Title: "Online Data Collection for Efficient Semiparametric Inference"

Abstract:

While many works have studied data fusion, they typically assume that the agent has already collected the requisite datasets. This perspective does not account for the difficult data collection decisions that pervade the practice of statistical estimation. A practitioner must determine the available data sources, their costs, and decide how many samples to collect. Moreover, data acquisition is often an iterative process: the data collected at a given time informs the decisions in the future steps. In our setup, the agent has access to multiple data sources, and under budget constraints, they must sequentially decide which data source to query to efficiently estimate a target parameter. We formalize this task using Online Moment Selection, a semi- parametric framework that applies to any parameter identified by a set of moment conditions. We propose two online data collection policies, Explore-then-Commit and Explore-then-Greedy, that use the estimates of the moments at a given time to optimally allocate the remaining budget in the future steps. We prove that both policies achieve zero regret (assessed by asymptotic MSE) relative to an oracle policy. We empirically validate our methods on both synthetic and real-world causal effect estimation tasks, demonstrating that the online data collection policies outperform their fixed counterparts.