Convergence of FL

Chaoqian Cheng

2024 年 3 月 11 日

1 FEDERATED AVERAGING (FEDAVG)

1.1 Notation.

Let N be the total number of user devices. Let T be the total number of every device's SGDs, E be the number of local iterations performed in a device between two communications, and thus $\frac{T}{E}$ is the number of communications.

1.2 Problem formulation.

We consider the following distributed optimization model:

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

where N is the number of devices, and p_k is the weight of the k-th device such that $p_k \geq 0$ and $\sum_{k=1}^{N} p_k = 1$. Suppose the k-th device holds the n_k training data: $x_{k,1}, x_{k,2}, \cdots, x_{k,n_k}$. The local objective $F_k(\cdot)$ is defined by

$$F_k(\mathbf{w}) \triangleq \frac{1}{n_k} \sum_{j=1}^{n_k} \ell(\mathbf{w}; x_{k,j}), \tag{2}$$

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

1.3 Algorithm description.

Here, we describe one round (say the t-th) of the *standard* FedAvg algorithm. First, the central server **broadcases** the latest model, $\mathbf{w}_t$, to all the devices. Secnond, every device (say the k-th) lets $\mathbf{w}_t^k = \mathbf{w}_t$ and then performs $E(\geq 1)$ **local updates:**

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

where η_{t+i} is the learning rate (*a.k.a* step size) and ξ_{t+1}^k is a sample uniformly chosen from the local data. Last, the server **aggregates** the local models, $\mathbf{w}_{t+E}^1, \cdots, \mathbf{w}_{t+E}^N$, to produce the new global model, $\mathbf{w}_{t+E}$. The aggregation step performs

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

1.4 Additional Notation.

In our analysis, we define a virual sequences $\bar{\mathbf{w}}_t = \sum_{k=1}^{N} p_k \mathbf{w}_t^k$. $\bar{\mathbf{w}}_{t+1}$ results from an single step of SGD from $\bar{\mathbf{w}}_t$. For convenience, we define $\bar{\mathbf{g}}_t = \sum_{k=1}^{N} p_k \nabla F_k(\mathbf{w}_t^k)$ and $\mathbf{g}_t = \sum_{k=1}^{N} p_k \nabla F_k(\mathbf{w}_t^k, \xi_t^k)$. Therefore, $\bar{\mathbf{w}}_{t+1} = \bar{\mathbf{w}}_t - \eta_t \mathbf{g}_t$ and $\mathbb{E}\mathbf{g}_t = \bar{\mathbf{g}}_t$.

2 常用假设

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

Remark 1 *以上不等式表明函数有一个二次函数的上界。它对函数的光滑性做出了适度的假设，使得许多梯度下降方法都可以在此假设下分析。*

L-smooth 的另一个形式: for any $\mathbf{v}$, $\mathbf{x}$, $\|\nabla F_k(\mathbf{v}) - \nabla F_k(\mathbf{x})\| \leq L\|\mathbf{v} - \mathbf{x}\|$.

Intuition 1 *$F_k^* \leq F_k(\mathbf{x}) - \frac{1}{2L}\|\nabla F_k(\mathbf{x})\|^2$, where $F_k^* = \min F_k$.*

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

Remark 2 *以上不等式表面函数有一个二次函数的下界。在实际问题中遇到纯粹的强凸函数的情况并不多，因此在学术研究中，强凸性假设通常被认为是理想化的*[2]。

Assumption 3 *(**Bounded variance**). 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: $\mathbb{E}\|\nabla F_k(\mathbf{w}_t^k, \xi_t^k) - \nabla F_k(\mathbf{w}_t^k)\|^2 \leq \sigma_k^2$*

Assumption 4 *(**Bounded stochastic gradient**). The expected squared norm of stochastic gradients is uniformly bounded,* i.e., $\mathbb{E}\|\nabla F_k(\mathbf{w}_t^k, \xi_t^k)\|^2 \leq G^2$.

Quantifying the degree of non-iid 1 *(**heterogeneity**) Let F^* and F_k^* be the minimum values of F and F_k, respectively. We use the term $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$ for quantifying the degree of non-iid.*

Remark 3 *Γ 量化了 non-iid 度，如果数据是 iid 的，随着样本增加显然 Γ 等于 0；如果数据是 non-iid 的，则 Γ 不为 0，其大小反映了数据分布的异构性。在 Lemma 1 的证明过程中，Γ 项是通过在 $F_k(\mathbf{w_k^t}) - F_k(\mathbf{w}^*)$ 中增加 $+F^* - F^*$ 项来主动构建的。*

3 收敛结果

Theorem is a mathematical statement that is proved using rigorous mathematical reasoning. In a mathematical paper, the term theorem is often reserved for the most important results. 定理是一个具有**结论性**的、用**数学陈述**的结果，它需要**严格的数学证明**。
下面的 Theorem 1与 Theorem 2 分别展示了 FedAvg 算法在凸模型和非凸模型上的收敛结果。

Theorem 1 *Let Assumptions 1 to 4 hold and* L, μ, σ_k, G *be defined there in. Choose* $\kappa = \frac{L}{\mu}, \gamma = \max\{5\kappa, E\}$ *and the learning rate* $\eta_t = \frac{2}{\mu(\gamma+t)}$*. Then FedAvg with full device participation satisfies*

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \frac{\kappa}{\gamma + T - 1}(\frac{2B}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}\|\mathbf{w}_1 - \mathbf{w}^*\|^2), \tag{3}$$

where

$$B = \sum_{k=1}^{N} p_k^2\sigma_k^2 + 6L\Gamma + 8(E-1)^2G^2. \tag{4}$$

Remark 4 *(凸函数算法的判定指标)*[3]。当 $F(\mathbf{w})$ 为凸函数时，判定指标选择为统计量 $R(T)$ *(Regret):*

$$R(T) = \sum_{t=1}^{T}[F(\mathbf{w}) - F(\mathbf{w}^*)].$$

当 $T \to \infty$, $R(T)$ 的均摊值 $R(T)/T \to 0$, 我们认为这样的算法是收敛的，即 $\mathbf{w} \to \arg\min_{\mathbf{w}} \sum_{t=1}^{T} F(\mathbf{w}) \triangleq \mathbf{w}^*$, 不仅趋于某个值，而且这个值使目标函数最小。

Theorem 2 *Let Assumptions 1, Assumptions 3, Assumptions 4 hold. Then FedAvg with full device participation satisfies*

$$\min_t \mathbb{E}\|\nabla F(\mathbf{w}_t)\|^2 \leq \frac{2}{\eta_t T}\mathbb{E}[F(\bar{\mathbf{w}}_0) - F^*] + \sum_{k=1}^{N} p_k^2\sigma_k^2 + (\eta_t L - 1)\sum_{k=1}^{N} p_k^2 G^2. \tag{5}$$

Remark 5 *(*非凸函数算法的判定指标*)*[4]。对于无限制条件的非 *convex* 优化问题，一般认为当目标函数的梯度消失时，算法收敛。由于目标函数非 *convex*, 不得不牺牲全局最优解，转而接受局部最优解。当 $F(\mathbf{w})$ 为非凸函数时，判定指标选择为 $E(T) = \min_{t=1,2,\cdots,T} \mathbb{E}\|\nabla F(\mathbf{w})\|_2^2$. 当 $T \to \infty$ 时，若 $E(T)$ 的均摊值 $E(T)/T \to 0$, 我们认为这样的算法是收敛的。
从表达式可以看出，$E(T)$ 是一系列梯度模值平方的期望的最小值，也就是说，只要有某一个 t 时刻梯度消失了，算法就收敛了。这个判定收敛的指标是比较弱的：它只要求存在时刻 t 使梯度消失，并没有要求当 t 大于某时刻 t_0 时，梯度消失；也就是说，如果任由算法无休止地运行下去，算法可能会发散。

4 关键引理

Lemma is a minor result whose sole purpose is to help in proving a theorem. It is a stepping stone on the path to proving a theorem. 引理是为了**证明定理**的一个**中间结果**。

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

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

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

Remark 6 最重要的引理。实际上是建立了 $\|\bar{\mathbf{w}}_t + 1 - \mathbf{w}^*\|^2$ 的递推关系。常通过 *L-smooth* 与 *Theorem* 的左边建立联系。其他的引理都是为了 *bound Lemma 1*中的项。*Lemma 1*证明的第一步就是代入 $\bar{\mathbf{w}}_{t+1}$ 的递推公式。

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

$$\mathbb{E}\|\mathbf{g}_t - \bar{\mathbf{g}}_t\|^2 \le \sum_{k=1}^{N} p_k^2 \sigma_k^2.$$

Remark 7 用于 *bound Lemma 1*的第二项，只要代入 $\mathbf{g}_t$ 与 $\bar{\mathbf{g}}_t$ 的定义即可直接证明。

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

$$\mathbb{E}[\sum_{k=1}^{N} p_k \|\bar{\mathbf{w}}_t - \mathbf{w}_t^k\|^2] \le 4\eta_t^2 (E-1)^2 G^2.$$

Remark 8 用于 *bound Lemma 1*的第三项。

以下引理用于非凸的收敛性证明:

Lemma 4 *Assume Assumption 3 holds. It follows that*

$$\nabla F(\bar{\mathbf{w}}_t) = \bar{\mathbf{g}}_t.$$

Proof. From the Problem Formulation, we have

$$\nabla F(\bar{\mathbf{w}}_t) = \sum_{k=1}^{N} p_k \nabla F_k(\mathbf{w}_t^k) = \bar{\mathbf{g}}_t.$$

Lemma 5 *Assume Assumption 4 holds. It follows that*

$$\mathbb{E}\|\mathbf{g}_t\|^2 \le \sum_{k=1}^{N} p_k^2 G^2.$$

5 重要结论

Fact 1 $\|\mathbf{a}+\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + 2 < \mathbf{a}, \mathbf{b} > + \|\mathbf{b}\|^2$. *Thus,* $< \mathbf{a}, \mathbf{b} >= \frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{2}\|\mathbf{b}\|^2 - \frac{1}{2}\|\mathbf{a}-\mathbf{b}\|^2$.

Fact 2 $< \mathbf{a}, \mathbf{b} >= \|\mathbf{a}\|\|\mathbf{b}\| cos\theta \ge -\|\mathbf{a}\|\|\mathbf{b}\|$.

Fact 3 $\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 \ge 2\|\mathbf{a}\|\|\mathbf{b}\|$.

Fact 4 $\mathbb{E}\|\mathbf{x} - \mathbb{E}\mathbf{x}\| = \mathbb{E}\|\mathbf{x}\|^2 - \|\mathbb{E}\mathbf{x}\|^2 \le \mathbb{E}\|\mathbf{x}\|^2$.

Fact 5 *(Cauchy-Schwarz inequality)* $\|\sum_{i=1}^{n} a_i b_i\|^2 \le \sum_{i=1}^{n} \|a_i\|^2 \sum_{i=1}^{n} \|b_i\|^2$.

6 定理 1 的证明

Proof. Let $\Delta_t = \mathbb{E}\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2$. From Lemma 1, Lemma 2, Lemma 3, it follows that

$$\Delta_{t+1} \le (1-\eta_t\mu)\Delta_t + \eta_t^2 B \tag{6}$$

where

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

这一步使用了引理 1-3 的结论，得到了 $\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2$ 的递推关系。

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 will prove $\Delta_t \leq \frac{v}{\gamma+t}$ where $v = \max\{\frac{\beta^2 B}{\beta\mu-1}, (\gamma+1)\Delta_1\}$.

这里对学习率 η_t 及相关参数进行了限制，并确定下一步的目标是证明 $\Delta_t \leq \frac{v}{\gamma+t}$，即摆脱递推关系。这是通过以下的数学归纳法证明的。

We prove it by induction. Firstly, the definition of v ensures that it holds for $t = 1$.

当 $t = 1$ 时，由于 $v = \max\{\frac{\beta^2 B}{\beta\mu-1}, (\gamma+1)\Delta_1\}$，无论两项中哪个更大，$\Delta_1 \leq \frac{v}{\gamma+1}$ 都成立。

Assume the conclusion holds for some t, it follows that

$$\begin{aligned}
\Delta_{t+1} &\leq (1 - \eta_t\mu)\Delta_t + \eta_t^2 B \\
&\leq (1 - \frac{\beta\mu}{t+\gamma})\frac{v}{t+\gamma} + \frac{\beta^2 B}{(t+\gamma)^2} \\
&= \frac{t+\gamma-1}{(t+\gamma)^2}v + [\frac{\beta^2 B}{(t+\gamma)^2} - \frac{\beta\mu-1}{(t+\gamma)^2}v] \\
&\overset{(a)}{\leq} \frac{v}{t+\gamma+1}.
\end{aligned}$$

(a): 因为 v 的定义，$v \geq \frac{\beta^2 B}{\beta\mu-1}$，前面限制 $\gamma > \frac{1}{\mu}$，分母移到左边即可证明 [] 中的项小于 0，可以放缩掉。至此 $\Delta_t \leq \frac{v}{\gamma+t}$ 证明结束。

Then by the L-smoothness of $F(\dot{})$,

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

这里使用 L-smooth 假设，将判定指标的上界与 Δ_t 相关联。由于 F^* 为最小的目标函数，认为它的梯度等于 0，则内积 $< \nabla F(\mathbf{w}^*), \bar{\mathbf{w}}_t - \mathbf{w}^* >$ 等于 0。之后设定具体的学习率，得到最后的收敛结果。

Specifically, if we choose $\beta = \frac{2}{\mu}$, $\gamma = \max\{8\frac{L}{\mu}, E\} - 1$ and denote $\kappa = \frac{L}{\mu}$, then $\eta_t = \frac{2}{\mu}\frac{1}{\gamma+t}$. Then, we have

$$v = \max\{\frac{\beta^2 B}{\beta\mu-1}, (\gamma+1)\Delta_1\} \leq \frac{\beta^2 B}{\beta\mu-1} + (\gamma+1)\Delta_1 \leq \frac{4B}{\mu^2} + (\gamma+1)\Delta_1,$$

and

$$\mathbb{E}[F(\bar{\mathbf{w}}_t) - F^*] \leq \frac{L}{2}\frac{v}{\gamma+t} \leq \frac{\kappa}{\gamma+t}\left(\frac{2B}{\mu} + \frac{\mu(\gamma+1)}{2}\Delta_t\right).$$

定理 1 及其关键引理的详细证明过程参见 Xiang Li 等人的工作[5]，也可以参考视频讲解[6]。

7 定理 2 的证明

定理 2 的证明参考了 SGD 在光滑非凸函数上的收敛性证明[7]。

Proof. The update rule of FedAvg is

$$\bar{\mathbf{w}}_{t+1} = \bar{\mathbf{w}}_t - \eta_t\mathbf{g}_t. \tag{7}$$

The Assumption 1 implies that

$$F(\bar{\mathbf{w}}_{t+1}) - F(\bar{\mathbf{w}}_t) - \langle\nabla F(\bar{\mathbf{w}}_t), \bar{\mathbf{w}}_{t+1} - \bar{\mathbf{w}}_t\rangle \leq \frac{L}{2}\|\bar{\mathbf{w}}_{t+1} - \bar{\mathbf{w}}_t\|^2. \tag{8}$$

从 L-smooth 开始证明。

Substitute (7) into (8) we get

$$F(\bar{\mathbf{w}}_{t+1}) - F(\bar{\mathbf{w}}_t) + \eta_t \langle \nabla F(\bar{\mathbf{w}}_t), \mathbf{g}_t \rangle \leq \frac{L}{2}\eta_t^2 \|\mathbf{g}_t\|^2. \tag{9}$$

将更新公式代入 L-smooth 的结果，消去 $\bar{\mathbf{w}}_{t+1} - \bar{\mathbf{w}}_t$ 项，然后使用一些结论展开内积项：

For the first term on the right side, applying **Fact 1**, we have

$$\eta \langle \nabla F(\bar{\mathbf{w}}_t), \mathbf{g}_t \rangle = \frac{\eta_t}{2}\|\nabla F(\bar{\mathbf{w}}_t)\|^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|^2 - \frac{\eta_t}{2}\|\nabla F(\bar{\mathbf{w}}_t) - \bar{\mathbf{g}}_t\|^2. \tag{10}$$

Substituting (10) into (9), we have

$$F(\bar{\mathbf{w}}_{t+1}) - F(\bar{\mathbf{w}}_t) + \frac{\eta_t}{2}\|\nabla F(\bar{\mathbf{w}}_t)\|^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|^2 \leq \frac{\eta_t}{2}\|\nabla F(\bar{\mathbf{w}}_t) - \mathbf{g}_t\|^2 + \frac{L}{2}\eta_t^2\|\mathbf{g}_t\|^2. \tag{11}$$

After shifting terms and multiplying $\frac{2}{\eta_t}$ on both sides, we have

$$\|\nabla F(\bar{\mathbf{w}}_t)\|^2 \leq \frac{2}{\eta_t}[F(\bar{\mathbf{w}}_t) - F(\bar{\mathbf{w}}_{t+1})] + \|\nabla F(\bar{\mathbf{w}}_t) - \mathbf{g}_t\|^2 + (\eta_t L - 1)\|\mathbf{g}_t\|^2. \tag{12}$$

之后就是使用引理和假设将右边的每一项放缩：

Taking expected values on both sides and applying Lemma (2), Lemma (4), Lemma (5), we have

$$\mathbb{E}\|\nabla F(\bar{\mathbf{w}}_t)\|^2 \leq \frac{2}{\eta_t}\mathbb{E}[F(\bar{\mathbf{w}}_t) - F(\bar{\mathbf{w}}_{t+1})] + \sum_{k=1}^{N} p_k^2\sigma_k^2 + (\eta_t L - 1)\sum_{k=1}^{N} p_k^2 G^2. \tag{13}$$

Summing over $t \in \{0, 1, \cdots, T-1\}$ and dividing both sides by T, we have

$$\min_t \mathbb{E}\|\nabla F(\bar{\mathbf{w}}_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\bar{\mathbf{w}}_t)\|^2 \leq \frac{2}{\eta_t T}\mathbb{E}[F(\bar{\mathbf{w}}_0) - F^*] + \sum_{k=1}^{N} p_k^2\sigma_k^2 + (\eta_t L - 1)\sum_{k=1}^{N} p_k^2 G^2. \tag{14}$$

参考文献

[1] Rocket. 强凸与光滑性[EB/OL]. 2021. https://zhuanlan.zhihu.com/p/369961290.

[2] 雍泰. 为什么在光滑凸优化研究中，Lipschitz gradient 比 strongly convex 更普遍？[EB/OL]. 2024. https://www.zhihu.com/question/459410340/answer/1888570770.

[3] 大厂推荐算法. 【科研喂饭】深度学习算法收敛性证明之 SGD[EB/OL]. 2021. https://zhuanlan.zhihu.com/p/338108328.

[4] 大厂推荐算法. 【科研喂饭】深度学习算法收敛性证明之拓展 SGD[EB/OL]. 2021. https://zhuanlan.zhihu.com/p/351682784.

[5] LI X, HUANG K, YANG W, et al. On the Convergence of FedAvg on Non-IID Data[J/OL]. ArXiv, 2019, abs/1907.02189. https://arxiv.org/abs/1907.02189.

[6] 丸一口. 【收敛性分析】Non-IID + FedAvg 收敛性分析「全设备参加」[EB/OL]. 2023. https://www.bilibili.com/video/BV1Av4y1E7Lg/?spm_id_from=333.999.0.0&vd_source=e63f08e3795a7d51a7cfc6c0294d87ee.

[7] 丸一口. 【收敛性分析】PL：使敛析变得更简单[EB/OL]. 2023. https://www.bilibili.com/video/BV1sP411i7Tc/?spm_id_from=333.999.0.0&vd_source=e63f08e3795a7d51a7cfc6c0294d87ee.