On Optimistic versus Randomized Exploration in Reinforcement Learning

Ian Osband
Google Deepmind
iosband@google.com

Benjamin Van Roy
Stanford University
bvr@stanford.edu

Abstract

We discuss the relative merits of optimistic and randomized approaches to exploration in reinforcement learning. Optimistic approaches presented in the literature apply an optimistic boost to the value estimate at each state-action pair and select actions that are greedy with respect to the resulting optimistic value function. Randomized approaches sample from among statistically plausible value functions and select actions that are greedy with respect to the random sample. Prior computational experience suggests that randomized approaches can lead to far more statistically efficient learning. We present two simple analytic examples that elucidate why this is the case. In principle, there should be optimistic approaches that fare well relative to randomized approaches, but that would require intractable computation. Optimistic approaches that have been proposed in the literature sacrifice statistical efficiency for the sake of computational efficiency. Randomized approaches, on the other hand, may enable simultaneous statistical and computational efficiency.

Keywords: Reinforcement learning, exploration, optimism, randomization, Thompson sampling.

Acknowledgements

This work was generously supported by a research grant from Boeing, a Marketing Research Award from Adobe, and a Stanford Graduate Fellowship, courtesy of PACCAR.

1 A Reinforcement Learning Problem

We consider the problem of learning to optimize a random finite-horizon MDP $\mathcal{M}$=($\mathcal{S}$,$\mathcal{A}$,$\mathcal{R}$,$\mathcal{P}$,H,ρ) over episodes of interaction, where $\mathcal{S} = \{1, .., S\}$ is the state space, $\mathcal{A} = \{1, .., A\}$ is the action space, H is the horizon, and ρ is the initial state distribution. At the start of each episode the initial state s_0 is drawn from the distribution ρ. In each time period $t = 0, \cdots, H-1$ within an episode, the agent observes state $s_t \in \mathcal{S}$, selects action $a_t \in \mathcal{A}$, receives a reward $r_{t+1} \sim \mathcal{R}_{t,s,a}$, and transitions to a new state $s_{t+1} \sim \mathcal{P}_{t,s,a}$. What we consider could be referred to as a Bayesian reinforcement learning setting, in which the unknown episodic nonstationary finite-horizon MDP $\mathcal{M}$ is taken to be a random variable.

A policy π is a mapping from a state $s \in \mathcal{S}$ and period $t = 0, .., H-1$ to an action $a \in \mathcal{A}$. For each MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, H, \rho)$ and policy π we define the state-action value function for each period t:

$$Q^{\mathcal{M}}_{\pi,t}(s,a) := \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{\tau=t}^{H-1} \overline{r}^{\mathcal{M}}(s_\tau, a_\tau)\middle| s_t = s, a_t = a\right], \tag{1}$$

where $\overline{r}^{\mathcal{M}}_t(s,a) = \mathbb{E}[r_{t+1}|\mathcal{M}, s_t = s, a_t = a]$. The subscript π indicates that actions over periods $t, \ldots, H-1$ are selected according to the policy π. Let $V^{\mathcal{M}}_{\pi,t}(s) := Q^{\mathcal{M}}_{\pi,t}(s, \pi(s,t))$. A policy $\pi^{\mathcal{M}}$ is optimal for the MDP $\mathcal{M}$ if $\pi^{\mathcal{M}} \in \arg\max_\pi V^{\mathcal{M}}_{\pi,t}(s)$ for all $s \in \mathcal{S}$ and $t = 0, \ldots, H-1$. We will use $\pi^{\mathcal{M}}$ to denote such an optimal policy.

Let $\mathcal{O}_\ell = (s^\ell_0, a^\ell_0, r^\ell_1, \ldots, s^\ell_{H-1}, a^\ell_{H-1}, r^\ell_H)$ be the sequence of observations made during episode ℓ. Let $\mathcal{H}_{L-1} = (\mathcal{O}_\ell : \ell = 1, \ldots, L-1)$ denote the history of observations made prior to episode L. The agent's behavior is governed by a reinforcement learning algorithm alg. Immediately prior to the beginning of episode L, the algorithm produces a policy $\pi^L = \text{alg}(\mathcal{S}, \mathcal{A}, \mathcal{H}_{L-1})$ based on the state and action spaces and the history $\mathcal{H}_{L-1} = (\mathcal{O}_\ell : \ell = 1, \ldots, L-1)$ of observations made over previous episodes. Note that alg may be a randomized algorithm, so that multiple applications of alg may yield different policies.

In episode ℓ, the agent enjoys a cumulative reward of $\sum_{t=1}^H r^\ell_t$. We define the *regret* over episode ℓ to be the difference between optimal expected value and the sum of rewards generated by algorithm alg. This can be written as $V^{\mathcal{M}}_{\pi^{\mathcal{M}},t}(s^\ell_0) - \sum_{t=0}^{H-1} r_{t+1}$, where actions are generated by a policy π^ℓ is produced by algorithm alg and state transitions and rewards are generate by MDP $\mathcal{M}$.

2 Optimism versus Randomization

In principle, given a history $\mathcal{H}_{L-1}$ of observations gathered over prior episodes, we can generate a point estimate $\hat{Q}_t = \mathbb{E}\left[Q^{\mathcal{M}}_{\pi^{\mathcal{M}},t}|\mathcal{H}_{L-1}\right]$ of the optimal state-action value function and apply a greedy policy with respect to this estimate over episode L. However, it is often essential to apply a policy that will explore beyond this to make discoveries that amplify expected rewards over subsequent episodes.

Optimistic approaches induce exploration by generating optimistic estimates $\overline{Q}_t$ of state-action values and following a greedy policy with respect to optimistic estimates. The idea is that an optimistic estimate $\overline{Q}_t(s,a)$ should represent the highest statistically plausible value of $Q^{\mathcal{M}}_{\pi^{\mathcal{M}},t}(s,a)$, given prior knowledge and observed history.

An alternative approach is to generate prior to each Lth episode the optimal value function $\tilde{Q}_t$ for an MDP sampled from the posterior distribution of $\mathcal{M}$ conditioned on the history $\mathcal{H}_{L-1}$. This is equivalent to sampling $\tilde{Q}_t$ from the posterior distribution of $Q^{\mathcal{M}}_{\pi^{\mathcal{M}},t}$. As discussed in [1, 2, 3], such a randomized approach can be analyzed through the study of confidence sets, similarly with how optimistic algorithms are typically studied, and offer performance similar to well-designed statistically efficient optimistic approaches. As we will discuss, the performance advantage of randomized approaches arises from the fact that optimistic approaches proposed and applied in the literature forgo statistical efficiency for computational tractability.

Empirical evidence suggests that randomization often leads to much faster learning than optimiism. For example, Figure 1, taken from [2], plots regret of UCRL2 [4] and PSRL [5, 2] applied to a variation of the *RiverSwim* problem from [6]. These are well-studied tabular model-based reinforcement learning algorithms that explore via optimism and randomization, respectively. For each algorithm, many trajectories are plotted, corresponding to independent simulations. For these computations, PSRL began with uninformative Dirichlet priors for transition probabilities and normal-gamma priors for transition rewards. It is clear from these results that, for this problem, PSRL learns much faster than UCRL2. In the next two sections, we present simple analytic examples that provide insight into why randomization offers more desirable behavior than common optimistic approaches.

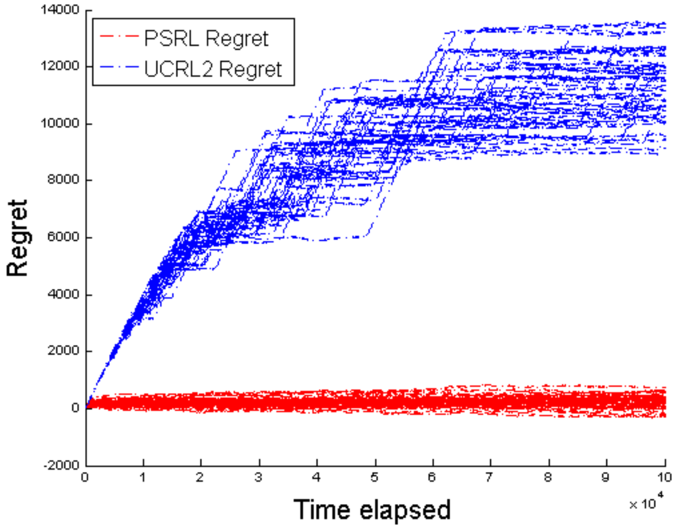


Figure 1: Cumulative regret of UCRL2 and PSRL in the *RiverSwim* environment.

3 Decision Coherence across Time Scales

Consider a simple example illustrated in Figure 2. An agent is at the left-most state and must select one of two actions. Action 1 takes the agent along the "high road" over which he knows that he will experience reward of 1 over the first transition and a reward of 0 over the following $H-1$ transitions. Action 2 takes the agent along the low road, where the agent is uncertain about mean rewards over the first τ transitions. According to the agent's posterior distribution, conditioned on the history of past observations, these mean rewards are independent and identically distributed zero-mean normal random variables with standard deviation $\epsilon/\sqrt{\tau}$.

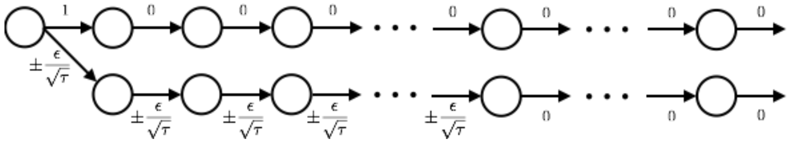

1
0
0
0
0
0
0
0
$\pm\frac{\epsilon}{\sqrt{\tau}}$
$\pm\frac{\epsilon}{\sqrt{\tau}}$
$\pm\frac{\epsilon}{\sqrt{\tau}}$
$\pm\frac{\epsilon}{\sqrt{\tau}}$
$\pm\frac{\epsilon}{\sqrt{\tau}}$
0
0
0

Figure 2: Influence of horizon on on exploration decision.

Table 1 quantifies the posterior distribution of mean value for each action, which is normal with a particular expectation and standard deviation. A well-designed optimistic approach should invest to explore if the standard deviation ϵ, which represents uncertainty in value of the second action, is sufficiently large relative to the difference in expectations, which is 1. In particular, the agent should select action 2 if and only if $c\epsilon > 1$, where c is a tuning parameter that represents the degree of optimism. However, ignoring logarithmic factors, optimistic approaches in the literature (e.g., [4, 7, 8]), are designed to apply an optimistic boost of the form $c\epsilon\sqrt{\tau}$, which results in selecting action 2 if and only if $c\epsilon\sqrt{\tau} > 1$. This is because these optimistic approaches aim to sum over future standard deviations, where one should more appropriately combine uncertainties by summing variances. This flaw in uncertainty quantification leads to an incoherence in decision making: for any fixed c, there are time scales τ for which the agent will explore when it is not sufficiently uncertain or fail to explore despite sufficient uncertainty.

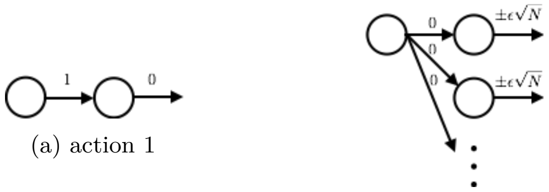
action	expected value	standard deviation	optimistic boost
1	1	0	0
2	0	ϵ	$c\epsilon\sqrt{\tau}$

Table 1: Expectation and standard distribution of action value, and a typical optimistic boost, as a function of horizon.

A typical randomized approach would, for this problem, explore in each episode with probability equal to the posterior probability that the value of action 2 exceeds that of action 1. In particular, randomized approaches allocate effort to exploration proportional to the chances of gaining actionable information. This probability does not depend on τ and therefore does not suffer from the same sort of incoherence with respect to scalings of τ.

4 Decision Coherence across Space Scales

Now consider an example illustrated in Figure 3. The diagrams focus on possible transitions from a single state. Action 1 generates an immediate reward of 1, and is known to transition to a state that leads to no subsequent value. Action 2 generates no immediate reward and is known to transition to one of N states, each with probability $1/N$. From each possible next state, the agent's posterior distribution models subsequent mean value as an independent zero-mean normal random variable with standard deviation $\epsilon\sqrt{N}$.



(a) action 1

(b) action 2

Figure 3: Influence of number of possible next states on on exploration decision.

Table 2 quantifies the posterior distribution of mean value for each action, which is normal with a particular expectation and standard deviation. A well-designed optimistic approach should invest to explore if the standard deviation ϵ, which represents uncertainty in value of the second action, is sufficiently large relative to the difference in expectations, which is 1. In particular, the agent should select action 2 if and only if $c\epsilon > 1$, where c is a tuning parameter that represents the degree of optimism. However, ignoring logarithmic factors, common optimistic approaches would apply the average among optimistic boosts $c\epsilon\sqrt{N}$ associated with possible next states, which results in selecting action 2 if and only if $c\epsilon\sqrt{N} > 1$. This is because these optimistic approaches average over standard deviations at possible next states, where one should more appropriately average variances. This flaw in uncertainty quantification leads to an incoherence in decision making: for any fixed c, there are values of N for which the agent will explore when it is not sufficiently uncertain or fail to explore despite sufficient uncertainty.

action	expected value	standard deviation	optimistic boost
1	1	0	0
2	0	ϵ	$c\epsilon\sqrt{N}$

Table 2: Expectation and standard distribution of action value, and a typical optimistic boost, as a function of the number of possible next states.

A typical randomized approach would again explore in each episode with probability equal to the posterior probability that the value of action 2 exceeds that of action 1. For our example, it is easy to see that this probability does not depend on N and therefore does not suffer from the same sort of incoherence with respect to scalings of the state space.

5 Closing Remarks

Reinforcement learning holds promise to provide the basis for an artificial intelligence that will manage a wide range of systems and devices to better serve society's needs. To date, its potential has primarily been assessed through learning in simulated systems, where data generation is relatively unconstrained and algorithms are typically trained over tens of millions to trillions of episodes. Migrating this technology to real systems where data collection is costly or constrained by the physical context calls for a focus on statistical efficiency. An important part of that lies in how agents explore when learning. Optimism and randomization offer guiding principles for efficient exploration. We have presented a couple analytic examples that shed light on sources of advantage in the efficiency of randomized approaches, relative to optimistic approaches that have been presented in the literature. In principle, it should be possible to design optimistic approaches that combine uncertainties in a more coherent manner and consequently perform at least as well as randomized approaches, but such approaches may be computationally intractable.

A recent area of intense research activity focusses on designing value function learning methods that efficiently explore intractably large state spaces. One thread of work develops count-based optimistic exploration schemes that operate with value function learning [7, 8]. Though these approaches may be effective for a range of problems, they suffer from incoherencies of the kind illustrated in our examples and therefore are likely to forgo a substantial degree of statistical efficiency. An alternative is offered by methods that sample statistically plausible parameterized value functions [9, 10, 11].

References

[1] Dan Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pages 2256–2264. Curran Associates, Inc., 2013.

[2] Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011. Curran Associates, Inc., 2013.

[3] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[4] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[5] Richard Dearden, Nir Friedman, and David Andre. Model based Bayesian exploration. In *UAI*, pages 150–159, 1999.

[6] Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC model-free reinforcement learning. In *ICML*, pages 881–888, 2006.

[7] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1471–1479. Curran Associates, Inc., 2016.

[8] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. #Exploration: A study of count-based exploration for deep reinforcement learning. *CoRR*, abs/1611.04717, 2016.

[9] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2377–2386, 2016.

[10] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances In Neural Information Processing Systems*, pages 4026–4034, 2016.

[11] Ian Osband, Daniel Russo, Benjamin Van Roy, and Zheng Wen. Deep exploration via randomized value functions, 2017. preprint.