Optimization Theory and Algorithm II	October 17, 2022

Lecture 11

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 Federated Optimization

Federated learning (FL) enables a large amount of edge computing devices to jointly optimize (learn) a model without data sharing. FL has three unique characters that distinguish it from the standard parallel optimization.

- 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.
- The FL system does not have control over user's device (stragglers).
- The training data are non-i.i.d.

Problem Formulation:

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}) \right\} \tag{1}$$

where K is the number of devices, and p_k is the weight of the kth device such that $p_k \geq 0$ and $\sum_k p_k = 1$. Suppose that kth device holds m_k training data: $\mathbf{z}_{k,1}, \ldots, \mathbf{z}_{k,m_k}$, then

$$f_k(\mathbf{x}) = \frac{1}{m_k} \sum_{j=1}^{m_k} \ell(\mathbf{x}; \mathbf{z}_{k,j}).$$

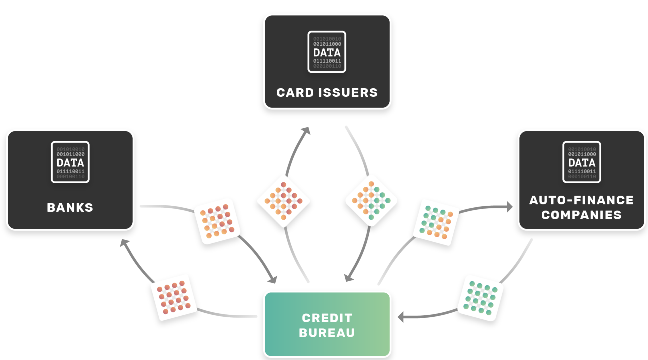

Figure 1: Federated Learning for Credit Scoring

Example 1 *(Federated Least Squares Problem) Suppose that we have K banks, they would like to jointly to train a model to predict the customer's income for "user profile" or to train a score system to estimate their* *financial credit (see Figure 1). They adopt a linear regression model, then*

$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2 = \frac{1}{2}\sum_{k=1}^{K}\|A_k\mathbf{x}-\mathbf{b}_k\|^2,$$

where

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_K \end{bmatrix} \in \mathbb{R}^{m\times n}, \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_K \end{bmatrix} \in \mathbb{R}^{m}.$$

However, we cannot combine the personal data set together due to the sensitive information and law regulations (E.g., GDPR). Then the idea is to transmit some information to a central server without sharing any dataset.

For the kth bank, it considers

$$\min_{\mathbf{x}} \frac{1}{2}\|A_k\mathbf{x}-\mathbf{b}_k\|^2.$$

Denote an operator $G_k(\mathbf{x}) = \mathbf{x} - s\nabla_{\mathbf{x}}(\frac{1}{2}\|A_k\mathbf{x}-\mathbf{b}_k\|^2) = (I - sA_k^\top A_k)\mathbf{x} + sA_k^\top \mathbf{b}_k$. *The federated gradient descent algorithm is*

$$\text{Step 1: } \mathbf{x}_k^{t+1/2} := G_k^E(\mathbf{x}_k^t), \tag{2}$$

$$\text{Step 2: } \mathbf{x}^{t+1} := \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k^{t+1/2}, \tag{3}$$

$$\text{Step 3: } \mathbf{x}_k^{t+1} := \mathbf{x}^{t+1}, \ \forall k \in [K], \tag{4}$$

where $G_k^E(\mathbf{x})$ *means that runs GD on the kth device E times.*

First, let us try to compute $G_k^2(\mathbf{x})$ *as*

$$\begin{aligned} G_k^2(\mathbf{x}) &= G_k(G_k(\mathbf{x})) = G_k((I - sA_k^\top A_k)\mathbf{x} + sA_k^\top \mathbf{b}_k) \\ &= (I - sA_k^\top A_k)((I - sA_k^\top A_k)\mathbf{x} + sA_k^\top \mathbf{b}_k) + sA_k^\top \mathbf{b}_k \\ &= (I - sA_k^\top A_k)^2\mathbf{x} + s[I + (I - sA_k^\top A_k)]A_k^\top \mathbf{b}_k. \end{aligned}$$

By induction, you can obtain that

$$G_k^E(\mathbf{x}) = (I - sA_k^\top A_k)^E\mathbf{x} + s[\sum_{e=0}^{E-1}(I - sA_k^\top A_k)^e]A_k^\top \mathbf{b}_k. \tag{5}$$

Thus,

$$\begin{aligned} \mathbf{x}^{t+1} &= \bar{\mathbf{x}}^{t+1/2} = \frac{1}{K}\sum_k \mathbf{x}_k^{t+1/2} = \frac{1}{K}\sum_k G_k^E(\mathbf{x}_k^t) \\ &= \frac{1}{K}\sum_k G_k^E(\mathbf{x}^t) = \frac{1}{K}[\sum_{k=1}^{K}(I - sA_k^\top A_k)^E]\mathbf{x}^t + \frac{s}{K}\sum_{k=1}^{K}\{[\sum_{e=0}^{E-1}(I - sA_k^\top A_k)^e]A_k^\top \mathbf{b}_k\}. \end{aligned}$$

That is $\mathbf{x}^{t+1} = B\mathbf{x}^t + C$, *where*

$$B = \frac{1}{K}\sum_k G_k^E(\mathbf{x}^t) = \frac{1}{K}[\sum_{k=1}^{K}(I - sA_k^\top A_k)^E]$$

and

$$C = \frac{s}{K}\sum_{k=1}^{K}\{[\sum_{e=0}^{E-1}(I - sA_k^\top A_k)^e]A_k^\top \mathbf{b}_k\}.$$

We konw that

$$\mathbf{x}^{t+1} = B^{t+1}\mathbf{x}^0 + (I + B + \cdots + B^t)C$$
$$= B^{t+1}\mathbf{x}^0 + (I - B)^{-1}(I - B^{t+1})C.$$

So,

$$\mathbf{x}^*_{FGD} = \lim_{t\to\infty} \mathbf{x}^t = (I - B)^{-1}C.$$

Compute that

$$I - B = \frac{1}{K}\sum_{k=1}^{K}[I - (I - sA_k^\top A_k)^E]$$
$$= \frac{1}{K}\sum_{k=1}^{K}(sA_k^\top A_k)\sum_{e=0}^{E-1}(I - sA_k^\top A_k)^e.$$

$$\mathbf{x}^*_{FGD} = [\sum_{k=1}^{K} A_k^\top A_k \sum_{e=0}^{E-1}(I - sA_k^\top A_k)^e]^{-1}\sum_{k=1}^{K}\{[\sum_{e=0}^{E-1}(I - sA_k^\top A_k)^e]A_k^\top \mathbf{b}_k\}. \tag{6}$$

We compare this result with

$$\mathbf{x}^*_{LS} = (A^\top A)^{-1}A^\top \mathbf{b} = [\sum_{k=1}^{K} A_k^\top A_k]^{-1}\sum_{k=1}^{K} A_k^\top \mathbf{b}_k.$$

If $E = 1$, then $\mathbf{x}^*_{FGD} = \mathbf{x}^*_{LS}$. Otherwise, $\mathbf{x}^*_{FGD} \neq \mathbf{x}^*_{LS}$.

1.1 FedAvg and Local SGD

FedAvg algorithm is proposed by [1] for training deep models distributed and efficiently. They used the mini-batch SGD as the algorithm for local training. Here, we present a slightly different setting called Local SGD which means that the SGD as the algorithm for local training.

Algorithm 1 Local Stochastic Gradient Descent

Input: Assumes that K clients index by k, E is the number of local iterations, s_t is the learning rate $\mathbf{x}^0 \in \mathbb{R}^n$, the total iteration number is T, and $t = 0$.
for $t = 1, E, 2E, \ldots, T$ **do**
  for $k = 1, \ldots, K$ **do**
    Local Update:
$$\mathbf{x}_k^{t+i+1} \leftarrow \mathbf{x}_k^{t+i} - s_{t+i}\nabla f_k(\mathbf{x}_k^{t+i}, \xi_k^{t+i}), i = 0, \ldots, E-1,$$
    where ξ_k^{t+i} is a sample uniformly chosen from the local data and s_{t+i} is the learning rate.
  end for
  Server Update by Aggregation:
$$\mathbf{x}^{t+E} \leftarrow \sum_{k=1}^{K} p_k \mathbf{x}_k^{t+E}.$$
  Update Local Parameter:
$$\mathbf{x}_k^{t+E} \leftarrow \mathbf{x}^{t+E}, \forall k = 1, \ldots, K.$$
end for
Output: $\mathbf{x}^T$.

Let us summary the local SGD algorithm as follows:

- Local Update:
$$\mathbf{x}_k^{t+i+1} \leftarrow \mathbf{x}_k^{t+i} - s_{t+i}\nabla f_k(\mathbf{x}_k^{t+i}, \xi_k^{t+i}), i = 0, \dots, E-1,$$
where ξ_k^{t+i} is a sample uniformly chosen from the local data and s_{t+i} is the learning rate.
- Server Update by Aggregation:
$$\mathbf{x}^{t+E} \leftarrow \sum_{k=1}^{K} p_k \mathbf{x}_k^{t+E}.$$
- Update Local Parameter:
$$\mathbf{x}_k^{t+E} \leftarrow \mathbf{x}^{t+E}, \forall k = 1, \dots, K.$$

Let T be the total interactions, then $[2T/E]$ is the communication number.

1.2 Convergence

Assumption 1 *(A1) f_k is β-smooth for all $k \in [K]$.*

Assumption 2 *(A2) f_k is α-strongly convex for all $k \in [K]$.*

Assumption 3 *(A3)*

Control variance:
$$\mathbb{E}\|\nabla f_k(\mathbf{x}_k^t, \xi_k^t) - \nabla f_k(\mathbf{x}_k^t)\|^2 \leq \sigma_k^2, \forall k \in [K].$$

Assumption 4 *(A4)*

Bounded Gradient:
$$\mathbb{E}\|\nabla f_k(\mathbf{x}_k^t, \xi_k^t)\|^2 \leq G, \forall k \in [K], t \in [T].$$

Let $\Gamma = f^* - \sum_{k=1}^{K} p_k f_k^*$ for quantifying the degree of non-i.i.d which reflects the heterogeneity of data distribution. If data is i.i.d., then Γ obviously goes to zero as $m \to \infty$.

Theorem 1 *[2] Assume that A1, A2, A3 and A4 hold. Let $\kappa = \beta/\alpha, \gamma = \max\{8\kappa, E\}, s_t = \frac{2}{\alpha(\gamma+t)}$, then*
$$\mathbb{E}[f(\mathbf{x}^T) - f^*] \leq \frac{\kappa}{\gamma + T - 1}(\frac{2B}{\alpha} + \frac{\alpha\gamma}{2}\mathbb{E}[\|\mathbf{x}^0 - \mathbf{x}^*\|^2]), \tag{7}$$
where $B = \sum_{k=1}^{K} p_k^2 \sigma_k^2 + 6\beta\Gamma + 8(E-1)^2 G^2$.

To justify the above theorem, let us define
$$\mathbf{v}_k^{t+1} = \mathbf{x}_k^t - s_t \nabla f_k(\mathbf{x}_k^t, \xi_k^t),$$
and
$$\mathbf{x}_k^{t+1} = \begin{cases} \mathbf{v}_k^{t+1}, & t+1 \notin \mathcal{I}_E, \\ \sum_{k=1}^{K} p_k \mathbf{v}_k^{t+1}, & t+1 \in \mathcal{I}_E, \end{cases}$$
where $\mathcal{I}_E = \{iE | i = 1, 2, \dots\}$. We further define two virtual sequences
$$\bar{\mathbf{v}}^t = \sum_{k=1}^{K} p_k \mathbf{v}_k^t, \quad \bar{\mathbf{x}}^t = \sum_{k=1}^{K} p_k \mathbf{x}_k^t.$$

$$\bar{\mathbf{g}}^t = \sum_{k=1}^K p_k \nabla f_k(\mathbf{x}_k^t), \quad \mathbf{g}^t = \sum_{k=1}^K p_k \nabla f_k(\mathbf{x}_k^t, \xi_k^t).$$

Thus, $\mathbb{E}\mathbf{g}^t = \bar{\mathbf{g}}^t$. If $t+1 \in \mathcal{I}_E$, then

$$\mathbf{x}_k^{t+1} = \sum_{k=1}^K p_k \mathbf{v}_k^{t+1} = \bar{\mathbf{v}}^{t+1} = \bar{\mathbf{x}}^{t+1}.$$

Lemma 1

$$\begin{aligned}\mathbb{E}\|\bar{\mathbf{v}}^{t+1} - \mathbf{x}^*\|^2 \leq (1 - s_t\alpha)\mathbb{E}\|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2 + s_t^2\mathbb{E}\|\mathbf{g}^t - \bar{\mathbf{g}}^t\|^2 \\ + 6\beta s_t^2 \Gamma + 2\mathbb{E}[\sum_{k=1}^K p_k \|\bar{\mathbf{x}}^t - \mathbf{x}_k^t\|^2].\end{aligned}$$

Lemma 2 *If A3 holds, then*

$$\mathbb{E}\|\mathbf{g}^t - \bar{\mathbf{g}}^t\|^2 \leq \sum_{k=1}^K p_k^2 \sigma_k^2.$$

Lemma 3 *If A4 holds and $s_t \leq 2s_{t+E}$, then*

$$\mathbb{E}[\sum_{k=1}^K p_k \|\bar{\mathbf{x}}^t - \mathbf{x}_k^t\|^2] \leq 4s_t^2(E-1)^2G^2.$$

Now we have all the materials to prove Theorem 1.

Proof 1 *According to above three lemmas, then*

$$\Delta_{t+1} \leq (1 - s_t\alpha)\Delta_t + s_t^2 B,$$

where $\Delta_t = \mathbb{E}\|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2$ and $B = \sum_{k=1}^K p_k^2\sigma_k^2 + 6\beta\Gamma + 8(E-1)^2G^2$.

For a diminishing learning rate $s_t = \frac{\ell}{t+\gamma}, \ell > 1/\alpha$ and $\gamma > 0$, such that $s_1 \leq \min\{1/\alpha, 1/4\beta\} = 1/4\beta$ and $s_t \leq 2s_{t+E}$. We will prove

$$\Delta_t \leq \frac{\nu}{\gamma + t}$$

where $\nu = \max\{\frac{\ell^2 B}{\ell\alpha - 1}, (\gamma+1)\Delta_1\}$, by induction.

For $t = 1$, it already holds, then assume the results holds for $t > 1$. We know that $(t+\gamma)^2 - 1 = (t+\gamma-1)(t+\gamma+1) \leq (t+\gamma)^2$ and $\ell^2 B - (\ell\alpha - 1)\nu < 0$, thus, it follows that

$$\begin{aligned}\Delta_{t+1} &\leq (1 - s_t\alpha)\Delta_t + s_t^2 B \\ &\leq (1 - \frac{\ell\alpha}{t+\gamma})\frac{\nu}{\gamma+t} + \frac{\ell^2 B}{(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2}\nu + \left[\frac{\ell^2 B}{(\gamma+t)^2} - \frac{\ell\alpha - 1}{(t+\gamma)^2}\nu\right] \\ &\leq \frac{\nu}{\gamma+t+1}.\end{aligned}$$

Moreover, by the β-smooth property,

$$\mathbb{E}[f(\bar{\mathbf{x}}^t] - f^* \leq \frac{\beta}{2}\Delta_t \leq \frac{\beta\nu}{2(\gamma+t)}.$$

Choose $\ell = 2/\alpha, \kappa = \beta/\alpha, \gamma = \max\{8\kappa, E\}, s_t = \frac{2}{\alpha(\gamma+t)}$, *then*

$$\mathbb{E}[f(\bar{\mathbf{x}}^t] - f^* \leq \frac{\kappa}{\gamma+t}(\frac{2B}{\alpha} + \frac{\alpha(\gamma+1)}{2}\Delta_1).$$

Let $\bar{\mathbf{x}}^t = \mathbf{x}^T$, *then we obtain the final results.*

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[2] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.