Pure Exploration in Episodic Fixed-Horizon Markov Decision Processes

(Extended Abstract)

Sudeep Raja Putta
Conduent Labs India
Bangalore, India
sudeepraja94@gmail.com

Theja Tulabandhula
University of Illinois Chicago
Chicago, USA
tt@theja.org

ABSTRACT

Multi-Armed Bandit (MAB) problems can be naturally extended to Markov Decision Processes (MDP). We extend the Best Arm Identification problem to episodic fixed-horizon MDPs. Here, the goal of an agent interacting with the MDP is to reach a high confidence on the optimal policy in as few episodes as possible. We propose Posterior Sampling for Pure Exploration (PSPE), a Bayesian algorithm for pure exploration in MDPs. We empirically show that PSPE achieves deep exploration and the number of episodes required by PSPE for reaching a fixed confidence value is exponentially lower than random exploration and lower than reward maximizing algorithms such as Posterior Sampling for Reinforcement Learning (PSRL).

Keywords

Reinforcement Learning, Pure Exploration, Multi-arm Bandit, Markov Decision Process

1. INTRODUCTION

In Pure Exploration (PE), the agent's goal is to explore the MDP such that it reaches a high confidence on the optimal policy in as few episodes as possible, i.e., maximize the probability of following an optimal policy. This is different from the classical Reinforcement Learning (RL) problem, where the goal is to maximize the rewards collected.

We consider episodic fixed-horizon MDPs with a finite states and actions. We are interested in model based Bayesian algorithms, where the agent maintains a prior distribution on the parameters of the MDP and computes posteriors based on the rewards and transitions observed. The algorithm uses these posteriors to pick actions according to the goal of the agent.

In this paper, we propose an algorithm for the PE problem in stochastic episodic fixed-horizon MDPs called PSPE. The following table captures our contribution.

	RL	PE
MAB	TS [4]	PTS[3]
MDP	PSRL [1]	PSPE

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

The Thompson Sampling (TS) [4] algorithm is for maximizing the cumulative rewards in MABs. The Pure exploration Thompson Sampling (PTS)[3] algorithm modifies TS by adding a re-sampling step . TS is not suitable for PE as it pulls the estimated best arm almost all the time, and it takes a very long time to ascertain that none of the other arms offer better rewards. The re-sampling step prevents pulling the estimated best arm too often and helps it in achieving higher confidence in lesser number of arm pulls. PSRL[1] extends TS to the complete RL problem on episodic fixed-horizon MDPs. We modify PSRL for PE by adding a re-sampling step.

2. EPISODIC FIXED HORIZON MDP

An episodic fixed horizon MDP M is given by the tuple $\langle \mathcal{S}, \mathcal{A}, R, P, H, \rho \rangle$. Here $\mathcal{S} = \{1, ..., S\}$ and $\mathcal{A} = \{1, ..., A\}$ are set of states and actions respectively. The agent interacts with the MDP in episodes of length H. ρ is the initial state distribution. In each step of an episode, the agent observes a state s_h and performs an action a_h. It receives a reward $r_h \sim R(s_h, a_h)$ and transitions to a new state $s_{h+1} \sim P(s_h, a_h)$. The average reward received for a particular state-action is $\bar{R}(s, a) = \mathbb{E}[r|r \sim R(s, a)]$.

A policy π is a mapping from $\mathcal{S}$ and time step $h = 1, ..., H$ to A. The value of a state s at step h under a policy π is $V_\pi(s, h) = \mathbb{E}[\sum_{i=h}^{H} \bar{R}(s_i, \pi(s_i, i))]$. π^* is an optimal policy for the MDP if $\pi^* \in \arg\max_\pi V_\pi(s, h)$ for all $s \in \mathcal{S}$ and $h = 1, ..., H$. For a MDP M, let Π_M be the set of optimal policies.

3. POSTERIOR SAMPLING FOR PE

PSPE modifies PSRL by adding a re-sampling step. This prevents it from following an estimated optimal policy too frequently. The algorithm depends on a parameter β, where $0 < \beta < 1$, which controls how often an optimal policy of a sampled MDP is followed. Let f be the prior density over the MDPs and $\mathcal{H}_t$ be the history of episodes seen until $t-1$. Algorithm 1 describes PSPE.

3.1 Computing the Confidence

Let M^* be the true underlying MDP and let Π^* be its set of optimal policies. The confidence of the agent α_t at episode t is the probability of sampling a MDP M_t and following one of its optimal policies π_t such that $\pi_t \in \Pi^*$.

The confidence of a set of policies Π is the probability of sampling a MDP M and following a policy from Π_M such that it is also in Π. Let $x_\Pi(M)$ denote the probability of

Algorithm 1 PSPE

$\mathcal{H}_1 = \{\}, t = 1$
for $t = 1, 2, ...$ **do**
  Sample $M_t \sim f(\cdot|\mathcal{H}_t)$
  Sample $B \sim Bernoulli(\beta)$
  if $B = 1$ **then**
    Choose a policy π_t at random from Π_{M_t}
  else
    repeat
      Re-sample $\widetilde{M}_t \sim f(\cdot|\mathcal{H}_t)$
    until $\Pi_{\widetilde{M}_t} - \Pi_{M_t} \neq \emptyset$
    Choose a policy π_t at random from $\Pi_{\widetilde{M}_t} - \Pi_{M_t}$
  end if
  Observe initial state $s_{1,t}$
  for $h = 1, ..., H$ **do**
    Perform action $a_{h,t} = \pi_t(s_{h,t}, h)$
    Observe reward $r_{h,t}$ and next state $s_{h+1,t}$
  end for
  $\mathcal{H}_{t+1} = \mathcal{H}_t \cup \{(s_{h,t}, a_{h,t}, r_{h,t}, s_{h+1,t}) | h = 1..H\}$
end for

picking a policy from Π_M which is also in Π.

$$x_\Pi(M) = \frac{|\Pi_M \cap \Pi|}{|\Pi_M|}$$

The confidence of Π, denoted by α_Π can be expressed as the expectation of $x_\Pi(M)$ computed over the current posterior distribution of MDPs.

$$\alpha_\Pi = \mathbb{E}_M[x_\Pi(M)] = \int_{M \in \mathcal{M}} x_\Pi(M) f(M|\mathcal{H}) dM$$

Due to the Law of Large Numbers, this expectation is the same as this summation in the limit.

$$\alpha_\Pi = \lim_{n \to \infty} \frac{\sum_{j=1}^{n} x_\Pi(M_j)}{n} \quad \text{where } M_j \sim f(\cdot|\mathcal{H})$$

At episode t, the confidence of the agent is $\alpha_t = \alpha_{\Pi^*}$. Our algorithm itself does not require the confidence value for its operation.

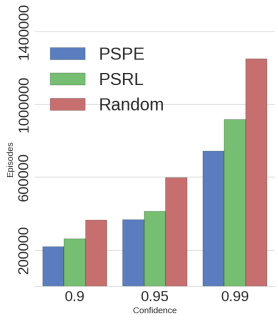
4. EMPIRICAL EVALUATION

We compare the performance of PSPE with PSRL and random exploration, by measuring the number of episodes required to reach a high confidence value. We use a uniform Dirichlet prior for the transition probabilities and a Gaussian prior ($\mathcal{N}(0, 1)$) for reward distribution. α_{Π^*} is calculated drawing 10000 i.i.d samples from the posterior. The experiment tracks the first time when the confidence value exceeds a fixed confidence of 0.90, 0.95 and 0.99. All the results are averaged across 50 trials. We choose $\beta = 1/2$ in PSPE.

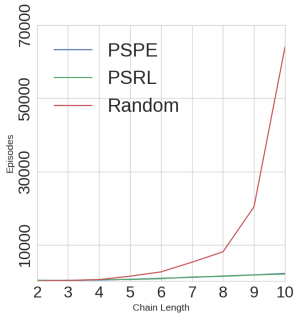
We generate a random MDP having 3 states, 3 actions and $H = 3$. Figure (a) displays the average number of episodes required to reach each fixed confidence. PSPE reaches a high confidence in lesser number of episodes than both PSRL and random exploration. The policy gap is small, resulting in requiring a large number of episodes(400000) to reach a high confidence.

Stochastic Chains (Figure (c)) proposed by Osband & Van Roy [2], is a family of MDPs which consist of chain states. There are two actions, left and right. The left action is deterministic, but the right action result in going right with probability $1 - 1/N$ or going left with probability $1/N$. The only two rewards in this MDP are obtained by choosing left in state 1 and choosing right in state N. These rewards are drawn from a normal distribution with unit variance. Each episode is of length $H = N$. The agent begins each episode at state 1. The optimal policy is to go right at every step to receive an expected reward of $(1 - \frac{1}{N})^{N-1}$.

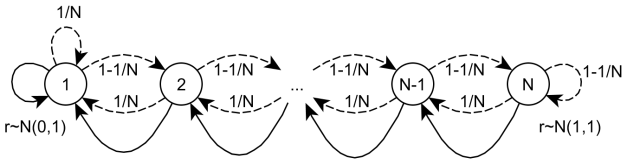
We consider stochastic chains of lengths 2 to 10 and measure the number of episodes required to reach a confidence of 0.95. For PSPE and PSRL, this number is practically the same and grows very slowly. For Random exploration, it grows exponentially, as seen in Figure (b). PSRL and PSPE are able to achieve "Deep Exploration".



(a) Episodes vs Confidence in Random MDP



(b) Episodes for 0.95 confidence vs Chain length



(c) Stochastic Chain MDP

5. CONCLUSION

We present Posterior Sampling for Pure Explorations as a Bayesian algorithm for the problem of Pure exploration under a fixed confidence setting in episodic fixed-horizon MDPs. We demonstrate that PSPE can achieve a high confidence in lesser number of episodes than PSRL or random exploration.

REFERENCES

[1] I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[2] I. Osband and B. Van Roy. Why is posterior sampling better than optimism for reinforcement learning. *arXiv preprint arXiv:1607.00215*, 2016.

[3] D. Russo. Simple bayesian algorithms for best arm identification. *Twenty ninth Annual Conference on Learning Theory*, pages 1417–1418, 2016.

[4] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.