Lower Bounds and Optimal Algorithms for Personalized Federated Learning

Filip Hanzely Peter Richtárik

King Abdullah University of Science and Technology

Classical Federated Learning (FL)

Goal (training):

$$\min_{z\in\mathbb{R}^d} \frac{1}{n}\sum_{j=1}^{n} f_j(z). \tag{1}$$

- Large number of devices n
- Device i owns f_i; f_i constructed from local data only
- Local data might differ greatly across devices
- The solution $z^\star$ is deployed to the clients after training
- **Communication is the bottleneck**
- Typical approach: Local (S)GD:

$$x_i^{k+1} = \begin{cases} x_i^k - \alpha\nabla f_i(x_i^k) & \text{if } i \bmod (T+1) \neq 0 \\ \frac{1}{n}\sum_{i=1}^n x_i^k & \text{if } i \bmod (T+1) = 0 \end{cases} \tag{2}$$

Issue 1: Local data differ greatly across devices – single model $z^\star$ is not enough $\Rightarrow$ need for personalization.

Issue 2: Local optimizers (i.e., Local SGD/FedAVG, FedProx) are strictly worse in theory compared to their non-local cousins (i.e., SGD) without assuming a similarity between the local problems.

Main contributions

New FL objective that solves Issue 1 and Issue 2 simultaneously.

- Instead of searching for a single global model (1), we seek for an implicit mixture of global and local models.
- Local (S)GD can be seen as a variant of non-uniform SGD applied to our objective.
- Doing more local steps improves the communication complexity. **This is the first time that locality was proven to be beneficial for FL with heterogeneous data.**
- Insights into the local methods: The role of local steps is not to reduce communication complexity (as is generally believed) but rather to provide an implicit personalization bias.
- Extensions: Local finite sum, local drift, variance reduction, partial participation.

Personalized FL

Allow different local models, penalize dissimilarity:

$$\min_{\substack{x=[x_1,\dots,x_n]\in\mathbb{R}^{nd}\\ \forall i:\, x_i\in\mathbb{R}^d}} \left\{ F(x) \overset{\text{def}}{=} \underbrace{\frac{1}{n}\sum_{i=1}^n f_i(x_i)}_{\overset{\text{def}}{=} f(x)} + \lambda \underbrace{\frac{1}{2n}\sum_{i=1}^n \|x_i - \bar{x}\|^2}_{\overset{\text{def}}{=}\psi(x)} \right\} \tag{3}$$

- Above, $\bar{x} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n x_i$ and $\lambda \geq 0$
- Local objective might be a finite sum:

$$f_i(x_i) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m \tilde{f}_{i,j}(x_i)$$

- $f_{i,j}$ is L-smooth and μ-strongly convex
- $x(\lambda) = [x_1(\lambda), \dots, x_n(\lambda))]$: unique solution of (1) as a function of λ
- Objective is not new [1], personalization was justified. However, we are the first to draw the connection with FL and local methods.

Extreme cases

$\lambda = 0$: Personalized FL objective (1) coincides with the classical objective (1).
$\lambda = \infty$: Dissimilarity of x_i's is not penalized $\Rightarrow x_i(0) = \text{argmin}_{z\in\mathbb{R}^d} f_i(z)$. No communication is required.

- Solution is similar to MAML [2]:

$$x_i(\lambda) = \overline{x}(\lambda) - \tfrac{1}{\lambda}\nabla f_i(x_i(\lambda))$$

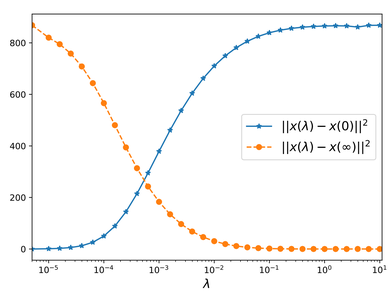

Figure 1:Distance of solution $x(\lambda)$ to pure local solution $x(0)$ and global solution $x(\infty)$ as a function of λ.

Local GD (LGD)

Local GD is a cyclic GD applied on (1):

1. Rewrite (1)

$$F(x) = \frac{1}{T+1}\sum_{i=1}^{T+1}\phi_i(x),\ \phi_i = \begin{cases} \frac{T+1}{T} f & i \leq T \\ (T+1)\lambda\psi & i = T+1 \end{cases} \tag{4}$$

2. Solve (4) with cyclic GD

$$x^{k+1} = x^k - \alpha\nabla\phi_{k \bmod T}(x^k), \tag{5}$$

3. Method (5) is (2) for a specific stepsize α

Loopless LGD (L2GD)

Non-uniform SGD applied on 2-sum

$$F(x) = f(x) + \lambda\psi(x) \tag{6}$$

Algorithm:

$$x^+ = \begin{cases} x - \frac{\alpha}{1-p}\nabla f(x) & \text{with probability } 1-p \\ x - \frac{\alpha}{p}\nabla\psi(x) & \text{with probability } p \end{cases}$$

- $\alpha \geq 0$ is the stepsize, $0 < p < 1$ is the probability
- Evaluating $\nabla f(x)$ requires local computation:

$$x_i^{k+1} = x_i^k - \frac{\alpha}{1-p}\nabla f_i(x_i^k)$$

- Evaluating $\nabla\psi(x)$ requires communication

$$x_i^{k+1} = (1 - \frac{\alpha\lambda}{np})x_i^k + \frac{\alpha\lambda}{np}\bar{x}^k$$

- On average, $\frac{1-p}{p}$ gradient steps per one communication

Convergence[3]

If $\alpha \leq \frac{1}{2\mathcal{L}}$, then

$$\mathbb{E}\left[\left\|x^k - x(\lambda)\right\|^2\right] \leq \left(1 - \frac{\alpha\mu}{n}\right)^k \left\|x^0 - x(\lambda)\right\|^2 + \frac{2n\alpha\sigma^2}{\mu},$$

where

$$\mathcal{L} \overset{\text{def}}{=} \frac{1}{n}\max\left\{\frac{L}{1-p}, \frac{\lambda}{p}\right\},$$

$$\sigma^2 \overset{\text{def}}{=} \frac{1}{n^2}\sum_{i=1}^n \left(\frac{1}{1-p}\|\nabla f_i(x_i(\lambda))\|^2 + \frac{\lambda^2}{p}\|x_i(\lambda) - \overline{x}(\lambda)\|^2\right).$$

Linear convergence to a neighborhood of the optimum.

L2GD+

- Non uniform SAGA applied on 2-sum (6)
- Linear convergence to the exact optimum
- Iteration complexity to reach ϵ-solution:

$$\mathcal{O}\left(\max\left\{\frac{L}{(1-p)\mu}, \frac{\lambda}{p\mu}\right\}\log\frac{1}{\epsilon}\right)$$

- Optimal communication frequency: $p^\star = \frac{\lambda}{L+\lambda}$
- Optimal communication complexity:

$$\mathcal{O}\left(\frac{\min\{L,\lambda\}}{\mu}\log\frac{1}{\epsilon}\right)$$

Small $\lambda \Rightarrow$ optimal to take many local gradient steps between communication rounds

- $\lambda = \infty$: recovered communication complexity of GD
- $\lambda = 0$: no communication required (as desired)

Issue 2 resolved: Local optimization $\Rightarrow$ faster convergence

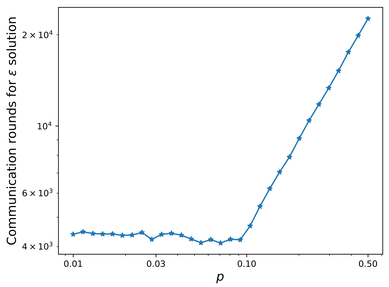

Figure 2:Communication rounds to get $\frac{F(x^k)-F(x^*)}{F(x^0)-F(x^*)} \leq 10^{-5}$ as a function of p with $p^* \approx 0.09$ ($\lambda = 0.1$).

References

[1] Sixin Zhang, Anna E Choromanska, and Yann LeCun.
Deep learning with elastic averaging sgd.
In *Advances in neural information processing systems*, pages 685–693, 2015.

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine.
Model-agnostic meta-learning for fast adaptation of deep networks.
In *International Conference on Machine Learning*, 2017.

[3] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik.
SGD: General analysis and improved rates.
In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209, Long Beach, California, USA, 09–15 Jun 2019. PMLR.