# NETWORK COMPRESSION IN FEDERATED MACHINE LEARNING
# 联邦学习中的网络压缩研究

电子信息与电气工程学院 信息安全专业

廖宁祎 517021910844

指导老师：邵硕 向立瑶

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

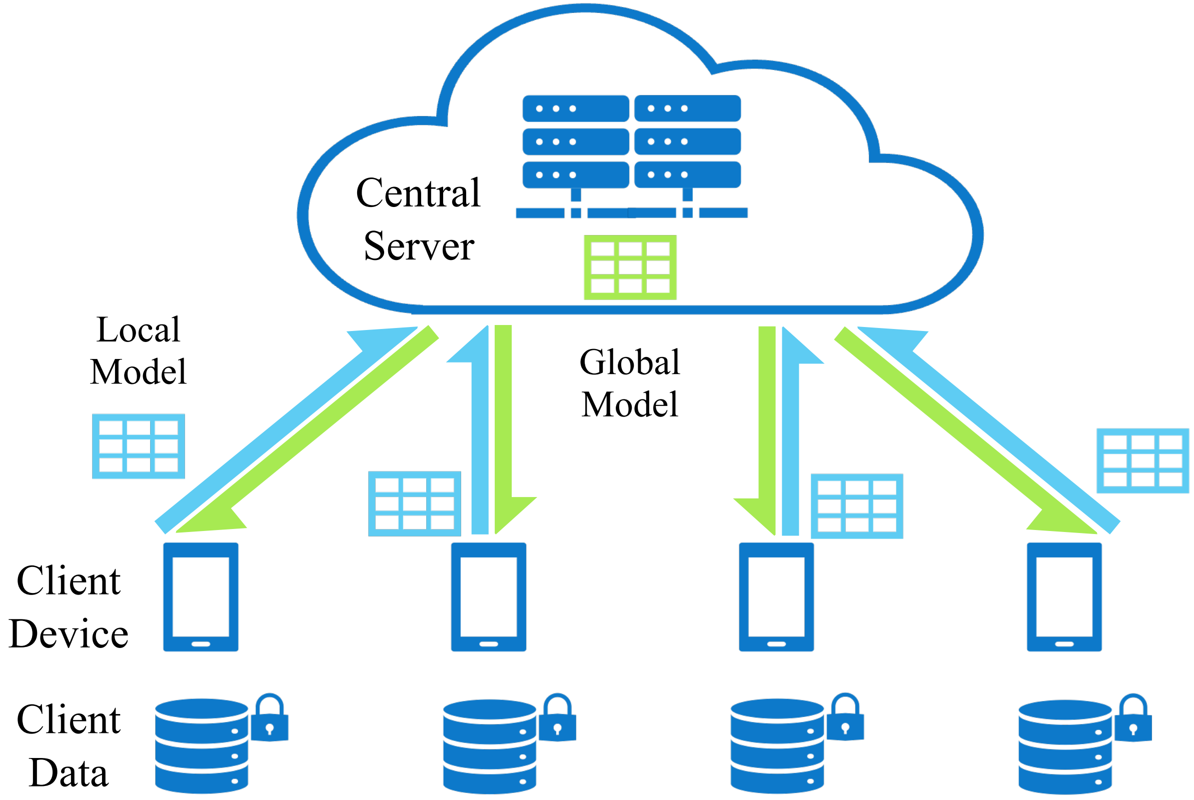
# OUTLINE

| | |
|---|---|
| Background | Introduction, Background |
| Contribution | Motivation, Contribution |
| Algorithm | Algorithmic Framework, Theoretical Results |
| Experiment | Experimental Results on Efficiency and Privacy |
| Conclusion | Conclusions, Future Work |

# Background

## Federated Learning (FL)

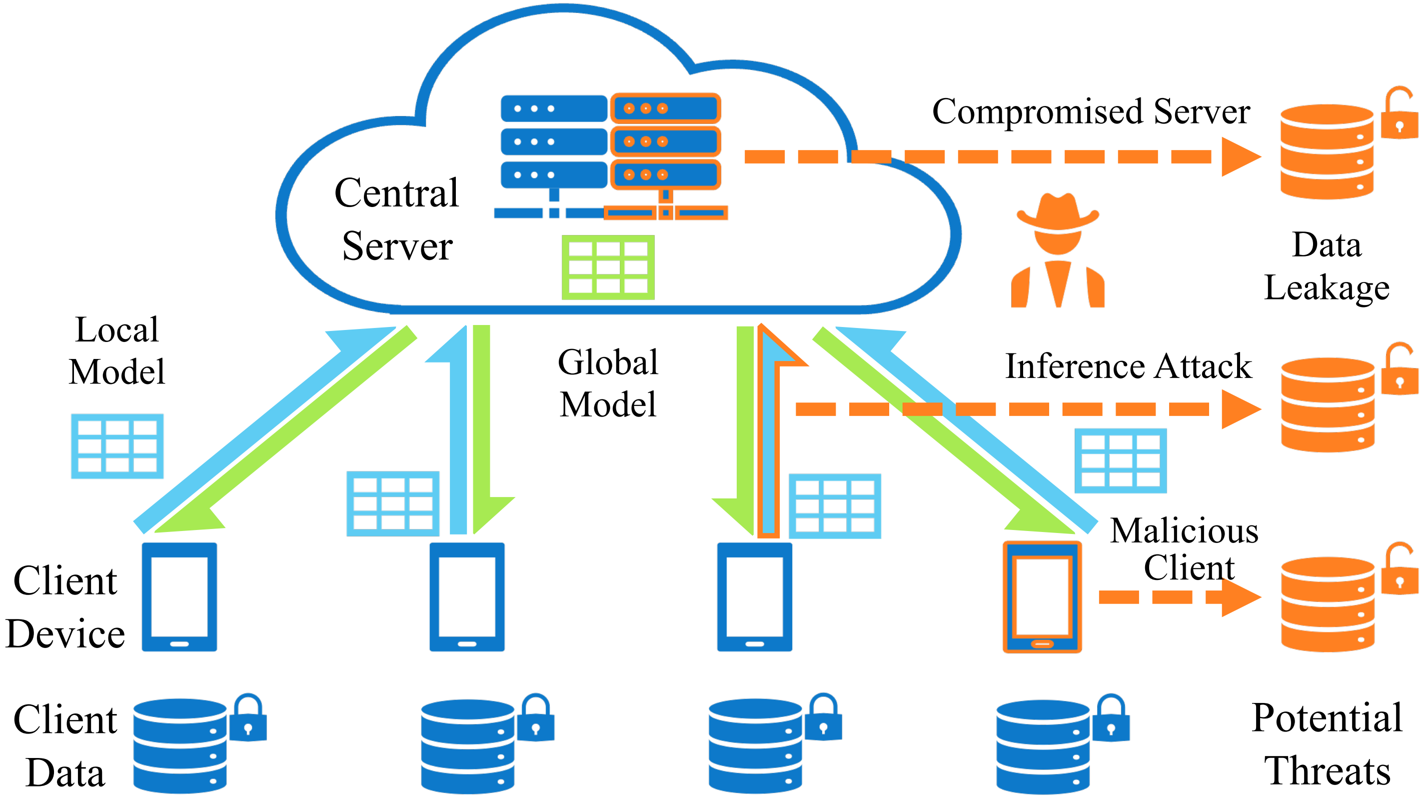
- Machine learning that decouples training model and data
- Key factors: communication overhead & data privacy
- *Client*: update local model; keep data private
- *Server*: aggregate global model



# BACKGROUND

## Federated Learning Privacy

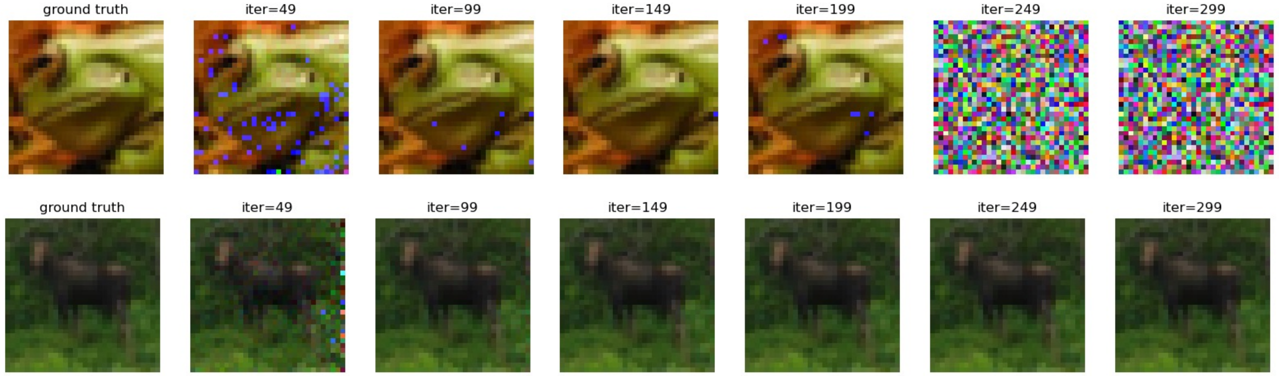
- Attackers exploit model uploads to infer client data
- *Membership inference*: distinguish client participation
- *Gradient inversion*: recover private training sample



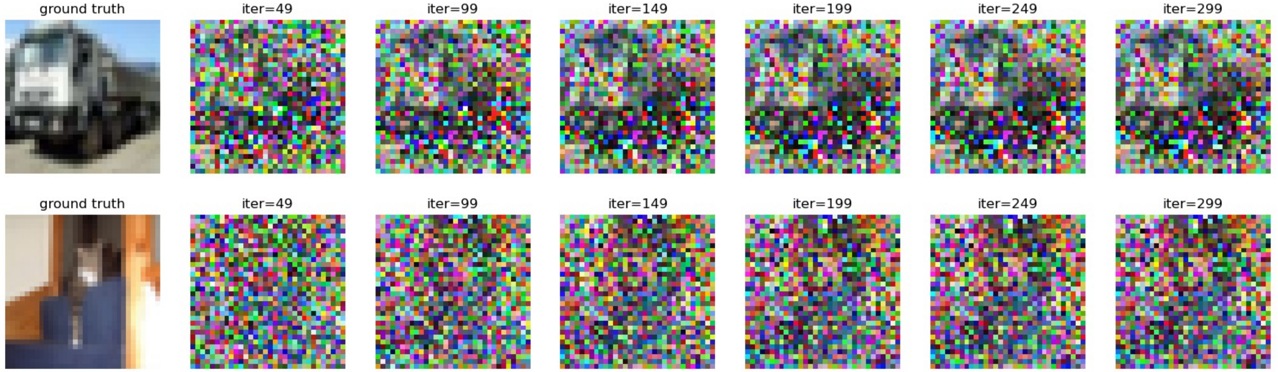
# Background

## Federated Learning Privacy

- Attackers exploit model uploads to infer client data
- *Membership inference*: distinguish client participation
- *Gradient inversion*: recover private training sample




truth DLG: FedSGD



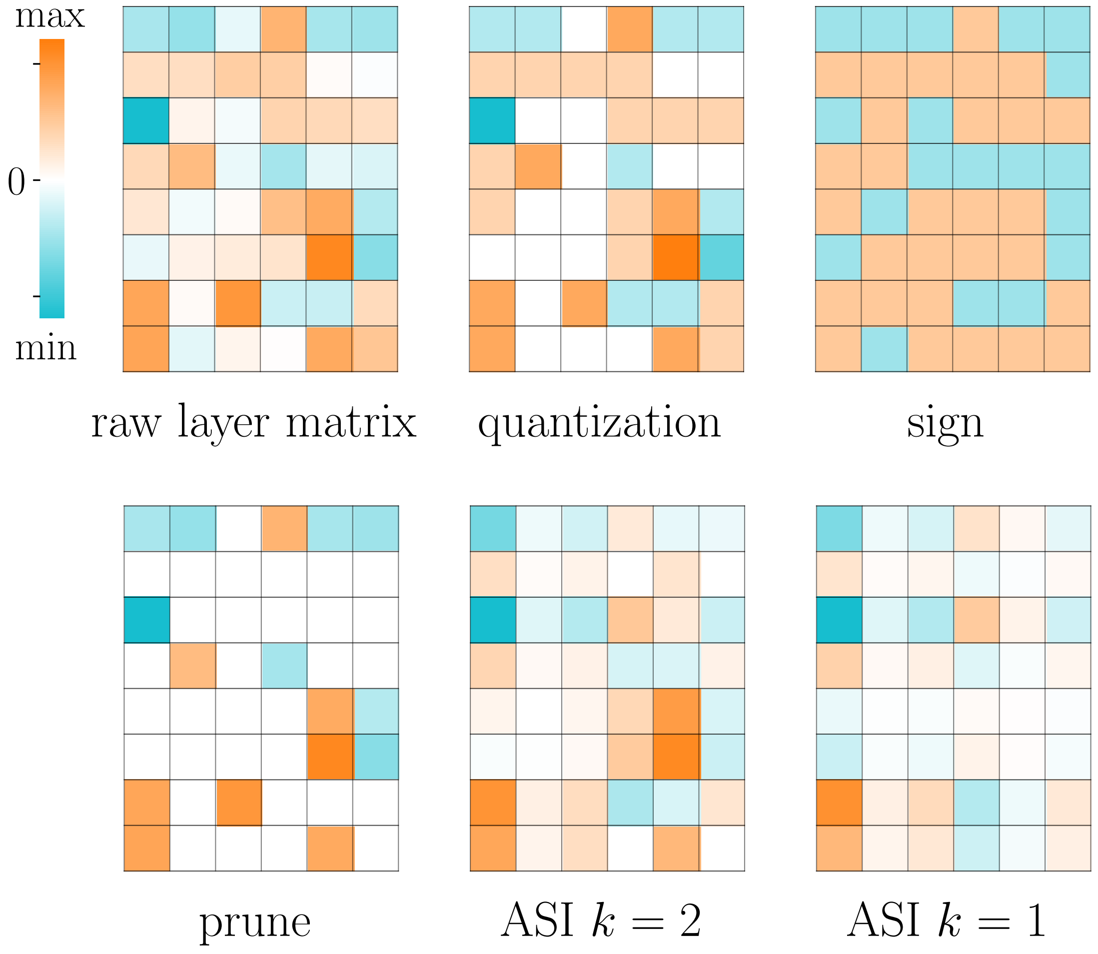

truth DLG: FedSGD + noise

Zhu L, Liu Z, Han S. Deep Leakage from Gradients[C]. in: *32nd Advances in Neural Information Processing Systems: NeurIPS*. 2019.

# BACKGROUND

## Federated Learning Communication

- Uplink transmits large-scale model weights/gradients tensors
- Protocol improvement: *client selection*, *local update*, etc.
- Parameter compression: *pruning*, *quantization*, etc.



# Contribution

## Question:

How to optimize transmission cost adaptively in the setting of differentially-private training, and design algorithms towards **communication-efficient** and **privacy-preserving** federated learning?

## Our Contribution:

**FedASI & FedALS**

Adaptive low-rank factorization

FL communication compression

Differential privacy guarantee

# CONTRIBUTION

## Our Contribution: **FedASI & FedALS**

Adaptive low-rank factorization

- Develop low-rank factorization based on ASI and ALS
- Solve matrix rank optimization with adaptive output
- Derive convergence proof on factorization iteration

FL communication compression

- Introduce low-rank compression to FL communication
- Derive convergence proof on learning
- Evaluate efficiency, accuracy, convergence by experiment

Differential privacy guarantee

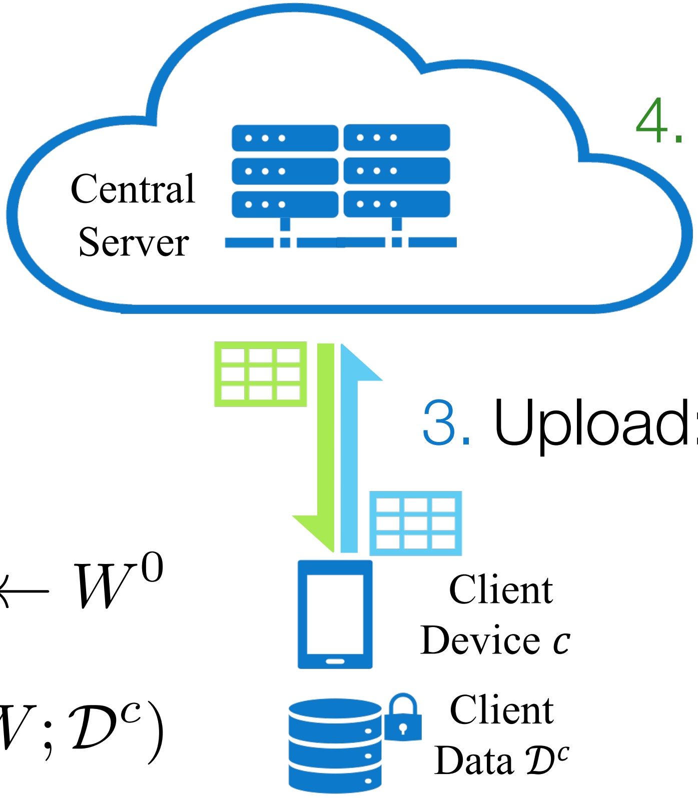
- Apply differential privacy mechanism with proof
- Link privacy with compression benefits

# FRAMEWORK DESIGN

## FL Algorithm Overview

- Objective:

$$\min_{W} \mathcal{L}(W) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathcal{L}^c(W; \mathcal{D}^c)$$




4. Aggregation:

$$W_e \leftarrow W_{e-1} + \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} M_e^c$$

3. Upload: $M \leftarrow W - W^0$

1. Download: $W \leftarrow W^0$

2. Update: $W \leftarrow W - \eta \nabla_W \mathcal{L}^c(W; \mathcal{D}^c)$

# FRAMEWORK DESIGN

## FL with Low-Rank Compression

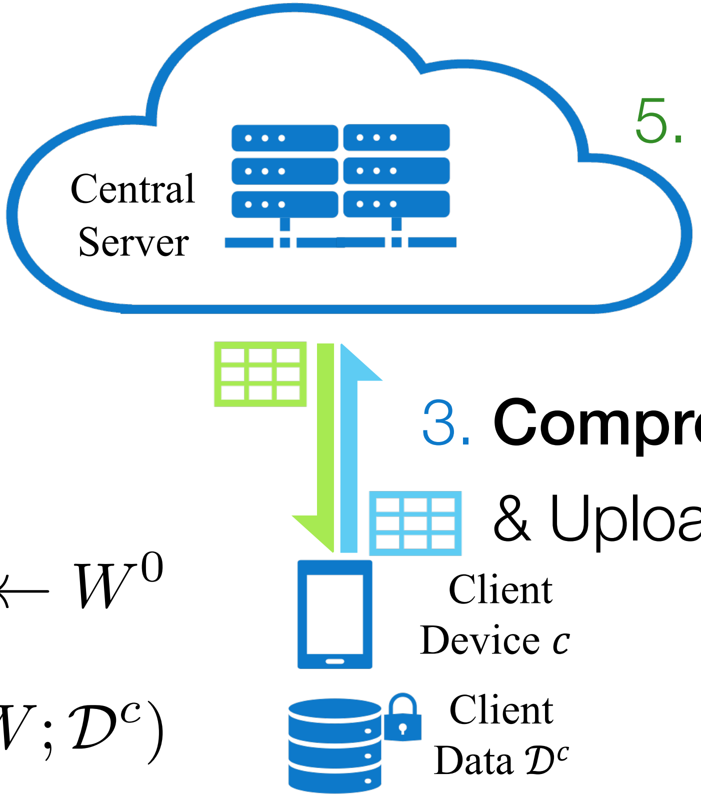
| | Efficiency: | Efficiency +Privacy: |
|---|---|---|
| Fixed: | **FedASI** | **FedASI-DP** |
| Adaptive: | **FedALS** | **FedALS-DP** |

4. **Decompress**:

$$M_e^c \leftarrow P_e^c Q_e^{cT}$$

5. Aggregation:

$$W_e \leftarrow W_{e-1} + \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} M_e^c$$




3. **Compress** $M \leftarrow W - W^0$
& Upload: $P, Q \leftarrow \mathrm{Factorize}(M)$

1. Download: $W \leftarrow W^0$

2. Update: $W \leftarrow W - \eta \nabla_W \mathcal{L}^c(W; \mathcal{D}^c)$

# Low-Rank Compression

## Low-Rank Matrix Factorization

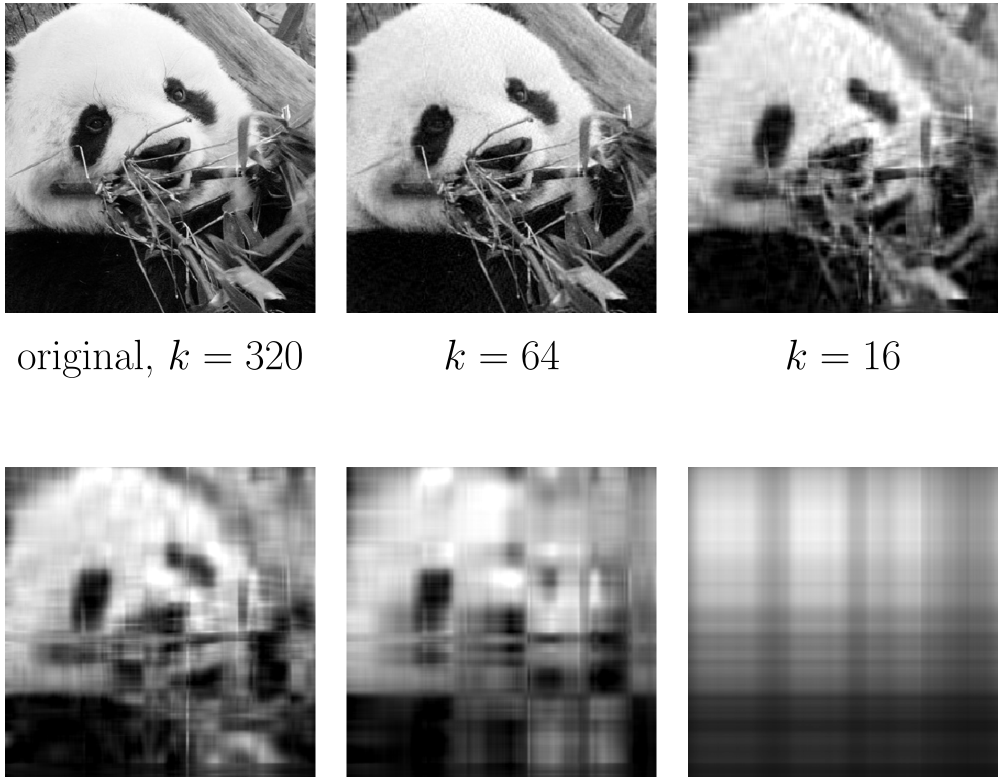
- Rank-$k$ Factorization:

$$M \approx M_k = PQ^T$$

- Rank-$k$ Restricted SVD:

$$M \approx M_k = U_k \Sigma_k V_k^T$$

- Other methods: eigenvalue decomposition, QR factorization, etc.



original, $k = 320$ | $k = 64$ | $k = 16$ | $k = 8$ | $k = 4$ | $k = 1$

# Low-Rank Compression

## Alternating Subspace Iteration (**ASI**)

- Rank-$k$ Factorization:

$$M \approx M_k = PQ^T$$

- Power Iteration:

$$\boldsymbol{u} \sim M^t \boldsymbol{u}$$

- Alternating Subspace Iteration:

$$\boldsymbol{p} \sim \left(MM^T\right)^t \boldsymbol{p} \qquad P = U_k$$

$$\boldsymbol{q} \sim \left(M^T M\right)^t \boldsymbol{q} \qquad Q = V_k \Sigma_k.$$

**Algorithm 3–2** Alternating Subspace Iteration

**Input:** matrix $M \in \mathbb{R}^{m \times n}$,
rank $k$, iteration $T$
**Output:** low-rank matrices $P \in \mathbb{R}^{m \times k}, Q \in \mathbb{R}^{n \times k}$

$Q_0 \leftarrow \mathcal{N}(0,1)^{n \times k}$
**for** *each iteration* $t = 1, 2, \cdots, T$ **do**
  $P_t \leftarrow M Q_{t-1}$
  $\hat{P}_t \leftarrow \texttt{orthonormalize}(P_t)$
  $Q_t \leftarrow M^T \hat{P}_t$
**end for**
**return** $\hat{P}_T, Q_T$

# Low-Rank Compression

## Alternating Least Squares (**ALS**)

- Rank-$k$ Factorization Optimization:

$$\min_{P,Q} \frac{1}{2} \left\|M - PQ^T\right\|_F^2, \quad \text{s.t. } \operatorname{rank}(PQ^T) = k$$

$$\Rightarrow \min_{P,Q} \frac{1}{2} \left\|M - PQ^T\right\|_F^2 + \lambda \operatorname{rank}(PQ^T)$$

$$\Rightarrow \min_{P,Q} \frac{1}{2} \left\|M - PQ^T\right\|_F^2 + \lambda \left\|PQ^T\right\|_*$$

$$\Rightarrow \min_{P,Q} \frac{1}{2} \left\|M - PQ^T\right\|_F^2 + \frac{\lambda}{2} \left(\|P\|_F^2 + \|Q\|_F^2\right)$$

$$P = U_k(\Sigma_k - \lambda I_k)_+^{\frac{1}{2}}, \; Q = V_k(\Sigma_k - \lambda I_k)_+^{\frac{1}{2}}$$

**Algorithm 3–3** Alternating Least Squares

**Input:** matrix $M \in \mathbb{R}^{m\times n}$,
regularization parameter $\lambda$, iteration $T$
**Output:** low-rank matrices $P \in \mathbb{R}^{m\times k}, Q \in \mathbb{R}^{n\times k}$

$Q_0 \leftarrow \mathcal{N}(0,1)^{n\times k}$
**for** *each iteration* $t = 1, 2, \cdots, T$ **do**
  $P_t \leftarrow M Q_{t-1} \left(Q_{t-1}^T Q_{t-1} + \lambda I\right)^{-1}$
  $Q_t \leftarrow M^T P_t \left(P_t^T P_t + \lambda I\right)^{-1}$
**end for**
**return** $P_T, Q_T$

# Private Low-Rank Compression

## Differential Privacy (DP)

- $(\epsilon, \delta)$-DP:

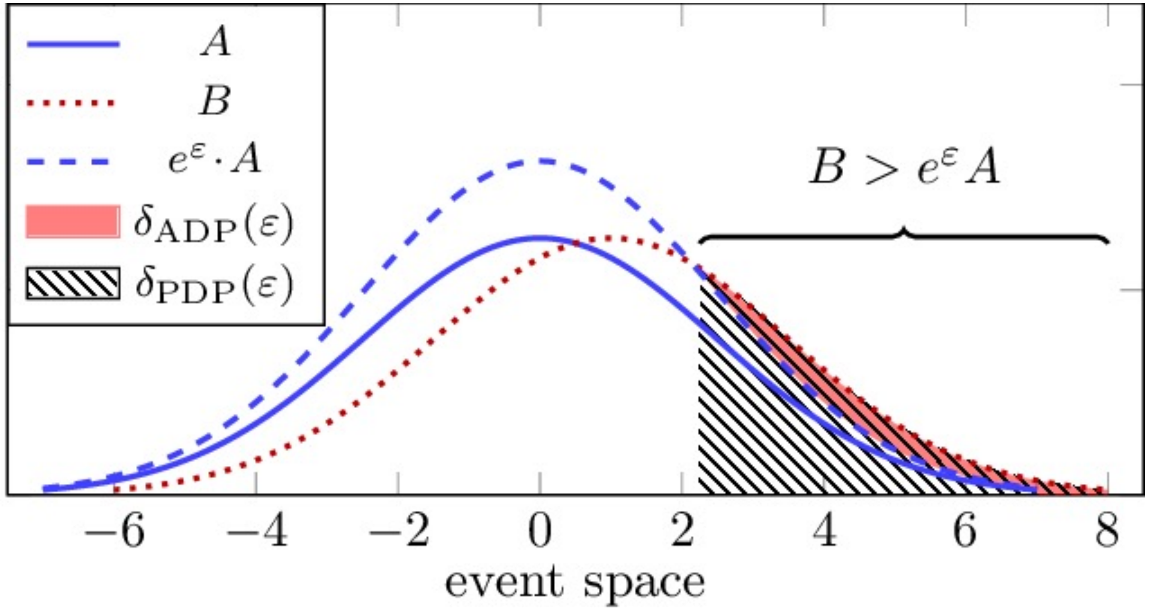
$$\Pr\left[\mathcal{M}(\mathcal{D}) \in \mathbb{S}\right] \leq e^{\epsilon} \cdot \Pr\left[\mathcal{M}(\mathcal{D}') \in \mathbb{S}\right] + \delta$$

- Gaussian Mechanism:

$$\mathcal{M}(\mathcal{D}) = f(\mathcal{D}) + G$$
$$G = \mathcal{N}\left(0, \mathcal{S}^2\sigma^2 I_M\right)$$
$$\sigma = \epsilon^{-1}\sqrt{2\ln 1.25\delta^{-1}}$$

- Sensitivity:

$$\mathcal{S}(f) = \max_{\mathcal{D},\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_F$$




David S, Sebastian M, Esfandiar M. Privacy Loss Classes: The Central Limit Theorem in Differential Privacy[C]. in: *Proceedings on Privacy Enhancing Technologies*. 2019.

# PRIVATE LOW-RANK COMPRESSION

Private ASI (**ASI-DP**)

Private ALS (**ALS-DP**)

- $(\epsilon, \delta)$-DP:

$$\mathcal{S}_Q = \|Q_{t-1}\|_\infty$$
$$\mathcal{S}_P = \|P_t\|_\infty$$
$$\sigma = \epsilon^{-1}\sqrt{4kT\ln(1/\delta)}$$

**Algorithm 4–1** ASI-DP

**Input:** matrix $M \in \mathbb{R}^{m\times n}$,
privacy budget $(\epsilon, \delta)$,
iteration $T$,
rank $k$

**Output:** low-rank matrices
$P \in \mathbb{R}^{m\times k}, Q \in \mathbb{R}^{n\times k}$

$Q_0 \leftarrow \mathcal{N}\left(0, I_{n\times k}\right)$
**for** *each iteration* $t = 1, 2, \cdots, T$ **do**
  $G_{P,t} \leftarrow \mathcal{N}\left(0, \mathcal{S}_Q^2\sigma^2 I_{m\times k}\right)$
  $P_t \leftarrow MQ_{t-1} + G_{P,t}$
  $\hat{P}_t \leftarrow \mathtt{orthonormalize}(P_t)$
  $G_{Q,t} \leftarrow \mathcal{N}\left(0, \mathcal{S}_P^2\sigma^2 I_{n\times k}\right)$
  $Q_t \leftarrow M^T\hat{P}_t + G_{Q,t}$
**end for**

**Algorithm 4–2** ALS-DP

**Input:** matrix $M \in \mathbb{R}^{m\times n}$,
privacy budget $(\epsilon, \delta)$,
iteration $T$,
regularization parameter $\lambda$

**Output:** low-rank matrices
$P \in \mathbb{R}^{m\times k}, Q \in \mathbb{R}^{n\times k}$

$Q_0 \leftarrow \mathcal{N}\left(0, I_{n\times k}\right)$
**for** *each iteration* $t = 1, 2, \cdots, T$ **do**
  $\check{Q}_{t-1} \leftarrow Q_{t-1}\left(Q_{t-1}^T Q_{t-1} + \lambda I\right)^{-1}$
  $G_{P,t} \leftarrow \mathcal{N}\left(0, \mathcal{S}_Q^2\sigma^2 I_{m\times k}\right)$
  $P_t \leftarrow M\check{Q}_{t-1} + G_{P,t}$
  $\check{P}_t \leftarrow P_t\left(P_t^T P_t + \lambda I\right)^{-1}$
  $G_{Q,t} \leftarrow \mathcal{N}\left(0, \mathcal{S}_P^2\sigma^2 I_{n\times k}\right)$
  $Q_t \leftarrow M^T\check{P}_t + G_{Q,t}$
**end for**

# Theoretical Result

## Convergence of Factorization

- ASI:

$$\boldsymbol{p}_k = \boldsymbol{u}_k + c\left(\frac{\varsigma_{k+1}}{\varsigma_k}\right)^{2t}\boldsymbol{u}_{k+1}$$

$$\boldsymbol{q}_k = \varsigma_k \boldsymbol{v}_k + c\left(\frac{\varsigma_{k+1}}{\varsigma_k}\right)^{2t}\boldsymbol{v}_{k+1}$$

- ALS:

$$\boldsymbol{p}_k = s_k^{\frac{1}{2}}\boldsymbol{u}_k + c\left(\frac{\varsigma_{k+1}}{\varsigma_k}\right)^{2t}\boldsymbol{u}_{k+1}$$

$$\boldsymbol{q}_k = s_k^{\frac{1}{2}}\boldsymbol{v}_k + c\left(\frac{\varsigma_{k+1}}{\varsigma_k}\right)^{2t}\boldsymbol{v}_{k+1}$$

- ASI-DP:

$$\left\|(I - P_t P_t^T)U_k\right\|_2 \le O\left(\frac{\sigma \max_t \|Q_t\|_\infty \sqrt{m \ln t}}{\varsigma_k - \varsigma_{k+1}} \frac{\sqrt{p}}{\sqrt{p} - \sqrt{k-1}}\right)$$

$$\left\|(\Sigma^2 - Q_t Q_t^T)V_k\right\|_2 \le O\left(\frac{\sigma \max_t \|P_t\|_\infty \sqrt{n \ln t}}{\varsigma_k - \varsigma_{k+1}} \frac{\sqrt{p}}{\sqrt{p} - \sqrt{k-1}}\right)$$

- ALS-DP:

$$\left\|(S - P_t P_t^T)U_k\right\|_2 \le O\left(\frac{\sigma \max_t \|Q_t\|_\infty \sqrt{m \ln t}}{\varsigma_k - \varsigma_{k+1}} \frac{\sqrt{p}}{\sqrt{p} - \sqrt{k-1}}\right)$$

$$\left\|(S - Q_t Q_t^T)V_k\right\|_2 \le O\left(\frac{\sigma \max_t \|P_t\|_\infty \sqrt{n \ln t}}{\varsigma_k - \varsigma_{k+1}} \frac{\sqrt{p}}{\sqrt{p} - \sqrt{k-1}}\right)$$

# THEORETICAL RESULT

## Convergence of Federated Learning

- FedASI/FedALS:

$$\mathbb{E}\left\|\nabla\mathcal{L}^c(\boldsymbol{w}_i^c)\right\|^2 \leq \left(\frac{\mathbb{E}[\mathcal{L}(\boldsymbol{w}_0)]-\mathcal{L}^*}{\bar{\eta}}+\frac{\bar{\eta}Lh^2}{b|\mathcal{C}|}\right)\frac{4}{\sqrt{T}}+8\frac{4-3\bar{\xi}^2}{\bar{\xi}^2}\frac{\bar{\eta}^2L^2H^2\mathcal{I}^2}{T}$$

- FedASI-DP/FedALS-DP:

$$\mathbb{E}\left\|\boldsymbol{w}_i-\boldsymbol{w}^*\right\|^2 \leq \frac{1}{\gamma+T}\left[\frac{4D}{\mu^2}+\left(\frac{8L}{\mu}+\mathcal{I}\right)\left\|\boldsymbol{w}_0-\boldsymbol{w}^*\right\|^2\right]$$

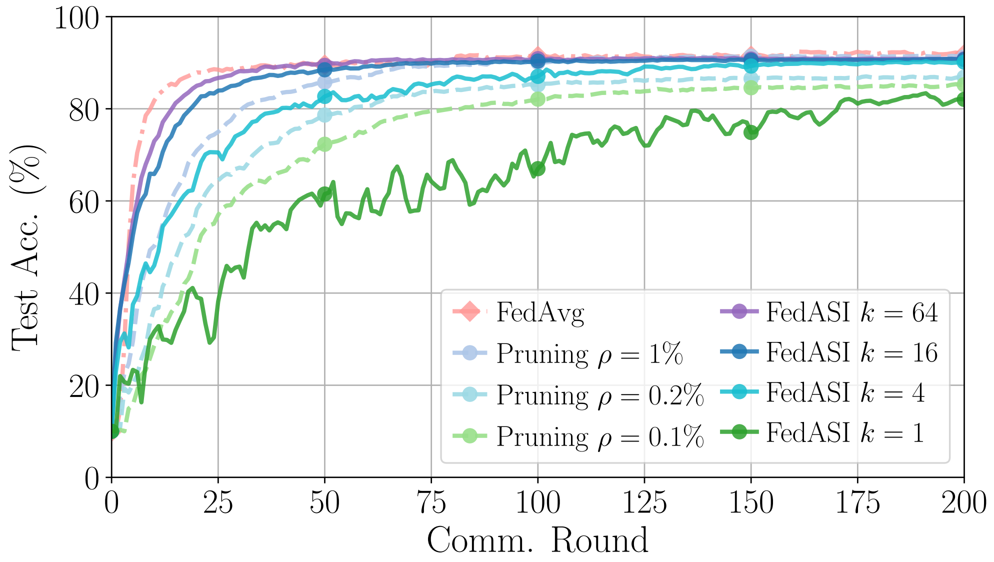
# Experiment on Efficiency

## Experiment Setting

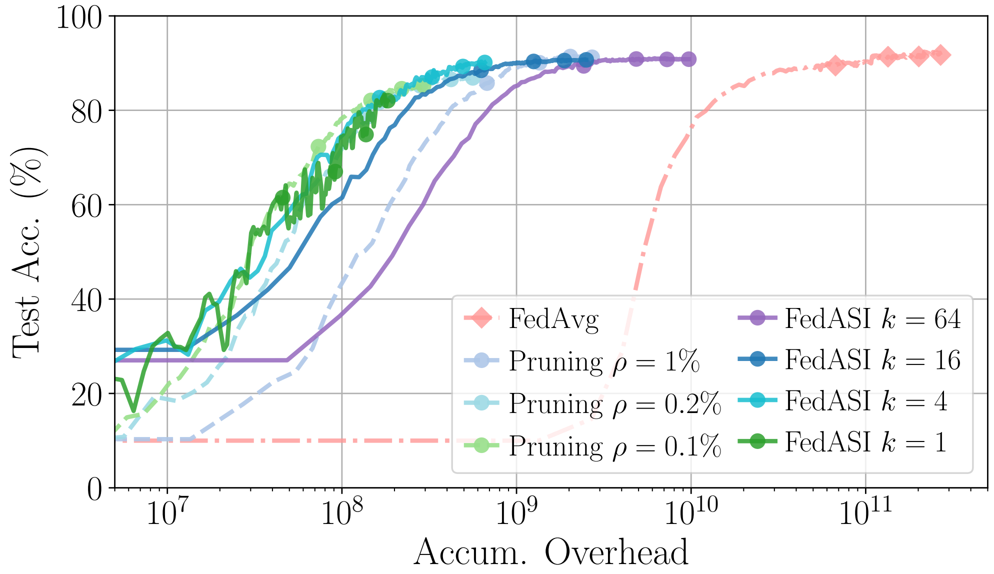
- Datasets: CIFAR-10 (IID)
- Model: VGG16, ResNet18

## FedASI Results

- Less than 0.5% accuracy loss
- Faster convergence




VGG16 Accuracy vs. Training Round

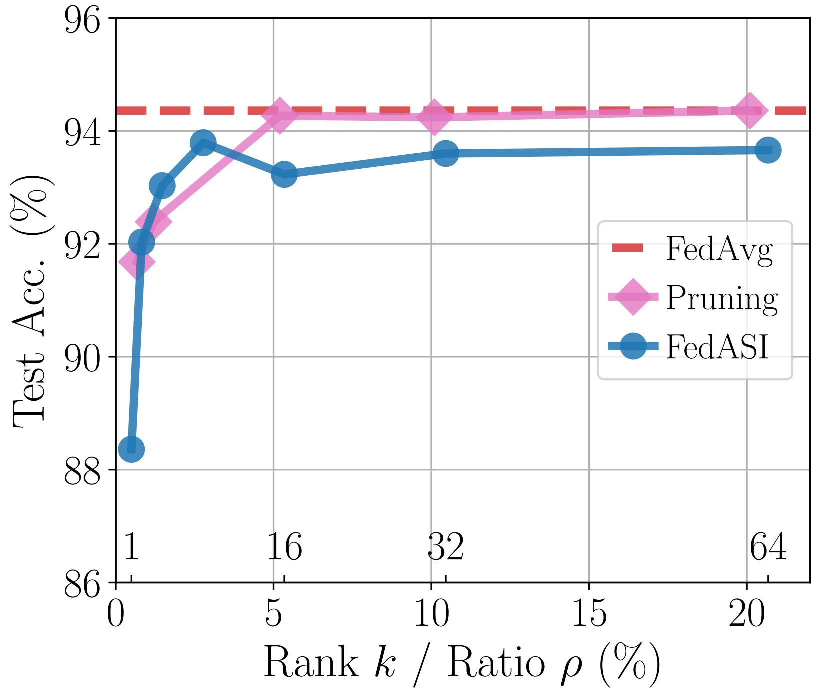



VGG16 Accuracy vs. Accum. Comm. Overhead

# Experiment on Efficiency

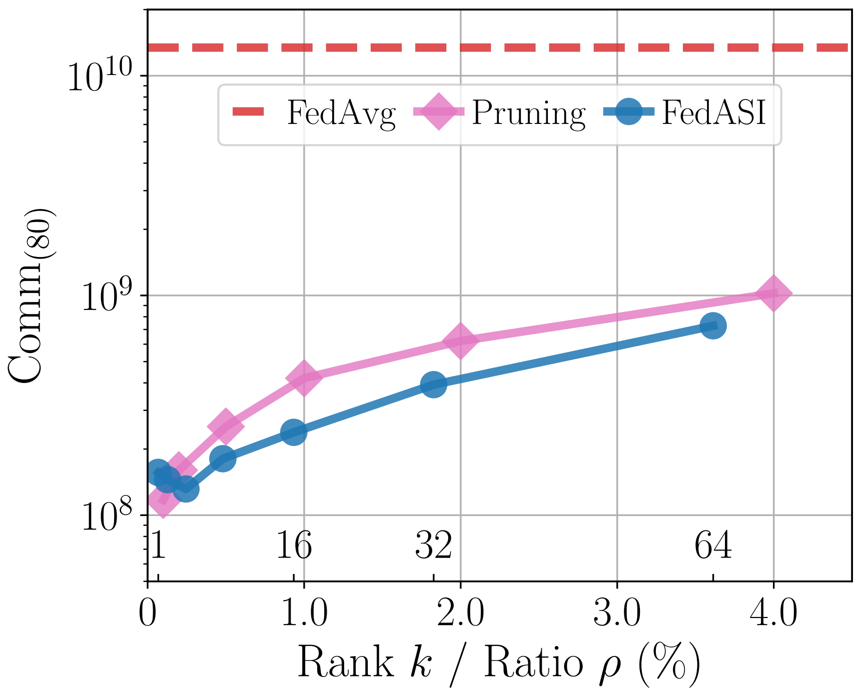
## FedASI Results

- Less than 0.5% accuracy loss




ResNet18 Accuracy vs. Rank

- Up to 1000x compression than FedAvg

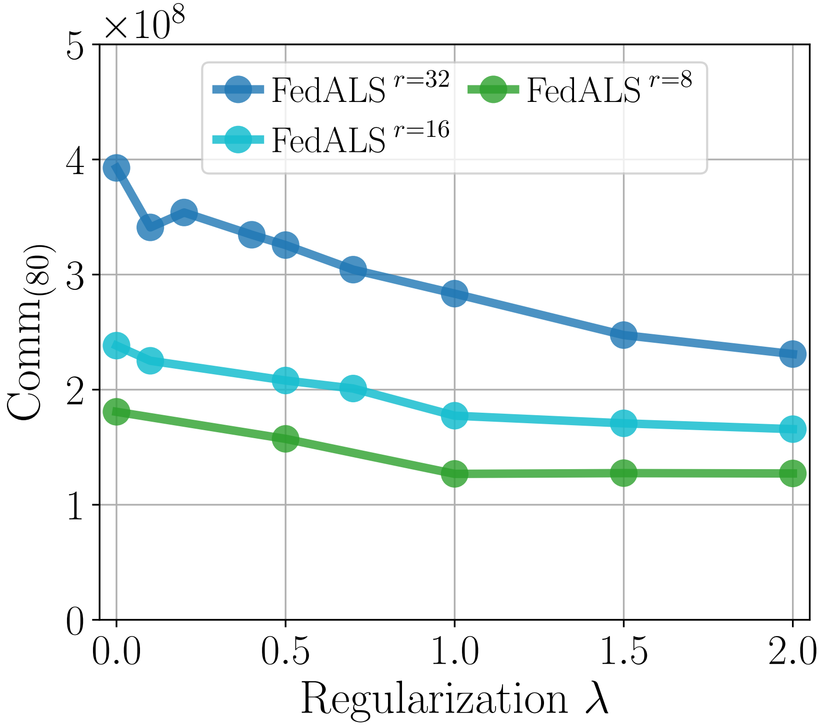



ResNet18 Converge Comm. vs. Rank

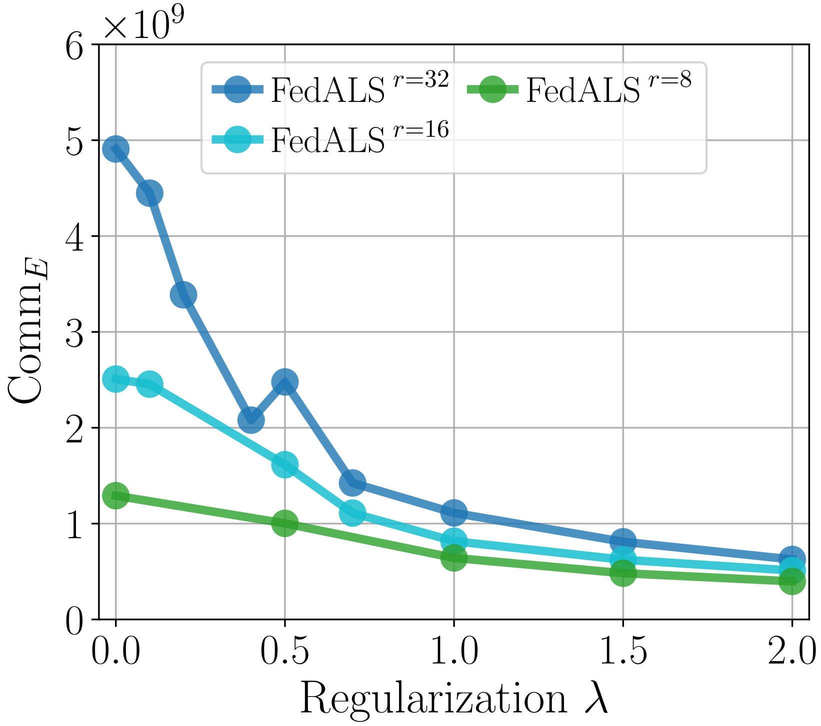
# Experiment on Efficiency

## FedALS Results

- Same accuracy and convergence
- Further 5x compression than FedASI




VGG16 Converge Comm. vs. Regularization

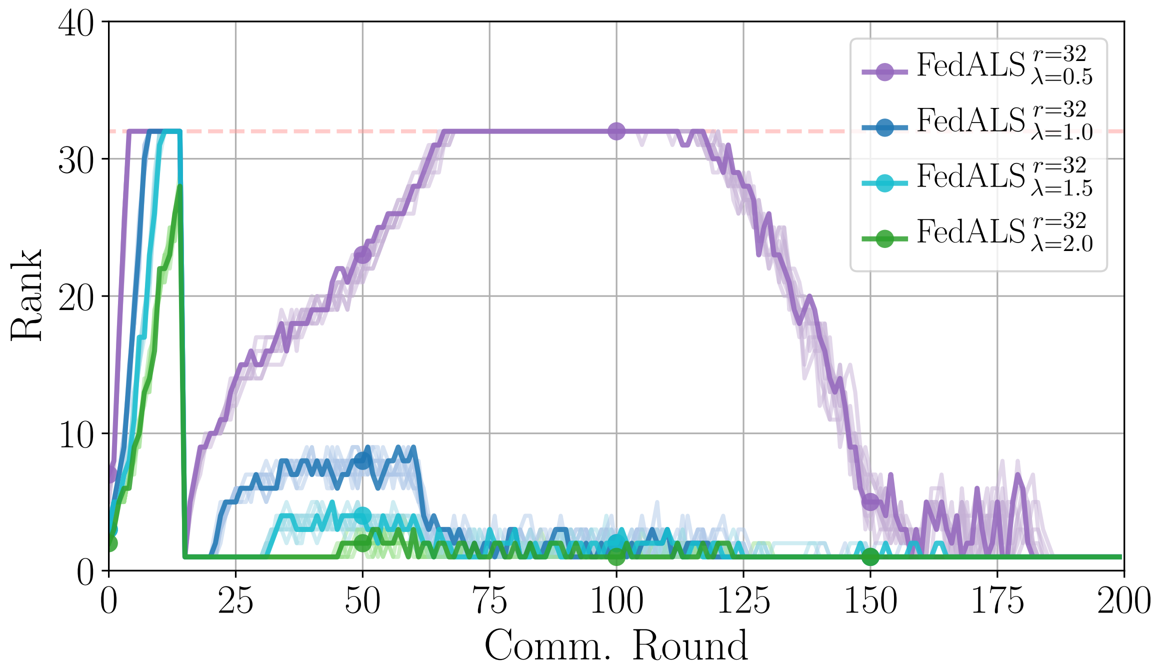



VGG16 Total Comm. vs. Regularization

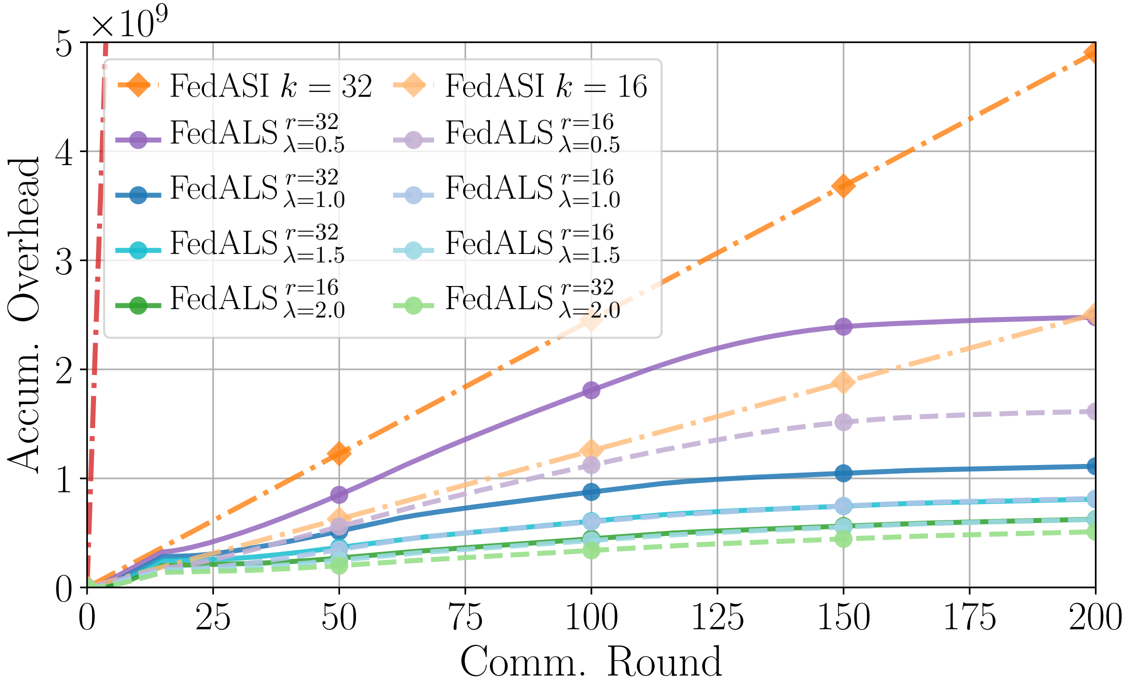
# Experiment on Efficiency

## FedALS Results

- Adaptive rank factorization




VGG16 Layer Rank vs. Round

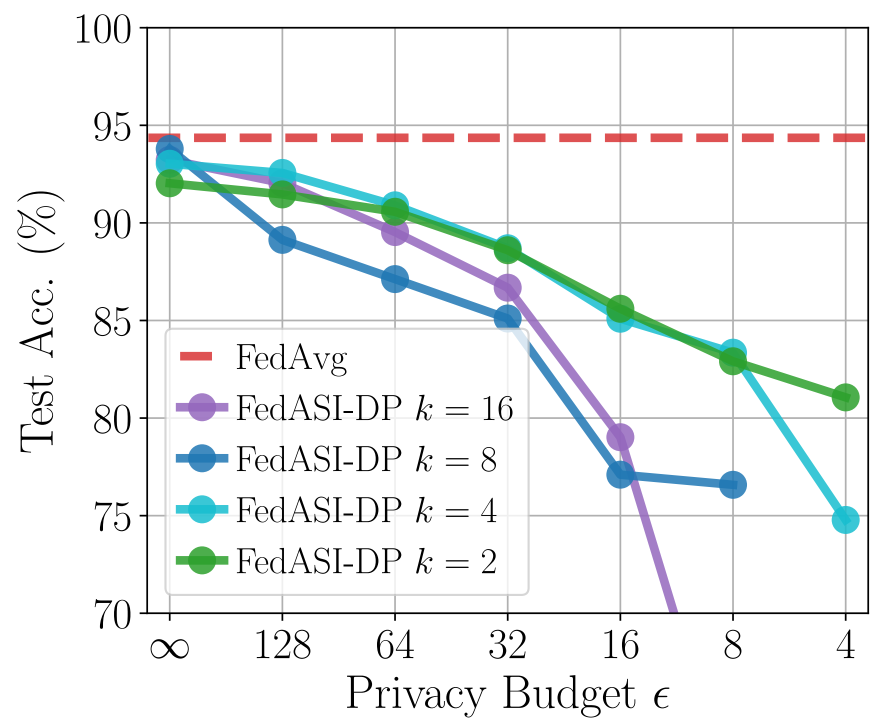



VGG16 Accum. Comm. Overhead vs. Round

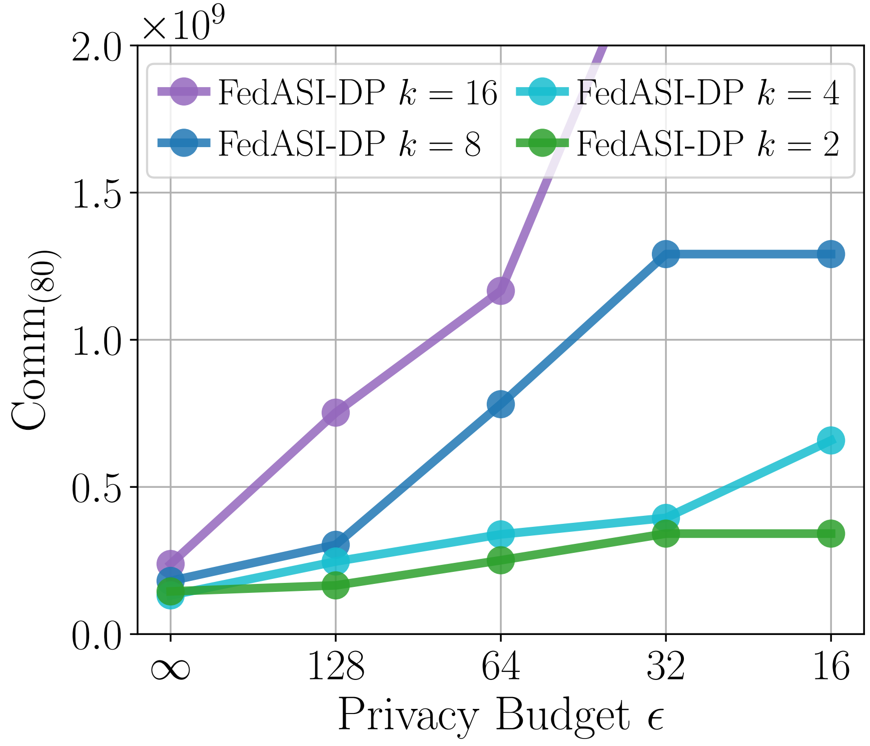
# Experiment on Privacy

## FedASI-DP Results

- Lower rank more robust to small $\epsilon$




ResNet18 Accuracy vs. Privacy Budget

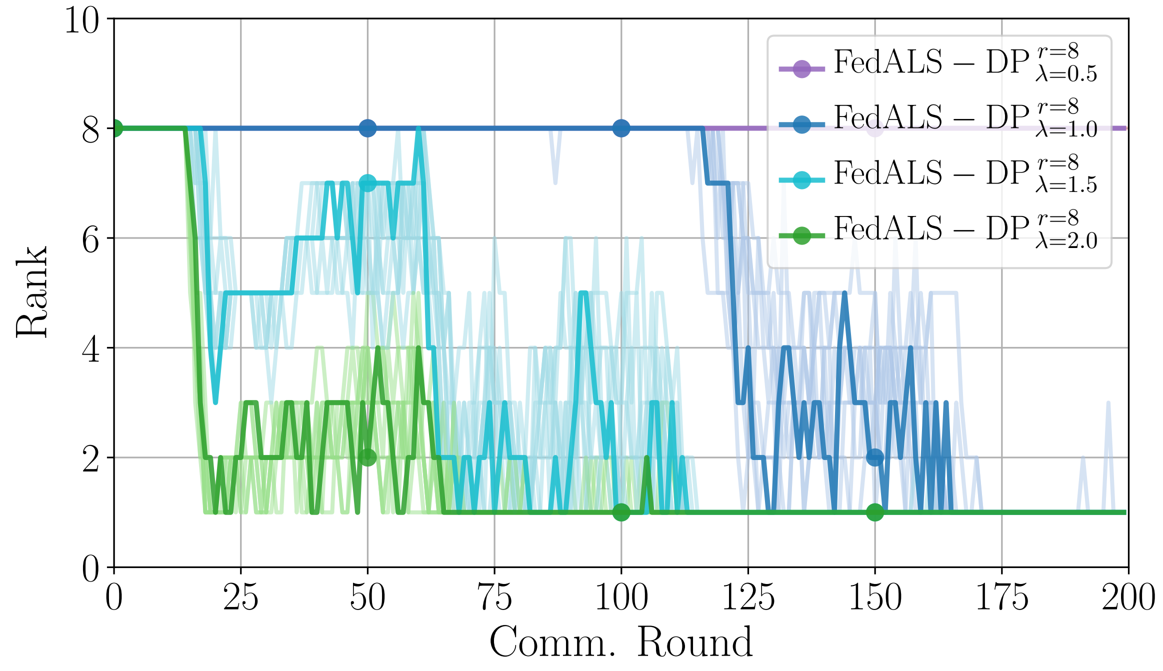



ResNet18 Converge Comm. vs. Privacy Budget

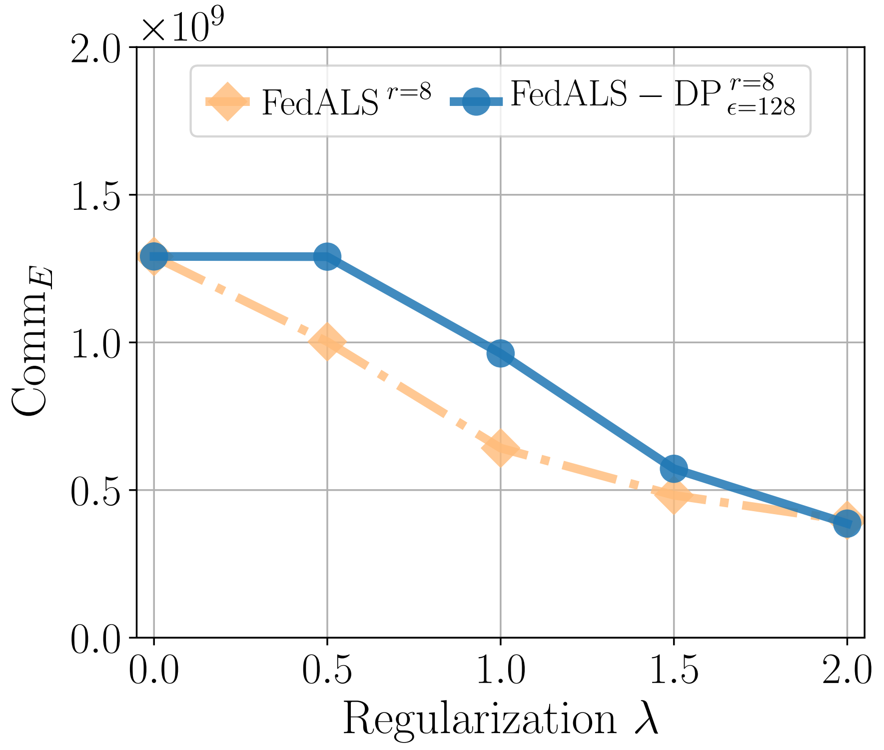
# Experiment on Privacy

## FedALS-DP Results

- Adaptive rank factorization




VGG16 Layer Rank vs. Round

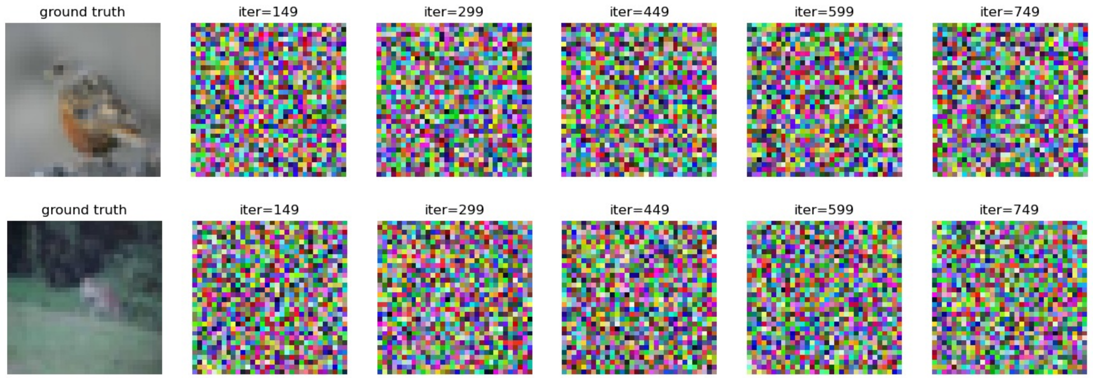



VGG16 Total Comm. vs. Regularization

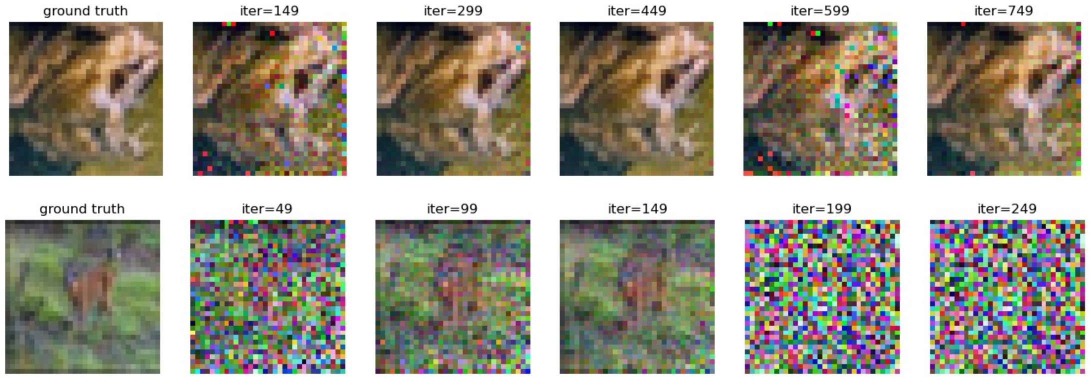
# Experiment on Privacy

## Defense to DLG Attack

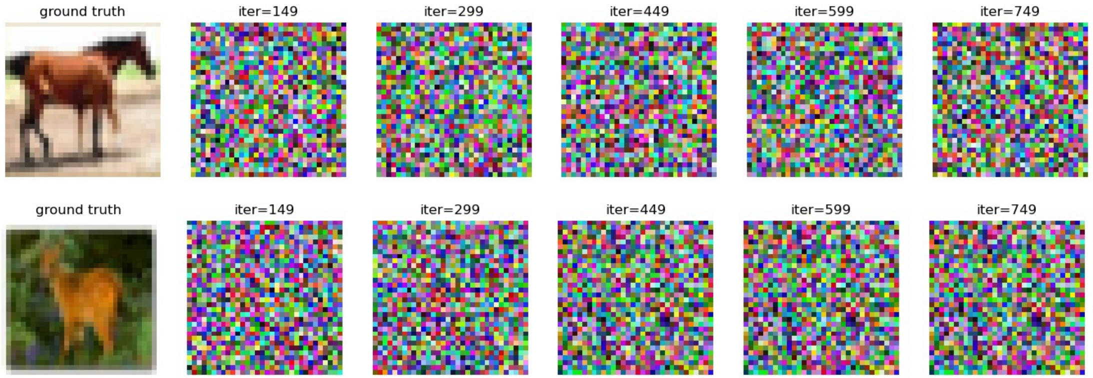
- Pruning: fail
- FedASI/FedALS: success
- FedASI-DP/FedALS-DP: success




truth DLG: FedASI-DP




truth DLG: Pruning




truth DLG: FedASI

# CONCLUSION

**FedASI & FedALS**

| | |
|---|---|
| Contribution | ▪ Adaptive low-rank factorization<br>▪ FL communication compression<br>▪ Differential privacy guarantee |
| Theory | ▪ Adaptive parameter compression by matrix rank optimization<br>▪ Differential privacy proof on algorithmic mechanism<br>▪ Convergence proof on low-rank factorization and FL training |
| Experiment | ▪ Up to 1000x compression without performance loss<br>▪ Benefit compression and privacy simultaneously |

THANK YOU