Personalized Federated Learning: A Unified Framework and Universal Optimization Techniques

Filip Hanzely \(^1\), Boxin Zhao \(^2\), and Mladen Kolar \(^2\)

\(^1\)Toyota Technological Institute at Chicago (TTIC) \(^2\)Booth School of Business, The University of Chicago

---

**Main takeaway**

- Optimization theory (by simple algorithms + convergence guarantees) applicable to all strongly convex personalized FL objectives.
- Tight convergence rates despite the generality: Matching best-known rates from the literature (in all but one case). Novel guarantees for new objectives.

A unified Personalized FL objective

Optimization problem of interest:

\[
\min_{\omega, \beta} \left\{ f(\omega, \beta) = \frac{1}{M} \sum_{m=1}^{M} f_m(\omega_m, \beta_m) \right\}
\]  

(1)

Notation: \( \omega \in \mathbb{R}^d \) shared parameters, \( \beta = (\beta_1, \ldots, \beta_M) \in \mathbb{R}^{m \times d} \).

Main idea: Choose \( f_m(\omega_m, \beta_m) \) to recover a particular personalized FL objective as an instance of (1) and apply our optimization theory.

**Detailed contributions**

- Universal (convex) optimization theory for personalized FL. We propose three algorithms for solving the general personalized FL objective (1):
  1. Local Stochastic Gradient Descent for Personalized FL (LSGD-PFL).
  2. Accelerated block Coordinate Descent for Personalized FL (ACD-PFL).
  3. Accelerated Stochastic Variance Reduced Coordinate Descent for Personalized FL (ASVRCD-PFL).
- Convergence rates. We provide lower complexity bounds for solving (1). ACD-PFL is always optimal in terms of the local computation and local computation when the full gradients are available, while ASVRCD-PFL can be optimal either in terms of the number of evaluations of the \( m \)-stochastic gradient or the \( \beta \)-stochastic gradient.
- Single personalized FL objective: We propose a single objective (1) capable of recovering all the existing personalized FL approaches by carefully constructing the local loss \( f_m(\omega_m, \beta_m) \). Surprisingly, the optimization guarantees for (1) yield fast convergence for individual special cases.
- Personalization and communication complexity. Our theory concludes that the personalization has positive effect on the communication complexity of training FL models.
- New personalized FL objectives. The universal personalized FL objective (1) enables us to obtain a range of novel personalized FL formulations as a special case.

**Algorithms**

- **LSGD-PFL**: Mixture between Local SGD and SGD. Local SGD step is taken with respect to \( \omega \)-variables or \( \beta \)-variables or both.
- **ACD-PFL**: An instance of the accelerated block coordinate descent with carefully designed non-uniform sampling of coordinate blocks (\( \omega \)-variables or \( \beta \)-variables).
- **ASVRCD-PFL**: ACD-PFL that subsamples the local finite sum combined with the variance reduction.

**Optimization guarantees for solving (1)**

<table>
<thead>
<tr>
<th>Alg.</th>
<th>Communication</th>
<th>( # \omega )</th>
<th>( # \beta )</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSGD-PFL</td>
<td>( \frac{\alpha}{\mu} ) + \frac{\beta}{\mu} \sigma w \sqrt{\frac{\alpha}{\mu}} )</td>
<td>( \frac{\alpha}{\mu} )</td>
<td>( \frac{\beta}{\mu} )</td>
<td>( \mu )</td>
</tr>
<tr>
<td>ACD-PFL</td>
<td>( \frac{\alpha}{\mu} ) + \frac{\beta}{\mu} \sigma w \sqrt{\frac{\alpha}{\mu}} )</td>
<td>( \frac{\alpha}{\mu} )</td>
<td>( \frac{\beta}{\mu} )</td>
<td>( \mu )</td>
</tr>
<tr>
<td>ASVRCD-PFL</td>
<td>( \frac{\alpha}{\mu} ) + \frac{\beta}{\mu} \sigma w \sqrt{\frac{\alpha}{\mu}} )</td>
<td>( \frac{\alpha}{\mu} )</td>
<td>( \frac{\beta}{\mu} )</td>
<td>( \mu )</td>
</tr>
</tbody>
</table>

---

**Special cases**

- **Traditional FL**: \( \min_{w, \omega \in \mathbb{R}^d} F(w) = \frac{1}{M} \sum_{m=1}^{M} f_m(w) \).
- **Fully personalized FL**: \( \min_{w, \beta \in \mathbb{R}^{m \times d}} F(w, \beta) = \frac{1}{M} \sum_{m=1}^{M} f_m(w, \beta_m) \).
- **Multi-task personalized FL**: \( \min_{w, \beta \in \mathbb{R}^{m \times d}} F_{\text{MT}}(w, \beta) = \frac{1}{M} \sum_{m=1}^{M} f_m(w, \beta_m) \).

**Smoothness and strong convexity for special cases**

<table>
<thead>
<tr>
<th>Objective</th>
<th>( \mu )</th>
<th>( L^\omega )</th>
<th>( L^\beta )</th>
<th>( L^\beta )</th>
<th>( \mu^\beta )</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully pers</td>
<td>( \mu )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>new</td>
<td>( \mu )</td>
</tr>
<tr>
<td>[3]</td>
<td>( \mu )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>new</td>
<td>( \mu )</td>
</tr>
<tr>
<td>[4]</td>
<td>( \mu )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>new</td>
<td>( \mu )</td>
</tr>
<tr>
<td>[7]</td>
<td>( \mu )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>new</td>
<td>( \mu )</td>
</tr>
</tbody>
</table>

**Experiments**

Setup: 3 personalized FL objectives, each applied to 3 datasets: MNIST, KMNIST, and FMNIST. Model: multiclass logistic regression.

Goal of experiment: Demonstrate the effect of the \( \mu^\beta \) rescaling of the \( w \)-space.

---

**References**


---

See full paper at https://arxiv.org/abs/2102.07423

ICLR Workshop Distributed and Private Machine Learning (DPML)