NEURAL INFORMATION PROCESSING SYSTEMS

Differentially Private Federated Low Rank Adaptation Beyond Fixed-Matrix

Ming Wen · Jiaqi Zhu · Yuedong Xu · Yipeng Zhou · DingDing Han

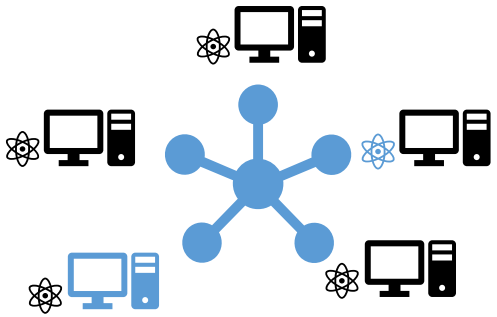
https://arxiv.org/abs/2507.09990

https://github.com/FLEECERmw/PrivacyFedLLM

Challenges in FedAvg-LoRA

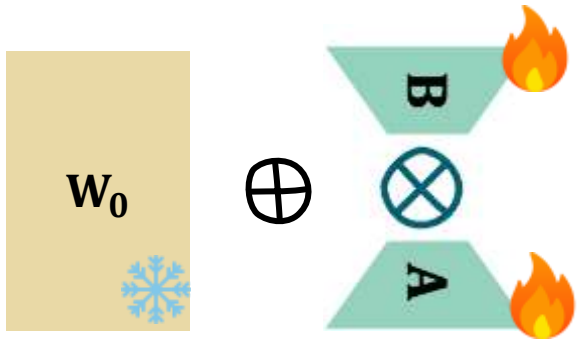
NEURAL INFORMATION PROCESSING SYSTEMS

Low-Rank Adaptation makes federated fine-tuning practical ☺



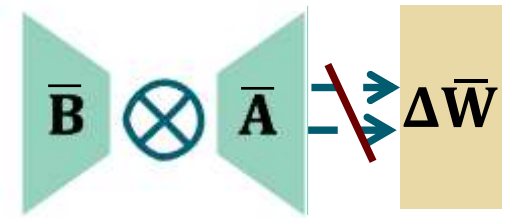
Communication Bottleneck & Limited Local Resources

Full-Para Finetune: Resources-intensive ✗



LoRA: Lower trainable parameters ✔ ***Compute & Communication & Memory Friendly***

Mot1: Imprecise aggregation



FedAvg introduces aggregation imprecision ☹

$$\bar{\mathrm{W}} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{W}_k = \frac{1}{K}\sum_{k=1}^{K}(\mathrm{W}_0 + \frac{\alpha}{r}\mathrm{B}_k\mathrm{A}_k) \neq \mathrm{W}_0 + \frac{\alpha}{K^2 r}\sum_{k=1}^{K}\mathrm{B}_k\sum_{k=1}^{K}\mathrm{A}_k.$$

Challenges in DP-LoRA

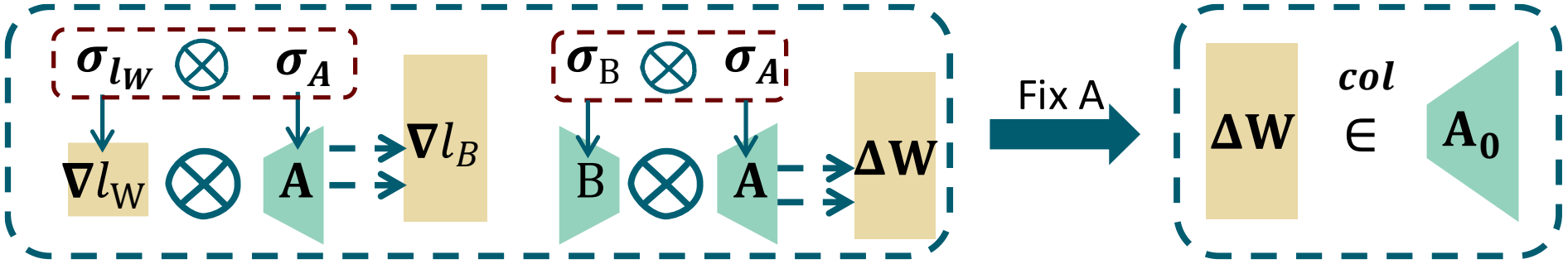
$$\nabla_A \ell = \frac{\partial \ell}{\partial A} = \frac{\alpha}{r} B^T \frac{\partial \ell}{\partial W}$$

$$\nabla_B \ell = \frac{\partial \ell}{\partial B} = \frac{\alpha}{r} \frac{\partial \ell}{\partial W} A^T$$

Fact 1: To apply DP-SGD on LoRA, both gradients are required to add noise.

Fact 2: Both gradients adding noise leads to noise amplification

$$\mathbb{E}[||\Delta W_{noise}||_F^2] \approx \underbrace{\eta^2 \frac{\sigma^2 C^2}{B_{size}^2} d_l r (||A_t||_F^2 + ||B_t||_F^2)}_{\text{Term 1: Linear Noise}} + \underbrace{\eta^4 \frac{\sigma^4 C^4}{B_{size}^4} d_l^2 r}_{\text{Term 2: Quadratic Noise}} + O(\eta^4 \sigma^2 + \eta^3 \sigma^2).$$



☹ AmplifiedNoise

☹ Limited Learning Ability

Our Goal

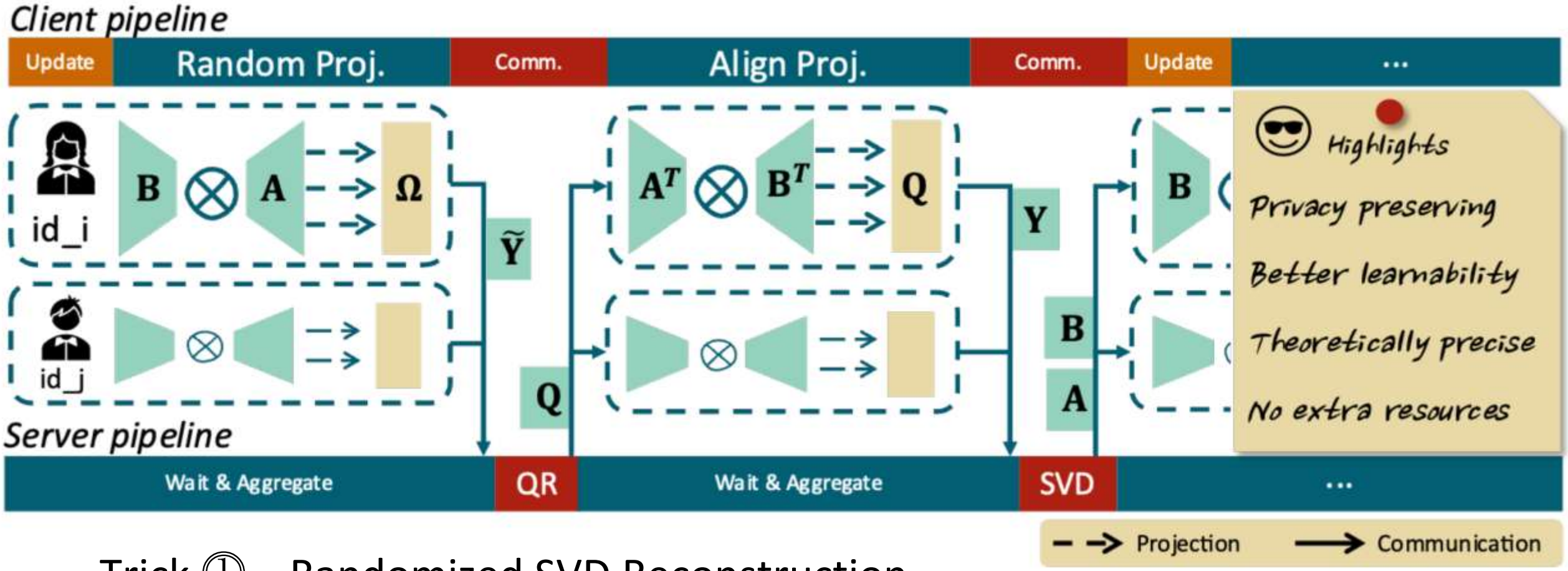
NEURAL INFORMATION PROCESSING SYSTEMS

*Can we design a federated LoRA algorithm that achieves **differential privacy guarantee**, **learnability**, **aggregation fidelity** and **efficiency** simultaneously?*

Table 1: Federated LoRA Comparison. DP (✓/✗): Supports DP. Agg. Type: Aggregation Precision. Client Init.: Parameter state before local training (Sync=Use Global Para, Keep B=Use Local B, Fixed A=Use Initial A, Rand A=Gaussian).

METHOD	DP	AGG. TYPE	AGG. MEMORY	COMMUNICATION	CLIENT INIT.
FEDAVG	✗	IMPRECISE	$O(d_l r)$	$O(K d_l r)$	SYNC A, B
FLORA	✗	PRECISE	$O(d_l^2 + K d_l r)$	$O(K^2 d_l r)$	$\mathrm{B} = \mathbf{0}, \mathrm{A}_k^0 = \mathrm{RAND}$
FEDSA	✗	IMPRECISE	$O(d_l r)$	$O(K d_l r)$	SYNC A, KEEP LOCAL B
FED-FFA	✓	PRECISE	$O(d_l r)$	$O(K d_l r)$	FIX A, SYNC B
FEDASK	✓	PRECISE	$O(d_l r)$	$O(K d_l r)$	SYNC A, B

FedASK Framework



Trick ① Randomized SVD Reconstruction

Trick ② $\mathbf{A}$ Freeze Locally But Update Globally

Analysis ---- Aggregation Fidelity

NEURAL INFORMATION PROCESSING SYSTEMS

Randomized Subspace Sketching *Global Alignment Projection* *Reconstruction Update*

$$\mathrm{Y}_k^{proj} = \mathrm{B}_k^t(\mathrm{A}_k^t\Omega). \qquad \tilde{\mathrm{Y}}_k^{proj} = (\mathrm{A}_k^t)^\top((\mathrm{B}_k^t)^\top\mathrm{Q}). \qquad \mathrm{B}^t = \mathrm{Q}\mathrm{U}\Sigma^{\frac{1}{2}}, \quad \mathrm{A}^t = \Sigma^{\frac{1}{2}}\mathrm{V}^\top.$$

Theorem 2 (Aggregate Precision of FedASK). *Let* $d_B = dim(span(\bigcup_{k=1}^K Range(\mathrm{B}_k)))$. *If the random projection matrix* $\Omega \in \mathbb{R}^{n\times(r+p)}$ *is a standard Gaussian random matrix with over-sketching parameter* p *satisfying* $p \geq d_B - r + 2$, *then FedASK guarantees that the global update before truncation* $\Delta W^t = \mathrm{B}^t\mathrm{A}^t$ *equals the exact average of local updates* $\Delta\bar{W} = \frac{1}{K}\sum_{k=1}^K \mathrm{B}_k\mathrm{A}_k$, *i.e.,*

$$\|\Delta W^t - \Delta\bar{W}\|_F = 0,$$

Analysis ---- Privacy Guarantee

NEURAL INFORMATION PROCESSING SYSTEMS

Standard DP-SGD Applied to Grad. *(Note! A is updated through reconstruction!)*

$$\mathrm{B}_k^{\tau+1} = \mathrm{B}_k^{\tau} - \frac{\gamma\alpha}{r}\left(\frac{\partial l}{\partial \mathrm{W}_k^{\tau}} \Big/ \max\left(1, \frac{\|\frac{\partial l}{\partial \mathrm{W}_k^{\tau}}\|_2}{C}\right) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)(\mathrm{A}^{t-1})^T,$$

Theorem 1 (Privacy Guarantee of FedASK). *Suppose the gradient sensitivity is $C = 1$. FedASK (Algorithm 1) guarantees that the final global LoRA matrices (A^T, B^T) are (ϵ, δ)-differentially private with respect to the joint dataset $D = \bigcup_{i=1}^{K} \mathcal{D}_k$ of K total clients, provided the variance σ^2 of the Gaussian noise added to local gradients satisfies:*

$$\sigma^2 = \mathcal{O}\left(\frac{q_{\mathcal{D}}^2 \cdot m \cdot q_{\mathcal{K}} \cdot T \cdot \ln(2/\delta) \cdot \ln(2T q_{\mathcal{K}}/\delta)}{\epsilon^2 \cdot K}\right),$$

where $q_{\mathcal{K}}$ is the client sampling ratio per communication round (total T rounds), and $q_{\mathcal{D}}$ is the data sampling ratio per local update (total m local updates per client per round).

Analysis ---- Efficiency

NEURAL INFORMATION PROCESSING SYSTEMS

Table 1: Communication Volume and Simual transfer time of Various Algorithms of 10 selected clients

	Llama-2-7b		Llama-2-13b	
	Comm. Time(s)	Total Comm. Vol.(MB)	Comm. Time(s)	Total Comm. Vol.(MB)
fedask	5.53	1200	34.60	7500
fedavg	4.43	960	27.68	6000
scaffold	8.86	1920	55.37	12000
flora	24.36	5280	152.26	33000

- *Communication complex.*

$$O(|\mathcal{K}_t| d_l \cdot (r+p))$$

- *Computation complex.*

$$O(d_l \cdot (r+p)^2)$$

- *Memory complex.*

$$O(d_l \cdot (r+p))$$

Table 4: End-to-End Wall-Clock Time per communication Round (s)

Selected Clients (K_t)	Algorithm	Client Compute (s)	Server Compute (CPU) (s)	Total Comm. Time (1Gbps) (s)	Total Round Time (1Gbps) (s)
2	**FedASK**	1.44	0.12	1.51	3.07
	FedAvg	1.64	0.07	1.21	2.92
10	**FedASK**	1.44	0.29	5.53	7.26
	FedAvg	1.64	0.35	4.84	6.79
30	**FedASK**	1.44	0.69	16.61	18.74
	FedAvg	1.64	0.71	13.29	15.64

Experiments ---- Privacy Settings

NEURAL INFORMATION PROCESSING SYSTEMS

Table 2: Performance Comparison of Different Algorithms with DP Budgets on Llama-2-7B.

Task	Priv. Budget	FedASK	FedAvg	FFA-LoRA	FedSA-LoRA	FedProx	Scaffold
MMLU	Non Private	**46.15**	45.13	45.98	45.19	44.98	45.65
	$\epsilon = 1$	**45.80**	42.07	42.76	42.9	41.99	43.41
	$\epsilon = 3$	**46.25**	41.49	42.72	41.13	43.17	42.47
	$\epsilon = 6$	**45.78**	43.34	42.82	42.84	43.70	43.80
DROP	Non Private	**32.09**	30.2	31.34	31.23	30.99	30.01
	$\epsilon = 1$	**31.23**	29.55	29.10	31.04	29.51	29.66
	$\epsilon = 3$	**32.08**	29.26	28.40	29.40	28.50	28.75
	$\epsilon = 6$	**31.36**	29.30	29.40	29.26	27.57	30.20
Human-Eval	Non Private	**15.24**	11.59	14.02	12.2	12.2	14.63
	$\epsilon = 1$	**15.24**	12.80	12.20	13.41	12.20	9.76
	$\epsilon = 3$	**15.24**	10.37	10.98	10.98	13.41	11.59
	$\epsilon = 6$	**15.85**	11.59	12.20	12.80	10.98	12.20

Experiments ---- Heterogenous Settings

NEURAL INFORMATION PROCESSING SYSTEMS

Table 4: Performance comparison across different data distributions on various tasks.

Task	Data Dist.	FedASK	FedAvg	FFA-LoRA	FedSA-LoRA	FedProx	Scaffold
MMLU	IID	**46.25**	41.49	42.72	41.13	43.17	42.47
	Dir(0.1)	**46.04**	42.69	42.54	44.27	42.61	43.05
	Dir(0.5)	**45.95**	42.11	41.46	42.72	42.98	41.97
	Dir(1)	**46.01**	42.96	43.23	41.04	42.98	41.71
DROP	IID	**32.08**	29.44	28.40	29.40	28.50	28.75
	Dir(0.1)	**31.01**	28.34	30.10	28.58	28.18	28.27
	Dir(0.5)	**31.15**	29.18	29.26	27.83	28.23	29.45
	Dir(1)	**31.58**	30.02	28.93	29.29	29.82	30.52
Human-Eval	IID	**15.24**	10.37	10.98	10.98	13.41	11.59
	Dir(0.1)	**13.41**	7.32	10.98	12.8	**13.41**	7.93
	Dir(0.5)	**14.63**	10.98	13.41	12.8	**14.63**	9.76
	Dir(1)	**14.02**	9.76	10.98	10.37	10.37	9.15

Experiments ---- Impact of Oversampling

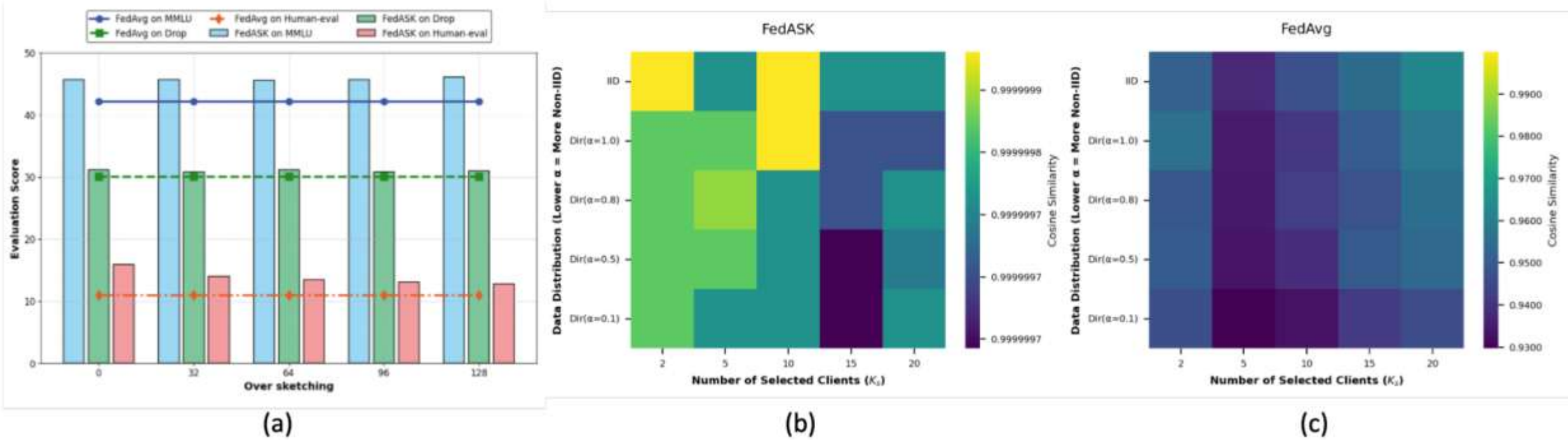


Figure 3: Impact of over-sketching p on FedASK. (a) MMLU score of FedASK versus p. (b) Aggregation fidelity of FedASK (with $p = 0$) and (c) FedAvg, measured by cosine similarity with the ideal mean of local updates (1.0 indicates perfect fidelity). Subplots (b) and (c) vary selected clients (K_s) and Non-IID degrees.

Thanks all :)

Email: *mwen25@m.fudan.edu.cn*