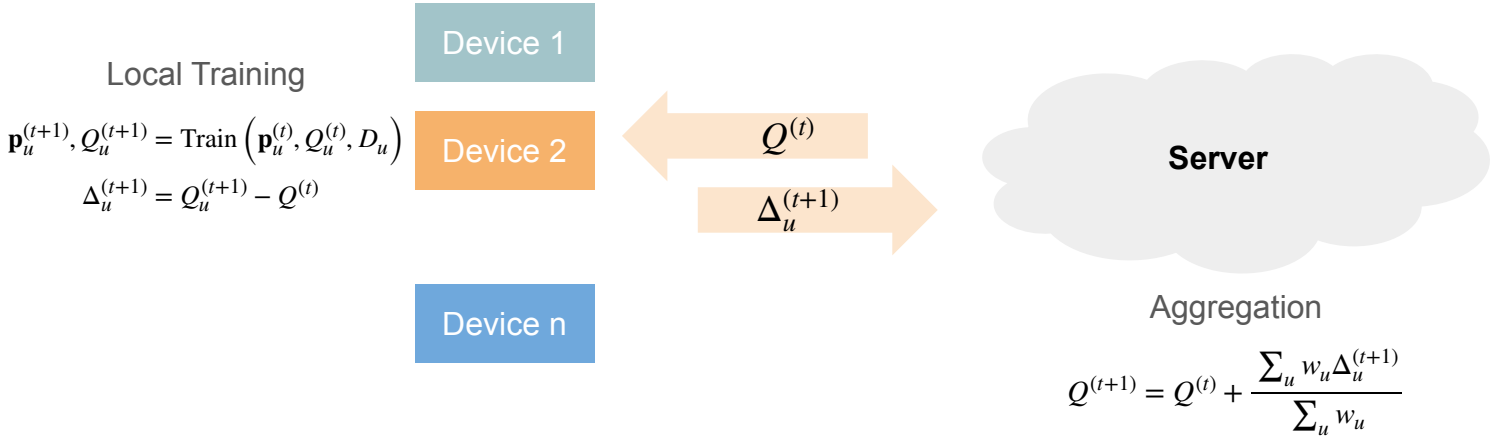
VINUNIVERSITY

Towards Efficient Communication and Secure Federated Recommendation System via Low-rank Training

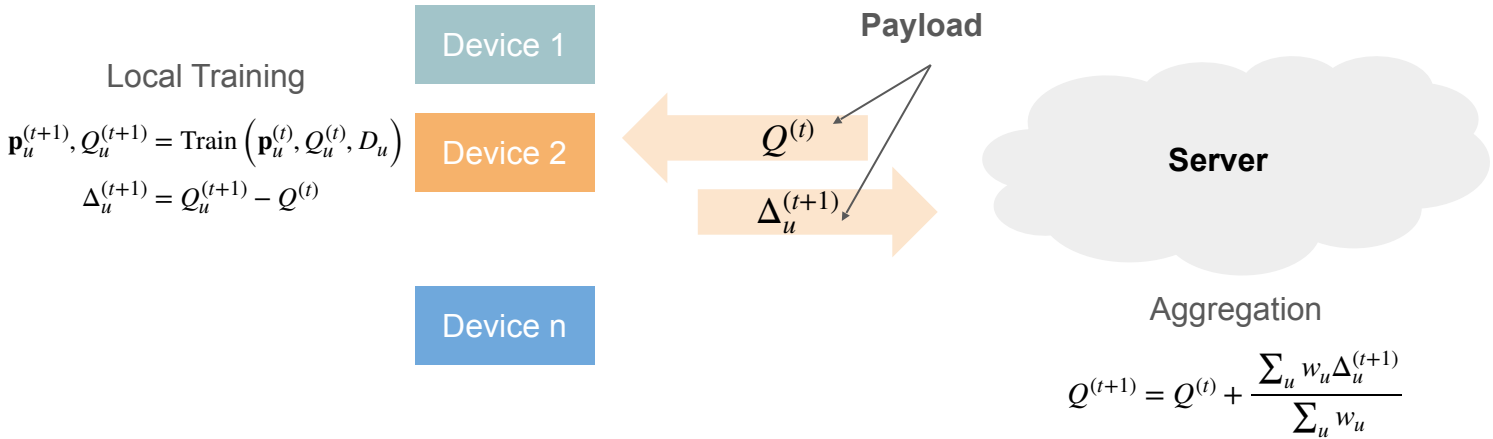
Ngoc-Hieu Nguyen (VinUni)
hieu.nn@vinuni.edu.vn

Joint work with Tuan-Anh Nguyen, Tuan Nguyen, Vu Tien Hoang, Dung D. Le, Kok-Seng Wong

Federated Recommendation Systems



Federated Recommendation Systems

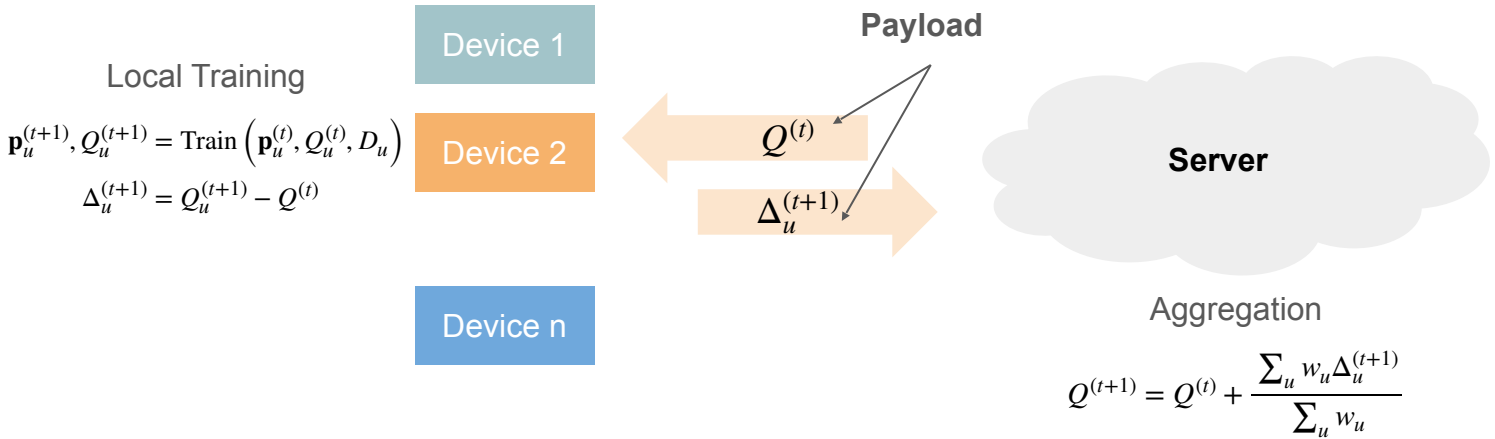


Communication cost

- The payload can be very large.
- Waiting for clients with slow network connection is a bottleneck.

VINUNIVERSITY

Federated Recommendation Systems



Communication cost

- The payload can be very large.
- Waiting for clients with slow network connection is a bottleneck.

Privacy concerns

- Curious servers can still infer users' data from updated parameters by inversion attacks.

VINUNIVERSITY

Challenges

Communication cost

- The payload can be very large.
- Waiting for clients with slow network connection is a bottleneck.

→ Message compression

Privacy concerns

- Curious servers can still infer users' data from updated parameters by inversion attacks.

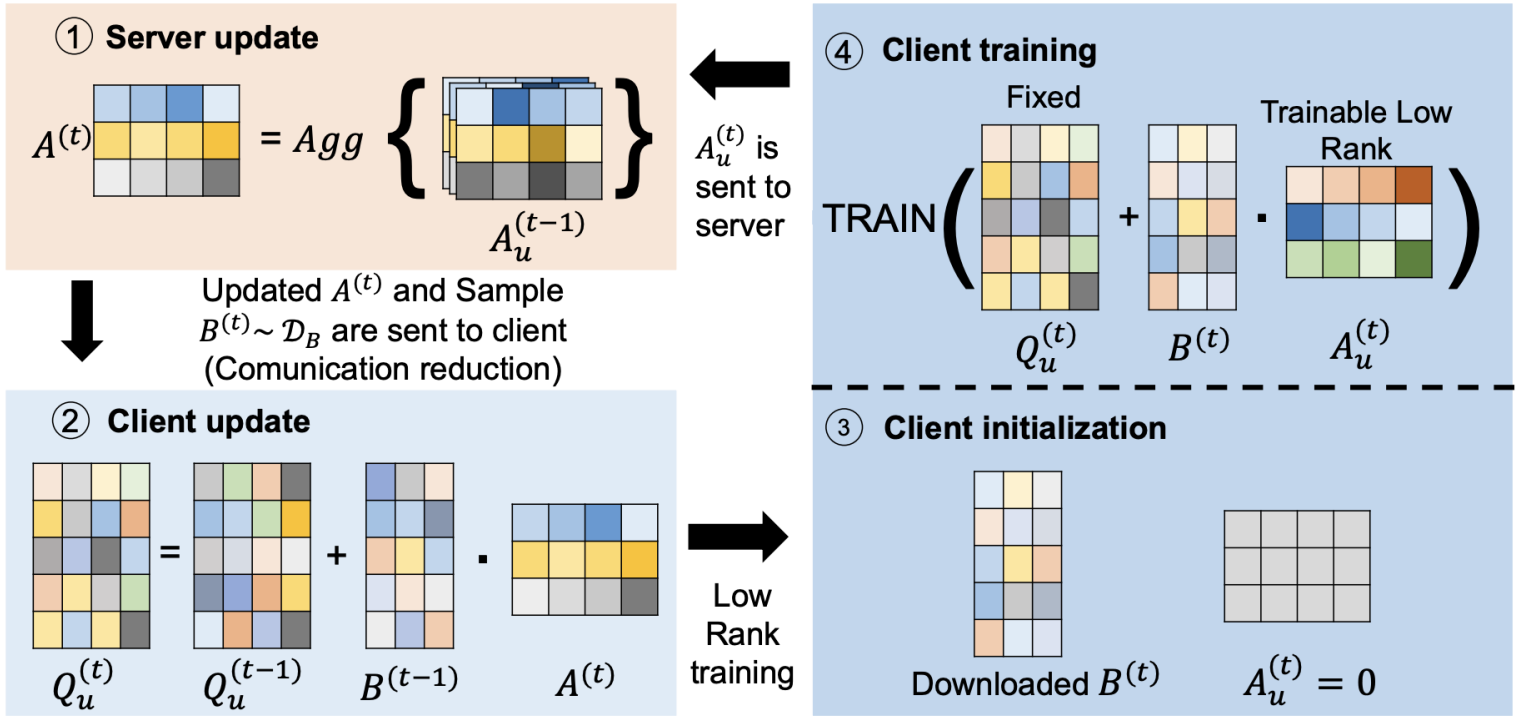
→ Homomorphic Encryption (HE)

Challenges

Tackling privacy and communication efficiency as separate concerns can result in suboptimal solutions

- **Homomorphic Encryption (HE) suffers from impractical overheads**
  - including encryption and decryption steps on the client side, as well as aggregation on the server side
- **Message compressions such as SVD or Top-K are not compatible with HE**
  - The SVD compression requires matrix multiplication on encrypted matrices which is costly
  - The Top-K compression requires the secured aggregation process on the server must be executed on each element in the top-K vector

Correlated Low-rank Structure update



Correlated Low-rank Structure update

Motivation

$$\mathbf{p}_u^{(t+1)}, Q_u^{(t+1)} = \text{Train}\left(\mathbf{p}_u^{(t)}, Q_u^{(t)}, D_u\right)$$

$$\Delta_u^{(t+1)} = Q_u^{(t+1)} - Q^{(t)}$$

$$= \eta\left[\left(\mathbf{m} * \left(\mathbf{r}_u - \hat{\mathbf{r}}_u\right)\right)\mathbf{p}_u^{(t)\top} - \lambda Q^{(t)}\right]$$

VINUNIVERSITY

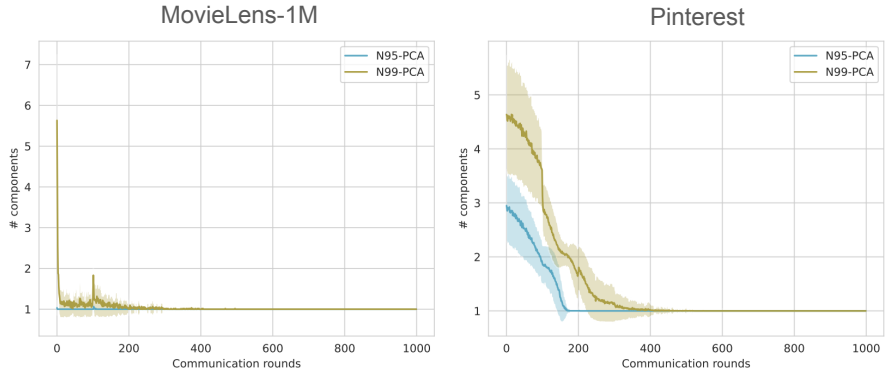
Correlated Low-rank Structure update

Motivation

$$\mathbf{p}_u^{(t+1)}, Q_u^{(t+1)} = \text{Train}\left(\mathbf{p}_u^{(t)}, Q_u^{(t)}, D_u\right)$$

$$\Delta_u^{(t+1)} = Q_u^{(t+1)} - Q^{(t)}$$

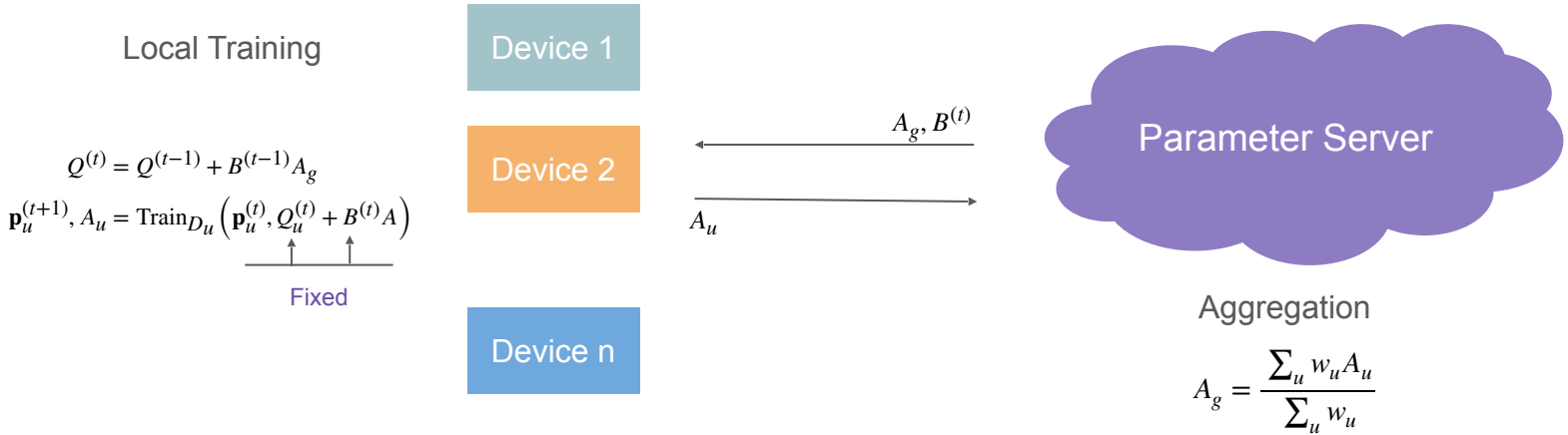
$$= \eta \left[\underbrace{\left(\mathbf{m} * \left(\mathbf{r}_u - \hat{\mathbf{r}}_u\right)\right) \mathbf{p}_u^{(t)\top}}_{\text{Rank 1}} - \underbrace{\lambda Q^{(t)}}_{\text{Typically small}} \right]$$



PCA components progression on 2 datasets

VINUNIVERSITY

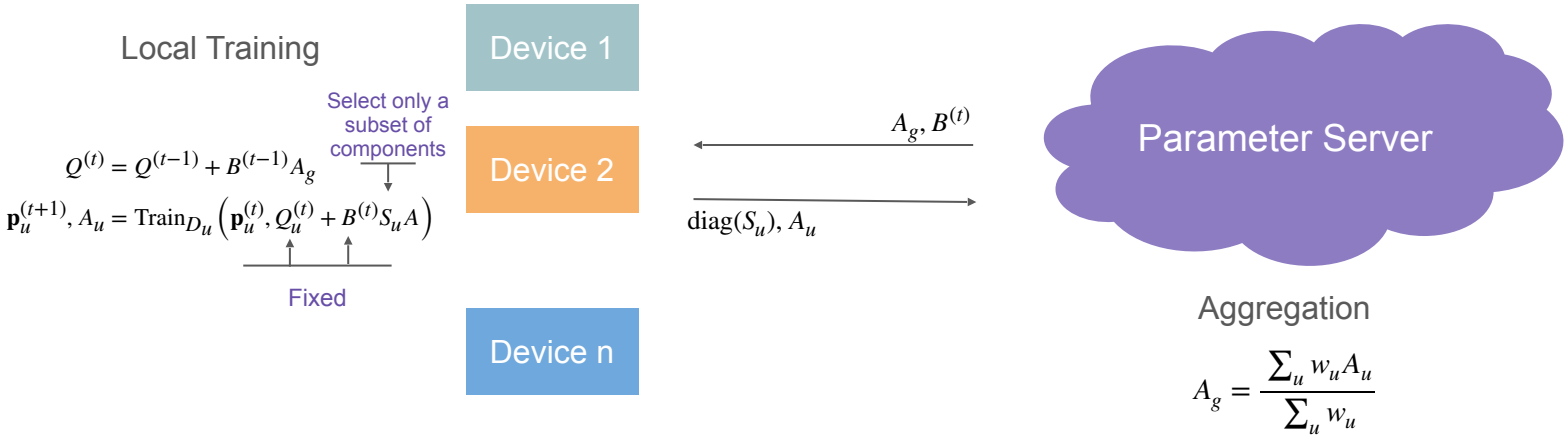
Correlated Low-rank Structure update



CoLR's advantages

1. Reduce communication cost for both uplink and downlink
2. Low computational overheads
3. Compatible with secure aggregation protocols

Subsampling Correlated Low-rank Structure update



SCoLR's advantages

1. Reduce communication cost for both uplink and downlink
2. Low computational overheads
3. Compatible with secure aggregation protocols
4. Suitable for heterogeneous network bandwidth settings

Experimental Settings

Datasets	# Users	# Items	# Interactions	Data Density
MovieLens-1M [11]	6, 040	3, 706	1, 000, 209	4.47%
Pinterest [9]	55, 187	9, 916	1, 500, 809	0.27%

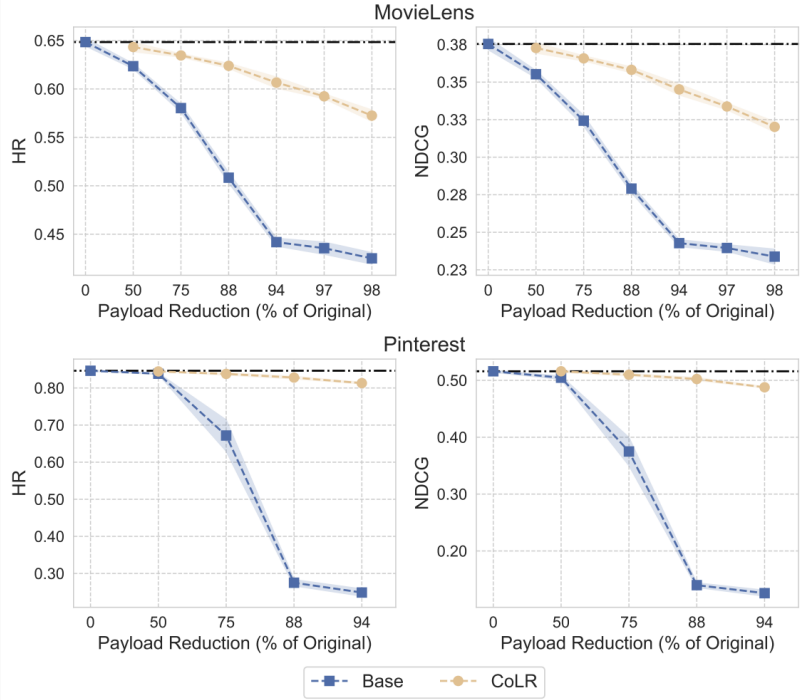
Leave-one-out evaluation protocol
Evaluation Measures
Hit Rate (HR) and Normalized Discounted Cumulative Gain (NDCG)

Comparison Methods
1. FedMF (Original)
2. FedMF-SVD
3. FedMF-TopK

Payload Reduction (%)
Recommendation performance is analyzed at 50%,75%, 88%,94%,97%,98%

Experimental Results

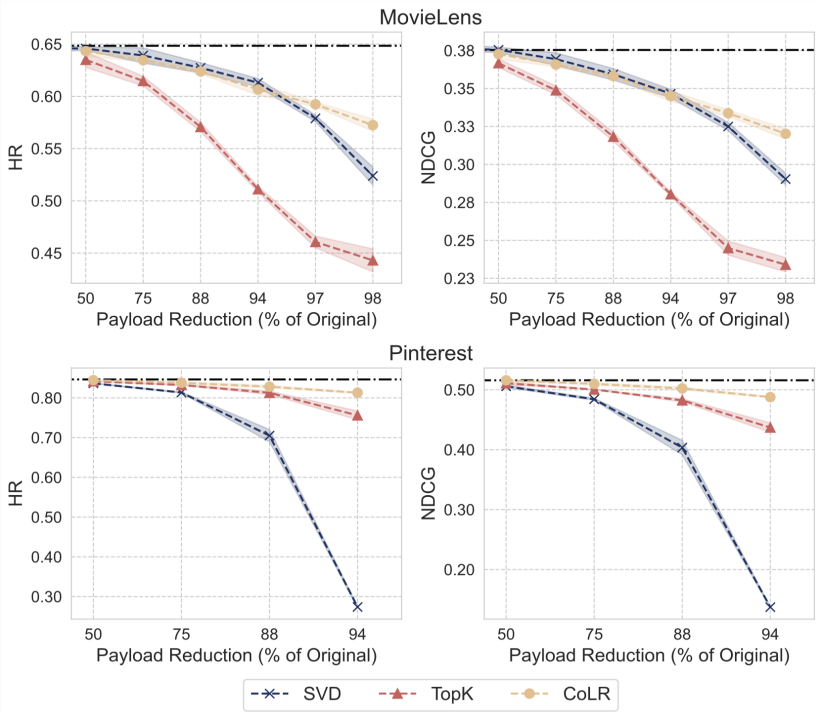
Our approach resulted in a reduction of up to 93.75% in payload size, with only an approximate 8% decrease in recommendation performance across datasets.



Performance on the MovieLens-1M dataset (Top) and the Pinterest dataset (Bottom)

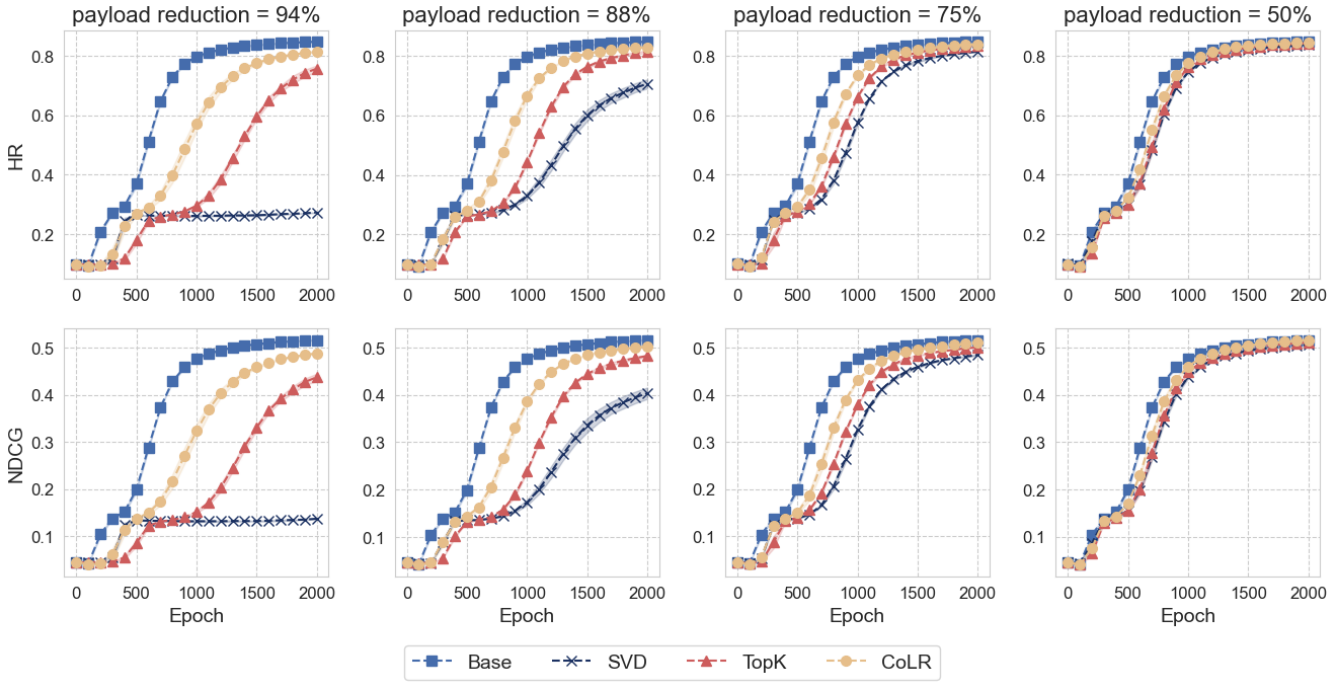
Experimental Results

CoLR consistently achieves favorable performance while outperforming other methods in scenarios with limited communication budgets.



Comparison between CoLR, SVD, and Top-K methods

Convergence Analysis



Experimental Results

Method	Client overheads	Server overheads	Ciphertext size	Plaintext size	Comm Ratio
FedMF	0.93 s	2.39 s	24,587 KB	927 KB	26.52
FedMF w/ Top-K@1/64	88.20 s	88.06 s	3,028 KB	29 KB	103.09
FedMF w/ Top-K@2/64	182.02 s	185.59 s	6,056 KB	58 KB	103.83
FedMF w/ Top-K@4/64	353.25 s	364.67 s	12,112 KB	116 KB	104.20
FedMF w/ Top-K@8/64	723.45 s	750.98 s	24,225 KB	232 KB	104.40
FedMF w/ Top-K@16/64	1449.90 s	1483.91 s	48,448 KB	464 KB	104.49
FedMF w/ CoLR@1	0.07 s	0.24 s	3,073 KB	15 KB	206.31
FedMF w/ CoLR@2	0.07 s	0.25 s	3,073 KB	29 KB	104.63
FedMF w/ CoLR@4	0.07 s	0.25 s	3,073 KB	58 KB	52.69
FedMF w/ CoLR@8	0.08 s	0.25 s	3,073 KB	116 KB	26.44
FedMF w/ CoLR@16	0.15 s	0.51 s	6,147 KB	232 KB	26.49
FedMF w/ CoLR@32	0.30 s	1.03 s	12,293 KB	464 KB	26.51

Overheads, and Communication ratios for MovieLens-1M dataset; Comm Ratio is calculated by file sizes of Ciphertext over file sizes of Plaintext

Experiments

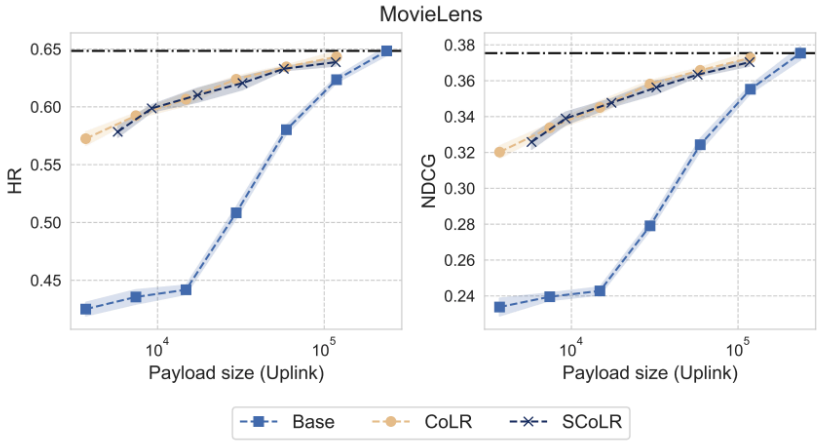
Settings

The global rank $r_g \in \{2,4,8,16,32,64\}$

Uniformly sample the local rank r_u for each device such that $1 \leq r_u \leq r_g$

Result

SCoLR is effective under the device heterogeneity setting and match the performance of CoLR under the sample uplink communication budget.



Summary

- We propose two novel frameworks, CoLR and SCoLR, designed to tackle the communication challenge in training FedRec systems.
- Experimental results showcase the effectiveness of our methods.
- CoLR and SCoLR is compatible with Homomorphic Encryption-based FedRec systems and, hence, reinforces the security of the overall systems.

Thank you