



# Appendix - ASYN2F: An Asynchronous Federated Learning Framework with Bidirectional Model Aggregation

Tien-Dung Cao , Nguyen T. Vuong, Thai Q. Le, Hoang V.N. Dao, Tram Truong-Huu  *Senior Member, IEEE*



## 1 MODEL CONVERGENCE ANALYSIS

In this section, we present a proof sketch that outlines the key points to prove the convergence of our aggregation algorithms at the server and training workers. As discussed in Section 3 of the paper, each training worker may perform multiple mini-batch SGD updates before a local aggregation step. The update flow follows this pattern: multiple SGD steps  $\rightarrow$  local aggregation  $\rightarrow$  multiple SGD steps  $\rightarrow$  local aggregation  $\rightarrow \dots$ . If the number of SGD steps significantly outweighs the frequency of local aggregation, the convergence of ASYN2F is largely dictated by that of SGD with non-IID FedAvg. Therefore, we focus primarily on scenarios where local aggregation occurs frequently.

**Setup and Notation.** Consider a fixed set of workers, denoted by  $\mathcal{K}$ , where  $|\mathcal{K}| = K$ . Each worker  $k \in \mathcal{K}$  is assigned a weight  $p_k > 0$  such that  $\sum_{k \in \mathcal{K}} p_k = 1$ . Each worker  $k$  has a local objective function defined as

$$f_k(W) = \mathbb{E}_{\xi \sim D_k} [F_k(W, \xi)],$$

with  $F_k(W, \xi)$  being the loss function used to evaluate the error of the model parameter  $W$  on the data sample  $\xi$ . The global objective is then given by

$$f(W) = \sum_{k \in \mathcal{K}} p_k f_k(W).$$

For any  $W \in \mathbb{R}^d$ , let  $[W]_r$  denote its  $r$ -th coordinate.

**Assumptions.** Similarly to previous model convergence analysis works [1], we make the following assumptions:

A1: *L-Smoothness.* For each worker  $k$ , the local objective  $f_k(W)$  (hence the global objective  $f(W)$ ) is  $L$ -smooth,

meaning for all  $W, W'$ :

$$f(W') \leq f(W) + \langle \nabla f(W), W' - W \rangle + \frac{L}{2} \|W' - W\|^2.$$

A2: *Unbiased Local Stochastic Gradient Estimator.* The mini-batch stochastic gradient used by worker  $k$  at iteration  $j$  satisfies

$$\mathbb{E} \left[ \frac{1}{|B_k^j|} \sum_{\xi \in B_k^j} \nabla F_k(W, \xi) \mid W \right] = \nabla f_k(W),$$

A3: *Unbiased Stochastic Gradient Estimator.* The worker gradients  $\nabla f_k(W_i^{\text{global}})$  are unbiased estimators of the true gradient  $\nabla f(W_i^{\text{global}})$  that is  $\mathbb{E}[\nabla f_k(W)] = \nabla f(W)$ .

A4: *Bounded Variance.* There exists  $\sigma > 0$  such that for all  $W \in \mathbb{R}^d$  and all  $k$ ,

$$\mathbb{E}[\|\nabla f_k(W) - \nabla f(W)\|^2] \leq \sigma^2.$$

A5: *Uniform Boundedness.* There exist constants  $0 \leq \alpha_{\min} \leq \alpha_{k,r}^{\text{local}} \leq \alpha_{\max} \leq 1$  that hold for all coordinate  $r$ , all worker  $k$ , and all updates.

A6: *Dynamic Participation.* At each communication round  $i$ , a random subset  $\mathcal{S}_i \subset \{1, \dots, K\}$  of workers is selected for aggregation.

A7: *Convexity.* The global objective  $f(W)$  is convex, i.e., for all  $W_1, W_2$ ,

$$f(W_1) \geq f(W_2) + \langle \nabla f(W_2), W_1 - W_2 \rangle.$$

A8: *Strong convexity.* The global objective function  $f(W)$  is strongly convex with parameter  $\mu > 0$  if, for all  $W_1, W_2$ , the following inequality holds:

$$f(W_2) \geq f(W_1) + \langle \nabla f(W_1), W_2 - W_1 \rangle + \frac{\mu}{2} \|W_2 - W_1\|^2.$$

**Theorem 1.** Let assumptions A1 to A7 hold and all parameters be defined therein. Let  $T$  be the number of global aggregation steps. Then, the proposed aggregation algorithms at the training workers and the server will converge to a neighborhood of the global optimum, with a convergence rate of  $O(1/\sqrt{T})$ . If A8 holds instead of A7 then the convergence rate is of  $O(1/T)$ .

- T.-D. Cao, N.T. Vuong, T.Q. Le and H.V.N. Dao are with the School of Information Technology, Tan Tao University, Vietnam. E-mail: dung.cao@ttu.edu.vn, nguyen.vuong1902060@std.ttu.edu.vn, thai.le1902097@std.ttu.edu.vn, hoang.dao1902093@std.ttu.edu.vn. T.-D. Cao is the corresponding author.
- T. Truong-Huu is with the Singapore Institute of Technology (SIT), Singapore. He is also with the Institute for Infocomm Research (I<sup>2</sup>R), Agency for Science, Technology and Research (A\*STAR), Singapore. E-mail: truonghuu.tram@singaporetech.edu.sg.

## 2 PROOF OF THEOREM 1

### 2.1 Analysis of a Single Worker's Local Update.

Let  $J_k$  denote the number of local iterations (i.e., batches) that worker  $k$  executes between two global aggregation rounds. We denote the  $j$ -th mini-batch by  $B_k^j$ , where  $|B_k^j|$  is its batch size. The update rule for mini-batch SGD, using a step size  $\eta_{i_k,j}$ , is

$$W_{i_k,j+1}^{\text{local}} \leftarrow W_{i_k,j}^{\text{local}} - \eta_{i_k,j} \frac{1}{|B_k^j|} \sum_{\xi \in B_k^j} \nabla F_k(W_{i_k,j}^{\text{local}}, \xi),$$

for  $j = 0, \dots, J_k - 1$ .

Let  $j^* + 1$  be the batch index at which worker  $k$  performs a local aggregation. The coordinate-wise local aggregation is presented in Algorithm 3 in Section 3 of the paper. Namely,

$$W_{i_k,j^*+1}^{\text{local}} \leftarrow (1 - \alpha_k^{\text{local}}) W_i^{\text{global}} + \alpha_k^{\text{local}} W_{i_k,j^*+1}^{\text{local}},$$

where the merging weight  $\alpha_k^{\text{local}}$  is determined either

$$\alpha_k^{\text{local}} = \beta \frac{Q_k^{\text{data}} \times s_k}{Q_k^{\text{data}} \times s_k + \bar{Q}^{\text{data}} \times S_i} + (1 - \beta) \frac{\bar{L}}{L_{i_k,j^*} + \bar{L}}. \quad (1)$$

or  $\alpha_k^{\text{local}} = 0$ . The case  $\alpha_k^{\text{local}} = 0$ , corresponds to setting  $W_{i_k,j^*+1}^{\text{local}} \leftarrow W_i^{\text{global}}$  in Algorithm 3 (Section 3 in the paper). Here, we assume all the quantities  $Q_k^{\text{data}}$ ,  $s_k$ ,  $\bar{Q}^{\text{data}}$ ,  $S_i$ ,  $\bar{L}$ ,  $L_{i_k,j^*}$  are strictly positive, ensuring that  $\alpha_k^{\text{local}} \in [0, 1]$ , and, in the coordinate-wise setting,  $\alpha_{k,r}^{\text{local}} \in [0, 1], \forall r$ .

For worker  $k$  and coordinate  $r$ , let

$$e_{k,j}^{(r)} \triangleq [W_{i_k,j}^{\text{local}}]_r - [W^*]_r,$$

where  $W^*$  is an optimal solution to  $f(W)$ . We now rewrite the SGD output, when  $j = j^*$ , as

$$[W_{i_k,j^*+1}^{\text{local}}]_r \leftarrow [W_{i_k,j^*}^{\text{local}}]_r - \eta_{i_k,j^*} [g_{k,j^*}]_r,$$

with  $[g_{k,j^*}]_r = \frac{1}{|B_k^{j^*}|} \sum_{\xi \in B_k^{j^*}} [\nabla F_k(W_{i_k,j^*}^{\text{local}}, \xi)]_r$ . Standard SGD analysis (using convexity and  $L$ -smoothness) shows that

$$\begin{aligned} & \mathbb{E} \left[ \left( e_{k,j^*}^{(r)} - \eta_{i_k,j^*} [g_{k,j^*}]_r \right)^2 \right] \\ &= \mathbb{E} \left[ \left( [W_{i_k,j^*}^{\text{local}}]_r - \eta_{i_k,j^*} [g_{k,j^*}]_r - [W^*]_r \right)^2 \right] \\ &\leq (e_{k,j^*}^{(r)})^2 - 2\eta_{i_k,j^*} \Delta_{k,j^*}^{(r)} + \eta_{i_k,j^*}^2 C \end{aligned} \quad (2)$$

where  $\Delta_{k,j^*}^{(r)}$  is the per-coordinate descent gap and  $C > 0$  is a constant. Next, the coordinate-wise local aggregation is

$$[W_{i_k,j^*+1}^{\text{local}}]_r \leftarrow (1 - \alpha_{k,r}^{\text{local}}) [W_i^{\text{global}}]_r + \alpha_{k,r}^{\text{local}} [W_{i_k,j^*+1}^{\text{local}}]_r.$$

Thus, the error satisfies

$$\begin{aligned} e_{k,j^*+1}^{(r)} &= (1 - \alpha_{k,r}^{\text{local}}) ([W_i^{\text{global}}]_r - [W^*]_r) \\ &\quad + \alpha_{k,r}^{\text{local}} (e_{k,j^*}^{(r)} - \eta_{i_k,j^*} [g_{k,j^*}]_r). \end{aligned} \quad (3)$$

By squaring and applying  $(a + b)^2 \leq 2a^2 + 2b^2$ , we have

$$\begin{aligned} \mathbb{E} \left[ (e_{k,j^*+1}^{(r)})^2 \right] &\leq 2(1 - \alpha_{k,r}^{\text{local}})^2 ([W_i^{\text{global}}]_r - [W^*]_r)^2 \\ &\quad + 2(\alpha_{k,r}^{\text{local}})^2 \mathbb{E} \left[ \left( e_{k,j^*}^{(r)} - \eta_{i_k,j^*} [g_{k,j^*}]_r \right)^2 \right]. \end{aligned} \quad (4)$$

Substituting the SGD bound (in Eq. (2)) into Ineq. (4) yields

$$\begin{aligned} \mathbb{E} \left[ (e_{k,j^*+1}^{(r)})^2 \right] &\leq 2(1 - \alpha_{k,r}^{\text{local}})^2 ([W_i^{\text{global}}]_r - [W^*]_r)^2 \\ &\quad + 2(\alpha_{k,r}^{\text{local}})^2 \left[ (e_{k,j^*}^{(r)})^2 - 2\eta_{i_k,j^*} \Delta_{k,j^*}^{(r)} + \eta_{i_k,j^*}^2 C \right]. \end{aligned}$$

Summing over all coordinates  $r = 1, \dots, d$ :

$$\begin{aligned} \sum_{r=1}^d \mathbb{E} \left[ (e_{k,j^*+1}^{(r)})^2 \right] &\leq \sum_{r=1}^d 2(1 - \alpha_{k,r}^{\text{local}})^2 ([W_i^{\text{global}}]_r - [W^*]_r)^2 \\ &\quad + \sum_{r=1}^d 2(\alpha_{k,r}^{\text{local}})^2 \left[ (e_{k,j^*}^{(r)})^2 - 2\eta_{i_k,j^*} \Delta_{k,j^*}^{(r)} + \eta_{i_k,j^*}^2 C \right]. \end{aligned} \quad (5)$$

Since the squared norm satisfies

$$\mathbb{E} \|W_{i_k,j}^{\text{local}} - W^*\|^2 = \sum_{r=1}^d \mathbb{E} [(e_{k,j}^{(r)})^2],$$

we can rewrite:

$$\begin{aligned} \mathbb{E} \|W_{i_k,j^*+1}^{\text{local}} - W^*\|^2 &\leq 2(1 - \alpha_{\min})^2 \|W_i^{\text{global}} - W^*\|^2 \\ &\quad + 2(\alpha_{\max})^2 \mathbb{E} \|W_{i_k,j^*}^{\text{local}} - W^*\|^2 - 4(\alpha_{\min})^2 \eta_{i_k,j^*} \sum_{r=1}^d \Delta_{k,j^*}^{(r)} \\ &\quad + 2(\alpha_{\max})^2 d \eta_{i_k,j^*}^2 C. \end{aligned} \quad (6)$$

Now telescoping over  $J_k$  local iterations leads to a bound of the form

$$\begin{aligned} \mathbb{E} \|W_{i_k,J_k}^{\text{local}} - W^*\|^2 &\leq \|W_{i_k,0}^{\text{local}} - W^*\|^2 - 2\eta_{\min} \alpha_{\min}^2 \sum_{j=0}^{J_k-1} \Delta_{k,j} \\ &\quad + O(J_k \alpha_{\max}^2 \eta_{\min}^2) + O\left(J_k (1 - \alpha_{\min})^2 \|W_i^{\text{global}} - W^*\|^2\right), \end{aligned} \quad (7)$$

where  $\eta_{\min} = \min_j \eta_{i_k,j}$ . Ineq. (7) indicates that the distance  $\|W_{i_k,j}^{\text{local}} - W^*\|$  contracts, with the contraction also depending on that of  $\|W_i^{\text{global}} - W^*\|$ .

### 2.2 Global Aggregation with Partially Dynamic Worker Participation

At a communication (global aggregation) round  $i + 1$ , a random subset  $\mathcal{S}_i$  of workers participates. The server aggregates via (non-IID) FedAvg as in Algorithm 2 (Section 3 of the paper):

$$W_{i+1}^{\text{global}} = \sum_{k \in \mathcal{S}_i} \alpha_k W_{i_k,J_k}^{\text{local}},$$

where  $\alpha_k$  is computed and normalized as in Algorithm 2 (Section 3 of the paper) (hence  $\sum_{k \in \mathcal{S}_i} \alpha_k = 1$ ). By convexity of the squared norm,

$$\|W_{i+1}^{\text{global}} - W^*\|^2 \leq \sum_{k \in \mathcal{S}_i} \alpha_k \|W_{i_k,J_k}^{\text{local}} - W^*\|^2. \quad (8)$$

Ineq. (8) shows that the distance  $\|W_{i+1}^{\text{global}} - W^*\|$  contracts as a function of the contraction in  $\|W_{i_k,J_k}^{\text{local}} - W^*\|$ . By combining Ineq. (7) and Ineq. (8), we conclude that  $W_{i_k,j}^{\text{local}}$  converges to  $W^*$  if and only if  $W_{i+1}^{\text{global}}$  does.

### 2.3 Convergence Analysis

We now prove that  $f(W_{i+1}^{\text{global}})$  converges (in expectation) to  $f(W^*)$  by applying a standard FedAvg analysis on a

non-IID data setting (see [1]) together with the above local progress result. Specifically, we can show that

$$\frac{1}{T} \sum_{i=0}^{T-1} \mathbb{E}[\|\nabla f(W_i^{\text{global}})\|^2] \leq \frac{f(W_0^{\text{global}}) - f(W^*)}{T(\eta J^{\text{eff}} - \frac{L}{2}\eta^2(J^{\text{eff}})^2)} + \frac{L(\eta J^{\text{eff}}\sigma^2)}{2(\eta J^{\text{eff}} - \frac{L}{2}\eta^2(J^{\text{eff}})^2)} \quad (9)$$

where  $J^{\text{eff}} = \sum_k \alpha_k J_k^{\text{eff}}$ ,  $J_k^{\text{eff}} = J_k \cdot df$ , (here,  $df$  is a damping factor influenced by  $\alpha_k^{\text{local}}$  and the frequency of local aggregations). We should choose  $\eta \leq \frac{2}{L J^{\text{eff}}}$ . Letting  $T \rightarrow \infty$  yields convergence (in expectation) to a neighborhood of the global optimum, with a typical convergence rate of  $O(1/\sqrt{T})$  in the convex setting (or linear convergence under strong convexity). The detail is as follows.

Because workers update their local models through a mix of multiple mini-batch SGD and local aggregations before sending the final result to the server for a global update, we need to account for these additional local aggregation steps. Over  $J_k$  SGD batches within a full local training round (which includes multiple local aggregations), we approximate the cumulative update of a client  $k$  as:

$$W_{i_k, J_k}^{\text{local}} = W_i^{\text{global}} - \eta_k \sum_{j=0}^{J_k-1} g_{i_k, j},$$

where the gradients  $g_{i_k, j} = \frac{1}{|B_k^j|} \sum_{\xi \in B_k^j} \nabla F_k(W_{i_k, j}^{\text{local}}, \xi)$ . According to the assumption (A2),

$$\mathbb{E}[g_{i_k, j}] = \nabla f_k(W_i^{\text{global}}).$$

After incorporating multiple local aggregations, the *effective gradient update* behaves like a dampened version of a standard local SGD step, leading to:

$$W_{i_k, J_k}^{\text{local}} \approx W_i^{\text{global}} - \eta_k J_k^{\text{eff}} \nabla f_k(W_i^{\text{global}}),$$

where  $J_k^{\text{eff}}$  accounts for the impact of local aggregation steps. Typically,  $J_k^{\text{eff}} \leq J_k$ . This is because merging the local model in the global model via local aggregation reduces the *effective number of SGD steps*, denoted by  $J_k^{\text{eff}}$ , since part of the learning gets overwritten by the global model. Mathematically, we can model this effect as:

$$J_k^{\text{eff}} = J_k \cdot df,$$

where the *damping factor*  $df$  is influenced by  $\alpha_k^{\text{local}}$  and the frequency of local aggregations. Since local aggregation pulls the model back towards  $W_i^{\text{global}}$ , the total drift of the local model is smaller compared to pure SGD.

After workers complete their multiple rounds of SGD and local aggregation, they send the final models to the server, which performs a weighted aggregation:

$$W_{i+1}^{\text{global}} = \sum_{k \in \mathcal{S}_i} \alpha_k W_{i_k, J_k}^{\text{local}}, \text{ with } \sum_{k \in \mathcal{S}_i} \alpha_k = 1.$$

Substituting our previous approximation:

$$W_{i+1}^{\text{global}} = \sum_{k \in \mathcal{S}_i} \alpha_k \left( W_i^{\text{global}} - \eta_k J_k^{\text{eff}} \nabla f_k(W_i^{\text{global}}) \right).$$

Expanding the sum:

$$W_{i+1}^{\text{global}} = W_i^{\text{global}} - \sum_{k \in \mathcal{S}_i} \alpha_k \eta_k J_k^{\text{eff}} \nabla f_k(W_i^{\text{global}}).$$

By defining  $g_i$  as

$$g_i = \sum_{k \in \mathcal{S}_i} \alpha_k J_k^{\text{eff}} \nabla f_k(W_i^{\text{global}})$$

and choosing  $\eta$  as a weighted average of the worker learning rates  $\eta_k$ , e.g.,  $\eta = \sum_{k \in \mathcal{K}} \alpha_k \eta_k$ , we approximate the final global update:

$$W_{i+1}^{\text{global}} = W_i^{\text{global}} - \eta g_i.$$

To prove that  $f(W_i^{\text{global}})$  converges (in expectation) to the optimal function value  $f(W^*)$ , we need to analyze the expected decrease in the function value at each global round and show that it diminishes over time. Using the *L-smoothness* of  $f(W)$  (the assumption (A1)), we have:

$$f(W_{i+1}^{\text{global}}) \leq f(W_i^{\text{global}}) - \eta \langle \nabla f(W_i^{\text{global}}), g_i \rangle + \frac{L}{2} \eta^2 \|g_i\|^2.$$

Taking expectation on both sides:

$$\begin{aligned} \mathbb{E}[f(W_{i+1}^{\text{global}})] &\leq \mathbb{E}[f(W_i^{\text{global}})] - \eta \mathbb{E}[\langle \nabla f(W_i^{\text{global}}), g_i \rangle] \\ &\quad + \frac{L}{2} \eta^2 \mathbb{E}[\|g_i\|^2]. \end{aligned} \quad (10)$$

The assumption (A3) gives

$$\mathbb{E}[\nabla f_k(W_i^{\text{global}})] = \nabla f(W_i^{\text{global}}).$$

So, we get

$$\mathbb{E}[g_i] = \sum_k \alpha_k J_k^{\text{eff}} \nabla f(W_i^{\text{global}}).$$

Now, defining the *effective number of local steps* as:

$$J^{\text{eff}} = \sum_k \alpha_k J_k^{\text{eff}},$$

we obtain

$$\mathbb{E}[g_i] = J^{\text{eff}} \nabla f(W_i^{\text{global}}),$$

and

$$\begin{aligned} \mathbb{E}[\langle \nabla f(W_i^{\text{global}}), g_i \rangle] &= \langle \nabla f(W_i^{\text{global}}), \mathbb{E}[g_i] \rangle \\ &= \langle \nabla f(W_i^{\text{global}}), J^{\text{eff}} \nabla f(W_i^{\text{global}}) \rangle \\ &= J^{\text{eff}} \|\nabla f(W_i^{\text{global}})\|^2. \end{aligned} \quad (11)$$

We now bound  $\mathbb{E}[\|g_i\|^2]$ . Using standard variance decomposition

$$\mathbb{E}[\|g_i\|^2] = \mathbb{E}[\|\mathbb{E}[g_i] + (g_i - \mathbb{E}[g_i])\|^2].$$

Expanding this using bias-variance decomposition

$$\mathbb{E}[\|g_i\|^2] = \|\mathbb{E}[g_i]\|^2 + \mathbb{E}[\|g_i - \mathbb{E}[g_i]\|^2]. \quad (12)$$

Since  $\mathbb{E}[g_i] = J^{\text{eff}} \nabla f(W_i^{\text{global}})$ , we obtain

$$\|\mathbb{E}[g_i]\|^2 = (J^{\text{eff}})^2 \|\nabla f(W_i^{\text{global}})\|^2. \quad (13)$$

Additionally, the assumption (A4) gives

$$\mathbb{E}[\|\nabla f_k(W_i^{\text{global}}) - \nabla f(W_i^{\text{global}})\|^2] \leq \sigma^2,$$

hence

$$\mathbb{E} [\|g_i - \mathbb{E}[g_i]\|^2] = \mathbb{E} \left[ \left\| \sum_{k \in \mathcal{S}_i} \alpha_k J_k^{\text{eff}} \nabla f_k(W_i^{\text{global}}) - J^{\text{eff}} \nabla f(W_i^{\text{global}}) \right\|^2 \right] \leq (J^{\text{eff}})^2 \sigma^2. \quad (14)$$

Eqs (12), (13) and (14) lead to

$$\mathbb{E}[\|g_i\|^2] \leq (J^{\text{eff}})^2 \mathbb{E}[\|\nabla f(W_i^{\text{global}})\|^2] + (J^{\text{eff}})^2 \sigma^2. \quad (15)$$

Plugging Eq. (11) and Eq. (15) into Ineq. (10):

$$\begin{aligned} \mathbb{E}[f(W_{i+1}^{\text{global}})] &\leq \mathbb{E}[f(W_i^{\text{global}})] - \eta J^{\text{eff}} \|\nabla f(W_i^{\text{global}})\|^2 \\ &\quad + \frac{L}{2} \eta^2 (J^{\text{eff}})^2 \mathbb{E}[\|\nabla f(W_i^{\text{global}})\|^2] \\ &\quad + \frac{L}{2} \eta^2 (J^{\text{eff}})^2 \sigma^2. \end{aligned} \quad (16)$$

Summing Ineq. (16) over  $i = 0, \dots, T-1$  implies

$$\begin{aligned} \mathbb{E}[f(W_T^{\text{global}})] &\leq f(W_0^{\text{global}}) - \eta J^{\text{eff}} \sum_{i=0}^{T-1} \|\nabla f(W_i^{\text{global}})\|^2 \\ &\quad + \frac{L}{2} \eta^2 (J^{\text{eff}})^2 \sum_{i=0}^{T-1} \mathbb{E}[\|\nabla f(W_i^{\text{global}})\|^2] \\ &\quad + \frac{L}{2} \eta^2 T (J^{\text{eff}})^2 \sigma^2. \end{aligned} \quad (17)$$

Noting that

$$f(W_0^{\text{global}}) - \mathbb{E}[f(W_T^{\text{global}})] \leq f(W_0^{\text{global}}) - f(W^*),$$

we now rearrange Ineq. (17) to bound the average gradient norm as follows. First,

$$\begin{aligned} &(\eta J^{\text{eff}} - \frac{L}{2} \eta^2 (J^{\text{eff}})^2) \sum_{i=0}^{T-1} \mathbb{E}[\|\nabla f(W_i^{\text{global}})\|^2] \\ &\leq f(W_0^{\text{global}}) - f(W^*) + \frac{L}{2} \eta^2 T (J^{\text{eff}})^2 \sigma^2. \end{aligned}$$

We then get

$$\begin{aligned} \frac{1}{T} \sum_{i=0}^{T-1} \mathbb{E}[\|\nabla f(W_i^{\text{global}})\|^2] &\leq \frac{f(W_0^{\text{global}}) - f(W^*)}{T \eta J^{\text{eff}} (1 - \frac{L}{2} \eta J^{\text{eff}})} \\ &\quad + \frac{L J^{\text{eff}} \eta \sigma^2}{2(1 - \frac{L}{2} \eta J^{\text{eff}})}. \end{aligned}$$

Choosing  $\eta$  sufficiently small such that  $\eta \leq \frac{2}{L J^{\text{eff}}}$ , and letting  $T \rightarrow \infty$  leads to

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{i=0}^{T-1} \mathbb{E}[\|\nabla f(W_i^{\text{global}})\|^2] = \frac{L J^{\text{eff}} \eta \sigma^2}{2(1 - \frac{L}{2} \eta J^{\text{eff}})}. \quad (18)$$

In Eq. (18), if  $\sigma^2 = 0$ , the above limit is zero, and the average gradient norm goes to zero, implying that the algorithm converges exactly to a stationary point  $W^*$ . If  $\sigma^2 > 0$ , the limit is close to 0, then the algorithm converges only to a neighborhood of a stationary point  $W^*$ . The size of this neighborhood is proportional to  $\eta$ ,  $J^{\text{eff}}$ ,  $L$  and  $\sigma^2$ . We can use a diminishing step size  $\eta$  to further drive this term down as  $T$  increases.

It is well-known that under the convexity assumption (A7),  $W^*$  is also a global minimum and the convergence rate is  $O(1/\sqrt{T})$ . While under the strong convexity assumption (A8),  $W^*$  is also a global minimum. However, the convergence rate is  $O(1/T)$ .

## REFERENCES

- [1] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the Convergence of FedAvg on Non-IID Data," 2020. [Online]. Available: <https://arxiv.org/abs/1907.02189>