Appendix

May 24, 2024

1 Optimization problem and assumptions

1.1 Fedrated Learning

This paper focuses on a parameter server (PS) architecture, comprising a centralized PS for global aggregation and a set of N distributed local devices represented as $\mathcal{N}$. The PS maintains a global model $\mathbf{w} \in \mathbb{R}^d$. Each device, represented by $i \in \mathcal{N}$, possesses a model weight $\mathbf{w}_i$ and a local dataset D_i. Within this dataset, there exist $|D_i|$ data samples, expressed as $\xi_i = [\xi_{i,1}, \xi_{i,2}, \cdots, \xi_{i,|D_i|}]$, which are utilized for local training. We define the loss function for each data sample $\xi_{i,j}$ ($j \in [1, |D_i|]$), as $f(\mathbf{w}_i, \xi_{i,j})$, and denote the local loss function of device i as:

$$F_i(\mathbf{w}_i) := \frac{1}{|D_i|} \sum_{j=1}^{|D_i|} f(\mathbf{w}_i, \xi_{i,j}), \tag{1}$$

The target of this system is to train a global model $\mathbf{w}$ that minimizes the global loss function defined as:

$$F(\mathbf{w}) = \frac{1}{N} \sum_{i \in \mathcal{N}} F_i(\mathbf{w}) \tag{2}$$

In each round of training, participating devices perform local training on their respective datasets using the current global model. During this local training, each device computes the gradient of the loss function with respect to its local data samples. After local training, devices communicate with the parameter server and send their computed gradients. The parameter server aggregates these gradients to obtain a new global model. The newly aggregated global model is then distributed back to the participating devices, replacing their previous local models. This iterative update process continues for multiple rounds, allowing the global model to be refined and improved over time.

1.2 Asynchronous FL with Periodic Aggregation

We specifically targets the domain of asynchronous federated learning with periodic aggregation. The fundamental concept underlying this approach is to enable independent training processes across different devices, where the server periodically aggregates the received updates from devices that have finished their computations, while allowing other devices to continue their local training uninterrupted. Considering the diversity in computational capabilities and communication capabilities among devices, once a device completes its local training, it transmits its local gradient to the server. Specifically, the entire training process comprises a total of T periods, *(i.e., a global round)*. Within each period, every local device undertakes k local iterations. At each local iteration j, ranging from 0 to $k-1$, local device i updates its local model following the prescribed rule:

$$\mathbf{w}_i^{t,j+1} = \mathbf{w}_i^{t,j} - \eta_l \nabla F_i(\mathbf{w}_i^{t,j}, \xi_i^{t,j}) \tag{3}$$

where η_l is the learning rate of local device, and $\mathbf{w}_i^{t,j}$ is the model of j-th local iteration of device i training with global model $\mathbf{w}^t$.

When a local device i has finished its k local updates to train the global model $\mathbf{w}^t$, it computes the overall gradient $\mathbf{g}_i^t$ in local training, that is:

$$\mathbf{g}_i^t = \mathbf{w}_i^{t,0} - \mathbf{w}_i^{t,k} \tag{4}$$

Note that when a local device i receives the t-th global model $\mathbf{w}^t$, it will be initialized with $\mathbf{w}_i^{t,0} = \mathbf{w}^t$. Then, the local device will uploads a compressed update $\tilde{\mathbf{g}}_i^t = C(\mathbf{g}_i^t)$ to the parameter server. Therefore, the total time for training and communication of device i is:

$$d_i = k\alpha_i + \delta\beta_i \tag{5}$$

where we define the number of local iterations is k and the compression rate of compressor C is δ. Let α_i denote the computation time required for one local iteration on device i, and β_i represent the communication time for

transmitting a full model on the same device. Given that the download bandwidth is typically bigger than the upload bandwidth[**?**, 4], our attention is primarily directed towards the communication time involved in transmitting the models from devices to the parameter server (PS) during the model exchange process.

At the same time the server continuously receives gradients from local devices. We define $\mathbf{S}^t$ as the set of local devices to which the server has received gradients in the t-th global round. The parameter server then aggregate the received local gradient from $\mathbf{S}^t$ and updates the global model according to:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \frac{\eta_g}{|\mathbf{S}^t|} \sum_{i \in \mathbf{S}^t} \tilde{\mathbf{g}}_i^{t-\tau_i} \tag{6}$$

where η_g is the global learning rate. Due to the asynchronous nature, the gradient may be stale. That is, the gradient of device $\tilde{\mathbf{g}}_i^{t-\tau_i}$ is generated by device i to train the global model $\mathbf{w}^{t-\tau_i}$. But when the parameter server received $\tilde{\mathbf{g}}_i^{t-\tau_i}$, it is executing the t-th round of aggregation. Based on the above content, the staleness of device i to train the global model $\mathbf{w}^{t-\tau_i}$ is τ_i. And τ_i is computed by the following rule:

$$\tau_i = t - \underset{t^{'}<t}{max}\{t^{'} | i \in \mathbf{S}^{t^{'}}\} \tag{7}$$

In the context of this paper, the staleness τ_i can also be calculated with $\tau_i = \lceil \frac{d_i}{\tilde{T}} \rceil$, where $\tilde{T}$ is the clock time of one period. When the number of local iterations and compression rate are unchanging, the staleness τ_i is always equal to $\tau_i = \lceil \frac{d_i}{\tilde{T}} \rceil$, *i.e.* $\tau_i = \tau_i = \lceil \frac{d_i}{\tilde{T}} \rceil, \forall t \leq T$.

Asynchronous FL with periodic aggregation is summarized in Algorithm 1.

Algorithm 1 Asynchronous FL with periodic aggregation

Server:
Broadcast $\mathbf{w}^0$ to all devices and start them
for $t = 0, 1, \cdots, T-1$: **do**
  Continuously receive $\tilde{\mathbf{g}}_i^{t-\tau_i}$ from local devices set $\mathbf{S}^t$
  $\mathbf{w}^{t+1} = \mathbf{w}^t - \frac{\eta_g}{|\mathbf{S}^t|} \sum_{i \in \mathbf{S}^t} \tilde{\mathbf{g}}_i^{t-\tau_i}$
  for $i \in \mathbf{S}^t$ **do**
    Send new global model $\mathbf{w}^{t+1}$ to i
  end for
end for
Notice all devices to *STOP*

Device:
while not *STOP*: **do**
  Receive $\mathbf{w}^t$ from server
  Set $\mathbf{w}_i^{t,0} \leftarrow \mathbf{w}^t$
  for each local iteration $j \in 0, 1, \cdots, k-1$ **do**
    $\mathbf{w}_i^{t,j+1} \leftarrow \mathbf{w}_i^{t,j} - \eta_l \nabla F_i(\mathbf{w}_i^{t,j}, \xi_i^{t,j})$
  end for
  Compute gradient $\mathbf{g}_i^t \leftarrow \mathbf{w}_i^{t,0} - \mathbf{w}_i^{t,k}$
  Compute compressed gradient $\tilde{\mathbf{g}}_i^t \leftarrow C(\mathbf{g}_i^t)$
  Send $\tilde{g}_i^t$ to server
end while

1.3 Assumptions

Assumption 1. (Bounded local variance). *There exists a constant σ, such that the variance of each the local estimator is bounded by:*

$$\mathbb{E}_{\xi \sim \mathcal{D}_i} \left[||\nabla F_i(\mathbf{w}, \xi) - \nabla F_i(\mathbf{w})|| \right] \leq \sigma, \forall i \in \mathcal{N}, \forall \mathbf{w} \in \mathbb{R}^d \tag{8}$$

Assumption 2. (Bounded function heterogeneity). *There exists N constants $\zeta_i^2 \geq 0, i \in \mathcal{N}$, such that the variance of the model gradients is bounded by:*

$$||\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})||^2 \leq \zeta_i^2, \forall \mathbf{w} \in \mathbb{R}^d \tag{9}$$

and we define $\zeta^2 := \frac{1}{N} \sum_{i \in \mathcal{N}} \zeta_i^2$.

Assumption 3. (L-smooth). *The loss functions F and F_i are L-smooth with a constant $L \geq 0$ such that:*

$$||\nabla F_i(\mathbf{y}) - \nabla F_i(\mathbf{x})|| \leq L||\mathbf{y} - \mathbf{x}||, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{10}$$

Assumption 4. (Bounded gradient). *There exists a constant $G \geq 0$ such that the norm of local gradient is bounded by:*

$$||\nabla F_i(\mathbf{w})||^2 \leq G^2, \forall \mathbf{w} \in \mathbb{R}^d \tag{11}$$

1.4 Useful inequalities and equalities

$$< \mathbf{a}, \mathbf{b} > \leq \frac{||\mathbf{a}||^2 + ||\mathbf{b}||^2}{2} \tag{12}$$

$$||\sum_{i=1}^{N} \mathbf{a}_i||^2 \leq N \sum_{i=1}^{N} ||\mathbf{a}_i||^2 \tag{13}$$

$$||\mathbf{a} + \mathbf{b}||^2 \leq (1+\alpha)||\mathbf{a}||^2 + (1+\alpha^{-1})||\mathbf{b}||^2 \tag{14}$$

$$< \mathbf{a}, \mathbf{b} >= \frac{1}{2}(||\mathbf{a}||^2 + ||\mathbf{b}||^2 - ||\mathbf{a} - \mathbf{b}||^2) \tag{15}$$

2 Theoretical Analysis

In order to analyze the convergence of Algorithm 1, and give the parameter optimization method. We analyzed the convergence of Algorithm 1.

The function is $L-$smooth, we have:

$$\mathbb{E}\big[F(\mathbf{w}^{t+1})\big] \leq F(\mathbf{w}^t) \underbrace{-\eta_g \mathbb{E}\big[\big\langle \nabla F(\mathbf{w}^t), \sum_{i \in \mathbf{S}^t} \frac{1}{|\mathbf{S}^t|} \tilde{\mathbf{g}}_i^{t-\tau_i} \big\rangle\big]}_{:=X_1} + \frac{L\eta_g^2}{2} \underbrace{\mathbb{E}\big[|| \sum_{i \in \mathbf{S}^t} \frac{1}{|\mathbf{S}^t|} \tilde{\mathbf{g}}_i^{t-\tau_i} ||^2\big]}_{:=X_2} \tag{16}$$

We define the expectation as :

$$\mathbb{E}[\cdot] = \mathbb{E}_{i \sim \mathcal{N}} \mathbb{E}_{\xi|i} \mathbb{E}_{C|\xi,i}[\cdot]$$

We first derive the first term X_1:

$$\begin{aligned}
X_1 &= -\eta_g \mathbb{E}\big[\big\langle \nabla F(\mathbf{w}^t), \sum_{i \in \mathbf{S}^t} \frac{1}{|\mathbf{S}^t|} \tilde{\mathbf{g}}_i^{t-\tau_i} \big\rangle\big] \\
&= -\eta_g \mathbb{E}\big[\sum_{i \in \mathbf{S}^t} \frac{1}{|\mathbf{S}^t|} \big\langle \nabla F(\mathbf{w}^t), C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k}) \big\rangle\big] \\
&= -\eta_g \mathbb{E}\big[\sum_{i \in \mathbf{S}^t} \frac{1}{|\mathbf{S}^t|} \big\langle \nabla F(\mathbf{w}^t), C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k}) \big\rangle\big] \\
&= -\frac{\eta_g}{2} \mathbb{E}\big[\sum_{i \in \mathbf{S}^t} \frac{1}{|\mathbf{S}^t|} (||\nabla F(\mathbf{w}^t)||^2 + ||C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2) - ||\nabla F(\mathbf{w}^t) - C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big] \\
&= -\frac{\eta_g}{2} ||\nabla F(\mathbf{w}^t)||^2 - \frac{\eta_g}{2} \mathbb{E}\big[\frac{1}{|\mathbf{S}^t|} \sum_{i \in \mathbf{S}^t} ||C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big] \\
&\quad + \frac{\eta_g}{2} \mathbb{E}\big[\frac{1}{|\mathbf{S}^t|} \sum_{i \in \mathbf{S}^t} ||\nabla F(\mathbf{w}^t) - C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big]
\end{aligned}$$

where the result of the penultimate equal sign is obtained according to inequality (15)

For the last term X_2, we have:

$$\begin{aligned}
X_2 &= \mathbb{E}\big[||\frac{1}{|\mathbf{S}^t|} \sum_{i \in \mathbf{S}^t} \tilde{\mathbf{g}}_i^{t-\tau_i}||^2\big] \\
&= \mathbb{E}\big[||\frac{1}{|\mathbf{S}^t|} \sum_{i \in \mathbf{S}^t} C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big] \\
&\overset{(13)}{\leq} \mathbb{E}\big[\frac{1}{|\mathbf{S}^t|} \sum_{i \in \mathbf{S}^t} ||C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big]
\end{aligned}$$

where the result of the last equal sign is obtained according to Inequality (13)
Combine X_1 and X_2 to the original inequality (16):

$$\begin{aligned}
\mathbb{E}\big[F(\mathbf{w}^{t+1})\big] &\leq F(\mathbf{w}^t) - \frac{\eta_g}{2}||\nabla F(\mathbf{w}^t)||^2 - \frac{\eta_g}{2}\mathbb{E}\big[\frac{1}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}||C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big] \\
&\quad + \frac{\eta_g}{2}\mathbb{E}\big[\frac{1}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}||\nabla F(\mathbf{w}^t) - C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big] + \frac{L\eta_g^2}{2}\mathbb{E}\big[\frac{1}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}||C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big] \\
&= F(\mathbf{w}^t) - \frac{\eta_g}{2}||\nabla F(\mathbf{w}^t)||^2 + \frac{\eta_g}{2}(L\eta_g - 1)\mathbb{E}\big[\frac{1}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}||C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big] \\
&\quad + \underbrace{\frac{\eta_g}{2}\mathbb{E}\big[\frac{1}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}||\nabla F(\mathbf{w}^t) - C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big]}_{X_3}
\end{aligned}$$

Then, we simplify $\xi_i^{t-\tau_i,j}$ as ξ. To derive the upper bound of the above inequality, we focus on the X_3 term:

$$\begin{aligned}
X_3 &= \mathbb{E}\big[\frac{1}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}||\nabla F(\mathbf{w}^t) - C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big] \\
&= \mathbb{E}\big[\frac{1}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}||\nabla F(\mathbf{w}^t) - C(\eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j},\xi))||^2\big] \\
&= \mathbb{E}\big[\frac{1}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}||(1-\eta_l k)\nabla F(\mathbf{w}^t) + \eta_l\sum_{j=0}^{k-1}\nabla F(\mathbf{w}^t) \pm \eta_l\sum_{j=0}^{k-1}\nabla F(\mathbf{w}^{t-\tau_i}) \\
&\quad \pm \eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}^{t-\tau_i}) \pm \eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j}) \pm \eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j},\xi) - C(\eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j},\xi))||^2\big] \\
&\leq \mathbb{E}\big[\frac{6}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}((1-\eta_l k)^2||\nabla F(\mathbf{w}^t)||^2 + ||\eta_l\sum_{j=0}^{k-1}\nabla F(\mathbf{w}^t) - \eta_l\sum_{j=0}^{k-1}\nabla F(\mathbf{w}^{t-\tau_i})||^2 \\
&\quad + ||\eta_l\sum_{j=0}^{k-1}\nabla F(\mathbf{w}^{t-\tau_i}) - \eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}^{t-\tau_i})||^2 + ||\eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}^{t-\tau_i}) - \eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j})||^2 \\
&\quad + ||\eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j}) - \eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j},\xi)||^2 + ||\eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j},\xi) - C(\eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j},\xi))||^2)\big] \\
&\leq \mathbb{E}\big[\frac{6}{|\mathbf{S}^t|}\sum_{i\in\mathbf{S}^t}((1-\eta_l k)^2||\nabla F(\mathbf{w}^t)||^2 + L^2\eta_l^2 k^2||\mathbf{w}^t - \mathbf{w}^{t-\tau_i}||^2 \\
&\quad + \eta_l^2 k^2\zeta_i^2 + L^2\eta_l^2 k\sum_{j=0}^{k-1}||\mathbf{w}^{t-\tau_i} - \mathbf{w}_i^{t-\tau_i,j}||^2 + \eta_l^2 k^2\sigma^2 + (1-\delta)||\eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_i^{t-\tau_i,j},\xi)||^2)\big]
\end{aligned}$$

We derive the above formula according to inequality (13) and assumption 1, 2 respectively. Next, to establish the upper bound of X_3, we employ two lemmas to facilitate our derivation process.

Lemma 1. The difference between the current global model and stale global model.

$$||\mathbf{w}^t - \mathbf{w}^{t-\tau_i}||^2 \leq \tau_i\sum_{h=t-\tau_i}^{t-1}\frac{1}{|\mathbf{S}^h|}\sum_{c\in\mathbf{S}^h}||C(\eta_l\sum_{j=0}^{k-1}\nabla F_i(\mathbf{w}_c^{c-\tau_c,j},\xi)||^2 \tag{17}$$

Proof.

$$\begin{aligned}
||\mathbf{w}^t - \mathbf{w}^{t-\tau_i}||^2 &= ||\sum_{h=t-\tau_i}^{t-1}(\mathbf{w}^{h+1} - \mathbf{w}^h)||^2 \\
&= ||\sum_{h=t-\tau_i}^{t-1}\frac{1}{|\mathbf{S}^h|}\sum_{c\in\mathbf{S}^h}C(\eta_l\sum_{j=0}^{k-1}\nabla F_c(\mathbf{w}_c^{t-\tau_c,j},\xi)||^2
\end{aligned}$$

$$\leq \tau_i \sum_{h=t-\tau_i}^{t-1} \frac{1}{|\mathbf{S}^h|} \sum_{c\in\mathbf{S}^h} ||C(\eta_l \sum_{j=0}^{k-1} \nabla F_i(\mathbf{w}_c^{c-\tau_c,j}, \xi)||^2$$

Lemma 2. The upper bound of statistic local gradient of device i:

$$\mathbb{E}||\nabla F_i(\mathbf{w}, \xi)||^2 \leq 3(\sigma^2 + \zeta_i^2 + G^2) \tag{18}$$

Proof.

$$\begin{aligned}
\mathbb{E}||\nabla F_i(\mathbf{w}, \xi)||^2 &\leq \mathbb{E}||\nabla F_i(\mathbf{w}, \xi) - \nabla F_i(\mathbf{w}) + F_i(\mathbf{w}) - F(\mathbf{w}) + F(\mathbf{w})||^2 \\
&\leq 3(\mathbb{E}||\nabla F_i(\mathbf{w}, \xi) - \nabla F_i(\mathbf{w})||^2 + \mathbb{E}||F_i(\mathbf{w}) - F(\mathbf{w})||^2 + \mathbb{E}||F(\mathbf{w})||^2) \\
&\leq 3(\sigma^2 + \zeta_i^2 + G^2)
\end{aligned}$$

which is similar to **Lemma 1.** in [3]. Next, let's focus on X_3 again.

$$\begin{aligned}
X_3 \leq\ & \mathbb{E}\big[\frac{6}{|\mathbf{S}^t|} \sum_{i\in\mathbf{S}^t} ((1-\eta_l k)^2 ||\nabla F(\mathbf{w}^t)||^2 + L^2 ||\mathbf{w}^t - \mathbf{w}^{t-\tau_i}||^2 \\
&+ \eta_l^2 k^2 \zeta_i^2 + L^2 \eta_l^2 k \sum_{j=0}^{k-1} ||\mathbf{w}^{t-\tau_i} - \mathbf{w}_i^{t-\tau_i,j}||^2 + \eta_l^2 k^2 \sigma^2 + (1-\delta)||\eta_l \sum_{j=0}^{k-1} \nabla F_i(\mathbf{w}_i^{t-\tau_i,j}, \xi)||^2)\big] \\
\leq\ & 6(1-\eta_l k)^2 ||\nabla F(\mathbf{w}^t)||^2 + \mathbb{E}\big[\frac{6}{|\mathbf{S}^t|} \sum_{i\in\mathbf{S}^t} (L^2 \tau_i \sum_{h=t-\tau_i}^{t-1} \frac{1}{|\mathbf{S}^h|} \sum_{c\in\mathbf{S}^h} ||C(\eta_l \sum_{j=0}^{k-1} \nabla F_i(\mathbf{w}_c^{c-\tau_c,j}, \xi)||^2 \\
&+ \eta_l^2 k^2 \zeta_i^2 + L^2 \eta_l^2 k \sum_{j=0}^{k-1} ||\eta_l \sum_{\rho=0}^{j-1} \nabla F(\mathbf{w}_i^{t-\tau_i,\rho}, \xi)||^2 + \eta_l^2 k^2 \sigma^2 + 3(1-\delta)\eta_l^2 k^2 (\sigma^2 + \zeta_i^2 + G^2))\big] \\
\leq\ & 6(1-\eta_l k)^2 ||\nabla F(\mathbf{w}^t)||^2 + 6\eta_l^2 k^2 (\sigma^2 + \zeta^2) + \mathbb{E}\big[\frac{6}{|\mathbf{S}^t|} \sum_{i\in\mathbf{S}^t} (2L^2 \tau_i (2-\delta) \sum_{h=t-\tau_i}^{t-1} \frac{1}{|\mathbf{S}^h|} \sum_{c\in\mathbf{S}^h} ||\eta_l \sum_{j=0}^{k-1} \nabla F_i(\mathbf{w}_c^{c-\tau_c,j}, \xi)||^2 \\
&+ 3L^2 \eta_l^4 k^4 (\sigma^2 + \zeta_i^2 + G^2) + 3(1-\delta)\eta_l^2 k^2 (\sigma^2 + \zeta_i^2 + G^2))\big] \\
\leq\ & 6(1-\eta_l k)^2 ||\nabla F(\mathbf{w}^t)||^2 + 6\eta_l^2 k^2 (\sigma^2 + \zeta^2) + 18L^2 \eta_l^4 k^4 (\sigma^2 + \zeta^2 + G^2) + 18(1-\delta)\eta_l^2 k^2 (\sigma^2 + \zeta^2 + G^2) \\
&+ 36L^2 \eta_l^4 k^4 \tau_{max}^2 (2-\delta)(\sigma^2 + \zeta^2 + G^2)
\end{aligned}$$

where $\tau_{max} = max_{i\in\mathcal{N}}\ \tau_i$ is the max staleness of all devices. Then we can get:

$$\begin{aligned}
\mathbb{E}\big[F(\mathbf{w}^{t+1})\big] \leq\ & F(\mathbf{w}^t) - \frac{\eta_g}{2} ||\nabla F(\mathbf{w}^t)||^2 + \frac{\eta_g}{2} (L\eta_g - 1) \mathbb{E}\big[\frac{1}{|\mathbf{S}^t|} \sum_{i\in\mathbf{S}^t} ||C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big] \\
&+ \frac{\eta_g}{2} \underbrace{\mathbb{E}\big[\frac{1}{|\mathbf{S}^t|} \sum_{i\in\mathbf{S}^t} ||\nabla F(\mathbf{w}^t) - C(\mathbf{w}_i^{t-\tau_i,0} - \mathbf{w}_i^{t-\tau_i,k})||^2\big]}_{X_3} \\
\leq\ & F(\mathbf{w}^t) - \frac{\eta_g}{2} ||\nabla F(\mathbf{w}^t)||^2 + 3\eta_g (L\eta_g - 1)\eta_l^2 k^2 (2-\delta)(\sigma^2 + \zeta^2 + G^2) \\
&+ \frac{\eta_g}{2} \big[6(1-\eta_l k)^2 ||\nabla F(\mathbf{w}^t)||^2 + 6\eta_l^2 k^2 (\sigma^2 + \zeta^2) + 18L^2 \eta_l^4 k^4 (\sigma^2 + \zeta^2 + G^2) + 18(1-\delta)\eta_l^2 k^2 (\sigma^2 + \zeta^2 + G^2) \\
&+ 36L^2 \eta_l^4 k^4 \tau_{max}^2 (2-\delta)(\sigma^2 + \zeta^2 + G^2)\big]
\end{aligned}$$

We choose global learning rate and local learning rate satisfy $\eta_l, \eta_g \leq \frac{1}{L}$. Besides, we suppose $Q_l = 6(1-\eta_l k)^2 - 1 \geq 0$. We arrange this convergence bound:

$$\begin{aligned}
\mathbb{E}\big[F(\mathbf{w}^{t+1})\big] - F(\mathbf{w}^t) \leq\ & -\frac{\eta_g}{2} Q_l ||\nabla F(\mathbf{w}^t)||^2 \\
&+ \frac{\eta_g}{2} \big[6\eta_l^2 k^2 (\sigma^2 + \zeta^2) + 18\eta_l^2 k^4 (\sigma^2 + \zeta^2 + G^2) + 18(1-\delta)\eta_l^2 k^2 (\sigma^2 + \zeta^2 + G^2) \\
&+ 36\eta_l^2 k^4 \tau_{max}^2 (2-\delta)(\sigma^2 + \zeta^2 + G^2)\big]
\end{aligned}$$

To simplify the expression, we assume $B_1 = 18(\sigma^2 + \zeta^2 + G^2)$ and $B_2 = 6(\sigma^2 + \zeta^2)$. Then we sum up t in the above inequality from 0 to $T-1$ and arrange the result, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}||\nabla F(\mathbf{w}^t)||^2 \leq \frac{2[F(\mathbf{w}^0) - F(\mathbf{w}^*)]}{\eta_g Q_l T} + \eta_l^2 k^2 \frac{(1-\delta)B_1 + B_2}{Q_l} + \eta_l^2 k^4 \frac{[2\tau_{max}^2(2-\delta)+1]B_1}{Q_l} \tag{19}$$

To get the convergence rate, we choose $\eta_g = \mathcal{O}(\sqrt{\frac{k}{T}})$ and $\eta_l = \mathcal{O}(T^{-1/4}k^{-5/2}\delta^{-1/2})$. Then we have the convergence rate:

$$\frac{1}{T}\sum_{t=0}^{T-1}||\nabla F(\mathbf{w}^t)||^2 \leq \mathcal{O}(\frac{F^*}{\sqrt{kT}}) + \mathcal{O}(\frac{(1-\delta)B_1 + B_2}{k^3\sqrt{T}\delta}) + \mathcal{O}(\frac{(\tau_{max}^2(2-\delta)+1)B_1}{\sqrt{T}k\delta}) \tag{20}$$

Insight.

Then we focus on the third term and discuss how to get the fastest convergence. We name the third term as *domain term*. To get the fastest convergence, we should minimize the domain term. According to the definition of staleness, we expand τ_{max} as $k\alpha + \delta\beta$, where $\alpha = \alpha_m, \beta = \beta_m, m = arg\ max_i \tau_i$. Then we get the domain term as :

$$\phi(k,\delta) = \frac{(k\alpha + \delta\beta)^2(2-\delta)+1}{k\delta} \tag{21}$$

where $k \in [k_{min}, k_{max}], \delta \in [\delta_{min}, \delta_{max}]$. Then we have the optimization equation:

$$\begin{aligned} min_{k,\delta}\ \ & \phi(k,\delta) \\ s.t.\ & k \in [k_{min}, k_{max}] \\ & \delta \in [\delta_{min}, \delta_{max}] \end{aligned}$$

References

[1] Chen M, Mao B, Ma T. Fedsa: A staleness-aware asynchronous federated learning algorithm with non-iid data[J]. Future Generation Computer Systems, 2021, 120: 1-12.

[2] Stich S U, Cordonnier J B, Jaggi M. Sparsified SGD with memory[J]. Advances in Neural Information Processing Systems, 2018, 31.

[3] Nguyen J, Malik K, Zhan H, et al. Federated learning with buffered asynchronous aggregation[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2022: 3581-3607.

[4] Xu Y, Liao Y, Xu H, et al. Adaptive control of local updating and model compression for efficient federated learning[J]. IEEE Transactions on Mobile Computing, 2022.