FedRAC: Rolling Submodel Allocation for Collaborative Fairness in Federated Learning

Supplementary Material

8. Proof of Theorem 1

We quantify the distance between the local submodel θ_i^t and the global model θ_g^t as $\delta_i^t := ||\theta_g^t - \theta_i^t||$. For any two clients i and j where $r_i \geq r_j$, our allocation strategy ensures a hierarchical containment relationship (i.e., $\theta_j^t \in \theta_i^t \in \theta_g^t$). This nested structure implies that the larger submodel θ_i^t inherently preserves more information from the global model, leading to a tighter alignment with θ_g^t. Consequently, this structure dictates the inequality $\delta_i^t \leq \delta_j^t$.

To establish the relationship $F(\theta_i^t) \leqslant F(\theta_j^t)$, we leverage the inequality $\delta_i^t \leqslant \delta_j^t$ and some regularity conditions of $F()$. Specifically, we assume the objective $F(\cdot)$ as an L-smooth and μ-strongly convex function and $L \geqslant \mu$. We present the foundational definitions of these properties below:

Assumption 1 (L-smooth F) *If F is L-smooth, then $\forall \theta_i, \theta_j \in \theta$,*

$$F(\theta_i) \leqslant F(\theta_j) + \nabla F(\theta_j)^T(\theta_i - \theta_j) + \frac{L}{2}||\theta_i - \theta_j||^2. \quad (11)$$

Assumption 2 (μ-strongly convex F) *If F is μ-strongly convex, then $\forall \theta_i, \theta_j \in \theta$,*

$$F(\theta_i) \geq F(\theta_j) + \nabla F(\theta_j)^T(\theta_i - \theta_j) + \frac{\mu}{2}||\theta_i - \theta_j||^2. \quad (12)$$

By leveraging the property of L-smoothness, the objective function $F(\theta_i^t)$ can be upper-bounded as follows:

$$F(\theta_i^t) \leqslant \underbrace{F(\theta_g^t) + \nabla F(\theta_g^t)^T(\theta_i^t - \theta_g^t) + \frac{L}{2}\delta_{i,t}^2}_{R_L}. \quad (13)$$

By leveraging the property of μ-strongly convex, we can establish the following lower bound:

$$F(\theta_j^t) \geq \underbrace{F(\theta_g^t) + \nabla F(\theta_g^t)^T(\theta_j^t - \theta_g^t) + \frac{\mu}{2}\delta_{j,t}^2}_{R_\mu}. \quad (14)$$

To establish the inequality $F(\theta_i^t) \leqslant F(\theta_j^t)$, the problem reduces to verifying that $R_L \leqslant R_\mu$.

$$R_L - R_\mu = \underbrace{\nabla F(\theta_g^t)^T(\theta_i^t - \theta_j^t)}_{R_1} + \underbrace{\frac{1}{2}(L\delta_{i,t}^2 - \mu\delta_{j,t}^2)}_{R_2 \leqslant 0}. \quad (15)$$

We characterize the proximity of θ_g^t to a stationary point of $F(\cdot)$ by bounding its gradient norm:

$$||\nabla F(\theta_g^t)|| \leqslant \frac{L|\delta_{i,t}^2 - \delta_{j,t}^2|}{2||\theta_i^t - \theta_j^t||}. \quad (16)$$

Based on $\delta_{i,t} \leqslant \delta_{j,t}$ and $L \leqslant \mu$, we have

$$R_2 = \frac{1}{2}(L\delta_{i,t}^2 - \mu\delta_{j,t}^2) \leqslant \frac{L}{2}(\delta_{i,t}^2 - \delta_{j,t}^2) \leqslant 0. \quad (17)$$

Based on equation (15), (16) and (17), we obtain the following relationship:

$$\begin{aligned} |R_1| &\triangleq |\nabla F(\theta_g^t)^T(\theta_i^t - \theta_j^t)| \\ &\leqslant ||\nabla F(\theta_{N,t})|| \times ||(\theta_i^t - \theta_j^t)|| \\ &\leqslant \frac{L|\delta_{i,t}^2 - \delta_{j,t}^2|}{2} \\ &\leqslant |R_2|, \end{aligned} \quad (18)$$

where the first inequality follows from the Cauchy-Schwarz, while the second inequality arises from substituting the bound in equation (16).

Finally, the constraints $|R_1| \leqslant |R_2|$ and $R_2 \leqslant 0$ imply $R_1 + R_2 \leqslant 0$. This relationship guarantees $R_L - R_\mu \leqslant 0$, which subsequently confirms that $F(\theta_i^t) \leqslant F(\theta_j^t)$.

9. Proof of Theorem 2

Let $I_E \triangleq \{nE\}_{n=1}^{\infty}$ be the set of global synchronization steps. We define the intermediate local update as $v_i^{t+1} = \theta_i^t - \eta_t \nabla F_i(\theta_i^t, \xi_i^t)$. $\bar{g}_t = \sum_{i=1}^N \frac{\nabla F_i(\theta_i^t)}{p_i}$ and $g_t = \sum_{i=1}^N \frac{\nabla F_i(\theta_i^t, \xi_i^t)}{p_i}$. Then, we have $\bar{v}_{t+1} = \bar{\theta}_t - \eta_t g_t$ and $E[g_t] = \bar{g}_t$.

Assumption 3 *Let ξ_i^t be a sample drawn uniformly at random from the i-th client. The variance of the local stochastic gradients is bounded by σ^2:*

$$E\|\nabla F_i(\theta_i^t, \xi_i^t) - \nabla F_i(\theta_i^t)\| \leqslant \sigma_i^2, \quad (19)$$

Assumption 4 *The expected squared norm of the stochastic gradients is uniformly bounded by G^2:*

$$E\|\nabla F_i(\theta_i^t, \xi_i^t)\| \leqslant G^2, \quad (20)$$

where $i \in \{1, 2, ..., N\}$ and $t \in \{1, 2, ..., T-1\}$.

Assumption 5 *FedRAC guarantees that each neuron in the global model is trained with uniform frequency over T*

rounds. Consequently, the parameters of submodel θ_i maintain a linear scaling relationship with the global model θ_g (i.e., $\theta_i^{t+1} = p_i\theta_g^t$). Here, p_i ($0 \leqslant p_i \leqslant 1$) represents the long-term expected size ratio between the submodel and the aggregate model. The updated global model is reconstructed by: $\theta_g^{t+1} = \sum_{i=1}^{N} \frac{\theta_i^{t+1}}{p_i}$.

Lemma 1 *(Result of one step SGD). Assume ASSUMPTION 1 and ASSUMPTION 2. If $\eta_t \leqslant \frac{1}{4L}$, we have*

$$\begin{aligned} E\|\bar{v}_{t+1} - \theta^*\|^2 \leqslant & (1-\eta_t\mu)E\|\bar{\theta}_t - \theta^*\|^2 \\ & + \eta_t^2 E\|g_t - \bar{g}_t\|^2 + 6L\eta_t^2\Gamma \\ & + 2E\sum_{i=1}^{N}\frac{\|\bar{\theta}_t - \theta_i^t\|^2}{p_i}, \end{aligned} \tag{21}$$

where $\Gamma = F^* - \sum_{i=1}^{N} \frac{F_i^*}{p_i}$. Lemma 1 follows from [17].

Since FedRAC requires communication every E steps, we assume $\eta_t \leqslant 2\eta_{t+E}$. Consequently, for any $t \geq 0$, there exists an index t_0 satisfying $t - t_0 \leqslant E-1$ and $\theta_i^{t_0} = \bar{\theta}_{t_0}$ for all $i = 1, 2, ..., N$. It follows that

$$\begin{aligned} E\sum_{i=1}^{N}\frac{1}{p_i}\|\bar{\theta}_t - \theta_i^t\|^2 &\leqslant E\sum_{i=1}^{N}\frac{1}{p_i}\|\theta_i^t - \bar{\theta}_{t_0}\|^2 \\ &\leqslant E\sum_{t=t_0}^{t-1}(E-1)\eta_t^2\|\nabla F_k(\theta_i^t, \xi_i^t)\|^2 \\ &\leqslant \sum_{t=t_0}^{t-1}(E-1)\eta_{t_0}^2 G^2 \\ &\leqslant 4\eta_t^2(E-1)^2G^2. \end{aligned} \tag{22}$$

We use $E\|X - EX\|^2 \leqslant E\|X\|^2$ where $X = \theta_i^t - \bar{\theta}_{t_0}$ with probability $\frac{1}{p_i}$. We use Jensen inequality:

$$\begin{aligned} \|\theta_i^t - \bar{\theta}_{t_0}\| &= \|\sum_{t=t_0}^{t-1}\eta_t\nabla F_i(\theta_i^t, \xi_i^t)\|^2 \\ &\leqslant (t-t_0)\sum_{t-t_0}^{t-1}\sum_{t-t_0}^{t-1}\eta_t^2\|\nabla F_i(\theta_i^t, \xi_i^t)\|^2. \end{aligned} \tag{23}$$

Here, we utilize $\eta_t \leqslant \eta_{t_0}$ for $t \geq t_0$ and $E\|\nabla F_k(\theta_i^t, \xi_i^t)\|^2 \leqslant G^2$ for $i = 1, 2, ..., N$. We $\eta_{t_0} \leqslant 2\eta_{t_0+E} \leqslant 2\eta_t$ for $t_0 \leqslant t \leqslant t_0 + E$.

According to Assumption 3, the variance of the stochastic gradients within client i is upper-bounded by a constant σ_i^2. Consequently,

$$\begin{aligned} E\|g_t - \bar{g}_t\|^2 &= E\|\sum_{i=1}^{N}\frac{1}{p_i}\left(\nabla F_k(\theta_i^t, \xi_i^t) - \nabla F_i(\theta_i^t)\right)\|^2 \\ &\leqslant \sum_{i=1}^{N}\frac{1}{p_i^2}\sigma_i^2. \end{aligned} \tag{24}$$

Let $\triangle_t = E\|\bar{\theta}_t - \theta^*\|$. From equations 21-24, we have

$$\begin{aligned} \triangle_{t+1} &\leqslant (1-\eta_t\mu)\triangle_t + \eta_t^2\sum_{i=1}^{N}\frac{\sigma^2}{p_i^2} + 6L\eta_t^2\Gamma + 8\eta_t^2(E-1)^2G^2 \\ &\leqslant (1-\eta_t\mu)\triangle_t + \eta_t^2\underbrace{\left(\sum_{i=1}^{N}\frac{\sigma^2}{p_i^2} + 6L\Gamma + 8(E-1)^2G^2\right)}_{B}. \end{aligned} \tag{25}$$

For a diminishing step-size $\eta_t = \frac{\kappa}{t+\gamma}$ (where $\kappa > \frac{1}{\mu}$, $\gamma > 0$, $\eta_1 \leqslant min\{\frac{1}{\mu}, \frac{1}{4L}\} = \frac{1}{4L}$, and $\eta_t \leqslant 2\eta_{t+E}$). We aim to prove $\triangle \leqslant \frac{v}{\gamma+t}$ by induction, where $v = max\{\frac{\kappa^2 B}{\kappa\mu - 1}, (\gamma+1)\triangle_1\}$. The definition of v ensures that the inequality holds for the base case $t = 1$. Assuming the validity of the hypothesis for an arbitrary $t \geq 1$, we have

$$\begin{aligned} \triangle_{t+1} &\leqslant (1-\eta_t\mu)\triangle_t + \eta_t^2 B \\ &\leqslant (1 - \frac{\kappa\mu}{t+\gamma})\frac{v}{t+\gamma} + \frac{\kappa^2 B}{(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2}v + [\frac{\kappa^2 B}{(t+\gamma)^2} - \frac{\kappa\mu - 1}{(t+\gamma)^2}v] \\ &\leqslant \frac{t+\gamma-1}{(t+\gamma)^2}v \underbrace{- \frac{\kappa\mu - 1}{(t+\gamma)^2(\gamma+1)\triangle_1}}_{\leqslant 0} \\ &\leqslant \frac{v}{t+\gamma-1}. \end{aligned} \tag{26}$$

By the L-smoothness of F (ASSUMPTION 1), we have

$$\begin{aligned} E[F(\bar{\theta}_T)] - F^* &\leqslant (\bar{\theta}_T - \theta^*)^T\underbrace{\nabla F_i(\theta^*)}_{=0} \\ &\quad + \frac{L}{2}\|\bar{\theta}_T - \theta^*\|_2^2 \\ &\leqslant \frac{L}{2}\frac{v}{\gamma+T}. \end{aligned} \tag{27}$$

With $\kappa = \frac{2}{\mu}$ and $\gamma = max\{8\frac{L}{\mu}, E\} - 1$, it follows that

$$\begin{aligned} v &= max\{\frac{\kappa^2 B}{\kappa\mu - 1}, (\gamma+1)\triangle_1\} \\ &\leqslant \frac{\kappa^2 B}{\kappa\mu - 1} + (\gamma+1)\triangle_1 \\ &\leqslant \frac{4B}{\mu^2} + (\gamma+1)\triangle_1. \end{aligned} \tag{28}$$

By applying equation 28 into equation 27, it follows that

$$\begin{aligned} 0 &\leqslant \lim_{T\to\infty} E[F(\bar{\theta}_T)] - F^* \\ &\leqslant \lim_{T\to\infty}\left[\frac{L}{\gamma+T}\left(\frac{2B}{\mu^2} + \frac{\gamma+1}{2}\triangle_1\right)\right] \\ &= 0. \end{aligned} \tag{29}$$

Therefore, $\lim_{T\to\infty} E[F(\bar{\theta}_T)] - F^* = 0$.

10. Setup Supplement

10.1. Data Splits

POW (imbalanced dataset sizes). We apply a power-law distribution to randomly assign the entire dataset to multiple clients, ensuring imbalanced data volume. In these settings, clients with larger datasets are expected to attain superior predictive performance.

CLA (imbalanced class numbers). We vary the number of classes while keeping the total data volume unchanged. In a five-client setup using CIFAR10, $Client_1$, $Client_2$, $Client_3$, $Client_4$, and $Client_5$ are assigned local training data containing 1, 3, 5, 7, and 10 classes, respectively.

DIR (imbalanced data size and class numbers). We adopt a Dirichlet distribution $DIR(\alpha)$ to provide each client with data of varying volumes and classes. Specifically, we sample $p_i^l \sim DIR(\alpha)$ from a Dirichlet distribution with parameter α, and assign class l to client i according to the sampled proportions p_i^l.

10.2. Baselines Methods

We provide detailed descriptions of all baseline methods considered in our experiments. FedAvg distributes the same model to all clients and then trains it locally for several rounds at each communication round. CFFL computes a reputation score based on local accuracy and data size (or label diversity), allocating more gradients to clients with higher reputations. CGSV rewards clients whose local model gradients are more similar to the global gradient. FedAVE compares the similarity between local and ideal data distributions, assigning more gradients to clients with higher similarity. IAFL allocates more gradient updates to highly contributive clients while distributing a reference model randomly to all clients. FedSAC evaluates neuron importance in the global model and assigns more important neurons to clients with higher contributions. Standalone denotes the case where each client trains its model independently without federated aggregation. To ensure fairness, we require all algorithms to allocate rewards according to client contributions, rather than relying on their reputation mechanism.

10.3. The Impact of β and λ on Experiment

In Figure 3, we present the performance of FedRAC under different values of β and λ across the POW, DIR(3.0), and DIR(7.0) scenarios of CIFAR10. As shown in Figure 3(a), larger β widen reputation gaps, enhancing fairness while slightly degrading model performance. In Figure 3(b), larger λ suppress the reputation r, reducing clients' rewards and lowering performance.

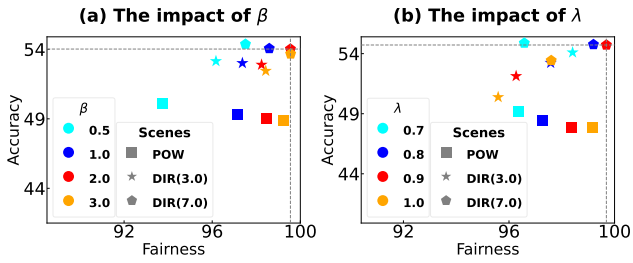

Figure 3. Comparison results of FedRAC across diverse scenarios with varying β and λ. Results in the upper-right region represent superior performance.

10.4. Overall Performance in Collaborative Fairness

To compare all methods comprehensively, we present their overall performance in Figure 4. The figure illustrates the trade-offs among fairness, accuracy, and rate on the EMNIST dataset. FedRAC maintains a dominant lead in fairness, accuracy, and rate metrics across various heterogeneous scenarios in the figure. Furthermore, we verify FedRAC's extensibility to complex datasets (i.e., Tiny-ImageNet) and models (i.e., ResNet18) in Figure 5. The results confirm FedRAC's generalizability in both fairness and accuracy.

10.5. Ablation Study

To evaluate the effectiveness of the two proposed modules in FedRAC, we conducted ablation studies on the CIFAR10 dataset. $w/o\ reputation$ denotes removing the Dynamic Reputation Calculation module, where each client's reputation remains fixed throughout training. As shown in Table 7 and Table 8 of the CIFAR10-DIR(3.0) scenario, the fairness decreases from 98.76% to 88.17%, and accuracy drops from 53.14% to 50.22%. These results demonstrate that the reputation calculation module is essential for maintaining both fairness and model accuracy. $w/o\ allocation$ refers to removing the dynamic submodel allocation, which randomly constructs client submodels. As illustrated in Table 8, accuracy in the CIFAR10-POW scenario decreases from 49.37% to 47.37% without this module. Furthermore, Table 9 indicates a general degradation in the rate metric across all data splits when this component is excluded. This confirms that dynamic submodel allocation further improves the model's fairness and stability. Overall, the ablation results show that both designed modules are indispensable.

11. Complexity and Communication Cost Analysis

In this section, we analyze the time complexity and communication costs of FedRAC as follows. **For the time complexity**, the primary computational demand in FedRAC

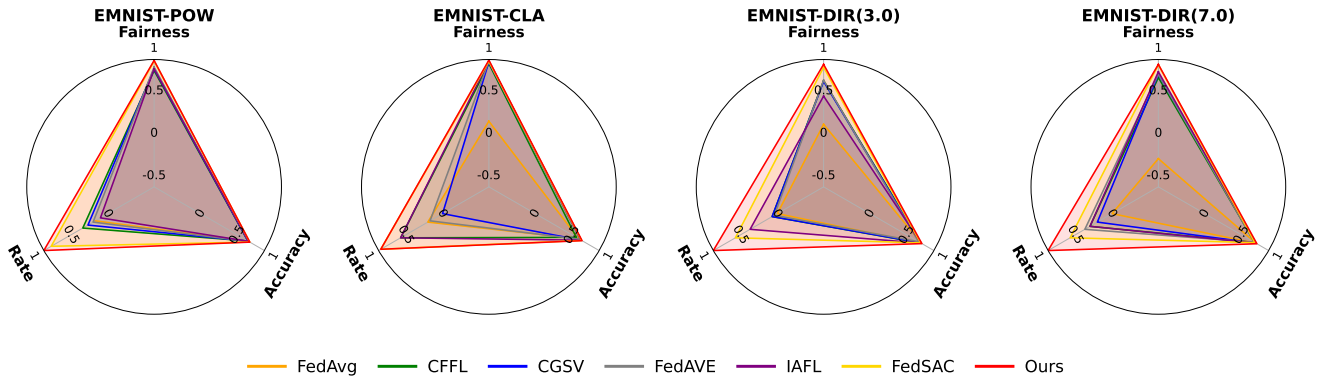

Figure 4. The visualization results of fairness, accuracy, and rate in EMNIST. The larger the area of the triangle, the better the algorithm performs.

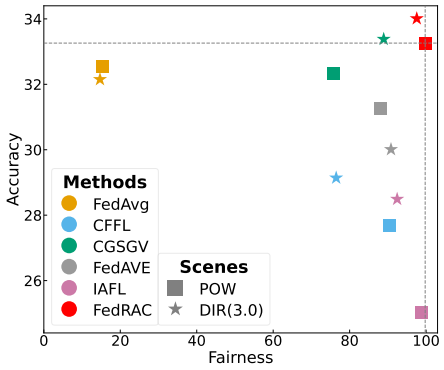

Figure 5. Overall performance comparison between FedRAC and existing methods for achieving CF on Tiny-ImageNet. Results in the upper-right region represent superior performance.

Dataset	CIFAR10			
Scene	POW	CLA	DIR(3.0)	DIR(7.0)
w/o reputation	99.06	99.51	88.17	98.81
w/o allocation	98.59	98.43	76.67	88.03
FedRAC (Ours)	**99.41**	**99.72**	**98.76**	**99.53**

Table 7. Ablation studies on FedRAC for the fairness on CIFAR10.

arises from submodel allocation (as defined in Formula (9)). The time complexity of this process is $O(M)$, where M is proportional to the number of hidden-layer neurons in the global model. **For the communication cost**, the submodel transmission enables FedRAC to reduce excess communication overhead. The per-round communication complexity of FedRAC is $O(M \times p)$, where $p \leq 1$ represents the average proportion of submodel parameters relative to the global model. This design effectively reduces unnecessary client-server data exchange, thereby boosting overall communication efficiency.

Dataset	CIFAR10			
Scene	POW	CLA	DIR(3.0)	DIR(7.0)
w/o reputation	48.21	43.92	50.22	50.29
w/o allocation	47.37	44.05	51.81	52.03
FedRAC (Ours)	**49.37**	**44.28**	**53.14**	**54.71**

Table 8. Ablation studies on FedRAC for the maximum test accuracy on CIFAR10.

Dataset	CIFAR10			
Scene	POW	CLA	DIR(3.0)	DIR(7.0)
w/o reputation	0.5	0.8	1.0	1.0
w/o allocation	0.2	0.9	0.4	0.5
FedRAC (Ours)	**1.0**	**0.97**	**1.0**	**1.0**

Table 9. Ablation studies on FedRAC for the rate on CIFAR10.

12. Limitation

In Tables 1 to 3, it can be observed that FedRAC achieves the best performance (in terms of fairness, model accuracy, and rate) compared to existing methods. Nevertheless, the submodel allocation (refer to Section 4.2.2) may incur a modest degree of additional computational overhead. This minor limitation is primarily more noticeable in the case of large-scale models. In future work, we will explore the methods for the construction of submodels rapidly, aiming to efficiently accommodate large-scale model scenarios.

13. The Impact of α on Experiments

To verify the impact of different α values in Definition 1 (α-Bounded Collaborative Fairness) on the experimental results, we present the maximum accuracy of FedRAC as α increases in Table 10, under the constraints of a fairness threshold $\gamma > 95$ and rate > 0.8. In Table 10, it can be observed that the performance of FedRAC gradually improves as α increases. For example, in a POW

Dataset	CIFAR10			
Scene	POW	CLA	DIR(3.0)	DIR(7.0)
FedRAC(α=0.1)	47.52	40.86	50.74	52.51
FedRAC(α=0.3)	47.68	42.29	50.95	52.73
FedRAC(α=0.5)	**49.37**	**44.28**	**53.14**	**54.71**

Table 10. The maximum test accuracy (%) of FedRAC across different α values, under the constraints of a fairness threshold $\gamma >$ 95 and rate $>$ 0.8, on the CIFAR10 dataset.

scenario, FedRAC achieves a performance score of 47.52 at $\alpha = 0.1$ and 49.37 at $\alpha = 0.5$. This is attributed to $((1-\alpha)\frac{(c_i)}{\max(c)} + \alpha) * \max(\theta^*)$ increasing with α in Definition 1, which raises the reward upper bound for low-contribution clients. This boosts all clients' rewards, enabling more thorough training of the aggregated model and thus improving the performance of FedRAC.