THE UNIVERSITY OF CHICAGO | THE COMMITTEE ON COMPUTATIONAL AND APPLIED MATHEMATICS

SPECIAL COLLOQUIUM

INDERJIT S. DHILLON

Department of Computer Science, University of Texas at Austin;
Google

CASPR: Combining Axes Preconditioners through Kronecker Approximation for Training Large Neural Networks

TUESDAY, April 7th at 4:00 PM
Kent Chemical Laboratory, Room 107, 1020 E. 58th St.

ABSTRACT

Most large neural networks, including Large Language Models, are trained using adaptive regularization methods such as Adam, which can be regarded as diagonally preconditioned stochastic gradient descent. This diagonal preconditioner comes from a diagonal approximation of the gradient outer product matrix. However, a recent open competition called "AlgoPerf: Training Algorithms benchmark competition" revealed an intriguing discovery: a non-diagonal preconditioning method called Shampoo, which uses a Kronecker product approximation of the outer-product matrix, was found to be the best method on a varied suite of benchmark problems. In this talk, I will introduce adaptive methods and show how Kroencker products can be used to get a computationally efficient preconditioner. I will then discuss a general technique called CASPR, which optimizes matrix-shaped parameters by finding preconditioners for each mode/axis of the parameter and combines them using a Kronecker sum/product based approximation, yielding Shampoo as a special case. Experimental results demonstrate that CASPR shows improved training and generalization performance.

Organizers:
Guillaume Bal, Department of Statistics (CCAM), guillaumebal@uchicago.edu & Nisha Chandramoorthy, Department of Statistics (CCAM), nishac@uchicago.edu, Daniel Sanz-Alonso, Department of Statistics (CCAM), sanzalonso@uchicago.edu

CAM Colloquium URL: https://cam.uchicago.edu/events/cam-colloquium/