An Optimistic Posterior Sampling Strategy for Bayesian Reinforcement Learning

Raphael Fonteneau
University of Liège, Belgium
raphael.fonteneau@ulg.ac.be

Nathan Korda
University of Oxford, England
nathaniel.korda@eng.ox.ac.uk

Rémi Munos
Inria Lille - Nord Europe, France / Microsoft Research New England, USA
remi.munos@inria.fr

Abstract

We consider the problem of decision making in the context of unknown Markov decision processes with finite state and action spaces. In a Bayesian reinforcement learning framework, we propose an optimistic posterior sampling strategy based on the maximization of state-action value functions of MDPs sampled from the posterior. First experiments are promising.

Introduction. The design of algorithms for planning in the context of unknown Markov Decision Processes (MDPs) remains challenging. In particular, one of the main difficulties is to address the so-called Exploration versus Exploitation (E/E) dilemma: at every time-step, the algorithm must both (i) take a decision which is of good quality regarding information that has been collected so far (the *exploitation* part) and (ii) open the door to collecting new information about the (unknown) underlying environment in order to take better decisions in the future (the *exploration* part). At the end of the eighties, the popularization of Reinforcement Learning (RL) [20] gave a new impulse to the research community working on this old problem, and the E/E dilemma was re-discovered in the light of the RL paradigm. Among the approaches that have been proposed to address the E/E dilemma in the RL field, one can mention approaches based on optimism in the face of uncertainty [12, 3, 4, 13, 6, 15] and Bayesian approaches [7, 19, 17, 9, 8]. In the last few years, posterior sampling approaches have received a lot of attention, in particular for solving multi-armed bandits problems [5, 11, 10]. Very recently, posterior sampling has also been proved theoretically and empirically to be efficient for solving MDPs in [16].

Our contribution lies at the crossroads between posterior sampling approaches and optimistic approaches. We propose a strategy based on two main assumptions: (i) a posterior distribution can be maintained over the set of all possible transition models, and (ii) one can easily sample and solve MDPs drawn according to this posterior. These two conditions are easily satisfied in the context of finite state and action space MDPs. Inspired from the principle of the Bayes-UCB algorithm proposed in the context of multi-armed bandit problems [10], our strategy works as follows: at each time-step, a pool of MDPs is drawn from the posterior distribution, and each MDP is solved. We finally take an action whose value is maximized over the set of state-action value functions of sampled MDPs. After observing a new transition, the posterior distribution is updated according to the Bayes rule. We illustrate empirically the performances of our approach on a standard benchmark.

Model-based Bayesian Reinforcement Learning. Let $M = (\mathcal{S}, \mathcal{A}, T, R)$ be a Markov Decision Process (MDP), where the set $\mathcal{S} = \left\{s^{(1)}, \ldots, s^{(n_{\mathcal{S}})}\right\}$ denotes the finite state space and the set $\mathcal{A} = \left\{a^{(1)}, \ldots, a^{(n_{\mathcal{A}})}\right\}$ the finite action space of the MDP. When the MDP is in state $s_t \in \mathcal{S}$ at time $t \in \mathbb{N}$, an action $a_t \in \mathcal{A}$ is selected and the MDP moves toward a new state $s_{t+1} \in \mathcal{S}$, drawn according to a probability

$$T(s_t, a_t, s_{t+1}) = P(s_{t+1}|s_t, a_t) .$$

It also produces an instantaneous deterministic scalar reward $r_t \in [0, 1]$: $r_t = R(s_t, a_t, s_{t+1})$. In this paper, we assume that the transition model T is unknown. For simplicity, we assume that the value $R(s, a, s') \in [0, 1]$ is known for any possible transitions $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

Let $\pi : \mathcal{S} \rightarrow \mathcal{A}$ be a deterministic policy, i.e. a mapping from states to actions. A standard criterion for evaluating the performance of π is to consider its expected discounted return J^π defined as follows:

$$\forall s \in \mathcal{S}, \quad J^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t), s_{t+1}) \big| s_0 = s\right]$$

where $\gamma \in [0, 1)$ is the so-called discount factor. An optimal policy is a policy π^* such that, for any policy π, $J^{\pi^*}(s) \geq J^\pi(s)$,$\forall s \in \mathcal{S}$.

Since the actual transition model T is initially unknown, one needs to address the exploration/exploitation (E/E) trade-off for efficiently acquiring knowledge about the it. Model-based Bayesian RL proposes to address such a trade-off by representing the knowledge about the unknown transition model using a probability distribution over all possible transition models $\boldsymbol{\mu}$. An initial prior distribution $\boldsymbol{b}_0$ is given and iteratively updated according to the Bayes rule as new samples of the actual transition model are generated. At any time-step t, the so-called posterior distribution $\boldsymbol{b}_t$ depends on the prior distribution $\boldsymbol{b}_0$ and the history $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$ observed so-far. The Markovian property implies that the posterior $\boldsymbol{b}_{t+1}$: $\boldsymbol{b}_{t+1} = P(\boldsymbol{\mu}|h_{t+1}, \boldsymbol{b}_0)$ can be updated sequentially: $\boldsymbol{b}_{t+1} = P(\boldsymbol{\mu}|(s_t, a_t, s_{t+1}), \boldsymbol{b}_t)$. The goal is to efficiently exploit the posterior distribution $\boldsymbol{b}_t$ for guiding exploration in order to generate a sequence of policies which maximizes a given E/E criterion. Such a criterion can be, for instance, the expected (either finite or discounted) sum of rewards collected, or the performance of the policy found after a given phase.

Optimistic Posterior Sampling. For a given MDP μ drawn according to the posterior distribution $\mu \sim \boldsymbol{b}_t$, we denote by Q^μ its optimal state-action value function:

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, Q^\mu(s, a) = \sum_{s' \in \mathcal{S}} T^\mu(s, a, s') \left(R(s, a, s') + \gamma \max_{a' \in \mathcal{A}} Q^\mu(s', a') \right) .$$

where $T^\mu(s, a, s')$ denotes the probability to move from state s to state s' when taking action a in MDP μ. Our optimistic posterior sampling (OPS) strategy works according to the following procedure: At time $t \in \mathbb{N}$, for a given state $s_t \in \mathcal{S}$ and a posterior $\boldsymbol{b}_t$:

1. draw a pool of $n_t \in \mathbb{N}$ MDPs $\{\mu_i\}_{i=1}^{n_t}$ according to $\boldsymbol{b}_t$:
$$\forall i \in \{1, \ldots, n_t\}, \mu_i \sim \boldsymbol{b}_t$$
2. obtain the values $\{Q^{\mu_i}(s_t, a)\}_{i=1 \ldots n_t, a \in \mathcal{A}}$ using value iteration
3. apply a decision $a_t \in \mathcal{A}$ such that:
$$a_t \in \arg\max_{a \in \mathcal{A}} \left\{ \max_{i \in \{1, \ldots, n_t\}} Q^{\mu_i}(s_t, a) \right\}$$
(ties are broken arbitrarily)
4. observe a new state s_{t+1}, and update the posterior $\boldsymbol{b}_{t+1} = P(\boldsymbol{\mu}|(s_t, a_t, s_{t+1}), \boldsymbol{b}_t)$.

Note that the second step of OPS can be parallelized. The OPS strategy is illustrated in Figure 1.

Illustration. We compare our approach with other model-based Bayesian RL algorithms on the vanilla 5-state chain problem [19] which is one of the most usual benchmarks for evaluating BRL algorithms. In this benchmark, with probability 0.8, action $a^{(1)}$ sends state $s^{(i)}$ to state $s^{(\min\{i+1,5\})}$, receiving a reward of 1 when starting from state $s^{(5)}$, and 0 otherwise; with probability 0.8, action $a^{(2)}$ sends state $s^{(i)}$ to $s^{(1)}$, receiving a reward of 0.2; with probability 0.2 the behaviours of the actions are reversed. The optimal strategy is to take action 1 whatever the state. In our experiments, we use Dirichlet distributions, and consider a full prior which means that we do not incorporate any specific prior knowledge (all transitions are possible).

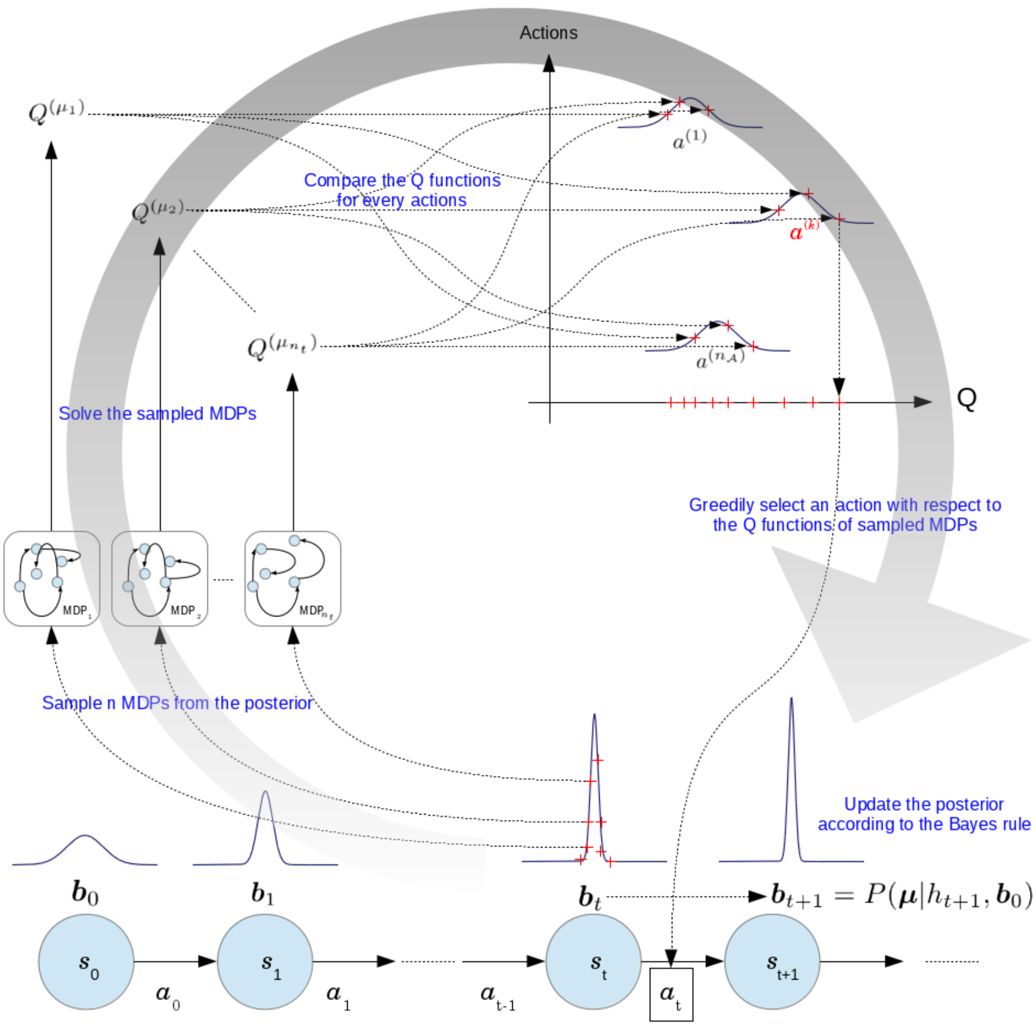

Figure 1: Illustration of the OPS strategy.

Algorithm	Performance
BEETLE [17]	175.4
BOLT ($\eta = 150, \eta = 7$) [1]	278.7, 289.6
BOSS [2]	300.3
f EXPLOIT [17]	307.8
BOP ($n = 500$) [8]	308.8
OPS ($n = 1, 2, 3, 5, 10, 20, \mathbf{30}, 50, 100$)	259.7, 288.5, 301.1, 310.2, 321.1, 325.5, **326.2**, 323.8, 322.2
BEB ($\beta = 150, \beta = 1$) [14]	165.2, 343.0
BVR [18]	**346.5**
Optimal strategy	**367.7**

Table 1: Performance of OPS compared with other model-based BRL approaches on the full-prior 5-state chain MDP problem. Radiuses of 95% confidence intervals are between 2.6 (for $n = 100$) and 6.1 (for $n = 1$).

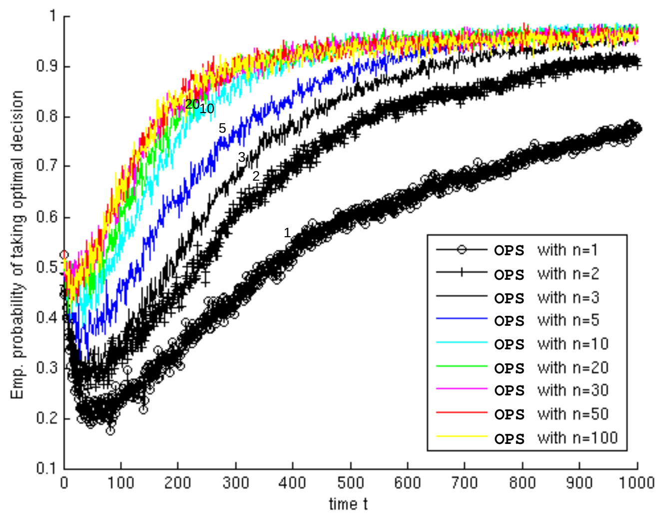

Figure 2: Empirical probability of taking the optimal decision (action $a^{(1)}$) over time (note that action $a^{(1)}$ is optimal for all states in this benchmark).

We ran our algorithm 500 times starting from state $s_0 = s^{(1)}$, each time for 1000 time-steps. We used the parameters $n_t = t$ when $t \leq n$ and $n_t = n$ otherwise, for different values of the threshold parameter $n \in \{1, 2, 3, 5, 10, 20, 30, 50, 100\}$. The empirical average performance (in terms of cumulative undiscounted received rewards) are given in Table 1. We also display in Table 1 the performances obtained by other BRL algorithms in the very same benchmark (obtained from the literature).

We first observe that OPS performs worst when $n = 1$, which corresponds to a simple posterior sampling approach, referred to as "Thompson sampling" in the multi-armed bandit literature, and for which a theoretical analysis of the Bayesian regret is already known [16]. We then observe that the performance of OPS increases with n until $n \sim 30$. Thus, the optimistic strategy offered by the maximization over several sampled MDPs shows an improved empirical performance compared to the Thompson Sampling benchmark. This should be theoretically investigated in future works.

OPS also performs well compare to other standard algorithms, except those using exploration bonuses such as BEB (with a tuned value of its parameter β) and BVR, which outperform OPS on this benchmark. Furthermore, OPS performs better that BOSS, another posterior sampling algorithm wich samples MDPs and combines them into a merged MDP from which a decision is greedily selected. Finally, OPS outperforms BOP [8], which is another algorithm using the optimism principle in a Bayesian RL setting. We also display in Figure 2 the evolution over time of the empirical probability (computed over the 500 runs) that the OPS algorithm takes optimal decision for $n \in \{1, 2, 3, 5, 10, 20, 30, 50, 100\}$.

Conclusions. This paper proposes a new, promising Bayesian RL approach based on an optimistic posterior sampling strategy. We plan to investigate some theoretical aspects of this approach in future research, in particular, analyzing the benefits of optimism in a posterior sampling framework.

Acknowledgment. Raphael Fonteneau is a postdoctoral fellow of the F.R.S-FNRS. We also thank the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements no 270327 (CompLACS) and the Belgian IAP DYSCO.

References

[1] M. Araya, V. Thomas, and O. Buffet. Near-optimal BRL using optimistic local transitions. In *International Conference on Machine Learning (ICML)*, 2012.

[2] J. Asmuth, L. Li, M.L. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Uncertainty in Artificial Intelligence (UAI)*, pages 19–26, 2009.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of multiarmed bandit problems. *Machine Learning*, 47:235–256, 2002.

[4] R.I. Brafman and M. Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2003.

[5] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In *Neural Information Processing Systems (NIPS)*, 2011.

[6] R. Coulom. Efficient selectivity and backup operators in Monte-Carlo tree search. *Computers and Games*, pages 72–83, 2007.

[7] R. Dearden, N. Friedman, and S. Russell. Bayesian Q-learning. In *National Conference on Artificial Intelligence*, pages 761–768, 1998.

[8] R. Fonteneau, L. Busoniu, and R. Munos. Optimistic planning for belief-augmented Markov decision processes. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2013.

[9] A. Guez, D. Silver, and P. Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Neural Information Processing Systems (NIPS)*, 2012.

[10] E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[11] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: an asymptotically optimal finite-time analysis. In *Twenty-third International Conference on Algorithmic Learning Theory (ALT)*, 2012.

[12] M. Kearn and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49:209–232, 2002.

[13] L. Kocsis and C. Szepesvári. Bandit based Monte-Carlo planning. *Machine Learning: ECML 2006*, pages 282–293, 2006.

[14] J.Z. Kolter and A.Y. Ng. Near-Bayesian exploration in polynomial time. In *International Conference on Machine Learning (ICML)*, pages 513–520, 2009.

[15] R. Munos. The optimistic principle applied to games, optimization and planning: Towards Foundations of Monte-Carlo Tree Search. *Foundations and Trends in Machine Learning*, 2013.

[16] I. Osband, D. Russo, and B. Van Roy. (More) Efficient reinforcement learning via posterior sampling. In *Neural Information Processing Systems (NIPS)*, 2013.

[17] P. Poupart, N. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete Bayesian reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 697–704, 2006.

[18] J. Sorg, S. Singh, and R.L. Lewis. Variance-based rewards for approximate Bayesian reinforcement learning. *Uncertainty in Artificial Intelligence (UAI)*, 2010.

[19] M. Strens. A Bayesian framework for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 943–950, 2000.

[20] R.S. Sutton. Learning to predict by the methods of temporal difference. *Machine Learning*, 3:9–44, 1988.