Federated Learning with Partial Model Personalization

October 19th, 2022  @ FLOW Seminar

Krishna Pillutla

University of Washington → Google Research
Joint work with

Kshitiz Malik
Abdelrahman Mohamed
Mike Rabbat
Maziar Sanjabi
Lin Xiao

ICML 2022
Data is decentralized and private
Federated Learning

Percentage of world population with a smartphone

Data Credit: Business Wire
Federated Learning

Percentage of world population with a smartphone

Data Credit: Business Wire
Federated Learning

Percentage of world population with a smartphone

Year

Percentage


Data Credit: Business Wire
Federated Learning

Communication cost > computation cost!

Percentage of world population with a smartphone

Data Credit: Business Wire
Challenge

models are deployed on clients with heterogeneous data
THE ACCENT GAP

We tested Amazon's Alexa and Google's Home to see how people with accents are getting left behind in the smart-speaker revolution.

GOOGLE HOME
Overall accuracy 83%
- Western U.S. +3.0
- Midwest U.S. +2.5
- Eastern U.S. +0.5
- Southern U.S. +0.1
- Indian langs. -0.3
- Chinese -3.2
- Spanish -3.2

AMAZON ECHO
Overall accuracy 86%
- Southern U.S. +3.1
- Eastern U.S. +2.7
- Western U.S. +2.0
- Midwest U.S. +1.0
- Indian langs. -1.8
- Chinese -4.2
- Spanish -2.7
Challenge

models are deployed on clients with heterogeneous data

Personalization: Adapt (a part of) the model to each client
Challenge

models are deployed on clients with heterogeneous data

Partial Personalization: Adapt a part of the model to each client
How to personalize?

Federated Learning with Personalization Layers

- **Modeling**: Personalize the output layer
- **Optimization**: Train personal and shared parameters simultaneously

Manoj Ghuhan Arivazhagan
Adobe Research

Vinay Aggarwal
Indian Institute of Technology, Roorkee, India

Aaditya Kumar Singh
Indian Institute of Technology, Kharagpur, India

Sunav Choudhary
Adobe Research

2019
How to personalize?

Think Locally, Act Globally: Federated Learning with Local and Global Representations

Paul Pu Liang\textsuperscript{1}, Terrance Liu\textsuperscript{1}, Liu Ziyin\textsuperscript{2}, Nicholas B. Allen\textsuperscript{3}, Randy P. Auerbach\textsuperscript{4}, David Brent\textsuperscript{5}, Ruslan Salakhutdinov\textsuperscript{1}, Louis-Philippe Morency\textsuperscript{1}

\textsuperscript{1}School of Computer Science, Carnegie Mellon University
\textsuperscript{2}Department of Physics, University of Tokyo
\textsuperscript{3}Department of Psychology, University of Oregon
\textsuperscript{4}Department of Psychiatry, Columbia University
\textsuperscript{5}Department of Psychiatry, University of Pittsburgh

\{pliang, terrancl, morency\}@cs.cmu.edu

July 15, 2020

**Modeling**: Personalize the \textit{input} layer

**Optimization**: Train personal and shared parameters \textit{simultaneously}
How to personalize?

Exploiting Shared Representations for Personalized Federated Learning

Liam Collins\textsuperscript{1}  Hamed Hassani\textsuperscript{2}  Aryan Mokhtari\textsuperscript{1}  Sanjay Shakkottai\textsuperscript{1}

ICML 2021

**Modeling:** Personalize the output layer

**Optimization:** Train personal and shared parameters \textit{alternatively}
How to personalize?

**Federated Reconstruction:**
Partially Local Federated Learning

Karan Singhal  
Google Research  
karansinghal@google.com

Hakim Sidahmed  
Google Research  
hsidahmed@google.com

Zachary Garrett  
Google Research  
zachgarrett@google.com

Shanshan Wu  
Google Research  
shanshanw@google.com

Keith Rush  
Google Research  
krush@google.com

Sushant Prakash  
Google Research  
sush@google.com

NeurIPS 2021

**Optimization:** Train personal and shared parameters *alternatively*
So, how do we personalize a federated model?

**Design decisions:**

- Modeling
- Optimization
Our contributions

1. **Theory**: Analysis of both these optimization algorithms

2. **Extensive experiments**: text, vision, and speech settings
Outline

1. Setup and review

2. Convergence Analysis

3. Experiments
Outline

1. Setup and review

2. Convergence Analysis

3. Experiments
(Non-personalized) federated learning

\[ \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} F_i(w) \]

where \[ F_i(w) = \mathbb{E}_{z \sim p_i} \left[ f(w; z) \right] \]

loss on client \( i \)

[McMahan et al. AISTATS (2017), Kairouz et al. (2021)]
**Personalized federated learning**

\[ \text{Objective: } \min_{u, v_1, \ldots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i) \]

Model on client \( i = (u, v_i) \)

- \( u \): shared parameters
- \( v_i \): personal parameters
Personalization architectures

![Diagram of personalization architectures](image)

**Personalized output layer**
- Pred.
- Personal
- Shared
- Input

Arivazhagan et al. (2019)
Collins et al. ICML (2021)

**Personalized input layer**
- Pred.
- Personal
- Shared
- Input

Liang et al. (2019)

**Combined predictions**

\[ F(u, v_i) = E_{iX,Y \sim p_i}(\phi_p(X; u) + \phi_l(X; v_i) - Y)^2 \]

Agarwal et al. (2020)

**Personalized adapters**
- Pred.
- Output
- Adapter
- Norm+MLP
- Adapter
- Norm+Attn
- Embed
- Input

\( \times N \)

Personalization architectures

**Personalized output layer**

**Personalized input layer**

**Combined predictions**

**Personalized adapters**

**Personalization architectures**

**Personalized output layer**

- Pred.
- Personal
- Shared
- Input

**Personalized input layer**

- Pred.
- Shared
- Personal
- Input

**Combined predictions**

- Pred.
- Shared
- Personal
- Input

\[ F(u, v_i) = \mathbb{E}_{(X,Y) \sim p_i} (\phi_g(X; u) + \phi_l(X; v_i) - Y)^2 \]

**Personalized adapters**

- Pred.
- Output
- ×N
- Adapter
- Norm+MLP
- Adapter
- Norm+Attn
- Embed
- Input

Other forms of personalization

\[ \text{pFedMe:} \quad \min_{u,v_1,\ldots,v_n} \frac{1}{n} \sum_{i=1}^{n} \left( f_i(v_i) + \frac{\lambda}{2} \| v_i - u \|_2^2 \right) \]

[Dinh et. al (NeurIPS 2020)]

Ditto, MAML, APFL, .... [Hanzely et al. (2021)]
Non-personalized (FedAvg)

\[
\min_w \frac{1}{n} \sum_{i=1}^{n} F_i(w)
\]

FedAvg [MacMahan et al. AISTATS (2017)]

Parallel Gradient Distribution [Mangasarian. SICON (1995)]
Iterative Parameter Mixing [McDonald et al. ACL (2009)]
BMUF [Chen & Huo. ICASSP (2016)]
Local SGD [Stich. ICLR (2019)]

Personalized (FedAlt/FedSim)

\[
\min_{u,v_1,\ldots,v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)
\]
Non-personalized (FedAvg)

\[
\min_w \frac{1}{n} \sum_{i=1}^{n} F_i(w)
\]

Personalized (FedAlt/FedSim)

\[
\min_{u,v_1,\cdots,v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)
\]

Step 1 of 3: Server samples in clients and broadcasts global model
Step 2 of 3: Clients perform $\tau$ local SGD steps on their local data.
Non-personalized (FedAvg)

\[
\min_w \frac{1}{n} \sum_{i=1}^{n} F_i(w)
\]

Step 2 of 3: Clients perform \( \tau \) local SGD steps on their local data

Personalized (FedAlt/FedSim)

\[
\min_{u,v_1,\ldots,v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)
\]

FedAlt (alternating update)

\[
\begin{align*}
v_i^+ &= v_i - \gamma \nabla_v F_i(u, v_i) \\
u_i^+ &= u - \gamma \nabla_u F_i(u, v_i^+)
\end{align*}
\]

FedSim (simultaneous update)

\[
\begin{align*}
v_i^+ &= v_i - \gamma \nabla_v F_i(u, v_i) \\
u_i^+ &= u - \gamma \nabla_u F_i(u, v_i)
\end{align*}
\]
Step 3 of 3: Aggregate (shared components) of client updates

Non-personalized (FedAvg)

\[
\min_w \frac{1}{n} \sum_{i=1}^{n} F_i(w)
\]

Personalized (FedAlt/FedSim)

\[
\min_{u,v_1,\ldots,v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)
\]

\[
w^+ = \frac{1}{m} \sum_{i} w_i^+
\]

\[
u^+ = \frac{1}{m} \sum_{i} u_i^+
\]

\(v_i\) stays on client \(i\)
Outline

1. Setup and review

2. Convergence Analysis

3. Experiments
Assumptions

Model on client $i = (u, v_i)$

Objective: $\min_{u, v_1, \ldots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$

$u$: shared parameters

$v_i$: personal parameters

1. Smoothness

$\nabla_u F_i$ is

\[
\begin{cases}
L_u \text{-Lipschitz w.r.t. } u \\
L_{uv} \text{-Lipschitz w.r.t. } v_i
\end{cases}
\]

$\nabla_v F_i$ is

\[
\begin{cases}
L_v \text{-Lipschitz w.r.t. } v_i \\
L_{uv} \text{-Lipschitz w.r.t. } u
\end{cases}
\]

$\chi^2 := \frac{L_{uv}^2}{L_u L_v}$ quantifies cross-dependence
Assumptions

Model on client $i = (u, v_i)$

Objective: \[
\min_{u, v_1, \ldots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)
\]

- stochastic gradients of $\nabla_u F_i$ and $\nabla_v F_i$ have bounded variance $\sigma_u^2$ and $\sigma_v^2$ respectively

- bounded gradient diversity:
  \[
  \frac{1}{n} \sum_{i=1}^{n} \| \nabla_u F_i(u, v) - \nabla_u F(u, v_{1:n}) \|^2 \leq \delta^2
  \]
Theorem [P., Malik, Mohamed, Rabbat, Sanjabi, Xiao]

Under the smoothness and bounded variance assumptions, we have the bounds

\[
\text{FedAlt} \quad \frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2 \right) \leq \sqrt{\frac{\sigma_1^2}{T}} + \left( \frac{\tilde{\sigma}_1^2}{T} \right)^{2/3} + O \left( \frac{1}{T} \right)
\]

\[
\text{FedSim} \quad \frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2 \right) \leq \sqrt{\frac{\sigma_2^2}{T}} + \left( \frac{\tilde{\sigma}_2^2}{T} \right)^{2/3} + O \left( \frac{1}{T} \right)
\]

\[\sigma_1^2, \sigma_2^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2 \text{ are linear combinations of } \sigma_u^2, \sigma_v^2, \delta^2\]
Theorem [P., Malik, Mohamed, Rabbat, Sanjabi, Xiao]

Under the smoothness and bounded variance assumptions, we have the bounds

\[
\text{FedAlt} \quad \frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \mathbb{E} \left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2 \right) \leq \sqrt{ \frac{\sigma_1^2}{T} + \left( \frac{\tilde{\sigma}_1^2}{T} \right)^{2/3} } + O \left( \frac{1}{T} \right)
\]

\[
\text{FedSim} \quad \frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \mathbb{E} \left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2 \right) \leq \sqrt{ \frac{\sigma_2^2}{T} + \left( \frac{\tilde{\sigma}_2^2}{T} \right)^{2/3} } + O \left( \frac{1}{T} \right)
\]

\( \sigma_1^2, \sigma_2^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2 \) are linear combinations of \( \sigma_u^2, \sigma_v^2, \delta^2 \)
Theorem [P., Malik, Mohamed, Rabbat, Sanjabi, Xiao]

Under the smoothness and bounded variance assumptions, we have the bounds

**FedAlt**
\[
\frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \mathbb{E} \left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2 \right) \leq \sqrt{\frac{\sigma_1^2}{T}} + \left( \frac{\tilde{\sigma}_1^2}{T} \right)^{2/3} + O \left( \frac{1}{T} \right)
\]

**FedSim**
\[
\frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \mathbb{E} \left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2 \right) \leq \sqrt{\frac{\sigma_2^2}{T}} + \left( \frac{\tilde{\sigma}_2^2}{T} \right)^{2/3} + O \left( \frac{1}{T} \right)
\]

\[\sigma_1^2, \sigma_2^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2 \text{ are linear combinations of } \sigma_u^2, \sigma_v^2, \delta^2\]
Theorem \([\textit{P.}, \text{Malik, Mohamed, Rabbat, Sanjabi, Xiao}]

Under the smoothness and bounded variance assumptions, we have the bounds

\[
\text{FedAlt} \quad \frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \mathbb{E} \left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2 \right) \leq \sqrt{\frac{\sigma_1^2}{T}} + \left( \frac{\tilde{\sigma}_1^2}{T} \right)^{2/3} + O \left( \frac{1}{T} \right)
\]

\[
\text{FedSim} \quad \frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \mathbb{E} \left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2 \right) \leq \sqrt{\frac{\sigma_2^2}{T}} + \left( \frac{\tilde{\sigma}_2^2}{T} \right)^{2/3} + O \left( \frac{1}{T} \right)
\]

\[\sigma_1^2, \sigma_2^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2 \text{ are linear combinations of } \sigma_u^2, \sigma_v^2, \delta^2\]
Theorem [P., Malik, Mohamed, Rabbat, Sanjabi, Xiao]

Under the smoothness and bounded variance assumptions, we have the bounds

\[
\text{FedAlt} \quad \frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \mathbb{E} \left\| \nabla_u F(u_t, v_{1:t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i:t}) \right\|^2 \right) \leq \sqrt{\frac{\sigma_1^2}{T}} + \left( \frac{\bar{\sigma}_1^2}{T} \right)^{2/3} + O \left( \frac{1}{T} \right)
\]

\[
\text{FedSim} \quad \frac{1}{T} \sum_{t=0}^{T-1} \left( \frac{1}{L_u} \mathbb{E} \left\| \nabla_u F(u_t, v_{1:t}) \right\|^2 + \frac{1}{nL_v} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla_v F_i(u_t, v_{i:t}) \right\|^2 \right) \leq \sqrt{\frac{\sigma_2^2}{T}} + \left( \frac{\bar{\sigma}_2^2}{T} \right)^{2/3} + O \left( \frac{1}{T} \right)
\]

\[\sigma_1^2, \sigma_2^2, \bar{\sigma}_1^2, \bar{\sigma}_2^2 \text{ are linear combinations of } \sigma_u^2, \sigma_v^2, \delta^2\]
FedAlt is better than FedSim when

$$\frac{\sigma_v^2}{L_v} \left(1 - \frac{2m}{n}\right) < \frac{\sigma_u^2}{mL_u} + \frac{\delta^2}{mL_u} \left(1 - \frac{m}{n}\right)$$

True if \(\delta^2 \gg \max\{\sigma_u^2, \sigma_v^2\}\)

Better by a factor of \((1 + \chi^2)^{1/2}\)

- \(m\): number of clients per round
- \(n\): total number of clients
- \(\sigma_u^2, \sigma_v^2, \delta^2\): noise variances
- \(\chi^2 = \frac{L_{uv}^2}{L_u L_v}\): cross-dependency
Assume $\sigma_u^2 = 0 = \sigma_v^2$ and single local gradient step per client

For FedAlt, apply smoothness for $u$-step (assuming $v$-step is complete) to get

$$
F(u_{t+1}, v_{t+1}) - F(u_t, v_{t+1}) \leq \langle \nabla_u F(u_t, v_{t+1}), u_{t+1} - u_t \rangle + \frac{L_u}{2} \|u_{t+1} - u_t\|^2
$$

both depend on sampling of clients

**first-order term is biased!**
For **FedSim**, no such difficulties

\[
F(u_{t+1}, v_{t+1}) - F(u_t, v_t) \leq \langle \nabla_u F(u_t, v_t), u_{t+1} - u_t \rangle + \frac{L_u}{2} \|u_{t+1} - u_t\|^2
\]

- \text{\textit{u-update starts from}} \ (u_t, v_t)
- \text{\textit{only dependence on sampling of clients}}

\textbf{first-order term is unbiased!}
For *FedAlt*, apply smoothness for $u$-step (assuming $v$-step is complete) to get

$$F(u_{t+1}, v_{t+1}) - F(u_t, v_{t+1}) \leq \langle \nabla_u F(u_t, v_{t+1}), u_{t+1} - u_t \rangle + \frac{L_u}{2} \|u_{t+1} - u_t\|^2$$

both depend on sampling of clients

**first-order term is biased!**
Virtual full participation

Let $\tilde{v}_t$ denote the (virtual) personal parameters if all clients had run the $\nu$-step, not just the selected clients.
For **FedAlt**, apply smoothness for $u$-step (assuming $v$-step is complete) to get

$$F(u_{t+1}, v_{t+1}) - F(u_t, v_{t+1}) \leq \langle \nabla_u F(u_t, \tilde{v}_{t+1}), u_{t+1} - u_t \rangle + \frac{L_u}{2} \|u_{t+1} - u_t\|^2 + \text{Error}_t$$

independent of sampling of clients

depends on sampling of clients

**first-order term is unbiased again!**
To complete the proof, suffices to bound

$$\mathbb{E}[\text{Error}_t] \leq O(L_u\gamma_u^2 + \chi^2 L_v\gamma_v^2)$$

and can be made smaller by controlling the learning rates $\gamma_u, \gamma_v$.
Outline

1. Setup and review

2. Convergence Analysis

3. Experiments
- **Next word prediction**
  - Mobile keyboard
  - StackOverflow (~1K clients)
  - 4-layer transformer (6M param)
  - vocabulary size: 10K

- **Speech recognition**
  - Mobile assistant
  - LibriSpeech dataset (~1K clients)
  - 6-layer transformer (15M param)
  - CTC Loss (dynamic programming)

- **Landmark detection**
  - Mobile camera app
  - GLDv2 dataset (~1K clients)
  - ResNet-18 (12M param)
  - ~2K classes: only 30/client
Question 1: Modeling

Which form of personalization do I use?
Next word prediction

Speech recognition

Landmark detection

\( y\)-axis shows error: lower is better
Next word prediction

Speech recognition

Landmark detection

y-axis shows error: lower is better
Next word prediction

Speech recognition

Landmark detection

y-axis shows error: lower is better
**Takeaway:** Best personalization architecture depends on the task’s statistical heterogeneity
Partial personalization vs. full personalization

StackOverflow

GLDv2

Δ Accuracy

# Personalized Params.

1% params

10% params

Δ Accuracy
Question 2: Optimization

Which optimization algorithm do I use?
Next word prediction

![Bar chart for Next word prediction](chart1.png)

- FedAlt: 75.6
- FedSim: 75.3

**y-axis shows error: lower is better**

Landmark detection

![Bar chart for Landmark detection](chart2.png)

- Input: 48
- Output: 46
- Adapter: 33.75

**y-axis shows error: lower is better**
Takeaway: FedAlt \approx FedSim
But FedAlt is slightly better
1. **Theory**: Analysis of both these optimization algorithms

2. **Extensive experiments**: text, vision, and speech settings