Federated Learning with Partial Model Personalization

ICML 2022

Krishna Pillutla
Kshitiz Malik
Abdelrahman Mohamed
Mike Rabbat
Maziar Sanjabi
Lin Xiao
Personalized federated learning

Server

\[ \sum \]

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
Our contributions

1. **Theory**: Analysis of 2 popular optimization algs

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

**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

Code:
Personalization architectures

**Personalization architectures**

![Diagram of personalization architectures]

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

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

- **Combined predictions**
  - Pred.
  - Shared
  - Personal
  - Input

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

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

Liang et al. (2019)

Agarwal et al. (2020)

**Personalization architectures**

- **Personalized output layer**
- **Personalized input layer**
- **Combined predictions**
- **Personalized adapters**


$$F(u, v_i) = E_{i(X,Y) - p} \left( \phi_k(X; u) + \phi_l(X; v) - Y \right)^2$$

Liang et al. (2019)

Arivazhagan et al. (2019)

Collins et al. ICML (2021)

Agarwal et al. (2020)
Personalization architectures

### Personalized output layer
- Output
- Personal
- Shared
- Input

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

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

Liang et al. (2019)

### Combined predictions
- Input
- Pred.
- Personal
- Shared

F(u, v_i) = E_{i;X,Y \sim p}(\phi_g(X; u) + \phi_l(X; v_i) - Y)^2

Agarwal et al. (2020)

### Personalized adapters
- Input
- Output
- Adapter
- Norm+MLP
- Adapter
- Norm+Attn
- Embed
- × N

Optimization

- Server samples $m$ clients and broadcast global model $u$

- **Local updates** on client $i$: 
  $(u_i^+, v_i^+) = \text{LocalUpdate}_i(u, v_i)$

- Aggregate updates to global part of the model:
  
  $u^+ = \frac{1}{m} \sum_i u_i^+$

---

**Alternating update**

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

**Simultaneous update**

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

---

Liang et al. (2019)
Arivazhagan et al. (2019)
Collins et al. ICML (2021)
Singhal et al. NeurIPS (2021)
Optimization

- Server samples $m$ clients and broadcast global model $u$

- **Local updates** on client $i$: $(u_i^+, v_i^+) = \text{LocalUpdate}_i(u, v_i)$

- Aggregate updates to global part of the model:
  \[
  u^+ = \frac{1}{m} \sum_i u_i^+
  \]

**Alternating update**

- $v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$
- $u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$

**Simultaneous update**

- $v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$
- $u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$

---

Liang et al. (2019)
Arivazhagan et al. (2019)
Collins et al. ICML (2021)
Singhal et al. NeurIPS (2021)
Optimization

- Server samples $m$ clients and broadcast global model $u$

- **Local updates** on client $i$: 
  \[(u_i^+, v_i^+) = \text{LocalUpdate}_i(u, v_i)\]

- Aggregate updates to global part of the model:
  \[u^+ = \frac{1}{m} \sum_i u_i^+\]

**Alternating update**

- \[v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)\]
- \[u_i^+ = u - \gamma \nabla_u F_i(u, v_i^+)\]

**Simultaneous update**

- \[v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)\]
- \[u_i^+ = u - \gamma \nabla_u F_i(u, v_i)\]

References:
- Liang et al. (2019)
- Arivazhagan et al. (2019)
- Collins et al. ICML (2021)
- Singhal et al. NeurIPS (2021)
Contribution 1: Theory

**Theorem** [P., Malik, Mohamed, Rabbat, Sanjabi, Xiao]

For smooth, nonconvex functions and client sampling, we have the rates:

Alternating update: \( \frac{\sigma^2}{\sqrt{t}} \)

Simultaneous update: \( \frac{\sigma^2}{\sqrt{t}} \)

where \( \sigma^2_1 < \sigma^2_2 \) under typical scenarios

**Experimentally**, small but consistent trend of alternating > simultaneous

**Alternating update**

\[ v_i^+ = v_i - \gamma \nabla v F_i(u, v_i) \]

\[ u_i^+ = u_i - \gamma \nabla u F_i(u, v_i^+) \]

**Simultaneous update**

\[ v_i^+ = v_i - \gamma \nabla v F_i(u, v_i) \]

\[ u_i^+ = u_i - \gamma \nabla u F_i(u, v_i) \]
Contribution 1: Theory

**Key technical challenge**: Dependent random variables in alternating update algorithm due to random sampling of clients

**Methodology**: Developed technique of virtual full participation
Contribution 2: Experiments

Next word prediction

Speech recognition

Landmark detection

y-axis shows error: lower is better
Contribution 2: Experiments

Next word prediction

Speech recognition

Landmark detection

\[ \text{y-axis shows error: lower is better} \]
Contribution 2: Experiments

Next word prediction

Speech recognition

Landmark detection

y-axis shows error: lower is better
Contribution 2: Experiments

**Takeaway:** Best personalization architecture depends on the task’s statistical heterogeneity

$y$-axis shows error: lower is better
Federated Learning with Partial Model Personalization


Code: https://github.com/krishnap25/FL_partial_personalization