On the Convergence of Local Stochastic Compositional Gradient Descent with Momentum

Hongchang Gao[1], Junyi Li[2], Heng Huang[2]

[1]Department of Computer and Information Sciences, Temple University, USA
[2] Department of Electrical and Computer Engineering, University of Pittsburgh, USA

Outline

Federated Stochastic Compositional Optimization Problem

- Federated Stochastic Compositional Optimization Problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\zeta\sim\mathcal{D}_f^{(k)}}\left[f^{(k)}\left(\mathbb{E}_{\xi\sim\mathcal{D}_g^{(k)}}[g^{(k)}(x;\xi)];\zeta\right)\right]. \tag{1}$$

  - $f^{(k)}(y) \triangleq \mathbb{E}_{\zeta\sim\mathcal{D}_f^{(k)}}[f^{(k)}(y;\zeta)] \in \mathbb{R}$ is the outer-level function on the k-th device
  - $g^{(k)}(x) \triangleq \mathbb{E}_{\xi\sim\mathcal{D}_g^{(k)}}[g^{(k)}(x;\xi)] \in \mathbb{R}^{d'}$ is the inner-level function on the k-th device

- Examples:
  - Policy evaluation
  - Sparse additive models
  - Model-agnostic meta-learning

Model-agnostic meta-learning

- Model-agnostic meta-learning

$$\min_{x\in\mathbb{R}^d} \frac{1}{K}\sum_{k=1}^{K} F^{(k)}(x) \triangleq \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(g^{(k)}(x)) \ ,$$
$$\text{where } g^{(k)}(x) = \mathbb{E}_{\xi^{(k)}\sim\mathcal{D}_{i,\text{train}}^{(k)}}\left[x - \lambda\nabla\mathcal{L}_i^{(k)}\left(x;\xi^{(k)}\right)\right] \ , \tag{2}$$
$$f^{(k)}(x) = \mathbb{E}_{i\sim\mathcal{P}_{\text{task}}^{(k)},\zeta^{(k)}\sim\mathcal{D}_{i,\text{test}}^{(k)}}\left[\mathcal{L}_i^{(k)}\left(y;\zeta^{(k)}\right)\right] \ .$$

  - $\mathcal{P}_{\text{task}}^{(k)}$ is the task distribution
  - λ is the learning rate
  - $\mathcal{D}_{i,\text{train}}^{(k)}$ is the training set and $\mathcal{D}_{i,\text{test}}^{(k)}$ is the testing set of the i-th task.

Challenges

- Challenges: Stochastic gradient is a biased estimation for the full gradient

$$\mathbb{E}_{\xi,\zeta}[\nabla g(x;\xi)^T \nabla_g f(g(x;\xi);\zeta)] \neq \nabla g(x)^T \nabla_g f(g(x)) \ . \tag{3}$$

- Existing methods
  - Federated learning methods:
    - Local-BSGD [Huang et al., 2021]: $O(1/\epsilon^8)$ sample complexity
    - Local-MOML [Wang et al., 2021]: $O(1/\epsilon^5)$ sample complexity
  - Single-machine methods:
    - NASA [Ghadimi et al., 2020]: $O(1/\epsilon^4)$ sample complexity
- Q: Can we have a better sample complexity under the federated learning setting?

Local-SCGDM

Algorithm 1 Local-SCGDM

for $t = 0, \cdots, T-1$ **do**
  if $t == 0$ **then**
    $u_1^{(k)} = g^{(k)}(x_0^{(k)}; \xi_0^{(k)})$, $z_0^{(k)} = \nabla g^{(k)}(x_0^{(k)}; \xi_0^{(k)})^T \nabla_g f^{(k)}(u_1^{(k)}; \zeta_0^{(k)})$, $m_1^{(k)} = z_0^{(k)}$,
  else
    $u_{t+1}^{(k)} = (1-\gamma\eta)u_t^{(k)} + \gamma\eta g^{(k)}(x_t^{(k)}; \xi_t^{(k)})$,
    $z_t^{(k)} = \nabla g^{(k)}(x_t^{(k)}; \xi_t^{(k)})^T \nabla_g f^{(k)}(u_{t+1}^{(k)}; \zeta_t^{(k)})$, //stochastic compositional gradient
    $m_{t+1}^{(k)} = (1-\alpha\eta)m_t^{(k)} + \alpha\eta z_t^{(k)}$, // momentum
  end if
  $x_{t+1}^{(k)} = x_t^{(k)} - \beta\eta m_{t+1}^{(k)}$,
  if $\mathrm{mod}(t+1, p) == 0$ **then**
    $u_{t+1}^{(k)} = \frac{1}{K}\sum_{k'=1}^{K} u_{t+1}^{(k')}$, $m_{t+1}^{(k)} = \frac{1}{K}\sum_{k'=1}^{K} m_{t+1}^{(k')}$, $x_{t+1}^{(k)} = \frac{1}{K}\sum_{k'=1}^{K} x_{t+1}^{(k')}$,
  end if
end for

Convergence Rate of Local-SCGDM

Theorem

Suppose Assumption 3.1-3.3 hold, by setting $\eta = T^{-1/2}$, $p = T^{1/4}$, *Algorithm 1 has the following convergence rate:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] \leq \frac{2(F(x_0)-F(x_*))}{\beta\sqrt{T}} + \frac{6(C_g^2\sigma_f^2 + C_g^2L_f^2\delta^2 + C_f^2\sigma_g^2)}{\alpha\sqrt{T}} + \frac{4C_g^2L_g^2\delta^2}{\gamma\sqrt{T}} + \frac{2\gamma C_g^2L_f^2\delta_g^2}{\sqrt{T}} + \frac{2\alpha(C_g^2\sigma_f^2 + C_f^2\sigma_g^2)}{\sqrt{T}} + \frac{2\gamma^2C_g^2L_f^2\delta_g^2}{T} + \frac{32\beta^2L_F^2(16C_g^2L_f^2\delta_g^2 + C_g^2\sigma_f^2 + 3C_f^2\sigma_g^2)}{\sqrt{T}}. \tag{4}$$

Convergence Rate of Local-SCGDM

- To make $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\bar{x}_t)\| \leq \epsilon$, T should be as large as $O(1/\epsilon^4)$.

Methods	Iteration	Batch size	Sample	Period	Commuication
Local-BSGD	$\mathcal{O}(\frac{1}{\epsilon^4})$	$\mathcal{O}(\frac{1}{\epsilon^4})$	$\mathcal{O}(\frac{1}{\epsilon^8})$	$\mathcal{O}(1)$	$\mathcal{O}(\frac{1}{\epsilon^4})$
Local-MOML	$\mathcal{O}(\frac{1}{\epsilon^5})$	$\mathcal{O}(1)$	$\mathcal{O}(\frac{1}{\epsilon^5})$	$\mathcal{O}(\frac{1}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon^3})$
Ours	$\mathcal{O}(\frac{1}{\epsilon^4})$	$\mathcal{O}(1)$	$\mathcal{O}(\frac{1}{\epsilon^4})$	$\mathcal{O}(\frac{1}{\epsilon})$	$\mathcal{O}(\frac{1}{\epsilon^3})$

Sinewave Regression

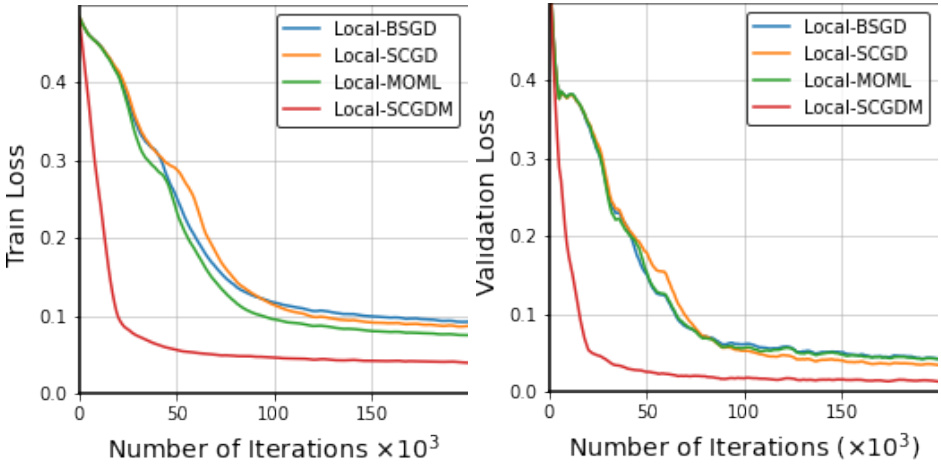

Figure: Train (Left) and Validation(Right) loss for our method and baselines.

References I

[Ghadimi et al., 2020] Ghadimi, S., Ruszczynski, A., and Wang, M. (2020).
A single timescale stochastic approximation method for nested stochastic optimization.
SIAM Journal on Optimization, 30(1):960–979.

[Huang et al., 2021] Huang, F., Li, J., and Huang, H. (2021).
Compositional federated learning: Applications in distributionally robust averaging and meta learning.
arXiv preprint arXiv:2106.11264.

[Wang et al., 2021] Wang, B., Yuan, Z., Ying, Y., and Yang, T. (2021).
Memory-based optimization methods for model-agnostic meta-learning.
arXiv preprint arXiv:2106.04911.