Faculty of Engineering

LEARNING PRIVATELY OVER DISTRIBUTED FEATURES: AN ADMM SHARING APPROACH

Yaochen Hu (yaochen@ualberta.ca) Peng Liu (P.Liu@kent.ac.uk) Keshi Ge (gekeshi@nudt.edu.cn)
Linglong Kong (lkong@ualberta.ca) Bei Jiang (bei1@ualberta.ca) Di Niu (dniu@ualberta.ca)

NEURAL INFORMATION PROCESSING SYSTEMS

UNIVERSITY OF ALBERTA

Introduction

- Distributed machine learning has been widely studied in order to handle exploding amount of data.
- We study an important yet less visited distributed learning problem where features are vertically partitioned among multiple parties.
- Sharing of raw data or model parameters among parties is prohibited due to privacy concerns.

Motivating example

Firestone and Ford tire controversy

Ford collects information about vehicles. Firestone collects information about tires. Vehicles can be linked to tires. In 2001, numerous accidents due to tread separation were reported.

Initially both companies blamed each other. It turned out that it was only Ford Explorers with Firestone tires from the Decatur, Illinois plant, in specific situations that had these problems. If found out earlier, much loss could have been avoided.

However, their data are not shared due to commercial concerns.

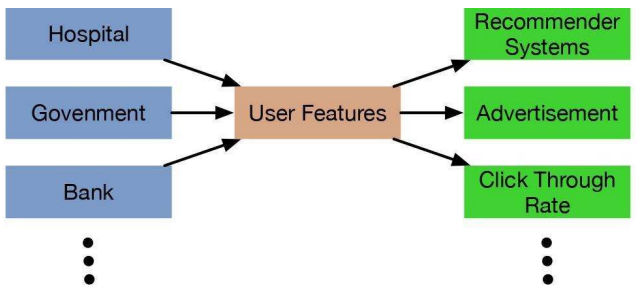


Figure 1: Applying Sensitive User Features to Other Applications

Objective

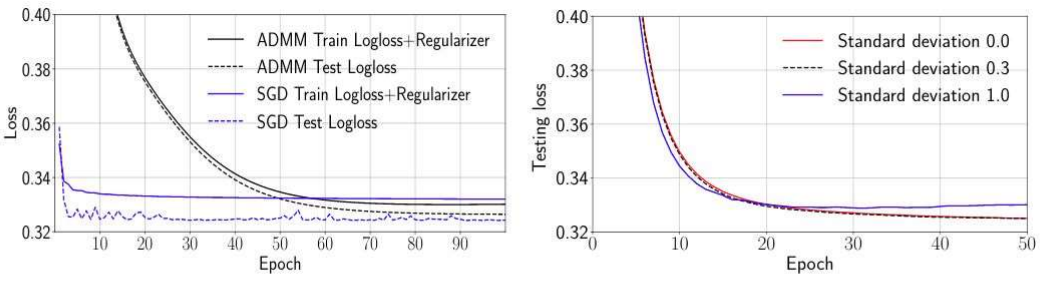
We want to develop a method that each party can share data while individual can not be identified from the “shared data”.

Method

We propose an ADMM sharing framework to approach risk minimization over distributed features, where each party only needs to share a single value for each sample in the training process, thus minimizing the data leakage risk, and we perturbed this value to achieve ε-δ differential privacy.

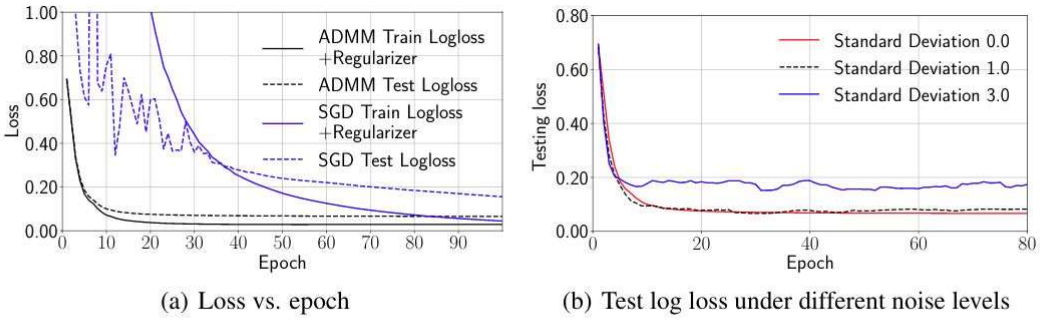
Results

We introduce a novel differentially private ADMM sharing algorithm and bound the privacy guarantee with carefully designed noise perturbation.



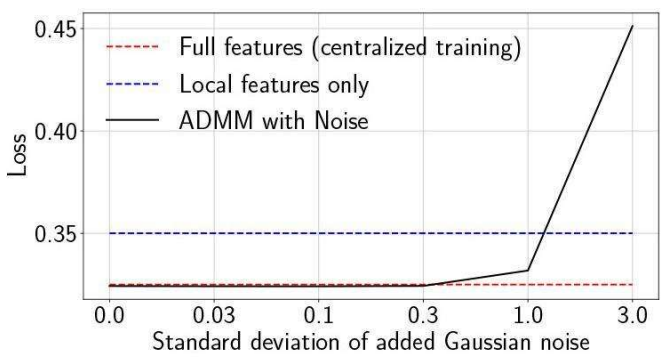
(a) Loss vs. epoch (b) Test log loss under different noise levels

Figure 1: Performance over the *a9a* data set with 32561 training samples, 16281 testing samples and 123 features.

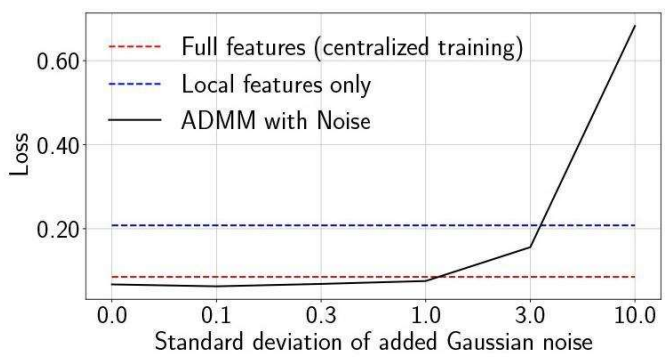


(a) Loss vs. epoch (b) Test log loss under different noise levels

Figure 2: Performance over the *gisette* data set with 6000 training samples, 1000 testing samples and 5000 features.



(a) *a9a* data set



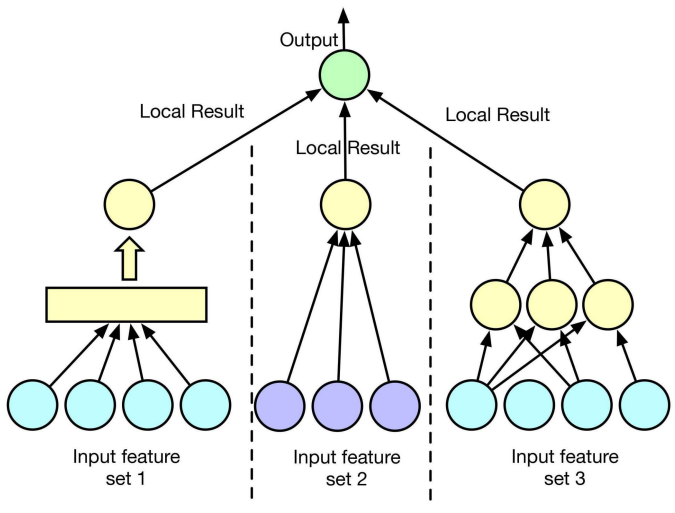
(b) *gisette* data set

Figure 3: Test performance for ADMM under different levels of added noise.

Conclusions

We study learning over distributed features where none of the parties shall share the local data.

- We propose the parallel ADMM sharing algorithm to solve this challenging problem where only intermediate values are shared, without even sharing model parameters.
- We have shown the convergence for convex and non-convex loss functions.
- To further protect the data privacy, we apply the differential privacy technique in the training procedure to derive a privacy guarantee within T epochs.
- The result shows that the ADMM sharing algorithm converges efficiently, especially on dataset with large number of features.
- Furthermore, the differentially private ADMM algorithm yields better prediction accuracy than model trained from only local features while ensuring a certain level of differential privacy guarantee.



References

1 Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Improving the privacy and accuracy of admm-based distributed algorithms. In International Conference on Machine Learning, pages 5791–5800, 2018.

2 Zonghao Huang, Rui Hu, Yanmin Gong, and Eric Chan-Tin. Dp-admm: Admm-based distributed learning with differential privacy. arXiv preprint arXiv:1808.10101, 2018.

3 Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. Foundations and Trends R in Theoretical Computer Science, 9(3–4):211–407, 2014.

www.ualberta.ca/engineering

Donadeo Innovation Centre for Engineering, University of Alberta, Edmonton, Alberta, Canada, T6G 1H9