Thesis: Coded Matrix Multiplication

Abstract: Matrix multiplication is a fundamental building block in many machine learning models. As the input matrices may be too large to be multiplied on a single server, it is common to split input matrices into multiple sub-matrices and execute the multiplications on different servers. However, in a distributed infrastructure, it is common to observe stragglers whose performance is significantly lower than other servers at some time. Compared to replicating each task on multiple servers, coded matrix multiplication, i.e., a combination of coding theoretic techniques and distributed matrix multiplication, can tolerate the same number of stragglers with much fewer servers. The recent years have witnessed the fast development of research in coded matrix multiplication. Besides alleviating the stragglers effect, there are many new improvements and applications in the coded matrix multiplication area.

In this proposal, we first describe the fundamental of coded matrix multiplication. Then we review Spinner, a coded matrix multiplication scheme that can leverage resources from homogeneous and heterogeneous workers. Furthermore, we present a general coding framework for the matrix chain multiplication where multiple matrix multiplication tasks can be finished with just one round. Finally, we propose a new coding scheme called epsilon-Local Coding that can change the locality of the parity data and original data in the encoding process.

Committee:

- Professor Jun Li, Mentor, Queens College
- Professor Ping Ji, John Jay College
- Professor Kaliappa Ravindran, The City College

Committee:

- Professor Amlan Nayak, Meta Platforms, Outside Member