My great thesis

Me

CMU-RI-TR-YY-NN

October 25, 1985

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Distinguished professor, *chair*
Distinguished professor
Distinguished professor
Distinguished professor, *Distinguished University*

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Robotics.

To my favorite robot.

Abstract

In modern machine learning, the abundance of data generated across diverse and distributed sources has made distributed training a central paradigm, particularly in large-scale applications such as Federated Learning. However, two key challenges arise in distributed training: ensuring communication efficiency and preserving the privacy of sensitive data used during training. This thesis addresses these challenges by exploring the interplay between communication efficiency, differential privacy, and optimization algorithms—key elements for enabling scalable, efficient, and privacy-preserving distributed learning.

We first address communication efficiency in distributed optimization by introducing Rand-Proj-Spatial, a sparsification-based communication efficient estimator for distributed vector mean estimation that leverages cross-client correlation through random subspace projections using the Subsampled Randomized Hadamard Transform (SRHT), achieving significant improvements over conventional sparsification methods. Next, focusing on differential privacy in prediction tasks, we propose DaRRM, a unified framework for private majority ensembling that optimizes a data-dependent noise function to improve model utility under fixed privacy guarantees and demonstrates strong empirical performance in private image classification. Finally, examining the interplay between privacy and optimization, we analyze the limitations of differentially private shuffled gradient methods (DP-ShuffleG), a practical optimization algorithm for solving private empirical risk minimization (ERM), and introduce Interleaved-ShuffleG, a hybrid algorithm that incorporates public data to reduce empirical excess risk, supported by novel theoretical insights and superior empirical performance across diverse datasets and baselines.

Together, these contributions advance the understanding and design of communication-efficient and privacy-preserving optimization algorithms critical for scalable and secure distributed learning.

Acknowledgments

These people are awesome.