W20: Learning from Logged Implicit Exploration Data

Alexander L. Strehl, John Langford, Lihong Li, Sham Kakade
Offline Policy Learning Problem

Can we find a better policy given click logs from each policy?
Formalization

• Given logged data from a production system $\pi$, evaluate the performance of a new and different policy.
• Our training examples are of the form (Input, Action, Reward), where actions were chosen by $\pi$, and rewards for unknown actions are not revealed.
• This is different (and harder) than supervised learning or contextual karmed bandits.
Solution, Results, Experiments

• Our **offline policy estimation** takes two steps:
  – Model the production system by conditional probability estimation.
  – Using this model, evaluate a new policy via importance sampling to correct for bias of production system.

• We prove that it converges to a biased estimate and provide sample complexity bounds for using it for policy optimization.

• We present experiment results on its application to **Yahoo!’s Front Page** and **Online Advertising** products.