Mathematics, Statistics, and Computer Science **@ UIC**

Statistics and Data Science Seminar

Minimax Off-Policy Evaluation for Multi-Armed Bandits

Cong Ma (University of Chicago)

Abstract: This talk is concerned with the problem of off-policy evaluation in the multi-armed bandit model with bounded rewards. We develop minimax rate-optimal procedures under three different settings. First, when the behavior policy is known, we show that the Switch estimator, a method that alternates between the plug-in and importance sampling estimators, is minimax rate-optimal for all sample sizes. Second, when the behavior policy is unknown, we analyze performance in terms of the competitive ratio, thereby revealing a fundamental gap between the settings of known and unknown behavior policies. When the behavior policy is unknown, any estimator must have mean-squared error larger—relative to the oracle estimator equipped with the knowledge of the behavior policy—by a multiplicative factor proportional to the support size of the target policy. Moreover, we demonstrate that the plug-in approach achieves this worst-case competitive ratio up to a logarithmic factor. Third, we initiate the study of the partial knowledge setting in which it is assumed that the minimum probability taken by the behavior policy is known. We show that the plug-in estimator is optimal for relatively large values of the minimum probability, but is sub-optimal when the minimum probability is low. In order to remedy this gap, we propose a new estimator based on approximation by Chebyshev polynomials that provably achieves the optimal estimation error. This is a joint work with Banghua Zhu, Jiantao Jiao and Martin Wainwright.

Wednesday, December 1 at 4:00 PM in Zoom