THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

PhD Dissertation Proposal Presentation

Jeonghwan Lee
Department of Statistics
The University of Chicago

"Topics in modern statistical learning: Distribution shift and learning with synthetic data"

December 9, 2025, at 2:30 PM
Jones 304, 5747 S. Ellis Ave.

Abstract

his proposal investigates two fundamental challenges in modern statistical learning: (1) learning reliably under distribution shift, and (2) understanding the behavior of AI models trained on recursively generated synthetic data.

(1) Statistical learning under distribution shift: The first part of the proposal develops new non-asymptotic theory for two central problems arising when the training and test distributions differ. We begin with off-policy estimation under adaptive data collection, establishing finite-sample guarantees for a broad class of augmented inverse propensity weighting (AIPW) estimators. These guarantees hinge on a sequentially weighted estimation error of the treatment effect. To control this estimation error, we introduce a general reduction scheme to produce a sequence of estimates that minimizes the sequentially weighted estimation error via online non-parametric regression. Complementary instance-dependent local minimax lower bounds demonstrate the optimality of the proposed methods. We then study covariate shift adaptation, providing structure-agnostic finite-sample learning bounds for doubly-robust (DR) estimators. In parametric settings, we show that DR estimators achieve a fast O(1/n)-rate of convergence of the excess target risk, remarkably independent of the statistical accuracies of the pilot estimates.

(2) Training AI models on synthetic data: The second part of this proposal focuses on model collapse, a phenomenon in which training generative models on their own synthetic outputs leads to progressive performance degradation. Two workflows are typically considered: the discard-workflow, where each generation trains solely on synthetic data from the previous model, and the augment-workflow, which continually accumulates all real data while adding synthetic data

For information about building access for persons with disabilities, please contact Keisha Prowoznik at 773.702-0541 or send an email to kprowoznik@uchicago.edu. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/info/statseminars.

over generations. While the discard-workflow generally induces collapse in parameter estimation, the augment-workflow provably avoids it across linear models, exponential families, and iterative maximum likelihood estimation. This proposal aims to these insights to binary classification. We investigate two key questions: (a) Does the augment-workflow effectively avoid the model collapse for classification tasks? (b) Does the augment-workflow effectively avoid the model collapse? The final part of the proposal presents some theoretical and empirical findings to address these two questions.

For information about building access for persons with disabilities, please contact Keisha Prowoznik at 773.702-0541 or send an email to kprowoznik@uchicago.edu. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/info/statseminars.