On Sample Optimality in Personalized Collaborative and Federated Learning

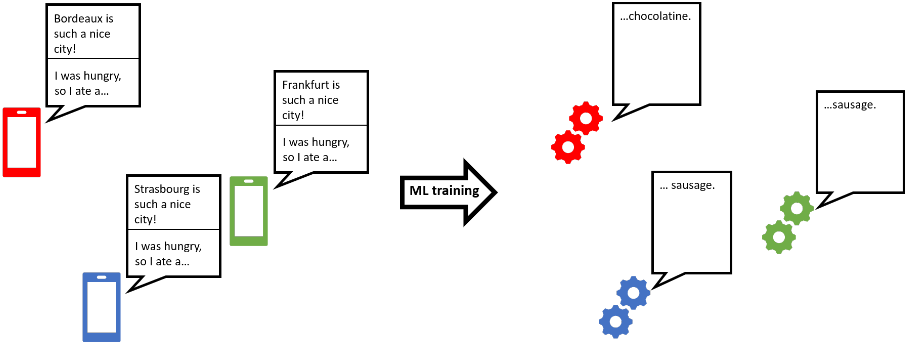
Neurips 2022

Mathieu Even, Laurent Massoulié, Kevin Scaman

Argo team, Inria Paris

Personalized Federated Learning (motivation)

- **Objective:** Train ML models from multiple data sources.
- One **local** model is learnt for each user, depending on its past activity.
- User datasets can be small, need to **collaborate**.



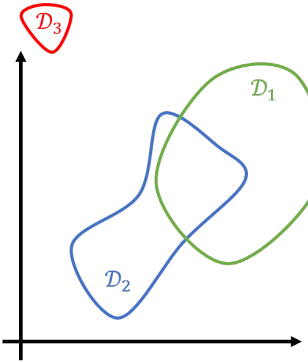
Personalized Federated Learning (setup)

Setup

- Let $(\mathcal{D}_i)_{i\in[\![1,N]\!]}$ be N data distributions on a space Ξ, and $\ell : \mathbb{R}^d \times \Xi \to \mathbb{R}$ a str. convex and smooth loss function. Our goal is to **minimize** the **local objective functions**

$$\forall i \in [\![1, N]\!], \qquad \min_{x\in\mathbb{R}^d} \quad f_i(x) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[\ell(x, \xi_i)\right]$$

- All agents receive a sample $\xi_{i,k} \sim \mathcal{D}_i$ at iteration $k > 0$.
- Agent i may compute and communicate **gradients** $g_i^k(x) = \nabla_x \ell(x, \xi_i^k)$ for any $x \in \mathbb{R}^d$.
- We focus on **sample complexity**.



Our objectives in this work

Theoretical questions

- How fast can we train our models?
- How does it depend on the data distributions?
- How to encode data dissimilarity?

Our objectives in this work

Theoretical questions

- How fast can we train our models?
- How does it depend on the data distributions?
- How to encode data dissimilarity?

Our contributions

- Lower and upper bounds on the **optimal sample complexity**
- IPMs can capture the **data dissimilarity** w.r.t. the optimization objective.
- **Gradient filtering** approaches are optimal while communication efficient!

Distances between distributions (1)

How to encode data dissimilarity in an optimization context?

Definition (Integral Probability Metrics, Muller, 1997)

For $\mathcal{H}$ a set of functions from Ξ to $\mathbb{R}^d$ and $\mathcal{D}, \mathcal{D}'$ two probability distributions on Ξ, let

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = \sup_{h \in \mathcal{H}} \left\| \mathbb{E}\left[h(\xi) - h(\xi') \right] \right\|$$

where $\xi \sim \mathcal{D}$ and $\xi' \sim \mathcal{D}'$. $d_{\mathcal{H}}$ is a pseudo-distance on the set of probability measures on Ξ.

Distances between distributions (1)

How to encode data dissimilarity in an optimization context?

Definition (Integral Probability Metrics, Muller, 1997)

For $\mathcal{H}$ a set of functions from Ξ to $\mathbb{R}^d$ and $\mathcal{D}, \mathcal{D}'$ two probability distributions on Ξ, let

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = \sup_{h \in \mathcal{H}} \left\| \mathbb{E}\left[h(\xi) - h(\xi')\right] \right\|$$

where $\xi \sim \mathcal{D}$ and $\xi' \sim \mathcal{D}'$. $d_{\mathcal{H}}$ is a pseudo-distance on the set of probability measures on Ξ.

Intuition

- Contains many standard distances for distributions, such as the Wasserstein (or earth mover's) distance, total variation, or maximum mean discrepancies.
- Measures how much a function class can distinguish the two distributions.

Distances between distributions (2)

Application to model training and optimization

- Most optimization algorithms rely on gradients to perform training.
- We want to measure how much **gradients see the two distributions as different**.
- We can take the function class $\mathcal{H}$ as our **knowledge on the gradients** $\nabla_x \ell(x, \xi)$!
- For example, for a quadratic models, the gradient is **linear**.

Assumption (Distribution-based dissimilarities)

Let $\mathcal{H}$ be such that, $\forall i = 1, \ldots, N$, and $x_i^\star$ a minimizer of f_i, we have

$$\left(\xi \in \Xi \mapsto \nabla_x \ell(x_i^\star, \xi)\right) \in \mathcal{H}$$

Moreover, there exists $(b_{ij})_{1 \leqslant i,j \leqslant N}$ such that, $\forall (i,j) \in [\![1, N]\!]^2$, $d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j) \leqslant b_{ij}$.

Our results (1)

Lower bound on the sample complexity

- Let $(b_{ij})_{ij}$ fixed non negative weights, $\varepsilon > 0$ target precision, and $i \in [\![1, N]\!]$ fixed.
- There exists "difficult" instantiations of our problem based on distributions $\mathcal{D}_1, \ldots, \mathcal{D}_N$ that verify the dissimilarity assumption for weights (b_{ij}), such that any "reasonable algorithm" that outputs a model x_i for user i using K_ε samples per agent, must verify:

$$K_\varepsilon \geqslant \frac{C}{\mathcal{N}_i^\varepsilon(b^2)},$$

where C is a constant that depends on the variance of local gradients noise and functions regularity assumptions, and $\mathcal{N}_i^\varepsilon(b^2)$ is the number of agents j that verify $b_{ij}^2 \leqslant \varepsilon$

Our results (2)

The All-for-all algorithm

Let $(W_{ij})_{1\leqslant i,j\leqslant N}$ be a $n \times n$ matrix with non negative entries and $\eta > 0$. Consider the iterates generated with $x^{k+1} = x^k - \eta W g^k$ *i.e.*,

$$x_i^{k+1} = x_i^k - \eta \sum_{j=1}^{N} W_{ij} \nabla_x \ell(x_j^k, \xi_j^k)$$

$\implies$ **Optimal collaboration speedup in average amongst clients, provided that η, W_{ij} tuned with b_{ij} from the IPM-based data-dissimilarity assumptions.**

Our results (3)

The estimation $d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j)$ based on S **samples** of each local distributions can be done up to a statistical precision that depends on the complexity of the function space $\mathcal{H}$: $1/\sqrt{S}$ for finite-dimensional $\mathcal{H}$ and some MMDs, $1/S^{1/d}$ for Wassertein distances, etc.

Case of quadratic linear regression

For a number S of samples $(\xi_i^s)_{i\in[n],s\in[S]}$, use the following estimates $\hat{\mu}_i, \hat{b}_{ij}$ and weights $W_{ij} = \hat{\lambda}_{ij}$ in the All-for-all algorithm:

$$\hat{\mu}_i = \frac{1}{S}\sum_{s=1}^{S} \xi_{i,s}\,, \quad \hat{b}_{ij} = \|\hat{\mu}_i - \hat{\mu}_j\|\,, \quad \hat{\lambda}_{ij} = \frac{\mathbb{1}_{[\![\hat{b}_{ij}^2 \leqslant u]\!]}}{\sum_{\ell=1}^{N} \mathbb{1}_{[\![\hat{b}_{i\ell}^2 \leqslant u]\!]}}\,.$$

$\Longrightarrow$ **Still optimal collaboration speedup under structural assumptions on the agents.**

Take home message

Conclusion

- Communication to **neighboring agents** w.r.t. $d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j)$ is sufficient, with a neighorhood radius that decreases with the desired precision ε.
- Best speedup **proportional to the number of neighbors** $\mathcal{N}_i^{\varepsilon}(b^2)$.
- This speedup can be achieved with **limited communication and local storage** with the All-for-all algorithm.
- In this setup, **no asymptotic speedup is possible** when all local distributions $\mathcal{D}_i$ differ (when $\varepsilon < \min_{ij} d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j)$, we have $\mathcal{N}_i^{\varepsilon}(b^2) = 1$).

For more details

Come at our poster and read our paper!

Thank you for your attention!