Causal Bandits

Author's Name

Editor: Editor's name

Abstract

The abstract
Keywords: Causality, Bandits, Regret Bounds

1. Introduction

The aim of this paper is to demonstrate theoretically the utility of having causal knowledge when solving multi-armed bandit (MAB) problems.

We do this by:

1. Providing some motivating examples as to why it is reasonable to expect to have some knowledge of causal structures when solving a sequential decision making problem such as MAB problems.

2. Formalising what we mean by "causal information" for a simple class of stochastic multi-armed bandit (MAB) problems.

   We can formalize a connection between causal graphs and fairly general stochastic MAB problems (section 4 of my TPR)

3. Demonstrating that knowledge of causal structure can induce interesting dependencies between the rewards of different MAB arms.

   Described in section 4 of my TPR

4. Introducing a specific causal MAB problem demonstrating this structure and proposing a simple explore-then-exploit algorithm that makes use of the causal information and yields significant improvements over the corresponding lower bounds for algorithms that do not use the causal information.

   Given the causal structure in figure 1 we get an upper bound $R(T) = m^{1/3}T^{2/3}log(KT)^{1/3}$ (section 1 of observe-then-best, section 1.1 if $\boldsymbol{q}$ is not known in advance). Where m can be interpreted as the number of arms corresponding to variable settings that occur only rarely if not explicitly set. The corresponding lower bound for algorithms that do not leverage causal structure is $\sqrt{KT}$ (Auer1995). This demonstrates the causal algorithm will do significantly better where there are large number of arms, provided m is not too large.

5. Deriving a matching lower bound for this problem for algorithms that make use of the causal structure.

We currently have (lowerbounds-auer section 3) a lower bound of $m^{1/3}T^{2/3}$ for algorithms that use information provided by purely observing. Ie the structure introduced by equation 6 in my TPR. This bound does not apply to algorithms using information in the form provided by equation 7 in my TPR. (ie we have not yet shown that we are fully utilizing the causal information available).

6. Demonstrating we can also use the causal structure to gain significant improvement for the closely related best-arm identification problem.

   We obtain simple regret $R_s(T) = \sqrt{m/T}$ (observe-then-best section 1.1) versus $R_s(T) = \sqrt{K/T}$ for the non-causal version of the problem (Bubeck et al 2009)

We identify a problem-dependent constant that appears in the upper and lower bounds that can be roughly interpreted as a measure of how much actions reveal about other actions via the causal structure.

1.1. Related Work

- Bareinboim, Forney & Pearl, *Bandits with unobserved confounders*, NIPS 2015
- Salomon, Audibert & Alaoui, *Lower Bounds and Selectivity of Weak-Consistent Policies in Stochastic Multi-Armed Bandit Problems*, JMLR 2013.
- Alon, Cesa-Bianchi, Dekel, Koren, *Online Learning with Feedback Graphs: Beyond Bandits*, COLT 2015.

Figure 1: Assumed Causal Structure