Accelerated Federated Learning with Decoupled Adaptive Optimization

Jiayin Jin[1], Jiaxiang Ren[1], Yang Zhou[1], Lingjuan Lyu[2], Ji Liu[3], Dejing Dou[3,4]

[1]Auburn University
[2]Sony AI
[3]Baidu Research
[4]University of Oregon

ICML | 2022

Federated Adaptive Optimization

- Federated adaptive optimization
  - Apply centralized adaptive optimization methods to federated learning (FL) for faster convergence and higher accuracy, e.g., SGDM, Adam, AdaGrad

- Related work
  - FedOPT uses adaptive optimizers as server optimizers
  - Local AdaAlter and Mime use the same optimizer states on all clients, which is similar to server adaptive methods in FedOpt
  - FedLocal is the first real client adaptive approach with correction techniques

Problem Definition

- Our goal
  - Build theoretical principles on where to and how to design and utilize adaptive optimization methods in federated settings
  - Develop novel adaptive optimization methods for FL from the perspective of dynamics of ordinary differential equations (ODEs) of of centralized optimizers

Connect Federated Optimization to Decompositions of ODEs of Centralized Optimizers

- Centralized SGDM

$$m(t+1) = \beta * m(t) + (1-\beta) * g(W(t)),$$
$$W(t+1) = W(t) - \eta * m(t+1),$$

- Corresponding ODE

$$\eta \frac{d}{d\tau} m(\tau) = -(1-\beta) m(\tau) + (1-\beta) g(W(\tau)),$$
$$\frac{d}{d\tau} W(\tau) = -m(\tau).$$

- Decomposition of SGDM on clients

$$m^i(t+1) = \beta * m^i(t) + (1-\beta) * g^i(W(t)),$$
$$W^i(t+1) = W^i(t) - \eta * m^i(t+1),$$
$$m(t+1) = \sum_{i=1}^{M} \frac{N^i}{N} m^i(t),$$
$$W(t+1) = \sum_{i=1}^{M} \frac{N^i}{N} W^i(t).$$

Momentum Decoupling Adaptive Optimization

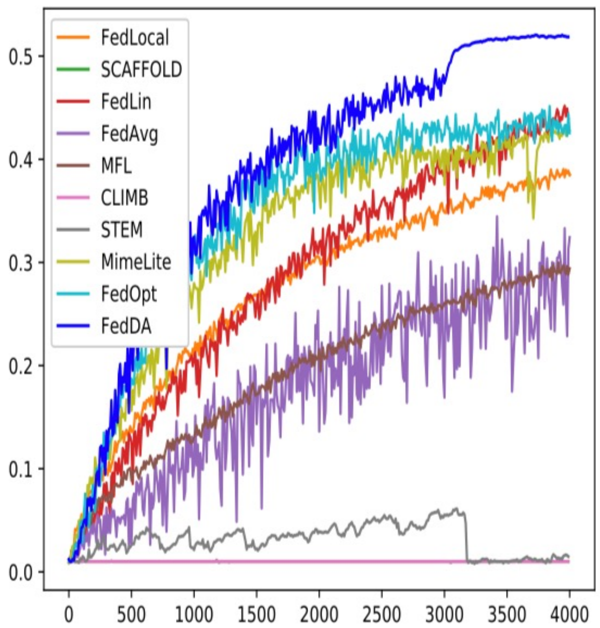
- Local update is independent of global momentum

$$\frac{d}{d\tau}W^i(\tau) = -g^i(W^i(\tau)),$$
$$\eta\frac{d}{d\tau}m^i(\tau) = -(1-\beta)m^i(\tau) + (1-\beta)g^i(W^i(\tau)),$$
$$\frac{d}{d\tau}W(\tau) = -\alpha * m(\tau).$$

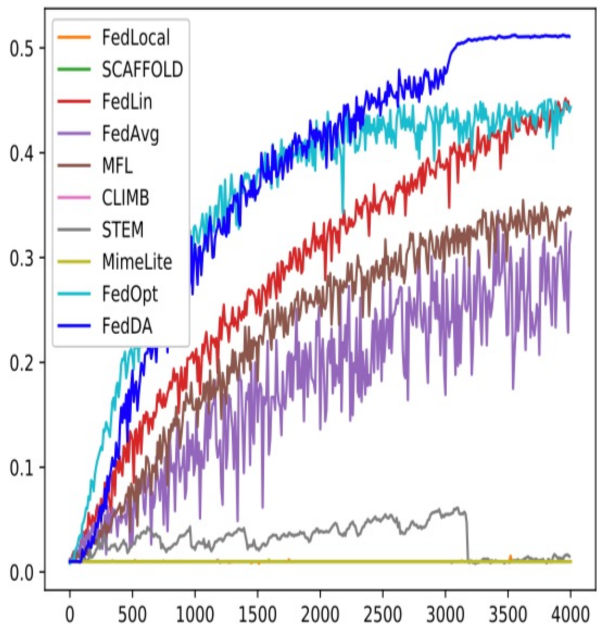
- Numerical solution to the above ODE

$$W^i(t+1) = W^i(t) - g^i(W^i(t)) * \eta,$$
$$m^i(t+1) = \beta * m^i(t) + (1-\beta) * g^i(W^i(t))\eta,$$
$$m(t+1) = \sum_{i=1}^{M} \frac{N^i}{N} m^i(t),$$
$$W(t+1) = W(t) - m(t+1) * \alpha * \eta.$$

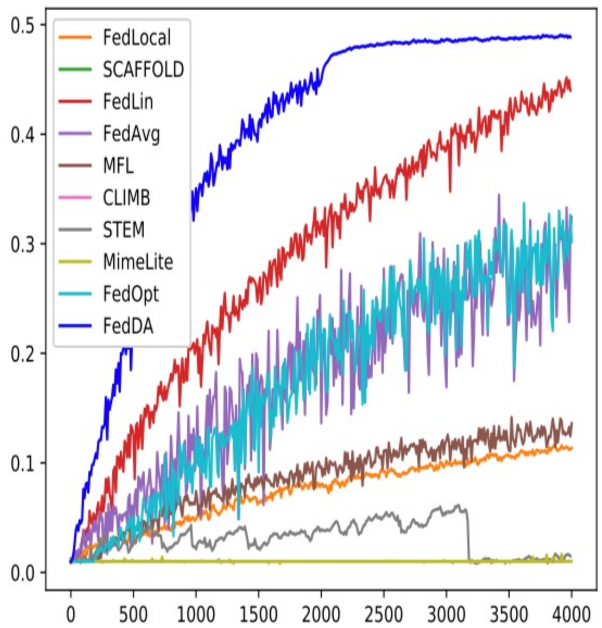
Convergence on CIFAR-100 with Three Optimizers



(a) SGDM



(b) Adam



(c) AdaGrad