Thompson Sampling (PSRL) under bias span constraint

Pierre Bizeul, Jules Kozolinsky

December 14, 2017

Contents

1 Goal of project

The Exploration-Exploitation trade-off is a fundamental dilemma in on-line reinforcement learning. Algorithms addressing this dilemma with theoretical performance guarantees have been proposed for the discounted, average and finite horizon settings. In the finite horizon setting, the usual criterion of performance is the notion of "regret" as in MAB. One of the current state-of-the-art algorithms PSRL [2] has a regret scaling linearly with the "diameter" of the unknown MDP in the worst-case (the diameter being a measure of how easy it is to navigate between any two states of the MDP). Although it has been shown that the dependency in the diameter is unavoidable in the worst-case, when additional properties on the optimal policy are known beforehand, the regret can, in theory, be drastically improved [1]. Unfortunately, exploiting this additional prior knowledge on the optimal policy requires solving an optimization problem (namely, "planning under bias span constraint") and no algorithm has been derived so far to solve it. We recently started investigating a new Bellman operator converging to the solution of this problem for many MDPs. In this research project, the student(s) is/are expected to:

1. Implement PSRL,
2. Integrate the modified Bellman operator to PSRL,
3. Compare the empirical regret of the two.

2 Introduction

2.1 Notations

Policy

A policy π is defined as $\pi : X \to A$.

Long-term average reward for a stationnary policy

Long-term average reward $\rho_\pi(M)$ of a stationnary policy π is defined as :

$$\rho_\pi(M) = \lim_{T\to\infty} \frac{1}{T} E(T) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r_t$$

where T is the number of time steps and $E(T)$ is the total expected reward after T steps.

(For proof see Proposition 8.1.1 of [3])

Optimal policy

A policy is called *optimal* if it maximizes the long-term average reward :

$$\pi^*(M) = \underset{\pi}{\operatorname{argmax}}\ \rho_\pi(M)$$

We note: $\rho^*(M) = \rho_{\pi^*(M)}(M)$.

2.2 Exploration-exploitation dilemma

Explore the environment to estimate its parameters vs Exploit the estimates to collect reward (See Figure 1).

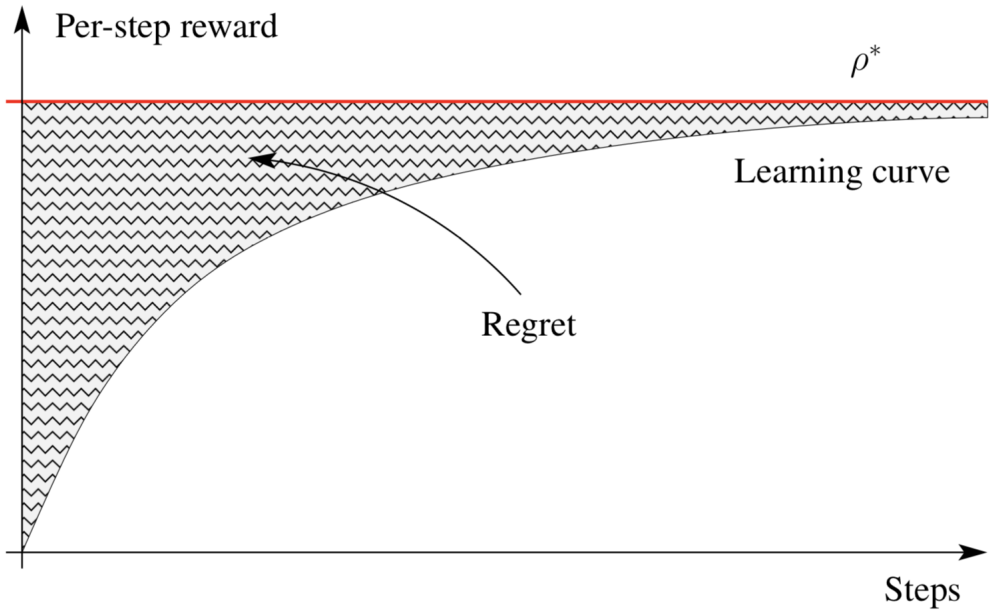

Figure 1: Exploration-Exploitation in RL

Cumulative regret

Cumulative regret $R_\pi(T)$ of a policy π is defined as :

$$R_\pi(T) = T\rho^* - \sum_{t=1}^{T} r_t$$

3 Extended Value iteration

4 Upper-confidence Bound for RL

4.1 Upper-confidence bound

4.2 UCRL2 Algorithm

5 Thompson Sampling for RL

We will use a normal distribution for the rewards and a Dirichlet distribution for the transitions.

References

[1] Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press, 2009.

[2] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[3] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.