Robust Exploration with Tight Bayesian Plausibility Sets

UNH

Abstract

Optimism about the poorly understood states and actions is the main driving force of exploration in provably-efficient reinforcement learning algorithms. We propose optimism in the face of sensible value functions (OFVF)- a novel reinforcement learning algorithm designed to explore robustly minimizing the worst case exploration cost, where the optimism is driven by tighter Bayesian bounds. OFVF proceeds in an episodic manner, where the duration of the episode is fixed and known. OFVF relaxes the requirement for the set of plausible MDPs to be represented by a confidence interval. It also optimizes the size and location of the plausibility set. Our algorithm is inherently Bayesian and can leverage prior information. Our theoretical analysis shows the robustness of OFVF, and the empirical results demonstrate its practical promise.

1 Introduction

Markov decision processes (MDPs) provide a versatile methodology for modeling dynamic decision problems under uncertainty [Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998; Puterman, 2005]. A perfect MDP model for many reinforcement learning problems is not known precisely in general. Instead, a reinforcement learning agent tries to maximize its cumulative payoff by interacting in an unknown environment with an effort to learn the underlying MDP model. It is important for the agent to explore sub-optimal actions to accelerate the MDP learning task which can help to optimize long-term performance. But it is also important to pick actions with highest known rewards to maximize short-run performance. So the agent always needs to balance between them to boost the performance of a learning algorithm during learning.

Optimism in the face of uncertainty (OFU) is a common principle for most reinforcement learning algorithms encouraging exploration [Jaksch *et al.*, 2010; Brafman and Tennenholtz, 2001; Kearns and Singh, 1998]. The idea is to assign a very high exploration bonus to poorly understood states and actions. The agent chooses a policy under this very "optimistic" model of the environment. As the less understood states-actions are incentivized, they seem lucrative to the agent encouraging exploration. As the agent visits and gathers statistically significant evidence for these states-actions, the uncertainty and optimism decreases converging to reality. Many RL algorithms including *Explicit Explore or Exploit* (E^3) [Kearns and Singh, 1998], *R-MAX* [Brafman and Tennenholtz, 2001], *UCRL2* [Auer, 2006] build on the idea of optimism guiding the exploration. These algorithms provide strong theoretical guarantees with polynomial bound on sample complexity.

The performance of these OFU algorithms greatly depends on the methods to implement optimism, which can often be complex in nature. Dealing with a family of plausible environments can sometime become expensive as well. With OFU exploration, it is possible for an agent to end up in a catastrophic situation paying an extremely high price (e.g. a self driving car hits a wall, a robot falls off the cliff etc.). Exploring and learning such a situation may not payoff the price. It can be wise for the agent to be robust and avoid those situations minimizing the worst-case exploration cost− which we call robust exploration. OFU algorithms are optimistic by definition and cannot guarantee robustness while exploring.

Probability matching class of algorithms like *Thompson sampling* [Thompson, 1933] performs exploration with a proportional likelihood to the underlying true parameters and has been successfully applied to multi-armed bandit problems [Agrawal and Goyal, 2012a;b]. *Posterior Sampling for reinforcement learning (PSRL)* [Osband and Van Roy, 2017; Osband *et al.*, 2013; Strens, 2000] applies the same idea in the context of reinforcement learning. PSRL algorithm samples a single instance of the environment from the posterior distribution, then solves and executes the policy optimal for that sampled environment over the episode. Selection of a policy in PSRL is proportional to the probability of that being optimal and exploration is guided by the variance of sampled policies as opposed to optimism. PSRL algorithm is simple, computationally efficient and can utilize any prior structural information to improve exploration. Strong theoretical analysis and practical applications for PSRL are also established in the literature. But similar to OFU algorithms, PSRL cannot handle worst case exploration penalty and performs poorly in such situations.

The main contribution of this paper is OFVF, a new *data-*

driven Bayesian approach to constructing *Plausibility* sets for MDPs. The method computes policies with tighter robust estimates for exploration by introducing two new ideas. First, it is based on Bayesian posterior distributions rather than distribution-free bounds. Second, OFVF does not construct plausibility sets as simple confidence intervals. Confidence intervals as plausibility sets are a sufficient but not a necessary condition. OFVF uses the structure of the value function to optimize the *location* and *shape* of the plausibility set to guarantee upper bounds directly without necessarily enforcing the requirement for the set to be a confidence interval.

2 Problem Statement

We consider the problem of episodically learning and solving a finite-horizon MDP: $M = (S, A, P^M, R^M, H, \rho)$ where $S = \{1, ..., S\}$ is the state space, $A = \{1, ..., A\}$ is the action space, $P^M(s'|s, a)$ is the believe distribution over the true transition probability $P(s'|s, a)$ of transitioning to state s' when take action a at state s, $R^M(s, a, s')$ is the believe distribution over the true reward $R(s, a, s')$ when take action a at state s and transition to state s', H is the time horizon, and ρ is the initial state distribution. In each episode, the initial state s_0 is sampled from ρ, and in time period $h = 1, 2, ..., H$, if an action a_h is taken in state s_h, then a next state s_{h+1} is sampled from $P(s_{h+1}|s_h, a_h)$ and a reward r_h is sampled from $R(s_h, s_a, s_{h+1})$. The episode terminates when s_H is reached.

A policy μ is a function mapping S to A. For each MDP M and policy μ, we define a value function for each time period $h = 1, 2, ..., H$:

$$V_h^\mu(s) := \mathbb{E}[\sum_{\tau=h}^{H} r_\tau | s_h = s, a_\tau = \mu(s_\tau)]$$

The optimal value function is defined by $V_h^*(s) = \max_\mu V_h^\mu(s)$. A policy μ^* is said to be optimal if $V^{\mu^*} = V^*$ for all $s \in S$ and $h = 1, ..., H$.

⟨FIX: TALK ABOUT STATE-ACTION OPTIMAL VALUE FUNCTION? SEE RLSVI⟩

A reinforcement learning algorithm interacts with the MDP over episodes from time period $\tau = 1$ to $\tau = H$. At each time τ, the algorithm selects an action a_τ, realizes a reward r_τ, and then transitions to state $s_{\tau+1}$. Over each episode k, the algorithm realizes reward $\sum_{\tau=1}^{H}$. One way to quantify the performance of a reinforcement algorithm is in terms of the expected cumulative regret up to time T over K episodes, defined by

$$\begin{aligned} Regret(T, M) &= \sum_{k=1}^{T/H} \mathbb{E}_M[V_1^*(s_{k1}) - \sum_{\tau=1}^{H} r_{k\tau}] \\ &= \sum_{k=1}^{T/H} \mathbb{E}_M[V_1^*(s_{k1}) - V_1^{\mu_k}(s_{k1})] \end{aligned}$$

3 Bayesian Optimism in the Face of Uncertainty

- Talk a little about UCRL
- explain the idea of Bayes UCRL
- Explain why it outperforms UCRL (using variance of the distribution instead of distribution free Hoeffding bound)
- Explain why it doesn't outperform PSRL (Rectangularity)

Algorithm 1: Bayes UCRL

Input: Desired confidence level δ and prior distribution
Output: Policy with an optimistic return estimate

1 **repeat**
2 Initialize MDP: M;
3 Compute posterior: $\tilde{p} \leftarrow$ compute_posterior(prior, samples) ;
4 **foreach** $s \in \mathscr{S}, a \in \mathscr{A}$ **do**
5 $\bar{p}_{s,a}, \psi_{s,a} \leftarrow$ Invoke Algorithm 3 with $\tilde{p}$, δ;
6 $M \leftarrow$ add transition with $\bar{p}_{s,a}, \psi_{s,a}$;
7 Compute policy by solving MDP: $\hat{\pi} \leftarrow$ Solve M;
8 Collect samples by executing the policy: samples $\leftarrow$ execute $\hat{\pi}$;
9 prior $\leftarrow$ posterior;
10 **until** *num episodes*;
11 **return** $(\pi_k, p_0^\mathsf{T} v_k)$;

4 OFVF: Optimism in the Face of sensible Value Functions

OFVF uses samples from a posterior distribution, similar to a Bayesian confidence interval, but it relaxes the safety requirement as it is sufficient to guarantee for each state s and action a that:

$$\max_{v \in \mathscr{V}} \mathbb{P}_{P^\star}\left[\min_{p \in \mathscr{P}_{s,a}} (p - p^\star_{s,a})^\mathsf{T} v \leqslant 0 \,\middle|\, \mathscr{D}\right] \geqslant 1 - \frac{\delta}{SA}\,, \quad (1)$$

with $\mathscr{V} = \{\hat{v}^\star_{\mathscr{P}}\}$. To construct the set $\mathscr{P}$ here, the set $\mathscr{V}$ is not fixed but depends on the robust solution, which in turn depends on $\mathscr{P}$. RSVF starts with a guess of a small set for $\mathscr{V}$ and then grows it, each time with the current value function, until it contains $\hat{v}^\star_{\mathscr{P}}$ which is always recomputed after constructing the ambiguity set $\mathscr{P}$.

In lines 4 and 8 of Algorithm 2, $\mathscr{P}_i$ is computed for each state-action $s, a \in \mathscr{S} \times \mathscr{A}$. Center $\bar{p}$ and set size $\psi_{s,a}$ are computed from Eq. (3) using set $\mathscr{V}$ & optimal g_v computed

Algorithm 2: OFVF: Optimism in the Face of sensible Value Functions

Input: Desired confidence level δ and posterior distribution $\mathbb{P}_{P^\star}[\cdot \mid \mathscr{D}]$

Output: Policy with a maximized safe return estimate

Initialize current policy $\pi_0 \leftarrow \arg\max_\pi \rho(\pi, \mathbb{E}_{P^\star}[P^\star \mid \mathscr{D}])$;

Initialize current value $v_0 \leftarrow v^{\pi_0}_{\mathbb{E}_{P^\star}[P^\star \mid \mathscr{D}]}$;

Initialize value robustness set $\mathscr{V}_0 \leftarrow \{v_0\}$;

Construct $\mathscr{P}_0$ optimal for $\mathscr{V}_0$;

Initialize counter $k \leftarrow 0$;

while *Eq.* (1) *is violated with* $\mathscr{V} = \{v_k\}$ **do**

  Include v_k that violates Eq. (1): $\mathscr{V}_{k+1} \leftarrow \mathscr{V}_k \cup \{v_k\}$;

  Construct $\mathscr{P}_{k+1}$ optimized for $\mathscr{V}_{k+1}$;

  Compute robust value function v_{k+1} and policy π_{k+1} for $\mathscr{P}_{k+1}$;

  $k \leftarrow k + 1$;

return $(\pi_k, p_0^\mathsf{T} v_k)$;

by solving Eq. (2). When the set $\mathscr{V}$ is a singleton, it is easy to compute a form of an optimal ambiguity set.

$$g = \max\left\{k \;:\; \mathbb{P}_{P^\star}[k \leqslant v^\mathsf{T} p^\star_{s,a}] \geqslant 1 - \delta/(SA)\right\} \quad (2)$$

When $\mathscr{V}$ is a singleton, it is sufficient for the ambiguity set to be a subset of the hyperplane $\{p \in \Delta^S \;:\; v^\mathsf{T} p = g^\star\}$ for the estimate to be safe. When $\mathscr{V}$ is not a singleton, we only consider the setting when it is discrete, finite, and relatively small. We propose to construct a set defined in terms of an L_1 ball with the minimum radius such that it is safe for every $v \in \mathscr{V}$. Assuming that $\mathscr{V} = \{v_1, v_2, \ldots, v_k\}$, we solve the following linear program:

$$\psi_{s,a} = \min_{p \in \Delta^S} \left\{ \max_{i=1,\ldots,k} \|q_i - p\|_1 \;:\; v_i^\mathsf{T} q_i = g_i^\star, q_i \in \Delta^S, i \in 1, \ldots, k \right\} \quad (3)$$

In other words, we construct the set to minimize its radius while still intersecting the hyperplane for each v in $\mathscr{V}$. Algorithm 2, as described, is not guaranteed to converge in finite time as written. It can be readily shown the value functions in the individual iterations are non-increasing. It is easy to just stop once the value function becomes smaller (and that is more conservative) than BCI.

5 Empirical Evaluation

In this section, we evaluate the worst-case estimates computed by Bayes UCRL and OFVF empirically. We assume a true model of each problem and generate a number of simulated data sets for the known distribution. We compute the tightest optimistic estimate for the optimal return and compare it with the optimal return for the true model. We compare our results with "UCRL2" and "PSRL" algorithms. The value ξ represents the predicted regret, which is the absolute difference between the *true* optimal value and the robust estimate: $\xi = |\rho(\pi^\star_{P^\star}, P^\star) - \hat{\rho}(\hat{\pi}^\star)|$, a smaller regret is better. All of our experiments use a 95% confidence for safety unless otherwise specified.

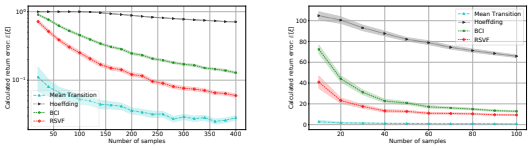

Figure 1: Return error with a Gaussian prior with 95% confidence, Left: Single state, Right: Full MDP, X-axis is the number of samples per state-action.

5.1 Single-state Bellman Update

We initially consider simple problems where transition from a single non-terminal state following a single action leads to multiple terminal states. The value function for the terminal states are fixed and assumed to be provided. We evaluate different priors over the transition probabilities: i) uninformative Dirichlet prior and ii) informative Gaussian prior. Note that RSVF is optimal in this simplistic setting, as **??** (left) and Fig. 1 (left) shows. As expected, the mean estimate provides the tightest bound, but **??** (right) illustrates that it does not provide any meaningful safety guarantees.

5.2 River Swim Problem

5.3 Mountain Car Problem

References

S. Agrawal and N. Goyal. Thompson Sampling for contextual bandits with linear payoffs. *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 28:127–135, 2012.

Shipra Agrawal and N Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Annual Conference on Learning Theory (COLT)*, pages 39.1–39.26, 2012.

Peter Auer. Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning. *Advances in Neural Information Processing Systems (NIPS)*, 2006.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. 1996.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal re-

inforcement learning. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

Karina V. Delgado, Leliane N. De Barros, Daniel B. Dias, and Scott Sanner. Real-time dynamic programming for Markov decision processes with imprecise probabilities. *Artificial Intelligence*, 230:192–223, 2016.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. 3rd edition, 2014.

GA Hanasusanto and Daniel Kuhn. Robust Data-Driven Dynamic Programming. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, may 2005.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(1):1563–1600, 2010.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. In *International Conference on Machine Learning*, volume 49, pages 260–268. Morgan Kaufmann, 1998.

Shie Mannor, O Mebel, and H Xu. Lightning does not strike twice: Robust MDPs with coupled uncertainty. In *International Conference on Machine Learning*, 2012.

Ian Osband and Benjamin Van Roy. Why is Posterior Sampling Better than Optimism for Reinforcement Learning? *International Conference on Machine Learning (ICML)*, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) Efficient Reinforcement Learning via Posterior Sampling? *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Marek Petrik and Dharmashankar Subramanian. RAAM : The benefits of robustness in approximating aggregated MDPs in reinforcement learning. In *Neural Information Processing Systems (NIPS)*, 2014.

Marek Petrik, Mohammad Ghavamzadeh, and Yinlam Chow. Safe Policy Improvement by Minimizing Robust Baseline Regret. In *Advances in Neural Information Processing Systems*, 2016.

Marek Petrik. Approximate dynamic programming by minimizing distributionally robust bounds. In *International Conference of Machine Learning*, 2012.

Martin L Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 2005.

a. L. Strehl and M. L. Littman. An empirical evaluation of interval estimation for markov decision processes. (April 2007):128–135, 2004.

Malcolm Strens. A Bayesian Framework for Reinforcement Learning. *International Conference on Machine Learning (ICML)*, 2000.

Richard S Sutton and Andrew Barto. *Reinforcement learning*. 1998.

Majid Alkaee Taleghan, Thomas G. Dietterich, Mark Crowley, Kim Hall, and H. Jo Albers. PAC Optimal MDP Planning with Application to Invasive Species Management. *Journal of Machine Learning Research*, 16(1):3877–3903, 2015.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling Up Robust MDPs using Function Approximation. In *International Conference of Machine Learning (ICML)*, 2014.

W.R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Oxford University Press*, 25(3):285–294, 1933.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L_1 deviation of the empirical distribution. jun 2003.

Wolfram Wiesemann, Daniel Kuhn, and Berc Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Huan Xu and Shie Mannor. The robustness-performance tradeoff in Markov decision processes. *Advances in Neural Information Processing Systems*, 2006.

Huan Xu and Shie Mannor. Parametric regret in uncertain Markov decision processes. *Proceedings of the IEEE Conference on Decision and Control*, pages 3606–3613, 2009.

A Technical Results

A.1 Computing Bayesian Confidence Interval

Algorithm 3: Bayesian Confidence Interval (BCI)

Input: Distribution θ over $p^{\star}_{s,a}$, confidence level δ, sample count m

Output: Nominal point $\bar{p}_{s,a}$ and L_1 norm size $\psi_{s,a}$

Sample $X_1, \ldots, X_m \in \Delta^S$ from θ: $X_i \sim \theta$;

Nominal point: $\bar{p}_{s,a} \leftarrow (1/m) \sum_{i=1}^{m} X_i$;

Compute distances $d_i \leftarrow \|\bar{p}_{s,a} - X_i\|_1$ and sort *increasingly*;

Norm size: $\psi_{s,a} \leftarrow d_{(1-\delta)\,m}$;

return $\bar{p}_{s,a}$ *and* $\psi_{s,a}$;