Talk 8: pFedMac

WEN Hao

2021-9-16

pFedMac - Motivations and Outlines

- personalization
  - via "maximizing correlation"

- communication efficiency
  - via "sparse" and/or "hierarchical" graphs

[1]Y. Li, X. Liu, X. Zhang, Y. Shao, Q. Wang, and Y. Geng, "Personalized Federated Learning via Maximizing Correlation with Sparse and Hierarchical Extensions," *arXiv preprint arXiv:2107.05330*, 2021

Previous Work - pFedMe

pFedMe (Personalized Federated Learning with Moreau Envelopes (or proximity operator)) is formulated as the following bi-level optimization problem in [2]

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

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

which is closely related to

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

[2]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

Previous Work - 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

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

which is a nonconvex sharing problem.

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

Maximizing Correlation - pFedMac

Personalization is formulated in pFedMac [1] as

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

$$\text{where} \quad F_i(x) = \min\left\{ \boxed{f_i(x_i) - \underbrace{\lambda\langle x_i, x\rangle}_{\text{correlation}}} + \frac{\lambda}{2}\|x\|^2 \right\}$$

$$= \min\left\{ \underbrace{f_i(x_i) + \frac{\lambda}{2}\|x_i - x\|^2}_{\text{pFedMe}} - \frac{\lambda}{2}\|x_i\|^2 \right\}$$

[1]Y. Li, X. Liu, X. Zhang, Y. Shao, Q. Wang, and Y. Geng, "Personalized Federated Learning via Maximizing Correlation with Sparse and Hierarchical Extensions," *arXiv preprint arXiv:2107.05330*, 2021

Maximizing Correlation - pFedMac

$$
\begin{aligned}
F_i(x) &= \min_{x_i} \{f_i(x_i) - \lambda\langle x_i, x\rangle\} + \frac{\lambda}{2}\|x\|^2 \\
&= \frac{\lambda}{2}\|x\|^2 - \max_{x_i} \{\langle x_i, \lambda x\rangle - f_i(x_i)\} \\
&= \frac{\lambda}{2}\|x\|^2 - f_i^*(\lambda x)
\end{aligned}
$$

where f_i^* is the conjugate function of f_i.

Comparison

Rewrite pFedMac

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

Comparison

Rewrite pFedMac

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

Mixture FL $\quad \frac{\lambda}{2}\|x_i\|^2 - \frac{\lambda}{2}\|x\|^2$

pFedMe $\quad \frac{\lambda}{2}\|x_i - x\|^2$

Sparsity Extension - pFedMac-S

By adding ℓ_1 penalty of local model parameters, one has the sparse extension of pFedMac (pFedMac-S)

$$\text{minimize} \sum_{i=1}^{N} \left(f_i(x_i) - \lambda \langle x_i, x \rangle + \frac{\lambda}{2} \|x\|^2 \boxed{+\gamma \|x_i\|_1} \right)$$

or

$$\text{minimize} \sum_{i=1}^{N} \left(f_i(x_i) + \frac{\lambda}{2} \|x_i - x\|^2 - \frac{\lambda}{2} \|x_i\|^2 \boxed{+\gamma \|x_i\|_1} \right)$$

Sparsity Extension - pFedMac-S

By adding ℓ_1 penalty of local model parameters, one has the sparse extension of pFedMac (pFedMac-S)

$$\text{minimize} \sum_{i=1}^{N} \left(f_i(x_i) - \lambda \langle x_i, x \rangle + \frac{\lambda}{2} \|x\|^2 \boxed{+\gamma \|x_i\|_1} \right)$$

or

$$\text{minimize} \sum_{i=1}^{N} \left(f_i(x_i) + \frac{\lambda}{2} \|x_i - x\|^2 - \frac{\lambda}{2} \|x_i\|^2 \boxed{+\gamma \|x_i\|_1} \right)$$

Sort of elastic net

Hierarchical Extension - pFedMac-SH

The hierarchy is

Client $\rightarrow$ **Edge** $\rightarrow$ **Central(Cloud)**

hence the problem is further split into 3 levels

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

$$\text{where} \quad F_i(x) = \min_{x_i} \left\{ \frac{1}{J} \sum_{j=1}^{J} F_{i,j}(x_i) - \lambda_2 \langle x_i, x \rangle + \frac{\lambda_2}{2} \|x\|^2 \right\}$$

$$\text{and} \quad F_{i,j}(x_i) = \min_{x_{i,j}} \left\{ f_{i,j}(x_{i,j}) - \lambda_1 \langle x_{i,j}, x_i \rangle + \frac{\lambda_1}{2} \|x_i\|^2 \boxed{+\gamma \|x_{i,j}\|_1} \right\}$$

Algorithm 1: pFedMac-SH

Cloud server executes:

for $t = 0, 1, \cdots, T-1$ **do**

  for $i = 1, 2, \cdots, N$ *in parallel* **do**

    $x_i^{t+1} \leftarrow$ **EdgeUpdate**(i, x^t)

  $\mathcal{S}^t \leftarrow$ (random set of S edge servers)

  $x^{t+1} \leftarrow (1-\beta)x^t + \beta \frac{1}{S} \sum_{i \in \mathcal{S}^t} x_i^{t+1}$ averaging

EdgeUpdate(i, x^t):

$y_i^{t,0} = x_i^{t,0} = x^t$

for $r = 0, 1, \cdots, R-1$ **do**

  for $j = 1, 2, \cdots, J$ **do**

    $y_{i,j}^{t,r+1} \leftarrow$ **ClientUpdate**(j, $y_{i,j}^{t,r}$)

  $y_i^{t,r+1} \leftarrow \frac{1}{J} \sum_{j=1}^{J} y_{i,j}^{t,r+1}$

  $x_i^{t,r+1} \leftarrow x_i^{t,r} - \eta_2 \lambda_2 (x_i^{t,r} - y_i^{t,r+1})$ (S)GD

  return $x_i^{t,R}$

pFedMac - Algorithm Continued

Algorithm 1: pFedMac-SH

ClientUpdate(j, $y_{i,j}^{t,r}$):

$y_{i,j}^{t,r,0} \leftarrow y_{i,j}^{t,r}$

for $k = 0, 1, \cdots, K-1$ **do**

$\mathcal{D}_{i,j}^{k} \leftarrow$ (sample a mini batch with size D)

$\widetilde{\theta}_{i,j}^{t,r,k} \leftarrow$

$$\arg\min_{\theta_{i,j}} \left\{ \widetilde{f}_{i,j}(\theta_{i,j}; \mathcal{D}_{i,j}^{k}) - \lambda_1 \langle \theta_{i,j}, y_{i,j}^{t,r,k} \rangle + \gamma_1 \boxed{\phi_\rho(\theta_{i,j})} \right\}$$

$y_{i,j}^{t,r,k+1} \leftarrow y_{i,j}^{t,r,k} - \eta_1 \lambda_2 (y_{i,j}^{t,r,k} - \widetilde{\theta}_{i,j}^{t,r,k}) \quad (S)GD$

return $y_{i,j}^{t,r,K}$

$\widetilde{f}_{i,j}(\theta_{i,j}; \mathcal{D}_{i,j}^{k})$ denotes objective function evaluated using data $\mathcal{D}_{i,j}^{k}$. ϕ_ρ is a twice continuously differentiable approximation of the ℓ_1-norm:

$$\phi_\rho(z) = \rho \sum_{n=1}^{d} \log\cosh\left(\frac{z_n}{\rho}\right) = \rho \sum_{n=1}^{d} \log\left(\frac{\exp(z_n/\rho) + \exp(-z_n/\rho)}{2}\right)$$

pFedMac - Algorithm Continued

pFedMac

Personalization

Communication

Problems

The inner-most optimization problem

$$\arg\min_{\theta_{i,j}} \left\{ \widetilde{f}_{i,j}(\theta_{i,j}; \mathcal{D}_{i,j}^{k}) - \lambda_1 \langle \theta_{i,j}, y_{i,j}^{t,r,k} \rangle + \gamma_1 \phi_\rho(\theta_{i,j}) \right\}$$

can be solved using first-order methods, ref. [4]

[4]Z. Lin, H. Li, and C. Fang, *Accelerated Optimization for Machine Learning*. Springer, 2020

The inner-most optimization problem

$$\arg\min_{\theta_{i,j}} \left\{ \widetilde{f}_{i,j}(\theta_{i,j}; \mathcal{D}_{i,j}^{k}) - \lambda_1 \langle \theta_{i,j}, y_{i,j}^{t,r,k} \rangle + \gamma_1 \phi_\rho(\theta_{i,j}) \right\}$$

can be solved using first-order methods, ref. [4]

An alternative to smoothing ℓ_1-norm with ϕ_ρ, one can use proximal gradient method via updating by

$$\theta_{i,j}^{(s+1)} = \mathrm{prox}_{h,\alpha_s}(\theta_{i,j}^{(s)} - \alpha_s \nabla g_{i,j}(\theta_{i,j}^{(s)}))$$

where

$$h(\theta_{i,j}) = \gamma_1 \|\theta_{i,j}\|_1$$

$$g_{i,j}(\theta_{i,j}) = \widetilde{f}_{i,j}(\theta_{i,j}; \mathcal{D}_{i,j}^{k}) - \lambda_1 \langle \theta_{i,j}, y_{i,j}^{t,r,k} \rangle$$

Or using the Huber regularization.

[4]Z. Lin, H. Li, and C. Fang, *Accelerated Optimization for Machine Learning*. Springer, 2020

Problems

- vanilla pFedMac: guarantee of finiteness of local (inner) problem $F_i(x) = \frac{\lambda}{2}\|x\|^2 - f_i^*(\lambda x)$, or finiteness of f_i^* for arbitrary f_i?
- theoretical comparison (convergence, performance(?), etc.) of personalization methods (models), including "mixture FL", "pFedMe", "pFedMac"
- hierarchical ADMM (if split)?
- hierarchical decentralized version (i.e. no central cloud server), and combination with techniques including gradient tracking and compression, etc.

References I

[1] Y. Li, X. Liu, X. Zhang, Y. Shao, Q. Wang, and Y. Geng, "Personalized Federated Learning via Maximizing Correlation with Sparse and Hierarchical Extensions," *arXiv preprint arXiv:2107.05330*, 2021.

[2] 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.

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

[4] Z. Lin, H. Li, and C. Fang, *Accelerated Optimization for Machine Learning*.

Springer, 2020.