Federated Optimization for Heterogeneous Networks

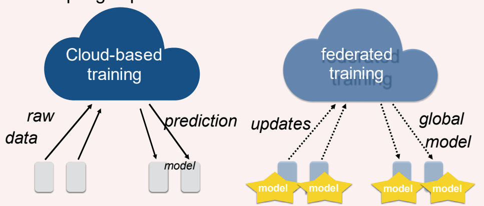
Carnegie Mellon University [a]

Tian Li [a], Anit Kumar Sahu [b], Manzil Zaheer [c], Maziar Sanjabi [d], Ameet Talwalkar [a], Virginia Smith [a]

[b] [c] USC [d]

Federated Learning

- **Training machine learning models in heterogeneous networks of edge devices**
- **Example Applications**
  - voice recognition on mobile phones
  - predicting low blood sugar via wearable devices
  - adapting to pedestrian behavior in autonomous vehicles



Challenges

- **Expensive communication: local updating methods**
  - millions of distributed devices, slower networks
- **Systems heterogeneity: low participation of devices**
  - variability in hardware, network connection, and power;
  - dropped nodes
- **Privacy: sending models instead of raw data**
  - local data is important
- **Statistical heterogeneity (focus of this work)**
  - data may be non-identically distributed across devices

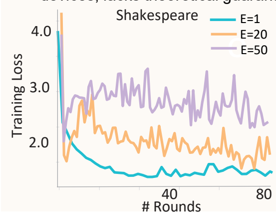
Contributions

- **FedProx, a novel federated optimization framework**
- **Convergence guarantees under a heterogeneity characteristic**
- **More robust and stable empirical performance**

Previous: Federated Averaging [1]

Objective: $\min_w f(w) = \sum_{k=1}^{N} p_k F_k(w) = \mathbb{E}_k[F_k(w)]$

- Simple method: averaging local SGD
- Works well in many settings (e.g., non-convex)
- Can diverge when there is significant heterogeneity across devices; lacks theoretical guarantees



- Select a subset of devices S_t
- Each selected device k runs SGD for ***E*** number of epochs locally to find w_k^{t+1} to solve inexactly: $F_k(w)$
- Server aggregates:

$$w^{t+1} = \frac{1}{K}\sum_{k \in S_t} w_k^{t+1}$$

Proposed: FedProx

Key idea: Modified local subproblem

$$\min_w F_k(w) + \frac{\mu}{2}||w - w^t||^2$$

The current global model (starting point)

- Introduces proximal term to limit local updates from diverging
- Generalization of FedAvg (μ = 0)

Convergence Results

B-dissimilarity to measure statistical heterogeneity $B(w) = \sqrt{\frac{\mathbb{E}_k[||\nabla F_k(w)||^2]}{||\nabla f(w)||^2}}$

Data IID: B(w)=1; larger dissimilarity, larger B(w)

(Assume B is bounded)

[Theorem] Obtain suboptimality ε, after ***T*** iterations, with:

$$T = O\left(\frac{f(w^0) - f^*}{\rho\varepsilon}\right)$$

ρ incorporates B, μ

- Same asymptotic convergence as SGD
- Both convex & non-convex; any local solver

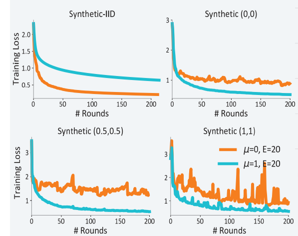
Experiments

LEAF [2]: A Benchmark for Learning in Federated Settings

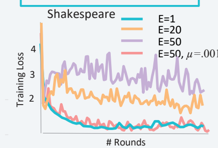
- Suite of open-source datasets
- Evaluation framework with statistical and systems metrics to assess competing solutions
- Online presence to encourage participation and reproducibility

leaf.cmu.edu

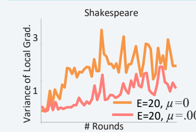
Increasing heterogeneity leads to worse convergence; Setting μ > 0 helps to combat this



Increasing μ helps to prevent divergence



B is correlated with convergence in practice



Moving Forward

- **Hyper-parameter optimization for federated settings (e.g., μ)**
- **Diagnostics and leverage systems heterogeneity**
- **Combining with compression schemes**

Paper & code: *cs.cmu.edu/~litian*

[1] McMahan, H. Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." arXiv preprint arXiv:1602.05629 (2016).
[2] Caldas, et al. "Leaf: A benchmark for federated settings." arXiv preprint arXiv:1812.01097 (2018).