Reinforcement Learning for Parameterized Markov Decision Processes using Posterior Sampling

Mohammad Zhalechian

April, 2021

Overview

1. Introduction
2. Literature Review
3. Algorithm Description
4. Performance Analysis
5. Numerical Results

Introduction

- Consider a reinforcement learning (RL) problem in which an agent interacts with an unknown environment and the aim is to minimize the cost.
- Finding the optimal policy in an unknown environment brings the difficulty of dealing with the exploration-exploitation trade-off.
- Optimism in the face of uncertainty (OFU) and posterior sampling (PS) are two popular methods to deal with this trade-off.

Introduction

- **OFU method:** Agent constructs confidence sets for the unknown model parameters and finds the optimal policy based on the optimistic estimates.
- **PS method:** Agent imposes prior distributions over the unknown model parameters and finds the optimal policy based on the estimates of the unknown model parameters sampled from the prior distributions.
- **Advantage of PS over OFU**: In each episode, PS-based reinforcement learning (PSRL) algorithms need to solve a sampled MDP rather than solving all MDPs that lie within the confidence sets (computationally more efficient).

Literature on PSRL Algorithms

- **Osband et al. 2013:** Proposed a PSRL algorithm for an *episodic problem* with fixed-length episodes that admits a Bayesian regret of $\widetilde{O}(\tau S\sqrt{AT})$, where T is time, τ is the episode length and S and A are the sizes of the state and action spaces.

  Many real-world problems are non-episodic with a continuing and non-resetting nature (e.g., sequential recommendations).

Literature on PSRL Algorithms

- **Ouyang et al. 2017b:** Proposed a PSRL algorithm for a *non-episodic problem* that admits a Bayesian regret of $\widetilde{O}(HS\sqrt{AT})$, where S and A are the sizes of the state and action spaces, and H is the bound of the span.
- **Agrawal an Jia, 2017:** Proposed a PSRL algorithm for a *non-episodic problem* that admits a high-probability regret of $\widetilde{O}(D\sqrt{SAT})$ for any communicating MDP with S states, A actions and diameter D.

Neither of them use generalization. That is, they learn separate parameters corresponding to each state-action pair.

In such a non-parametric case, an observed feedback corresponding to a state-action pair does not help to improve the estimations for other state-action pairs.

Literature on PSRL Algorithms

- **Theocharous et al. 2018:** Considered a parametric setting in which the structure of the MDP can be determined with an scalar parameter. They Proposed a PSRL algorithm, called DS-PSRL, for a *non-episodic problem* that admits a Bayesian regret bound of $\widetilde{O}(C\sqrt{C'T})$, which does not depend on the sizes of A and S.

We will discuss this algorithm and the proof sketch.

High Level Description of DS-PSRL

- Deterministic-schedule PRSL (DS-PSRL) Algorithm is episodic
- Actual experience is one long trajectory:

$$(s_1, a_1, r_1), (s_2, s_2, r_2), \ldots, (s_T, a_T, r_T)$$

  generated during interaction with a parametrized MDP $(\mathcal{S}, \mathcal{A}, \ell, P^{\theta^*})$, where the state and action sets $\mathcal{S}$ and $\mathcal{A}$ can be infinite or even continuous.
- The loss function $\ell(s, a)$ is known and the transition function $P(s^{'}|s, a, \theta^*)$ is unknown.
- The agent starts with a prior belief P_0 on θ^*.

High Level Description of Algorithm

In every episode beginning at time t:

- Sample $\tilde{\theta}_t$ from P_t (posterior distribution on θ^* at time t).
- Find the optimal policy based on $\tilde{\theta}_t$.
- Follow the optimal policy until we switch to another episode.
- Update P_t using the tuples (s_i, a_i, s_{i+1}) obtained by interacting with the MDP during the episode.

Switching Rule: If the length of the current episode is L, the length of the next episode would be $2L$.

This switching rule ensures that the total number of switches is controlled.

Average Loss Criterion

- The long term average loss

$$J(\theta, \pi, s) = \mathbb{E}\left[\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \ell(s_t, a_t)\middle| s_1 = s\right]$$

- The goal is to develop an algorithm that adaptively selects an action a_t at every time step t based on the prior information and past observations to minimize the long term average loss.

Optimality equation and Bayesian regret

- The optimal $J(\theta^*, \pi^*)$ is well defined and independent of the starting state under the mild weakly communicating assumption.
- The optimal policy π^* with state-independent $J(\theta^*, \pi^*)$ satisfies:

$$J(\theta^*, \pi^*) + h(s_t) = \ell(s_t, a_t) + \int_{\mathcal{S}} P(s|s_t, a_t, \theta^*) h(s) \quad \forall\, s \in \mathcal{S},$$

  where h is the bias function and $s_{t+1} \sim P(.|s_t, a_t, \theta^*)$.
- The Bayesian regret R_T can be computed as:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \Big(\ell(s_t, a_t) - J(\theta^*, \pi^*)\Big)\right].$$

Policy $\pi^*(s, \tilde{\theta}_t)$

- Within each episode k, the near-optimal action at time t can be calculated as:
$$a_t \leftarrow \pi^*(s, \tilde{\theta}_t),$$
where $\tilde{\theta}_t$ is the same for all time steps within the episode k.
- The policy $\pi^*(s_t, \tilde{\theta}_t)$ satisfies:
$$J(\tilde{\theta}_t) + h_t(s_t) = \ell(s_t, a_t) + \mathbb{E}\Big[h_t(\tilde{s}_{t+1})|\mathcal{F}_t, \tilde{\theta}_t\Big].$$
where $\mathcal{F}_t$ is a filteration containing historic information until time t and $\tilde{s}_{t+1} \sim P(.|s_t, a_t, \tilde{\theta}_t)$.
- We assume that there exists $H > 0$ such that $h_t(s) \in [0, H]$.

Assumptions

1. **(Lipschitz Dynamics)** The exists a constant C such that for any state s and action a and parameters $\theta, \theta' \in \Theta \subset \mathbb{R}$, we have:

$$\left\| P(.|s, a, \theta) - P(.|s, a, \theta') \right\|_1 \leq C|\theta - \theta'|$$

   This implies that dynamics are parameterized by a scalar parameter and satisfy a smoothness condition.

2. **(Concentrating Posterior)** Let N_k be one plus the number of steps in the first k episodes. Let $\tilde{\theta}_k$ be sampled from the posterior at the current episode k. Then, there exists a constant C' such that:

$$\max_k \mathbb{E}\left[N_{k-1} \left| \theta^* - \tilde{\theta}_k \right|^2 \right] \leq C' \log(T)$$

   This implies that the variance of the posterior decreases given more data (i.e., the problem is learnable).

Decomposing the regret term

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \mathbb{E}\left[\Big(\ell(s_t, a_t) - J(\theta^*, \pi^*)\Big)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\Big(\ell(s_t, a_t) - J(\tilde{\theta}_t, \pi^*)\Big)\right] \quad \text{by } \mathbb{E}[g(\theta^*)|\mathcal{F}_t] = \mathbb{E}[g(\tilde{\theta}_t)|\mathcal{F}_t] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[h_t(s_t) - \mathbb{E}\big[h_t(\tilde{s}_{t+1}|\mathcal{F}_t, \tilde{\theta}_t)\big]\right] \quad \text{by optimality equation} \\
&= \sum_{t=1}^{T} \mathbb{E}\left[h_t(s_t) - h_t(\tilde{s}_{t+1})\right] \\
&= \mathbb{E}\underbrace{\left[h_1(s_1) - h_{T+1}(\tilde{s}_{T+1})\right]}_{\leq H \text{ since } h_1(s_1) \leq H} + \sum_{t=1}^{T} \mathbb{E}\left[h_{t+1}(s_{t+1}) - h_t(\tilde{s}_{t+1})\right].
\end{aligned}
$$

Decomposing the regret term

$$
\begin{aligned}
R_T &\leq \sum_{t=1}^{T} \mathbb{E}\left[h_{t+1}(s_{t+1}) - h_t(\tilde{s}_{t+1})\right] + H \\
&= \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[h_{t+1}(s_{t+1}) - h_t(s_{t+1})\right]}_{\text{Term (I)}} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[h_t(s_{t+1}) - h_t(\tilde{s}_{t+1})\right]}_{\text{Term (II)}} + H.
\end{aligned}
$$

Bounding term (I)

- This term is related to sequential changes in $h_{t+1} - h_t$.
- Note that $h_{t+1} - h_t = 0$ as long as $t + 1$ and t belong to the same episode.
- It is already controlled by following the switching rule.
- Let A_t be the event that the algorithm has changed its policy at time t. Thus, we have:

$$\sum_{t=1}^{T} \mathbb{E}\left[h_{t+1}(s_{t+1}) - h_t(s_{t+1})\right] \leq H \sum_{t=1}^{T} \mathbb{E}[1\{A_t\}]$$

$$\sum_{t=1}^{T} 1\{A_t\} \leq \log_2(T) \Rightarrow \sum_{t=1}^{T} \mathbb{E}[1\{A_t\}] \leq \log_2(T).$$

Bounding term (II)

- Let K be the total number of episodes up to time T.
- **Claim 1:** Under Assumption 1, the following can be shown (proof is given in the Appendix):

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(h_t(s_{t+1}) - h_t(\tilde{s}_{t+1})\right)\right] \leq CH\sqrt{T\mathbb{E}\left[\sum_{k=1}^{K} M_k\left|\theta^* - \tilde{\theta}_k\right|^2\right]},$$

  where M_k is the number of steps in the k^{th} episode.
- **Claim 2:** Under Assumption 2, the following can be shown:

$$\mathbb{E}\left[\sum_{k=1}^{K} M_k\left|\theta^* - \tilde{\theta}_k\right|^2\right] \leq 2C' \log^2(T).$$

Bayesian regret bound

$$R_T \leq \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[h_{t+1}(s_{t+1}) - h_t(s_{t+1})\right]}_{\text{Term (I)}} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[h_t(s_{t+1}) - h_t(\tilde{s}_{t+1})\right]}_{\text{Term (II)}} + H$$

$$\leq H \underbrace{\log_2(T)}_{\text{Term (I)}} + \underbrace{CH\sqrt{2C' T \log^2(T)}}_{\text{Term (II)}} + H \; .$$

Theorem

Under Assumptions 1 and 2, the Bayesian regret of DS-PSRL is bounded:

$$R_T = O\left(CH\sqrt{C'T}\log(T)\right)$$

Numerical Results

- RiverSwim problem
- An agent is swimming in a river and can choose to swim either left or right (two actions). The river current is from right to left.
- The MDP consists of $K = 50$ states. The agent starts from the leftmost state ($s = 1$).
- Reward function:

$$r(1, \mathit{left}) = 5;\ \ r(50, \mathit{right}) = 10,000;\ \ r(s, a) = 0 \text{ otherwise.}$$

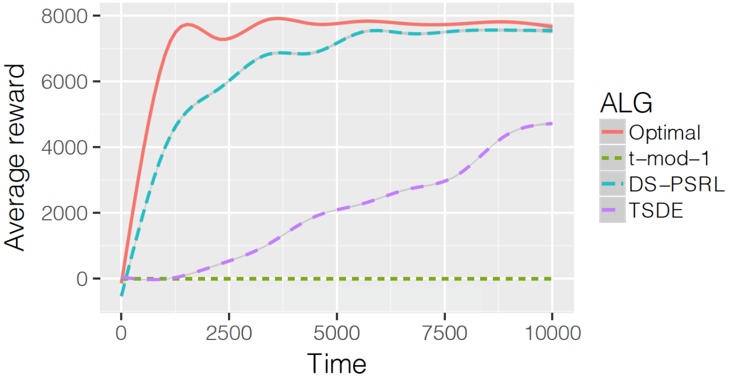
- Transition function:
  - If the agent decides to move left (toward the river current), the agent is always successful.
  - If the agent decides to move right (against the river current), the agent may fail (there is a fail probability).

Benchmarks

- Using simulation, the average reward of different benchmarks are evaluated.
- Four benchmarks are considered:
  - DS-PSRL: The algorithm we discussed today.
  - TSDE: A non-parametric PSRL algorithm proposed by Ouyang et al. 2017.
  - t-mod-1: A policy that switches the action every time-step.
  - Optimal: The optimal policy obtained using the ground truth model.

Average reward

The DS-PSRL algorithm outperforms TSDE and t-mod-1, and it performs comparably to the optimal policy.



Questions?

References

- Osband, I., Russo, D., Van Roy, B. (2013). (More) efficient reinforcement learning via posterior sampling. arXiv preprint arXiv:1306.0940.
- Agrawal, S., Jia, R. (2017). Posterior sampling for reinforcement learning: worst-case regret bounds. arXiv preprint arXiv:1705.07041.
- Ouyang, Y., Gagrani, M., Nayyar, A., Jain, R. (2017). Learning unknown markov decision processes: A thompson sampling approach. arXiv preprint arXiv:1709.04570.
- Theocharous, G., Wen, Z., Abbasi-Yadkori, Y., Vlassis, N. (2017). Posterior sampling for large scale reinforcement learning. arXiv preprint arXiv:1711.07979.

Appendix: Proof of claim 1

First, we have:

$$\sum_{t=1}^{T}\Big(h_t(s_{t+1}) - h_t(\tilde{s}_{t+1})\Big)$$
$$\leq \sqrt{T\sum_{t=1}^{T}\Big(h_t(s_{t+1}) - h_t(\tilde{s}_{t+1})\Big)^2} \quad \text{by Cauchy-Schwarz} .$$

Next, we have:

$$\begin{aligned} h_t(s_{t+1}) - h_t(\tilde{s}_{t+1}) &\leq \Big\| P(.|s_t, a_t, \theta^*) - P(.|s_t, a_t, \tilde{\theta}_t) \Big\|_1 \|h_t\|_\infty \\ &\leq CH|\theta^* - \tilde{\theta}_t| \quad \text{by Assumption 1.} \end{aligned}$$

Appendix: Proof of claim 1

Recall that $\tilde{\theta}_{t+1} = \tilde{\theta}_t$ as long as $t+1$ and t belong to the same episode k. Let T_k be the length of episode k. Accordingly, we have:

$$\begin{aligned}\sum_{t=1}^{T}\Big(h_t(s_{t+1}) - h_t(\tilde{s}_{t+1})\Big) &\leq \sqrt{T\sum_{t=1}^{T}\Big(CH|\theta^* - \tilde{\theta}_t|\Big)^2} \\ &= CH\sqrt{T\sum_{k=1}^{K}\sum_{s=1}^{T_k}|\theta^* - \tilde{\theta}_k|^2} \\ &= CH\sqrt{T\sum_{k=1}^{K}M_k|\theta^* - \tilde{\theta}_k|^2}.\end{aligned}$$

Taking expectation on both sides:

$$\mathbb{E}\left[\sum_{t=1}^{T}\Big(h_t(s_{t+1}) - h_t(\tilde{s}_{t+1})\Big)\right] \leq CH\sqrt{T\mathbb{E}\left[\sum_{k=1}^{K}M_k|\theta^* - \tilde{\theta}_k|^2\right]}.$$