Decentralized Learning via Clustering and Diverse Subgroup

Shizhou Xu

December 2020

1 Introduction

Federated learning is first proposed in (McMahan, 2016) as a solution to both the privacy concern of centralized data analysis or learning and the computational limit of a centralized work station. In particular, federated learning train a machine learning model by first letting each of the agent to compute the model parameter gradients via a shared current model parameter, local data, local device, and then merely collecting the gradients from each of the agent to update the shared model, and finally iterate the procedure until the training terminates. 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 the local data.

2 The decentralized training framework utilize the agents' device and therefore save the cost of centralized work station to perform training.

However, the recent works on deep leakage from gradients (2019, 2020) provides us numerical evidence that the idea of sharing gradients does not fully protect the local data. Particularly, the papers 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. Thereby, the attacker is able to retrieve the original data structure from the leakage of gradients.

On the other hand, although the data structure retrieval via gradient leakage is not theoretically guaranteed, the information extraction from gradients is useful to both adversary and the central server. For instance, recently works on federated cluster learning 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 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 the model personalization and sample diversification case, 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. Also, it would be computationally expansive to perform clustering independently for each of the two purposes.

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

- Study the information retrieval form gradient sharing and the corresponding solution to the leakage.
- Study the dual problem of clustering via transfer learning: generate the diverse subgroups on data set by randomly selecting the data points from the each of the generated clusters.

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 a brief review for the state-of-the-art clustering methods K-means and spectral clustering, section 3 gives one of the motivations of gradient clustering from the adversarial perspective, section 4 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.

2 Preliminary

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

3 Gradient Information Leakage/Extraction

In this section, we first use improved deep leakage via gradient (iDLG) 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.

3.1 Label Leakage via Gradient

In [2], Zhao et. al. showed that sharing gradient actually reveal the true labels of the corresponding agent, when the training model is a neural network with cross-entropy loss and over one-hot labels. More specifically, 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.

3.2 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.

More importantly, the

4 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 result, we first have to talk about the clustering methods we use. In this project, we use K-means and spectral clustering.

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

Setting: 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 left 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:

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

□

2. WTS:

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

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

Proof.

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

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

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

we have

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

□

References

[1] Jianyu Wang, Gauri Joshi, Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms, available online: https://arxiv.org/abs/1808.07576

[2] Jianyu Wang, Gauri Joshi, Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms, available online: https://arxiv.org/abs/1808.07576

[3] Jianyu Wang, Gauri Joshi, Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms, available online: https://arxiv.org/abs/1808.07576

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

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