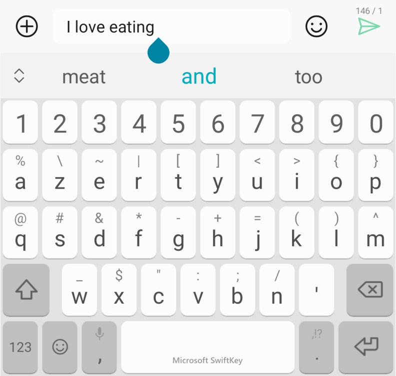
FEDERATED MULTI-TASK LEARNING UNDER A MIXTURE OF DISTRIBUTIONS

Aurélien Bellet (Inria)
Joint work with O. Marfoq, G. Neglia (Inria), L. Kameni, R. Vidal (Accenture Labs)

Google Federated Learning Workshop
November 8-10, 2021

PERSONALIZED FEDERATED LEARNING



- Personalized models are a necessity in many Federated Learning (FL) applications
- **Key questions:** how to model the relations between local data distributions? How to design efficient FL algorithms that exploit these relations?

BRIEF OVERVIEW OF RELATED WORK

- Local fine-tuning of a global model: [Jiang et al., 2019], [Fallah et al., 2020]...
- Interpolation of global and local model: [Deng et al., 2020], [Mansour et al., 2020]...

$\Rightarrow$ works only if local distributions are close from the global distribution

- Clustered FL: [Sattler et al., 2020], [Ghosh et al., 2020]...

$\Rightarrow$ no knowledge transfer across clusters

BRIEF OVERVIEW OF RELATED WORK

- Multi-task learning via task relationships [Smith et al., 2017], [Vanhaesebrouck et al., 2017] or simpler penalization terms [Hanzely et al., 2020], [Dinh et al., 2020]...

  $\Rightarrow$ limited to linear models or lose ability to model complex relationships

- Hypernetworks [Shamsian et al., 2021]

  $\Rightarrow$ flexible but potential blow up in the number of parameters

Overall: conditions under which users benefit from collaboration are not well understood

SUMMARY OF CONTRIBUTIONS

1. A flexible statistical assumption for personalized FL: local distributions are mixtures of underlying components
2. Federated EM-like algorithms with convergence guarantees, both in server-client and fully decentralized settings
3. A general federated surrogate optimization framework that can be used to analyze other FL algorithms
4. Higher accuracy and fairness than SOTA algorithms, even for users not present at training time

- A (countable) set $\mathcal{T}$ of tasks representing the set of possible users
- A data distribution $\mathcal{D}_t$ over $\mathcal{X} \times \mathcal{Y}$ for each user $t \in \mathcal{T}$ with $p_t(x, y)$ the joint density and $p_t(x)$, $p_t(y)$ the marginal densities
- User t wants to learn hypothesis $h_t \in \mathcal{H}$ minimizing the expected risk over $\mathcal{D}_t$:

$$\min_{h_t \in \mathcal{H}} \mathcal{L}_{\mathcal{D}_t}(h_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}[l(h_t(x), y)]$$

- A set of T users $[T] = \{1, \ldots, T\} \subseteq \mathcal{T}$ participate to the training phase
- Local dataset $\mathcal{S}_t = \{(x_t^{(i)}, y_t^{(i)})\}_{i=1}^{n_t}$ at user $t \in T$ drawn i.i.d. from $\mathcal{D}_t$

AN IMPOSSIBILITY RESULT

- Assume $p_t(x)$ is identical across $t \in T$, but $p_t(y|x)$ can be arbitrarily different
- FL with T users is then equivalent to T semi-supervised learning (SSL) problems
- With no assumptions on the data distribution, SSL does not improve sample complexity [Ben-David et al., 2008, Darnstädt et al., 2013, Göpfert et al., 2019]

$\Rightarrow$ some assumptions on local data distributions are needed for FL to be beneficial

PROPOSED ASSUMPTION

- For any user $t \in \mathcal{T}$, the local distribution $\mathcal{D}_t$ is a mixture of underlying distributions $\tilde{\mathcal{D}}_1, \dots, \tilde{\mathcal{D}}_M$ defined by weights $\pi^*_{t1}, \dots, \pi^*_{tm}$

Assumption

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.

$$z_t \sim \mathcal{M}(\pi^*_t), \quad ((x_t, y_t) | z_t = m) \sim \tilde{\mathcal{D}}_m, \quad \forall t \in \mathcal{T},$$

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

GENERALIZING EXISTING FRAMEWORKS

- Our assumptions generalizes previous personalized FL formulations
- Clustered FL [Sattler et al., 2020, Ghosh et al., 2020] with C clusters: set $M = C$ and $\pi^*_{tc} = 1$ if task (user) t is in cluster c and $\pi^*_{tc} = 0$ otherwise
- We also recover model interpolation [Deng et al., 2020, Mansour et al., 2020] and Fed-MTL with task relationships [Smith et al., 2017, Vanhaesebrouck et al., 2017] as special cases

Proposition (informal)

Let $\breve{\Theta} = [\breve{\theta}_1, \ldots, \breve{\theta}_M]$ and $\breve{\Pi} = [\breve{\pi}_1, \ldots, \breve{\pi}_{\mathcal{T}}]$ be a solution of

$$\arg\min_{\Theta,\Pi} \mathop{\mathbb{E}}_{t\sim D_{\mathcal{T}}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_t} [-\log p_t(x, y|\Theta, \pi_t)]$$

Then, for any $t \in \mathcal{T}$, we have:

$$h_t^* = \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m} \tag{1}$$

- We can estimate $\breve{\Theta}$ and $\breve{\Pi}$ by minimizing

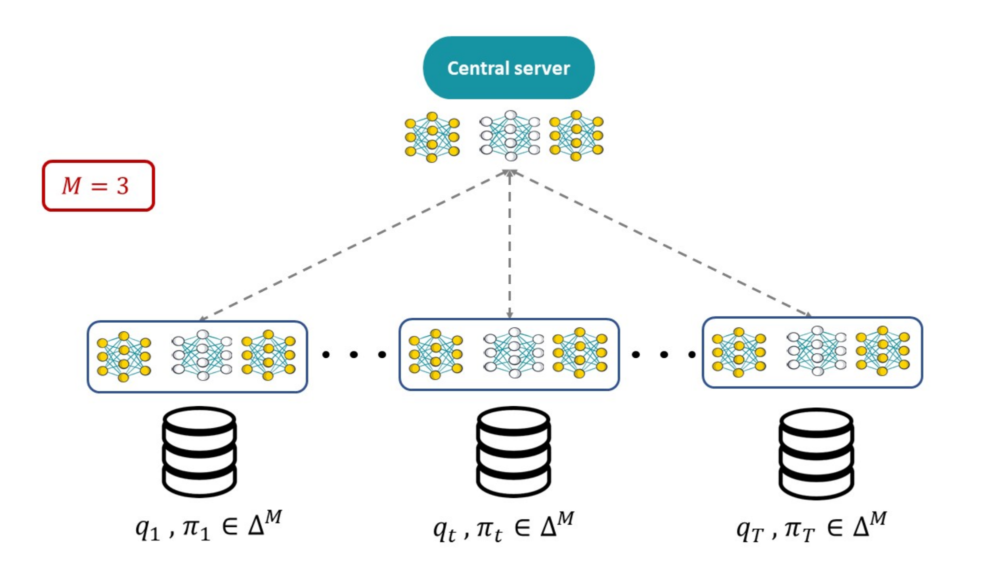
$$f(\Theta, \Pi) \triangleq -\frac{\log p(\mathcal{S}_{1:T}|\Theta, \Pi)}{n} \triangleq -\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n_t} \log p(s_t^{(i)}|\Theta, \pi_t),$$

- For a user t' not seen at training time: learn $\pi_{t'}$ in a single shot, and use (1)

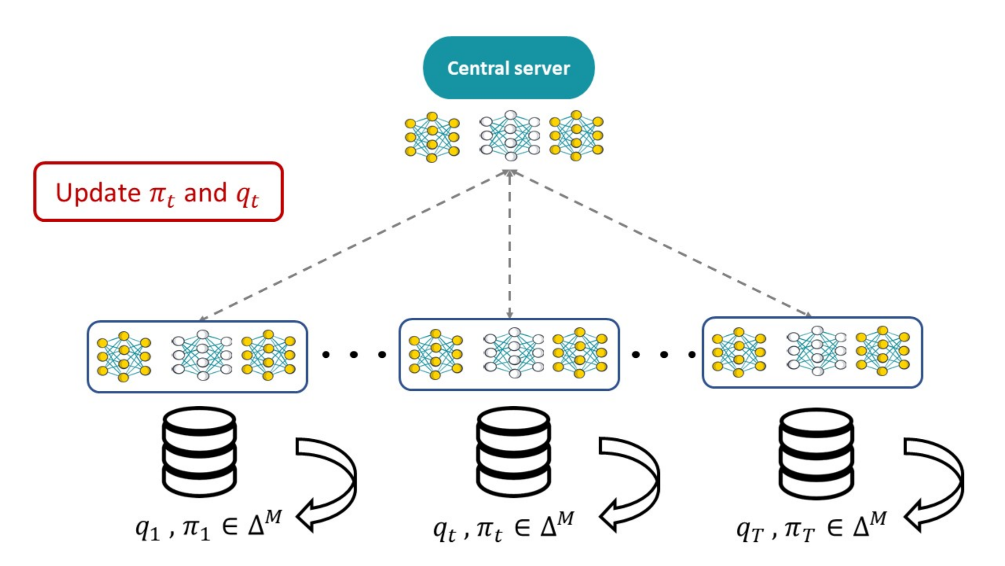
- Natural approach: Expectation-Maximization (EM) algorithm
- We denote by q_t the distribution over the latent variables $z_t^{(i)}$
- **E-step**: $q_t^{k+1}(z_t^{(i)} = m) \propto \pi_{tm}^k \cdot \exp\left(-l(h_{\theta_m^k}(\vec{x}_t^{(i)}), y_t^{(i)})\right)$
- **M-step:**

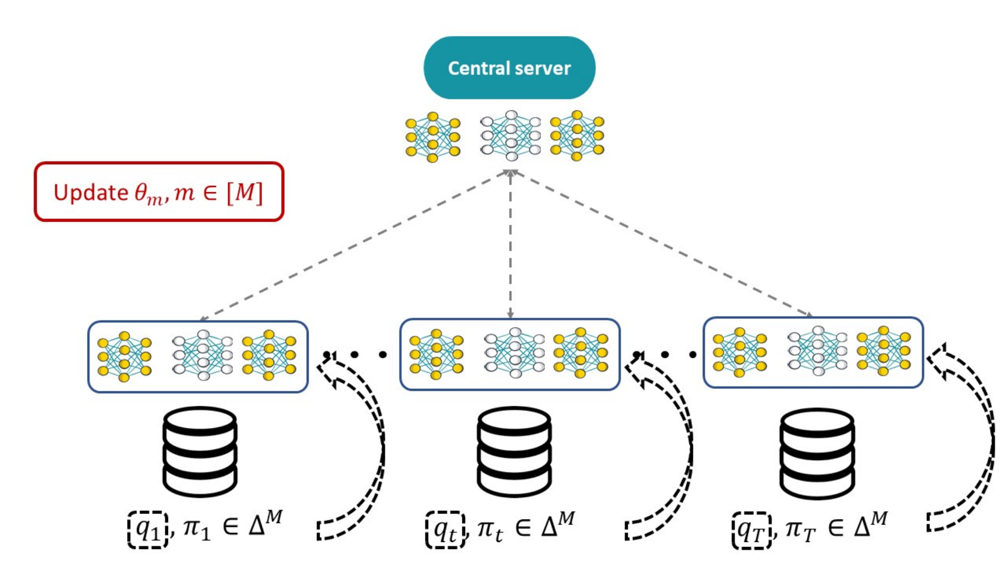
$$\pi_{tm}^{k+1} = \frac{\sum_{i=1}^{n_t} q_t^{k+1}(z_t^{(i)} = m)}{n_t}$$

$$\theta_m^{k+1} \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} \sum_{i=1}^{n_t} q_t^{k+1}(z_t^{(i)} = m) \cdot l\left(h_\theta(\vec{x}_t^{(i)}), y_t^{(i)}\right)$$


Central server
$M = 3$
$q_1 , \pi_1 \in \Delta^M$
$q_t , \pi_t \in \Delta^M$
$q_T , \pi_T \in \Delta^M$

FEDERATED EXPECTATION-MAXIMIZATION




Central server
Update $\theta_m, m \in [M]$
$q_1, \pi_1 \in \Delta^M$
$q_t, \pi_t \in \Delta^M$
$q_T, \pi_T \in \Delta^M$

Theorem (Informal)

With local SGD as the local solver, the iterates of `FedEM` *satisfy:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_{\Theta}f\left(\Theta^{k},\Pi^{k}\right)\right\|_{F}^{2}\leq\mathcal{O}\left(\frac{1}{\sqrt{K}}\right),$$

$$\frac{1}{K}\sum_{k=1}^{K}\Delta_{\Pi}f(\Theta^{k},\Pi^{k})\leq\mathcal{O}\left(\frac{1}{K^{3/4}}\right),$$

where the expectation is over the random batches samples, and

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

- `FedEM` can be seen as a particular instance of a more general framework that we call federated surrogate optimization, extending the centralized framework of [Mairal, 2013]
- This framework minimizes an objective function of the form $\sum_{t=1}^{T} \omega_t f_t\left(\vec{u}, \vec{v}_t\right)$
- Each user $t \in [T]$ can compute a partial first order surrogate of f_t
- Our framework can be used to analyze the convergence of other FL algorithms, such as `pFedMe` [Dinh et al., 2020] (see paper for details)

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 users / bottom decile.

Dataset	FedAvg	FedAvg+	FedEM
FEMNIST	78.3 (80.9)	74.2 (84.2)	**79.1** (81.5)
EMNIST	83.4 (82.7)	83.7 (92.9)	**84.0** (83.3)
CIFAR10	77.3 (77.5)	80.4 (80.5)	**85.9** (90.7)
CIFAR100	41.1 (42.1)	36.5 (55.3)	**47.5** (46.6)
Shakespeare	**46.7** (47.1)	40.2 (93.0)	**46.7** (46.6)
Synthetic	68.6 (70.0)	69.1 (72.1)	**73.0** (74.1)

Table 2: Average test accuracy across users unseen at training (train accuracy in parenthesis).

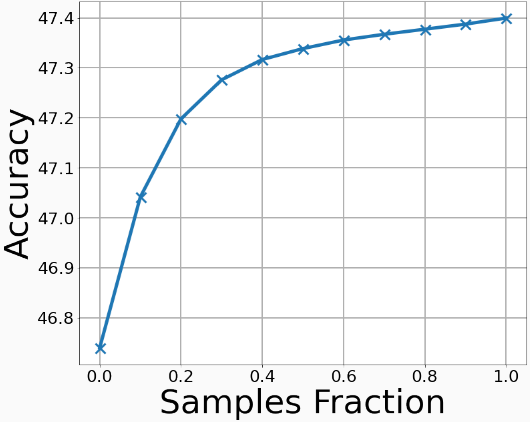


Figure 1: Effect of local dataset size on the average test accuracy across unseen users for CIFAR100.

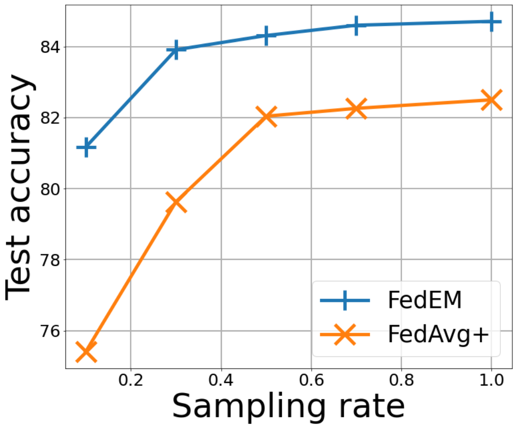


Figure 2: Effect of user sampling rate on the test accuracy for CIFAR10.

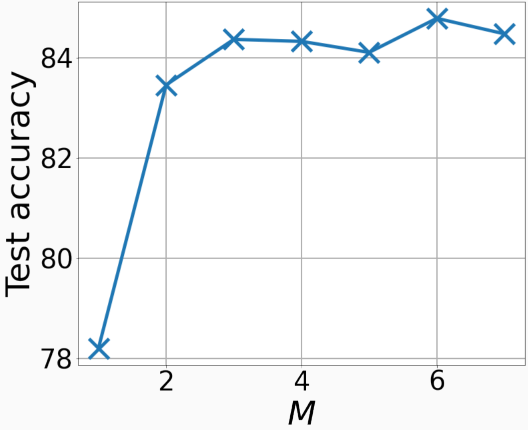


Figure 3: Effect of number of mixture components M on the test accuracy for CIFAR10.

THANK YOU FOR YOUR ATTENTION!

TO APPEAR AT NEURIPS 2021

ARXIV LINK: `https://arxiv.org/abs/2108.10252`

CODE: `https://github.com/omarfoq/fedem`

[Ben-David et al., 2008] Ben-David, S., Lu, T., and Pál, D. (2008).
Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning.
In *COLT*.

[Darnstädt et al., 2013] Darnstädt, M., Simon, H. U., and Szörényi, B. (2013).
Unlabeled data does provably help.
In *STACS*.

[Deng et al., 2020] Deng, Y., Kamani, M. M., and Mahdavi, M. (2020).
Adaptive personalized federated learning.
arXiv preprint arXiv:2003.13461.

[Dinh et al., 2020] Dinh, C. T., Tran, N. H., and Nguyen, T. D. (2020).
Personalized Federated Learning with Moreau Envelopes.
In *NeurIPS*.

[Fallah et al., 2020] Fallah, A., Mokhtari, A., and Ozdaglar, A. (2020).
Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach.
In *NeurIPS*.

[Ghosh et al., 2020] Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. (2020).
An efficient framework for clustered federated learning.
In *NeurIPS*.

[Göpfert et al., 2019] Göpfert, C., Ben-David, S., Bousquet, O., Gelly, S., Tolstikhin, I., and Urner, R. (2019).
When can unlabeled data improve the learning rate?
In *COLT*.

[Hanzely et al., 2020] Hanzely, F., Hanzely, S., Horváth, S., and Richtarik, P. (2020).
Lower Bounds and Optimal Algorithms for Personalized Federated Learning.
In *NeurIPS*.

[Jiang et al., 2019] Jiang, Y., Konečný, J., Rush, K., and Kannan, S. (2019).
Improving federated learning personalization via model agnostic meta learning.
arXiv preprint arXiv:1909.12488.

[Mairal, 2013] Mairal, J. (2013).
Optimization with first-order surrogate functions.
In *ICML*.

[Mansour et al., 2020] Mansour, Y., Mohri, M., Ro, J., and Suresh, A. T. (2020).
Three approaches for personalization with applications to federated learning.
arXiv preprint arXiv:2002.10619.

[Sattler et al., 2020] Sattler, F., Müller, K.-R., and Samek, W. (2020).
Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints.
IEEE Transactions on Neural Networks and Learning Systems.

REFERENCES III

[Shamsian et al., 2021] Shamsian, A., Navon, A., Fetaya, E., and Chechik, G. (2021).
Personalized federated learning using hypernetworks.
In *ICML*.

[Smith et al., 2017] Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. (2017).
Federated Multi-Task Learning.
In *NIPS*.

[Vanhaesebrouck et al., 2017] Vanhaesebrouck, P., Bellet, A., and Tommasi, M. (2017).
Decentralized collaborative learning of personalized models over networks.
In *AISTATS*.