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

COLLOQUIUM

Yijun Dong

Courant Institute of Mathematical Sciences
New York University

Understanding Post-training through the Lens of Intrinsic Dimension.

THURSDAY, January 15th at 4:00 PM
Jones 303, 5747 S. Ellis Ave. Chicago, IL 60637

ABSTRACT

Post-training is becoming the primary interface between powerful pre-trained models and challenging real-world problems, where we aim to adapt large pre-trained models via limited, heterogeneous data while preserving their capabilities and reliability. In this talk, we introduce a step toward a unified theoretical and algorithmic framework for post-training through the lens of intrinsic dimensions. In particular, we focus on an emerging post-training phenomenon, weak-to-strong (W2S) generalization, in which a strong pre-trained student model fine-tuned only with supervision from a weaker teacher model can often outperform its teacher. Theoretically, we explain when and why W2S generalization occurs from a sample-efficiency perspective, reveal the value of teacher-student discrepancy for W2S, and investigate the effects of systematic biases on W2S. Algorithmically, we propose a practical, theory-inspired remedy for W2S under spurious correlation. The talk will conclude with an outlook on the broad applications of random matrix tools for understanding and improving post-training.

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/