Communication-Efficient Federated Bilevel Optimization with Global and Local Lower Level Problems

Junyi Li[1], Feihu Huang[2], Heng Huang[1]

[1]University of Maryland College Park, [2]University of Pittsburgh,

January 15, 2024

Federated Bilevel Optimization Problems

- **Problem Formulation**:

$$\min_{x\in\mathbb{R}^p} h(x) := \frac{1}{M}\sum_{m=1}^{M} f^{(m)}(x, y_x), \text{ s.t. } y_x = \arg\min_{y\in\mathbb{R}^d} \frac{1}{M}\sum_{m=1}^{M} g^{(m)}(x, y)$$

- **Global Hyper-gradient**:

$$\Phi(x,y) = \nabla_x f(x,y) - \nabla_{xy} g(x,y) \times [\nabla_{y^2} g(x,y)]^{-1} \nabla_y f(x,y),$$

- **Local Hyper-gradient**:

$$\Phi^{(m)}(x,y) = \nabla_x f^{(m)}(x,y) - \nabla_{xy} g^{(m)}(x,y) \times [\nabla_{y^2} g^{(m)}(x,y)]^{-1} \nabla_y f^{(m)}(x,y),$$

Hyper-gradient Evaluation as a quadratic FL problem

Challenge: $\Phi(x, y_x) \neq \frac{1}{M}\sum_{m=1}^{M} \Phi^{(m)}(x, y_x)$

Quadratic FL problem:

$$\min_{u\in\mathbb{R}^d} l(u) = \frac{1}{M}\sum_{m=1}^{M} u^T(\nabla_{y^2} g^{(m)}(x, y))u - \langle \nabla_y f^{(m)}(x, y), u\rangle$$

Suppose that we denote the solution of the above problem as u^*, then we have the following linear operation to get the hypergradient:

$$\nabla h(x) = \frac{1}{M}\sum_{m=1}^{M}\left(\nabla_x f^{(m)}(x, y_x) - \nabla_{xy} g^{(m)}(x, y_x)u^*\right)$$

The FedBiO Algorithm

Perform alternative update of upper level variable $x_t^{(m)}$, the lower level variable $y_t^{(m)}$ and hyper-gradient computation variable $u_t^{(m)}$

$$y_{t+1}^{(m)} = y_t^{(m)} - \gamma_t \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y)$$

$$u_{t+1}^{(m)} = u_t^{(m)} - \tau_t \left(\nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,2}) u_t^{(m)} - \nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,2}) \right)$$

$$x_{t+1}^{(m)} = x_t^{(m)} - \eta_t \left(\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,1}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,1}) u_t^{(m)} \right)$$

Accelerate FedBiO with momentum-based variance reduction (FedBiOAcc)

Convergence Theorem

Theorem

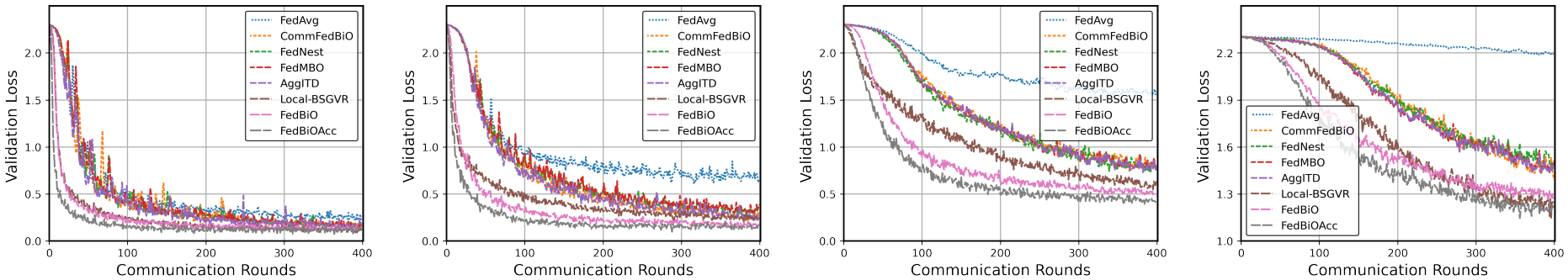
Suppose we choose learning rate $\alpha_t = \frac{\delta}{(u+t)^{1/3}}, t \in [T]$, *for some constant* δ *and* u, *and let* c_ν, c_ω, c_u *choose some value,* η, γ *and* τ *be some small values decided by the Lipschitz constants of* $h(x)$, *we choose the minibatch size to be* $b_x = b_y = b$ *and the first batch to be* $b_1 = O(Ib)$, *then we have:*

$$\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\left[\|\nabla h(\bar{x}_t)\|^2\right] = O\left(\frac{\kappa^{19/3}I}{T} + \frac{\kappa^{16/3}}{(bMT)^{2/3}}\right)$$

To reach an ϵ*-stationary point, we need* $T = O(\kappa^8(bM)^{-1}\epsilon^{-1.5})$, $I = O(\kappa^{5/3}(bM)^{-1}\epsilon^{-0.5})$.

Federated Data Cleaning

Experimental Results



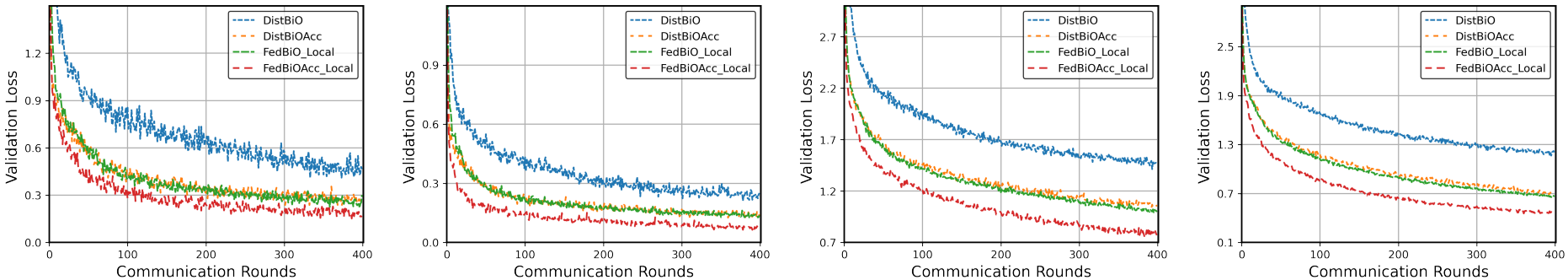
- Validation Error vs Communication Rounds. From Left to Right: $\rho = 0.1, 0.4, 0.8, 0.95$. The local step I is set as 5 for FedBiO, FedBiOAcc and FedAvg.

Federated Bilevel Optimization with Local Lower Level Problems

Problem Formulation:

$$\min_{x\in\mathbb{R}^p} h(x) := \frac{1}{M}\sum_{m=1}^{M} f^{(m)}(x, y_x^{(m)}), \text{ s.t. } y_x^{(m)} = \arg\min_{y\in\mathbb{R}^d} g^{(m)}(x, y)$$

Federated Hyper-Representation Learning:



- Validation Error vs Communication Rounds for the Omniglot Dataset. From Left to Right: 5-way-1-shot, 5-way-5-shot, 20-way-1-shot, 20-way-5-shot. The local step I is set to 5