THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Xueran Tao
Department of Statistics
The University of Chicago

"Finite-Sample Analysis of Prediction-Powered Inference for Linear Regression"

May 7, 2026, at 2:00 PM
Jones 111, 5747 S. Ellis Avenue

Abstract

Prediction-powered inference (PPI) is a recently proposed framework that leverages black-box machine learning predictions together with a small labeled dataset to improve estimation efficiency while correcting for bias introduced by imperfect predictors. In this thesis, we study PPI for linear regression. While existing asymptotic theory guarantees that PPI with an optimally tuned weighting parameter outperforms classical ordinary least squares (OLS), its finite-sample behavior remains less well understood.

We derive a non-asymptotic error bound for the PPI estimator that makes explicit its dependence on the tuning parameter, the labeled and unlabeled sample sizes, the dimension, and the quality of the black-box predictive model. The bound reveals two mechanisms through which PPI can improve over OLS: stabilization of the empirical covariance through unlabeled data, and variance reduction via a control-variate effect when the predictor captures structure beyond the best linear approximation.

Building on this bound, we further identify regimes in which PPI may fail to outperform OLS, arising from finite-sample effects that are invisible to asymptotic theory. Motivated by these observations, we characterize a non-asymptotic oracle tuning parameter that explicitly balances variance reduction against the cost of using unlabeled data and provides theoretical guidance for finite-sample tuning. We further discuss practical implications, including a corrected plug-in estimator. Numerical experiments support the qualitative predictions of the theory.
