To split or not to split that is the question: From cross validation to debiased machine learning

Morgane Austern
Harvard, United States
morgane.austern@gmail.com

Data splitting is an ubiquitous method in statistics with examples ranging from cross validation to cross-fitting. However, despite its prevalence, theoretical guidance regarding its use is still lacking. In this talk we will explore two examples and establish an asymptotic theory for it. In the first part of this talk, we study the cross-validation method, a ubiquitous method for risk estimation, and establish its asymptotic properties for a large class of models and with an arbitrary number of folds. Under stability conditions, we establish a central limit theorem and Berry-Esseen bounds for the cross-validated risk, which enable us to compute asymptotically accurate confidence intervals. Using our results, we study the statistical speed-up offered by cross validation compared to a train-test split procedure. We reveal some surprising behavior of the cross-validated risk and establish the statistically optimal choice for the number of folds. In the second part of this talk, we study the role of cross fitting in the generalized method of moments with moments that also depend on some auxiliary functions. Recent lines of work show how one can use generic machine learning estimators for these auxiliary problems, while maintaining asymptotic normality and root-n consistency of the target parameter of interest. The literature typically requires that these auxiliary problems are fitted on a separate sample or in a cross-fitting manner. We show that when these auxiliary estimation algorithms satisfy natural leave-one-out stability properties, then sample splitting is not required. This allows for sample re-use, which can be beneficial in moderately sized sample regimes

Joint work with Wenda Zhou (NYU and Flatiron), Jane Chen (Harvard) and Vasilis Syrgkanis (Stanford).