PSFL: Parallel-Sequential Federated Learning with Convergence Guarantees

IEEE INFOCOM 2025
May 20, 2025

Jinrui Zhou[1] , Yu Zhao[1], Yin Xu[1], Mingjun Xiao[1], Jie Wu[2], Sheng Zhang[3]

1 School of Computer Science and Technology & Suzhou Institute for Advanced Study, University of Science and Technology of China
2 Department of Computer and Information Sciences, Temple University, USA
3 Department of Computer Science and Technology, Nanjing University, China

CONTENTS

CONTENTS

Introduction

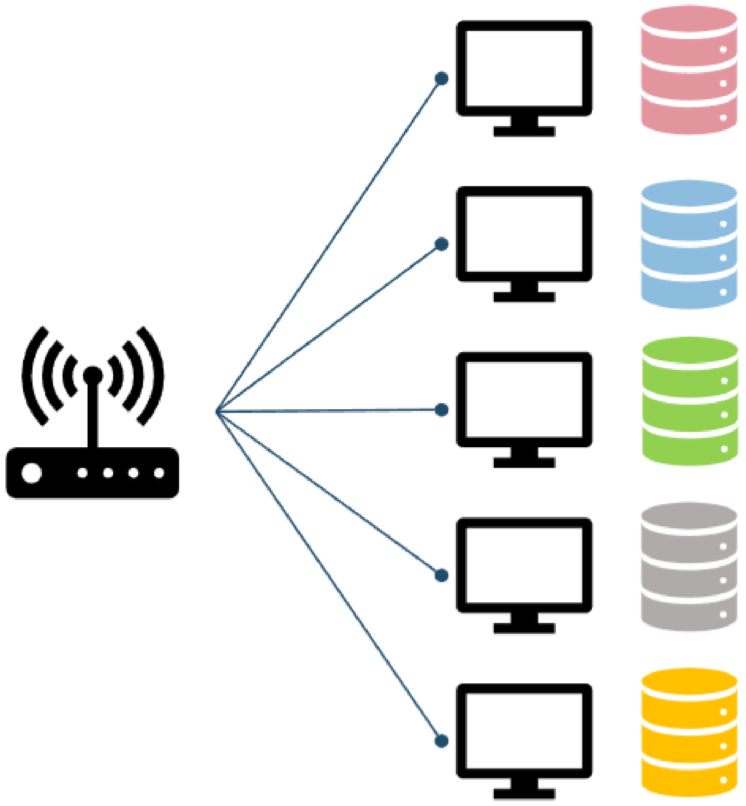
Federated Learning (FL)

Concept of FL

A novel **distributed learning paradigm** which can coordinate multiple clients to jointly train a machine learning model by using their local data samples.

Procedure of Parallel FL (PFL)

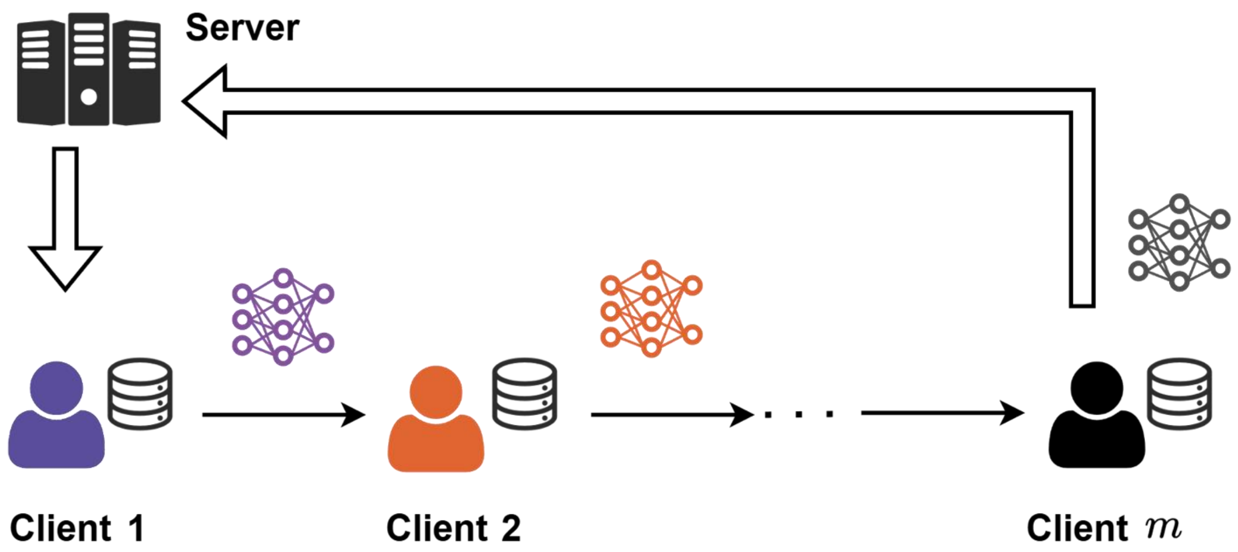
- ✓ Data stay locally on clients
- ✓ Clients train models locally **in parallel**
- ✓ Clients send models or updates to server
- ✓ Server aggregate local models



Introduction

Sequential FL (SFL)

✓ Clients send models or updates to **next client**

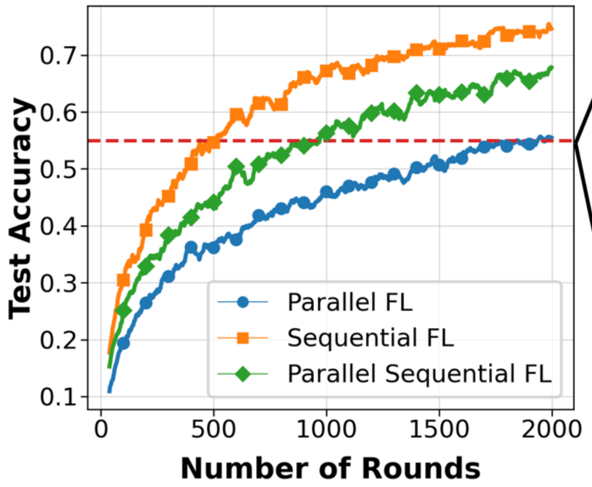


Sequential Federated Learning Training Process.

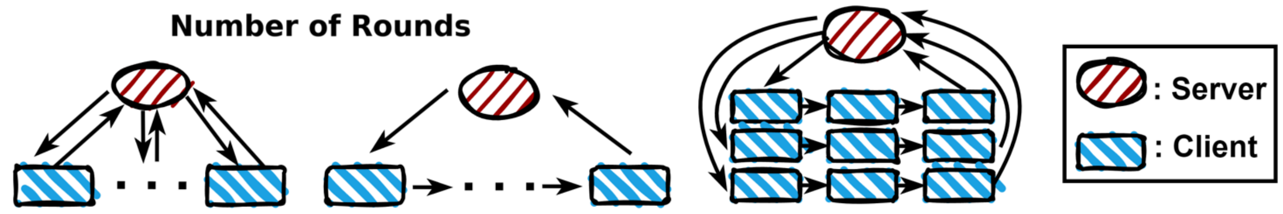
Introduction

Motivations

- ✓ PFL significantly reduces the training time per round, but it typically requires many more rounds to reach the target accuracy.
- ✓ SFL achieves faster accuracy improvement in fewer rounds, but each round takes much longer due to sequential updates.



	(a) PFL	(b) SFL	(c) PSFL
Training rounds	2000	500	900
Training time per round	5.97	26.09	9.58
Total time	11940	13045	8622



(a). Parallel FL (PFL) (b). Sequential FL (SFL) (c). Parallel Sequential FL (PSFL)

CONTENTS

System, Modeling, and Problem

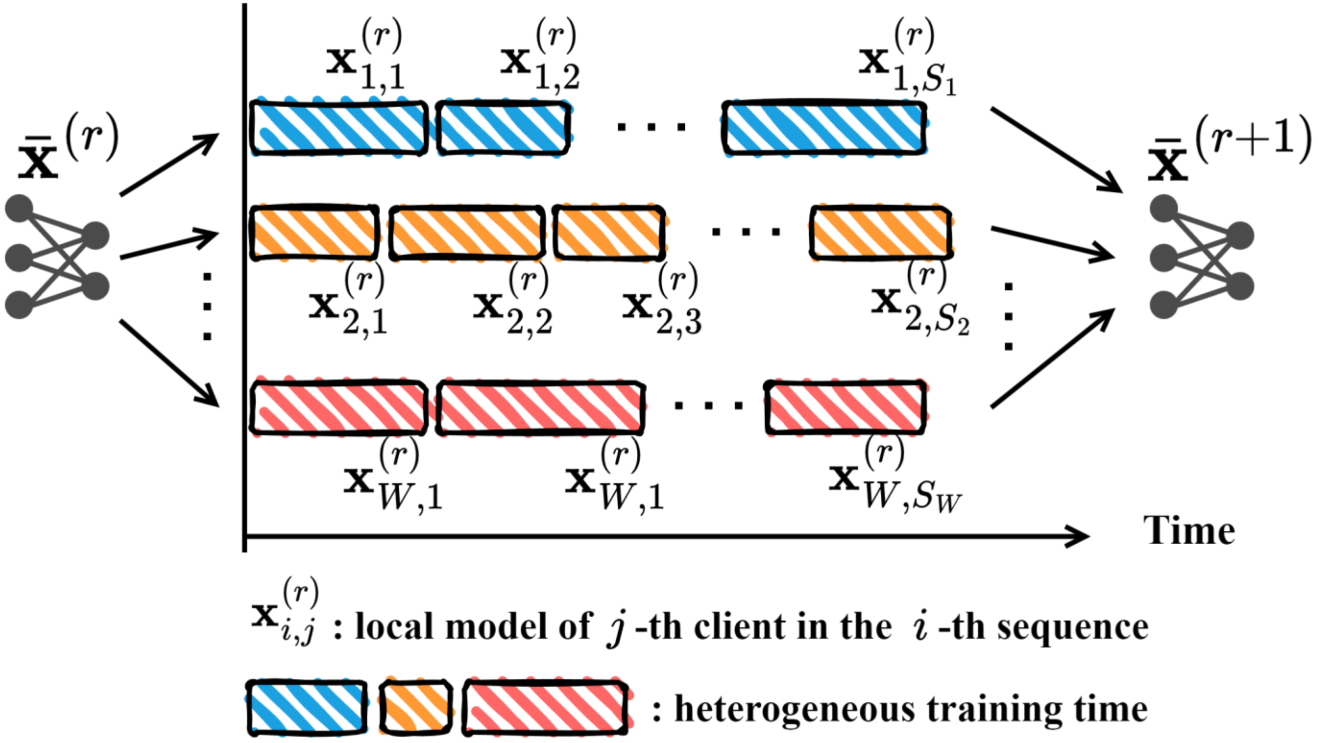
Parallel-Sequential FL (PSFL)

Training Structure:

- ✓ Denoted by $\mathcal{A} \triangleq (\mathcal{W}, \{S_w\}_{w\in\mathcal{W}})$
- ✓ $\mathcal{W}$ represents a set of sequences
- ✓ $W = |\mathcal{W}|$: parallel width
- ✓ S_w : sequence leagth

Client sampling strategy:

$$\Pi_{\mathcal{A}}^{(r)} = \{\pi_1^w, \pi_2^w, \ldots, \pi_{S_w}^w\}_{w\in\mathcal{W}}$$



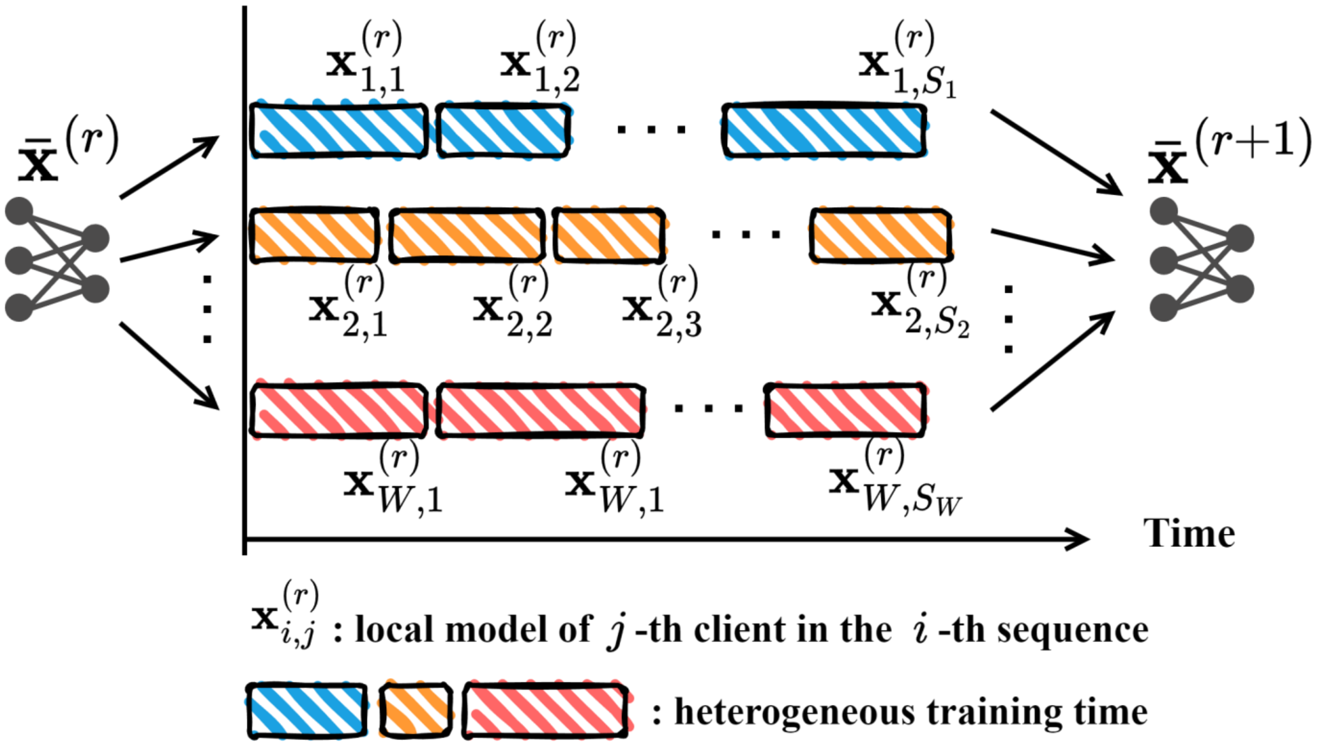
System, Modeling, and Problem

Parallel-Sequential FL (PSFL)

- ✓ **Statistical heterogeneity:** the training data are distributed in an unbalanced and non-iid fashion among clients
- ✓ **System heterogeneity**: clients exhibit heterogeneous capabilities in both computing and communication.

Training time: $t_n^{(r)}$

$$T_{total}^{(r)} = \max_{w \in \mathcal{W}} \sum_{m=1}^{S_w} t_{\pi_m^w}^{(r)}$$



System, Modeling, and Problem

Problem formulation

Our goal is to determine the optimal client sampling strategy based on a training structure, so as to minimize the expected total training time, while ensuring that the expected global loss convergences to the optimal value with an ϵ precision.

$$\textbf{P1:} \quad \min_{\Pi_{\mathcal{A}}} \quad \mathbb{E}[\sum_{r=0}^{R-1} T_{total}^{(r)}],$$

$$s.t. \quad \mathbb{E}[F(\bar{\mathbf{x}}^{(R)})] - F(\mathbf{x}^*) \le \epsilon,$$

$$\sum_{w \in \mathcal{W}} S_w \le \widetilde{N}.$$

← Expected total training time.

← Convergence guarantee.

← The number of selected clients in each round is bounded.

CONTENTS

Theorem, Optimization, and Algorithm

Convergence Analysis

Theorem 1. *Let Assumptions 1 to 4 hold, and the values of* L, σ^2, $\mathcal{A}$ *are given. If the client sampling strategy* $\Pi_{\mathcal{A}}$ *is unbiased in the PSFL framework, and the learning rate satisfies* $\eta \leq \frac{c_0}{LS}$, *where* $0 < c_0 < \frac{1}{5}$ *is a constant, then the weighted average of the global parameters* $\widetilde{\mathbf{x}} = \frac{1}{R+1}\sum_{r=0}^{R} \bar{\mathbf{x}}^{(r)}$ *satisfies:*

$$\mathbb{E}[F(\widetilde{\mathbf{x}}) - F(\mathbf{x}^*)] \leq \frac{r_0}{b\tilde{\eta}R} + \frac{\tilde{\eta}}{b}(\alpha W + \beta), \tag{8}$$

where $b = \frac{16c_0^3 - 28c_0^2 - 24c_0 + 6}{3(1-2c_0^2)}$, $\tilde{\eta} = \frac{\eta N_0}{W}$, $r_0 = \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2$, $\alpha = \frac{4c_0(1+2c_0)}{1-2c_0^2}\frac{(\sigma^2+B)}{N_0} + \frac{4B}{N_0}$, $\beta = \frac{4\sigma^2}{N_0}$, *and* $N_0 = \sum_{w\in\mathcal{W}} S_w$.

Theorem, Optimization, and Algorithm

Convergence Analysis

Corollary 1. *By choosing an appropriate learning rate* $\tilde{\eta} = \min\{\sqrt{\frac{r_0}{R(\alpha W+\beta)}}, \frac{c_0 N_0}{LSW}\}$, *we can obtain the convergence bound:*

$$\mathbb{E}[F(\widetilde{\mathbf{x}}) - F(\mathbf{x}^*)] \leq \mathcal{O}\left(\frac{1}{R}\frac{r_0 LSW}{bc_0 N_0} + \frac{1}{b}\sqrt{\frac{r_0(\alpha W + \beta)}{R}}\right), \quad (9)$$

where $b = \frac{16c_0^3 - 28c_0^2 - 24c_0 + 6}{3(1-2c_0^2)}$, $r_0 = \|\bar{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2$, $N_0 = \sum_{w \in \mathcal{W}} S_w$, $\alpha = \frac{4c_0(1+2c_0)}{1-2c_0^2}\frac{(\sigma^2 + B)}{N_0} + \frac{4B}{N_0}$, *and* $\beta = \frac{4\sigma^2}{N_0}$.

Theorem, Optimization, and Algorithm

Convergence Bound

✓ $S = \max_{w \in \mathcal{W}}\{S_w\}$, longest sequence length

$$\mathbb{E}[F(\widetilde{\mathbf{x}}) - F(\mathbf{x}^*)] \le \mathcal{O}\left(\frac{1}{R}\frac{r_0 L S W}{b c_0 N_0} + \frac{1}{b}\sqrt{\frac{r_0(\alpha W + \beta)}{R}}\right)$$

✓ The number of training rounds.

✓ $N_0 = \sum_{w \in \mathcal{W}} S_w$

✓ The parallel width.

Theorem, Optimization, and Algorithm

Bound for the Expected Training Time

✓ Subgaussian training time

Theorem 2. *Let Assumption 5 hold, and assume that the client sampling strategy* $\Pi_{\mathcal{A}}$ *is unbiased, then the expected total training time is bounded as follows:*

$$\mathbb{E}[\sum_{r=0}^{R-1} T_{total}^{(r)}] \geq R\frac{N_0}{W}\frac{1}{N}\sum_{n=1}^{N} t_n, \quad (10)$$

$$\mathbb{E}[\sum_{r=0}^{R-1} T_{total}^{(r)}] \leq R[S\frac{1}{N}\sum_{n=1}^{N} t_n + \sqrt{2\kappa^2 S \log W}], \quad (11)$$

where $N_0 = \sum_{w\in\mathcal{W}} S_w$ *and* κ *is a constant.*

Theorem, Optimization, and Algorithm

Problem Transformation

$$\textbf{P2:}\quad \min_{S,W} \quad R\Big(S\frac{1}{N}\sum_{n=1}^{N} t_n + \sqrt{2\kappa^2 S \log W}\Big), \tag{15}$$

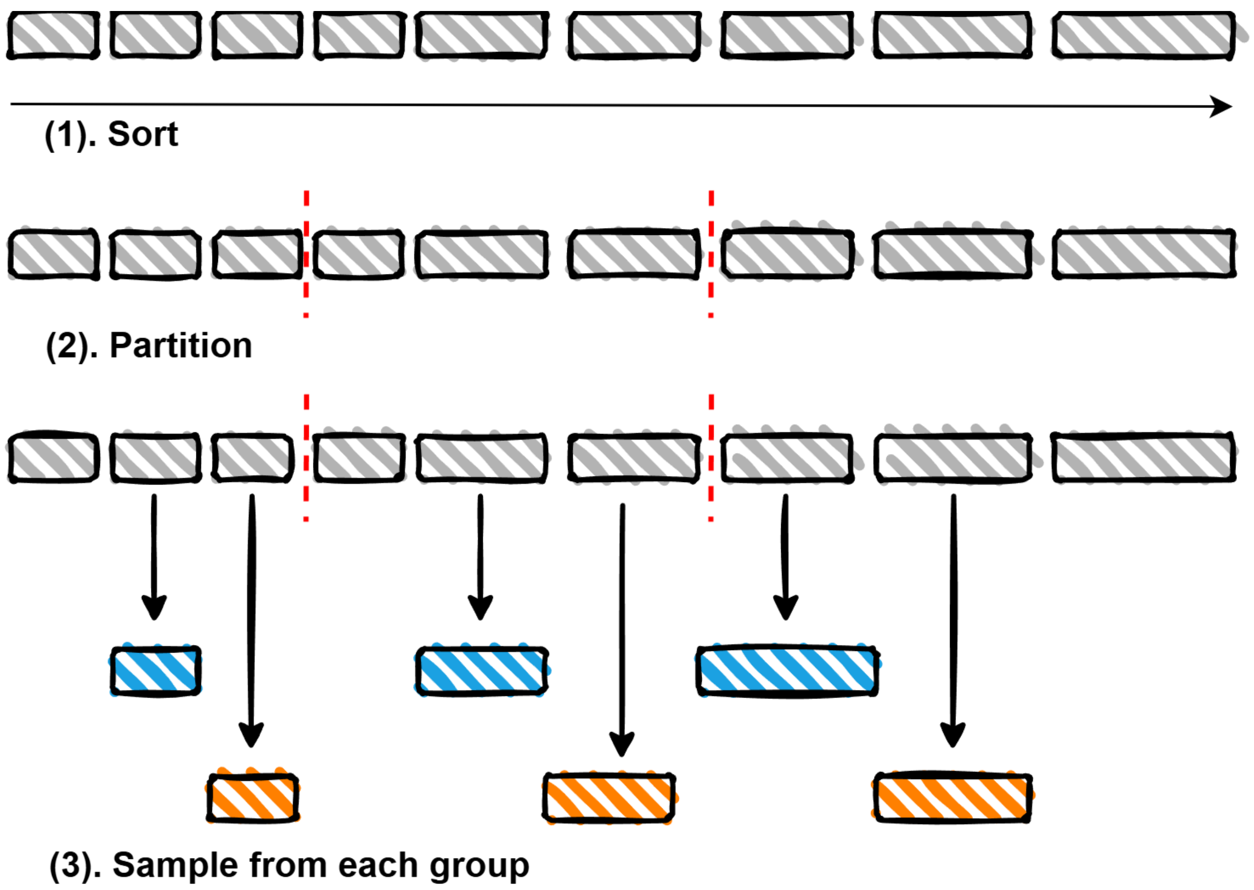
$$s.t. \quad \frac{1}{R}(\alpha W + \beta) \leq \epsilon', N_0 \leq \widetilde{N}, \tag{16}$$

$$\alpha = \frac{4c_0(1+2c_0)(\sigma^2+B)}{(1-2c_0^2)N_0} + \frac{4B}{N_0}, \beta = \frac{4\sigma^2}{N_0}, \tag{17}$$

Theorem, Optimization, and Algorithm

Client Sampling Strategy

- ✓ Unbiased
- ✓ Any unbiased sampling strategy cannot reduce the training time of a sequence.
- ✓ Minimize the variance between sequences



Theorem, Optimization, and Algorithm

Algorithm 1: Parallel-Sequential Federated Learning

input : number of total training rounds R.
output: aggregated global model $\bar{\mathbf{x}}^{(R)}$.
1 **//Warm-up Phase:**
2 Initialize: the global model $\mathbf{x}^0$;
3 **for** *training round* $k = 0, 1, \ldots, K-1$ **do**
4   **for** *client* $n = 1, \ldots, N$ *in parallel* **do**
5   Initialize: $\mathbf{x}_n^k = \mathbf{x}^k$;
6   Local update: $\mathbf{x}_n^{k+1} = \mathbf{x}_n^k - \eta \nabla F_n(\mathbf{x}^k)$;
7   Estimate $\hat{\sigma}_{n,k}^2 = \mathbb{E}[\|\nabla f(\mathbf{x}^k, \xi_n) - \nabla F_n(\mathbf{x}^k)\|^2]$;
8   Global aggregation: $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \frac{1}{N}\sum_{n=1}^{N} \nabla F_n(\mathbf{x}^k)$;
9   Estimate $\hat{B}_k = \mathbb{E}[\|\nabla F_n(\mathbf{x}^k) - \nabla F(\mathbf{x}^k)\|^2]$;
10 Estimate $\hat{\sigma}^2$, $\hat{B}$, $\hat{\bar{t}}$, and $\hat{\kappa}^2$ based on Eq. (24) $\sim$ Eq. (26);
11 Solve Eq. (21) to get optimal sequence length S and optimal parallel width W;
12 **//Training Phase:**
13 Initialize: $\bar{\mathbf{x}}^{(0)}$ and the estimates of training time $\hat{t}_n$;
14 **for** *training round* $r = 0, 1, \ldots, R-1$ **do**
15   Sort the clients according to estimate $\hat{t}_n$;
16   Sample clients $\{\pi_1^w, \pi_2^w, \ldots, \pi_S^w\}_{w \in \mathcal{W}}$ based on time-based partitioning and sampling strategy;
17   **for** *sequence* $w = 1, \ldots, W$ *in parallel* **do**
18   Initialize: $\mathbf{x}_{w,0}^{(r)} = \bar{\mathbf{x}}^{(r)}$;
19   **for** *client* $m = 1, \ldots, S$ *in sequence* **do**
20   Local update: $\mathbf{x}_{w,m}^{(r)} = \mathbf{x}_{w,m-1}^{(r)} - \eta \mathbf{g}_{\pi_m^w}^{(r)}$;
21   Update the estimate of training time $\hat{t}_{\pi_m^w}$;
22   Global aggregation: $\bar{\mathbf{x}}^{(r+1)} = \frac{1}{W}\sum_{w=1}^{W} \mathbf{x}_{w,S}^{(r)}$.

Lines 1- 11:
Warm-up Phase: Estimate some parameters and solve the optimization problem to get optimal training structure

Lines 12- 22:
Training Phase: Based on the optimal training structure, sample clients and train the models.

CONTENTS

Experimental Evaluation

Evaluation Setup

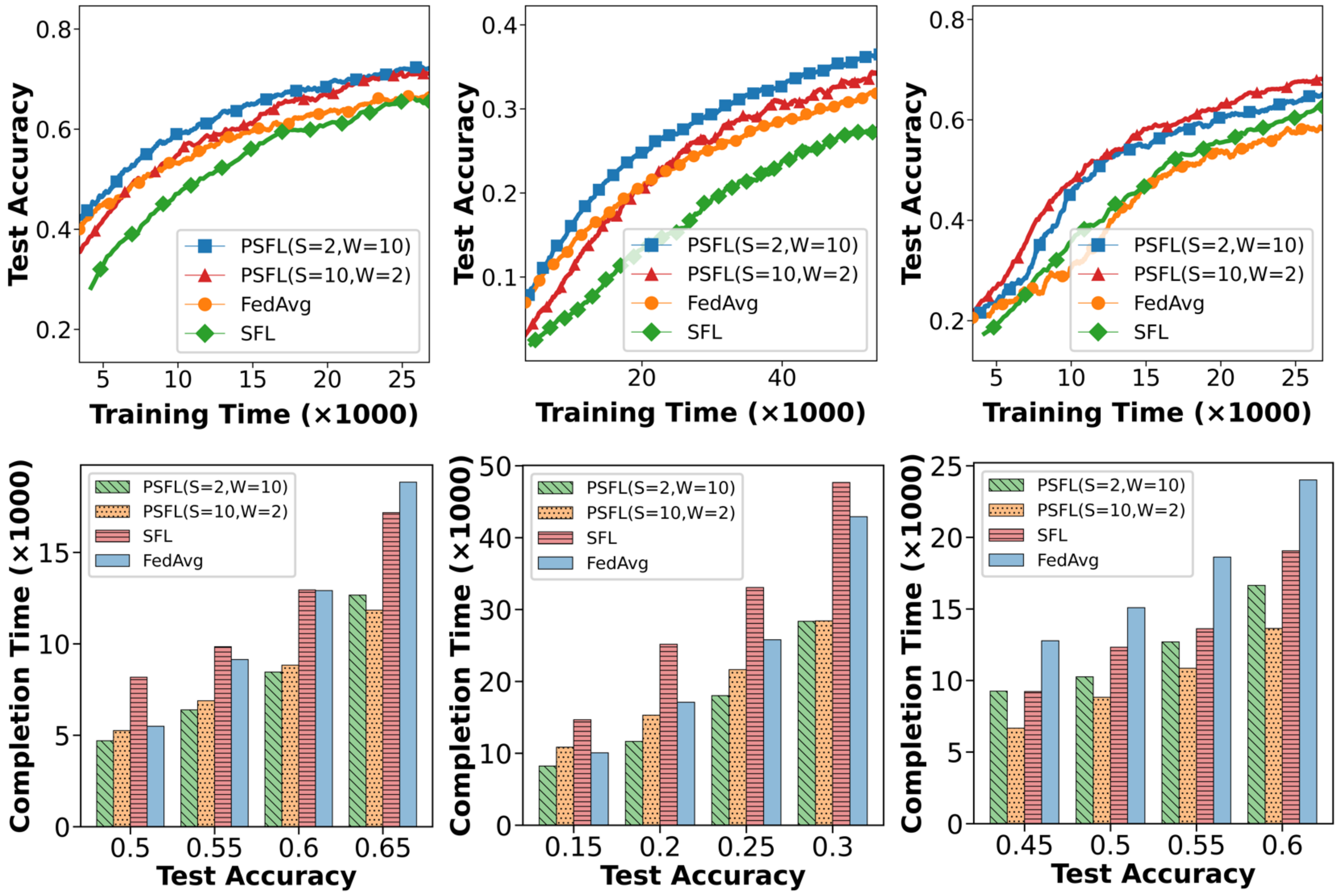
Parameter Name	Range
number of clients N	500
number of selected clients $\boldsymbol{N_0}$	20, 50, 100, 200
Heterogeneous data	ExDir(2,10), ExDir(1,10), ExDir(2,5), Dir(0.2)
Heterogeneous system, $\boldsymbol{t_n}$	{0.5, 1, 2, 4, 5}, Gaussian

- ✓ **Dataset:**
  CIFAR-10, CIFAR-100, HAM10000
- ✓ **Baselines:**
  PFL (FedAvg), SFL
- ✓ **Metrics:**
  test accuracy, total training time

Experimental Evaluation

Comparing to baselines

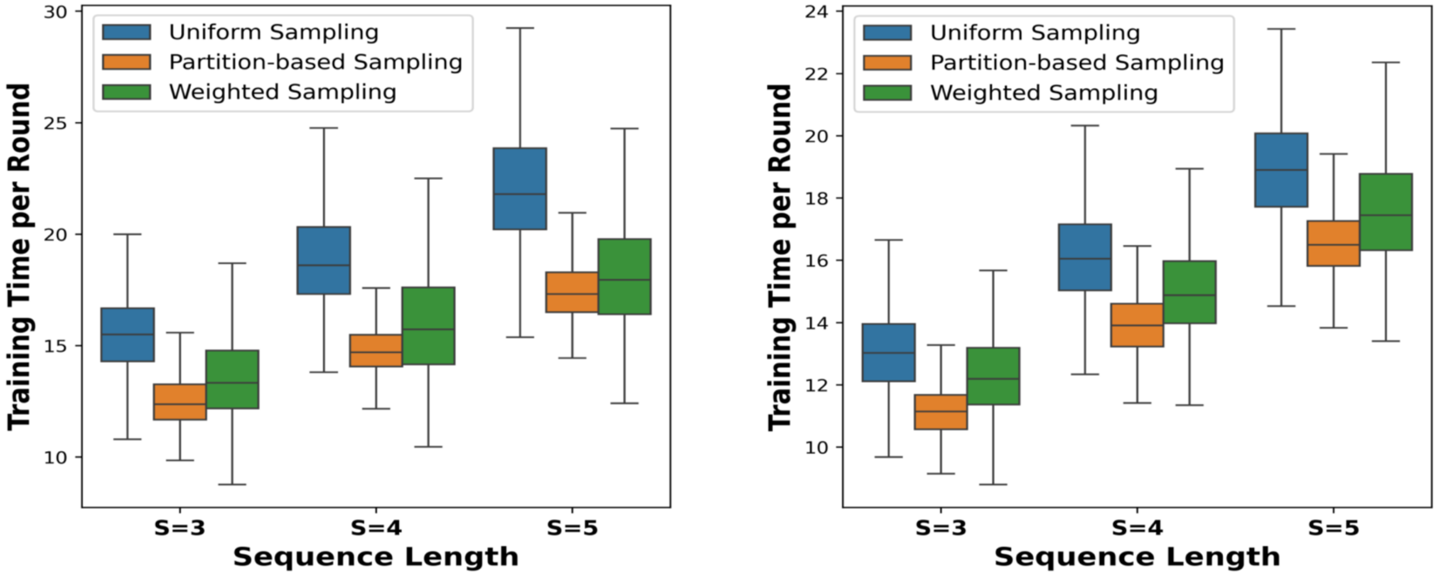
- ✓ $N_0 = 20$
- ✓ ExDir(2,10)
- ✓ PSFL achieves better convergence performance under the same training time
- ✓ PSFL achieves the same target test accuracy with significantly less training time



Experimental Evaluation

The efficiency of sampling strategy

- ✓ Compared strategies:
  - uniform sampling
  - weighted sampling
- ✓ Sequence lenngth: S = 3, 4, 5
- ✓ time setting:
  - ✓ discrete distribution
  - ✓ gaussian distribution

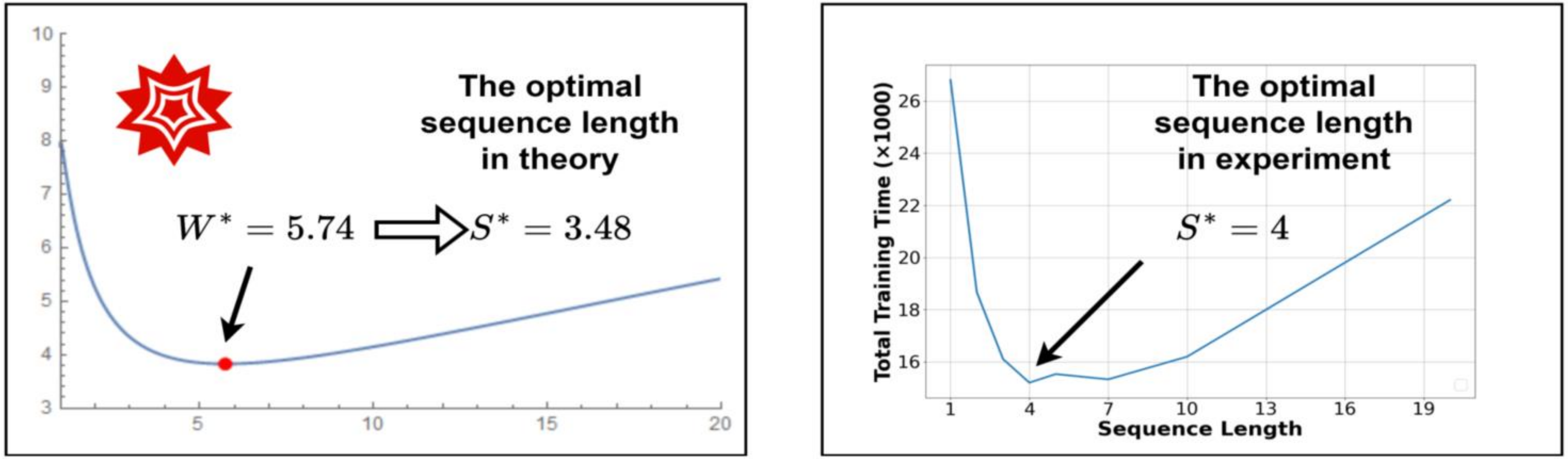


(a) Discrete Distribution (b) Gaussian Distribution

Fig. 6. Comparison of sampling strategies under different distributions.

Experimental Evaluation

The efficiency of optimal structure



(a) Theoretical Result (b) Experimental Result

Fig. 7. Comparison of theoretical and experimental optimal sequence length.

CONTENTS

Conclusion

- **Propose a novel hybrid PSFL framework by integrating the parallel and sequential training modes together.**
- **Provide a theoretical analysis to derive the upper bounds of the model convergence and the expected total training time for the PSFL framework.**
  - Solve the optimization problem and get the optimal training structure.
- **The performance is demonstrated on extensive simulations.**

Thank you for your attention!

Jinrui Zhou, Yu Zhao, Yin Xu, Mingjun Xiao, Jie Wu, Sheng Zhang

zzkevin@mail.ustc.edu.cn