6. Implementation Details

The number of communication rounds is 2000 for most methods on CIFAR-10/100-LT and 1000 on ImageNet-LT. On the CIFAR-100-LT, the learning rate is initialized as 0.5 and decayed by 0.05 at round 1600 for all methods except PaCo [5], where the initial learning rate is 0.3. On the CIFAR-10-LT [2] with imbalance ratio 50 and 100, the initial learning rate is 0.1 for FedAvg [23], Ratio Loss [36], CLIMB [32], Focal Loss [21], CRT [15], LDAM [2], BSM [**?**], LADE [13], and GBME. It is 0.2 for RIDE [37] and 0.3 for PaCo. The learning rate is decayed by 0.1 for all methods. With RIDE, PaCo, and GBME, the model is trained for 2500 communication rounds on CIFAR-10-LT due to their complex model structures. For training on ImageNet-LT [22], we set the training round as 1000 and learning rate as 0.1, which is decayed at 800th round by 0.1. For CRT, we retrain the classifier on the last 200 rounds. For LADE, we set the weight of LADER as 0.01 for CIFAR-10-LT and 0.1 for CIFAR-100-LT and ImageNet-LT. We set client selection ratio $\alpha = 0.6$ in GBME on all datasets.

7. Comparison with More FL Methods

Table 12: Top-1 accuracy of various FL algorithms for dealing with data heterogeneity on CIFAR-100-LT.

IR	FedAvg		FedProx		SCAFFOLD	
	w/o BSM	w/ BSM	w/o BSM	w/ BSM	w/o BSM	w/ BSM
50	36.84	41.91	38.77	42.33	40.44	42.27
100	34.48	37.19	34.16	37.78	34.70	38.55
IR	FedAlign		Ditto		FedRep	
	w/o BSM	w/ BSM	w/o BSM	w/ BSM	w/o BSM	w/ BSM
50	39.80	43.66	37.03	39.29	37.69	39.17
100	35.36	39.21	33.45	35.18	32.23	35.04

We further adpoted FedProx [18], SCAFFOLD [16], FedAlign [25], Ditto [17] and FedRep [4] as baselines on the CIFAR100-LT. They are recent FL algorithms for dealing with data heterogeneity. When combined with BSM loss [**?**], for non-personalized FL methods (i.e., FedProx [18], SCAFFOLD [16], FedAlign [25]), we use GPI as the class prior of local re-balance, and for personalized FL methods (i.e., Ditto [17], FedRep [4]).

As shown in Table 12, the non-personalized FL methods are effective in FL with relatively mild class imbalance (imbalance ratio = 50), while are less effective when class imbalance is severe (imbalance ratio = 100). More importantly, when combined with our re-balance strategy, these methods can obtain a significant performance improvement for various imbalance ratios.

However, the personalized FL methods Ditto [17] and FedRep [4] perform worse on the federated long-tailed problem, even with more training rounds. We believe the main reason is the distribution shift. Under the general setting of personalized FL without the long-tailed problem, the distribution of test data and training data are identical. However, considering the long-tailed problem, the training data is imbalanced and the test data is balanced for each client, resulting in a distribution shift. The personalized model focuses more on fitting the local training data distribution and therefore generalizes poorly to different data distributions. Compared with the personalized FL methods, non-personalized FL methods with a well-trained global model have a stronger generalization ability, so they can achieve better performance on the federated long-tailed problem.

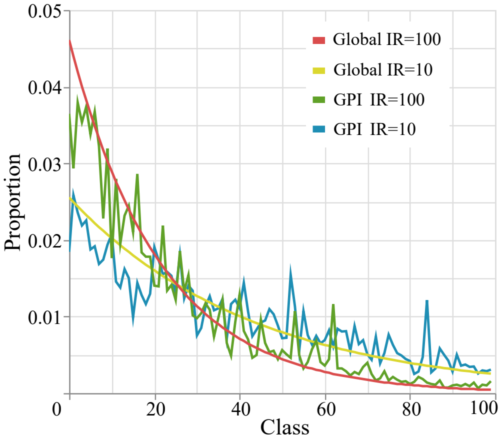


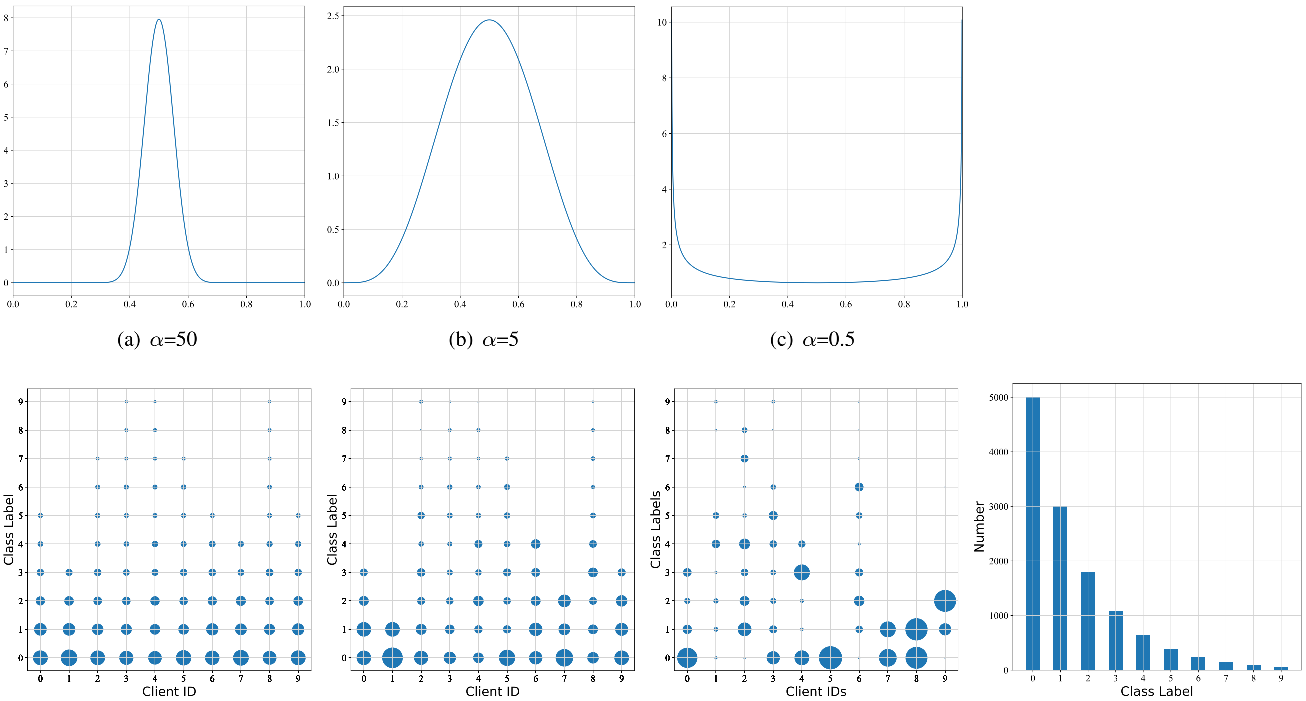
Figure 8: Comparison between the global label distribution and GPI on CIFAR-100-LT with IR = 10 and IR = 100.

8. GPI Visualization

As shown in Figure 8, we visualize the class-wise GPI curves on CIFAR-100-LT with IR = 10 and IR = 100. Under the same IR, GPI curves exhibit a similar tendency with the ground-truth label distribution, which can as a global balanced prior for existing re-balance strategies. Besides, GPI can maintain the performance of majority classes by using smaller GPI values for re-balancing.

9. Visualization of Data Distribution

To simulate the data heterogeneity, for each class in the global dataset, we partition the samples into different clients according to the Dirichlet distribution. In this section, we partition the CIFAR10-LT into 10 clients with different values of α_{dir}, *i.e.*, the hyper-parameter of the Dirichlet distribution. In Figure 9, (d, e, f) show the data distributions of local clients with different values of α_{dir}, and (a, b, c) show the corresponding probability density of the two-dimensional Dirichlet distribution. The point size indicates the sample number. Small α_{dir} results in higher heterogeneity. The global label distribution of the CIFAR10-LT with IR = 100 is shown in Figure 9(g).



(a) α=50 (b) α=5 (c) α=0.5

(d) α=50 (e) α=5 (f) α=0.5 (g) Global label distribution

Figure 9: The heterogeneous long-tailed data distributions of local datasets with different Dirichlet settings.

10. Proofs

10.1. Proof to Theorem 1

We denote the global objective of FL model with global re-balance and local re-balance as $G_g(\theta)$ and $G_l(\theta)$, respectively, then we have:

$$\begin{aligned} G_g(\theta) &= \sum_{k\in\{0,1\}} \frac{n_k}{n_0+n_1} \sum_{c\in\{+,-\}} \frac{n_0+n_1}{(n_0^c+n_1^c)n_k} \sum_{(\boldsymbol{x},y^c)\in\mathcal{D}^k} f(\boldsymbol{x},y^c,\theta) \\ &= \sum_{c\in\{+,-\}} \frac{1}{n_0^c+n_1^c} \sum_{(\boldsymbol{x},y^c)\in\mathcal{D}} f(\boldsymbol{x},y^c,\theta) \ , \end{aligned} \tag{9}$$

and

$$\begin{aligned} G_l(\theta) &= \sum_{k\in\{0,1\}} \frac{n_k}{n_0+n_1} \sum_{c\in\{+,-\}} \frac{1}{n_k^c} \sum_{(\boldsymbol{x},y^c)\in\mathcal{D}^k} f(\boldsymbol{x},y^c,\theta) \\ &= \sum_{c\in\{+,-\}} \frac{1}{n_0+n_1}\left(\frac{n_0}{n_0^c}+\frac{n_1}{n_1^c}\right) \sum_{(\boldsymbol{x},y^c)\in\mathcal{D}} f(\boldsymbol{x},y^c,\theta) \ , \end{aligned} \tag{10}$$

where c is the class index. Then, the objective gap is:

$$\begin{aligned} \Delta &= G_l(\theta) - G_g(\theta) \\ &= \sum_{c\in\{+,-\}} \frac{n_1(n_0^c)^2+n_0(n_1^c)^2}{n_0^c n_1^c(n_0+n_1)(n_0^c+n_1^c)} \sum_{(\boldsymbol{x},y^c)\in\mathcal{D}} f(\boldsymbol{x},y^c,\theta). \end{aligned} \tag{11}$$

For sample number $n_k^c > 0$, we have $\Delta > 0$, indicating that the objective function value of local re-balance is always larger than global re-balance on the same dataset. Thus, let $\mathcal{E}$ be the approximate estimation of global label distribution, taking $\mathcal{E}$ as the prior for the global re-balance strategy e, the strategy e based global objective G_e satisfies:

$$G_g(\theta) \le G_e(\theta) < G_l(\theta). \tag{12}$$

Proof is completed.

10.2. Proof to Theorem 2

Let v be any vector satisfying $\|v\| \le \Delta$. For the one dimensional case, we seek conditions on σ to bound the privacy loss:

$$\left| \ln \frac{e^{(-1/2\sigma^2)\|x-\mu\|^2}}{e^{(-1/2\sigma^2)\|x+v-\mu\|^2}} \right| \tag{13}$$

where x is sampled from $\mathcal{N}(0,\Sigma)$ and Σ is a diagonal matrix with entries σ and $\mu = (0,\ldots,0)$.

$$\begin{aligned} \left| \ln \frac{e^{(-1/2\sigma^2)\|x-\mu\|^2}}{e^{(-1/2\sigma^2)\|x+v-\mu\|^2}} \right| &= \left| \ln e^{(-1/2\sigma^2)\left[\|x-\mu\|^2-\|x+v-\mu\|^2\right]} \right| \\ &= | \frac{1}{2\sigma^2} \Big(\|x\|^2 - \|x+v\|^2) \Big) | \ . \end{aligned} \tag{14}$$

Considering the right triangle with base $v+x^{[1]}$ and edge $\sum_{i=2}^{m} x^{[i]}$ orthogonal to v, following [8], the hypotenuse of this triangle is $x+v$.

$$\begin{aligned} \|x+v\|^2 &= \left\|v+x^{[1]}\right\|^2 + \sum_{i=2}^{m} \left\|x^{[i]}\right\|^2 \\ \|x\|^2 &= \sum_{i=1}^{m} \left\|x^{[i]}\right\|^2 . \end{aligned} \tag{15}$$

Assuming without loss of generality that $x^{[1]}$ is parallel to v, we have $\left\|v+x^{[1]}\right\|^2 = \left(\|v\| + \|x\|^{[1]}\right)^2$. Thus, $\|x+v\|^2 - \|x\|^2 = \|v\|^2 + 2x^{[1]} \cdot \|v\|$. Because of $\|v\| \leq \Delta$ and $x^{[1]} \sim \mathcal{N}(0,\sigma^2)$, we have by writing $x^{[1]}$ as λ:

$$\left| \frac{1}{2\sigma^2} \left(\|x\|^2 - \|x+v\|^2 \right) \right| \leq \left| \frac{1}{2\sigma^2} \left(2\lambda\Delta + \Delta^2 \right) \right| \tag{16}$$

This quantity is bounded by ϵ whenever $\lambda < \sigma^2\epsilon/\Delta - \Delta/2$. To ensure privacy loss bounded by ϵ with probability at least $1-\delta$, we require

$$\Pr\left[|\lambda| \geq \sigma^2\epsilon/\Delta - \Delta/2 \right] < \delta \tag{17}$$

We use the tail bound

$$\Pr[\lambda > t] \leq \frac{\sigma}{\sqrt{2\pi}} e^{-t^2/2\sigma^2} \tag{18}$$

Thus we have

$$\delta < \Pr[\lambda > t] \leq \frac{\sigma}{\sqrt{2\pi}} e^{-t^2/2\sigma^2} \tag{19}$$

We require

$$\begin{aligned} & \frac{\sigma}{\sqrt{2\pi}} \frac{1}{t} e^{-t^2/2\sigma^2} < \delta/2 \\ \Leftrightarrow & \sigma \frac{1}{t} e^{-t^2/2\sigma^2} < \sqrt{2\pi}\delta/2 \\ \Leftrightarrow & \frac{t}{\sigma} e^{t^2/2\sigma^2} > 2/\sqrt{2\pi}\delta \\ \Leftrightarrow & \ln(t/\sigma) + t^2/2\sigma^2 > \ln(2/\sqrt{2\pi}\delta) \end{aligned} \tag{20}$$

Taking $t = \sigma^2\epsilon/\Delta - \Delta/2$, we get

$$\begin{aligned} & \ln\left(\left(\sigma^2\epsilon/\Delta - \Delta/2 \right)/\sigma \right) + \left(\sigma^2\epsilon/\Delta - \Delta/2 \right)^2 / 2\sigma^2 \\ & > \ln(2/\sqrt{2\pi}\delta) = \ln\left(\sqrt{\frac{2}{\pi}} \frac{1}{\delta} \right) . \end{aligned} \tag{21}$$

Let us write $\sigma = c\Delta/\epsilon$; we wish to bound c. We begin by finding the conditions under which the first term is non-positive.

$$\frac{1}{\sigma} \left(\sigma^2 \frac{\epsilon}{\Delta} - \frac{\Delta}{2} \right) = c - \frac{\epsilon}{2c} \tag{22}$$

Since $0 < \epsilon < 1$ and $c \geq 1$, we have $c - \frac{\epsilon}{2c} \geq c - 1/2$. Thus $\frac{1}{\sigma}\left(\sigma^2 \frac{\epsilon}{\Delta} - \frac{\Delta}{2}\right) \geq 1$ provided $c \geq 3/2$. We can therefore focus on the t^2/σ^2 term.

$$\begin{aligned} \left(\frac{1}{2\sigma^2}\right) \left(\frac{\sigma^2\epsilon}{\Delta} - \frac{\Delta}{2} \right)^2 &= \left(\frac{1}{2\sigma^2}\right) \left[\Delta^2 \left(\frac{c^2}{\epsilon} - \frac{1}{2} \right) \right]^2 \\ &= \left[\Delta^2 \left(\frac{c^2}{\epsilon} - \frac{1}{2} \right) \right]^2 \left[\frac{\epsilon^2}{c^2\Delta^2} \right] \frac{1}{2} \\ &= \frac{1}{2} \left(\frac{c^2}{\epsilon} - \frac{1}{2} \right)^2 \frac{\epsilon^2}{c^2} \\ &= \frac{1}{2} \left(c^2 - \epsilon + \epsilon^2/4c^2 \right) \end{aligned} \tag{23}$$

In the range $c \geq 1$, the derivative of $\left(c^2 - \epsilon + \epsilon^2/4c^2\right)$ with respect to c is positive. Then we have $c^2 - \epsilon + \epsilon^2/4c^2 > c^2 - 8/9$ and it suffices to ensure

$$c^2 - 8/9 > 2\ln\left(\sqrt{\frac{2}{\pi}} \frac{1}{\delta} \right) \tag{24}$$

which is satisfied $c^2 > 2\ln\left(\frac{1.25}{\delta}\right)$.

Now considering $\mathbb{R} = R_1 \cup R_2$, where $R_1 = \{x \in \mathbb{R} : |x| \leq c\Delta/\epsilon\}$ and $R_2 = \{x \in \mathbb{R} : |x| > c\Delta\epsilon\}$. Fix any subset $S \subseteq \mathbb{R}$, and define

$$\begin{aligned} S_1 &= \{\mathcal{M}(x) + x \mid x \in R_1\} \\ S_2 &= \{\mathcal{M}(x) + x \mid x \in R_2\} \end{aligned} \tag{25}$$

Then we have

$$\begin{aligned} & \underset{x\sim\mathcal{N}(0,\sigma^2)}{Pr} [\mathcal{M}(x) + x \in S] \\ & = \underset{x\sim\mathcal{N}(0,\sigma^2)}{Pr} [\mathcal{M}(x) + x \in S_1] + \underset{x\sim\mathcal{N}(0,\sigma^2)}{Pr} [\mathcal{M}(x) + x \in S_2] \\ & \leq e^{\epsilon} \underset{x\sim\mathcal{N}(0,\sigma^2)}{Pr} [\mathcal{M}(x) + x \in S_1] + \delta \end{aligned} \tag{26}$$

yielding (ϵ, δ)-DP for the Gaussian mechanism.