FedQS: Optimizing Gradient and Model Aggregation for Semi-Asynchronous Federated Learning

Yunbo Li[1], Jiaping Gui[1], Zhihang Deng[1], Fanchao Meng[1], Yue Wu[1]
[1]Shanghai Jiao Tong University

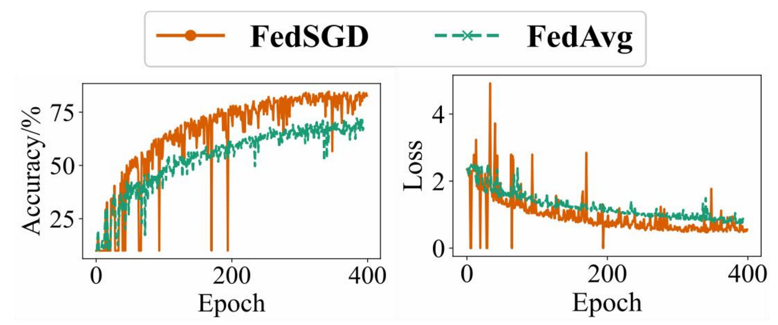
Background

Semi-Asynchronous Federated Learning:

- SAFL sits between synchronous and fully asynchronous FL, balancing training delay, stability, and efficiency.
- Two dominant aggregation modes:
  - **Gradient aggregation** (FedSGD): fast convergence but oscillatory.
  - **Model aggregation** (FedAvg): more stable but slower and sub-optimal.

Challenge:

- **Lack of theoretical guidance**:
  - existing work mostly empirical
- **Inherent Aggregation Disparity:**
  - Gradients: continuous in loss space.
  - Models: stale and interrupted.
- **Server- or Client-Centric Limitations:**
  - Server-centric: one-sided aggregation.
  - Client-centric: insufficient global information



Q: How to jointly stabilize gradient aggregation and accelerate model aggregation under SAFL while guaranteeing convergence?

Analysis

Key Factors Affecting Aggregation Gap

- **Staleness (Factor 1):**
  - Gradient aggregation:
    - Continuous but unstable updates.
  - Model aggregation:
    - Reset optimization path.
- **Data Heterogeneity (Factor 2):**
  - Frequent clients dominate, causing bias & overfitting.

Activated Factors		Average Best Acc. (%)		Gap (%)
Factor 1	Factor 2	Gradient Aggregation	Model Aggregation	
○	○	90.93	91.05	0.12
●	○	90.73	90.51	0.22
○	●	86.79	87.29	0.50
●	●	82.63	71.11	**11.52**

Theoretical Insights

- FedQS under **gradient aggregation**:

$$\mathbb{E}[F(w_g^t)] - F^* \leq L\mathcal{V}^t\mathbb{E}[||w_g^0 - w^*||^2] + \mathcal{U} + \mathcal{W}$$

FedQS-SGD adds momentum to improve convergence stability.

- FedQS under **model aggregation**:

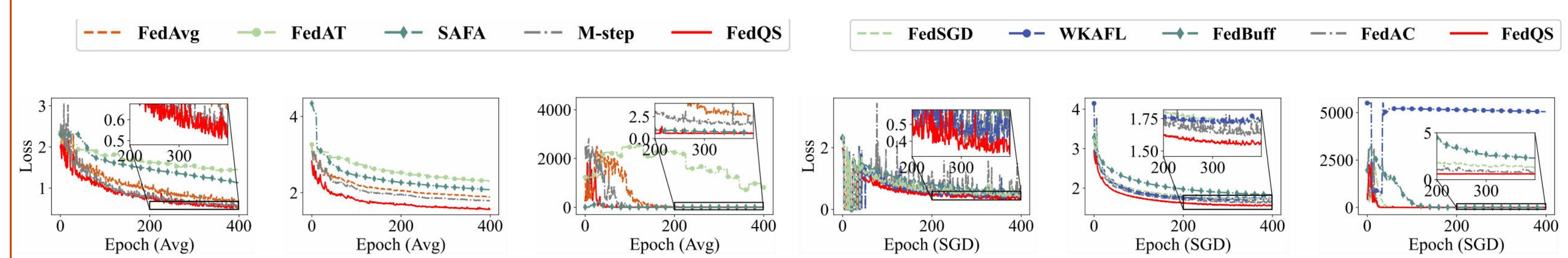
$$\mathbb{E}[F(w_g^t)] - F^* \leq (3LpK^2 + L)\mathcal{V}^t\mathbb{E}[||w_g^0 - w^*||^2] + \mathcal{U} + \mathcal{W}$$

FedQS-Avg uses feedback mechanism to address suboptimal convergence.

- **Term $\mathcal{V}^t \in (0, 1)$**: exponentially decaying and guaranteeing convergence.
- **Term $\mathcal{U} = \mathcal{O}(\delta^2)$**: highlighting the impact of data heterogeneity.
- **Term $\mathcal{W} = \mathcal{O}(G_c^2)$**: capturing gradient norm effects.

Evaluation

Overall Performance:



(a) CV Task (b) NLP Task (c) RWD Task (d) CV Task (e) NLP Task (f) RWD Task

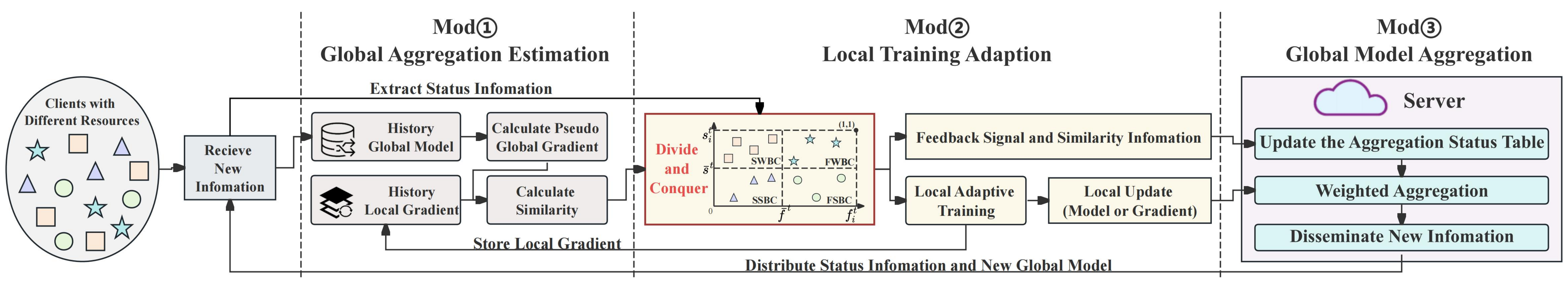
FedQS achieves the lowest loss, fastest convergence, and the most stable training under both gradient- and model-aggregation strategies.

Scalability & Robustness:

Scenario 1: Dynamic Resource Scale.
Scenario 2: Unstable Resource per Client.
Scenario 3: Client Dropout.

Dynamic Scenario	Algorithm	x=0.1			x=0.5			x=1		
		Accuracy (%)	Conv. Speed	Runtime (s)	Accuracy (%)	Conv. Speed	Runtime (s)	Accuracy (%)	Conv. Speed	Runtime (s)
Scenario 1	FedAvg	52.77	322	46488	77.29	288	47732	76.79	227	48003
	FedQS-Avg	**66.47**	**264**	53986	**82.90**	**235**	54575	**83.22**	**201**	55529
	FedSGD	68.29	244	47443	85.66	211	47728	87.86	147	48280
	FedQS-SGD	**76.17**	**206**	54095	**88.12**	**177**	55268	**88.80**	**115**	55257
Scenario 2	FedAvg	53.12	310	47397	76.74	294	47156	79.90	286	47788
	FedQS-Avg	**64.52**	**268**	55535	**83.53**	**241**	55816	**84.35**	**252**	56253
	FedSGD	65.13	268	47613	85.37	207	47096	88.63	159	47634
	FedQS-SGD	**69.65**	**223**	55037	**87.96**	**185**	55201	**89.14**	**133**	55808
Scenario 3	FedAvg	44.39	343	58993	68.44	332	59362	72.47	296	59447
	FedQS-Avg	**53.22**	**297**	62623	**79.33**	**272**	63308	**81.08**	**253**	63776
	FedSGD	58.40	292	58345	81.73	288	59487	84.43	262	59732
	FedQS-SGD	**60.08**	**263**	62870	**83.39**	**245**	63992	**86.18**	**241**	64703

Methodology



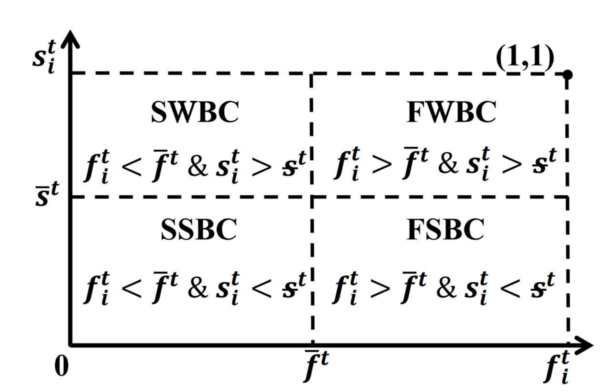
Overview of FedQS:

- **Mod ① (Global Aggregation Estimation):**
  - Compute pseudo global gradient: $L_g(w_g^t) = w_g^t - w_g^{t-1}$.
- **Mod ② (Local Training Adaption):**
  - Divide clients into four types based on (f_i^t, s_i^t).
- **Mod ③ (Global Model Aggregation):**
  - Aggregate after receiving K updates with adaptive weight.

$$p_i \propto \frac{n_i}{n} \; or \; \frac{exp(\phi - F)}{2^{\phi-F}} \cdot \frac{(1+\mathcal{G})^2}{K}$$

Divide and Conquer Strategy:

- **Fast-and-Weakly-Biased:**
  - Frequent, mild drift -> Decrease LR + momentum
- **Fast-but-Strongly-Biased:**
  - Frequent, strong drift -> Fixed LR + feedback weighting
- **Straggling-but-Weakly-Biased:**
  - Rare, mild drift -> Increase LR + momentum
- **Straggling-and-Strongly-Biased:**
  - Rare, strong drift -> Adaptively switch: momentum vs. feedback weighting



Key Takeaways

- Bridges the gap between gradient and model aggregation in Semi-Asynchronous Federated Learning through a theoretically governed design.
- Ultilize Divide-and-Conquer strategy to enhances training stability, accuracy, and speed under resource and data heterogeneity.
- Demonstrates strong scalability and robustness under fluctuating resources and client dropouts.

More Details of the Paper:

Contact:

Yunbo Li: li-yun-bo@sjtu.edu.cn
Jiaping Gui: jgui@sjtu.edu.cn
Zhihang Deng: dzh1227@sjtu.edu.cn
Fanchao Meng: mactavishmeng@sjtu.edu.cn
Yue Wu: wuyue@sjtu.edu.cn