# Personalized Federated Learning with Parameter Propagation


**Jun Wu[1], Wenxuan Bao[1], Elizabeth Ainsworth[1,2], Jingrui He[1]**

[1]University of Illinois at Urbana-Champaign, [2]USDA ARS Global Change and Photosynthesis Research Unit

junwu3@illinois.edu, wbao3@illinois.eu, ainswort@illinois.edu, jingrui@illinois.edu

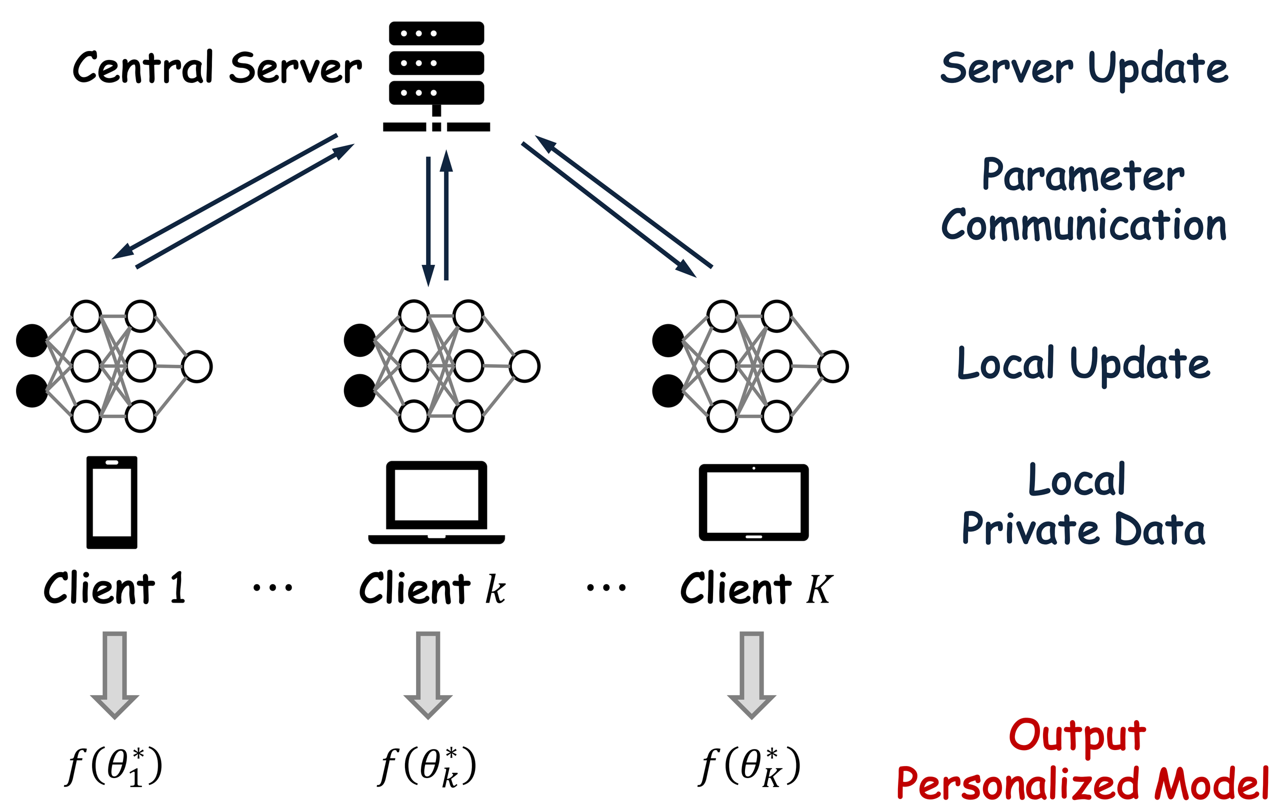

KDD2023 AUGUST 6 - 10 LONG BEACH, CA

AIFARMS Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability

## Problem Definition

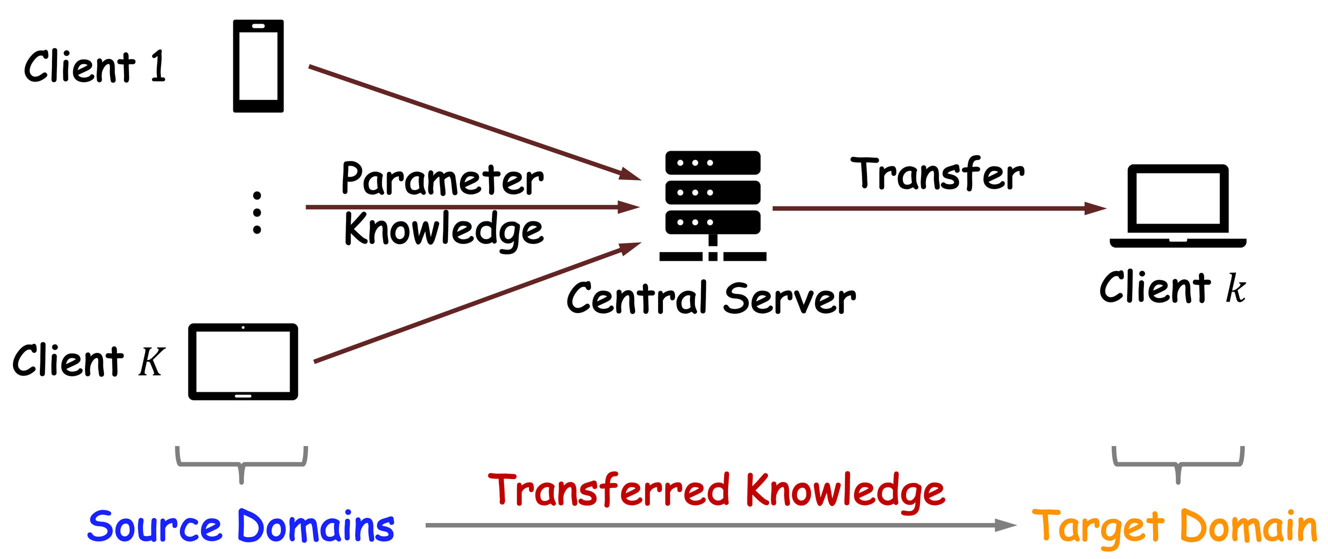
### ❑ Personalized Federated Learning

- Input: (i) a set of private clients ; (ii) a learning algorithm $f(\cdot)$
- Goal: Learn a personalized model for each client

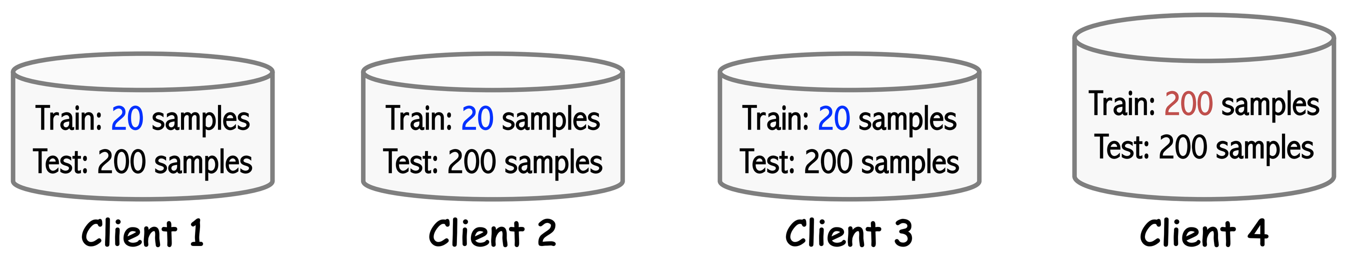



## A Transfer Learning Perspective

### ❑ Knowledge Transfer across Clients




### ❑ Negative Transfer




(a) Heterogeneous clients with imbalanced training samples

| Model | Accuracy | | | | Average Accuracy |
|---|---|---|---|---|---|
| | Client 1 | Client 2 | Client 3 | Client 4 | |
| LOCAL | 0.5270 | 0.4840 | 0.4980 | 0.8110 | 0.5800 |
| FedAvg | 0.3755 | 0.4420 | 0.6455 | 0.7965 | 0.5649 |
| LG-FedAvg | 0.5440 | 0.5115 | 0.5430 | 0.8095 | 0.6020 |
| Ditto | 0.4095 | 0.4810 | 0.6465 | 0.8095 | 0.5866 |
| FedAMP | 0.5300 | 0.5210 | 0.5415 | 0.8105 | 0.6008 |

(b) Results of personalized federated learning

## Federated Parameter Propagation (FEDORA)

### ❑ Overall Objective Function

$$\min_{\theta_k,\hat{\theta}_k} \underbrace{\sum_{k=1}^{K}\frac{1}{\lambda_k n_k}\sum_{i=1}^{n_k}\ell(x_i^k,y_i^k;\theta_k)}_{①} + \underbrace{\sum_{k=1}^{K}\left\|\theta_k-\hat{\theta}_k\right\|_2^2}_{②} + \underbrace{\frac{\alpha}{2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\frac{w_{kk'}}{D_{kk}}\left\|\hat{\theta}_k-\hat{\theta}_{k'}\right\|_2^2}_{③}$$

① **Class Membership**: Each client updates its local parameters $\theta_k$ w.r.t. private data

② **Consistency**: Each client approximates the received auxiliary parameters $\hat{\theta}_k$

③ **Parameter Smoothness**: Two clients share similar auxiliary parameters, if they are distributionally similar

### ❑ Overall Training Procedures

- **Step 0 (Pre-processing):** Estimate distribution similarity $w_{kk'}$

(i) Truncated SVD: $X_k = U_k\Sigma_k V_k^T$ (Client Update)

(ii) Principal Angles: (Server Update)

$$\zeta_1^{kk'} = \min_{a_1\in\mathcal{U}_k,\, b_1\in\mathcal{U}_{k'}} \arccos\left(\frac{\langle a_1,b_1\rangle}{\|a_1\|\cdot\|b_1\|}\right)$$

$$\vdots$$

$$\zeta_p^{kk'} = \min_{\substack{a_1\in\mathcal{U}_k,\, b_1\in\mathcal{U}_{k'} \\ a_p\perp a_1,\cdots,a_{p-1} \\ b_p\perp b_1,\cdots,b_{p-1}}} \arccos\left(\frac{\langle a_1,b_1\rangle}{\|a_1\|\cdot\|b_1\|}\right)$$

(iii) Similarity: $w_{kk'} = \sum_{i=1}^{p}\cos\zeta_i^{kk'}$ (Server Update)

- **Step 1 (Client Update):** Update personalized parameters $\theta_k$

$$\min_{\theta_k}\frac{1}{n_k}\sum_{i=1}^{n_k}\ell(x_i^k,y_i^k;\theta_k)+\lambda_k\left\|\theta_k-\hat{\theta}_k\right\|_2^2 \quad \text{(Fix } \hat{\theta}_k\text{, update } \theta_k\text{)}$$

  - ➢ Selective regularization: $\lambda_k = \max\left(\epsilon, \ell_k\left(\theta_k; D_k^{val}\right) - \ell_k\left(\hat{\theta}_k; D_k^{val}\right)\right)$

- **Step 2 (Communication):** Upload personalized parameters $\theta_k$ to the central server

- **Step 3 (Server Update):** Update auxiliary parameters $\{\hat{\theta}_k\}_{k=1}^K$

$$\min_{\{\hat{\theta}_k\}_{k=1}^K}\sum_{k=1}^{K}\left\|\theta_k-\hat{\theta}_k\right\|_2^2 + \frac{\alpha}{2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\frac{w_{kk'}}{D_{kk}}\left\|\hat{\theta}_k-\hat{\theta}_{k'}\right\|_2^2 \quad \text{(Fix } \theta_k\text{, update } \hat{\theta}_k\text{)}$$

  - ➢ Adaptive parameter propagation:

$$\hat{\theta}_k^{(m)} = \frac{\alpha}{(1+\alpha)D_{kk}}\sum_{k'=1}^{K} w_{kk'}\hat{\theta}_{k'}^{(m-1)} + \frac{1}{1+\alpha}\theta_k$$

- **Step 4 (Communication):** Sent auxiliary parameters $\hat{\theta}_k$ back to client $k$

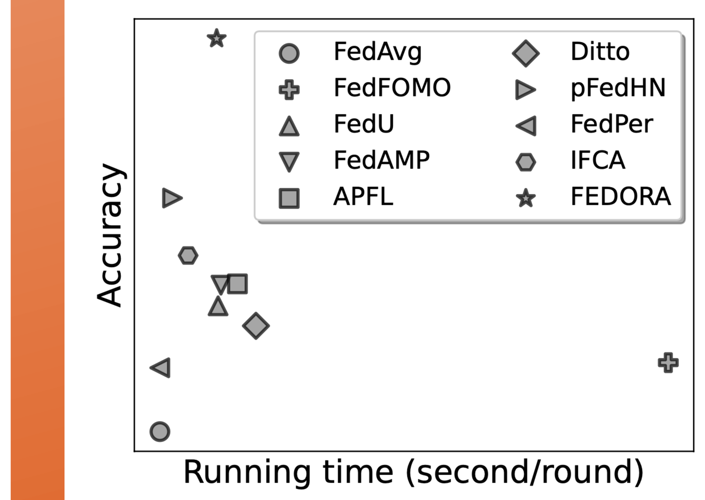
## Evaluation

### ❑ Performance Evaluation

| Model | Rotated MNIST | | | Rotated Fashion-MNIST | | | CIFAR-10 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc ↑ | R-Acc ↑ | PTR ↑ | Acc ↑ | R-Acc ↑ | PTR ↑ | Acc ↑ | R-Acc ↑ | PTR ↑ |
| LOCAL | 0.7642 | - | - | 0.7057 | - | - | 0.7617 | - | - |
| FedAvg [25] | 0.6889 | -0.0976 | 0 | 0.6441 | -0.0847 | 0.1250 | 0.6531 | -0.1382 | 0.3000 |
| FedAvg+FT | 0.7411 | -0.0293 | 0.3056 | 0.6848 | -0.0283 | 0.3472 | 0.7992 | 0.0513 | 0.9000 |
| FedProx [21] | 0.5375 | -0.2962 | 0 | 0.5968 | -0.1521 | 0 | 0.6984 | -0.0799 | 0.2000 |
| FedProx+FT | 0.6893 | -0.0973 | 0.0278 | 0.6788 | -0.0358 | 0.3056 | 0.7953 | 0.0460 | 0.9000 |
| LG-FedAvg [23] | 0.7804 | 0.0214 | 0.9444 | 0.7137 | 0.0115 | 0.7361 | 0.7656 | 0.0054 | 0.8000 |
| FedPer [1] | 0.7741 | 0.0135 | 0.6389 | 0.6725 | -0.0457 | 0.1389 | 0.8352 | 0.0990 | **1.0000** |
| pFedHN [33] | 0.8004 | 0.0486 | 0.8611 | 0.7215 | 0.0249 | 0.6944 | 0.7766 | 0.0221 | 0.6000 |
| APFL [6] | 0.7871 | 0.0303 | 0.8889 | 0.7134 | 0.0112 | 0.7639 | 0.8258 | 0.0866 | 0.9000 |
| Ditto [20] | 0.7806 | 0.0220 | 0.7222 | 0.7212 | 0.0232 | 0.7361 | 0.8078 | 0.0630 | 0.9000 |
| IFCA [9] | 0.7915 | 0.0365 | 0.6944 | 0.7305 | 0.0370 | 0.7639 | 0.8227 | 0.0828 | 0.9000 |
| FeSEM [46] | 0.7720 | 0.0110 | 0.6111 | 0.7074 | 0.0051 | 0.5278 | 0.8547 | 0.1255 | **1.0000** |
| FedFOMO [47] | 0.7749 | 0.0140 | 0.9167 | 0.7110 | 0.0076 | 0.7639 | 0.8242 | 0.0797 | **1.0000** |
| FedU [7] | 0.7837 | 0.0260 | 0.8889 | 0.7208 | 0.0225 | 0.8056 | 0.7836 | 0.0295 | 0.9000 |
| FedAMP [13] | 0.7869 | 0.0298 | **1.0000** | 0.7203 | 0.0213 | 0.8056 | 0.7953 | 0.0457 | 0.8000 |
| **FEDORA** | **0.8251** | **0.0806** | **1.0000** | **0.7433** | **0.0548** | **0.9028** | **0.8570** | **0.1288** | **1.0000** |

**Acc: accuracy, R-Acc: Relative accuracy, PTR: positive transferability ratio**

### ❑ Efficiency & Communication Cost




| Model | Cost | # params |
|---|---|---|
| FedAvg | $2KRd_\theta$ | 118,282,000 |
| FEDORA | $2KRd_\theta + Kpd_{in}$ | 118,282,784 |

Communication costs on Rotated MNIST

## Conclusion

❑ **Problem**: Analyze personalized federated learning from a transfer learning perspective

❑ **Algorithm**: Propose federated parameter propagation (FEDORA) with adaptive parameter propagation and selective regularization

❑ **Evaluation**: Demonstrate the effectiveness and efficiency of FEDORA in mitigating negative transfer

## References


[1] Brendan McMahan, et al. "Communication-efficient learning of deep networks from decentralized data." AISTATS 2017.

[2] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. "Ditto: Fair and robust federated learning through personalization." ICML 2021.

[3] Yutao Huang, et al. "Personalized cross-silo federated learning on non-IID data." AAAI 2021.


ILLINOIS