A Proof of Proposition 1

Proof. Assume $\mathcal{C}$ contains M clients. A client $c_i \in \mathcal{C}$ has N_i examples locally. With FEDSGD, a client c_i only train model for a single step for each training round, with a batch size of N_i. Since the model is trained with a batch insensitive loss ℓ_{BI}, with Equation (1), we derive that the local model gradient of client c_i at the end of training round k is

$$\begin{aligned} \nabla \ell_{C_i} &= \nabla \frac{1}{N_i} \sum_{j=0}^{N_i} \ell_{BI}(f(\mathbf{X}_j), g(\mathbf{Y}_j)) \\ &= \frac{1}{N_i} \sum_{j=0}^{N-i} \nabla \ell_{BI}(f(\mathbf{X}_j), g(\mathbf{Y}_j)) \end{aligned} \tag{3}$$

The local model gradient of all the clients are aggregated to update the server model. Therefore, the server model update at step k is

$$\begin{aligned} \Delta_{k,fedsgd}(\mathcal{C}|\ell_{BI}, \Theta) &= \eta_{s,fedsgd} \cdot \frac{\sum_{i=0}^{M} \nabla \ell_{C_i}}{\sum_{i=0}^{M} N_i} \\ &= \eta_{s,fedsgd} \cdot \frac{\sum_{i=0}^{M} \sum_{j=0}^{N_i} \nabla \ell_{BI}(f(\mathbf{X}_j), g(\mathbf{Y}_j))}{\sum_{i=0}^{M} N_i} \end{aligned} \tag{4}$$

Let $N_{\mathcal{E}}$ be the number of total examples in $\mathcal{E}$, we have $N_{\mathcal{E}} = \sum_{i=0}^{M} N_i$. Then the server model update becomes

$$\Delta_{k,fedsgd}(\mathcal{C}|\ell_{BI}, \Theta) = \eta_{s,fedsgd} \cdot \frac{1}{N_{\mathcal{E}}} \sum_{i=0}^{N_{\mathcal{E}}} \nabla \ell_{BI}(f(\mathbf{X}_i), g(\mathbf{Y}_i)) \tag{5}$$

For centralized training with SGD with all examples in $\mathcal{E}$ in a batch, the model update at step k is

$$\begin{aligned} \Delta_{k,sgd}(\mathcal{E}|\ell_{BI}, \Theta) &= \eta_{s,sgd} \cdot \nabla \frac{1}{N_{\mathcal{E}}} \sum_{i=0}^{N_{\mathcal{E}}} \ell_{BI}(f(\mathbf{X}_j), g(\mathbf{Y}_j)) \\ &= \eta_{s,sgd} \cdot \frac{1}{N_E} \sum_{i=0}^{N_{\mathcal{E}}} \nabla \ell_{BI}(f(\mathbf{X}_j), g(\mathbf{Y}_j)) \end{aligned} \tag{6}$$

Note that $\eta_{s,fedsgd} = \eta_{s,sgd}$. Therefore, with Equation (5) and Equation (6), we prove that

$$\Delta_{k,fedsgd}(\mathcal{C}|\ell_{BI}, \Theta) \equiv \Delta_{k,sgd}(\mathcal{E}|\ell_{BI}, \Theta)$$

□