JDS

Anytime-valid off-policy inference for contextual bandits

IAN WAUDBY-SMITH, Dept. of Statistics & Data Science, Carnegie Mellon University, USA
LILI WU, Microsoft, USA
AADITYA RAMDAS, Dept. of Statistics & Data Science, Carnegie Mellon University, USA
NIKOS KARAMPATZIAKIS, Microsoft, USA
PAUL MINEIRO, Microsoft, USA

PROBLEM STATEMENT

Contextual bandits and adaptive experimentation are becoming increasingly commonplace in the tech industry and health sciences. The problem setting consists of (at each time t) observing a context X_t, taking a randomized action A_t drawn from a policy h_t, and observing a stochastic reward R_t. The goal of *off-policy evaluation* is to provide an answer to the counterfactual question: "What would the average reward (or its distribution) have been like, had we used a different policy π instead of h?".

METHODS

Several solutions to this problem have been provided over the years in the form of importance weighted and doubly robust estimators alongside confidence intervals that are typically provided via versions of the central limit theorem or finite-sample concentration inequalities. Our paper takes a different approach by deriving *confidence sequences* — sequences of confidence intervals that are uniformly valid over time — for various off-policy functionals. On a more fundamental level, our approach uses nonnegative (super)martingales that we tailor to different off-policy inference problems.

RESULTS

Our confidence sequences enable "anytime-valid inference" in the sense that the data can be continuously monitored *while* the online contextual bandit algorithm is running, the experiment can be adaptively stopped or extended, and yet the confidence intervals are valid at the eventual stopping time. We not only focus on estimating average off-policy rewards (that is, mean rewards under a target policy π), but also arbitrary functionals of the distribution of rewards under π. The latter is enabled by constructing confidence sequences for the cumulative distribution function of the off-policy reward distribution.

SIGNIFICANCE

All of our results satisfy four desirable properties: they are (1) nonasymptotic, meaning they are valid in finite samples, (2) nonparametric, not imposing any parametric assumptions on the distributions of contexts, actions or rewards, (3) valid under adaptive sampling, meaning the logging policies $(h_t)_{t=1}^{\infty}$ can be chosen based on previous data (e.g. from online learning algorithm), (4) anytime-valid, uniformly valid for all sample sizes and hence at arbitrary data-dependent stopping times, and (5) do not need to know the maximal importance weight $w_{\max} := \operatorname{ess\,sup}_{x,a,t} \pi(a \mid x)/h_t(a \mid x)$, which itself need not be finite. As far as we know, no prior work satisfying properties (1)–(5) exists even in the fixed-n (not anytime-valid) regime. Nevertheless, we provide solutions for both fixed-n and anytime-valid settings with excellent empirical performance in both.

Keywords: Confidence sequences, sequential testing, doubly robust, importance weighting

Authors' addresses: Ian Waudby-Smith, ianws@cmu.edu, Dept. of Statistics & Data Science, Carnegie Mellon University, USA; Lili Wu, liliwu@microsoft.com, Microsoft, USA; Aaditya Ramdas, aramdas@cmu.edu, Dept. of Statistics & Data Science, Carnegie Mellon University, USA; Nikos Karampatziakis, nikosk@microsoft.com, Microsoft, USA; Paul Mineiro, pmineiro@microsoft.com, Microsoft, USA.

Research Article