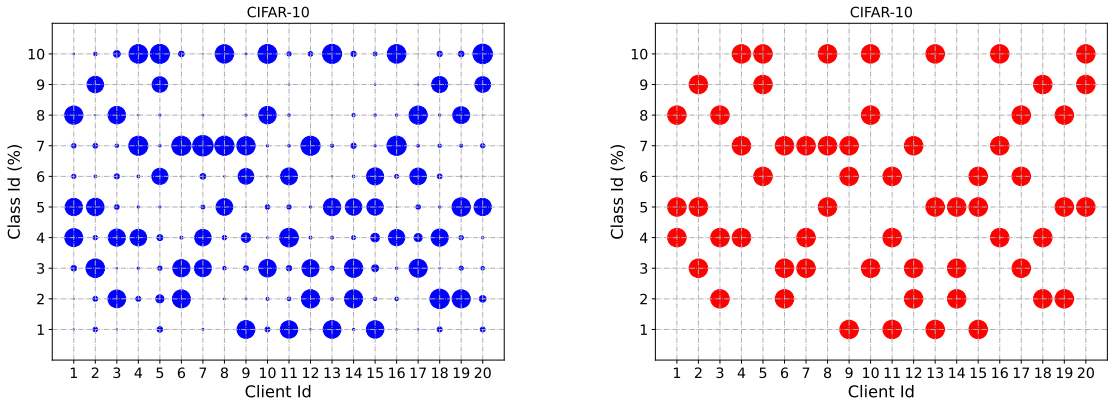
C.2. Estimated Priors of Clients

Here we take the CIFAR-10 as an example and plot the estimated label distributions of 20 clients by three methods, respectively. From the results in Fig. 4-Fig. 6, it can be seen that all methods can achieve a relatively precise estimate of the unknown priors and the rectification can further remove the estimate noises to a large extent. We also notice that the prior estimates of some clients may still be biased even after rectification, e.g., client 5 in FedCal-PL and client 7 in FedCal-FM.



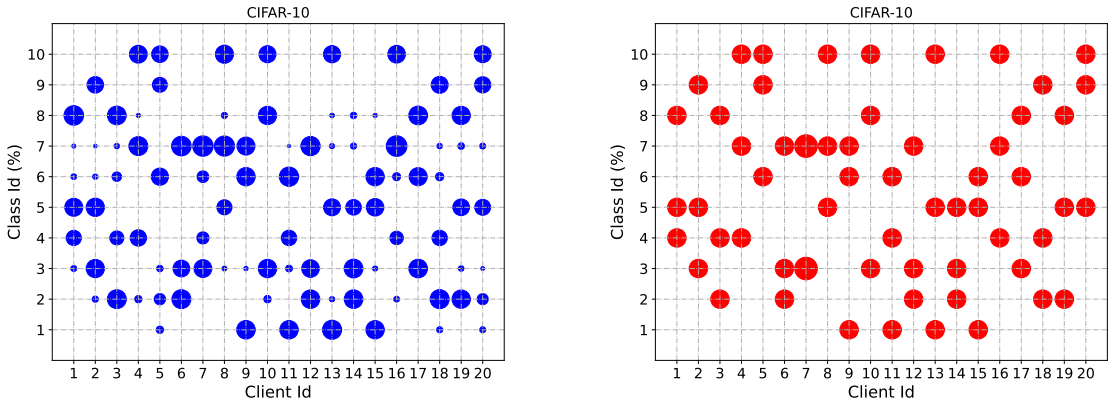
(a) Pseudo labeling w/o rectification (b) Pseudo labeling w/ rectification

Figure 4. Estimated test priors on CIFAR-10 with pseudo labeling method. (a) Original prior estimate; (b) Rectified prior estimate.



(a) Maximal likelihood w/o rectification (b) Maximal likelihood w/ rectification

Figure 5. Estimated test priors on CIFAR-10 with maximal likelihood estimation. (a) Original prior estimate; (b) Rectified prior estimate.



(a) Feature matching w/o rectification (b) Feature matching w/ rectification

Figure 6. Estimated test priors on CIFAR-10 with feature matching method. (a) Original prior estimate; (b) Rectified prior estimate.

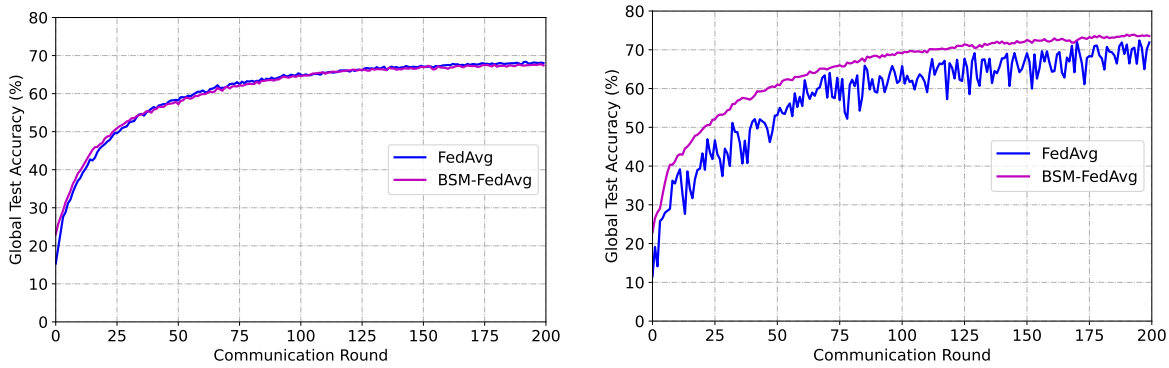


Figure 7. Global model performance on CIFAR-10. (a) 20 clients with 100% participation; (b) 100 clients and 10% participating ratio.

C.3. Global Model Performance with Full and Partial Participation

We also provide the global model accuracy on global test set over the training rounds in Fig. 7. It can be find that when all clients participate the federated training in each round, the aggregated model is relatively balanced and the balanced-softmax loss function will has little or no effect. In contrast, when the number of clients becomes larger and the participating ratio is relatively low, the model performance will fluctuate over time and the clients that only have a short active period will only receive a intermediate biased global model. Therefore, when those clients encounter with time-varying test set, the estimate of unknown priors and the prediction calibration will become unreliable and lead to low test accuracy as shown in Table 2.

C.4. More results on generalizing to new clients

There, we provide more empirical evaluation results on test-time adaptation to new clients with unknown label distribution shift. We consider two typical cases in the non-IID FL literature, including the one studied in the main text where each client only has partial but balanced categories, and the other one named Dirichlet distribution based sampling. For the first case, the results on MNIST/Fashion-MNIST/CIFAR-10/CINIC-10 are also provided in Fig. 8. Moreover, we use the concentration parameter $\beta \in \{0.01, 0.1, 0.5, 1.0\}$ to simulate a more complicated label distribution shift scenario, where the available labels in the test set is not balanced. In such cases, the assumption of balanced test samples of available categories does not hold anymore, therefore we also omit the rectification operation accordingly. From the results in Fig. 9, it can also be found that our method still outperforms the baseline without test-time adaptation.

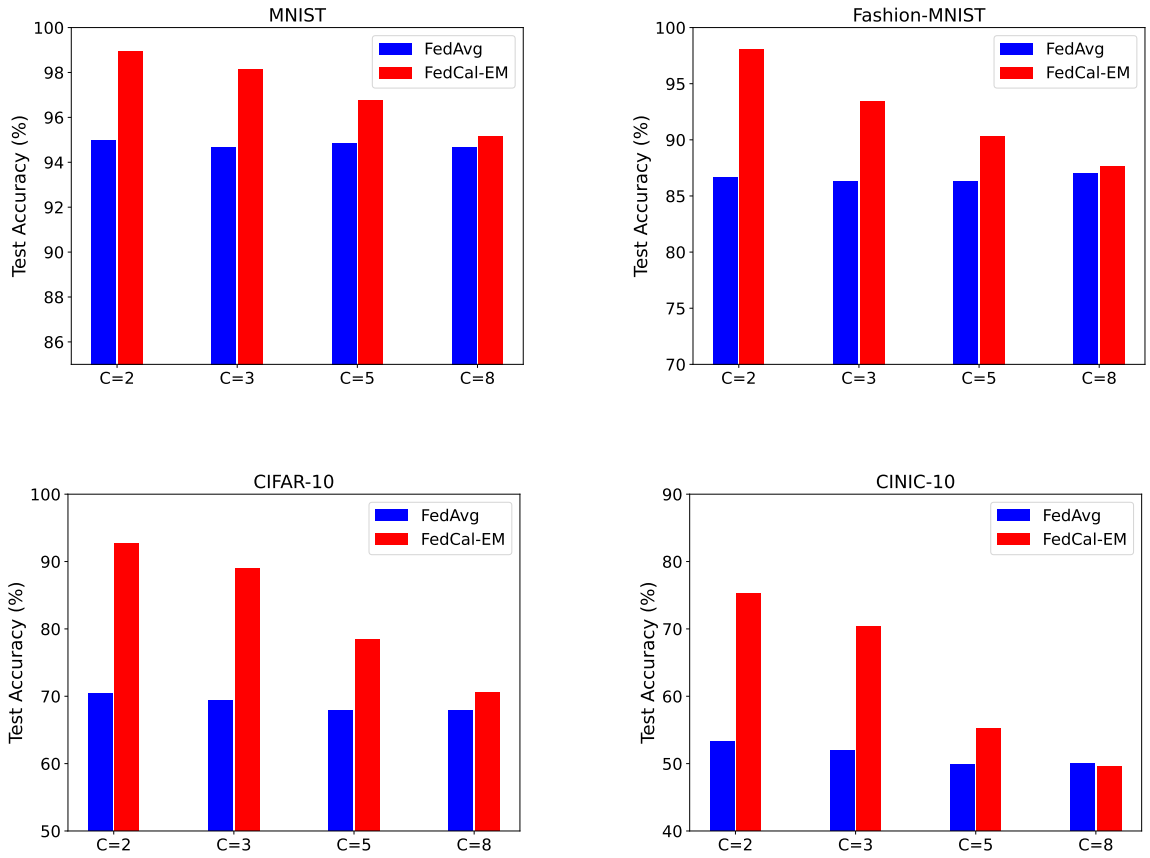


Figure 8. Generalization on new clients with various numbers of target categories. Each client has C randomly selected classes in the test set. Rectification is applied.

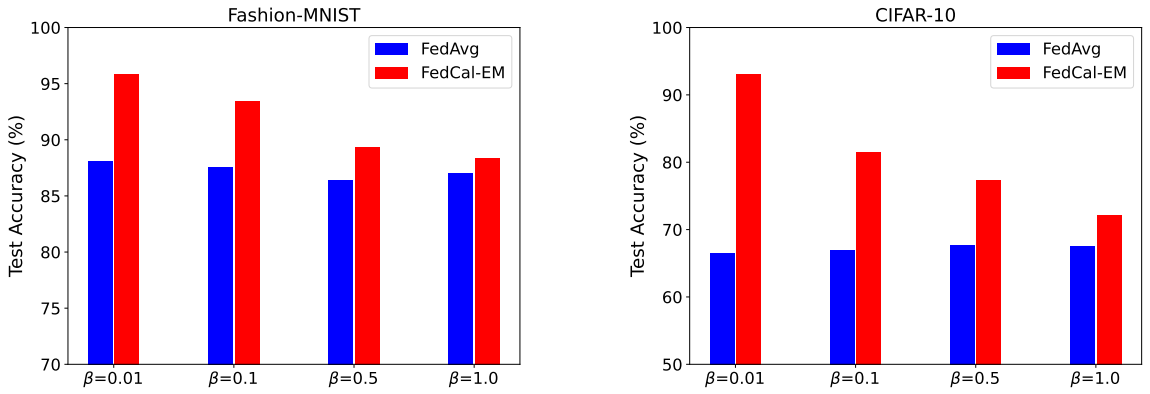


Figure 9. Generalization on new clients with various levels of label distribution shift. The label distributions among clients are sampled according to the Dirichlet distribution with a concentration parameter β. Rectification is **not** applied.