Talk 3: Personalization in Federated Learning

WEN Hao

2021-5-27

Personalization for FL

When does one need personalization?

— When data across clients are “enough” non-IID, which is more realistic.

When does one need personalization?

— When data across clients are “enough” non-IID, which is more realistic.

Means of personalization:

- Federated Multi-Task Learning (+ regularization / proximal term), e.g. [1]
- Model-Agnostic Meta Learning, e.g. [2]
- Local Fine-tuning.

Personalization for FL

- Mixture of global and local [3]:

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2} \sum_{i=1}^{N} \|x_i - \overline{x}\|^2$$

- pFedMe (bi-level) [4] (and similarly EASGD[5]):

$$\text{minimize} \quad \sum_{i=1}^{N} F_i(x),$$

$$\text{where} \quad F_i(x) = \min \left\{ f_i(x_i) + \frac{\lambda}{2} \|x_i - x\|^2 \right\}$$

- FedU [6]:

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \|x_i - x_j\|^2$$

Mixture FL

The "weak" consensus problem (originally stated as "mixture" FL problem)

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2} \sum_{i=1}^{N} \|x_i - \bar{x}\|^2$$

can be reformulated as constrained optimization problems

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2} \sum_{i=1}^{N} \|x_i - z\|^2 \\ \text{subject to} \quad & Nz - \sum_{i=1}^{N} x_i = 0 \end{aligned}$$

Mixture FL

or equivalently as the following problem,

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2}\sum_{i=1}^{N}\|x_i\|^2 - \frac{\lambda N}{2}\|z\|^2 \\ \text{subject to} \quad & Nz - \sum_{i=1}^{N} x_i = 0 \end{aligned}$$

which is a nonconvex sharing problem considered in [7] (Eq. (3.2)). Note the difference of between formulations of a sharing problem in [7] (Section 3) and in [8] (Section 7.3)

The algorithm "Flexible ADMM" proposed in [7] (Algorithm 4) updates x_i using Gauss-Seidel method, which is non-trivial (or impossible) for parallelization. On the other hand, Jacobi method seems to have no guarantee of convergence.

Mixture FL

Under certain assumptions, this problem is a (split?) DC (difference-of-convex) programming problem with linear constraints.

$$
\begin{aligned}
\text{minimize} \quad & \boxed{\sum_{i=1}^{N}\left(f_i(x_i) + \frac{\lambda}{2}\|x_i\|^2\right)} - \lambda \boxed{\frac{N}{2}\|z\|^2} \\
\text{subject to} \quad & Nz - \sum_{i=1}^{N} x_i = 0
\end{aligned}
$$

One writes $\widetilde{f}_i(x_i) = f_i(x_i) + \dfrac{\lambda}{2}\|x_i\|^2$, and $r(z) = \frac{N}{2}\|z\|^2$.

Mixture FL

The original unconstrained problem is studied in [3] using the so-called looplesss local gradient descent (L2GD) method, with the assumptions that

- f_i are Lipschitz L-smooth

$$f(y) \leqslant f(x) + \langle \nabla f(x), (y-x) \rangle + \frac{L}{2}\|x-y\|^2$$

- f_i are μ-strongly convex

$$f(y) \geqslant f(x) + \langle \nabla f(x), (y-x) \rangle + \frac{\mu}{2}\|x-y\|^2$$

Looplessness is the one of the key contribution of [3], in which **inner (local) loops are replaced with probabilistic gradient updates**.

Mixture FL

Rewrite $\sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2}\sum_{i=1}^{N}\|x_i - \bar{x}\|^2$ as $f(x) + \psi(x)$ with $x = (x_1, \cdots, x_N)$, the local step of L2GD at a client i is

$$x^{k+1} = x^k - \alpha G(x^k)$$

where

$$G(x^k) = \begin{cases} \dfrac{\nabla f(x^k)}{1-p} & \text{with probability } 1-p \\ \dfrac{\lambda \nabla \psi(x^k)}{p} & \text{with probability } p \end{cases}$$

Locally, one has

$$x_i^{k+1} = x_i^k - \beta \nabla f_i(x_i^k), \quad x_i^{k+1} = (1-\gamma)x_i^k + \gamma \bar{x}^k$$

with probabilities $1-p$ and p respectively.

Mixture FL

Questions

1. Assumptions on the objective functions can be loosened?
2. DCA with linear constraints? (augmented) Lagrangian is

$$\mathcal{L}_\rho(x,z,y) = \sum_{i=1}^{N} \widetilde{f}_i(x_i) - \lambda r(z) + \langle y, Nz - \sum_{i=1}^{N} x_i \rangle + \boxed{\frac{\rho}{2}\|Nz - \sum_{i=1}^{N} x_i\|^2}$$

3. DCA (or stochastic, accelerated variants) can have better convergence?
4. more

Mixture FL — Research

Because of the existence of the boxed term $\frac{\rho}{2}\|Nz - \sum_{i=1}^{N} x_i\|^2$, the only choice to fit in the distributed settings is to update using the Jacobi method, as follows:

$$x_i^{k+1} = \arg\min_{x_i} \left\{ \widetilde{f}_i(x_i) - \langle y_i^k, x_i \rangle + \boxed{\frac{\rho}{2}\|Nz^k - \sum_{j \neq i}^{N} x_j^k - x_i\|^2} \right\}$$

$$z^{k+1} = \arg\min_{z} \left\{ \langle y^k, Nz \rangle + \frac{\rho}{2}\|Nz - \sum_{i=1}^{N} x_i^{k+1}\|^2 - \lambda r(z) \right\}$$

$$y^{k+1} = y^k + \beta(Nz^{k+1} - \sum_{i=1}^{N} x_i^{k+1})$$

Mixture FL — Research II

One can add more intermediate variables and reformulates the DC-like sharing problem as

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} \widetilde{f}_i(x_i) - \lambda r(z) \\ \text{subject to} \quad & x_i' = x_i, \quad i = 1, \cdots, N \\ & Nz = \sum_{i=1}^{N} x_i' \end{aligned}$$

Augmented Lagrangian of the above problem is

$$\sum_{i=1}^{N} \widetilde{f}_i(x_i) - \lambda r(z) + \sum_{i=1}^{N} \langle y_i, x_i' - x_i \rangle + \sum_{i=1}^{N} \frac{\rho_i}{2} \| x_i' - x_i \|^2 + \langle y, Nz - \sum_{i=1}^{N} x_i' \rangle + \frac{\rho}{2} \left\| Nz - \sum_{i=1}^{N} x_i' \right\|^2$$

ADMM iterations of the above problem are

$$
\begin{aligned}
x_i^{k+1} &= \arg\min_{x_i} \left\{ \widetilde{f}_i(x_i) + \frac{\rho_i}{2}\|x_i - (x_i')^k\|^2 - \langle y_i^k, x_i \rangle \right\} \\
(x_i')^{k+1} &= \arg\min_{x_i'} \left\{ \langle y_i^k - y^k, x_i' \rangle + \frac{\rho_i}{2}\|x_i^{k+1} - x_i'\|^2 \right. \\
&\qquad\qquad \left. + \boxed{\frac{\rho}{2}\|Nz^k - \sum_{j\neq i}^{N}(x_j')^k - x_i'\|^2} \right\} \\
z^{k+1} &= \arg\min_{z} \left\{ \langle y^k, Nz \rangle + \frac{\rho}{2}\|Nz - \sum_{i=1}^{N}(x_i')^{k+1}\|^2 - \lambda r(z) \right\} \\
y_i^{k+1} &= y_i^k + \beta((x_i')^{k+1} - x^{k+1}) \\
y^{k+1} &= y^k + \beta(Nz^{k+1} - \sum_{i=1}^{N}(x_i')^{k+1})
\end{aligned}
$$

Mixture FL — Research III

to update....

References I

[1] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated Multi-Task Learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4427–4437, 2017.

[2] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *International Conference on Machine Learning*, pp. 1126–1135, PMLR, 2017.

[3] F. Hanzely and P. Richtárik, "Federated Learning of a Mixture of Global and Local Models," *arXiv preprint arXiv:2002.05516*, 2020.

[4] C. T. Dinh, N. H. Tran, and T. D. Nguyen, "Personalized Federated Learning with Moreau Envelopes," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, (Red Hook, NY, USA), Curran Associates Inc., 2020.

References II

[5] S. Zhang, A. Choromanska, and Y. LeCun, "Deep Learning with Elastic Averaging SGD," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 685–693, 2015.

[6] C. T. Dinh, T. T. Vu, N. H. Tran, M. N. Dao, and H. Zhang, "FedU: A Unified Framework for Federated Multi-Task Learning with Laplacian Regularization," *arXiv preprint arXiv:2102.07148*, 2021.

[7] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.

[8] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers.*

Now Publishers Inc., 2011.