Few-for-Many Personalized Federated Learning Supplementary Material

Ping Guo[1,6], Tiantian Zhang[2], Xi Lin[3], Xiang Li[4], Zhi-Ri Tang[5], Qingfu Zhang[1,6]
[1]City University of Hong Kong; [2]Hong Kong Metropolitan University;
[3]Xi'an Jiaotong University; [4]Southeast University; [5]Jinan University;
[6]CityU Shenzhen Research Institute

1. Theoretical Analysis

This supplementary material provides detailed proofs for the theorems presented in the main paper. We organize the material as follows: (A) proof of convergence of K-for-M framework (Theorem 3.1 from Section 3.2), (B) proof of uniform smooth approximation (Theorem 4.1 from Section 4.4), and (C) proof of Pareto properties (Theorem 4.2 from Section 4.4), including both Pareto optimality and Pareto stationarity.

1.1. Notation

We begin by establishing the notation used for all proofs. Table 1 summarizes the key symbols and their definitions.

1.2. Theorem 3.1: K-for-M Convergence

We now provide the complete proof of Theorem 3.1 from the main paper. Following the notation in the theorem statement, we use $\Theta^{(K)} = \{\theta_1, \ldots, \theta_K\}$ to denote the K-for-M solution obtained by minimizing empirical losses over finite samples.

Proof roadmap. The proof proceeds in three steps. First, we establish that each model in the K-for-M solution lies on the Pareto frontier (Lemma 1.1). Second, we bound the Pareto coverage gap, which measures how well K models approximate M personalized optima (Lemma 1.2). This bound depends on the maximum heterogeneity across clients. Third, we bound the statistical error arising from finite-sample learning (Lemma 1.3). Combining these two independent error sources yields the final convergence rate.

To quantify the approximation quality, we introduce the notion of maximum heterogeneity:

Definition 1.1 (Maximum Heterogeneity)**.** The maximum pairwise heterogeneity is defined as:

$$\Delta_{het} = \max_{i,j \in [M]} \left[L_i(\theta_j^*) - L_i(\theta_i^*) \right] \tag{1}$$

This measures the worst-case loss degradation when a client uses another client's optimal model instead of its own.

Table 1. Key notation for theoretical analysis. Stars (*) denote optimal or population-level quantities, while hats ($\hat{\cdot}$) denote empirical estimators.

Symbol	Description
M	number of heterogeneous clients
K	number of shared models in the K-for-M framework
$L_i(\theta)$	expected (population) loss for client i with model θ
$\hat{L}_i(\theta)$	empirical loss for client i based on finite samples
θ_i^*	optimal personalized model for client i, defined as $\arg\min_\theta L_i(\theta)$
$\Theta_*^{(K)}$	optimal K-for-M solution (minimizing population losses)
$\hat{\Theta}^{(K)}$	empirical K-for-M solution (minimizing empirical losses)
n	average sample size per client
d	VC dimension of hypothesis class Θ

1.2.1. Pareto Coverage Analysis

We analyze how well K models can approximate the Pareto set. Following the Pareto optimality definition in the main paper, let $\mathcal{P} = \{\theta : \nexists \theta' \text{ s.t. } L_i(\theta') \leq L_i(\theta)\, \forall i \text{ with } L_j(\theta') < L_j(\theta) \text{ for some } j\}$ denote the set of all Pareto optimal models.

The K-for-M solution $\Theta_*^{(K)}$ is defined as:

$$\Theta_*^{(K)} = \arg \min_{\theta_1, \ldots, \theta_K} \sum_{i=1}^{M} \min_{k \in [K]} L_i(\theta_k) \tag{2}$$

Lemma 1.1 (Pareto Optimality of K-for-M Solution)**.** Each model $\theta_k^* \in \Theta_*^{(K)}$ is Pareto optimal, i.e., $\theta_k^* \in \mathcal{P}$.

Proof. Suppose for contradiction that some $\theta_k^* \notin \mathcal{P}$. Then there exists a Pareto-dominating model $\theta' \in \mathcal{P}$ with

$L_i(\theta') \leq L_i(\theta_k^*)$ for all i and $L_j(\theta') < L_j(\theta_k^*)$ for some j. Replacing θ_k^* with θ' in $\Theta_*^{(K)}$ strictly decreases the objective, contradicting optimality. □

Model Capacity Gap. Define the Pareto endpoints as the M extreme points on the Pareto frontier:

$$\theta_i^* = \arg\min_{\theta} L_i(\theta), \quad i = 1, \ldots, M \tag{3}$$

The model capacity gap measures how well K Pareto points approximate these M endpoints. For client i, the gap is defined as:

$$\text{Gap}_i(K) = \min_{k \in [K]} L_i(\theta_k^*) - L_i(\theta_i^*) \tag{4}$$

Lemma 1.2 (Pareto Coverage Bound)**.** Under a clustering assumption where clients can be approximately partitioned into K groups with balanced sizes, the average model capacity gap is bounded by:

$$\frac{1}{M}\sum_{i=1}^{M} \text{Gap}_i(K) \leq \left(1 - \frac{K}{M}\right) \cdot \Delta_{het} \tag{5}$$

Proof. Assume clients partition into K groups $\mathcal{G}_1, \ldots, \mathcal{G}_K$ with $|\mathcal{G}_k| \approx M/K$. Under the clustering assumption, the K-for-M solution aligns with this partition: each group $\mathcal{G}_k$ has a representative client r_k whose optimal model $\theta_{r_k}^*$ is included in $\Theta_*^{(K)}$.

For each group $\mathcal{G}_k$:

- **Representative clients** (K clients): For $i = r_k$, we have $\text{Gap}_{r_k}(K) = 0$ since $\theta_{r_k}^* \in \Theta_*^{(K)}$.
- **Non-representative clients** ($M - K$ clients): For $i \in \mathcal{G}_k \setminus \{r_k\}$:

$$\text{Gap}_i(K) = L_i(\theta_{r_k}^*) - L_i(\theta_i^*) \leq \Delta_{het} \tag{6}$$

Averaging over all M clients:

$$\frac{1}{M}\sum_{i=1}^{M} \text{Gap}_i(K) = \frac{1}{M}\left[\sum_{k=1}^{K} 0 + \sum_{k=1}^{K} \sum_{i \in \mathcal{G}_k \setminus \{r_k\}} \text{Gap}_i(K)\right] \tag{7}$$

$$\leq \frac{1}{M} \cdot (M - K) \cdot \Delta_{het} \tag{8}$$

$$= \left(1 - \frac{K}{M}\right) \cdot \Delta_{het} \tag{9}$$

Note that when $K = M$, every client is a representative, yielding zero gap. □

1.2.2. Statistical Convergence Analysis

We now analyze the statistical error arising from learning with finite samples.

With finite samples, we optimize empirical losses $\{\hat{L}_i(\theta)\}_{i=1}^M$ instead of population losses $\{L_i(\theta)\}_{i=1}^M$. The empirical K-for-M solution is:

$$\Theta^{(K)} = \arg\min_{\theta_1, \ldots, \theta_K} \sum_{i=1}^{M} \min_{k \in [K]} \hat{L}_i(\theta_k) \tag{10}$$

1.2.3. Convergence Bound

Lemma 1.3 (Statistical Convergence)**.** For a hypothesis class Θ with VC dimension d, with probability at least $1-\delta$:

$$\frac{1}{M}\sum_{i=1}^{M} \left| \min_k L_i(\theta_k) - \min_k L_i(\theta_k^*) \right| \leq \mathcal{O}\left(\sqrt{\frac{Kd + \log(M/\delta)}{n}}\right) \tag{11}$$

where n is the average sample size per client.

Proof. The K-for-M problem optimizes over the product space Θ^K with VC dimension $\mathcal{O}(Kd)$. By standard ERM analysis, for each client i:

$$\begin{aligned} &\min_k L_i(\theta_k) - \min_k L_i(\theta_k^*) \\ &= \underbrace{\left[\min_k L_i(\theta_k) - \min_k \hat{L}_i(\theta_k)\right]}_{\leq \epsilon(n, Kd)} \\ &\quad + \underbrace{\left[\min_k \hat{L}_i(\theta_k) - \min_k \hat{L}_i(\theta_k^*)\right]}_{\leq 0} \\ &\quad + \underbrace{\left[\min_k \hat{L}_i(\theta_k^*) - \min_k L_i(\theta_k^*)\right]}_{\leq \epsilon(n, Kd)} \end{aligned} \tag{12}$$

where the first and third terms are bounded by uniform convergence (Theorem 6.8 in [3]), and the second term is non-positive since $\Theta^{(K)}$ minimizes the empirical objective.

Applying a union bound over M clients and absorbing logarithmic factors yields the stated bound. □

Remark 1.1 (Union Bound)**.** The union bound (Boole's inequality) states that $P[\bigcup_{i=1}^M E_i] \leq \sum_{i=1}^M P[E_i]$. For M clients each with failure probability δ/M, the total failure probability is at most δ. The exact probability is $1 - (1 - \delta/M)^M \approx \delta$ for small δ/M.

1.2.4. Combining the Bounds

The total error for the empirical K-for-M solution decomposes as:

$$\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}\left[\mathbb{E}[L_i(\theta_{k_i})] - L_i(\theta_i^*)\right] \\
&= \underbrace{\frac{1}{M}\sum_{i=1}^{M}\left[L_i(\theta_{k_i}^*) - L_i(\theta_i^*)\right]}_{\text{Pareto coverage gap}} \\
&\quad + \underbrace{\frac{1}{M}\sum_{i=1}^{M}\left[\mathbb{E}[L_i(\theta_{k_i})] - L_i(\theta_{k_i}^*)\right]}_{\text{statistical error}} \qquad (13)
\end{aligned}$$

where $k_i = \arg\min_k L_i(\theta_k)$ for each client i.

Combining Lemmas 1.2 and 1.3:

$$\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}\left\{\mathbb{E}\left[\min_{k\in[K]} L_i(\theta_k)\right] - L_i(\theta_i^*)\right\} \\
&\le \frac{M-K}{M}\cdot\Delta_{het} + \mathcal{O}\left(\sqrt{\frac{Kd}{n}}\right) \qquad (14)
\end{aligned}$$

This completes the proof of Theorem 3.1.

Remark 1.2 (Interpretation of the Bound)**.** The bound reveals a fundamental trade-off in the K-for-M framework:

- **Pareto coverage gap** $(M-K)/M\cdot\Delta_{het}$: Decreases with K, vanishes when $K = M$
- **Statistical error** $\mathcal{O}(\sqrt{Kd/n})$: Increases with K due to larger hypothesis class
- **Optimal** K: Balances model expressiveness against sample efficiency
- **Asymptotic behavior**: As $n \to \infty$, statistical error vanishes, leaving only the coverage gap

1.3. Theorem 4.1: Smooth Approximation

We prove that $g^{\text{STCH-Set}}(\Theta_K)$ uniformly approximates $g^{\text{TCH-Set}}(\Theta_K)$ by deriving tight upper and lower bounds using standard log-sum-exp approximation properties.

Proof. The log-sum-exp function provides well-known smooth approximations for max and min operators [1, 2]. For any $y_1, \ldots, y_n$ and smoothing parameter $\mu > 0$:

$$\begin{aligned}
&\mu\log\sum_{i=1}^{n} e^{y_i/\mu} - \mu\log n \\
&\le \max\{y_1,\ldots,y_n\} \\
&\le \mu\log\sum_{i=1}^{n} e^{y_i/\mu}, \qquad (15)
\end{aligned}$$

$$\begin{aligned}
&-\mu\log\sum_{i=1}^{n} e^{-y_i/\mu} \\
&\le \min\{y_1,\ldots,y_n\} \\
&\le -\mu\log\sum_{i=1}^{n} e^{-y_i/\mu} + \mu\log n. \qquad (16)
\end{aligned}$$

For $g^{\text{TCH-Set}}(\Theta_K) = \max_{i\in[M]}\min_{k\in[K]} L_i(\theta_k)$, we apply (16) to the inner minimization and (15) to the outer maximization. Define the smooth inner minimum:

$$\tilde{m}_i := -\mu\log\left(\sum_{k=1}^{K} e^{-L_i(\theta_k)/\mu}\right). \qquad (17)$$

By (16), we have $\tilde{m}_i - \mu\log K \le \min_{k\in[K]} L_i(\theta_k) \le \tilde{m}_i$. Applying (15) to $\max_{i\in[M]}\tilde{m}_i$ and noting that $g^{\text{STCH-Set}}(\Theta_K) = \mu\log(\sum_{i=1}^{M} e^{\tilde{m}_i/\mu})$:

$$\begin{aligned}
g^{\text{TCH-Set}}(\Theta_K) &= \max_{i\in[M]}\min_{k\in[K]} L_i(\theta_k) \\
&\ge \max_{i\in[M]}(\tilde{m}_i - \mu\log K) \\
&\ge g^{\text{STCH-Set}}(\Theta_K) - \mu\log M - \mu\log K, \qquad (18)
\end{aligned}$$

$$g^{\text{TCH-Set}}(\Theta_K) \le \max_{i\in[M]}\tilde{m}_i \le g^{\text{STCH-Set}}(\Theta_K). \qquad (19)$$

Combining (18) and (19) yields the uniform approximation bound with error $\mathcal{O}(\mu\log M + \mu\log K)$, which vanishes as $\mu \to 0$. □

1.4. Theorem 4.2: Pareto Properties

We prove both parts of Theorem 4.2: Pareto optimality and Pareto stationarity.

1.4.1. Part 1: Pareto Optimality

Proof. The smooth Tchebycheff set scalarization objective:

$$g^{\text{STCH-Set}}(\Theta_K) = \mu\log\sum_{i=1}^{M}\left(\sum_{k=1}^{K}\exp\left(-\frac{L_i(\theta_k)}{\mu}\right)\right)^{-1} \qquad (20)$$

Let $\Theta_K^* = \{\theta_1^*, \ldots, \theta_K^*\}$ be an optimal solution set. We need to show that each $\theta_k^* \in \Theta_K^*$ is Pareto optimal.

Part 1: Weak Pareto Optimality. Suppose for contradiction that Θ_K^* is not weakly Pareto optimal. Then there

Table 2. Training hyperparameters for benchmark and medical datasets. Benchmark datasets use 2000 rounds with CNN/TextCNN backbones, while medical datasets use 1000 rounds with ResNet-18 to prevent overfitting on smaller institutional samples. All experiments employ single local epoch and full client participation for fair comparison across methods.

Type	Dataset	Model	Rounds	Local Epochs	Batch Size	Learning Rate	Join Ratio
Benchmark	CIFAR-10	CNN	2000	1	50	0.005	1.0
	CIFAR-100	CNN	2000	1	50	0.005	1.0
	TinyImageNet	CNN	2000	1	50	0.0005	1.0
	AG News	TextCNN	2000	1	100	0.005	1.0
	FEMNIST	CNN	2000	1	100	0.005	1.0
Medical	Kvasir	ResNet-18	1000	1	100	0.002	1.0
	FedISIC	ResNet-18	1000	1	50	0.005	1.0

exists another set Θ'_K such that $L_i(\theta'_{k(i)}) < L_i(\theta^*_{k(i)})$ for all $i \in [M]$, where $k(i) = \arg\min_k L_i(\theta_k)$.

Since all objectives strictly decrease:

$$\sum_{k=1}^{K} \exp\left(-\frac{L_i(\theta'_k)}{\mu}\right) > \sum_{k=1}^{K} \exp\left(-\frac{L_i(\theta^*_k)}{\mu}\right) \quad \forall i \quad (21)$$

This implies:

$$\left(\sum_{k=1}^{K} \exp\left(-\frac{L_i(\theta'_k)}{\mu}\right)\right)^{-1} < \left(\sum_{k=1}^{K} \exp\left(-\frac{L_i(\theta^*_k)}{\mu}\right)\right)^{-1} \quad \forall i \quad (22)$$

Therefore $g^{\text{STCH-Set}}(\Theta'_K) < g^{\text{STCH-Set}}(\Theta^*_K)$, contradicting the optimality of Θ^*_K.

Strong Pareto Optimality. Under either condition (unique optimal set or all positive preferences), we can strengthen the result to Pareto optimality. When the optimal set is unique, any Pareto-dominating solution would yield a strictly better objective value, contradicting uniqueness. When all preferences are positive, the scalarization ensures that improving any subset of objectives without harming others strictly decreases the objective, again contradicting optimality. □

1.4.2. Part 2: Pareto Stationarity

We prove that stationary points of STCH-Set are Pareto stationary for the original multi-objective problem.

Proof. Consider a point $\hat{\Theta} = \{\hat{\theta}_1, \ldots, \hat{\theta}_K\}$ where gradient descent has converged. The gradient of STCH-Set with respect to model $\hat{\theta}_k$ is:

$$\nabla_{\hat{\theta}_k} g^{\text{STCH-Set}} = \sum_{i=1}^{M} \alpha_i \cdot w_{ik} \cdot \nabla_{\hat{\theta}_k} L_i(\hat{\theta}_k) \quad (23)$$

where:

$$w_{ik} = \frac{\exp(-L_i(\hat{\theta}_k)/\mu)}{\sum_{j=1}^{K} \exp(-L_i(\hat{\theta}_j)/\mu)} \quad (24)$$

$$\alpha_i = \frac{S_i^{-1}}{\sum_{j=1}^{M} S_j^{-1}} \quad (25)$$

At a stationary point where $\nabla_{\hat{\theta}_k} g^{\text{STCH-Set}} = 0$ for all k, we have:

$$\sum_{i=1}^{M} \bar{w}_i \nabla_{\hat{\theta}_k} L_i(\hat{\theta}_k) = 0 \quad (26)$$

where $\bar{w}_i = \alpha_i \cdot w_{ik} \geq 0$ and $\sum_i \bar{w}_i = 1$ (forms a convex combination).

This is precisely the Pareto stationarity condition: the zero vector can be expressed as a convex combination of the individual gradients, meaning no common descent direction exists that improves all objectives simultaneously. □

2. Experimental Details

2.1. Training Hyperparameters

Table 2 presents the complete training configurations for all datasets evaluated in our experiments. We categorize the datasets into benchmark and medical imaging domains, each requiring tailored hyperparameter settings due to their distinct characteristics.

Benchmark datasets. We evaluate on five diverse benchmark datasets (CIFAR-10, CIFAR-100, TinyImageNet, AG News, FEMNIST) spanning vision and text domains. All benchmark datasets use 2000 communication rounds to ensure convergence across heterogeneous client distributions. For vision datasets, CIFAR-10 and CIFAR-100 use CNN backbones with batch size 50 and learning rate 0.005, balancing training stability with limited samples per client. TinyImageNet employs the same CNN architecture and batch size, but uses a reduced learning rate of 0.0005 to accommodate its higher resolution (64×64) and larger number of classes (200). For text classification, AG

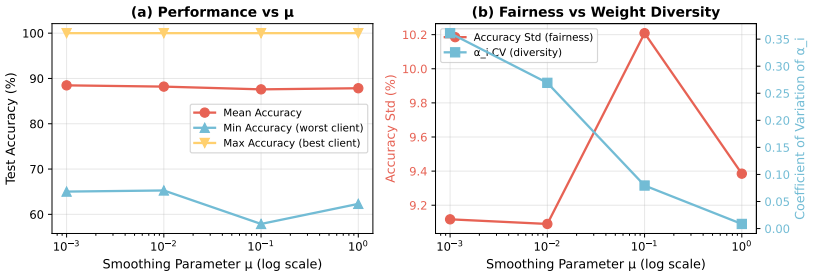

Figure 1. **Fairness and weight diversity analysis.** (a) Performance metrics across different μ values show that mean accuracy is relatively stable, but worst-case (minimum) accuracy drops significantly at $\mu = 0.1$, suggesting a phase transition region. (b) The relationship between fairness (accuracy standard deviation, red) and outer weight diversity (coefficient of variation of α_i, blue) reveals that both extreme values ($\mu \to 0$ and $\mu \to \infty$) achieve better fairness than intermediate values, with outer weight diversity decreasing monotonically as μ increases.

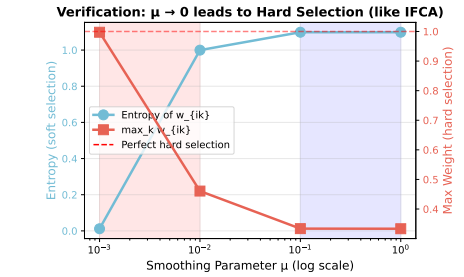

Figure 2. **Theoretical verification: $\mu \to 0$ recovers hard clustering.** Left axis (blue): entropy of inner weights w_{ik} increases with μ, indicating softer model selection. Right axis (red): maximum inner weight decreases with μ, moving away from one-hot assignments. The shaded regions indicate hard selection ($\mu < 0.01$, red) and soft selection ($\mu > 0.1$, blue) regimes.

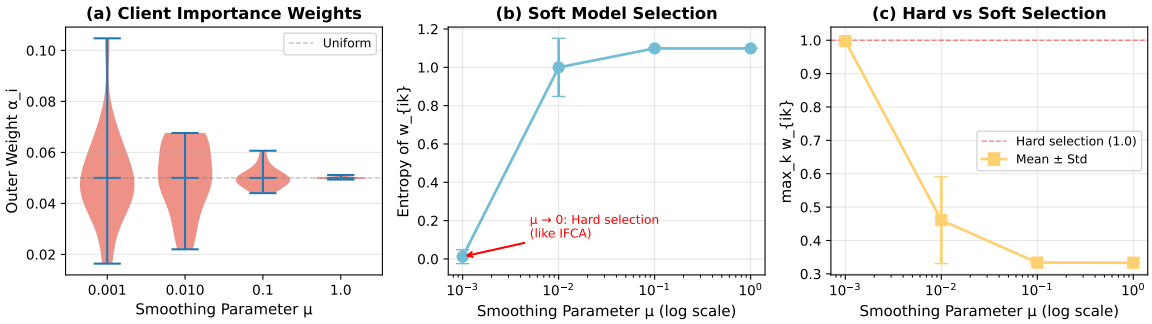

Figure 3. **Impact of μ on dual-layer weights.** (a) Outer weights α_i distribution across clients. (b) Inner weight entropy (higher = softer selection). (c) Max inner weight (closer to 1 = harder selection). As $\mu \to 0$, FedFew recovers hard clustering like IFCA.

News uses TextCNN with batch size 100 and learning rate 0.005, leveraging vocabulary diversity to mitigate overfitting. FEMNIST adopts CNN with batch size 100 and learning rate 0.005, benefiting from natural user partitioning that reduces overfitting tendencies.

Medical datasets. We include two medical imaging datasets (Kvasir, FedISIC) representing real-world healthcare scenarios. Both use ResNet-18 backbones and 1000 communication rounds, as medical data exhibits faster convergence and higher overfitting risks due to smaller sample sizes per institution. Kvasir, focusing on gastrointestinal disease classification, uses batch size 100 and a conservative learning rate of 0.002 to handle fine-grained categories while exploiting data augmentation. FedISIC, dealing with skin lesion classification from small medical centers, adopts batch size 50 and learning rate 0.005 to prevent overfitting on limited training samples.

Common settings. Across all datasets, we fix local epochs to 1 and join ratio to 1.0 (full client participation) to ensure fair comparison between different personalized FL methods. These settings align with standard practices in federated learning benchmarks.

Algorithm-specific hyperparameters. For FedFew and IFCA, we use $K = 3$ server models across all experiments, which provides a good balance between model expressiveness and optimization complexity as validated in Section 5.3.1. For FedFew specifically, we set the smoothing parameter $\mu = 0.01$ for the STCH-Set objective, which enables effective soft model selection while maintaining stable optimization (see Section 2.2 for sensitivity analysis).

2.2. Sensitivity Analysis on Smoothing Parameter

We investigate the impact of the smoothing parameter μ on both the dual-layer weight mechanism and overall performance. Figure 3 illustrates how μ controls the balance between hard and soft model selection. As theory predicts, when $\mu \to 0$, the inner weights w_{ik} approach one-hot assignments (entropy ≈ 0.012, max weight ≈ 0.997), recovering IFCA-style hard clustering. Conversely, for large μ (e.g., $\mu = 1.0$), the weights become nearly uniform (entropy $\approx 1.099 \approx \log 3$, max weight ≈ 0.333), enabling soft model selection. The outer weights α_i exhibit complementary behavior: smaller μ values lead to more diverse client importance weights (CV = 0.361 at $\mu = 0.001$), emphasiz-

ing adaptive up-weighting of harder clients, while larger μ yields nearly uniform weighting (CV = 0.009 at $\mu = 1.0$). On accuracy, we find that performance remains relatively stable across different μ values.

We provide additional analysis on the impact of the smoothing parameter μ beyond what is presented in the main paper. Figure 1 examines the relationship between weight diversity and fairness. Interestingly, both very small and very large μ values achieve similar fairness (accuracy std $\approx$ 9.1–9.4%), while the intermediate region ($\mu = 0.1$) exhibits significantly worse fairness (std $\approx$ 10.2%). This suggests a phase transition phenomenon: when μ is neither small enough for stable hard clustering nor large enough for effective soft selection, the optimization becomes unstable. The outer weight diversity (measured by coefficient of variation of α_i) is highest at small μ (CV = 0.36), indicating strong differentiation of client importance, and decreases monotonically as μ increases, approaching uniform weighting (CV = 0.009) at $\mu = 1.0$.

Figure 2 provides theoretical verification that $\mu \to 0$ recovers hard clustering behavior (similar to IFCA), while large μ enables soft model selection. The entropy of inner weights w_{ik} decreases from ≈ 1.099 (uniform distribution over K=3 models, corresponding to $\log 3$) at $\mu = 1.0$ to nearly zero at $\mu = 0.001$, while the maximum weight increases from ≈ 0.333 (uniform) to ≈ 0.997 (one-hot). This validates our theoretical prediction that the smoothing parameter interpolates between soft and hard selection regimes.

2.3. Communication Efficiency: Alternative Perspective

The main paper presents communication-computation trade-offs by plotting convergence against total local updates (Figure 4). Here we provide complementary analysis from the communication efficiency perspective.

Convergence vs communication rounds. Figure 4 replots the same convergence data against communication rounds rather than total updates. This perspective reveals the dramatic communication savings: while all configurations perform identical total computation (2000 local updates), LE=16 achieves convergence in merely 125 communication rounds compared to 2000 rounds for LE=1—a 16$\times$ reduction in network overhead. The convergence curves show that configurations with more local epochs not only reduce communication frequency but also exhibit smoother optimization trajectories, with LE=16 demonstrating the steepest and most stable descent in $g^{\text{STCH-Set}}$ values.

Accuracy stability across configurations. As shown in the main paper (Figure 3), despite the 16-fold difference in communication costs between LE=1 and LE=16, mean accuracies remain tightly clustered within 87.8–88.3%, with a maximum deviation of only 0.5 percentage points. This stability validates two key properties of our STCH-Set optimization: (1) robustness to different synchronization frequencies, and (2) insensitivity to the specific (local epochs, communication rounds) decomposition as long as total computation remains constant. The slight accuracy variation (LE=2 achieves 88.3% while LE=16 achieves 87.8%) is practically negligible compared to the substantial communication savings, making LE=8 or LE=16 compelling choices for bandwidth-constrained federated deployments.

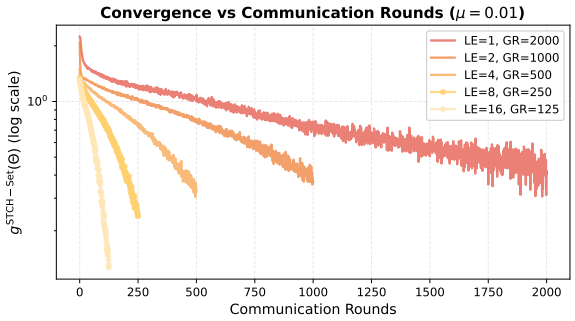

Figure 4. **Convergence vs communication rounds.** Re-plotting the main paper data against communication rounds instead of total updates reveals the communication efficiency perspective: LE=16 converges in 125 rounds while LE=1 requires 2000 rounds, achieving 16$\times$ communication reduction with comparable final $g^{\text{STCH-Set}}$ values.

2.4. Fairness Analysis

To evaluate the fairness of personalization across clients, we compute Jain's Fairness Index on per-client test accuracies. A higher index (maximum $J = 1$) indicates more equitable performance across clients.

Table 3. Jain's Fairness Index (higher is better, $J = 1$ is perfect fairness).

	CIFAR-10			CIFAR-100			Medical	
Method	Dir-10	Dir-20	Pat-10	Dir-10	Dir-20	Pat-10	Kvasir	FedISIC
FedAvg	0.981	0.981	0.982	0.995	0.982	0.977	**0.999**	0.945
FedProx	0.984	0.977	0.976	0.995	0.983	0.982	0.981	0.924
APFL	0.992	0.985	0.996	0.995	0.988	0.996	0.994	0.950
Ditto	0.984	0.982	0.997	0.994	0.987	0.996	0.994	0.951
FedRep	0.990	0.985	0.997	0.995	0.991	0.996	0.995	0.938
IFCA	0.992	0.974	0.985	0.973	0.984	0.993	0.934	0.873
FedFew	**0.992**	**0.990**	0.997	**0.996**	**0.992**	**0.997**	0.996	**0.958**

Table 3 shows that FedFew achieves the highest or near-highest Jain's Fairness Index across most settings. Notably, FedFew consistently outperforms IFCA (the other multi-model baseline) in fairness, demonstrating that the soft model selection via STCH-Set provides more equitable personalization than hard clustering.

2.5. Additional Ablation on K: AG News and Kvasir

To complement the K ablation on CIFAR-10 in the main paper, we conduct additional experiments on AG News ($K \in \{1, \ldots, 5\}$) and Kvasir ($K \in \{1, \ldots, 5\}$) to investigate whether the optimal K depends on dataset characteristics.

Table 4. Test accuracy (%) vs. number of server models K on AG News and Kvasir datasets.

	AG News			Kvasir		
K	Min	Mean	Max	Min	Mean	Max
1	84.5	96.4	100.0	79.5	91.6	99.5
2	73.8	95.7	100.0	82.4	92.2	99.8
3	83.9	96.0	100.0	**83.9**	**92.9**	100.0
4	78.0	95.7	100.0	82.9	92.6	100.0
5	**86.3**	**96.2**	**100.0**	81.2	92.5	**100.0**

Table 4 confirms that the optimal K depends on dataset characteristics: Kvasir peaks at $K = 3$, while AG News peaks at $K = 5$. This aligns with Theorem 3.1, which predicts that the optimal K balances the Pareto coverage gap (favoring larger K) against statistical error (favoring smaller K). Datasets with more heterogeneous client distributions benefit from a larger K to adequately cover the Pareto front.

2.6. Cost Analysis

We analyze the computational and communication overhead of maintaining K server models compared to single-model approaches.

- **Training**: Clients compute gradients for K models, resulting in a $K\times$ increase in local computation per round. For $K = 3$, this represents a modest $3\times$ overhead.
- **Inference**: Each client uses only its best-matching model (determined by w_{ik}), so inference cost is identical to single-model methods.
- **Communication**: The overhead scales linearly with K (a constant factor), not with M. For $K = 3$ and $M = 20$, FedFew achieves $> 6\times$ reduction in server-side model storage compared to maintaining M personalized models.
- **Server storage**: The server maintains K models instead of M, yielding an M/K storage reduction factor.

Given that $K \ll M$ and inference cost is unchanged, the trade-off is favorable: a modest increase in training computation yields significant personalization improvements with reduced server-side storage.

References

[1] Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*. Athena Scientific, 2003. 3

[2] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 3

[3] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. 2