A Why FL Gradients Can Be Larger Than CL?

Let $D = \{(x_i, y_i)\}_{i=1}^N$ be the set of training data points. Let $\mathbf{w}$ be the model weights. Let l be the loss function (e.g., cross-entropy loss).

A.1 Centralized Learning

Our optimization goal in CL is the following:

$$\min_{\mathbf{w}} \frac{1}{N} \cdot \sum_{i=1}^{N} l(x_i, y_i; \mathbf{w}).$$

If we run gradient descent (GD) to solve this optimization problem, the update step at the epoch t is

$$\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \eta \cdot \mathbf{g}^t,$$

where η is the learning rate and $\mathbf{g}$ is the gradient defined as

$$\mathbf{g}^t = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{\partial l(x_i, y_i; \mathbf{w}^{t-1})}{\partial \mathbf{w}^{t-1}}.$$

We are interested in the variance of the gradient $\mathbf{g}$, namely

$$[\mathbf{g}] = [\frac{1}{N} \cdot \sum_{i=1}^{N} \frac{\partial l(x_i, y_i; \mathbf{w})}{\partial \mathbf{w}}],$$

where we have dropped the epoch t from our notation to avoid clutter. Now define the random variable $Z_i = \frac{\partial l(x_i, y_i; \mathbf{w})}{\partial \mathbf{w}}$ for $i \in [N]$. Since the data points $\{(x_i, y_i)\}_{i=1}^N$ are i.i.d. samples, the random variables $\{Z_i\}_{i=1}^N$ are i.i.d. too. As a result, we can let $[Z_i] = \sigma^2$ be some constant for all $i \in [N]$. It then follows that

$$[\mathbf{g}] = [\frac{1}{N} \cdot \sum_{i=1}^{N} Z_i] = \frac{1}{N^2} \cdot \sum_{i=1}^{N} [Z_i] = \frac{\sigma^2}{N}.$$

A.2 Federated Learning

Suppose that we have K clients participating in FL training. Each client $k \in [K]$ owns a local training dataset $D_k = \{(x_{kj}, y_{kj})\}_{j=1}^{n_k}$ with size $n_k = |D_k|$. We have $N = \sum_{k=1}^K n_k$ as $D = \cup\{D_k\}_{k=1}^K$. Our optimization goal in FL then becomes

$$\min_{\mathbf{w}} \sum_{k=1}^{K} \frac{n_k}{N} \cdot L_k(\mathbf{w}),$$

where we define

$$L_k(\mathbf{w}) = \frac{1}{n_k} \cdot \sum_{j=1}^{n_k} l(x_{kj}, y_{kj}; \mathbf{w}).$$

Consider a client $k \in [K]$ and its local update step at the epoch t, which can be written as

$$\mathbf{w}_k^t \leftarrow \mathbf{w}_k^{t-1} - \eta \cdot \mathbf{g}_k^t,$$

where η is the same learning rate as in CL and $\mathbf{g}_k^t$ is the local gradient defined as

$$\mathbf{g}_k^t = \frac{1}{n_k} \cdot \sum_{j=1}^{n_k} \frac{\partial l(x_{kj}, y_{kj}; \mathbf{w}_k^{t-1})}{\partial \mathbf{w}_k^{t-1}}.$$

Similarly, we drop the epoch t in our notation to avoid clutter and study the variance of $\mathbf{g}_k$:

$$[\mathbf{g}_k] = [\frac{1}{n_k} \cdot \sum_{j=1}^{n_k} \frac{\partial l(x_{kj}, y_{kj}; \mathbf{w}_k)}{\partial \mathbf{w}_k}].$$

Now assume a simplified case where data is distributed to the clients in an i.i.d. manner. Under this assumption, we can similarly define the i.i.d. random variables $Z_{kj} = \frac{\partial l(x_{kj}, y_{kj}; \mathbf{w}_k)}{\partial \mathbf{w}_k}$ for $j \in [n_k]$. As a result, it follows that

$$[\mathbf{g}_k] = [\frac{1}{n_k} \cdot \sum_{j=1}^{n_k} Z_{kj}] = \frac{1}{n_k^2} \cdot \sum_{j=1}^{n_k} [Z_{kj}] = \frac{\sigma_k^2}{n_k},$$

where $[Z_{kj}] = \sigma_k^2$ is some constant due to the i.i.d. assumption.

We next consider the global model update step in FedAvg with model averaging, namely,

$$\mathbf{w}^t = \sum_{k=1}^{K} \frac{n_k}{N} \cdot \mathbf{w}_k^t = \sum_{k=1}^{K} \frac{n_k}{N} \cdot \left(\mathbf{w}_k^{t-1} - \eta \cdot \mathbf{g}_k^t\right).$$

Since

$$\mathbf{w}^{t-1} = \sum_{k=1}^{K} \frac{n_k}{N} \cdot \mathbf{w}_k^{t-1},$$

it follows that

$$\mathbf{w}^t = \mathbf{w}^{t-1} - \eta \cdot \sum_{k=1}^{K} \frac{n_k}{N} \cdot \mathbf{g}_k^t.$$

Comparing this to the CL update step, we can define (by dropping the epoch t to avoid clutter)

$$\mathbf{g}' = \sum_{k=1}^{K} \frac{n_k}{N} \cdot \mathbf{g}_k$$

and compute the variance of $\mathbf{g}'$, which results in

$$[\mathbf{g}'] = [\sum_{k=1}^{K} \frac{n_k}{N} \cdot \mathbf{g}_k].$$

A.2.1 (Optimistic) Case Study. If we assume $n_k = \frac{N}{K}$ for $k \in [K]$ (i.e., uniform distribution of data volume) and $\sigma_k^2 = \sigma^2$ (which is unlikely in practice), we have $[\mathbf{g}_k] = \frac{\sigma_k^2}{n_k} = \frac{K}{N} \cdot \sigma^2$ and thus

$$\begin{aligned}
[\mathbf{g}'] &= [\frac{1}{K} \cdot \sum_{k=1}^{K} \mathbf{g}_k] \\
&= \frac{1}{K^2}\Big(\sum_{k=1}^{K} [\mathbf{g}_k] + \sum_{i} \sum_{j,j\neq i} (\mathbf{g}_i, \mathbf{g}_j) \Big) \\
&= \frac{\sigma^2}{N} + \frac{1}{K^2} \cdot \sum_{i} \sum_{j,j\neq i} (\mathbf{g}_i, \mathbf{g}_j) \\
&= [\mathbf{g}] + \frac{1}{K^2} \cdot \sum_{i\in[K]} \sum_{j\in[K],j\neq i} (\mathbf{g}_i, \mathbf{g}_j).
\end{aligned}$$

Using the Cauchy–Schwarz inequality, we have

$$|(\mathbf{g}_i, \mathbf{g}_j)| \leq \sqrt{[\mathbf{g}_i] \cdot [\mathbf{g}_j]} = \frac{K}{N} \cdot \sigma^2.$$

As a result,

$$[\mathbf{g}'] \leq [\mathbf{g}] + \frac{K(K-1)}{K^2} \cdot \frac{K}{N}\sigma^2 = [\mathbf{g}] + \frac{K-1}{N} \cdot \sigma^2.$$

Therefore, in the worst situation, $[\mathbf{g}']$ can be larger than $[\mathbf{g}]$ by $\frac{K-1}{N} \cdot \sigma^2$ even for this highly simplified case study. That is, the relative ratio between $[\mathbf{g}']$ and $[\mathbf{g}]$ can be as large as

$$\frac{[\mathbf{g}']}{[\mathbf{g}]} \leq \frac{[\mathbf{g}] + \frac{K-1}{N} \cdot \sigma^2}{[\mathbf{g}]} = \frac{\frac{\sigma^2}{N} + \frac{K-1}{N} \cdot \sigma^2}{\frac{\sigma^2}{N}} = K,$$

which is proportional to the number of clients K. This implies a relative ratio between the standard deviations of $\mathbf{g}'$ and $\mathbf{g}$ as

$$\frac{S[\mathbf{g}']}{S[\mathbf{g}]} \leq \sqrt{K}.$$

Hence, the magnitude of $\mathbf{g}'$ (in FL) can be much larger than that of $\mathbf{g}$ (in CL) due to higher chance of large variation (i.e., larger variance or standard deviation), especially with larger number of clients.