On the Convergence of FedAvg on Non-iid Data

Xiang Li[*,1], Kaixuan Huang[*,1], Wenhao Yang[*,2], Shusen Wang[3], Zhihua Zhang[1]

[1]School of Mathematical Science, Peking University

[2]Academy for Advanced Interdisciplinary Studies, Peking University

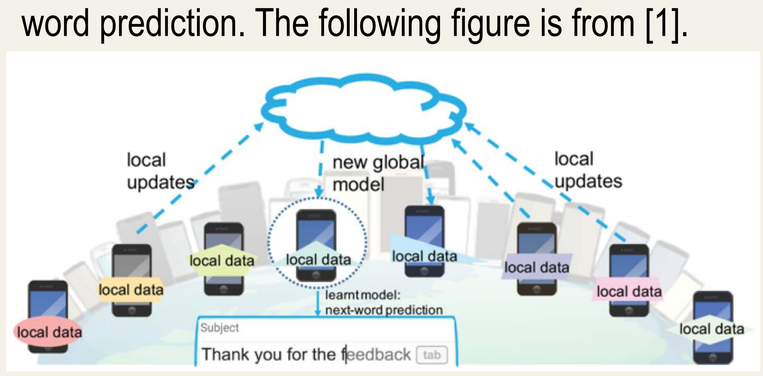
[3]Department of Computer Science, Stevens Institute of Technology

Introduction

Federated Learning (FL), also known as federated optimization, allows multiple parties to collaboratively train a model without data sharing.

FL lets the user devices (aka worker nodes) perform most of the computation and a central parameter server update the model parameters using the descending directions returned by the user devices.

A typical application is to learn user behaviors across mobile phones, where the task is the next-word prediction. The following figure is from [1].



[1] Li T, Sahu A K, Talwalkar A, et al. Federated learning: Challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

Three Unique Characters

First, the training data are massively distributed over an incredibly large number of devices, and the connection between the central server and a device is slow. Typically, the technique of local update is used to reduce communication frequency and thus to prove communication efficiency.

⟹ Communication Efficiency.

Second, unlike the traditional distributed learning systems, the FL system does not have control over users' devices. For example, unavailable WiFi access makes mobile phones offline. It is thus impractical to require all the devices to be active.

⟹ Partial Participation.

Third, the training data are non-i.i.d. (precisely meaning data are independent but not identically distributed.). Hence, the data available locally fail to represent the overall distribution.

⟹ Statistical Heterogeneity.

Previous Work

There have been much efforts on developing convergence guarantees for FL algorithms based on the assumptions that (1) the data are iid and (2) all the devices are active. The reference can be checked up in our paper.

These two assumptions obviously violate the second and third characters of FL, making previous analysis less practical and realistic.

Our work aim to provide analysis for a classic algorithm used in FL, given more realistic assumptions that live in harmony with FL.

Problem Setup

We consider the following distributed optimization model:

$$\min_{\mathbf{w}} \left\{ F(\mathbf{w}) \triangleq \sum_{k=1}^{N} p_k F_k(\mathbf{w}) \right\}$$

Where N is the number of devices, p_k is the weight of the k-th device. Suppose the k-th device holds n_k data points: $x_{k,1}, x_{k,2} \ldots, x_{k,n\_k}$. The local objective is given by

$$F_k(\mathbf{w}) \triangleq \frac{1}{n_k} \sum_{j=1}^{n_k} \ell\left(\mathbf{w}; x_{k,j}\right)$$

where l is a user-specified loss function. For example, the L2 loss for linear regression and the log loss for logistic regression.

The FedAvg Algorithm

First, the central samples a small portion of active device, denoted by S_t.

Then, the central server broadcasts the latest model, w_t, to the active devices.

Third, every active device (say, the k-th) lets $w_t^k = w_t$ and then performs E (>=1) local updates:

$$\mathbf{w}_{t+i+1}^k \longleftarrow \mathbf{w}_{t+i}^k - \eta_{t+i} \nabla F_k\left(\mathbf{w}_{t+i}^k, \xi_{t+i}^k\right)$$

$$i = 0, 1, \cdots, E-1$$

where η_{t+i} is the learning rate (a.k.a. the step size) and x_{t+i}^k is a sample uniformly chosen from the local data.

Lastly, the server aggregates the local models, w_{t+E}^1, w_{t+E}^2,..., w_{t+E}^K to produce the new global model w_{t+E}. Because of the non-iid and partial device participation issues, the aggregation step can vary.

Two Device Sampling Methods

We consider two device sampling methods that generate the active set S_t:

(1) S1: sample K indices randomly selected with replacement according to p_1, p_2,..., p_N.

(2) S2: (assuming $p_1 = \cdots = p_K$) sample K indices randomly from [N] without replacement.

Assumptions

We made the following typical assumptions. They are quite standard in optimization literature.

(A1) F_k are all L-smooth, i.e., it satisfies

$$f(\mathbf{v}) \leq \mathbf{f}(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T \nabla f(\mathbf{w}) + \frac{L}{2}\|\mathbf{v}-\mathbf{w}\|_2^2$$

(A2) F_k are all μ-strongly convex, i.e., it satisfies

$$f(\mathbf{v}) \geq \mathbf{f}(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T \nabla f(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v}-\mathbf{w}\|_2^2$$

(A3) Uniformly bounded variance:

$$\mathbb{E}\left\|\nabla F_k\left(\mathbf{w}_t^k, \xi_t^k\right) - \nabla F_k\left(\mathbf{w}_t^k\right)\right\|^2 \leq \sigma_k^2$$

(A4) Uniformly bounded second moments:

$$\mathbb{E}\left\|\nabla F_k\left(\mathbf{w}_t^k, \xi_t^k\right)\right\|^2 \leq G^2$$

Convergence Result

Let (A1)-(A4) hold and L, μ, σ_k, G be defined therein. Let $\kappa = \frac{L}{\mu}$, $\gamma = max(9\kappa, E)$ and the learning rate $\eta_t = \frac{2}{\mu(\gamma+t)}$. For Fedavg, we have

$$\mathbb{E}\left[F\left(\mathbf{w}_T\right)\right] - F^* \leq \frac{2\kappa}{\gamma+T}\left(\frac{B+C}{\mu} + 2L\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|^2\right)$$

where

$$B = \sum_{k=1}^{N} p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2$$

and C depends on the sampling methods.

For S1, $C = \frac{4}{K}E^2G^2$; for S2, $C = \frac{(N-K)4}{(N-1)K}E^2G^2$.

Some Finding From the Result

(1) About E: to find an ε accuracy solution, the requires round is about

$$\left(1+\frac{1}{K}\right)EG^2 + \frac{\sum_{k=1}^{N} p_k^2\sigma_k^2 + L\Gamma + \kappa G^2}{E}$$

(2) About K: weak dependence.

(3) About sampling methods: S2 should be better while S1 is more practical.

The Importance of Learning Rate Decay

Diminishing learning rates is crucial for the convergence of FedAvg in the non-iid setting.

Specifically, we can construct a ridge regression model (which is strongly convex and smooth) so that with full batch size, $E > 1$, and any fixed and sufficiently small step size, FedAvg will converge to sub-optimal points.

In particular, let $\widetilde{w}^*$ be the solution produced by FedAvg with a small enough and constant learning rate η and w^* be the true optimal points. Then

$$\left\|\tilde{\mathbf{w}}^* - \mathbf{w}^*\right\|_2 = \Omega((E-1)\eta) \cdot \left\|\mathbf{w}^*\right\|_2$$

Experiments