Federated Learning with Label Distribution Skew via Logits Calibration

Jie Zhang, Zhiqi Li, Bo Li,
Jianghe Xu, Shuang Wu, Shouhong Ding, Chao Wu

Zhejiang University, Youtu Lab, Tencent

FL with label distribution skew

- What is label distribution skew[1,2]?

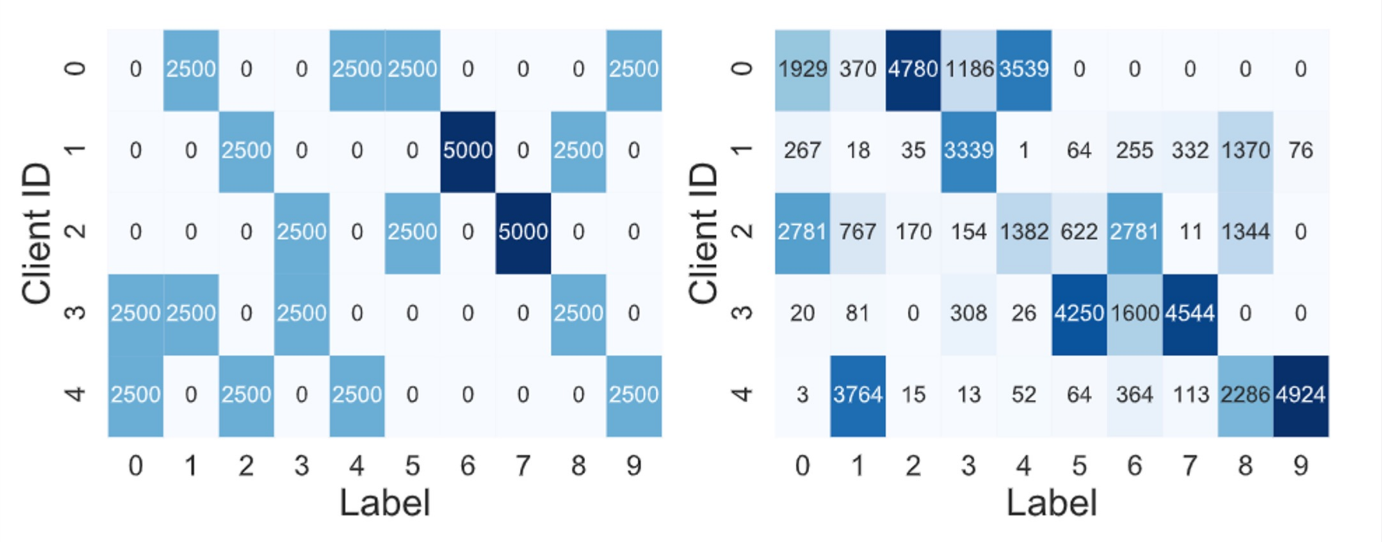
Suppose that client i can draw an example $(x, y) \sim P_i(x, y)$ from the local data, and the data distribution $P_i(x, y)$ can be rewritten as $P_i(x \mid y)P_i(y)$. For label distribution skew, the marginal distributions $P_i(y)$ varies across clients, while $P_i(y \mid x) = P_j(y \mid x)$ for all clients i and j.

Visualizations of skewed CIFAR10 on 5 clients.

★ Left: quantity-based label skew

★ Right: distribution-based label skew.

(The value in each rectangle is the number of data samples of a label belonging to a certain client.)



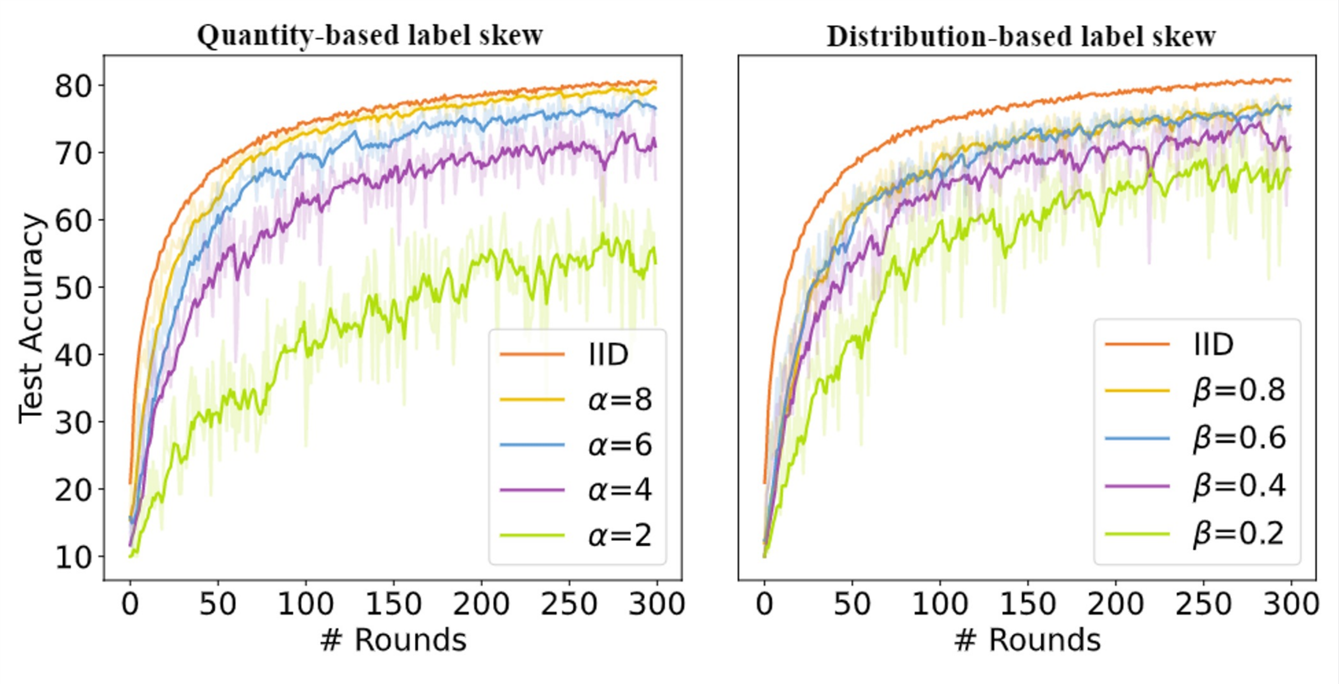
[1] Wang T, Zhu J Y, Torralba A, et al. Dataset distillation[J]. arXiv preprint, 2018
[2] Zhao B, Mopuri K R, Bilen H. Dataset condensation with gradient matching[J]. ICLR 2021.

FL with label distribution skew

- What is the problem of label distribution skew?

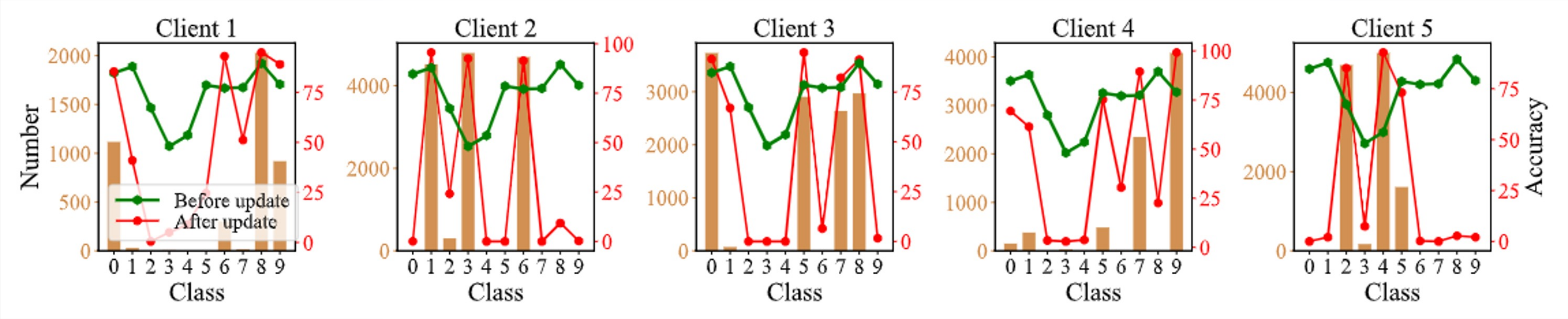
Test accuracy of FedAvg under various label skew settings on CIFAR10. The lower the α and β, the more skewed the distribution.

In comparison with IID settings, the accuracy is significantly decreased by 26.07% and 13.97% for α=2 and β=0.2, respectively.



FL with label distribution skew

- What is the problem of label distribution skew?

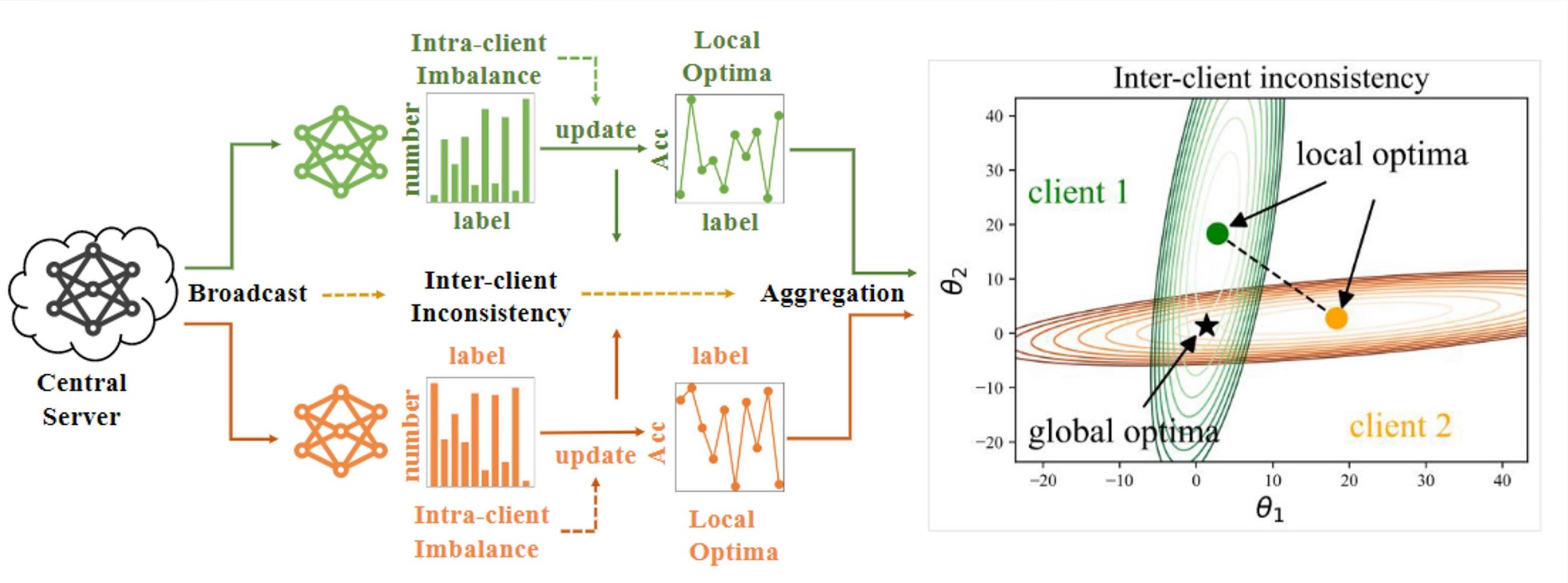


For skewed CIFAR10 dataset, the accuracy decreases heavily on minority classes, achieving an overall accuracy of zero for missing classes.

The histogram displays the number of samples for each class, while the red line represents the accuracy of each class.

FL with label distribution skew

- Why does FL perform poorly when the labels are skewed?



Heterogeneous data can result in inconsistent objective functions among clients, which leads the global model to converge to a stationary point that is far from global optima.

Furthermore, skewed data on the local client results in a biased model overfitting to minority classes and missing classes, which aggravates the objective inconsistency between clients.

Our proposed method: FedLC

- Learning objective

The goal of standard machine learning is to minimize the misclassification error from a statistical perspective:

$$P_{x,y}(y \neq \hat{y}), \text{and } P(y \mid x) \propto P(x \mid y)P(y).$$

However, we focus on label distribution skew in FL, which means P (y) is skewed.

Minority classes have a much lower probability of occurrence compared with majority classes, which means minimizing the misclassification error P(x | y) P(y) is no longer suitable [1].

When label distribution is skewed, we aim to minimize the misclassification error as follows:

$$\text{Calibrated error} = \min \frac{1}{k} \sum_{y \in K} \mathcal{P}_{x|y}(y \neq \hat{y}).$$

$$\begin{aligned} \arg\max_{y \in K} P^{Cal}(y \mid x) &= \arg\max_{y \in K} P(x \mid y) \\ &= \arg\max_{y \in K} \{P(y \mid x)/P(y)\}. \end{aligned}$$

Since softmax cross-entropy loss indicates that P(y |x) ∝ $e^{f_y(x)}$, then we have:

$$\arg\max_{y \in K} P^{Cal}(y \mid x) = \arg\max_{y \in K} \{f_y(x) - \log \gamma_y\},$$

Here γ_y is the estimate of the class prior P (y).

[1] Menon A K, Jayasumana S, Rawat A S, et al. Long-tail learning via logit adjustment[J]. ICLR 2021.

Our proposed method: FedLC

- Fine-grained Calibrated Cross-Entropy

$$\arg\max_{y \in K} P^{Cal}(y \mid x) = \arg\max_{y \in K}\{f_y(x) - \log \gamma_y\},$$

This formulation inspires us to calibrate the logits before softmax cross-entropy according to the probability of occurrence of each class. Then the modified cross-entropy loss can be formulated as:

$$\mathcal{L}_{\mathrm{Cal}}(y; f(x)) = -\log \frac{1}{\sum_{i \neq y} e^{-f_y(x)+f_i(x)+\Delta_{(y,i)}}},$$

Here $\Delta_{(y,i)} = \log(\frac{\gamma_i}{\gamma_y})$. It can be viewed as a pairwise label margin, which represents the desired gap between scores for y and i.

- For label skewed data, motivated by the interesting idea in [1], we aim to minimize the test error:

$$\mathcal{L}_{\mathrm{Cal}}(y; f(x)) = -\log \frac{e^{f_y(x)-\tau \cdot n_y^{-1/4}}}{\sum_{i \neq y} e^{f_i(x)-\tau \cdot n_i^{-1/4}}}.$$

➔ This loss function simultaneously minimizes the classification errors and forces the learning to focus on margins of minority classes to reach the optimal results.

[1] Cao K, Wei C, Gaidon A, et al. Learning imbalanced datasets with label-distribution-aware margin loss[J]. NeurIPS, 2019, 32.

Experiments

- Main results on SVHN, CIFAR10, and CIFAR100

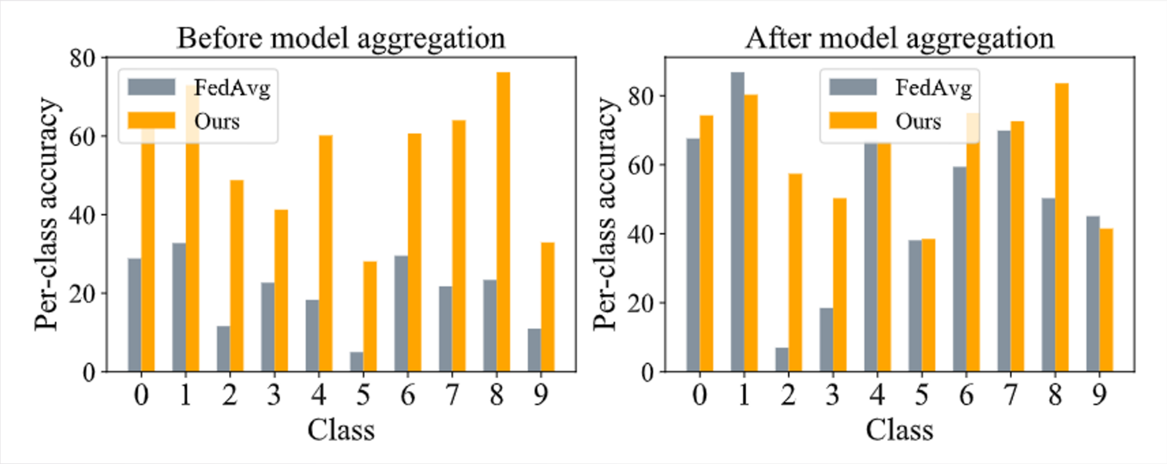
Table 2. Performance overview for different degrees of distribution-based label skew.

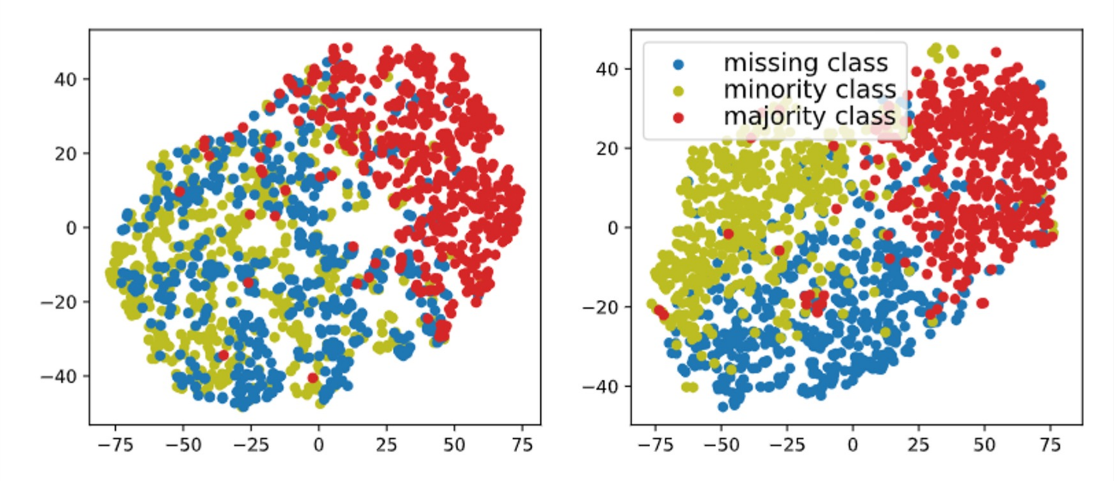
Dataset	SVHN				CIFAR10				CIFAR100			
Skewness	$\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$
FedAvg	$69.51_{\pm 1.45}$	$79.86_{\pm 1.46}$	$85.14_{\pm 0.83}$	$86.02_{\pm 1.15}$	$37.63_{\pm 1.36}$	$48.07_{\pm 1.38}$	$55.95_{\pm 0.83}$	$60.18_{\pm 1.78}$	$21.37_{\pm 0.87}$	$25.06_{\pm 1.04}$	$28.44_{\pm 1.51}$	$29.29_{\pm 1.32}$
FedProx	$71.42_{\pm 1.24}$	$81.39_{\pm 1.35}$	$86.30_{\pm 0.95}$	$87.53_{\pm 1.56}$	$39.03_{\pm 1.27}$	$49.57_{\pm 0.90}$	$57.88_{\pm 0.93}$	$62.13_{\pm 1.17}$	$22.92_{\pm 1.71}$	$26.44_{\pm 0.86}$	$30.16_{\pm 1.18}$	$31.20_{\pm 1.23}$
Scaffold	$71.23_{\pm 1.63}$	$81.80_{\pm 1.75}$	$86.32_{\pm 1.19}$	$87.13_{\pm 1.39}$	$38.84_{\pm 0.93}$	$49.12_{\pm 1.21}$	$57.39_{\pm 1.16}$	$61.54_{\pm 1.28}$	$22.61_{\pm 1.37}$	$26.30_{\pm 1.32}$	$29.96_{\pm 1.17}$	$31.26_{\pm 1.75}$
FedNova	$72.50_{\pm 1.21}$	$82.41_{\pm 1.40}$	$87.11_{\pm 1.38}$	$86.65_{\pm 1.25}$	$39.81_{\pm 1.18}$	$50.56_{\pm 1.42}$	$58.85_{\pm 0.93}$	$62.77_{\pm 0.86}$	$24.03_{\pm 0.91}$	$27.65_{\pm 0.99}$	$30.76_{\pm 0.95}$	$31.93_{\pm 0.98}$
FedOpt	$73.46_{\pm 1.07}$	$82.71_{\pm 1.13}$	$86.85_{\pm 0.85}$	$87.41_{\pm 1.72}$	$41.08_{\pm 1.01}$	$51.89_{\pm 0.86}$	$59.39_{\pm 1.68}$	$63.38_{\pm 1.62}$	$24.51_{\pm 1.71}$	$28.98_{\pm 1.08}$	$32.42_{\pm 1.66}$	$32.94_{\pm 1.28}$
FedRS	$75.97_{\pm 1.15}$	$83.27_{\pm 1.54}$	$87.01_{\pm 0.98}$	$87.40_{\pm 1.67}$	$44.39_{\pm 1.63}$	$54.04_{\pm 1.59}$	$62.40_{\pm 1.38}$	$66.39_{\pm 1.28}$	$27.93_{\pm 1.18}$	$32.89_{\pm 1.50}$	$36.58_{\pm 0.94}$	$38.98_{\pm 1.35}$
Ours	$\mathbf{82.36}_{\pm 0.67}$	$\mathbf{84.41}_{\pm 0.87}$	$\mathbf{88.02}_{\pm 1.19}$	$\mathbf{88.48}_{\pm 1.29}$	$\mathbf{54.55}_{\pm 1.70}$	$\mathbf{65.91}_{\pm 1.68}$	$\mathbf{72.18}_{\pm 0.86}$	$\mathbf{72.99}_{\pm 1.12}$	$\mathbf{38.08}_{\pm 0.84}$	$\mathbf{41.01}_{\pm 1.08}$	$\mathbf{44.23}_{\pm 1.70}$	$\mathbf{44.96}_{\pm 1.71}$

- As data heterogeneity increases (i.e.\ , smaller β), all competing methods struggle, whereas our method displays markedly improved accuracy on highly skewed data.
- For CIFAR-10 dataset with β=0.05, our method gets a test accuracy of 54.55%, which is much higher than that of FedRS by 10.16%.

Experiments

- Analysis of Experiments





- Average per-class accuracy before and after model aggregation. For fair comparisons, we use the same well-trained model for initialization and the same data partition on each client.

- TSNE visualizations on majority, minority and missing classes.

  Left: For FedAvg, the samples from the minority class and missing class are mixed together and indistinguishable.

  Right: For our method, the data from minority class and missing class can be distinguished well, which indicates our method can learn more discriminative features.

Thanks for listening!