Decentralized Learning via Clustering and Diverse Subgroup

Shizhou Xu

December 2020

1 Introduction

Federated learning was first proposed by McMahan in 2017 [1] as a solution to both the privacy concern and the computational limit of centralized data analysis. In particular, federated learning train a machine learning model by first sharing a the current model parameter with agents with their own local data to compute the parameter gradients on the agents' local devices, and then collecting the resulting gradients to update the shared model parameter, and finally iterating the procedure until the training termination condition satisfied. Here, we emphasize two importance of the concept of federated learning:

1 Federated learning protects agent's privacy because the central server does not have access to any local information, except for the parameter gradients.

2 The decentralized training framework perform training on the agents' devices and therefore save the cost of building powerful work stations in centralized training.

However, the recent works on information leakage via gradients, [3, 4], provide us numerical evidence that the idea of sharing gradients does not protect the local data in practice. Particularly, [3] borrows the concept of reinforcement learning to give feedback to the dummy data structure by comparing the generated dummy gradient to the leaked true gradient at each of the update. [4] gives theoretical guarantee for the leakage of the true label of local data. Thereby, the adversary is able to retrieve at least some information of the local data from the shared gradients.

On the other hand, although the data structure retrieval via gradient leakage is not theoretically guaranteed, the information extraction from gradients is still useful to both adversary and the central server. For instance, recently works on federated cluster learning, such as [6], tend to cluster agents with similar

gradients together and then perform more personalized learning on each of the clusters. Also, the cluster of agents with similar data structure also help sampling procedure to reduce computational cost. For example, corrected clustered agents give the central server a distribution of data and hence provides a more efficient way to generated approximately uniform samples while keep the sample size small.

Two examples to illustrate the above advantages of correctly clustered agents/local data are the followings:

1 In linear regression, if there are two clusters on both of the training and test data structure, generating two linear regression models would give better predictions.

2 In logistic regression model, the non-uniform sample would result in a poor categorical model.

Therefore, for both personalized model and sample diversification purpose, it is important to have agents clustering in federated or more broadly distributed learning. However, the two clustering methods are the opposite to each other and it would be computationally expansive to perform the two independently for each of the two purposes.

As a result, the goal of the project is the following:

- Study the information retrieval from gradient sharing. In particular, we try to answer if a good clustering of the shared gradients implies a good labeling of the local data.
- Study the diverse subgroup (clustering) problems via transfer learning: generate the diverse subgroup (clustering) on data set by randomly selecting the data points via a well-trained clustering (diverse subgroup).

In short, clustering and diverse subgroup provide solution to the non-i.i.d. local data distributions in distributed learning. The rest of the report is organized as the following: section 2 gives one of the motivations of gradient clustering from the adversarial perspective, section 3 develops the relationship between the clustering problem and the most diverse subgroups problem on the same data set.

The goal is to provide theoretical analysis for information leakage of sharing gradients in distributed learning and thereby provide good solution to prevent the information leakage. Specifically, if we develop quantitative description of the relationship between gradients and underlying data distribution in section 3 and achieve good transfer learning performance between clustering and diverse subgroups, then the combination of the two provides us a powerful quantitative description of clustering and diversity of local data distributions. That is, the

central server obtain the some level of 'distribution' information among the agents. Once one extract the maximum amount of distribution information from shared gradients, efficient defense strategy follows naturally.

In this report, the notation will follow the standard optimization, machine learning, and probability textbook. We denote the set $\{1, 2, ..., N\} = [N]$.

2 Gradient Information Leakage/Extraction

In this section, we first use improved deep leakage from gradient (iDLG) [4] as an example of theoretically guaranteed gradient information leakage, and then show how to use random mask to defend leakage of true label in the iDLG case. Thereafter, we extend the iDLG case and formulate the gradient clustering problem for distributed learning and show that there no defend strategy as we show in the iDLG case in general gradient clustering setting when the number of agents is large. The ultimate goal is to explore the relationship between gradient and the underlying data structure and thereby give theoretical guarantee to local data leakage via sharing gradients.

2.1 Label Leakage via Gradient

In [4], Zhao et al. showed the following: given a neural network model provided by the central server and a set of local sample data (x, k^*), each agent has his/her own loss function:

$$l(x, k^*) = -\log(\frac{e^{y_{k^*}}}{\sum_k e^{y_k}}).$$

Here, k^* is the true label of the agent's local data set. Therefore, we have

$$\frac{\partial l(x, k^*)}{\partial y_i} = -\frac{\partial \log e^{y_{k^*}} - \partial \log(\sum_k e^{y_k})}{\partial y_i} = \begin{cases} \frac{e^{y_i}}{\sum_k e^{y_k}} - 1, & \text{if } i = k^* \\ \frac{e^{y_i}}{\sum_k e^{y_k}}, & \text{otherwise} \end{cases}.$$

Notice that $\{\frac{\partial l(x,k^*)}{\partial y_i}\}_{i\in[k]}$ is a vector that has the one and only negative entry at the true label index k^*. Finally, by back propagation, we obtain

$$\begin{aligned} \nabla(W_o)_i &= \frac{\partial l(x, k^*)}{\partial y_i} \frac{\partial y_i}{\partial (W_o)_i} \\ &= \frac{\partial l(x, k^*)}{\partial y_i} \frac{\partial (W_o)_i^T u_{o-1} + b_i}{\partial (W_o)_i} \\ &= \frac{\partial l(x, k^*)}{\partial y_i} u_{o-1} \end{aligned}$$

Therefore, $\nabla W_o \in \mathbb{R}^{k \times N_{o-1}}$ is a rank 1 matrix with rows scaled differently. Since we also showed that the constant row scalars $\{\frac{\partial l(x,k^*)}{\partial y_i}\}_i$ all positive entries

except for the true-label index entry, by comparing the sign of rows of ∇W_o, one would be able to determine the true label of the agent's local data.

2.2 Random Mask

It is clear that the regular differential privacy is useless in this case. In fact, without any knowledge of the size range of ∇W, one could show that adding random noise does not stop the revealing with high probability. Moreover, since the training iterates many times, it would be likely for the adversary to have a finite but long sequence of such gradient and thereby reveal the true label by law of large number.

In order to defend such information leakage from the sharing gradient, which is necessary for distributed learning, it would be efficient to use random mask to report at most two entries in each column of ∇W. The argument is simple: since the revealing of true label depends on the comparison among the sigh of entries on the same column, if a random mask gives at most two entries on each column, there is not enough reference entries to differentiate true label index from the others.

The label leakage via gradient is merely a specific case of general data labeling. As we mentioned in the motivation part, central servers and adversaries could both benefit from general data labeling, no matter what the labels are. Therefore, the generalization of label leakage via gradient needs a more careful study. To that end, we start with developing transfer learning of diverse subgroups from a well-clustered set.

3 Clustering and Diverse Subgroups

In this section, we first formulate the quantitative relationship between clustering and the diverse subgroup problem, from which we hope to develop a theoretical guarantee to the transfer learning performance from well-trained clusters to diverse subgroups via random selection. Particularly, we give a probabilistic formulation of K-means and also the relationship between spectral clustering and ratio cuts, then develop the corresponding diverse subgroup problems, and give the relationship in between. Moreover, we hope to develop a quantitative measure between the optimal solution to the diverse subgroup problem and the suboptimal solution via transfer learning.

Before the main results, we first give a brief review to the two state-of-the-art clustering methods: K-means and spectral clustering, on a given data set $\{x_i\}_{i\in[N]}$.

1 K-means: Here, we don't differentiate the partition $P = \{p_i\}_{i\in[k]}$ and the corresponding label map $P(i) = j \iff x_i \in p_j$.

$$\min_{P} \sum_{j\in[k]} \sum_{i\in[N]} ||x_i - \frac{1}{\sum_{i\in[N]} \mathbb{1}_{P(i)=j}} \sum_{i\in[N]} x_i \mathbb{1}_{P(i)=j}||_{l^2}^2 \mathbb{1}_{P(i)=j}$$
$$\text{s.t. } P : [N] \longrightarrow [k] \text{ be surjective}$$

2 Spectral Clustering (minimum ratio-cut relaxation): we have to first find the similarity matrix $W_{i,j} := d(x_i, x_j)$ to form the similarity graph via some kernel $d : \{x_i\}_{i\in[N]} \times \{x_i\}_{i\in[N]} \to \mathbb{R}_+$, then we generate the graph Laplacian $L : D - W$, where $D := diag(\{\sum_j W_{i,j}\}_i)$. The minimum ratio-cut problem can be formulated as the following:

$$\min_{h} \ tr(H_P^T L H_P)$$
$$\text{s.t. } h_{i,j} = \begin{cases} \frac{1}{\sqrt{|p_j|}}, & \text{if } x_i \in p_j \\ 0, & \text{otherwise} \end{cases}$$
$$H^T H = I$$

It is not clear for now why the objective function above is related to ratio-cut of the similarity graph corresponding to W, explanation will be given later before our result regarding spectral clustering.

Now, we give the result for transfer learning of diverse subgroups from K-means:

Lemma 1.

$$\min_{P} \sum_{p\in P} Var(X_p) \iff \max_{P} \mathbb{E}_{\mathcal{Q}}[\sum_{q\in Q} Var(X_q)]$$

On the left hand side, P is a partition on $[N] = \{1, 2, ..., N\}$ where N is the cardinality of the given data set $\{x_i\}_{i=1}^N$. Also, we fix the size of partition to be k: $P = \{p_i\}_{i=1}^k$, and each $p \in P$ is a subset of $[N]$ with $|p| = \frac{k}{N}$. For convenience, we assume $\frac{k}{N} \in \mathbb{N}$. For each $p \in P$, we define $X_p \sim uniform(\{x_i\}_{i\in p})$ as a random variable with uniform distribution on $\{x_i\}_{i\in p}$. Also, we do not differentiate between the partition P and the corresponding label map $P : [N] \to [k]$. That is, $P(i) = P(j) \iff \exists p \in P$ such that $i, j \in p$.

On the right hand side, for each partition P, we define $D_P := \{Q : Q(i) = Q(j) \implies P(i) \neq P(j)\}$ being a set of partitions Q on $[N]$ such that data points in the same element in P cannot be clustered into the same element in Q. Also, we define $\mathcal{Q}$ to be the random partition with a distribution on D_P generated by uniformly choosing data points without replacement from each element in P to form each of its own element.

In short, the left hand side is the objective for K-means, the right hand side equivalent to find P such that the average variance on the resulting random partition is maximized.

Proof. For fixed P on $[N]$, we have:

$$\begin{aligned}
\sum_{p\in P} Var(X_p) &= \sum_{p\in P}(\frac{k}{N}\sum_{i\in p}||x_i - \frac{k}{N}\sum_{j\in p}x_j||_{l^2}^2) \\
&= \sum_{p\in P}(\frac{k}{N}\sum_{i\in [N]}||x_i||_{l^2}^2 - \frac{k^2}{N^2}\sum_{i,j\in p}\langle x_i, x_j\rangle_{l^2}) \\
&= \frac{k}{N}\sum_{i\in [N]}||x_i||_{l^2}^2 - \frac{k^2}{N^2}\sum_{p\in P}\sum_{i,j\in p}\langle x_i, x_j\rangle_{l^2} \\
&= \frac{k}{N}\sum_{i\in [N]}||x_i||_{l^2}^2 - \frac{k^2}{N^2}\sum_{i,j\in [N]}\langle x_i, x_j\rangle_{l^2}\mathbb{1}_{\{P(i)=P(j)\}}. \qquad (1)
\end{aligned}$$

Now, for right hand side, we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}[\sum_{q\in Q} Var(X_q)] &= \sum_{Q\in D_P}\mathbb{P}_{\mathcal{Q}}(Q)[\frac{1}{k}\sum_{i\in [N]}||x_i||_{l^2}^2 - \frac{1}{k^2}\sum_{q\in Q}\sum_{i,j\in q}\langle x_i, x_j\rangle_{l^2}] \\
&= \frac{1}{k}\sum_{i\in [N]}||x_i||_{l^2}^2 - \frac{1}{k^2}\sum_{Q\in D_P}\mathbb{P}_{\mathcal{Q}}(Q)[\sum_{q\in Q}\sum_{i,j\in q}\langle x_i, x_j\rangle_{l^2}] \\
&= \frac{1}{k}\sum_{i\in [N]}||x_i||_{l^2}^2 - \frac{1}{k^2}\sum_{Q\in D_P}\mathbb{P}_{\mathcal{Q}}(Q)[\sum_{i,j\in [N]}\langle x_i, x_j\rangle_{l^2}\mathbb{1}_{\{Q(i)=Q(j)\}}] \\
&= \frac{1}{k}\sum_{i\in [N]}||x_i||_{l^2}^2 - \frac{1}{k^2}\sum_{i,j\in [N]}\langle x_i, x_j\rangle_{l^2}[\sum_{Q\in D_P}\mathbb{P}_{\mathcal{Q}}(Q)\mathbb{1}_{\{Q(i)=Q(j)\}}].
\end{aligned}$$

But

$$\begin{aligned}
\sum_{Q\in D_P}\mathbb{P}_{\mathcal{Q}}(Q)\mathbb{1}_{\{Q(i)=Q(j)\}} &= \mathbb{P}_{\mathcal{Q}}(\{\mathcal{Q}(i)=\mathcal{Q}(j)\}|\mathcal{Q}\in D_P) \\
&= \begin{cases} 0 & \text{if } P(i)=P(j) \\ \frac{k}{N} & \text{if } P(i)\neq P(j) \end{cases}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{Q}}[\sum_{q\in Q} Var(X_q)] \\
=&\frac{1}{k}\sum_{i\in[N]} ||x_i||_{l^2}^2 - \frac{1}{kN}\sum_{i,j\in[N]} \langle x_i, x_j\rangle_{l^2} \mathbb{1}_{\{P(i)\neq P(j)\}} \\
=&\frac{1}{k}\sum_{i\in[N]} ||x_i||_{l^2}^2 - \frac{1}{kN}\sum_{i,j\in[N]} \langle x_i, x_j\rangle_{l^2} + \frac{1}{kN}\sum_{i,j\in[N]} \langle x_i, x_j\rangle_{l^2} \mathbb{1}_{\{P(i)=P(j)\}}
\end{aligned}
$$

Together with (1), we obtain:

$$
\min_P \sum_{p\in P} Var(X_p) \iff \max_P \mathbb{E}_{\mathcal{Q}}[\sum_{q\in Q} Var(X_q)].
$$

□

Corollary 1.

$$
\max_Q \sum_{q\in Q} Var(X_q) \iff \min_Q \mathbb{E}_{\mathcal{P}}[\sum_{p\in P} Var(X_p)]
$$

Proof. The setting is similar to the setting above, except for we fix Q first, then define D_Q, and generate random partition $\mathcal{P}$.

$$
\begin{aligned}
\mathbb{E}_{\mathcal{P}}[\sum_{p\in P} Var(X_p)] &= \sum_{P\in D_Q} \mathbb{P}_{\mathcal{P}}(Q)[\frac{k}{N}\sum_{i\in[N]} ||x_i||_{l^2}^2 - \frac{k^2}{N^2}\sum_{p\in P}\sum_{i,j\in q}\langle x_i, x_j\rangle_{l^2}] \\
&= \frac{k}{N}\sum_{i\in[N]} ||x_i||_{l^2}^2 - \frac{k^2}{N^2}\sum_{i,j\in[N]} \langle x_i, x_j\rangle_{l^2}[\sum_{P\in D_Q} \mathbb{P}_{\mathcal{P}}(P)\mathbb{1}_{\{P(i)=P(j)\}}] \\
&= \frac{k}{N}\sum_{i\in[N]} ||x_i||_{l^2}^2 - \frac{k}{N^2}\sum_{i,j\in[N]} \langle x_i, x_j\rangle_{l^2} \mathbb{1}_{\{Q(i)\neq Q(j)\}}
\end{aligned}
$$

The last equality follows from:

$$
\begin{aligned}
\sum_{P\in D_Q} \mathbb{P}_{\mathcal{P}}(P)\mathbb{1}_{\{P(i)=P(j)\}} &= \mathbb{P}_{\mathcal{P}}(\{\mathcal{P}(i)=\mathcal{P}(j)\}|\mathcal{P}\in D_Q) \\
&= \begin{cases} 0 & \text{if } Q(i)=Q(j) \\ \frac{1}{k} & \text{if } Q(i)\neq Q(j) \end{cases}.
\end{aligned}
$$

Now, since

$$\sum_{q\in Q} Var(X_q) = \sum_{q\in Q}(\frac{1}{k}\sum_{i\in q}||x_i - \frac{1}{k}\sum_{j\in q} x_j||_{l^2}^2)$$
$$= \frac{1}{k}\sum_{i\in[N]}||x_i||_{l^2}^2 - \frac{1}{k^2}\sum_{i,j\in[N]}\langle x_i, x_j\rangle_{l^2}\mathbb{1}_{\{Q(i)=Q(j)\}},$$

we have

$$\max_Q \sum_{q\in Q} Var(X_q) \iff \min_Q \mathbb{E}_{\mathcal{P}}[\sum_{p\in P} Var(X_p)].$$

□

The next step is to show that the sub-optimal solution via the transfer learning is a good estimation of the optimal solution to the diverse subgroup problem.

For the spectral clustering, we also achieved similar transfer learning results, but with more quantitative description.

Lemma 2.

$$\mathbb{E}_{\mathcal{Q}}[tr(H_Q^T L H_Q)] = \frac{1}{k}[tr(L) - tr(H_P^T L H_P)].$$

The setting is similar to the settings in the case of K-means, except for the matrix $H_P := \{(h_P)_{i,j}\}_{i,j} \in \mathbb{R}^{N\times k}$ is defined as the following:

$$(h_P)_{i,j} := \begin{cases} \frac{1}{\sqrt{|A_j|}}, & \text{if } v_i \in A_j \\ 0, & \text{otherwise} \end{cases}, \forall i \in [N], \forall j \in [k].$$

We define H_Q similarly.

Proof. On the one hand, we have:

$$tr(H_P^T L H_P) = \sum_{i\in[k]} h_i^T L h_i = \frac{1}{2}\sum_{i\in[k]}\sum_{j,k\in[N]} w_{j,k}((h_P)_{j,i} - (h_P)_{k,i})^2$$
$$= \frac{1}{2}(\frac{N}{k})^{-1}\sum_{j,k\in[N]} w_{j,k}\mathbb{1}_{P(j)\neq P(k)}$$
$$= \frac{k}{2N}(\sum_{i,j\in[N]} w_{i,j} - \sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{P(i)=P(j)}).$$

On the other,

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}}[tr(H_Q^T L H_Q)] &= \sum_{Q\in D_P} \mathbb{P}_{\mathcal{Q}}(Q)[\frac{1}{2k}(\sum_{i,j\in[N]} w_{i,j} - \sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{Q(i)=Q(j)})] \\
&= \frac{1}{2k}(\sum_{i,j\in[N]} w_{i,j} - \sum_{i,j\in[N]} w_{i,j} \sum_{Q\in D_P} \mathbb{P}_{\mathcal{Q}}(Q)\mathbb{1}_{Q(i)=Q(j)}) \\
&= \frac{1}{2k}(\sum_{i,j\in[N]} w_{i,j} - \sum_{i,j\in[N]} w_{i,j}\frac{k}{N}\mathbb{1}_{P(i)\neq P(j)}) \\
&= \frac{1}{k}(\frac{1}{2}\sum_{i,j\in[N]} w_{i,j} - \frac{k}{2N}\sum_{i,j\in[N]} w_{i,j}\frac{k}{N}\mathbb{1}_{P(i)\neq P(j)}) \\
&= \frac{1}{k}(tr(L) - tr(H_P^T L H_P))
\end{aligned}$$

□

By the similar argument, we also obtain:

Corollary 2.

$$\mathbb{E}_{\mathcal{P}}[tr(H_P^T L H_P)] = \frac{k}{N}[tr(L) - tr(H_Q^T L H_Q)].$$

Given a clustering P on the data set and let Q_* denotes the optimal diverse subgroup, we have $\mathbb{E}_{\mathcal{Q}(P)}[tr(H_Q^T L H_Q)] \leq tr(H_{Q_*}^T L H_{Q_*})$ by the definition of Q_*. Our goal is to obtain an upper bound for

$$tr(H_{Q_*}^T L H_{Q_*}) - \mathbb{E}_{\mathcal{Q}(P)}[tr(H_Q^T L H_Q)].$$

By semi-definite relaxation, we have

$$tr(H_{Q_*}^T L H_{Q_*}) \leq \sum_{i=N-\frac{N}{k}+1}^{N} \lambda_i(L),$$

where $\{\lambda_i(L) : i \in [N], 0 = \lambda_1(L) \leq ... \leq \lambda_N(L)\}$ is the spectrum of L. On the other hand, we have:

$$\mathbb{E}_{\mathcal{Q}(P)}[tr(H_Q^T L H_Q)] = \frac{1}{k}(tr(L) - tr(H_P^T L H_P)).$$

It follows that

$$\begin{aligned}
& tr(H_{Q_*}^T L H_{Q_*}) - \mathbb{E}_{\mathcal{Q}(P)}[tr(H_Q^T L H_Q)] \\
\leq & \sum_{i=N-\frac{N}{k}+1}^{N} \lambda_i(L) - [\frac{1}{k}(tr(L) - tr(H_P^T L H_P))] \\
= & (\frac{\sum_{i=N-\frac{N}{k}+1}^{N} \lambda_i(L)}{\sum_{i=1}^{N} \lambda_i(L)} - \frac{1}{k})tr(L) + \frac{1}{k}tr(H_P^T L H_P)
\end{aligned}$$

However, the above bound is merely a qualitative bound as it does not tell us how well the estimation is. Moreover, it is heavily dependent upon both the spectrum of the graph Laplacian and the proportion of total weights that are cut by the partition P.

Therefore, instead of using the relaxation to obtain an upper bound, we calculate the difference directly:

$$\begin{aligned}
&tr(H_{Q_*}^T L H_{Q_*}) - \mathbb{E}_{\mathcal{Q}(P)}[tr(H_Q^T L H_Q)] \\
=&\frac{1}{2k}\sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{Q_*(i)\neq Q_*(j)} - (\frac{1}{2k}\sum_{i,j\in[N]} w_{i,j} - \frac{1}{2N}\sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{P(i)\neq P(j)}) \\
=&\frac{1}{2N}\sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{P(i)\neq P(j)} - \frac{1}{2k}(\sum_{i,j\in[N]} w_{i,j} - \sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{Q_*(i)\neq Q_*(j)}) \\
=&\frac{1}{2N}\sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{P(i)\neq P(j)} - \frac{1}{2k}\sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{Q_*(i)=Q_*(j)} \\
=&\frac{1}{2k}(\frac{k}{N}\sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{P(i)\neq P(j)} - \sum_{i,j\in[N]} w_{i,j}\mathbb{1}_{Q_*(i)=Q_*(j)})
\end{aligned}$$

Now, assume we have obtained an optimal solution to the balanced minimum k-cut problem: P_* and let Q_* denote the partition that is the diverse subgroup on a graph generated by a stochastic block model, the following result tells us that the simple random selection from P_* generate a solution that is consistent (a consistent estimator) to the optimal solution Q_*.

Although the simple method works well on average and on typical toy models such as the stochastic block model, there exists pathological cases where the random selection performs poorly with certain probability. Let P_* denote the optimal solution to the minimum balanced cut problem, and Q_* the solution to the maximum balanced cut problem. It follows from the definition of P_*: for all $i, j \in [N]$, we have $P_*(i) \neq P_*(j)$ implies that

$$\sum_{k\in p_{P_*(j)}} w_{i,k} + \sum_{k\in p_{P_*(i)}} w_{j,k} \leq \sum_{k\in p_{P_*(i)}} w_{i,k} + \sum_{k\in p_{P_*(j)}} w_{j,k}.$$

Therefore, if there exists graphs where $\exists!\tilde{k} \in p_{P_*(j)}$ such that

- $w_{i,k} = \epsilon, \forall k \in p_{P_*(j)} \setminus \{\tilde{k}\}$,
- $w_{j,k} = \epsilon, \forall k \in p_{P_*(i)}$,
- $w_{i,\tilde{k}} = \sum_{k\in p_{P_*(i)}} w_{i,k} + \sum_{k\in p_{P_*(j)}} w_{j,k} - (2\frac{N}{k} - 1)\epsilon$.

In the case, the random selection would result in an arbitrarily poor partition for diversity purpose. An simple example is the following:

Let $G = ([4], E, W)$ where $w_{1,2} = 2, w_{1,3} = 4.9, w_{3,4} = 3$ and $w_{i,j} = 0$ otherwise. By inspection, we have $P_* = \{\{1,2\}\{3,4\}\}$ and $D_{P_*} = \{\{1,3\}\{2,4\}\} \bigcup \{\{1,4\}\{2,3\}\} =: Q_1 \bigcup Q_2$. But we have the random selections have significant deviation:

- $\frac{1}{2}\sum_{i,j} w_{i,j} \mathbb{1}_{Q_1(i)=Q_1(j)} = 4.9$,
- $\frac{1}{2}\sum_{i,j} w_{i,j} \mathbb{1}_{Q_2(i)=Q_2(j)} = 0 = \frac{1}{2}\sum_{i,j} w_{i,j} \mathbb{1}_{Q_*(i)=Q_*(j)}$.

Conclusion and Further Steps

To conclude, the project studies the followings:

1 the training data/information leakage from gradients with a goal to understand if clustering on gradients, or more specifically gradient trajectories, could give a good labeling on the local data distributions.

2 the transfer learning between clustering and diverse subgroups on gradients, or more specifically gradient trajectories, to generate effective labeling map on agents or local data distributions to improve the training efficiency, provided 1. is true.

The next steps are to further study

1 the fundamentals information leakage from gradients,

2 better bound for the difference between the optimal solution to the diverse subgroup problem and the random solution via transfer leaning from clustering,

3 the application of the clustering (diverse subgroup or both) of gradients and the labeling map on local data distributions via gradient clustering (diverse subgroup or both) to improve learning efficiency, privacy, and fairness.

References

[1] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Agüeray Arcas, Communication-Efficient Learning of Deep Networks from Decentralized Data, 2017, available online: https://arxiv.org/pdf/1602.05629.pdf

[2] Peter Kairouz, H. Brendan McMahan et al., Advances and Open Problems in Federated Learning, 2019, available online: https://arxiv.org/abs/1912.04977

[3] Ligeng Zhu Zhijian Liu Song Han, Deep Leakage from Gradients, 2019, available online: https://arxiv.org/pdf/1906.08935.pdf

[4] Bo Zhao, Konda Reddy Mopuri, Hakan Bilen, iDLG: Improved Deep Leakage from Gradients, 2020, available online: https://arxiv.org/pdf/2001.02610.pdf

[5] Tian Li, Federated Learning: Challenges, Methods, and Future Directions, 2019, available online: https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/

[6] Avishek Ghosh, Jichan Chung, Dong Yin, Kannan Ramchandran, An Efficient Framework for Clustered Federated Learning, 2020, available online: https://arxiv.org/pdf/2006.04088.pdf

[7] Chen Yu, Hanlin Tang, Cedric Renggli, Simon Kassing, Ankit Singla, Dan Alistarhk, Ce Zhang, and Ji Liu, Distributed Learning over Unreliable Networks, 2018, available online: https://arxiv.org/pdf/1810.07766.pdf