NEURAL INFORMATION PROCESSING SYSTEMS

Local Gradient Aggregation for Decentralized Learning from Non-IID data

Yasaman Esfandiari, Sin Yong Tan, Zhanhong Jiang*, Aditya Balu, Chinmay Hegde†, Soumik Sarkar
Iowa State University, Johnson Controls*, New York University†

Abstract

Distributed centralized learning refers to a class of learning algorithms that enable a group of agents to train models using a dataset distributed amongst them with the aid of a central parameter server. Recently, decentralized learning algorithms have demonstrated state-of-the-art results. However, an essential requirement to achieve such performance has been balanced distribution (among classes) of data among the agents, referred to as IID data. In real-life applications, having a precise IID distribution of data among the agents is often not feasible. We propose a Local Gradient Aggregation (*LGA*) where each agent collects the gradient information from its neighboring agents and updates its model with a projected gradient. By comparing against state-of-the-art decentralized algorithms, we show that our algorithm achieves the highest accuracy rate on non-IID data distribution while preserving the IID counterpart's performance.

Introduction

- Centralized learning algorithms (e.g., federated learning) have demonstrated state-of-the-art performance in learning collaboratively from numerous agents.
- In certain use cases such as learning over a robotic network, continuous communication with a central parameter server is often not feasible. To address this concern, several decentralized learning algorithms have been proposed.
- In this paper, we study two aspects of distributed deep learning: Decentralized learning, and learning from non-IID data distributions.
- We propose *Local Gradient Aggregation* (LGA) algorithm and show its effectiveness in learning models in a decentralized manner from both IID and non-IID data distributions.

Problem Formulation

- The standard (unconstrained) empirical risk minimization problem that we are solving in decentralized distributed learning can be represented as:

$$\min f(x) = \sum_{j=1}^{N}\sum_{i\in\mathcal{D}_j} f^i(x^j) := \frac{N}{n}\sum_{j=1}^{N} f_j(x^j); \text{s.t. } x^j = x^l \;\; \forall\, (j,l)\in\mathbb{C}$$

- Comparing the data distribution shifts in the continual learning with the non-IID data distributions in the decentralized learning, we can leverage the techniques into the decentralized learning framework by finding the local optimal gradient for each local model.

$$\langle g, g_t\rangle := \langle \frac{\partial f(s;x)}{\partial x}, \frac{\partial f(\mathcal{M}_t;x)}{\partial x}\rangle \geq 0,\ \forall t < k.$$

Algorithm

Algorithm 1 Local Gradient Aggregation

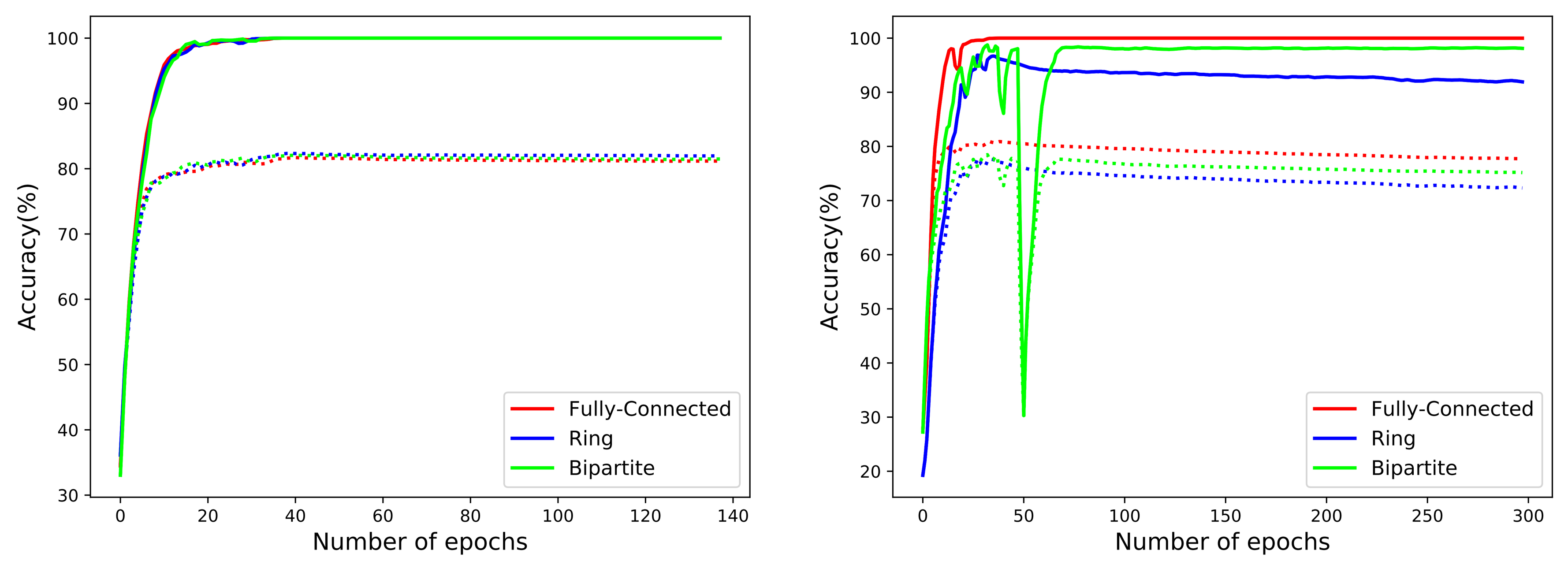
```
Initialization: D_j, x_0^j, v_0^j, (j = 1, 2, ..., N), α, K, a QP solver
for k = 0 : K do
    for j = 1 : N do
        Randomly shuffle the corresponding data subset D_j
        Compute g_k^{jj}
        G^j = {}
        for each agent l, s.t. (j, l) ∈ C do
            Compute g_k^{jl}
            G^j ← G^j ∪ g_k^{jl}
            w_k^j = Σ_l π_{jl} x_{k-1}^l
        end for
        g̃^j ← QP(g_k^{jj}, G^j)

        v_k^j = μ v_{k-1}^j − α g̃^j

        x_k^j = w_k^j + v_k^j
    end for
end for
```

Experimental Results

- Figure 1 shows that LGA can maintain the high accuracy when learning from both IID and non-IID data distributions.

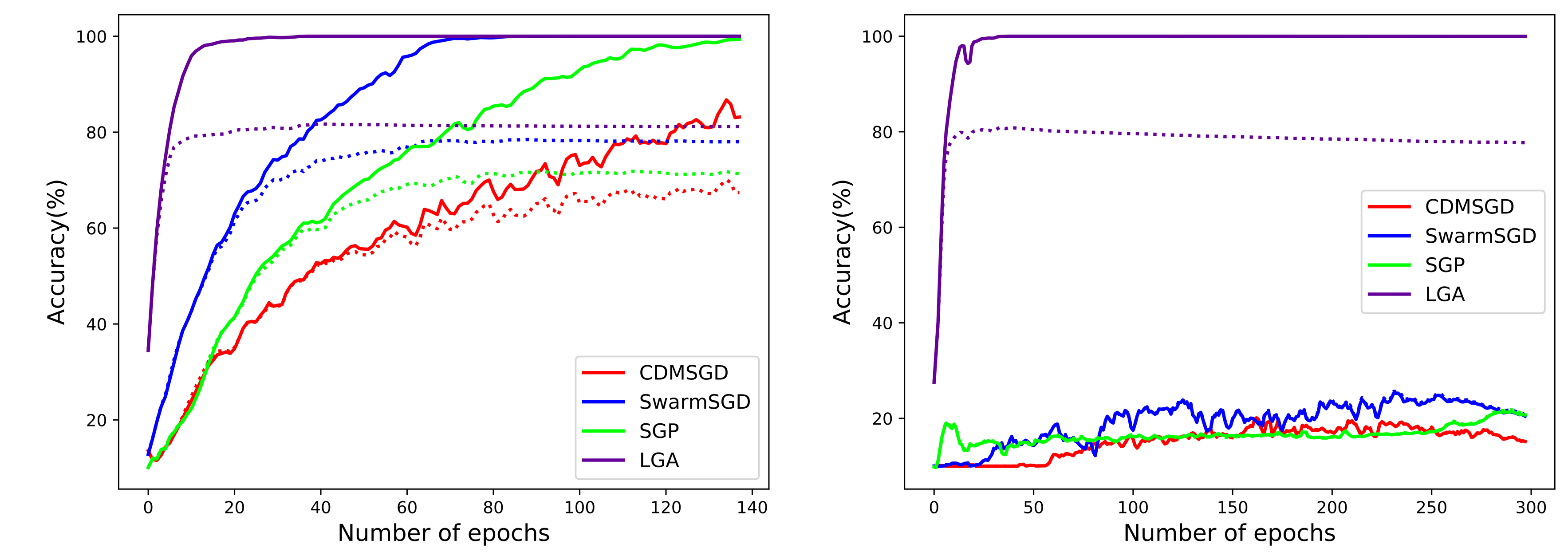


(a) (b)

Figure 1: *Average training and validation accuracy for LGA method on (a) IID (b) non-IID data*

Experimental Results

- Figure 2 shows that LGA achieves the highest accuracy compared to the state-of-the-art methods in less number of epochs smoothly and maintains it in both IID and non-IID scenarios.



(a) (b)

Figure 2: *Average training (solid line) and validation (dash line) accuracy for different methods on (a) IID (b) non-IID data distributions for fully connected graph topology with 5 agents*

- We have also tested our method on CIFAR-10 data set with 10 agents and MNIST data sets with bot 5 and 10 agents. Experimental results show that LGA achieves the highest accuracy compared to the state-of-the-art methods in all of those experiments as well and will preserve the performance when changing from IID to non-IID data distributions.

Conclusions

- In this paper, we propose a Local Gradient Aggregation (*LGA*) algorithm to effectively learn from non-IID data distributions in a decentralized manner that resolves the scalability and connectivity concerns associated with using a central parameter server.
- We present the convergence characteristics of the algorithm and investigate the effect of different topologies, with different combinations of agent numbers empirically.
- Also, we compare the performance of *LGA* algorithm with state-of-the-art decentralized learning algorithms as baseline methods.
- Future research directions include: (i) Computation analysis of *LGA* (ii) investigating projection methods other than QP (iii) empirical comparison between different extents of non-IIDness in the data distribution.

IOWA STATE UNIVERSITY

Self-aware Complex Systems Laboratory

References: [1] Esfandiari, Y., Tan, S.Y., Balu, A., Jiang, Z., Hegde, C. , Sarkar, S., Local Gradient Aggregation for Decentralized Learning from Non-IID data, OPT 2020 workshop

Acknowledgements: This work was partly supported by the National Science Foundation under grant number CAREER-1845969.

DICE Lab

NYU