Federated Multi-Task Learning under a Mixture of Distributions

Othmane Marfoq [1,2,3] Giovanni Neglia [1,2] Aurélien Bellet [1] Laetitia Kameni [3] Richard Vidal [3]

[1]Inria [2]Université Côte d'Azur [3]Accenture Labs

The problem

- A (countable) set $\mathcal{T}$ of classification (or regression) tasks which represent the set of possible clients.
- Data $\mathcal{S}_t = \{s_t^{(i)} \triangleq (\mathbf{x}_t^{(i)},\ y_t^{(i)})\}_{i=1}^{n_t}$ at client t is drawn from a local distribution $\mathcal{D}_t$ over $\mathcal{X} \times \mathcal{Y}$.
- Client t wants to learn hypothesis $h_t^* \in \mathcal{H} = \{h : \mathcal{X} \mapsto \mathcal{Y}\}$

$$\underset{h_t \in \mathcal{H}}{\text{minimize}}\ \mathcal{L}_{\mathcal{D}_t}(h_t) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[l\left(h_t\left(\mathbf{x}\right), y\right)\right]. \tag{1}$$

- Personalized models for each client are a necessity in many *federated learning* (FL) applications.

Our goal is to study personalized federated learning under the flexible assumption that the data distribution of each client is a mixture of M underlying distributions.

An impossibility result

Some assumptions on the local data distributions $\mathcal{D}_t,\ t \in \mathcal{T}$, are needed for federated learning to be beneficial, because

1. Federated learning with T clients is equivalent to T *semi-supervised learning* (SSL) problems.
2. With no assumption on data distributions, SSL is impossible [1].

Main Assumptions

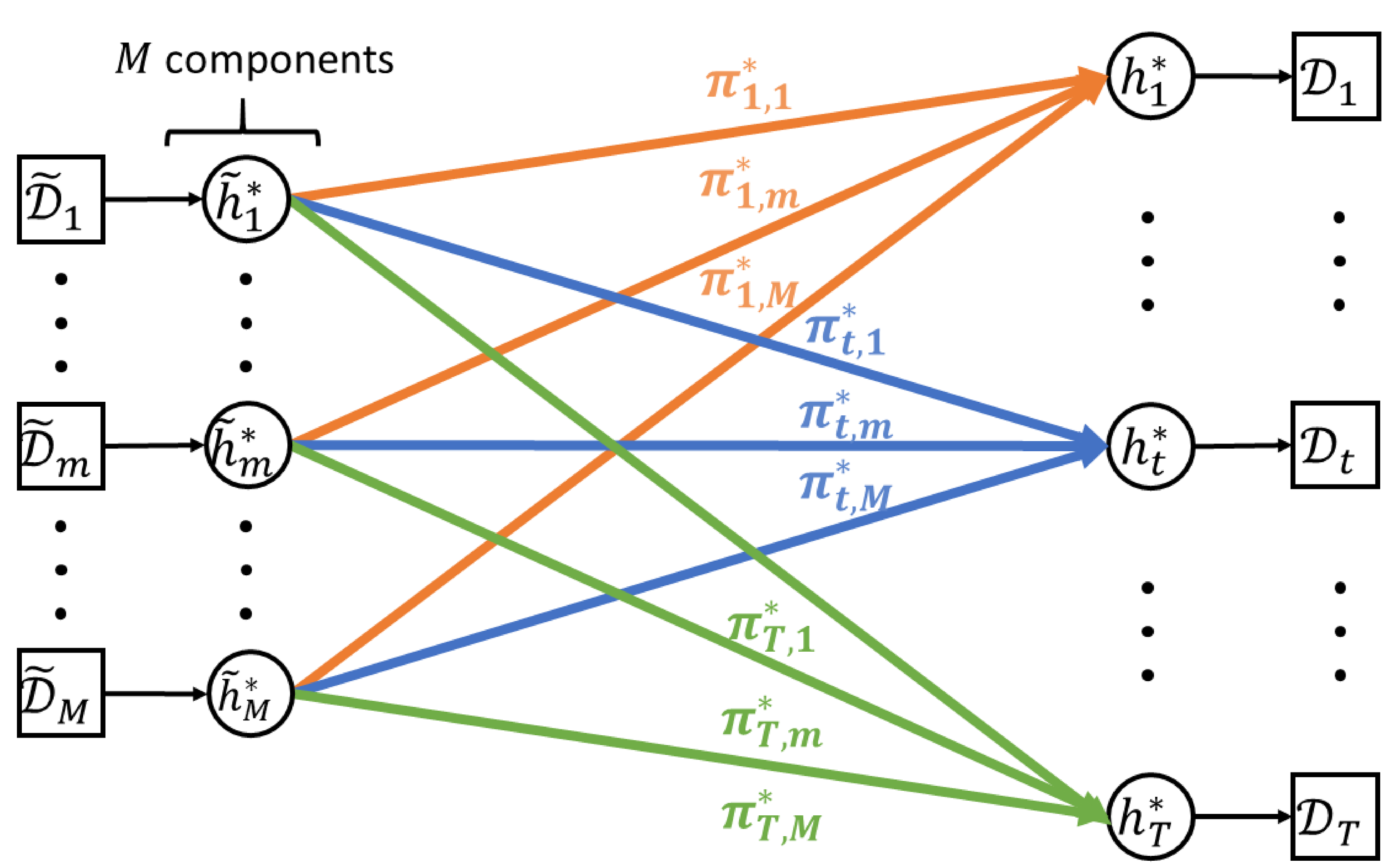
There exist M underlying (independent) distributions $\tilde{\mathcal{D}}_m,\ 1 \le m \le M$, such that for $t \in \mathcal{T}$, $\mathcal{D}_t$ is mixture of the distributions $\{\tilde{\mathcal{D}}_m\}_{m=1}^M$ with weights $\pi_t^* = [\pi_{t1}^*, \dots, \pi_{tM}^*] \in \Delta^M$, i.e.

$$z_t \sim \mathcal{M}(\pi_t^*), \quad ((\mathbf{x}_t, y_t)\,|z_t = m) \sim \tilde{\mathcal{D}}_m, \quad \forall t \in \mathcal{T}, \tag{2}$$

where $\mathcal{M}(\pi)$ is a multinomial (categorical) distribution with parameters π.

We consider d-dimensional parametric models:

$$\forall m \in [M],\ \exists \theta_m^* \in \mathbb{R}^d,\ l\left(h_{\theta_m^*}(\mathbf{x}), y\right) = -\log \tilde{\mathcal{D}}_m(y|\mathbf{x}) + c, \tag{3}$$



Remark

The generative model in Assumption 1 extends some popular multi-task/personalized FL formulation in the literature, including ***Clustered FL*** *[2],* ***Personalization via model interpolation*** *[3], and* ***Federated MTL via task relationships*** *[4].*

Main contributions

- Flexible assumption for personalized FL (mixtures of components).
- *Expectation-Maximization*-like learning algorithms with convergence guarantees (both in client-server and fully-decentralized settings).
- More general *federated surrogate optimization* framework.
- Higher accuracy and fairness than SOTA algorithms, even for clients not present at training time.

Learning under a Mixture Model

Proposition (informal). Let $\breve{\Theta}, \breve{\Pi} \in \arg\min_{\Theta,\Pi} \mathbb{E}_{t\sim\mathcal{D}_\mathcal{T}}\, \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[-\log \mathcal{D}_t(\mathbf{x}, y|\Theta, \pi_t)\right]$. Then,

$$h_t^* = \sum_{m=1}^M \breve{\pi}_{tm} h_{\breve{\theta}_m}, \quad \forall t \in \mathcal{T}. \tag{4}$$

This Proposition suggests the following approach to solve Problem (1).

- **First**, estimate $\breve{\Theta}$ and $\breve{\pi}_t,\ 1 \le t \le T$, by minimizing

$$f(\Theta, \Pi) \triangleq -\frac{\log \mathcal{D}_t(\mathcal{S}_{1:T}|\Theta, \Pi)}{n} \triangleq -\frac{1}{n}\sum_{t=1}^T \sum_{i=1}^{n_t} \log \mathcal{D}_t(s_t^{(i)}|\Theta, \pi_t). \tag{5}$$

- **Second**, use Eq. (4) to get the client predictor for the T clients present at training time.

Federated Expectation-Maximization

- A natural approach to solve problem (5) is via the Expectation-Maximization algorithm (EM), which alternates between two steps.

E-step:
$$q_t^{k+1}(z_t^{(i)} = m) \propto \pi_{tm}^k \cdot \exp\left(-l(h_{\theta_m^k}(\mathbf{x}_t^{(i)}), y_t^{(i)})\right), \quad t \in [T],\ m \in [M],\ i \in [n_t] \tag{6}$$

M-step:
$$\pi_{tm}^{k+1} = \frac{\sum_{i=1}^{n_t} q_t^{k+1}(z_t^{(i)} = m)}{n_t}, \quad t \in [T],\quad m \in [M] \tag{7}$$

$$\theta_m^{k+1} \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{t=1}^T \sum_{i=1}^{n_t} q_t^{k+1}(z_t^{(i)} = m)\, l\big(h_\theta(\mathbf{x}_t^{(i)}), y_t^{(i)}\big), \quad m \in [M] \tag{8}$$

- While the E-step (6) and the Π update (7) can be performed locally at each client, the Θ update (8) requires interaction with other clients.
- `FedEM` updates the local estimates of Θ through a solver which approximates the exact minimization in (8) using only the local dataset $\mathcal{S}_t$.

Theorem

When clients use SGD as local solver with learning rate $\eta = \frac{a_0}{\sqrt{K}}$, after a large enough number of communication rounds K, ***`FedEM`****'s iterates satisfy:*

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left\|\nabla_\Theta f\left(\Theta^k, \Pi^k\right)\right\|_F^2 \le \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \qquad \frac{1}{K}\sum_{k=1}^K \Delta_\Pi f(\Theta^k, \Pi^k) \le \mathcal{O}\left(\frac{1}{K^{3/4}}\right), \tag{9}$$

where the expectation is over the random batches samples, and

$$\Delta_\Pi f(\Theta^k, \Pi^k) \triangleq f\left(\Theta^k, \Pi^k\right) - f\left(\Theta^k, \Pi^{k+1}\right) \ge 0. \tag{10}$$

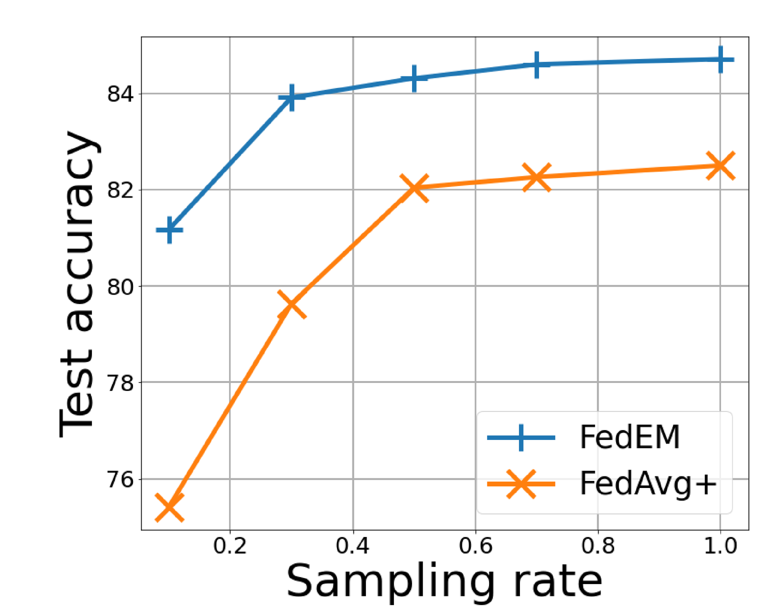
Surrogate Federated Optimization

- `FedEM` can be seen as a particular instance of a more general framework that we call *federated surrogate optimization*.
- This framework minimizes an objective function $\sum_{t=1}^T \omega_t f_t(\mathbf{u}, \mathbf{v}_t)$.
- Each client $t \in [T]$ can compute a partial first order surrogate of f_t.

Experiments

Dataset	Local	`FedAvg`	`FedProx`	`FedAvg+`	`Clustered FL`	`pFedMe`	`FedEM` (Ours)
FEMNIST	71.0 / 57.5	78.6 / 63.9	78.9 / 64.0	75.3 / 53.0	73.5 / 55.1	74.9 / 57.6	**79.9** / **64.8**
EMNIST	71.9 / 64.3	82.6 / 75.0	83.0 / 75.4	83.1 / 75.8	82.7 / 75.0	83.3 / 76.4	**83.5** / **76.6**
CIFAR10	70.2 / 48.7	78.2 / 72.4	78.0 / 70.8	82.3 / 70.6	78.6 / 71.2	81.7 / 73.6	**84.3** / **78.1**
CIFAR100	31.5 / 19.9	40.9 / 33.2	41.0 / 33.2	39.0 / 28.3	41.5 / 34.1	41.8 / 32.5	**44.1** / **35.0**
Shakespeare	32.0 / 16.6	**46.7** / 42.8	45.7 / 41.9	40.0 / 25.5	46.6 / 42.7	41.2 / 36.8	**46.7** / **43.0**
Synthetic	65.7 / 58.4	68.2 / 58.9	68.2 / 59.0	68.9 / 60.2	69.1 / 59.0	69.2 / 61.2	**74.7** / **66.7**

Table 1:Test accuracy: average across clients / bottom decile.



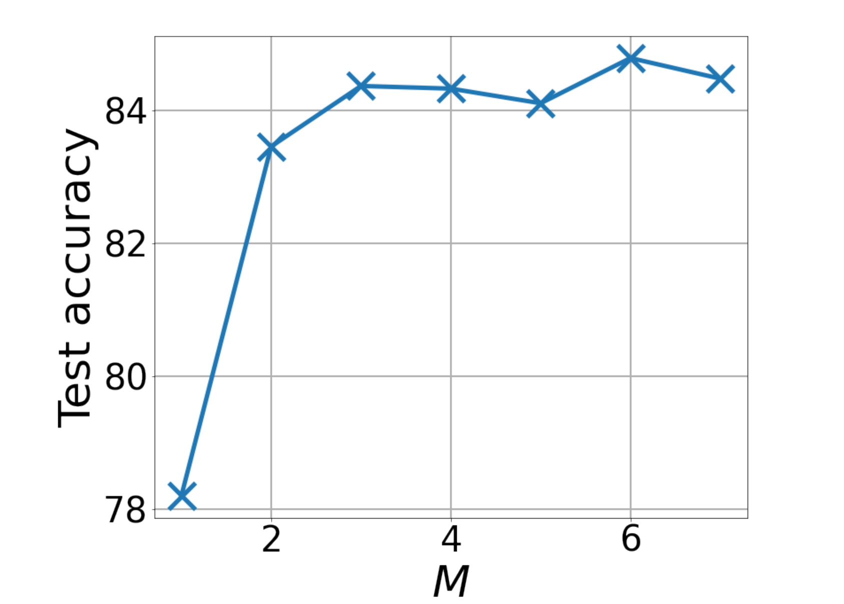


Figure 1:Effect of client sampling rate (left) and number of mixture components M (right) on test accuracy for CIFAR-10.

References

[1] Shai Ben-David, Tyler Lu, and D. Pál. ``Does Unlabeled Data Provably Help? Worst-case Analysis of the Sample Complexity of Semi-Supervised Learning''. In: *COLT*. 2008.

[2] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. ``Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints''. In: *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[3] Yishay Mansour et al. ``Three approaches for personalization with applications to federated learning''. In: *arXiv preprint arXiv:2002.10619* (2020).

[4] Virginia Smith et al. ``Federated Multi-Task Learning''. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4427–4437. ISBN: 9781510860964.