Robust Exploration with Tight Bayesian Plausibility Sets

Reazul H. Russel
Department of Computer Science
University of New Hampshire
Durham, NH 03824
rrussel@cs.unh.edu

Tianyi Gu
Department of Computer Science
University of New Hampshire
Durham, NH 03824
gu@cs.unh.edu

Marek Petrik
Department of Computer Science
University of New Hampshire
Durham, NH 03824
mpetrik@cs.unh.edu

Abstract

Optimism about the poorly understood states and actions is the main driving force of exploration for many provably-efficient reinforcement learning algorithms. We propose optimism in the face of sensible value functions (OFVF)- a novel *data-driven* Bayesian algorithm to constructing *Plausibility* sets for MDPs to explore robustly minimizing the worst case exploration cost. The method computes policies with tighter optimistic 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. OFVF proceeds in an episodic manner, where the duration of the episode is fixed and known. 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.

Keywords: Reinforcement Learning, Markov Decision Process, Exploration in RL,
Bayesian Learning, Multi-armed bandits.

Acknowledgements

This work was supported by the National Science Foundation under Grant No. IIS-1717368 and IIS-1815275.

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 [Auer *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. 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; Auer *et al.*, 2010], *MBIE* [Strehl and Littman, 2008, 2004b,a; Wiering and Schmidhuber, 1998] build on the idea of optimism guiding the exploration. Probability matching class of algorithms like *Posterior Sampling for reinforcement learning (PSRL)* [Osband and Van Roy, 2017; Osband *et al.*, 2013; Strens, 2000] performs exploration with a proportional likelihood to the underlying true parameters. PSRL algorithm is simple, computationally efficient and can utilize any prior structural information to improve exploration. These algorithms provide strong theoretical guarantees with polynomial bound on sample complexity.

During exploration, it is possible for an agent to be overly optimistic about a potentially catastrophic situation and end up there 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 and PSRL algorithms are optimistic by definition and cannot guarantee robustness while exploring. The main contribution of this paper is OFVF, an optimistic counter part of RSVF [Russel and Petrik, 2018]. OFVF is a Bayesian approach of constructing plausibility sets for robust exploration.

The paper is organized as follows: Section 2 formally defines the problem setup and goals of the paper. Section 3 reviews some existing methods to construct the plausibility sets and their extension to Bayesian setting. OFVF is proposed and analyzed in Section 4. Finally, Section 5 presents empirical performance on several problem domains.

2 Problem Statement

We consider the problem of learning a finite horizon Markov Decision Process $\mathscr{M}$ with states $\mathscr{S} = \{1, \ldots, S\}$ and actions $\mathscr{A} = \{1, \ldots, A\}$. $p : \mathscr{S} \times \mathscr{A} \to \Delta^{\mathscr{S}}$ is a transition function, where $p_{ss'}^a$ is interpreted as the probability of ending in state $s' \in \mathscr{S}$ by taking an action $a \in \mathscr{A}$ from state $s \in \mathscr{S}$. We omit s' when the next state is not deterministic and denote the transition probability as $p_{sa} \in \mathbb{R}^S$. $R : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ is a reward function and $R_{ss'}^a$ is the reward for taking action $a \in \mathscr{A}$ from state $s \in \mathscr{S}$ and reaching state $s' \in \mathscr{S}$. Each MDP $\mathscr{M}$ is associated with a discount factor $0 \leqslant \gamma \leqslant 1$ and a distribution of initial state probabilities p_0. We consider an episodic learning process where L is the number of episodes and H is the number of periods in each episode. A policy $\pi = (\pi_0, \ldots, \pi_{H-1})$ is a set of functions mapping a state $s \in \mathscr{S}$ to an action $a \in \mathscr{A}$. We define a value function for a policy π as:

$$V_h^\pi(s) := \sum_{s'} P_{ss'}^{\pi(s)}[r_h + V(s')] \tag{1}$$

The optimal value function is defined by $V_h^\star(s) = \max_\pi V_h^\pi(s)$ and the optimal policy is defined by $\pi^\star(s) = \arg\max_{a \in \mathscr{A}} p_{ss'}^a V(s'), \forall s' \in \mathscr{S} : p_{ss'}^a > 0$.

Optimistic algorithms encouraging exploration find the probability distribution $\tilde{P}_{sa}$ for each state and action within an interval of the empirically derived distribution $\bar{p}_{sa} = \mathbb{E}[\cdot|s, a]$, which defines the plausible set $\mathscr{P}_{sa}$ of MDPs. They then solve an optimistic version of Eq. (1) within $\mathscr{P}_{sa}$ that leads to the policy with highest reward.

$$V_h^\star(s, a) := \max_{p_{sa} \in \mathscr{P}_{sa}} \sum_{s'} p_{ss'}^{\pi(s)}[r_h + V^\star(s')] \tag{2}$$

We evaluate the performance of the agent in terms of worst-case *cumulative regret*, which is the maximum total regret incurred by the agent upto time T for a policy $\pi_l^\star$:

$$Regret(T, \pi_l^\star) = \sum_{l=0}^{T/H-1} \sup \left[\sum_{s \in \mathscr{S}} p_0(s) \big(V^\star(s) - V^{\pi_l^\star}(s) \big) \right] \tag{3}$$

Where $V^\star(s)$ is the true value w.r.t $\mathscr{M}^*$.

3 Interval Estimation for Plausibility Sets

In this section, we first describe the standard approach to constructing plausibility sets as distribution free confidence intervals. We then propose its extension to Bayesian setting and present a simple algorithm to serve that purpose. It is important to note that distribution-free bounds are subtly different from the Bayesian bounds, the Bayesian safety guarantee holds conditional on a given dataset $\mathscr{D}$ while the distribution-free hold across the sets. This makes the guarantees qualitatively different and difficult to compare.

3.1 Plausibility Sets as Confidence Intervals

It is common in the literature to use L_1 norm as the distribution-free bound. This bound is constructed around the empirical mean of the transition probability $\bar{p}_{s,a}$ by applying the Hoeffding inequality [Auer *et al.*, 2010; Petrik *et al.*, 2016; Wiesemann *et al.*, 2013; Strehl and Littman, 2004b].

$$\mathscr{P}_{sa} = \left\{ \|\tilde{p}_{sa} - \bar{p}_{sa}\|_1 \leqslant \sqrt{\frac{2}{n_{s,a}} \log \frac{SA2^S}{\delta}} \right\}$$

where $\bar{p}_{sa}$ is the mean transition computed from D, $n_{s,a}$ is the number of times the agent arrived in state s' after taking action a in state s, δ is the required probability of the interval and $\|\cdot\|_1$ is the L_1 norm. An important limitation of this approach is that, the size of $\mathscr{P}_{sa}$ grows linearly with the number of states, which makes it practically useless in general.

3.2 Bayesian Plausibility Sets

The Bayesian plausibility sets take the same interval estimation idea and extend it into Bayesian setting, which is analogous to *credible intervals* in Bayesian statistics. Credible intervals are constructed with the posterior probability distributions and they are fixed $-$ not a random variable, given the data $\mathscr{D}$. Instead the estimated transition probabilities maximizing the rewards are random variables. To construct a plausibility set, we optimize for the smallest credible region around the mean transition probability with the assumption that a smaller region will lead to a tighter upper bound estimate. Formally, the optimization problem to compute $\psi_{s,a}$ for each state s and action a is:

$$\min_{\psi \in \mathbb{R}_+} \left\{ \psi \ : \ \mathbb{P}\left[\|\tilde{p}_{s,a} - \bar{p}_{s,a}\|_1 > \psi \mid \mathscr{D} \right] < \delta \right\} \ , \tag{4}$$

where nominal point is $\bar{p}_{s,a} = \mathbb{E}_{\tilde{P}}[\tilde{p}_{s,a} \mid \mathscr{D}]$. A Bayesian extension of the celebrated *UCRL* [Auer *et al.*, 2010] algorithm is *BayesUCRL*, which we consider for comparison. BayesUCRL algorithm uses a hierarchical Bayesian model that can be used to infer the posterior transition probability over $p^\star$. The plausibility set here is a function of the $\frac{1}{t}$-quantile of the posterior samples. We omit the details of BayesUCRL to conserve space.

4 OFVF: Optimism in the Face of sensible Value Functions

Algorithm 1: OFVF

```
Input: Desired confidence level $\delta$ 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 set $\mathscr{V}_0 \leftarrow \{v_0\}$ ;
Construct $\mathscr{P}_0$ optimal for $\mathscr{V}_0$;
Initialize counter $k \leftarrow 0$;
while Eq. (5) is violated with $\mathscr{V} = \{v_k\}$ do
    Include $v_k$ that violates Eq. (5): $\mathscr{V}_{k+1} \leftarrow \mathscr{V}_k \cup \{v_k\}$ ;
    Construct $\mathscr{P}_{k+1}$ optimized for $\mathscr{V}_{k+1}$;
    Compute optimistic value function $v_{k+1}$ and policy $\pi_{k+1}$ for $\mathscr{P}_{k+1}$;
    $k \leftarrow k + 1$ ;
return $(\pi_k, p_0^\mathsf{T} v_k)$ ;
```

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:

$$\min_{v \in \mathscr{V}} \mathbb{P}_{P^\star} \left[\max_{p \in \mathscr{P}_{s,a}} (p - p^\star_{s,a})^\mathsf{T} v \leqslant 0 \;\middle|\; \mathscr{D} \right] \geqslant 1 - \frac{\delta}{SA} \ , \tag{5}$$

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 optimistic solution, which in turn depends on $\mathscr{P}$. OFVF 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 $\mathscr{P}$.

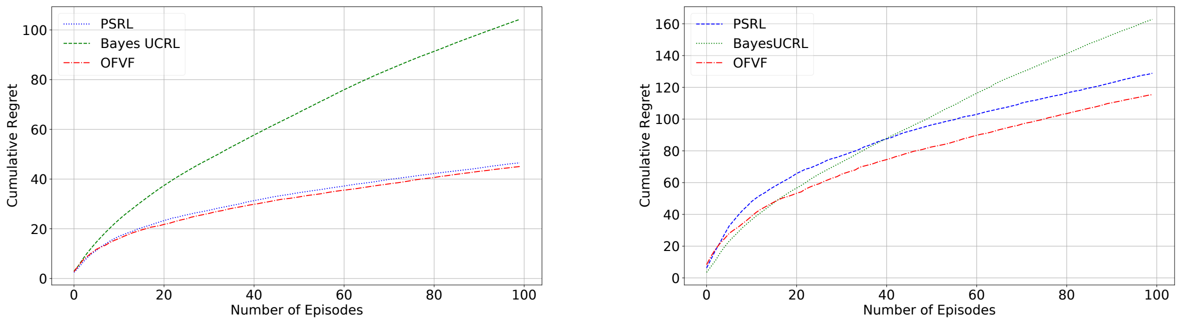

Figure 1: Cumulative regret for the single-state simple problem. Left) average-case, Right) worst-case.

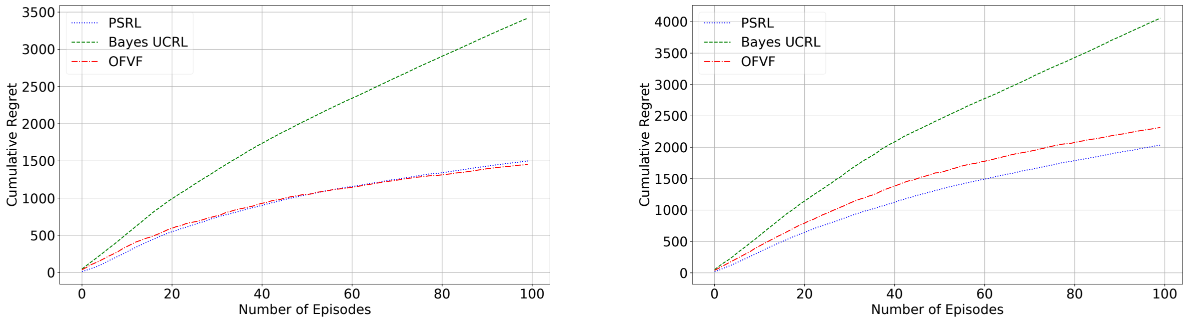

Figure 2: Cumulative regret for the RiverSwim problem. Left) average-case, Right) worst-case.

In lines 4 and 8 of Algorithm 1, $\mathscr{P}_i$ 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. (7) using set $\mathscr{V}$ & optimal g_v computed by solving Eq. (6). When the set $\mathscr{V}$ is a singleton, it is easy to compute a form of an optimal plausibility set.

$$g = \max\left\{k \;:\; \mathbb{P}_{P^\star}[k \leqslant v^\top p^\star_{s,a}] \geqslant 1 - \delta/(SA)\right\} \tag{6}$$

For a singleton $\mathscr{V}$, it is sufficient for the plausibility set to be a subset of the hyperplane $\{p \in \Delta^S \;:\; v^\top p = g^\star\}$ for the estimate to be sufficiently optimistic. 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:

$$\psi_{s,a} = \min_{p \in \Delta^S}\left\{\max_{i=1,\ldots,k} \|q_i - p\|_1 \;:\; v_i^\top q_i = g_i^\star, q_i \in \Delta^S, i \in 1, \ldots, k\right\} \tag{7}$$

In other words, we construct the set to minimize its radius while still intersecting the hyperplane for each v in $\mathscr{V}$.

5 Empirical Evaluation

In this section, we empirically evaluate the estimated returns over episodes. 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. To judge the performance of the methods, we evaluate both the absolute error of the worst case estimates from optimal, as well the average case estimate from optimal.

We compare our results with BayesUCRL and PSRL algorithms. We omit UCRL from comparison because it performs too poorly compared to other methods. PSRL performs very well in both average and worst case, and as we will see in the experiments, OFVF outperforms BayesUCRL and performs competitively with PSRL. For all the experiments, we use an uninformative Dirichlet prior for the transition probabilities, and run experiments for 100 episodes each containing 100 runs, unless otherwise specified.

Single-state Bellman Update We initially consider a simple problem with one single non-terminal state. The agent can take three different actions on that state. Each action leads to one of three terminal states with different transition probabilities. The value function for the terminal states are fixed and assumed to be known. Fig. 1 compares the average-case and worst-case returns computed by different methods. Note that OFVF outperforms all other methods in this simplistic setting. OFVF is able to explore in a robust way maximizing the worst and average case returns.

RiverSwim Problem We compare the performance of different methods in standard example of RiverSwim [Osband *et al.*, 2013; Strehl and Littman, 2004b]. The problem is designed requiring hard exploration to find the optimal policy, we omit the full description of the problem to preserve space. Fig. 2 compares the average and worst case regrets of different methods. Among optimistic methods, OFVF performs better than BayesUCRL both in average and worst case scenario. But the stochastically optimistic PSRL outperforms all other methods. This is due to the fact that, BayesUCRL and OFVF constructs a plausibility set for each state and action. Even if the plausibility sets are tight, the resulting optimistic MDP is simultaneously optimistic in each state-action, yielding a way too optimistic overall MDP model [Osband and Van Roy, 2017]. Thus OFVF can construct tighter plausibility sets for exploration, but still may not match the statistical efficiency of PSRL. This performance however shows that, as an OFU algorithm, OFVF can be reasonably optimistic and can offer competitive performance.

6 Summary and Conclusion

In this paper, we proposed OFVF, a Bayesian algorithm capable of constructing plausibility sets with better shapes and sizes. Beside the fact that our proposed Bayesian methods are computationally demanding than other distribution free methods, our theoretical and experimental analysis furnished that they can pay-off with much tighter return estimates. We showed that, OFU algorithms can be useful and can be competitive to stochastically optimistic algorithm like PSRL.

References

P Auer, Thomas Jaksch, and R Ortner. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(1):1563–1600, 2010.

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*. Athena Scientific, 1996.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *International Conference on Machine Learning (ICML)*, 1998.

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, Yinlam Chow, and Mohammad Ghavamzadeh. Safe Policy Improvement by Minimizing Robust Baseline Regret. *Advances in Neural Information Processing Systems (NIPS)*, 2016.

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

Reazul Hasan Russel and Marek Petrik. Tight Bayesian Ambiguity Sets for Robust MDPs. *Infer to Control, Workshop on Probabilistic Reinforcement Learning and Structured Control, Advances in Neural Information Processing Systems (NIPS)*, 2018.

Alexander. L. Strehl and Michael L. Littman. An empirical evaluation of interval estimation for markov decision processes. *IEEE International Conference on Tools with Artificial Intelligence*, 2004.

Alexander L Strehl and Michael L Littman. Exploration via Model-based Interval Estimation. *International Conference on Machine Learning (ICML)*, 2004.

Alexander L Strehl and Michael L Littman. An Analysis of Model-Based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74:1309–1331, 2008.

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

Richard S Sutton and Andrew Barto. Reinforcement Learning: An Introduction. 1998.

Marco Wiering and Jurgen Schmidhuber. Efficient Model-Based Exploration. *International Conference on Simulation of Adaptive Behavior (SAB)*, pages 223–228, 1998.

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