Robust Dynamic Distributed Learning

Lab Report: Development and Application of Data Mining and Learning Systems: Machine Learning and Data Mining, 2019

Pratika Kochar

Data is often generated by physically distributed devices, such as mobile phones, cars, sensors. Training models using this data can provide useful insights. One approach is to centralise the data, prior to training. This would not only involve high centralisation costs but also raise privacy concerns. To decouple model training from the direct access to the user data, federated learning is used. Model parameters are aggregated instead of data centralisation. Currently, the arithmetic mean is used for aggregation. However, the mean is not robust to noise. At the same time, noise plays a crucial role for enforcing privacy, thus, mandates the need of a robust aggregator. We suggest to use geometric median because of its suitable properties - robustness, ability to deal with large scale high dimensional data, translation and scale invariant. We compare the performance of geometric median with arithmetic mean under different setups. These experiments demonstrate that geometric median outperforms arithmetic mean in the presence of noise.

1 Introduction

Cars, mobile phones are sources of ample data on which machine learning can be performed. Apart from being large scale, this data on the other hand is highly private. Centralising this privacy-sensitive data at a server is something which users would not want.McMahan et al. [2016] proposed an approach called Federated Learning that enables training models without centralising data. Instead, only the local updates are communicated from the client nodes to the global shared model at the server. Updates received from the client nodes are aggregated using averaging. This aggregated result is then used to update the parameters of the global shared model.
Above approach involves periodic communication between the client nodes to synchronise models. Kamp et al. [2014] introduce first data-dependent distributed prediction protocol that dynamically adjusts the amount of communication by performing model synchronizations only in system states that show a high variance among the local models.

Although privacy-preserving by design, federated learning is vulnerable to noise corruption of local agents as mentioned in the paper by Han and Zhang [2019].Even a single noise-corrupted agent can bias the model training resulting in invalid results. In certain cases, noise is added deliberately. Updates sent from devices can still contain some personal data and o differential privacy is used to add gaussian noise to data shared by devices [JOHNSON, 2019].
Realising the impact of noise in federated learning, we wanted to incorporate a robust aggregator for performing updates. Geometric median has desirable properties especially for handling large scale high dimensional data and potential outliers. Geometric median can handle larger sample sizes recursively and provides estimations that are much more accurate than static estimation procedures as demonstrated by Cardot et al. [2011]. In static aggregation, synchronisation is done in fixed intervals (after certain number of rounds), whereas in dynamic aggregation, it is done when certain set of global and local conditions are violated. As a result, communication between local learners is less in dynamic aggregation. Averaged updates might not represent the overall learning state of all the local learners. Geometric median has the normalisation effect and thus, the ability to give better estimates. Therefore, it makes more sense to use, geometric median in dynamic aggregation. Given these properties, geometric median is a strong candidate to be considered an aggregator.
Therefore, in this work, geometric median is used to perform aggregates both in peridic (static) and dynamic aggregation schemes. Experiments demonstrate that, without noise, geometric median is at par with arithmetic mean with slightly more communication in dynamic aggregation scheme. On inclusion of minute noise, arithmetic mean breaks, resulting in high loss. It also leads to the exploding gradients in certain scenarios. On the contrary, geometric median handles noise gracefully and thus, proves to be more robust than arithmetic mean.

2 Robust Aggregation Using Geometric Median

2.1 Preliminaries

We extend our approach on the static averaging and dynamic synchronization protocol as mentioned in Kamp et al. [2014]. Therefore, we consider a distributed learning environment with k client nodes (local leaners) with linear models $w_{t,1}, .., w_{t,k} \in R^{n}$. t $\in [T]$ represents discrete time, with $T \in$ N total time horizon with respect to which we analyze the system's performance. (x_{tl}, y_{tl}) represent training examples at learner l at time t , $p_{t,l}$ represents prediction score by the model. Local updates $w_{t+1,l} \in R^n$ are computed using an update rule φ of the form $\varphi(\mathrm{w}_{t,l}, x_{t,l}, y_{t,l})$. $W_t \in R^{kxn}$ denotes the complete model configuration of all local models at time t and $\sigma : R^{kxn} \rightarrow \mathrm{R}^{\,kxn}$ performs synchronisation by resetting the whole model configuration to new state after completion of all the local updates.

Definition 2 Aggregation: Synchronisation operator $\sigma : R^{kxn} \rightarrow \mathrm{R}^{\,kxn}$ transfers the

current model configuration $W_t \in R^{kxn}$ into a single strong global model $\sigma(\mathrm{W}_t)$ replacing the local models. Kamp et al. [2014]

Definition 3 Given $w_{t,1}, .., w_{t,k} \in R^n$, we define $W_t = 1/\mathrm{k} \sum_{l=1}^{k} w_{t,l}$ as the (arithmetic) mean model at time t Kamp et al. [2014].

First we define geometric median and then present the adaptation of the state-of-the-art static averaging and dynamic synchronization protocol Kamp et al. [2014] for geometric median.

2.1 Geometric Median

The geometric median of a discrete set of sample points in a Euclidean space is the point minimizing the sum of distances to the sample points [Wikipedia contributors, 2019]

Definition 1 Given $x_1, x_2, ..., x_m$ for all $x_i \in R^d$ then $\mathrm{y}^* = \underset{y \in R^d}{\mathrm{argmin}} \sum_{i=1}^{m} ||x_i - y||_2$ is called the Geometric Median [Wikipedia contributors, 2019]

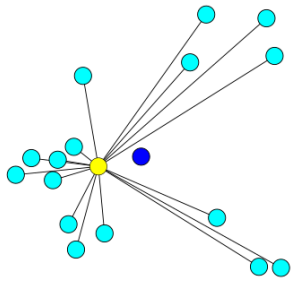


Figure 1: Geometric median represented by the yellow ball[Wikipedia contributors, 2019]

2.3 Modifications to Static Aggregation Protocol

Here we replace the aggregation of the model weights done using mean by the geometric median. Static aggregation operator is now formally given by $\sigma(\mathrm{W}_t) = ((\mathrm{W}_t)_{gm}, ..., (W_t)_{gm})$ if t mod b = 0 and $\sigma(\mathrm{W}_t) = W_t$, otherwise where $(W_t)_{gm}$ represents model configuration aggregated using geometric median at time t and b is the batch size.

Algorithm 1 Static Aggregation for Geometric Median

Initialization:
local models $w_{1,1}, .., w_{1,k} \longleftarrow (0, ..., 0)$

Round t at node l:
observe $x_{t,l}$ and provide service based on $p_{t,l}$
observe $y_{t,l}$ and update $w_{t+1,l} \longleftarrow \varphi(w_{t,l}, x_t, y_t)$
if t mod b == 0 then
send $w_{t,l}$ to coordinator

At coordinator every b rounds:
receive local models $\{w_{t,l} : l \in [k]\}$
for all l ∈ [k] set $w_{t,l} \leftarrow \sigma(w_{t,1}, ..., w_{t,k})_l$

2.4 Modifications to Dynamic Synchronization Protocol

During the static aggregation, we can end up in the situation where communication is done even after the models have already converged an optimum. To make communication efficient Kamp et al. [2014] introduced dynamic synchronization protocol that achieves the same performance as static averaging but with lesser communication. We adapt their algorithm as follows: We replace averaging with geometric median and modify the local and globals conditions.Conditions are modified as such because the distance of each local model to the actual aggregate will always be smaller than to any common reference point.

Algorithm 2 Dynamic Synchronisation Protocol for Geometric Median

Initialization:
local models $w_{1,1}, .., w_{1,k} \longleftarrow (0, ..., 0)$
reference vector $r \longleftarrow (0, ..., 0)$
violation counter $v \longleftarrow 0$

Round t at node l:
observe $x_{t,l}$ and provide service based on $p_{t,l}$
observe $y_{t,l}$ and update $w_{t+1,l} \longleftarrow \varphi(w_{t,l}, x_t, y_t)$
if t mod b == 0 and $||w_{t,l} - r^2|| > \Delta$ then
send $w_{t,l}$ to coordinator (violation)

At coordinator on violation:
let B be the set of nodes with violation
v $\longleftarrow v+ \mid B \mid$
if v = k then $B \longleftarrow$ [k], $v \longleftarrow 0$
while $B \neq [k] and \frac{1}{B} \sum_{l \in B} ||w_{t,l} - r^2|| > \Delta$
do
  augment B by augmentation strategy
  receive models from nodes added to B
end
send model $w_{gm} = geometricmedian(w_{t,l})$ to nodes in B
if B = [k] also set new reference vector r $\longleftarrow$ w_{gm}

3 Experiments

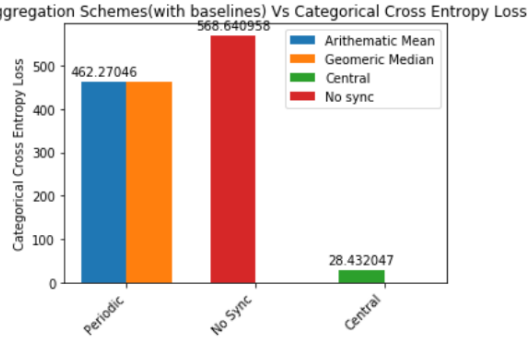
We simulate distributed learning environment where the nodes are processed in round robin fashion. Experiments are conducted to compare the performance - loss, communication of different aggregation techniques - geometric median, arithmetic mean and aggregation schemes - periodic, dynamic. Two datasets are used for training - MNIST and CIFAR-10. Here, Stochastic Gradient Descend has been applied to all the nodes along with categorical cross-entropy loss. Basically, experiments can be categorised as - with noise and without noise. Our findings and insights are discussed in the following sections.

3.1 Experiments conducted without introduction of noise

Without noise, arithmetic mean and geometric median are equal in model performance

Figure 2 depicts categorical cross entropy loss for both - geometric mean and averaging along with no synchronisation and central baselines. It can be observed that both the techniques reach similar loss values. As expected, both he methods perform better than when there is no synchronisation at all. Central and no sync baselines have been trained with the ideal parameters in this setup. In dynamic aggregation, geometric median

Figure 2: Performance with respect to categorical cross entropy loss tested on MNIST dataset

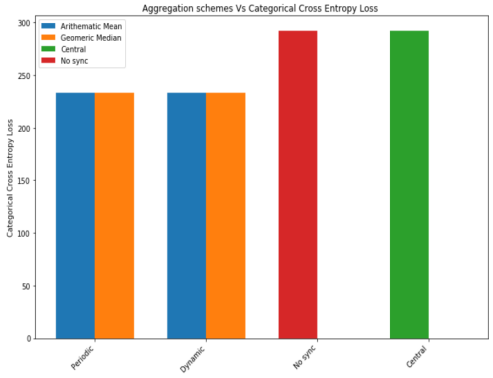


communicates more than arithmetic mean as depicted in Figure 3b. Both the techniques achieve same loss but it can be observed that in dynamic aggregation, geometric median communicates slightly more than arithmetic mean. Similar behavior can be noticed in Figure 4 for MNIST dataset.

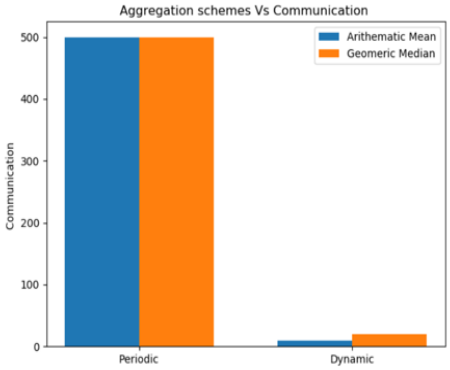
Experiments were executed to capture impact on categorical cross entropy loss of different learning rates, batch sizes - macro and mini. First individual parameters were varied and then finally learning rate and mini batch sizes were increased each time with a factor k. For both the techniques - geometric median and arithmetic mean loss with miniscule differences were observed in all of the above setups. Figure 5 summarises the outputs in each case.

3.2 Experiments conducted with introduction of noise

We noticed significant change in performance of arithmetic mean on introduction of noise.

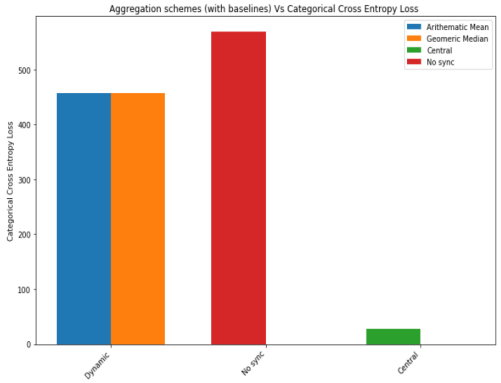


(a) Plot depicting loss for different aggregation schemes

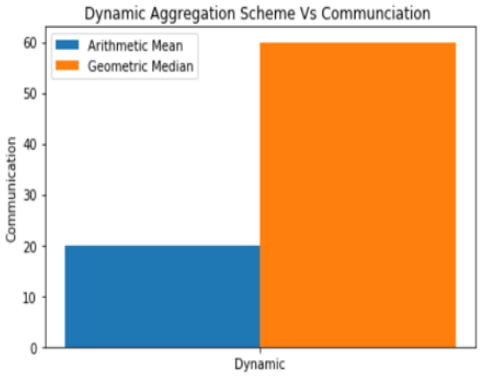


(b) Plot depicting communication for different aggregation schemes

Figure 3: Performance with respect to communication tested on CIFAR-10 dataset Geometric median communicates more often as compared to arithmetic mean in dynamic case



(a) Plot depicting loss for different aggregation schemes



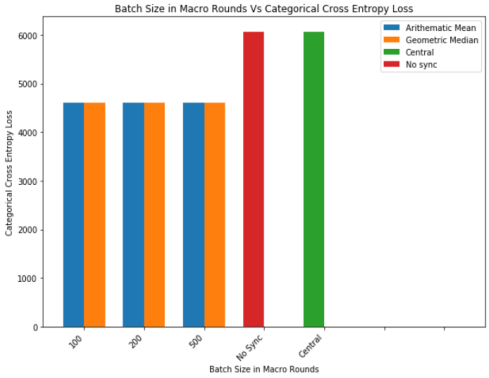
(b) Plot depicting communication for different aggregation schemes

Figure 4: Performance with respect to communication tested on MNIST dataset Geometric median communicates more often as compared to arithmetic mean in dynamic case

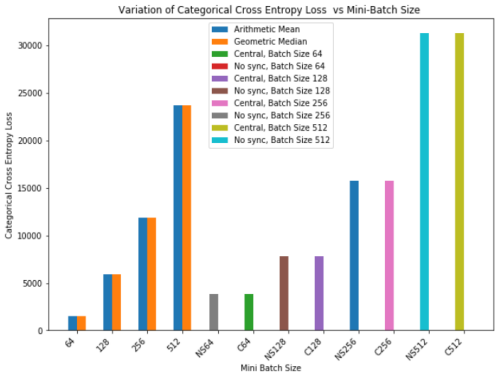
3.2.1 Types of noise

Two types of noise: gaussian noise which is added for differential privacy and other by perturbing model weights.

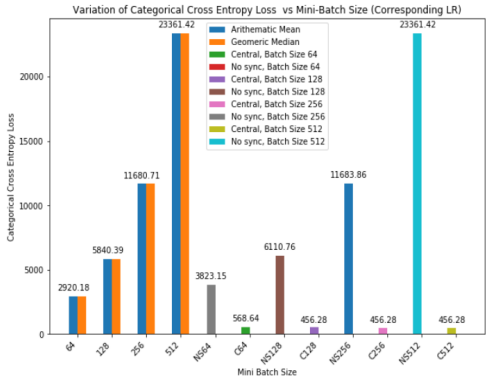
(a) Plot depicting loss for different learning rates



(b) Plot depicting loss for different macro batch sizes



(c) Plot depicting loss for different mini batch sizes



(d) Plot depicting loss for different mini batch sizes and learning rates

Figure 5: Performance with respect to categorical cross entropy loss tested on MNIST dataset [Periodic aggregation]

Gaussian Noise

We add gaussian noise with a small sigma value of 0.1 only to two nodes. This is done to check if adding a miniscule amount of noise modifies the performance of both the aggregation techniques.In Figure 6, arithmetic mean results in a very high loss around 43000 whereas geometric median reaches till 3000. Geometric median on the other hand is robust to this noise.

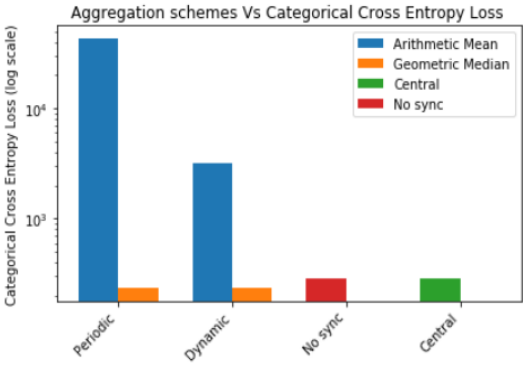
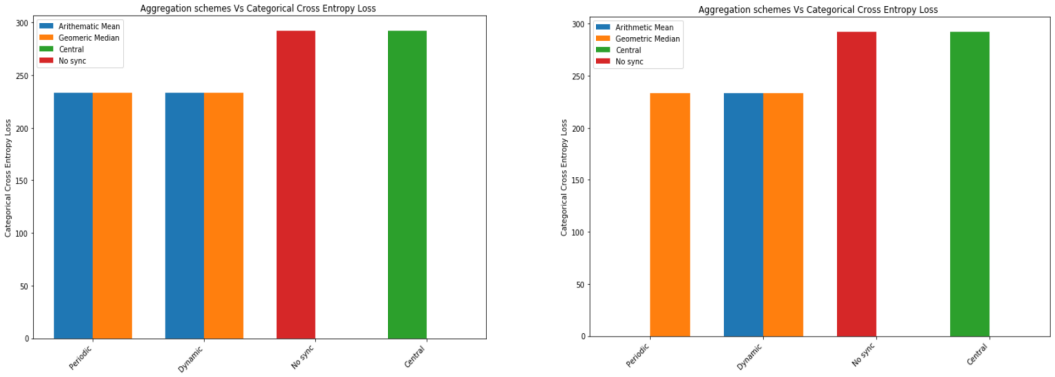


Figure 6: Performance with respect to categorical cross entropy loss tested on MNIST dataset

Noise by perturbing model weights

Here we modify weights of the existing models (only two). In case of arithmetic mean, we observe exploding gradient scenario resulting in Nan loss Figure 7b but geometric median works in this case as well Figure 7 .



(a) Plot depicting categorical cross entropy for less number of rounds

(b) Plot depicting categorical cross entropy for more number of rounds

Figure 7: Performance with respect to categorical cross entropy loss tested on CIFAR-10 dataset for different number of rounds

4 Conclusion

We began with discussing the need of federated learning and how is noise is inevitable in such cases. Because of the robustness of the geometric median, we sugested to use it as an aggregator instead of the state-of-the-art arithmetic mean. Experiments were conducted on different datasets demonstrating that geometric median has similar be-

havior as arithmetic mean with more communication in dynamic aggregation scheme in absence of noise. But on introduction of noise, we noticed arithmetic mean results in high loss and sometimes exploding gradients too whereas geometric median works even in the dynamic aggregation. In real world scenario, noise is something which cannot be overlooked, therefore, we conclude that in the presence of noise geomteric median is better than arithmetic mean.

Aggregation techniques / Parameters	Arithmetic Mean	Geometric Median
Loss	Similar	Similar
Robustness	Periodic aggregation: **Not robust** High loss, exploding gradients Dynamic aggregation: **Not robust**	Periodic aggregation: **Robust** Dynamic aggregation: **Robust**
Communication	Periodic aggregation: Similar Dynamic aggregation: Lower than periodic	Periodic aggregation: Similar Dynamic aggregation: Lower than periodic, higher than arithmetic mean

Table 1: Table summarizing experimental results

References

Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19, 01 2011. doi: 10.3150/11-BEJ390.

Yufei Han and Xiangliang Zhang. Robust federated training via collaborative machine teaching using trusted instances. *ArXiv*, abs/1905.02941, 2019.

KHARI JOHNSON. How federated learning could shape the future of ai in a privacy-obsessed world. 2019.

Michael Kamp, Mario Boley, Daniel Keren, Assaf Schuster, and Izchak Sharfman. Communication-efficient distributed online prediction by dynamic model synchronization. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 623–639, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016. URL http://arxiv.org/abs/1602.05629.

Wikipedia contributors. Geometric median — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Geometric_median&oldid=914151483, 2019. [Online; accessed 7-September-2019].