Federated Conditional Stochastic Optimization

Xidong Wu*[1], Jianhui Sun*[2], Zhengmian Hu[3], Junyi Li[1], Aidong Zhang[2], Heng Huang[3]
[1] University of Pittsburgh [2] University of Virginia [3] University of Maryland

NEURAL INFORMATION PROCESSING SYSTEMS

Background

Conditional stochastic optimization has applications in a wide range of machine learning tasks, such as invariant learning, AUPRC maximization, and meta-learning.

- Model-Agnostic Meta-Learning (MAML)

In meta-learning, we attempt to train models that can efficiently adapt to unseen tasks via learning with meta data from similar tasks. When the tasks are distributed at different clients, a federated version of MAML would be beneficial to leverage information from all workers to improve the model performance on the downstream tasks.

$$\min_x \mathbb{E}_{i\sim\mathcal{P}_{\text{task}},a\sim D^i_{\text{query}}} \mathcal{L}_i\left(\mathbb{E}_{b\sim D^i_{\text{support}}}\left(x-\lambda\nabla\mathcal{L}_i(x,b)\right),a\right)$$

- Online AUPRC Maximization In AUPRC maximization, AP loss is used as the loss function, instead of cross-entropy loss since AP is the surrogate function of AUPRC and cross-entropy is corresponding to accuracy. By directly optimizing AP loss, model performance is improved with the metric of AUPRC.

$$\hat{\text{AP}} = \mathbb{E}_{\xi^+\sim\mathcal{D}^+}\frac{\mathbb{E}_{\xi\sim\mathcal{D}}\mathbf{I}\left(y=1\right)\ell\left(x;z^+,z\right)}{\mathbb{E}_{\xi\sim\mathcal{D}}\quad\ell\left(x;z^+,z\right)}$$

Challenges

To address the large-scale distributed data challenges across multiple clients with communication-efficient distributed training, federated learning (FL) is gaining popularity. Many optimization algorithms for CSO problems have been developed in the centralized setting. Nonetheless, the algorithm for CSO problems under FL is still underexplored.

$$\min_{x\in\mathcal{X}} F(x) := \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\xi^n} f^n_{\xi^n}\left(\mathbb{E}_{\eta^n|\xi^n} g^n_{\eta^n}(x,\xi^n)\right), \tag{1}$$

Federated conditional stochastic optimization subsumes the standard federated learning optimization as a special case when the inner-layer function $g^n_{\eta^n}(x,\xi^n)=x$. In addition, federated stochastic compositional optimization is similar to federated conditional stochastic optimization given that both problems contain two-layer nested expectations. However, they are fundamentally different. In federated stochastic compositional optimization, the inner randomness η and the outer randomness ξ are independent and data samples of the inner layer are available directly from η (instead of a conditional distribution as in Problem (1)). Therefore, when randomnesses η and ξ are independent and $g^n_{\eta^n}(x,\cdot)=g^n_{\eta^n}(x)$, (1) is converted into federated stochastic compositional optimization.

Existing Methods

Summary of complexity results of proposed federated conditional stochastic optimization algorithms to reach an ε-stationary point. Sample complexity is defined as the number of calls to the First-order Oracle (IFO) by clients to reach an ε-stationary point. Communication complexity denotes the total number of back-and-forth communication rounds between each client and the central server required to reach an ε-stationary point.

Algorithm	Sample	Communication
FCSG	$O\left(\epsilon^{-6}\right)$	$O\left(\epsilon^{-3}\right)$
FCSG-M	$O\left(\epsilon^{-6}\right)$	$O\left(\epsilon^{-3}\right)$
Theoretical Lower Bound	$O\left(\epsilon^{-5}\right)$	-
Acc-FCSG-M	$O\left(\epsilon^{-5}\right)$	$O\left(\epsilon^{-2}\right)$

FedSGDA+ Algorithm

Algorithm 1 FCSG and FCSG-M Algorithm

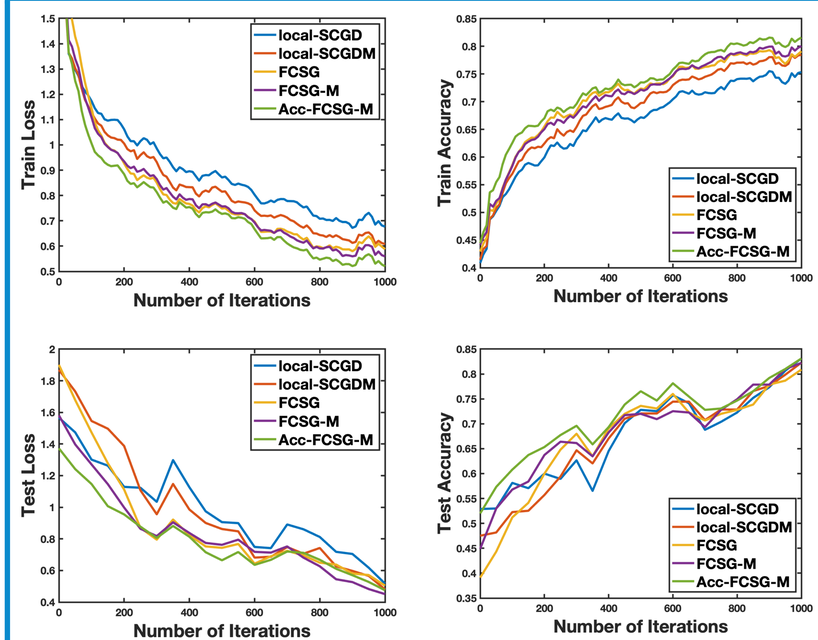
Input: Parameters: T, momentum weight β, learning rate α, the number of local updates q, inner batch size m and outer batch size b, as well as the initial outer batch size B ;
Initialize: $x_0^n = \bar{x}_0 = \frac{1}{N}\sum_{k=1}^N x_0^n$. Draw B samples of $\{\xi^n_{t,1},\cdots,\xi^n_{t,B}\}$ and draw m samples $\mathcal{B}^n_{0,i} = \left\{\eta^n_{ij}\right\}_{j=1}^m$ from $P(\eta^n \mid \xi^n_{0,i})$ for each $\xi^n_{0,i} \in \{\xi^n_{t,1},\cdots,\xi^n_{t,B}\}$; $u_1^n = \frac{1}{B}\sum_{i=1}^B \nabla\hat{F}^n(x_0^n;\xi^n_{0,i},\mathcal{B}^n_{0,i})$.
for $t=1,2,\ldots,T$ **do**
  for $n=1,2,\ldots,N$ **do**
    if $\mod(t,q)=0$ **then**
      Server Update:
      $u_t^n = \bar{u}_t = \frac{1}{N}\sum_{i=1}^N u_t^n$
      $x_t^n = \bar{x}_t = \frac{1}{N}\sum_{n=1}^N(x_{t-1}^n - \alpha u_t^n)$
    else
      $x_t^n = x_{t-1}^n - \alpha u_t^n$
    end if
    Draw b samples of $\{\xi^n_{t,1},\cdots,\xi^n_{t,b}\}$
    Draw m samples $\mathcal{B}^n_{t,n} = \left\{\eta^n_{ij}\right\}_{j=1}^m$ from $P(\eta^n \mid \xi^n_{t,i})$ for each $\xi^n_{t,i} \in \{\xi^n_{t,1},\cdots,\xi^n_{t,b}\}$,
    $u_{t+1}^n = \frac{1}{b}\sum_{i=1}^b \nabla\hat{F}^n(x_t^n;\xi^n_{t,i},\mathcal{B}^n_{t,i})$
    $u_{t+1}^n = (1-\beta)u_t^n + \frac{\beta}{b}\sum_{i=1}^b \nabla\hat{F}^n(x_t^n;\xi^n_{t,i},\mathcal{B}^n_{t,i})$
  end for
end for
Output: x chosen uniformly random from $\{\bar{x}_t\}_{t=1}^T$.

FedSGDA-M Algorithm

Algorithm 2 Acc-FCSG-M Algorithm

Initialize: Draw B samples of $\{\xi^n_1,\cdots,\xi^n_B\}$ and draw m samples $\mathcal{B}^n_{0,i} = \left\{\eta^n_{ij}\right\}_{j=1}^m$ from $P(\eta^n \mid \xi^n_i)$ for each $\xi^n_i \in \{\xi^n_1,\cdots,\xi^n_B\}$, then $u_1^n = \frac{1}{B}\sum_{i=1}^B \nabla\hat{F}^n(x_0^n;\xi^n_{0,i},\mathcal{B}^n_{0,i})$ for $n\in[N]$.
for $t=1,2,\ldots,T$ **do**
  for $n=1,2,\ldots,N$ **do**
    if $\mod(t,q)=0$ **then**
      $u_t^n = \bar{u}_t = \frac{1}{N}\sum_{i=1}^N u_t^n$, $x_t^n = \bar{x}_t = \frac{1}{N}\sum_{n=1}^N(x_{t-1}^n - \alpha u_t^n)$
    else
      $x_{t,n} = x_{t-1}^n - \alpha u_t^n$
    end if
    Draw b samples of $\{\xi^n_{t,1},\cdots,\xi^n_{t,b}\}$, draw m samples $\mathcal{B}^n_{t,n} = \left\{\eta^n_{ij}\right\}_{j=1}^m$ from $P(\eta^n \mid \xi^n_{t,i})$ for each $\xi^n_{t,i} \in \{\xi^n_{t,1},\cdots,\xi^n_{t,B}\}$,
    $u_{t+1}^n = \frac{1}{b}\sum_{i=1}^b \nabla\hat{F}^n(x_t^n;\xi^n_{t,i},\mathcal{B}^n_{t,i}) + (1-\beta)(u_t^n - \frac{1}{b}\sum_{i=1}^b \nabla\hat{F}^n(x_{t-1}^n;\xi^n_{t,i},\mathcal{B}^n_{t,i}))$
  end for
end for

Experiments



The 5-way-1-shot case over Omniglot Dataset in MAML (Train loss; Train accuracy; Test loss; Test accuracy)