University of Pittsburgh

UNIVERSITY OF CENTRAL FLORIDA

ICCV23 PARIS

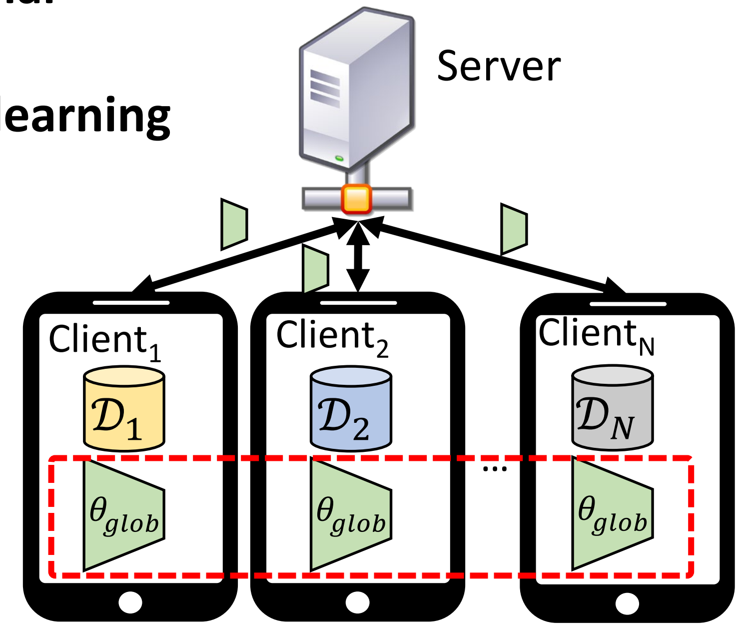
PGFed: Personalize Each Client's Global Objective for Federated Learning

Jun Luo[†], Matias Mendieta[‡], Chen Chen[‡], and Shandong Wu[†]

[†] University of Pittsburgh, Pittsburgh, PA, USA
[‡] University of Central Florida, Orlando, FL, USA

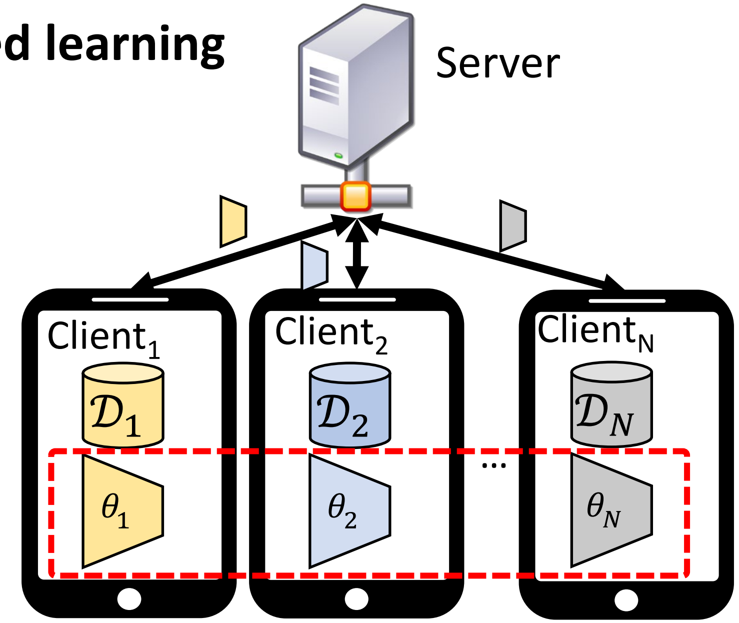
Background

Conventional (Global) federated learning



$$\min_{\theta_{glob}} F(\theta_{glob}) = \min_{\theta_{glob}} \sum_i p_i F_i(\theta_{glob})$$

Personalized federated learning



$$\min_{\Theta} F(\Theta) = \min_{\theta_1,\theta_2,\ldots,\theta_N} \sum_i p_i F_i(\theta_i)$$

Background

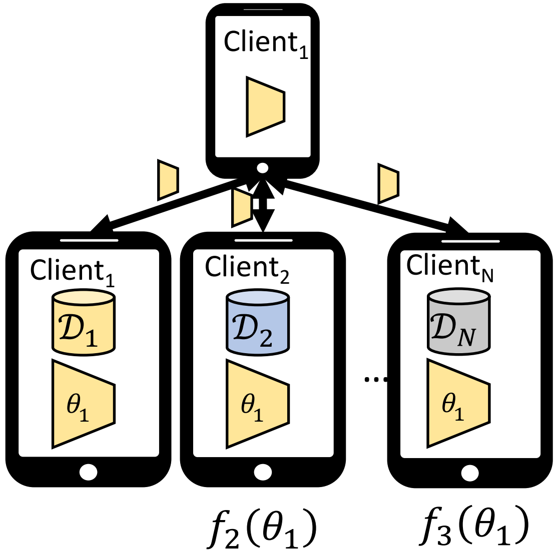
- In existing personalized FL (pFL) algorithms (with heterogeneous data), *the way in which the* ***collaborative knowledge*** *transfers from the server to the clients is* ***implicit*****.**

  - **Collaborative knowledge**: non-local information
    - E.g., Global FL's objective: $F(\theta_{glob}) = \sum_i p_i F_i(\theta_{glob})$

  - **Explicitness** (as opposite of **implicitness**): Direct engagement of multiple clients' empirical risks (explicit since not embed non-local info into model weights or regularization)
    - E.g., Global FL's objective: $F(\theta_{glob}) = \sum_i p_i F_i(\theta_{glob})$ where $F_i(\theta_{glob}) = f_i(\theta_{glob})$

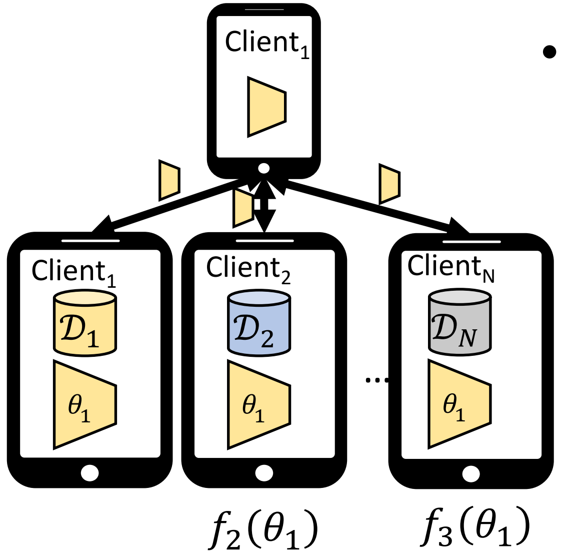
Motivation

- Why explicit (especially for personalized model update)?
  - (**Explicitness**: Direct engagement of multiple clients' empirical risks)
  - Intuition/motivation: facilitate the generalizability of θ_i directly by penalizing its performance over other clients' empirical risks.

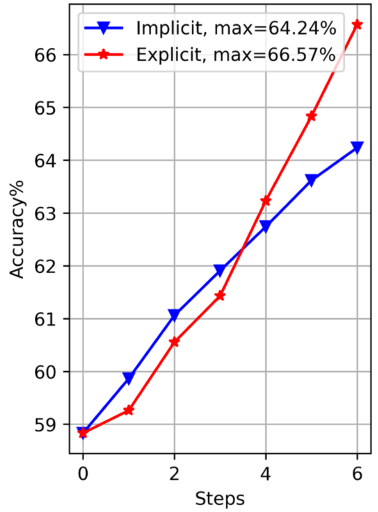


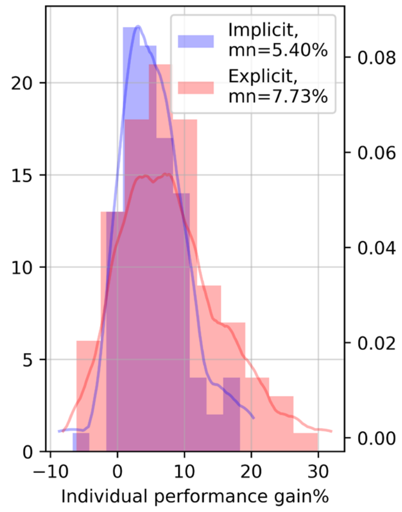
Motivation

- Why explicit (especially for personalized model update)?
  - (**Explicitness**: Direct engagement of multiple clients' empirical risks)
  - Intuition/motivation: facilitate the generalizability of θ_i directly by penalizing its performance over other clients' empirical risks.



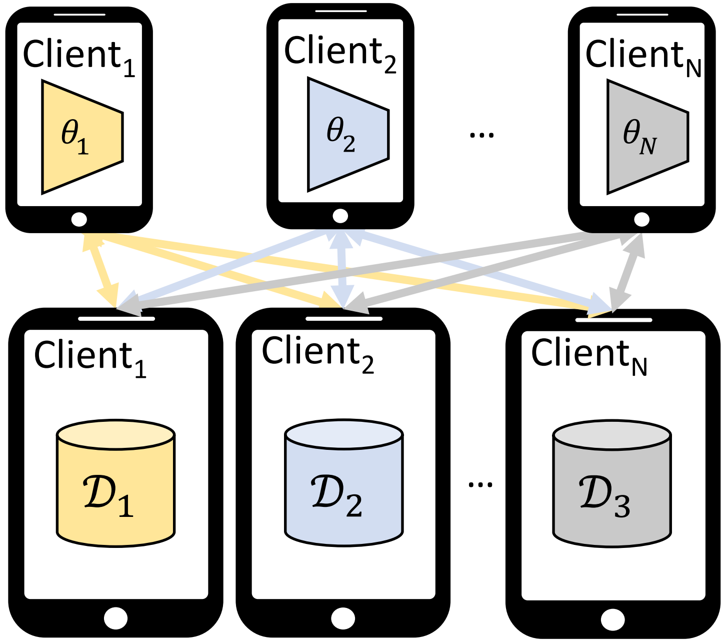
- Toy experiment on exemplar design
  - Cifar10, 100 heterogeneous clients
  - Explicit: $F_i(\theta_i) = f_i(\theta_i) + \frac{\mu}{N-1}\sum_{j\neq i} f_j(\theta_i)$
  - Implicit: $F_i(\theta_i) = f_i(\theta_i)$ (local model of FedAvg)





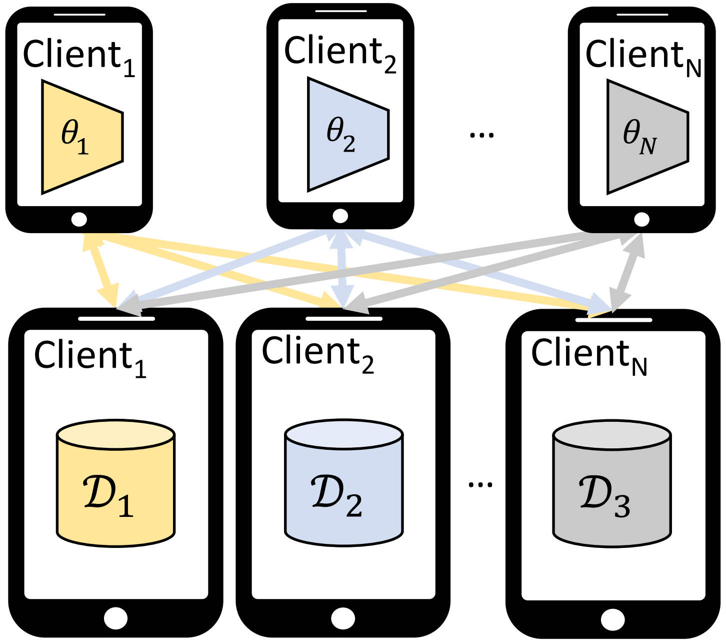
Motivation

- Difficulty to achieve explicitness
  - $O(N^2)$ communication overhead
  - Proper coefficient for each non-local risk



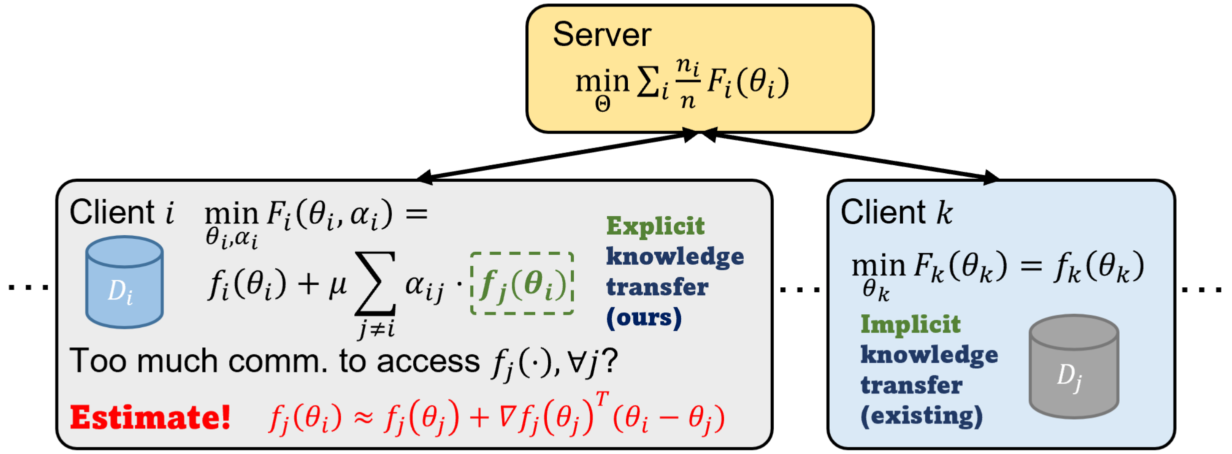
Motivation

- Difficulty to achieve explicitness
  - $O(N^2)$ communication overhead
  - Proper coefficient for each non-local risk



Method

- Proposed solution: **PGFed**
  - ✓ Estimate $f_j(\theta_i) \approx f_j(\theta_j) + \nabla f_j(\theta_j)^T(\theta_i - \theta_j)$, $O(N^2) \rightarrow O(N)$
  - ✓ Use adaptive coefficient $\alpha_{ij} \forall i, j \in [N]$



Method

- Objectives of Personalized Global Federated Learning (**PGFed**)
  - Global objective: $\min_{\boldsymbol{\Theta},\boldsymbol{A}} F(\boldsymbol{\Theta},\boldsymbol{A}) = \min_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N,\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_N} \sum_{i=1}^{N} p_i F_i(\boldsymbol{\theta}_i,\boldsymbol{\alpha}_i)$
  - Local objective: $F_i(\boldsymbol{\theta}_i,\boldsymbol{\alpha}_i) = f_i(\boldsymbol{\theta}_i) + \mu \sum_{j\in[N]} \alpha_{ij} f_j(\boldsymbol{\theta}_i)$
  - Plugging $f_j(\theta_i) \approx f_j(\theta_j) + \nabla f_j(\theta_j)^T(\theta_i - \theta_j)$ into Local objective, we have

$$F_i(\boldsymbol{\theta}_i,\boldsymbol{\alpha}_i) \approx f_i(\boldsymbol{\theta}_i) + \mathcal{R}_{aux}^{[N]}(\boldsymbol{\theta}_i,\boldsymbol{\alpha}_i)$$

$$\mathcal{R}_{aux}^{[N]}(\boldsymbol{\theta}_i,\boldsymbol{\alpha}_i) = \mu \sum_{j\in[N]} \alpha_{ij} \left(f_j(\boldsymbol{\theta}_j) + \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)\right)$$

($\mathcal{R}_{aux}^{[N]}(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) = \mu \sum_{j \in [N]} \alpha_{ij} \left(f_j(\boldsymbol{\theta}_j) + \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \right)$)

Method

- Gradient-based update
  - W.r.t θ_i:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_i} F_i(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) &= \nabla_{\boldsymbol{\theta}_i} f_i(\boldsymbol{\theta}_i) + \nabla_{\boldsymbol{\theta}_i} \mathcal{R}_{aux}^{[N]}(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) \\ &= \nabla_{\boldsymbol{\theta}_i} f_i(\boldsymbol{\theta}_i) + \underbrace{\mu \sum_{j \in [N]} \alpha_{ij} \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)}_{\tilde{\boldsymbol{g}}_{[N]}} . \end{aligned}$$

    - $\tilde{g}_{[N]}$ can be computed by the server with:
      - Client i uploading α_i
      - Client j uploading local gradient

($\mathcal{R}_{aux}^{[N]}(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) = \mu \sum_{j \in [N]} \alpha_{ij} \left(f_j(\boldsymbol{\theta}_j) + \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \right)$)

Method

- Gradient-based update
  - W.r.t α_{ij}:

$$\nabla_{\alpha_{ij}} F_i(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) = \mu \left(f_j(\boldsymbol{\theta}_j) + \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \right)$$
$$= \underbrace{\mu \left(f_j(\boldsymbol{\theta}_j) - \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T \boldsymbol{\theta}_j \right)}_{g_\alpha^{(1)}} + \underbrace{\mu \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j)^T \boldsymbol{\theta}_i}_{g_\alpha^{(2)}} .$$

    - $g_\alpha^{(1)}$ (a scalar) can be computed and uploaded by the client j
    - $g_\alpha^{(2)}$ (exact value needs to transmit all gradients to client i (takes $O(N^2)$ comm.))
      - Estimate: $g_\alpha^{(2)} \approx \bar{\boldsymbol{g}}_{[N]}^T \boldsymbol{\theta}_i = \frac{\mu}{N} \left(\sum_{j \in [N]} \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j) \right)^T \boldsymbol{\theta}_i$
        - Client j uploading local gradient

Method

- To accommodate to M selected clients per round: $[N] \rightarrow S_t$ (selected set of clients in round t)

$$\tilde{\boldsymbol{g}}_{\mathcal{S}_t} = \mu \sum_{j \in \mathcal{S}_t} \alpha_{ij} \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j) \qquad \bar{\boldsymbol{g}}_{\mathcal{S}_t} = \frac{\mu}{M} \left(\sum_{j \in \mathcal{S}_t} \nabla_{\boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j) \right)$$

- To keep information from clients selected in previous round, use momentum (PGFedMo)

$$\tilde{\boldsymbol{g}}^i_{\mathcal{S}_t} = (1-\beta)\tilde{\boldsymbol{g}}^i_{\mathcal{S}_t}(\text{downloaded}) + \beta \tilde{\boldsymbol{g}}^i_{\mathcal{S}_t}(\text{previous})$$

Algorithm 1 PGFed and PGFedMo

Input: N clients, learning rates η_1, η_2, number of rounds T, coefficient μ(, momentum β for PGFedMo)
Output: Personalized models $\boldsymbol{\theta}_1^T, ..., \boldsymbol{\theta}_N^T$.

ServerExecute:

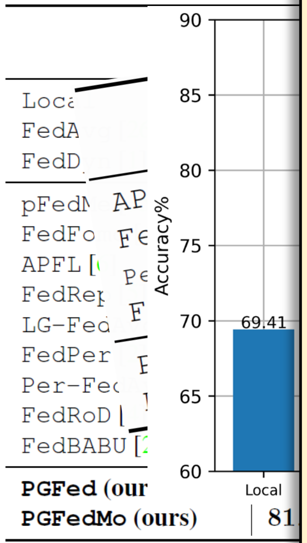
Initialize $\alpha_{ij} \leftarrow 1/M \; \forall i, j \in [N]$, global model $\boldsymbol{\theta}^0_{glob}$
$\mathbf{A}[i] \leftarrow \boldsymbol{\alpha}_i \; \forall i \in [N]$
for $t \leftarrow 1, 2, ..., T$ **do**
  Select a subset of M clients, $\mathcal{S}_t$
  $g_t^{(1)} \leftarrow \{\}; \nabla_t \leftarrow \{\}$ // built for next round
  for $i \in \mathcal{S}_t$ **in parallel do**
    if t=1 **then**
      $\boldsymbol{\theta}_i^t, g_\alpha^{(1)}, \nabla f(\boldsymbol{\theta}_i^t), \boldsymbol{\alpha}_i \leftarrow$ **ClientUpdate**$(\boldsymbol{\theta}^{t-1}_{glob}, t)$
    else
      $\tilde{\boldsymbol{g}}_{\mathcal{S}_{t-1}} \leftarrow \mu \sum_{j \in \mathcal{S}_{t-1}} \alpha_{ij} \nabla_{t-1}[j]$
      $\bar{\boldsymbol{g}}_{\mathcal{S}_{t-1}} \leftarrow \frac{\mu}{M} \sum_{j \in \mathcal{S}_{t-1}} \nabla_{t-1}[j]$
      $\boldsymbol{\theta}_i^t, g_\alpha^{(1)}, \nabla f(\boldsymbol{\theta}_i^t), \boldsymbol{\alpha}_i \leftarrow$ **ClientUpdate**$(\boldsymbol{\theta}^{t-1}_{glob}, t, \tilde{\boldsymbol{g}}_{\mathcal{S}_{t-1}}, \bar{\boldsymbol{g}}_{\mathcal{S}_{t-1}}, g^{(1)}_{t-1})$
    end if
    // the next line records the values for next round
    $\mathbf{A}[i] \leftarrow \boldsymbol{\alpha}_i; g_t^{(1)}[i] \leftarrow g_\alpha^{(1)}; \nabla_t[i] \leftarrow \nabla f(\boldsymbol{\theta}_i^t)$
    $\boldsymbol{\theta}^t_{glob} \leftarrow \sum_{i \in \mathcal{S}_t} p_i \boldsymbol{\theta}_i^t$
  end for
  for $i \in ([N] - \mathcal{S}_t)$ **in parallel do**
    $\boldsymbol{\theta}_i^t \leftarrow \boldsymbol{\theta}_i^{t-1}; \tilde{\boldsymbol{g}}_i^t \leftarrow \tilde{\boldsymbol{g}}_i^{t-1}$
  end for
end for
return $\boldsymbol{\theta}_1^T, ..., \boldsymbol{\theta}_N^T$

ClientUpdate$(\boldsymbol{\theta}^{t-1}_{global}, t \, (, \tilde{\boldsymbol{g}}, \bar{\boldsymbol{g}}, g^{(1)}_{t-1}))$:

if t=1 **then**
  $\boldsymbol{\theta}_i^t \leftarrow$ **ClientUpdate**$(\boldsymbol{\theta}^{t-1}_{global}, \eta_1)$ as in FedAvg
else
  $\boldsymbol{\theta}_i^t \leftarrow \boldsymbol{\theta}^{t-1}_{global}$
  $\tilde{\boldsymbol{g}}_i^t \leftarrow \tilde{\boldsymbol{g}}$ // without momentum
  $\tilde{\boldsymbol{g}}_i^t \leftarrow (1-\beta)\tilde{\boldsymbol{g}} + \beta \tilde{\boldsymbol{g}}_i^{t-1}$ // with momentum
  for Batch of data $\mathcal{B} \in \mathcal{D}_i$ **do**
    $\boldsymbol{\theta}_i^t \leftarrow \boldsymbol{\theta}_i^t - \eta_1(\nabla f(\boldsymbol{\theta}_i^t, \mathcal{B}) + \tilde{\boldsymbol{g}}_t^i)$
    $g^{(2)} = \bar{\boldsymbol{g}}^T \boldsymbol{\theta}_i$
    $\forall j \in g^{(1)}_{t-1}: \; \alpha_{ij} \leftarrow \alpha_{ij} - \eta_2(g^{(1)}_{t-1}[j] + g^{(2)})$
  end for
end if
$g_\alpha^{(1)} \leftarrow \mu \left(f(\boldsymbol{\theta}_i^t) - \nabla f(\boldsymbol{\theta}_i^t)^T \boldsymbol{\theta}_i^t \right)$ // for next round
return $\boldsymbol{\theta}_i^t, g_\alpha^{(1)}, \nabla f(\boldsymbol{\theta}_i^t), \boldsymbol{\alpha}_i$

Experiments & results

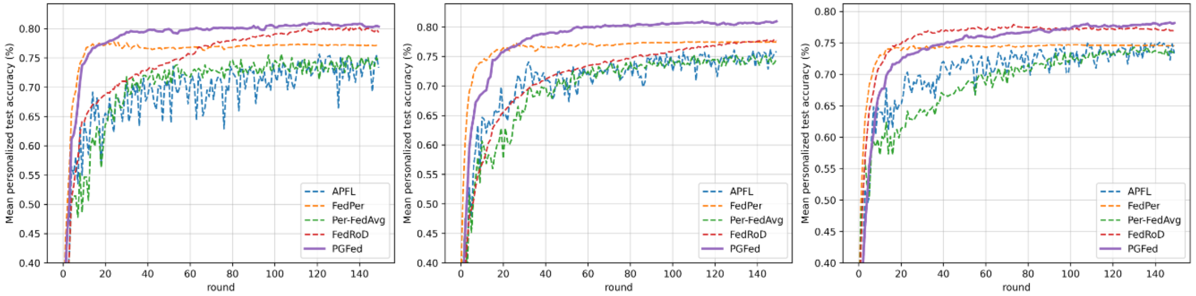
- Settings
  - Datasets: CIFAR10, CIFAR100, OrganAMNIST, Office-home
  - Heterogeneity: Dir $\alpha = 0.3, 1.0$
  - Number of clients: 20, 25, 50, 100 clients
  - Metrics
    - Mean local test accuracy
    - Mean individual gain over `Local`
    - #Rounds to reach 70% acc. & speedup
    - Accuracy of fine-tuning resulting global model on new clients
    - Throughput
    - Etc. (see full paper)



Main takeaways:

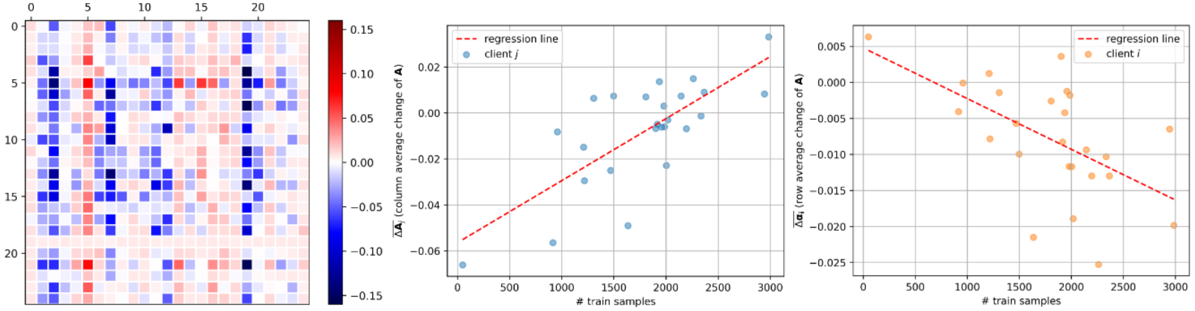
- ✓ **PGFed and PGFedMo boost the accuracy by up to 15.47%.**
- ✓ **PGFed and PGFedMo have 3.7× average speedup with highest individual gain.**
- ✓ **Global models of PGFed and PGFedMo have highest generailizability on new clients.**

More experiments & results



(a) 25 clients (b) 50 clients (c) 100 clients

Figure 1. Convergence behavior of the personalized FL approaches with top performance on CIFAR10. While achieving the highest accuracy performance, PGFed is also able to consistently converge faster than several of the baselines that reach high accuracies.



(a) (b) (c)

Figure 2. Visualization of the change in $\boldsymbol{A}$. Figure (a) is a heat map of the change in $\boldsymbol{A}$. For Figure (b) and (c), the Y-axis of Figure (b) represents the column average of the change in $\boldsymbol{A}$ (the average change of weights of client j's empirical risk on other clients). The Y-axis of Figure (c) is the row average of the change in $\boldsymbol{A}$ (the average change of weights of the auxiliary risk on client i). Through the regression line, we verify the positive correlation between $\overline{\Delta \boldsymbol{A}}_j$ and n_j in Figure (b), and the negative correlation between $\overline{\Delta \alpha}_i$ and n_i in Figure (c).

	Art	Clipart	Product	Real World	Mean
Local	17.16 ± 0.85	37.65 ± 0.47	43.83 ± 0.40	24.50 ± 0.21	30.79 ± 0.23
FedAvg	11.68 ± 1.26	41.29 ± 0.85	42.49 ± 1.28	19.14 ± 0.89	28.65 ± 0.49
APFL	19.11 ± 1.55	44.67 ± 0.61	**50.40 ± 0.56**	25.85 ± 0.88	35.00 ± 0.41
FedRep	20.24 ± 1.45	38.43 ± 1.02	43.70 ± 1.04	24.02 ± 0.81	31.60 ± 0.05
LGFedAvg	17.54 ± 0.45	38.75 ± 0.13	44.59 ± 0.62	25.79 ± 0.61	31.67 ± 0.21
FedPer	17.83 ± 1.07	38.97 ± 0.35	45.87 ± 0.13	25.01 ± 0.52	31.92 ± 0.24
Per-FedAvg	14.62 ± 0.40	39.94 ± 1.29	44.40 ± 1.32	21.58 ± 0.65	30.13 ± 0.07
FedRoD	19.67 ± 1.23	42.44 ± 0.77	44.34 ± 2.07	24.28 ± 1.69	32.68 ± 0.69
FedBABU	18.18 ± 3.54	42.10 ± 2.31	43.51 ± 0.91	**26.81 ± 1.86**	33.38 ± 0.29
PGFed	**22.40 ± 0.26**	**46.48 ± 1.00**	<u>49.86 ± 2.14</u>	26.04 ± 0.80	**36.19 ± 0.92**
PGFedMo	<u>22.16 ± 0.45</u>	<u>45.88 ± 0.83</u>	49.45 ± 0.19	<u>26.60 ± 0.99</u>	<u>36.02 ± 0.20</u>

Table 2. Mean and standard deviation over three trials of the mean personalized accuracy% of the four domains (5 clients/domain) and the average performance on Office-home dataset. The highest and second-highest accuracies under each setting are in **bold** and <u>underlined</u>, respectively.

	25 clients sample 50% Dir(1.0)	50 clients sample 25% Dir(0.3)	100 clients sample 25% Dir(0.3)
Local	90.45±0.19	90.63±0.07	87.14±0.10
FedAvg	99.11±0.03	98.74±0.04	98.47±0.08
APFL	97.49±0.05	97.53±0.06	96.19±0.11
FedRep	95.06±0.16	94.86±0.07	92.47±0.04
LGFedAvg	90.47±0.18	90.99±0.08	87.52±0.22
FedPer	97.89±0.06	97.55±0.08	95.56±0.33
Per-FedAvg	98.40±0.02	96.80±0.04	95.09±0.07
FedRoD	98.61±0.05	98.14±0.09	97.05±0.06
FedBABU	96.49±0.28	94.33±0.13	91.07±0.23
PGFed	99.20±0.04	**99.17±0.05**	**98.94±0.02**
PGFedMo	**99.21±0.04**	99.17±0.07	98.86±0.06

Table 1. Mean and standard deviation over three trials of the mean personalized test accuracy (%) on OrganAMNIST

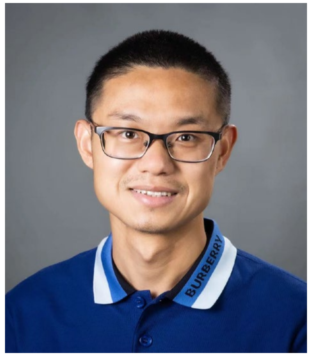
	Images/s	Relative speed	Accuracy
FedAvg	6917.1	100.00%	64.41±0.66
APFL	3389.8	48.99%	77.36±0.18
Per-FedAvg	3464.5	50.09%	76.27±0.50
FedRoD	6682.4	96.61%	79.61±0.22
PGFed	6120.0	88.48%	81.42±0.31
PGFedMo	6032.8	87.22%	81.48±0.32
PGFed-CE*	6175.5	89.28%	81.16±0.56

* A more communication-efficient variation of PGFed, introduced in Appendix D

Table 3. Computational speed (in terms of "images/s") and accuracy on CIFAR10 with 50 clients

More details in full paper…

Acknowledgements



<u>Jun Luo</u> Matias Mendieta Dr. Chen Chen Dr. Shandong Wu

Also check out our poster
Oct. 4th (Wed.)
02:30 PM-04:30 PM
Paper ID: 7061

Full paper

Code

NATIONAL CANCER INSTITUTE
NSF
aws
University of Pittsburgh
XSEDE

NIH/NIC 1R01CA218405
NSF CICI: SIVD: 2115082
NSF/NIH 1R01EB032896
NIH 3R01EB032896-03S1
Pitt Momentum Funds

Amazon AWS ML Research Reward
XSEDE by NSF ACI-1548562
Bridges-2 by NSF ACI-1928147
NSF/Intel Partnership on MLWiNS 2003198

Intelligent **C**omputing for **C**linical **I**maging (**ICCI**) Lab, University of Pittsburgh

Thank you!

Full paper

Code

Also check out our poster
Oct. 4th (Wed.)
02:30 PM-04:30 PM

Jun Luo
University of Pittsburgh
jul117@pitt.edu