Value Directed Exploration in Multi-Armed Bandits with Structured Priors

University of New Hampshire

Bence Cserna Marek Petrik Reazul Hasan Russel Wheeler Ruml

Contributions

- A Bayesian Multi Armed Bandit algorithm (ELSV) that utilizes value-function-driven technique in MDP/RL setup.
- Linearly-separable value functions based on UCB and Gittins index that consider expected return & the benefit of exploration to perform n-step lookahead.
- Sub-linear performance guarantee and strong simulation results for problems with structured priors.
- Simple and general, applicable in complex MAB problems.

Introduction

- Repeatedly choose one of N arms: $\mathcal{A} = \{a_1, \ldots, a_N\}$ & receive a reward $R_i \in \{0, 1\}$.
- Rewards are distributed according to a *Bernoulli distribution* with a mean μ_a which is not known in advance.
- Use *linearly-separable value functions* that considers expected return & the benefit of exploration to perform n-step lookahead.
- Aim is to maximize cumulative reward & minimize Bayesian regret:

$$\text{BayesRegret}(\pi, T) = \mathbb{E}_\mu \left[\text{Regret}(T, \mu) \right] .$$

Problem Formulation

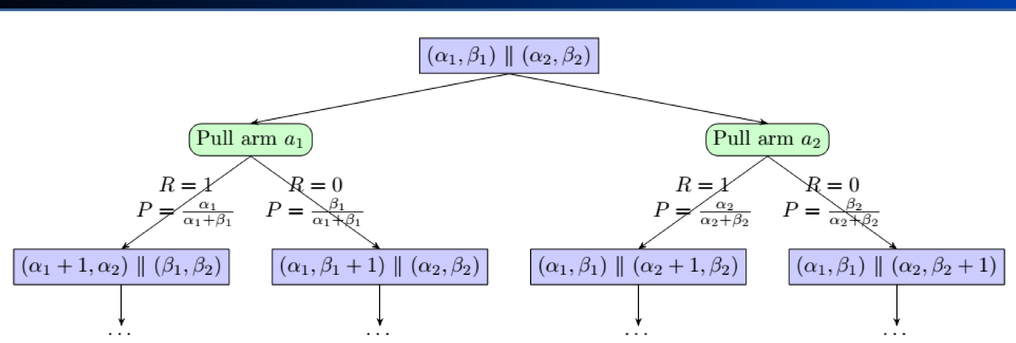


Figure 1: A single transition of the Bernoulli multi-armed bandit problem.

- Bayesian MAB problem is modeled as a Markov Decision Process (MDP).
- Any state $s \in \mathcal{S}$ can be represented with Beta distribution $\text{Beta}(\alpha, \beta)$.
- The actions $\mathcal{A}$ in this MDP are pulls of arms of the bandit.
- The rewards received in transitions are 1 or 0 based on success or failure respectively.

Constrained Problem Setup

- Optimize the level of *personalized* discount offers to customers in an e-commerce setting.
- Arm $i+1$ represents a *smaller* discount than arm i: $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_N$.
- The rewards satisfy: $r_1 < r_2 < \ldots < r_N$.

Index Policy

$$a^* \leftarrow \arg\max_{a \in \mathcal{A}} z_t(s_t^i, a_i)$$

- Index value z_t^{UCB} can be defined as:

$$z_t^{\text{UCB}}(s_t^i, a_i) = r(s_t^i, a_i) + \sqrt{\frac{\alpha \log t}{T_i(s_t^i)}}$$

Approach

- A linearly-separable value function for components $v_t^i : \mathcal{S}_t^i \to \mathbb{R}$ s.t. $t \in \mathcal{T}$ & $s_t \in \mathcal{S}_t$, holds

$$v_t(s_t) = \sum_{a \in \mathcal{A}} v_t^i(s_t^i) .$$

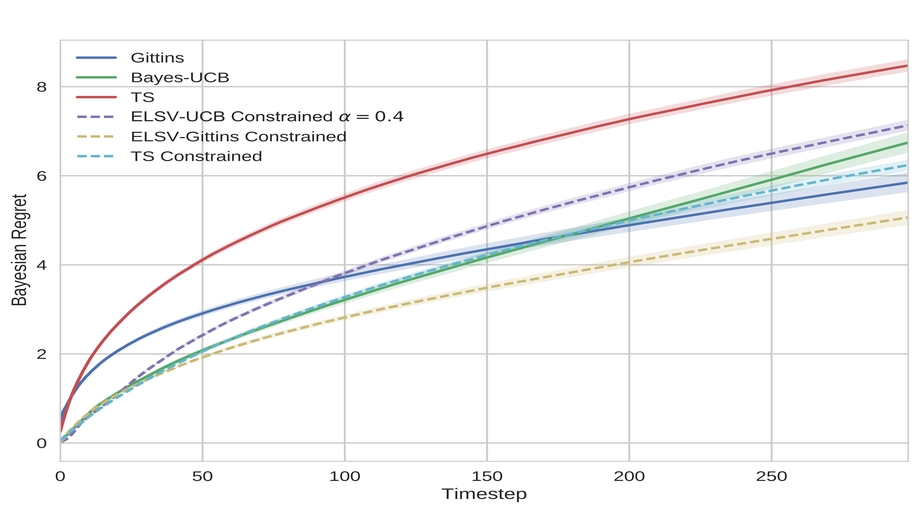
- With an arbitrary bandit index function z_t^i, corresponding exploration bonus function b_t^i, each component v_t^i must satisfy:

$$\mathbb{E}\left[v_{t+1}^i(S_{\tau+1}^i) \right] - v_{t+1}^i(s_\tau^i) = b_t(s_\tau^i, a_i) .$$

- Compute linearly separable value functions:

$$v_t^i(\alpha, \beta) \leftarrow p \cdot v_t^i(\alpha + 1, \beta) + q \cdot v_t^i(\alpha, \beta + 1) - b_t(\alpha, \beta)$$

Results



Above figure shows regret of our Exploration via Linearly Separable Value Functions (ELSV) algorithm with 1-step lookahead, compared with other algorithms on the constrained bandit problem (discount levels of $20\%, 10\%, 0\%$ for arms $1, 2, 3$ respectively) with 3 arms and averaged over 5300 problem instances and with 95% confidence intervals.

Conclusion

- A new linearly separable value function that takes expected return & benefit of exploration into consideration is proposed.
- Can be used in concert with MDP & RL.

Future Research

- ELSV is a good first step in developing more sophisticated value-directed methods.
- We hope to further extend this approach to contextual bandits problems.