(More) Efficient Reinforcement Learning via Posterior Sampling

Ian Osband, Daniel Russo and Benjamin Van Roy - Stanford University

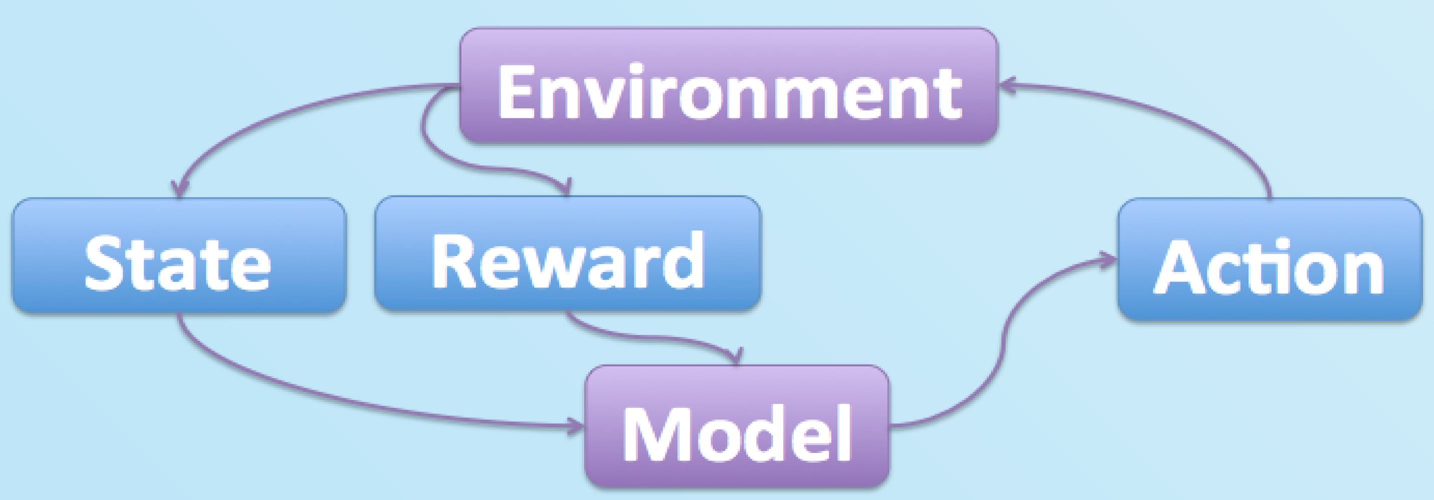
STANFORD ENGINEERING

Introduction

- We study efficient exploration in reinforcement learning.
- Most provably-efficient learning algorithms introduce optimism about poorly understood states and actions.
- Motivated by potential advantages relative to optimistic algorithms, we study an alternative approach: *posterior sampling for reinforcement learning (PSRL).*
- This is the extension of the Thompson sampling algorithm for multi-armed bandit problems to reinforcement learning.
- We establish the first regret bounds for this algorithm.

Problem Formulation

- We study learning to behave near optimally in a fixed but unknown (randomly drawn) MDP M^*.
- Repeated τ-length episodes of interaction with the MDP.
- In episode k, actions selected based on chosen policy μ_k.
- As a result of a_t, the reward r_t and next state s_{t+1} are drawn according to on M^*.
- Goal: Maximize cumulative reward earned.
- Requires managing exploration / exploitation tradeoff.



Algorithm - PSRL

Data: Prior distribution f, t=1

for *episodes* $k = 1, 2, \dots$ **do**

  sample $M_k \sim f(\cdot|H_{t_k})$

  compute $\mu_k = \mu^{M_k}$

  for *timesteps* $j = 1, \dots, \tau$ **do**

    sample and apply $a_t = \mu_k(s_t, j)$

    observe r_t and s_{t+1}

    $t = t + 1$

  end

end

*First introduced by Strens (2002) under the name "Bayesian Dynamic Programming."

Motivation - Advantages of PSRL

✓ Conceptually simple, separates *algorithm* from *analysis:*

- PSRL selects policies according to the probability they are optimal without need for explicit construction of confidence sets.
- UCRL2 bounds error in each (s, a) separately, which allows for worst-case mis-estimation to occur *simultaneously* in every (s, a).
- We believe this will make PSRL more statistically efficient.

✓ The algorithm is computationally efficient:

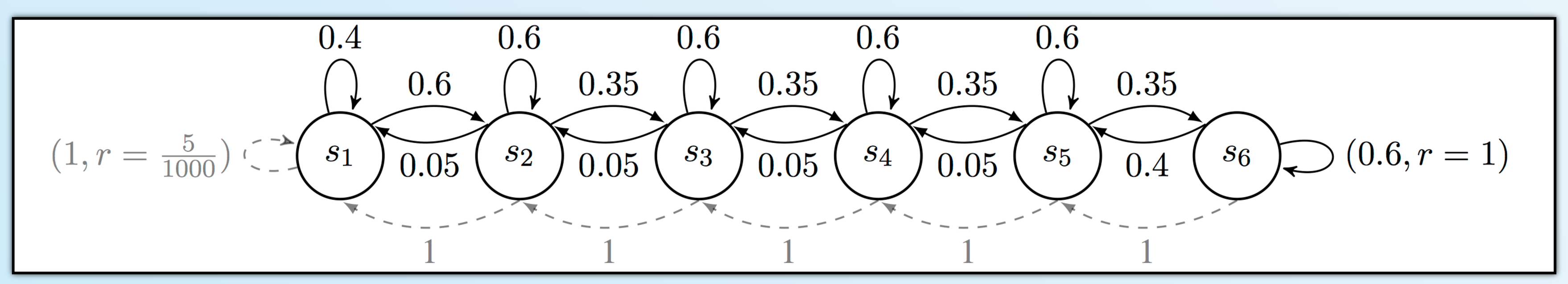
- Optimistic algorithms often require optimizing simultaneously over all policies and a family of plausible MDPs.
- PSRL computes the optimal policy under a *single* sampled MDP.

✓ Can naturally incorporate prior knowledge:

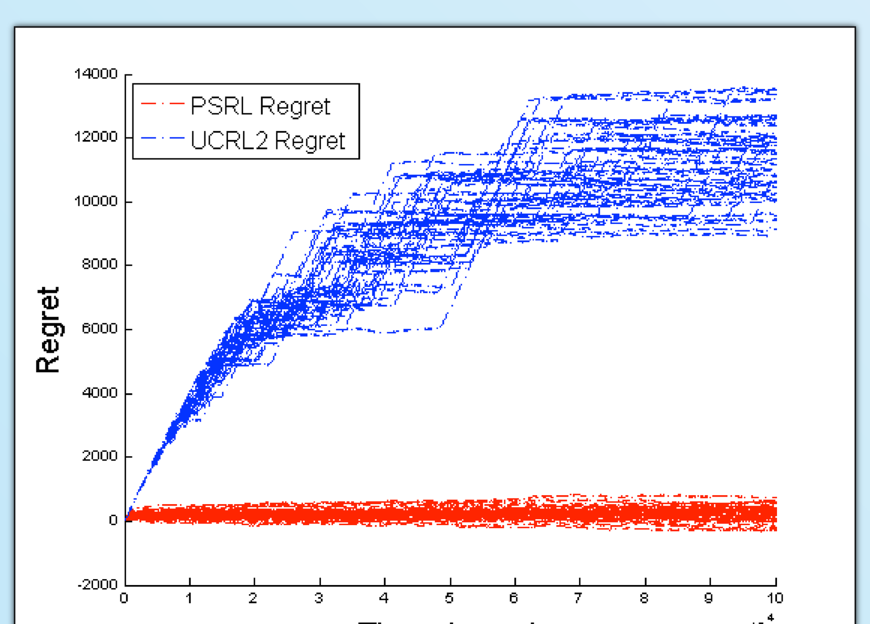
- Crucial for practical applications -Tabula Rasa is often unrealistic.
- Our bounds apply for any prior distribution over finite MDPs.
- PSRL can use *any* environment model, not just finite MDPs.

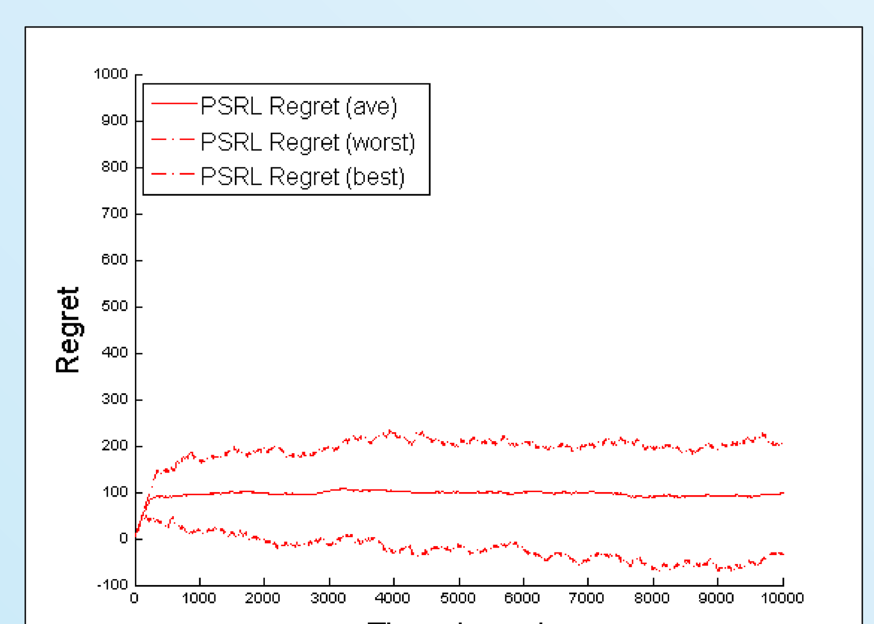
Experimental results

We compared the performance of PSRL to UCRL2 (an optimistic algorithm with similar regret bounds) on several MDP examples.



- We tested the algorithm on *RiverSwim* (an MDP designed to require efficient exploration) as well as random MDPs.
- We saw that PSRL outperforms UCRL2 by large margins.
- PSRL learns quickly even with a mis-specified prior.





Algorithm	Random MDP τ-episodes	Random MDP ∞-horizon	RiverSwim τ-episodes	RiverSwim ∞-horizon
PSRL	1.04×10^4	7.30×10^3	6.88×10^1	1.06×10^2
UCRL2	5.92×10^4	1.13×10^5	1.26×10^3	3.64×10^3

Key lemma - posterior sampling

The true and sampled MDPs are equal in distribution at the start of an episode (when the sample is taken).

$$\mathbb{E}[g(M^*)|H_{t_k}] = \mathbb{E}[g(M_k)|H_{t_k}].$$

Any H_{t_k}-measurable function of these MDPs must therefore be equal in expectation.

Regret bounds

The regret of an algorithm π at time T is the random variable equal to the cumulative reward of the optimal policy minus the realized rewards of π.

Our main result bounds expected regret under the prior:

$$\mathbb{E}\left[\text{Regret}(T, \pi_\tau^{\text{PS}})\right] = O\left(\tau S \sqrt{AT} \log(SAT)\right)$$

- This is not a worst-case MDP bound as per UCRL2 etc.
- But, the two bounds are related via Markov's inequality:

For any α > 0.5 : $\frac{\text{Regret}(T, \pi_\tau^{\text{PS}})}{T^\alpha} \xrightarrow{p} 0.$

- Corresponding results for UCRL2/REGAL deal with non-episodic learning, and replace τ with Diameter/Span.
- In the episodic case, all three give O($\tau S \sqrt{AT}$) bounds.
- These are close to the lower bounds in S,A and T of $\sqrt{SAT}$.

Summary

- PSRL is not just a heuristic but is provably efficient
- First regret bounds for an algorithm not driven by "OFU".
- Regret bounds are competitive with state of the art.
- Bounds allow for an arbitrary prior over finite MDPs.
- Conceptually simple, computationally efficient.
- Statistically efficient, separating *algorithm* from *analysis.*
- Performs well in simulation on benchmark MDPs.

References

Please consult arXiv:1306.0940 for a full list of references.

Simulation code is available at www.stanford.edu/~iosband