A Smoothed Analysis of the Greedy Algorithm for the Linear Contextual Bandit Problem

Steven Wu
Microsoft Research
10/18/2017 at 11:00 am
CoRE A 301

Abstract

Bandit learning models the common setting when the decisions of an algorithm feed back into its training data, and it cannot observe counter-factuals. These settings include criminal recidivism prediction (would an inmate have committed another crime had he been released?), lending (would an applicant denied a loan have paid it back?), and drug trials (how would a patient have responded to a different drug?) The main tension in bandit learning is balancing exploration with exploitation. However, exploration, which explicitly sacrifices welfare today in exchange for potential future benefit, can be distasteful when decisions pertain to individuals. For example, it might be considered unethical to give drug A to a patient while knowing drug B often cures their illness merely to learn about drug A's efficacy. In other settings (like predictive policing), a lack of exploration has been blamed for perpetuating unfairness. This motivates studying the performance of the greedy algorithm, which only exploits and never explores. Although this is not a no-regret algorithm, we show that when adversarially selected contexts are perturbed by a small amount of Gaussian noise, it recovers strong no-regret bounds.