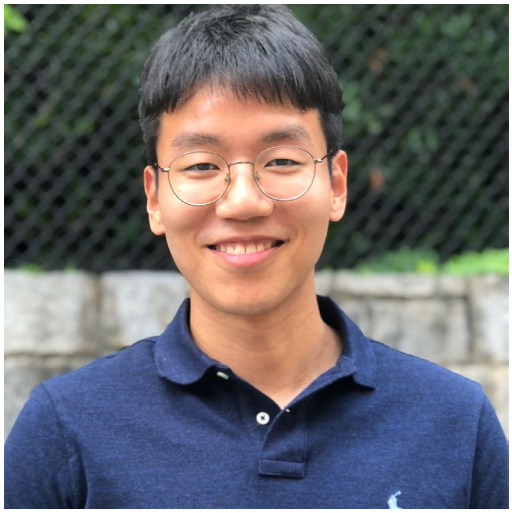
RUTGERS–NEW BRUNSWICK
School of Arts and Sciences

Department of Statistics

Department of Statistics Seminar Series



Heejune Sheen
Department of Statistics and Data Science
Yale University

Statistical Analysis of Synthetic Data with Verification for Fine-Tuning

Wednesday, September 9th, 2026
12:10-1:30 PM
Hill Center - 552

Zoom Meeting: Meeting ID: 926 0900 0053
Password: 565095

https://rutgers.zoom.us/j/92609000053?pwd=8GvwJfLzSeJSwacBNcPq7yYk6oYfdg.1

Light refreshments will be served in Hill Center 452 at 11:40am

Abstract:

Training on self-generated synthetic data has become increasingly common, yet its theoretical understanding remains limited. We study verified self-training in a linear regression model, where a pre-trained estimator generates synthetic covariate-response pairs, an oracle verifier retains pairs whose synthetic responses differ from the corresponding true downstream regression values by at most a verification threshold, and post-training refits an active subset of coordinates while fixing the remaining coordinates at their pre-trained values. We derive nonasymptotic risk bounds in four regimes determined by the relative magnitudes of the verification threshold, generation-noise scale, and pre-training error scale. These bounds characterize how verification affects prediction error across the four regimes. We also quantify the effects of generation noise, the effective accepted sample size, and the covariance structure of the accepted covariates on prediction risk.

Bio:

Heejune Sheen is a sixth-year Ph.D. student in the Department of Statistics and Data Science at Yale University, co-advised by Professors Harrison Zhou and Zhuoran Yang. His research interests include the theoretical foundations of large language models and empirical research on their behavior and capabilities.