ON THE CONVERGENCE OF FEDAVG ON NON-IID DATA

这篇论文给出了不需要分布数据集是IID的假设的一个收敛上界

背景

现有的联邦学习关于收敛性证明的工作需要有以下两个假设：

- 分布在不同的设备上的数据集是IID的
- 参与联邦学习的每一个设备都能与服务器保持稳定的连接

事实上这两个假设在现实部署的过程中是难以实现的，首先，保证不同设备上的数据集都是IID的显然是难以保证的，其次，我们也无法保证参与联邦学习的每一个设备都时刻保持有效的连接，当某一台设备关闭电源或者断开连接的是否，那么服务器和其他设备都需要等待这一台的重启或重新连接，这明显会造成很大的耗费

对于第二个问题的解决办法是每次服务器向设备广播数据集的时候都会选择其中保持连接的一部分设备进行广播然后进行训练，而不是向所有的设备进行广播

贡献

该论文主要有以下两个贡献：

- 给出了不需要以上两个假设的两个收敛上界，分别包括Full Device Participation(每次都选择所有的设备来训练)和Partial Device Participation(每次只选择一部分的设备进行训练)
- 提出全局模型更新的步长需要衰减，并给出解释

定义

损失函数：

$$\min_{\mathbf{w}} \left\{ F(\mathbf{w}) \triangleq \sum_{k=1}^{N} p_k F_k(\mathbf{w}) \right\}, \tag{2}$$

在某个设备上的损失函数：

$$F_k(\mathbf{w}) \triangleq \frac{1}{n_k} \sum_{j=1}^{n_k} \ell(\mathbf{w}; x_{k,j}), \tag{3}$$

where $\ell(\cdot;\cdot)$ is a user-specified loss function.

在某个设备上的local update:

$$\mathbf{w}_{t+i+1}^k \longleftarrow \mathbf{w}_{t+i}^k - \eta_{t+i} \nabla F_k(\mathbf{w}_{t+i}^k, \xi_{t+i}^k), i = 0, 1, \cdots, E-1$$

Full Device Participation下全局模型的更新：

$$\mathbf{w}_{t+E} \longleftarrow \sum_{k=1}^{N} p_k \mathbf{w}_{t+E}^k.$$

Partical Device Participation下全局模型的更新：

$$\mathbf{w}_{t+E} \longleftarrow \frac{N}{K} \sum_{k \in \mathcal{S}_t} p_k \mathbf{w}_{t+E}^k.$$

收敛上界

这篇论文给出的收敛上界需要满足以下4个假设：

Assumption 1. *$F_1, \cdots, F_N$ are all L-smooth: for all $\mathbf{v}$ and $\mathbf{w}$, $F_k(\mathbf{v}) \leq F_k(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{L}{2}\|\mathbf{v}-\mathbf{w}\|_2^2$.*

Assumption 2. *$F_1, \cdots, F_N$ are all μ-strongly convex: for all $\mathbf{v}$ and $\mathbf{w}$, $F_k(\mathbf{v}) \geq F_k(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v}-\mathbf{w}\|_2^2$.*

Assumption 3. *Let ξ_t^k be sampled from the k-th device's local data uniformly at random. The variance of stochastic gradients in each device is bounded: $\mathbb{E}\left\|\nabla F_k(\mathbf{w}_t^k, \xi_t^k) - \nabla F_k(\mathbf{w}_t^k)\right\|^2 \leq \sigma_k^2$ for $k = 1, \cdots, N$.*

Assumption 4. *The expected squared norm of stochastic gradients is uniformly bounded, i.e., $\mathbb{E}\left\|\nabla F_k(\mathbf{w}_t^k, \xi_t^k)\right\|^2 \leq G^2$ for all $k = 1, \cdots, N$ and $t = 1, \cdots, T-1$*

然后它给出的Full Device Participation收敛上界是：

Theorem 1. *Let Assumptions 1 to 4 hold and L, μ, σ_k, G be defined therein. Choose $\kappa = \frac{L}{\mu}$, $\gamma = \max\{8\kappa, E\}$ and the learning rate $\eta_t = \frac{2}{\mu(\gamma+t)}$. Then `FedAvg` with full device participation satisfies*

$$\mathbb{E}\left[F(\mathbf{w}_T)\right] - F^* \leq \frac{\kappa}{\gamma + T - 1}\left(\frac{2B}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}\|\mathbf{w}_1 - \mathbf{w}^*\|^2\right), \tag{4}$$

where

$$B = \sum_{k=1}^{N} p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2. \tag{5}$$

给出的Partial Device Participation收敛上界是：

Theorem 2. *Let Assumptions 1 to 4 hold and L, μ, σ_k, G be defined therein. Let κ, γ, η_t, and B be defined in Theorem 1. Let Assumption 5 hold and define $C = \frac{4}{K}E^2G^2$. Then*

$$\mathbb{E}\left[F(\mathbf{w}_T)\right] - F^* \leq \frac{\kappa}{\gamma + T - 1}\left(\frac{2(B+C)}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}\|\mathbf{w}_1 - \mathbf{w}^*\|^2\right), \tag{6}$$

Alternatively, we can select K indices from $[N]$ uniformly at random without replacement. As a consequence, we need a different aggregation strategy. Assumption 6 assumes the K indices are selected uniformly without replacement and the aggregation step is the same as in Section 2. However, to guarantee convergence, we require an additional assumption of balanced data.

证明

定理1

Theorem 1. *Let Assumptions 1 to 4 hold and L, μ, σ_k, G be defined therein. Choose $\kappa = \frac{L}{\mu}$, $\gamma = \max\{8\kappa, E\}$ and the learning rate $\eta_t = \frac{2}{\mu(\gamma+t)}$. Then FedAvg with full device participation satisfies*

$$\mathbb{E}\left[F(\mathbf{w}_T)\right] - F^* \leq \frac{\kappa}{\gamma + T - 1}\left(\frac{2B}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}\|\mathbf{w}_1 - \mathbf{w}^*\|^2\right), \tag{4}$$

where

$$B = \sum_{k=1}^{N} p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2. \tag{5}$$

证明：

定义一个辅助变量v_{t+1}^k：

$$\mathbf{v}_{t+1}^k = \mathbf{w}_t^k - \eta_t \nabla F_k(\mathbf{w}_t^k, \xi_t^k), \tag{9}$$

$$\mathbf{w}_{t+1}^k = \begin{cases} \mathbf{v}_{t+1}^k & \text{if } t+1 \notin \mathcal{I}_E, \\ \sum_{k=1}^N p_k \mathbf{v}_{t+1}^k & \text{if } t+1 \in \mathcal{I}_E. \end{cases} \tag{10}$$

这个辅助变量的作用可简单看作一个中间变量，（10）的第一个条件的意思是$t+1$次迭代训练不需要进行联邦平均，第二个条件的意思是$t+1$次迭代训练需要进行联邦平均

提出三个引理来辅助证明（这三个引理怎么得到的后面讨论）：

Lemma 1 (Results of one step SGD). *Assume Assumption 1 and 2. If $\eta_t \leq \frac{1}{4L}$, we have*

$$\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^\star\right\|^2 \leq (1 - \eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t - \mathbf{w}^\star\right\|^2 + \eta_t^2 \mathbb{E}\left\|\mathbf{g}_t - \overline{\mathbf{g}}_t\right\|^2 + 6L\eta_t^2\Gamma + 2\mathbb{E}\sum_{k=1}^{N} p_k \left\|\overline{\mathbf{w}}_t - \mathbf{w}_k^t\right\|^2$$

where $\Gamma = F^* - \sum_{k=1}^N p_k F_k^\star \geq 0$.

Lemma 2 (Bounding the variance). *Assume Assumption 3 holds. It follows that*

$$\mathbb{E}\left\|\mathbf{g}_t - \overline{\mathbf{g}}_t\right\|^2 \leq \sum_{k=1}^{N} p_k^2 \sigma_k^2.$$

Lemma 3 (Bounding the divergence of $\{\mathbf{w}_t^k\}$). *Assume Assumption 4, that η_t is non-increasing and $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$. It follows that*

$$\mathbb{E}\left[\sum_{k=1}^{N} p_k \left\|\overline{\mathbf{w}}_t - \mathbf{w}_k^t\right\|^2\right] \leq 4\eta_t^2 (E-1)^2 G^2.$$

证明过程及思路如下：

定理2

更新方式：

$$\mathbf{v}_{t+1}^k = \mathbf{w}_t^k - \eta_t \nabla F_k(\mathbf{w}_t^k, \xi_t^k), \tag{19}$$

$$\mathbf{w}_{t+1}^k = \begin{cases} \mathbf{v}_{t+1}^k & \text{if } t+1 \notin \mathcal{I}_E, \\ \text{samples } \mathcal{S}_{t+1} \text{ and average } \{\mathbf{v}_{t+1}^k\}_{k \in \mathcal{S}_{t+1}} & \text{if } t+1 \in \mathcal{I}_E. \end{cases} \tag{20}$$

两种更新方式：

(I) The server establishes $\mathcal{S}_{t+1}$ by i.i.d. **with replacement** sampling an index $k \in \{1, \cdots, N\}$ with probabilities $p_1, \cdots, p_N$ for K times. Hence $\mathcal{S}_{t+1}$ is a multiset which allows a element to occur more than once. Then the server averages the parameters by $\mathbf{w}_{t+1}^k = \frac{1}{K}\sum_{k\in\mathcal{S}_{t+1}} \mathbf{v}_{t+1}^k$. This is first proposed in (Sahu et al., 2018) but lacks theoretical analysis.

(II) The server samples $\mathcal{S}_{t+1}$ uniformly in a **without replacement** fashion. Hence each element in $\mathcal{S}_{t+1}$ only occurs once.Then server averages the parameters by $\mathbf{w}_{t+1}^k = \sum_{k\in\mathcal{S}_{t+1}} p_k \frac{N}{K} \mathbf{v}_{t+1}^k$. Note that when the p_k's are not all the same, one cannot ensure $\sum_{k\in\mathcal{S}_{t+1}} p_k \frac{N}{K} = 1$.

主要区别：

- 第一种方法是重复K次，每次选择一个设备，最终能选择出K个设备，并且一个设备可以被选择多次
- 第二种方法是直接从N个设备中选出K个设备，每个设备在本次训练中最多能选择一次

然后给出了这两种选择方法的$\overline{v}_{t-1}$和$\overline{w}_{t-1}$的差值上界：

Lemma 5 (Bounding the variance of $\overline{\mathbf{w}}_t$)**.** *For $t+1 \in \mathcal{I}$, assume that η_t is non-increasing and $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$. We have the following results.*

(1) For Scheme I, the expected difference between $\overline{\mathbf{v}}_{t+1}$ and $\overline{\mathbf{w}}_{t+1}$ is bounded by

$$\mathbb{E}_{\mathcal{S}_t} \left\|\overline{\mathbf{v}}_{t+1} - \overline{\mathbf{w}}_{t+1}\right\|^2 \leq \frac{4}{K}\eta_t^2 E^2 G^2.$$

(2) For Scheme II, assuming $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$, the expected difference between $\overline{\mathbf{v}}_{t+1}$ and $\overline{\mathbf{w}}_{t+1}$ is bounded by

$$\mathbb{E}_{\mathcal{S}_t} \left\|\overline{\mathbf{v}}_{t+1} - \overline{\mathbf{w}}_{t+1}\right\|^2 \leq \frac{N-K}{N-1}\frac{4}{K}\eta_t^2 E^2 G^2.$$

证明过程大致与证明定理1的一样，区别在于定理1不适用了：

$$\begin{aligned}\left\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\|^2 &= \left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1} + \overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2 \\ &= \underbrace{\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1}\right\|^2}_{A_1} + \underbrace{\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2}_{A_2} + \underbrace{2\left\langle \overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1}, \overline{\mathbf{v}}_{t+1} - \mathbf{w}^* \right\rangle}_{A_3}.\end{aligned}$$

When expectation is taken over $\mathcal{S}_{t+1}$, the last term (A_3) vanishes due to the unbiasedness of $\overline{\mathbf{w}}_{t+1}$.

If $t+1 \notin \mathcal{I}_E$, A_1 vanishes since $\overline{\mathbf{w}}_{t+1} = \overline{\mathbf{v}}_{t+1}$. We use Lemma 5 to bound A_2. Then it follows that

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t - \mathbf{w}^*\right\|^2 + \eta_t^2 B.$$

If $t+1 \in \mathcal{I}_E$, we additionally use Lemma 5 to bound A_1. Then

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\|^2 &= \mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1}\right\|^2 + \mathbb{E}\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|^2 \\ &\leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t - \mathbf{w}^\star\right\|^2 + \eta_t^2(B+C), \qquad (21)\end{aligned}$$

后面就是直接改变一些值便能得到结果：

> The only difference between eqn. (21) and eqn. (11) is the additional C. Thus we can use the same argument there to prove the theorems here. Specifically, for a diminishing stepsize, $\eta_t = \frac{\beta}{t+\gamma}$ for some $\beta > \frac{1}{\mu}$ and $\gamma > 0$ such that $\eta_1 \leq \min\{\frac{1}{\mu}, \frac{1}{4L}\} = \frac{1}{4L}$ and $\eta_t \leq 2\eta_{t+E}$, we can prove $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\|^2 \leq \frac{v}{\gamma+t}$ where $v = \max\left\{\frac{\beta^2(B+C)}{\beta\mu-1}, (\gamma+1)\|\mathbf{w}_1 - \mathbf{w}^*\|^2\right\}$.

还是证明定理1时的公式：

$$\mathbb{E}[F(\overline{\mathbf{w}}_t)] - F^* \leq \frac{L}{2}\Delta_t \leq \frac{L}{2}\frac{v}{\gamma+t}.$$

然后直接代入：

> Specifically, if we choose $\beta = \frac{2}{\mu}, \gamma = \max\{8\frac{L}{\mu}, E\} - 1$ and denote $\kappa = \frac{L}{\mu}$, then $\eta_t = \frac{2}{\mu}\frac{1}{\gamma+t}$ and
>
> $$\mathbb{E}[F(\overline{\mathbf{w}}_t)] - F^* \leq \frac{\kappa}{\gamma+t}\left(\frac{2(B+C)}{\mu} + \frac{\mu(\gamma+1)}{2}\|\mathbf{w}_1 - \mathbf{w}^*\|^2\right).$$
>
> □

结论

这篇论文给出了不需要数据集为IID的假设的收敛上界，并且关于步长对于收敛性的影响进行了讨论，我觉得是一篇比较有开创性的文章。相关的数学证明我还会继续去研究一下