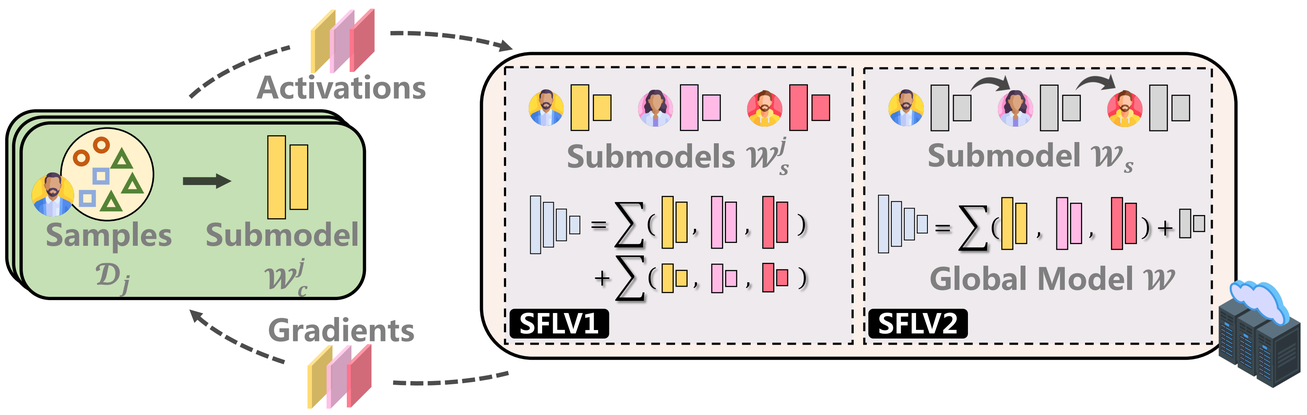
SMoFi: Step-wise Momentum Fusion for Split Federated Learning on Heterogeneous Data

Mingkun Yang, Ran Zhu, Qing Wang, Jie Yang
{m.yang-3, r.zhu-1, qing.wang, j.yang-3}@tudelft.nl

MOTIVATION



Split FL Partitioning model as $\mathcal{W} = [\mathcal{W}_c, \mathcal{W}_s]$, the server

- **Collaboratively** updates and aggregates with clients;
- **Directly** controls the learning pace of submodels $\mathcal{W}_s$.

😫 Inferior global $\mathcal{W}$ under data heterogeneity across $\mathcal{D}_j$

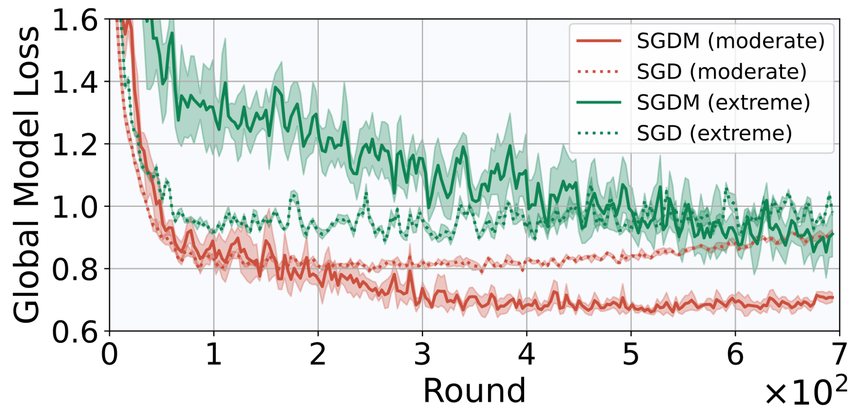
Comparisons	SFLV1	SFLV2	SMoFi
Server-side Updating	Parallel	Sequential	Parallel
Server-side Aggregation	$\bar{\tau} \in [1, N]$	No Aggregation	$\bar{\tau} = N$
Optimizer Resetting	$\bar{\tau}$-dependent	Each Step	Each Step

Can we impose constraints on model training by inherent client-server interaction in Split FL without introducing additional overheads or privacy risk?

Contributions Our SMoFi achieves

- **Plug-and-play:** integration with Split FL frameworks;
- **Transparency:** requiring no client-side modifications;
  😄 Minimal overhead ⛔ Same privacy guarantee
- **Performance Gains** in accuracy and convergence.

METHOD

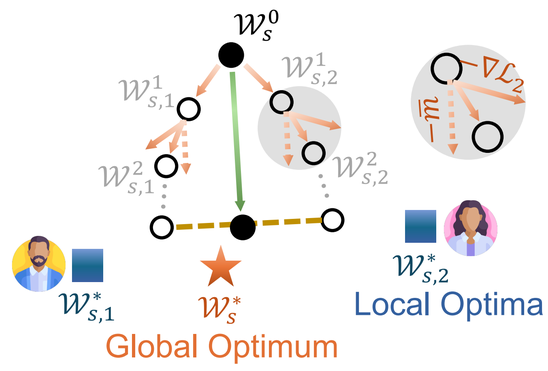


Observations SGD with momentum in (Split) FL

- **Facilitates** convergence toward local optima;
- **Exacerbates** divergence across local updates;

⬆ Higher accuracy (long-run) ⬇ Slower convergence

- More significant under higher data heterogeneity.



One-step SGDM

$$\mathcal{W}_{s,j}^{\tau+1} = \mathcal{W}_{s,j}^{\tau} - \eta m_{s,j}^{\tau+1}$$

(SFLV1) $m_{s,j}^{\tau+1} = \beta m_{s,j}^{\tau} + \nabla\mathcal{L}$

(SMoFi) $m_{s,j}^{\tau+1} = \beta \bar{m}^{\tau} + \nabla\mathcal{L}$

Alignment At each training step, the server performs

- **Fusion** of momentum buffers across optimizers:

$$\bar{m}^{\tau} = \frac{1}{|\mathcal{J}|}\left(\sum_{\mathcal{J}^{\tau}} m_{s,j}^{\tau+1} + \sum_{\mathcal{H}} S_{\alpha}^{\tau} m_{s,j}^{\tilde{\tau}}\right), \text{where } \mathcal{J} = \mathcal{J}^{\tau} \cup \mathcal{H};$$

- **Recording** of historical buffers into $\mathcal{H}$;
- **Staleness-aware** weighted averaging by factor S_{α}^{τ}:

$$S_{\alpha}^{\tau} = (\tau - \tilde{\tau} + 1)^{\alpha}, \alpha < 0.$$

EVALUATION

Setup Three image tasks CIFAR-10/-100 (ResNet-18) and Tiny-ImageNet (ResNet-34) are under a Dirichlet distribution with concentration parameter 0.2; a text task Shakespeare (stacked Transformers) is inherently non-IID data. We report Top-1 accuracy (**Acc.**) and Round-to-Accuracy (**R**) performance.

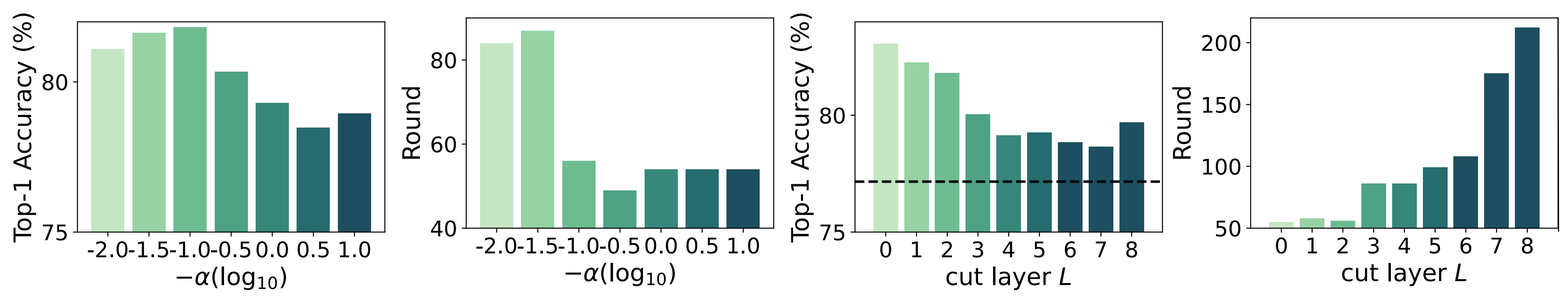
Setup	CIFAR-10		CIFAR-100		Tiny-ImageNet		Shakespeare	
Methods	Acc. (%)	R	Acc. (%)	R	Acc. (%)	R	Acc. (%)	R
FedAvg	77.16±0.11	258	48.10±0.36	183	33.43±0.12	161	46.08±0.53	170
FedAvgM	79.19±0.09	190	50.28±0.26	126	33.58±0.34	57	49.13±0.29	**62**
SlowMo	76.54±0.06	177	50.96±0.23	125	33.82±0.29	44	47.62±0.74	85
FedNAG	78.24±0.43	170	48.30±1.06	198	30.94±0.44	335	42.56±2.59	210
SFLV1 ($\bar{\tau} = 1$)	68.10±0.57	>1000	38.43±0.06	>600	21.81±0.98	>400	44.07±0.64	240
SFLV1 ($\bar{\tau} = E$)	77.84±0.17	69	46.68±0.21	**40**	35.47±0.12	44	48.69±0.63	162
SFLV2	79.42±0.04	278	53.64±0.51	143	34.72±0.95	310	45.84±1.39	99
MergeSFL	79.47±0.09	76	50.16±0.20	53	34.74±0.55	118	42.35±1.01	250
SMoFi	**81.82**±0.61	**56**	**53.83**±0.79	64	**39.73**±0.05	**16**	**51.83**±0.21	74

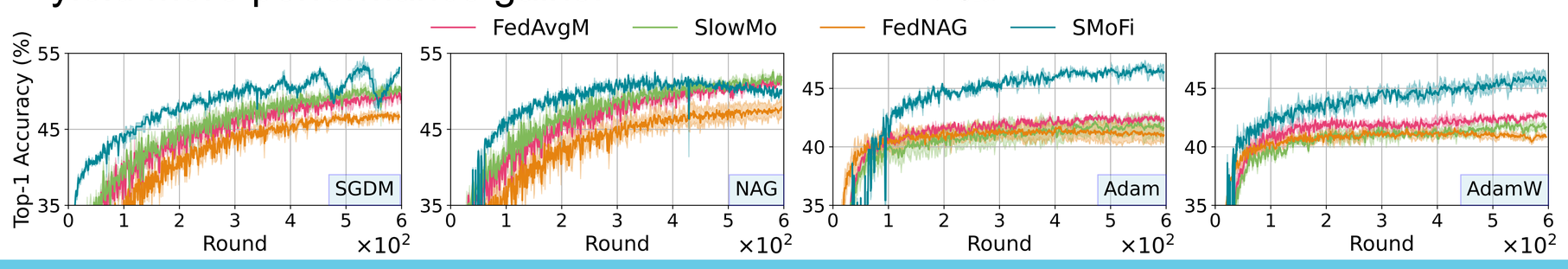
Effectiveness SMoFi consistently delivers

- **Accuracy**: Superior to momentum-based and Split FL counterparts;
- **Convergence**: Up to 10.25× faster than baseline FedAvg.

Sensitivity Two key hyper-parameters

- **Staleness α**: Trade-off between faster convergence and higher accuracy;
- **Cut Layer L**: Shallower model splits yield more performance gains.





Robustness Performance consistently benefits from momentum fusion across four widely used optimizers.