Federated Learning Framework with Personalized Model Compression and Privacy Protection

Chenlin Ding[1], Mingjun Xiao[1] , Yin Xu[1], and Jie Wu[2]

[1]University of Science and Technology of China
[2]Temple University

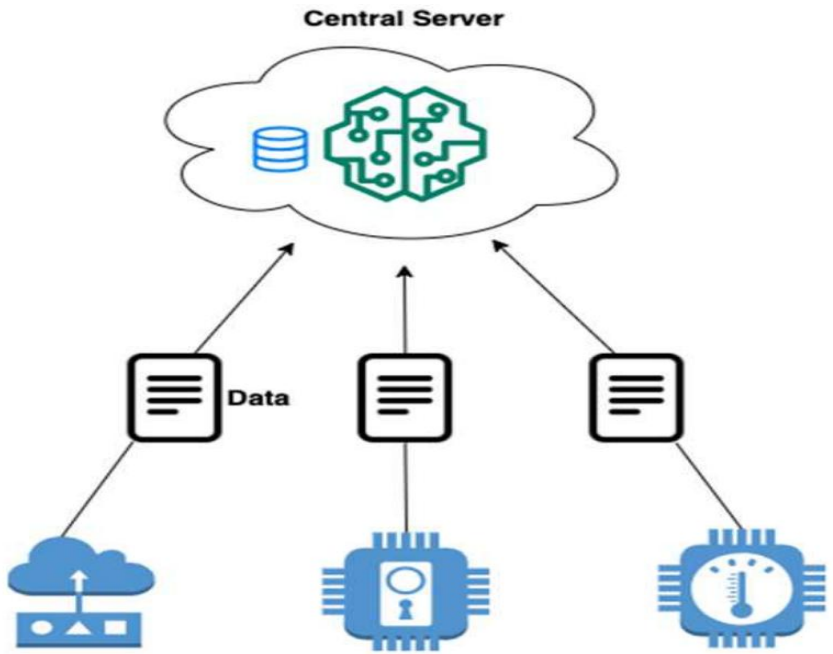
Outline

- **Background & Motivation**
- **Related Work & System Model**
- **PCM & OPM**
- **Evaluation & Conclusion**

Background

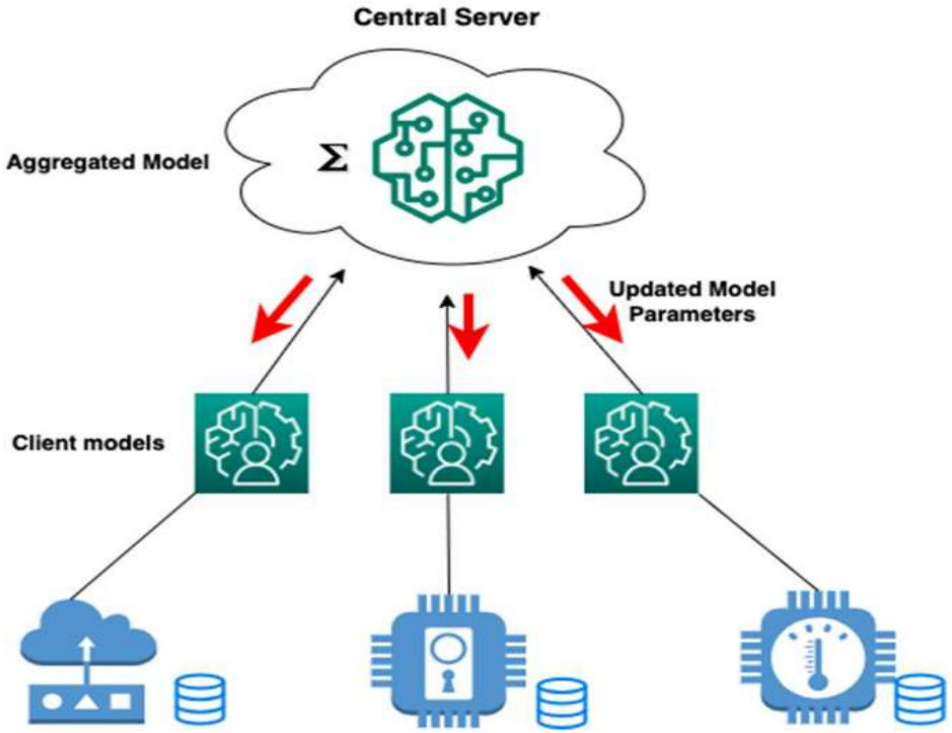
■Centralized learning

- High privacy risks
- High communication costs
- Not suitable for heterogeneous equipment



■Federated learning

- Protect data privacy
- Reduce communication overhead
- Adaptable to heterogeneous equipment



Motivation

Communication Overhead

- **High update cost:** Uploading gradients consumes significant bandwidth
- **Network heterogeneity:** Diverse client conditions hinder unified strategies

Privacy protection

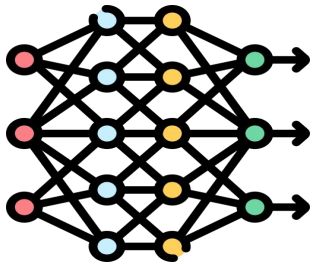
- **Security risks:** Updates must stay confidential despite potential attacks
- **Noise trade-off:** Noise ensures privacy but may reduce accuracy

Fewer model parameters

Noise disturbance

Reduce

Model Accuracy



Outline

- Background & Motivation
- Related Work & System Model
- PCM & OPM
- Evaluation & Conclusion

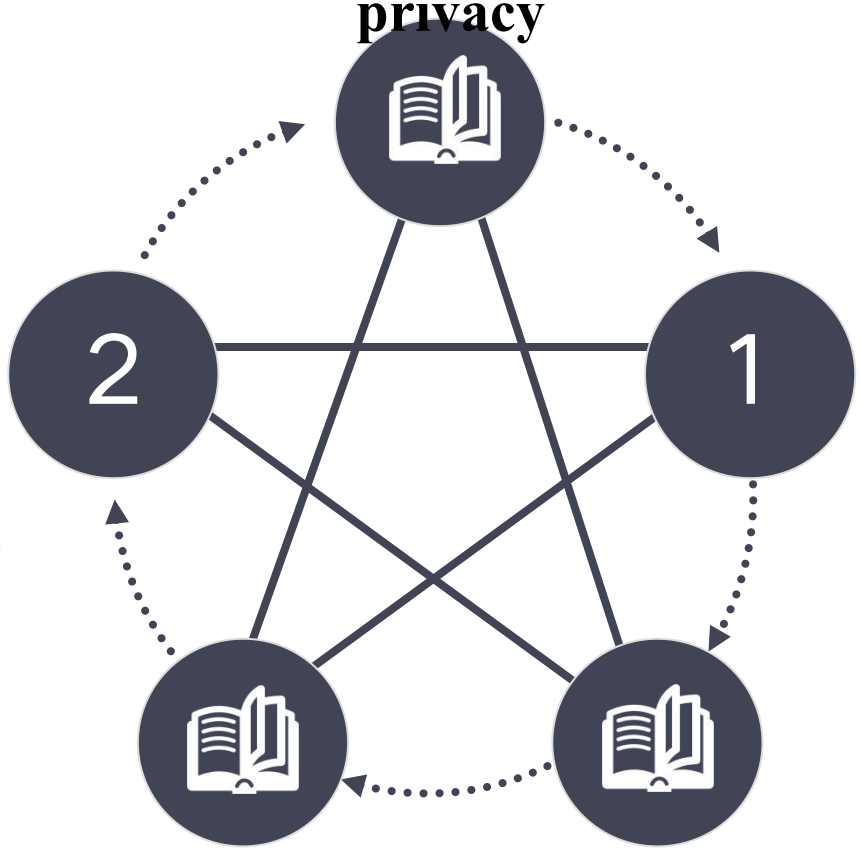
Related Work

Goal: To find the best balance between communication overhead and model accuracy while protecting user privacy

Privacy protection

[3] WANG W, LI Y, LIU T, et al. User-level privacy-preserving federated learning: Analysis and performance optimization[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 34. 2020: 7085-7092.

[4] HANG L, ZHU T, XIONG P, et al. A robust game-theoretical federated learning framework with joint differential privacy[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(4): 3333-3346.

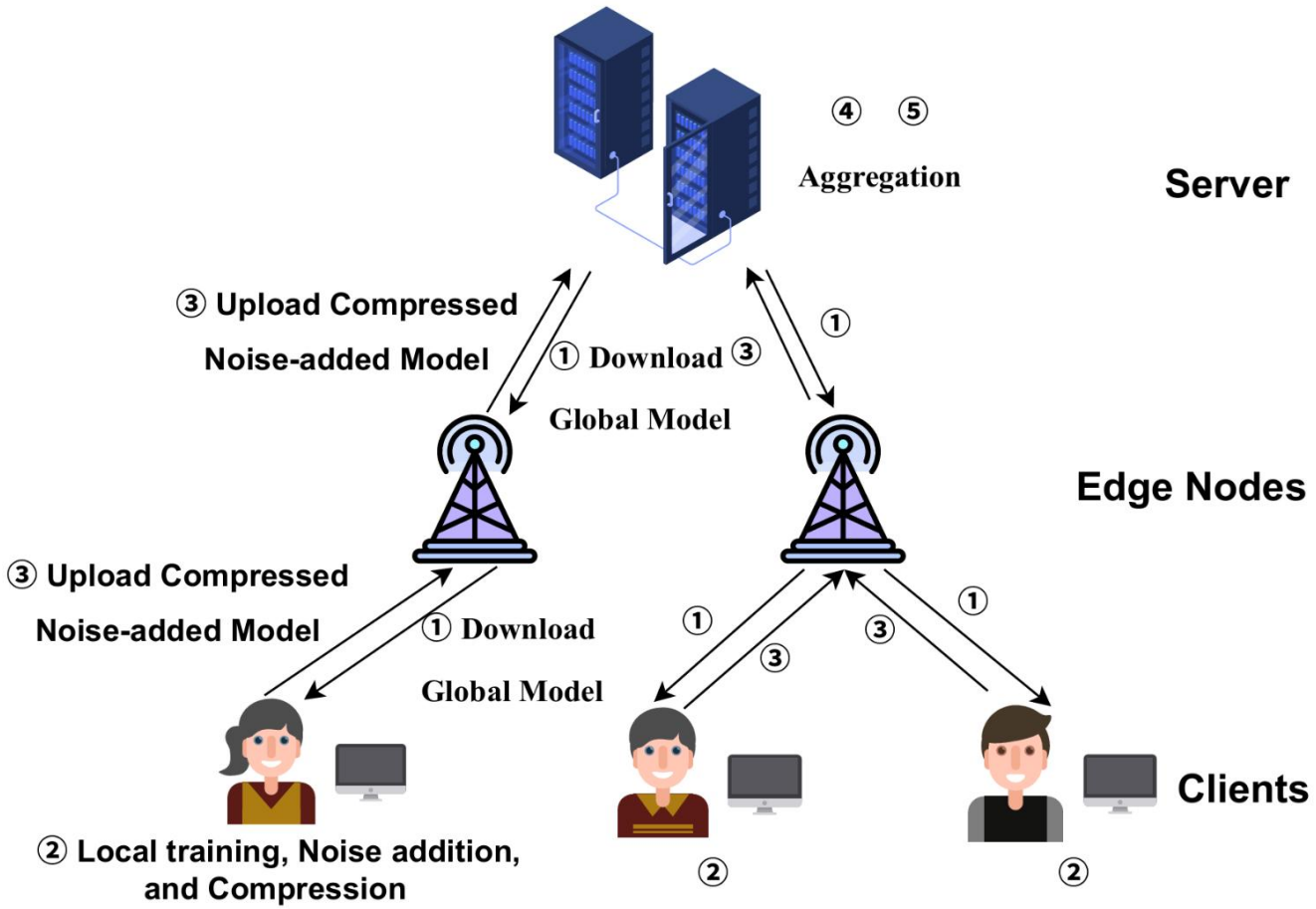


Communication Overhead

[1] ERNSTEIN J, WANG Y X, AZIZZADENESHELI K, et al. signsgd: Compressed optimisation for non-convex problems[C]//International Conference on Machine Learning. PMLR,2018: 560-569.

[2] IANG Z, XU Y, XU H, et al. Heterogeneity-aware federated learning with adaptive client selection and gradient compression[C]//IEEE INFOCOM 2023-IEEE Conference on Computer Communications. IEEE, 2023: 1-10.

System Model

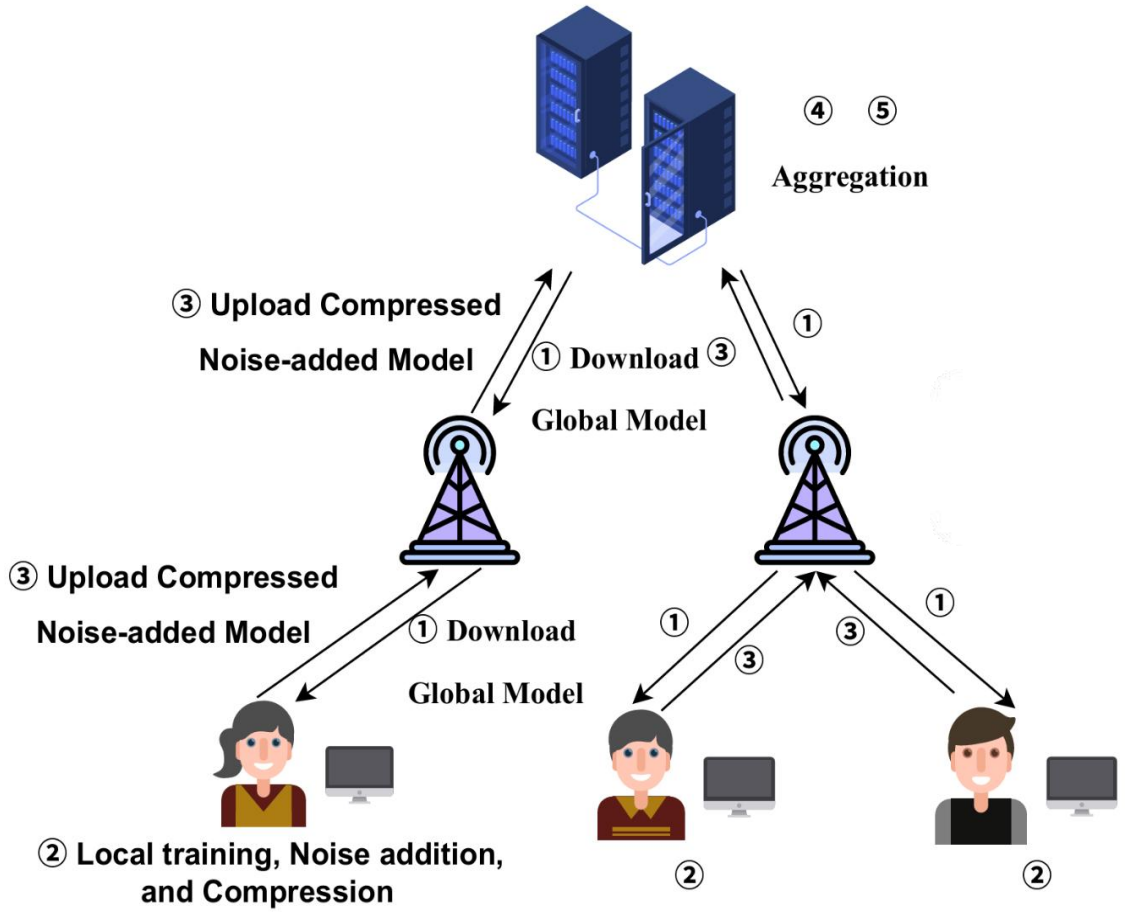


System Model

- **Global Model w_t**
- **Client i:** local dataset → local training
- **Edge Aggregation $\tilde{w}_j^t$**
- **Local Model w_i^t:** Client i's locally trained and compressed model
- **Noisy Model $\tilde{w}_i^t$:** Compressed model with added differential privacy noise
- **Compression Rate γ_i^t:** Dynamically adjusted compression ratio per client
- **Reward R_i^t, R^t:** Incentives for clients based on contribution and privacy

System Model

FedCP: Five Key Steps



1) Global Model Broadcast: The server broadcasts the current global model to selected clients.

2) Local Training & Model Processing: Each client trains the model on local data and applies personalized compression based on a two-stage Stackelberg game.

3) Optimized Privacy Noise Addition: Clients add noise to model updates using an optimized privacy mechanism.

4) Model Upload: Clients send the compressed and privacy-enhanced updates back to the server.

5) Global Aggregation: The server aggregates the received updates to form the new global model.

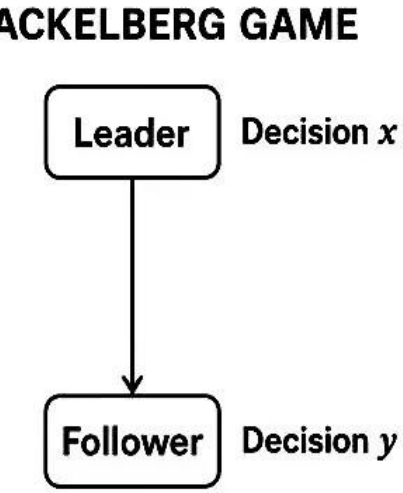
Outline

- Background & Motivation
- Related Work & System Model
- **PCM & OPM**
- Evaluation & Conclusion

Optimized Piecewise Noise mechanism

■ Optimization Objectives

- **Client-side objective:** To balance communication cost and economic reward.
- **Server-side objective**: To trade off between incentive cost and model convergence speed.



■ Implementation Steps

- **Step 1:** Design utility functions for both the server and clients to characterize their optimal decision-making behavior.
- **Step 2**: Derive the optimal strategy functions for the server and clients based on Stackelberg equilibrium.

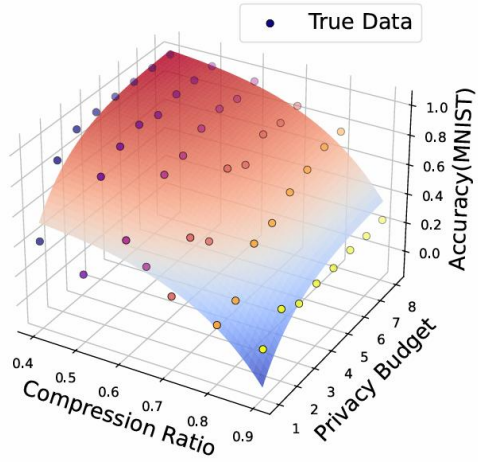
Optimized Piecewise Noise mechanism

■ Server Utility Design

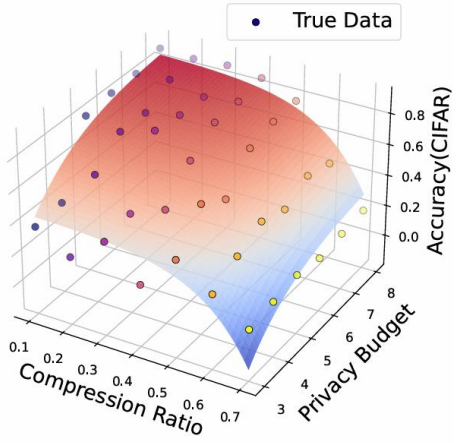
- **Utility factors:** Server utility depends on global model accuracy, client payments, and operational cost.
- **Server-side objective**: The local model accuracy of client i, denoted as Ω_i^t , can be approximately represented as a convex function. The server's utility function U^t can then be formulated as:

$$U^t(R^t, \gamma_i^t, \gamma_{-i}^t) = \Theta\Big(\sum_{i=1}^{N}\big(-\alpha e^{\lambda\gamma_i^t} - \mu e^{-\eta\epsilon_i} + \beta\big)\Big) - R^t - R_p^t$$

Model accuracy　Reward　Operating costs



(a) MNIST



(b) CIFAR

Optimized Piecewise Noise mechanism

■ Client Utility Design

➢ **Utility function** : The utility function $U_i^t(R^t, \gamma_i^t, \gamma_{(-i)}^t)$ of client i can be defined as follows:

$$U_i^t(R^t, \gamma_i^t, \gamma_{-i}^t) = R_i^t - c_i^{cm} - c_i^{pv} - c_i^{cp}$$

$$= \frac{\epsilon_i(1-\gamma_i^t)}{\sum_{i=1}^{N}\epsilon_i(1-\gamma_i^t)} R^t + \omega_1 \rho_i^{cm}(1-\gamma_i^t) + \omega_2 \rho_i^{pv}\epsilon_i - \omega_3 \rho_i^{cp}|D_i| f_i^2$$

The obtained reward | Communication costs | Privacy costs | Calculation costs

➢ **Definition:** $\omega_j (j \in \{1,2,3\})$ are positive adjustment coefficients.

Optimized Piecewise Noise mechanism

Determination of the optimal strategy group

Server

- Given a fixed payment amount R^t from the server, the optimal personalized compression rate for each client is determined:

$$\gamma_i^{t*} = \frac{(N-1)R^t[(N-1)\rho_i^{cm} - \epsilon_i \sum_{i=1}^{N}(\frac{\rho_i^{cm}}{\epsilon_i})]}{\omega_1 \epsilon_i^2 (\sum_{i=1}^{N}(\frac{\rho_i^{cm}}{\epsilon_i}))} + 1$$

- Once the leader determines the payment strategy R^t, the followers can directly compute and apply their optimal compression strategies.

γ_i^{t*}

R^{t*}

Client

- Then, based on the determined optimal compression rates γ_i^{t*}, the server's optimal payment R^{t*} satisfies the following equation:

$$\frac{\partial \Theta}{\partial \Omega_{sum}} \sum_{i=1}^{N} (a\lambda e^{\lambda \gamma_i^{t*}} \cdot \frac{\partial \gamma_i^{t*}}{\partial R^t}) + 1 = 0$$

- By applying the Newton-Raphson method, the server can efficiently solve for the optimal payment R^{t*} using the known values of γ_i^{t*} .

Globally communication-constrained scenario

■ Problem Formulation

➢ Under communication resource constraints, the two-stage Stackelberg game can be reformulated as:

$$\text{Server's side}: \text{Maximize } \mathcal{U}^t\left(R^t, \gamma_i^t, \gamma_{-\mathbf{i}}^{\mathbf{t}}\right),$$

$$\text{Client's side}: \text{Maximize } \mathcal{U}_i^t\left(R^t, \gamma_i^t, \gamma_{-\mathbf{i}}^{\mathbf{t}}\right),$$

$$\text{Subject to}: \sum_{i=1}^{N}\left(1-\gamma_i^t\right) \leq \mathcal{C}, 0 \leq \gamma_i^t < 1, t = \{1, 2, \ldots\}$$

■ Solution Approach

➢ Construct the Lagrangian function and use backward induction to derive the optimal strategies

➢ By introducing the parameter ϕ and a Lagrange multiplier ϖ under the global communication constraint C, the Lagrangian function for round t is formulated as:

$$L(R^t, \gamma_i^t, \gamma_{-i}^t, \varpi) \triangleq \mathrm{U}_i^t(R^t, \gamma_i^t, \gamma_{-i}^t) + \varpi(\phi C - \sum_{i=1}^{N} \epsilon_i(1-\gamma_i^t))$$

Globally communication-constrained scenario

➢ **KKT optimality conditions**

Condition 1

$$\left(\frac{\partial L}{\partial \gamma_i^t}\right)\Big|_{\gamma_i^t=\gamma_i^{t*}} = 0$$

Condition 2

$$\varpi \geq 0\,,$$
$$\varpi\left(\phi C - \sum_{i=1}^{N} \epsilon_i \left(1-\gamma_i^{t*}\right)\right) = 0$$

Condition 3

$$\sum_{i=1}^{N} \epsilon_i \left(1-\gamma_i^{t*}\right) \leq \phi C$$

➢ **Theorem:** Under any given payment amount R^t and communication resource cap C, the optimal personalized compression ratio γ_i^{t*} for each client i can be expressed as:

$$\gamma_i^{t*} = -\frac{\phi C}{\epsilon_i N} + \frac{\phi^2 C^2 \omega_1}{\epsilon_i R^t}\left(\frac{\rho_i^{cm}}{\epsilon_i} - \frac{K}{N}\right) + 1$$

$$K = \sum_{i=1}^{N} \frac{\rho_i^{cm}}{\epsilon_i}$$

Optimized Piecewise noise Mechanism

Data: Model parameters $\mathbf{w}_t^i$, central value C, maximum offset R, privacy parameter ϵ
Result: Perturbed model parameters $\tilde{\mathbf{w}}_t^i$

Initialize model parameters $\mathbf{w}_t^i$, central value C, and maximum offset R;
foreach *parameter* $\mathbf{w} \in \mathbf{w}_t^i$ **do**
  Compute offset $x = \mathbf{w} - C$;
  Compute noise modulation factors:
  - $L = \frac{(e^{\epsilon}+e^{\epsilon/3})(e^{\epsilon/3}+1)}{e^{\epsilon/3}(e^{\epsilon}-1)}$
  - $f(\epsilon, x) = \frac{(e^{\epsilon}+e^{\epsilon/3})(xe^{\epsilon/3}-R)}{e^{\epsilon/3}(e^{\epsilon}-1)}$
  - $g(\epsilon, x) = \frac{(e^{\epsilon}+e^{\epsilon/3})(xe^{\epsilon/3}+R)}{e^{\epsilon/3}(e^{\epsilon}-1)}$

  Randomly sample y from interval $[0, 1]$;
  if $y < 1/(e^{-2\epsilon/3}+1)$ **then**
    Randomly sample $\tilde{x}$ from interval $[f(\epsilon, x), g(\epsilon, x)]$;
  else
    Randomly sample $\tilde{x}$ from interval $[-RL, f(\epsilon, x)) \cup (g(\epsilon, x), RL]$;
  end
  Update perturbed parameter: $\tilde{\mathbf{w}} = C + \tilde{x}$;
end
return Perturbed model parameters $\tilde{\mathbf{w}}_t^i$;

These expressions ensure bounded variance and zero-mean noise under privacy guarantee.

High-probability region = Small, Centered
Low-probability region = Large noise.

After adding perturbation, parameter becomes $\tilde{x}$:

$$F[\tilde{x} = y \mid x] = \begin{cases} \dfrac{e^{\frac{4\epsilon}{3}}(e^{\epsilon}-1)}{2R\left(e^{\frac{\epsilon}{3}}+e^{\epsilon}\right)^2}, & y \in [f(\epsilon, x), g(\epsilon, x)] \\ \dfrac{e^{\frac{\epsilon}{3}}(e^{\epsilon}-1)}{2R\left(e^{\frac{\epsilon}{3}}+e^{\epsilon}\right)^2}, & y \in [-RL, f(\epsilon, x)) \cup (g(\epsilon, x), RL] \end{cases}$$

Outline

- Background & Motivation
- Related Work & System Model
- PCM & OPM
- Evaluation & Conclusion

Evaluation

Experimental Settings

- MNIST
- CIFAR10

Real Dataset

Compared Algorithms

- No Compression and No Privacy
- PM
- Laplace algorithm

- The total number of clients N is chosen from the range [40,120]
- Privacy budget is chosen from the range [1,10]
- $M = N/2$, $\omega_j = 1$ ($j \in \{1, 2, 3\}$), and $R_p^t = 600$

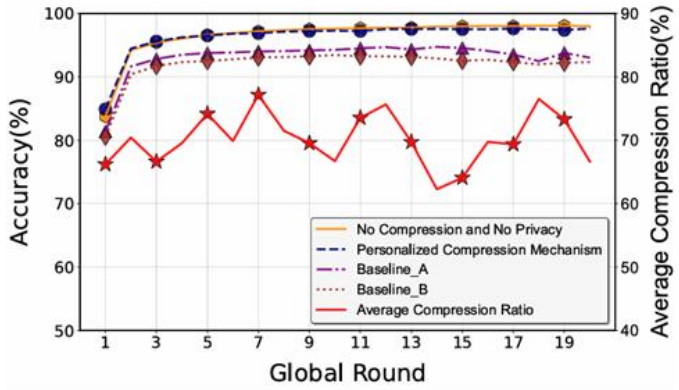
Parameter settings

Evaluation

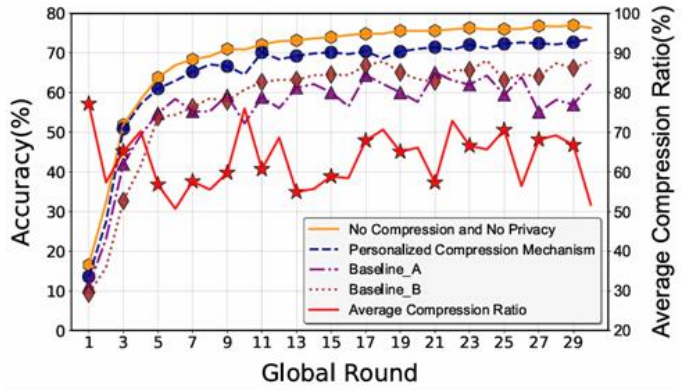
■ Simulation experiments of PCM

➢ Comparative Experiments

- Baseline_A: Uses average compression rate of all clients in the current round
- Baseline_B: Uses fixed global average compression rate across all rounds



(a) MNIST



(b) CIFAR

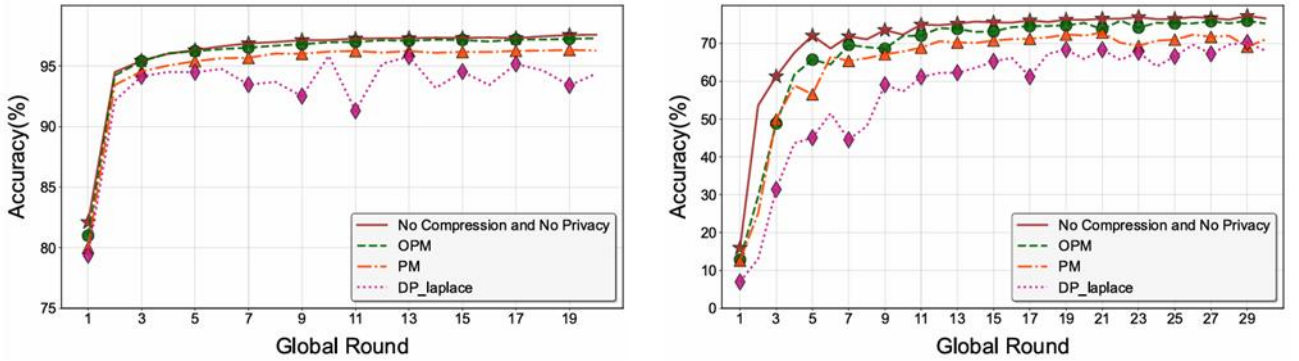
- Significant improvement in model performance and stability
- Achieves 70% avg. compression on MNIST, 63% on CIFAR-10
- Maintains accuracy with minimal loss

Evaluation

■ Simulation experiments of OPM

➢ Comparative Experiments (with compression strategy uniformly using PCM)

- PM: A piecewise noise injection method based on standard perturbation techniques for differential privacy.
- Laplace Mechanism: A privacy-preserving method based on Laplace noise addition.

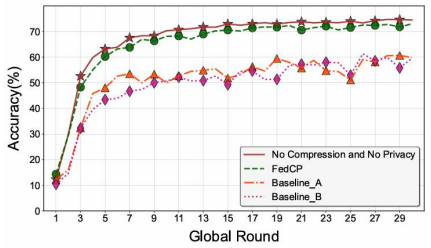


(a) MNIST (b) CIFAR

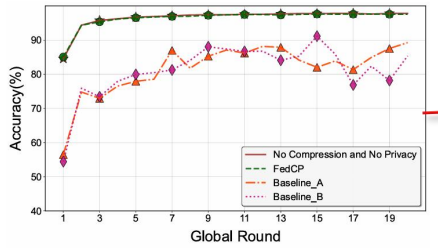
- Higher accuracy & faster convergence
- Noise has lower variance, stays closer to original data
- Reduces information loss, preserves model quality

Evaluation

■ Simulation experiments of FedCP

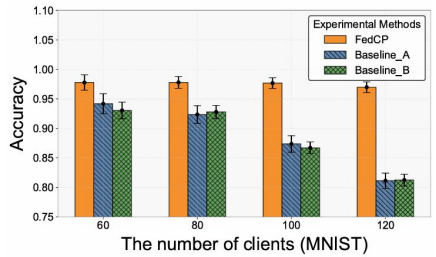


(a) MNIST

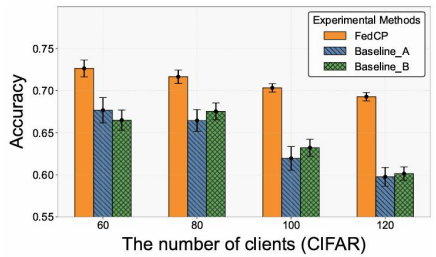


(b) CIFAR

Iid data

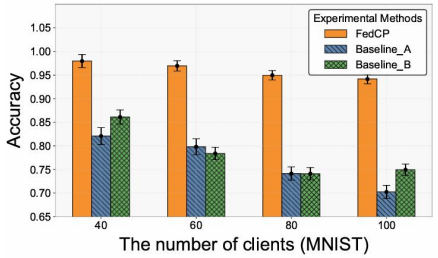


(a) MNIST

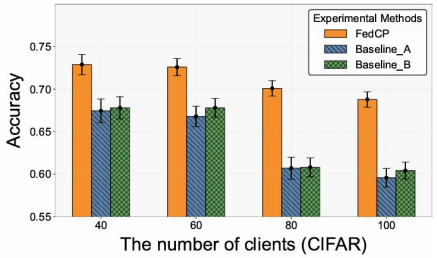


(b) CIFAR

Non-iid data



(a) MNIST

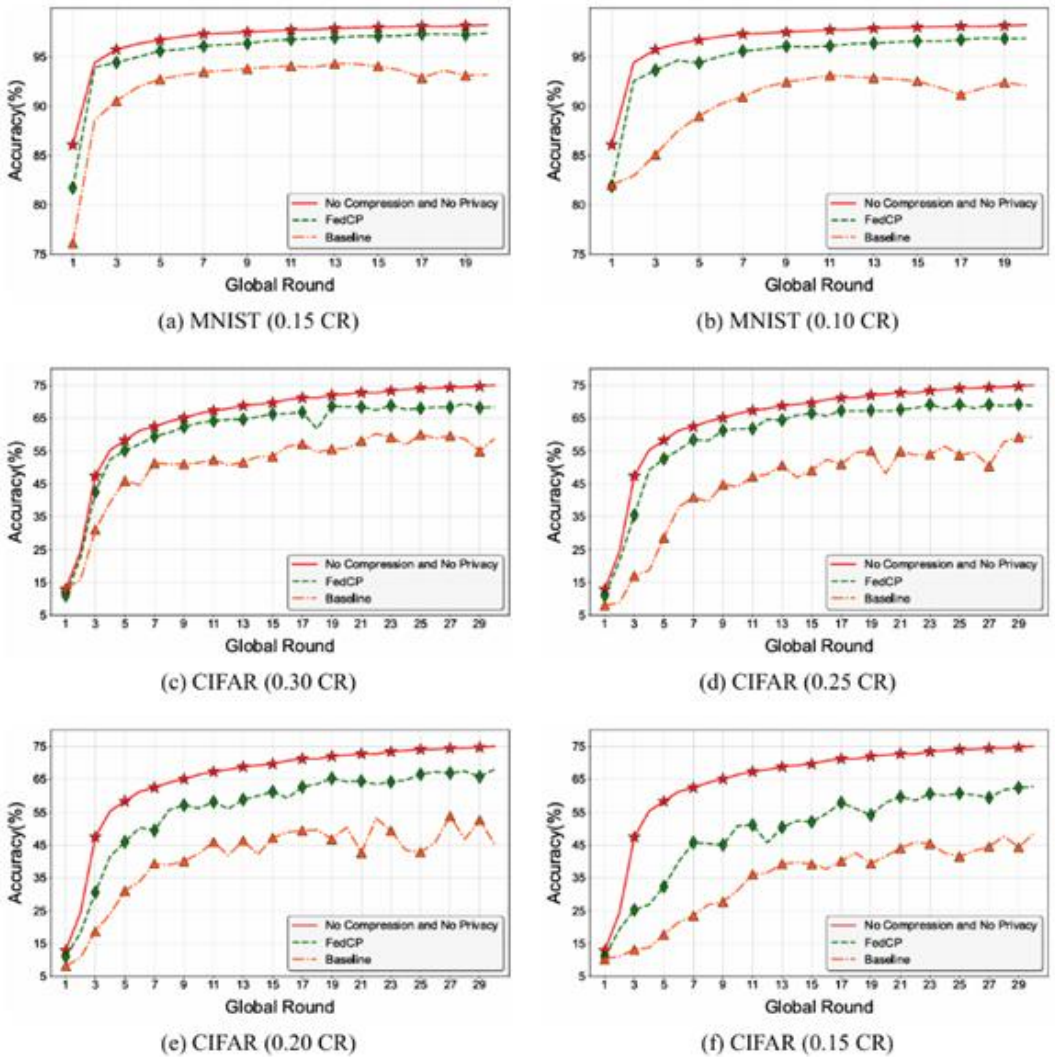


(b) CIFAR

- FedCP improves efficiency with minimal accuracy loss
- Performance gap with baseline widens as clients increase
- Larger gap under non-IID, showing robustness & adaptability

Evaluation

■ Simulation experiments of FedCP



(a) MNIST (0.15 CR) (b) MNIST (0.10 CR)

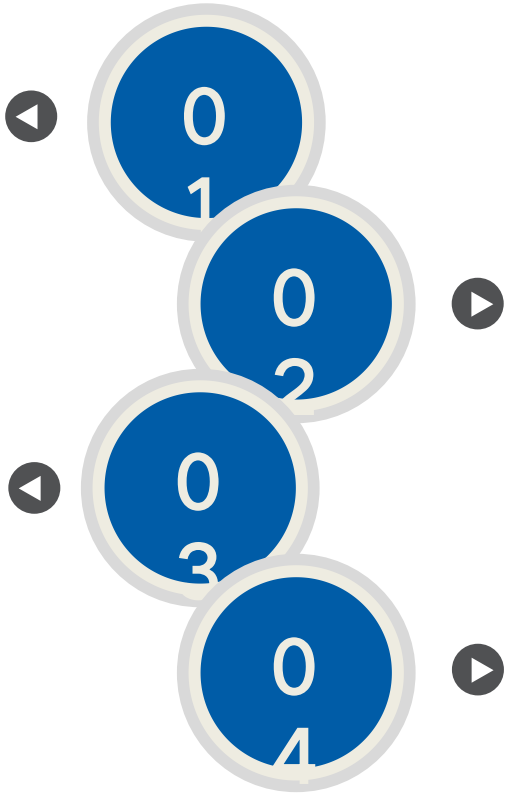
(c) CIFAR (0.30 CR) (d) CIFAR (0.25 CR)

(e) CIFAR (0.20 CR) (f) CIFAR (0.15 CR)

- Baseline: All clients use the same compression rate; PM mechanism is used for privacy.
- Higher final accuracy and faster convergence vs. baseline.
- Tighter communication constraints widen the performance gap.

Conclusion

- ✓ Address the dual challenges in federated learning: protecting user privacy, maintaining model accuracy, and reducing communication costs.
- ✓ Proposed an optimized piecewise differential privacy method for larger privacy budgets, which controls global model error while ensuring performance.



- ✓ Designed using a two-stage Stackelberg game model, allowing each client to adjust compression ratio based on local accuracy, bandwidth, and privacy needs for optimal balance.
- ✓ FedCP enables high-accuracy model training under extremely low communication overhead, demonstrating the efficiency and robustness of the framework across various scenarios.

中国科学技术大学
University of Science and Technology of China

Question?