Robust Exploration with Tight Bayesian Plausibility Sets

UNH

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 reinforcement learning algorithm designed to explore robustly minimizing the worst case exploration cost. 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], *MBIE* [Strehl and Littman, 2008; 2004; Wiering and Schmidhuber, 1998] 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 (e.g. Chernoff-Hoeffding’s inequality for UCRL2, Confidence Interval for MBIE), which can often be complex in nature. Dealing with a family of plausible environments can sometimes become expensive as well. With OFU 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 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.

The paper is organized as follows ...

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 $p^a_{ss'}$ 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 $R^a_{ss'}$ 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. L is the number of episodes and H is the number of periods in each episode, we consider an episodic learning environment throughout the paper. In each episode $l \in L$, an initial state $s_0 \in \mathscr{S}$ is sampled from p_0. In period $h = 0 \ldots H$ of each episode, for a state s_h and action a_h, a next state s_{h+1} is sampled following $p^{a_h}_{s_h s_{h+1}}$ and a reward is obtained from $R^{a_h}_{s_h s_{h+1}}$.

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 each policy π as:

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

The optimal value function is defined by $V^\star_h(s) = \max_\pi V^\pi_h(s)$ and the optimal policy is defined by $\pi^\star(s) = \arg\max_{a \in \mathscr{A}} p^a_{ss'} V(s'), \forall s' \in \mathscr{S} : p^a_{ss'} > 0$. We also define the state-action optimal value function for $h = 0, \ldots, H-1$ as:

$$Q^\star_h(s,a) := \sum_{s'} P^a_{ss'}[r_h + V^\star(s')] \tag{2}$$

The optimal policy $\pi^\star$ is defined as $\pi^\star(s) = \arg\max_{a \in \mathscr{A}} Q^\star_h(s,a), \forall s, h.$ 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. (2) within $\mathscr{P}_{sa}$ that leads to the policy with highest reward.

$$Q^\star_h(s,a) := \max_{p_{sa} \in \mathscr{P}_{sa}} \sum_{s'} p^a_{ss'}[r_h + V^\star(s')] \tag{3}$$

An RL agent interacts with the environment in an episodic setting. At each episode $l \in 0, \ldots, L-1$, the agent takes actions based on the policy $\pi^\star_l$ optimal in episode l and realizes reward $\sum_{h=0}^{H} r_{lh}$. $\pi^\star_l$ is computed from the experiences gathered by the agent in previous episodes $0, \ldots, l-1$. We evaluate the performance of the agent in terms of *expected cumulative regret*:

$$Regret(T) = \sum_{l=0}^{T/H-1} \mathbb{E}\Big[V^\star(s_0) - V^{\pi^\star_l}(s_0)\Big]$$

Where $s_0 \sim p_0$, and $V^\star(s_0)$ are the true values of initial states..

3 Interval Estimation for Plausibility Sets

When solving an MDP in reinforcement learning, the main type of uncertainties are encountered on the model parameters, namely in transition probabilities and rewards. Given data set $\mathscr{D}$, an interval estimation of the model parameters is able to acknowledge and quantify this uncertainty, which a maximum likelihood based point estimate cannot do. 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 $\mathscr{D}$. 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.*, 2009; 2010; Petrik and Luss, 2016; Wiesemann, 2013; Strehl and Littman, 2004].

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

where $\bar{P}_{sa}$ is the mean transition computed from D, $n_{s,a}$ is the number of times the agent arrived 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. They can take advantage of situation-specific prior knowledge unlike general confidence intervals as discussed above. However, a concise comparison between Bayesian and frequentist bounds is beyond the scope of this paper.

The credible region for the plausibility set can be constructed around the mean, median or mode of the sample posterior distribution. Among which, bound around the empirical mean of the transition probabilities is simple and more common. To construct such a region, we optimize for the smallest plausibility set around the mean transition probability with the assumption that a smaller plausibility set 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:

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

where nominal point is $\bar{p}_{s,a} = \mathbb{E}_{\tilde{P}}[\tilde{p}_{s,a} \mid \mathscr{D}]$. The nominal point is not included in the optimization to reduce computational complexity.

The optimization problem in Eq. (4) can be solved by the Sample Average Approximation (SAA) algorithm [Shapiro *et al.*, 2014]. The main idea is to sample from the posterior distribution and then choose the minimal size $\psi_{s,a}$ that satisfies the constrain. Algorithm 4, in the appendix, summarizes the simple sort-based method.

3.3 Bayesian Upper Confidence Reinforcement Learning (BayesUCRL)

We are now ready to describe a simple Bayesian optimistic algorithm, named as Bayes-UCRL. This builds directly on top of UCRL [Auer *et al.*, 2009; 2010], which constructs the plausibility set as distribution free bounds, similar to what is described in Section 3.1. Unlike UCRL, Bayes-UCRL constructs the plausibility set in a Bayesian setting, as described in Section 3.2. Algorithm 1 describes the idea in more detail. The algorithm uses a hierarchical Bayesian model that can be used to infer the posterior transition probability over $p^\star$. The details of the Bayesian model are largely irrelevant for our purpose. The model may have a simple analytical posterior, or may need to be sampled using MCMC methods like Stan [Gelman *et al.*, 2014]. The algorithm assumes that it is possible to draw enough samples from the posterior that the sampling error becomes negligible.

Algorithm 1: BayesUCRL

```
Input: Prior distribution $\theta^0$ over $p^\star_{s,a}$, number of
       episodes K
Output: Policy with an optimistic return estimate
$k = 1, X = \phi$;
repeat
    Confidence level: $\delta^k = 1 - \frac{1}{k+1}$;
    Apply union bound: $\delta^k = 1 - \frac{\delta^k}{SA}$;
    Initialize MDP: $\mathscr{M}^k$;
    Compute posterior $\theta^k$ from $\theta^{k-1}$ and samples $X$;
    for All states and actions: $s \in \mathscr{S}, a \in \mathscr{A}$ do
        Compute mean transition $\bar{p}_{s,a}$, and interval
          $\psi_{s,a}$ (Invoke Algorithm 4 with $\theta^k$ and $\delta^k$);
        Add transition for $s$ and $a$ in $\mathscr{M}^k$;
    Compute value function and policy: $v_k, \hat{\pi}_k \leftarrow$
      solve $\mathscr{M}^k$;
    $X \leftarrow$ execute $\hat{\pi}_k$ and collect samples;
    $k \leftarrow k + 1$
until $k \leqslant K$;
return $(\hat{\pi}_k, p_0^\mathsf{T} v_k)$ ;
```

Algorithm 1 proceeds in an episodic manner. At the beginning of each episode, it computes the posterior transition probabilities for each state and action. It then invokes Algorithm 4 to compute the nominal transition points and the interval estimations for each state and action. It solves the MDP to compute the optimal policy for current episode. The execution of this policy for a fixed horizon generates more samples for posterior estimation in the next episode. This algorithm requires some extra steps compared to UCRL. It is also computationally expensive. But it performs better than UCRL in terms of computed returns over episodes which we will see in Section 5.

4 OFVF: Optimism in the Face of sensible Value Functions

In this section, we describe OFVF, a new approach to constructing plausibility sets to drive exploration in reinforcement learning. OFVF uses samples from posterior distribution, similar to BayesUCRL. But OFVF can optimize the size and shape of the plausibility sets that can provide much better upper bounds for optimism. We first outline the algorithm and illustrate how each step is achieved.

OFVF relaxes the requirement for plausibility sets to be a confidence interval. It is sufficient to guarantee for each state s and action a that:

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

with value functions $\mathscr{V} = \{\hat{v}^\star_{\mathscr{P}}\}$. 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 $\mathscr{P}$. 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 $\mathscr{P}$.

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

```
Input: Desired confidence level $\delta$ and posterior
       distribution $\mathbb{P}_{P^\star}[\cdot \mid \mathscr{D}]$
Output: Policy with a maximized safe return estimate
1  Initialize current policy
     $\pi_0 \leftarrow \arg\max_\pi \rho(\pi, \mathbb{E}_{P^\star}[P^\star \mid \mathscr{D}])$;
2  Initialize current value $v_0 \leftarrow v^{\pi_0}_{\mathbb{E}_{P^\star}[P^\star \mid \mathscr{D}]}$;
3  Initialize value robustness set $\mathscr{V}_0 \leftarrow \{v_0\}$ ;
4  Construct $\mathscr{P}_0$ optimal for $\mathscr{V}_0$ (Algorithm 3);
5  Initialize counter $k \leftarrow 0$;
6  while Eq. (5) is violated with $\mathscr{V} = \{v_k\}$ do
7      Include $v_k$ that violates Eq. (5):
         $\mathscr{V}_{k+1} \leftarrow \mathscr{V}_k \cup \{v_k\}$ ;
8      Construct $\mathscr{P}_{k+1}$ optimized for $\mathscr{V}_{k+1}$
         (Algorithm 3);
9      Compute optimistic value function $v_{k+1}$ and
         policy $\pi_{k+1}$ for $\mathscr{P}_{k+1}$;
10     $k \leftarrow k + 1$ ;
11 return $(\pi_k, p_0^\mathsf{T} v_k)$ ;
```

To illustrate how OFVF computes a tighter upper bound than the standard Bayesian approach, consider the following simple example. Given a random variable X that is normally distributed with a 0 mean, two possible 95% confidence intervals are 1) the mean-centered interval $[-1.96, 1.96]$, and 2) left-biased interval $[-\infty, 1.64]$. The robust estimates on X with respect to the left-biased interval is a tighter 1.64 instead of 1.96. We apply this intuition in the context of the multi-dimensional transition probabilities, in which the effect can be more dramatic.

OFVF, as described in Algorithm 2, is not guaranteed to converge in finite time. 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. We discuss how to construct an ambiguity set in (5) below in Algorithm 3.

Now we describe how the plausibility set is constructed for each fixed set $\mathscr{V}$ in order to satisfy (5). This is needed in Lines 3 and 7 in Algorithm 2. Algorithm 3 describes this procedure. The algorithm is based on the two main insights. First, if the set $\mathscr{V}$ is a singleton, it is easy to construct an optimal ambiguity set.

$$g^\star_v = \max\left\{ g \;:\; \mathbb{P}_{P^\star}[g \leqslant v^\mathsf{T} p^\star_{s,a}] \geqslant 1 - \frac{\delta}{SA} \right\} \tag{6}$$

This optimization problem can be solved easily by SAA. That is, we sample points q_i from the probability distribution over $p^\star_{s,a}$ and then sort them by $v^\mathsf{T} q_i$ and let k be the $1 - \delta/(SA)$ quantile of this sample.

Algorithm 3: Compute ambiguity set for a posterior distribution with a set of possible value functions

```
Input: Discrete set of value functions $\mathscr{V}$, confidence
       level $\delta$, and distribution $\mathbb{P}_{P^\star}[\cdot \mid \mathscr{D}]$
Output: $L_1$ ambiguity set optimized for (5) with $\mathscr{V}$
for All states and actions: $s, a \in \mathscr{S} \times \mathscr{A}$ do
    Compute $g^\star_v$ for each $v \in \mathscr{V}$ by solving (6) ;
    Compute $L_1$ set size $\psi^\star_{s,a}(\mathscr{V})$ as the value of (7)
      and let $\bar{p}^\star_{s,a}$ be the minimizer;
return $\mathscr{P}_{s,a} = \{p \in \Delta^S \;:\; \|p - \bar{p}_{s,a}\|_1 \leqslant \psi_{s,a}\}$ ;
```

When $\mathscr{V}$ is a singleton, it is sufficient for the plausibility set to be a be a subset of the hyperplane $\{p \in \Delta^S \;:\; v^\mathsf{T} p = g^\star_v\}$. This is a very important difference compared with the confidence interval approach. In fact, for a continuous posterior distribution of $P^\star$, any ambiguity set that is a subset of the hyperplane described above will have a probability of 0. This is significant departure from the traditional methods for constructing ambiguity set which was first proposed and analyzed in detail in [Gupta, 2015]. It is important to note, however, that our setting is very different and the results of [Gupta, 2015] are not directly applicable.

Now for the case 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 satisfies for every $v \in \mathscr{V}$.

$$\begin{aligned} \psi^\star_{s,a}(\mathscr{V}) = \min_{p\in\Delta^S} \max_{v\in\mathscr{V}} &\|q_v - p\|_1 \\ \text{s.t. } v^\mathsf{T} q_v &= g^\star_v, q_v \in \Delta^S,\ \forall v \in \mathscr{V} \end{aligned} \tag{7}$$

The optimization in (7) can be represented as a linear program. 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 evaluate the estimated returns over episodes computed by Bayes UCRL and OFVF empiri-

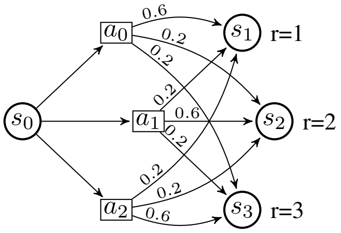

Figure 1: Simple MDP with one non-terminal state, three actions leading to three terminal states with different transition probabilities

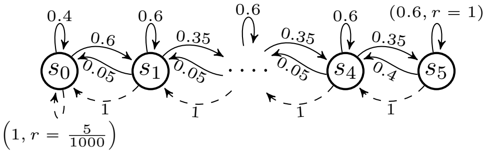

Figure 2: RiverSwim problem with six states $(s_0, \ldots, s_5)$ and 2 actions (left- dashed arrow, right- solid arrow)

cally. 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 "UCRL" and "PSRL" algorithms. UCRL simply constructs the distribution free plausibility set by applying Chernoff-hoeffding inequality. PSRL does not construct any explicit plausibility set. Rather, it is a stochastically optimistic algorithm and it proceeds in a manner similar to Thompson sampling. PSRL samples a single statistically plausible MDP in proportion to the likelihood from the posterior. Optimism at each state-action independently makes OFU algorithms far too optimistic. PSRL can overcome this problem [Osband and Van Roy, 2017] and outperforms all OFU algorithms including OFVF in average case performance. But as we will see in the experiments, OFVF performs robustly in the worst case scenario and outperforms all other methods.

Next in section Section 5.1, we compare the methods in a simplified setting in which we consider the problem of estimating the value of a single state from a Bellman update. Then in the following sections, we evaluate the methods on different problems involving challenging exploration in full MDP setup.

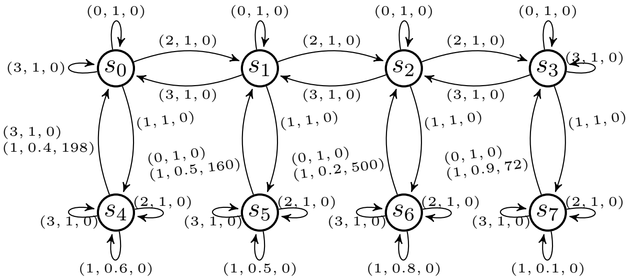

Figure 3: CasinoLand problem with eight states $(s_0, \ldots, s_7)$ and 3 actions (0,1,2). Transitions are shown with arrows, each labeled with (action, transition probability, reward).

5.1 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. The dynamics of the MDP is shown in Fig. 1. We consider an uninformative Dirichlet prior for the transition probabilities. We run the experiments for 100 episodes, each episode

Fig. 4 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-case return. As expected, PSRL outperforms all other methods in average case, but performs poorly in the worst-case scenario.

5.2 RiverSwim Problem

We compare the performance different methods in standard example of RiverSwim [Osband *et al.*, 2013; Strehl and Littman, 2004]. RiverSwim consists of a chain of six states as shown in Fig. 2. The agent starts at the left-most state s_0. The agent can take one of two action: left or right. Strong current flows in the river from right to left. Swimming left thus always successful and takes the agent to the state on the left with probability 1. Swimming right goes against the current and often fails. The agent receives a very small reward for reaching leftmost state. But the rightmost state gives a much higher reward and the optimal policy is to swim right. Efficient exploration is required in this problem to find the optimal policy.

We start with a uniform Dirichlet prior over the transition probabilities. We run the experiments for 500 episodes with each containing 100 Monte Carlo simulations. We compute the worst and average case cumulative regrets over the

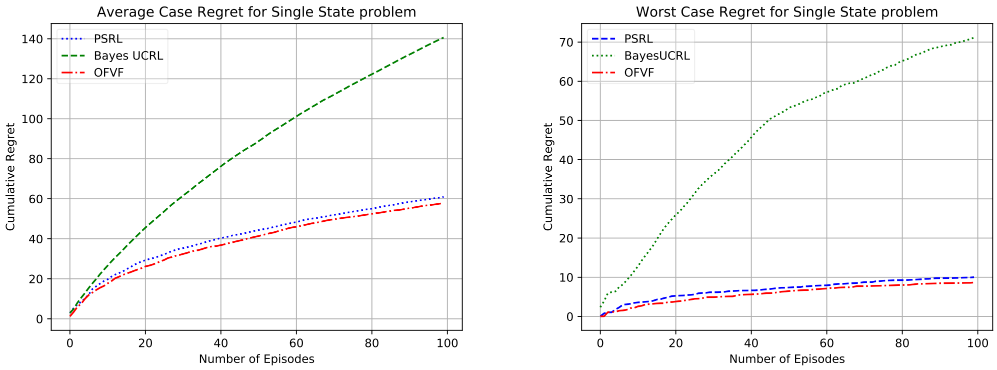

Figure 4: Cumulative regret for the single state simple problem. Left) average cumulative regret, Right) Worst-case cumulative regret.

episodes. Fig. 5 shows that, OFVF outperforms all other methods in terms of worst-case cumulative regret.

5.3 CasinoLand Problem

We compare the performance of the algorithms on another challenging exploration environment, CasinoLand [Strehl and Littman, 2004]. CasinoLand consists of eight rooms in a two-by-four grid as shown in Fig. 3. The transition model of the agent is shown in the picture. On the top row, the agent can move freely to any direction. Actions are less rewarding on the top row, with rewards basically zero for all the movements. On the bottom row, no horizontal movement is allowed. Taking an specific action (conceptually pulling a lever) yields a large reward with small probability. The MDP requires hard exploration because the lever with smallest probability of generating reward is the best lever to pull.

We start with a uniform Dirichlet prior over the transition probabilities. We run the experiments for 500 episodes with each containing 100 Monte Carlo simulations. We compute the worst and average case cumulative regrets over the episodes. [Plot reference] shows that, OFVF outperforms all other methods in terms of worst-case cumulative regret.

5.4 Mountain Car Problem

Next, we run experiments on the MountainCar problem from OpenAI Gym [Brockman *et al.*, 2016]. The problem is about driving an under-powered car to the top (position=0.5) of a one-dimensional hill on the right side of the car. There is another hill on the left side of the car. There are two actions, going left or right. Climbing the hill on the left of the car can give some potential energy and accelerate the car towards the target on the top of the right hill.

We start with a uniform Dirichlet prior for transition over the nearby states of each individual state. We compute the reference optimal solution by running Q-Learning for a very large number (10^5) of episodes. We run our algorithms for hundreds of episodes each containing hundreds of consequent runs. We compare both the average-case and worst-case cumulative regret. [reference] plot shows that, OFVF outperforms all other methods in the worst case scenario.

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, Thomas Jaksch, and Ronald Ortner. Near-optimal Regret Bounds for Reinforcement Learning. In D Koller, D Schuurmans, Y Bengio, and L Bottou, editors, *Advances in Neural Information Processing Systems 21*. 2009.

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*. 1996.

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

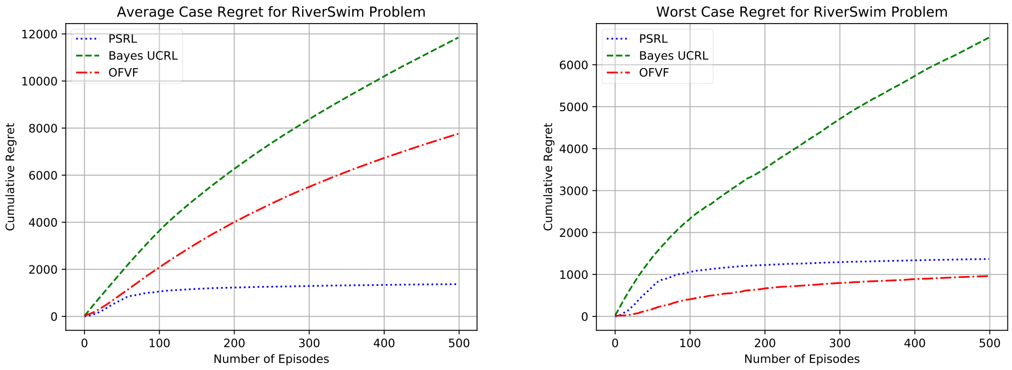

Figure 5: Cumulative regret for the RiverSwim problem. Left) average cumulative regret, Right) Worst-case cumulative regret.

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

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

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

Vishal Gupta. Near-Optimal Bayesian Ambiguity Sets for Distributionally Robust Optimization. 2015.

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.

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 Ronny Luss. Interpretable Policies for Dynamic Product Recommendations. In *Uncertainty in Artificial Intelligence (UAI)*, 2016.

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

A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming: Modeling and theory*. 2014.

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

Alexander L Strehl and Michael L Littman. *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*. 1998.

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.

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

Wolfram Wiesemann. Robust Markov decision processes. *. . . of Operations Research*, pages 1–52, 2013.

A Technical Results

A.1 Computing Bayesian Confidence Interval

Algorithm 4: Bayesian Confidence Interval (BCI)

```
Input: Distribution $\theta$ over $p^{\star}_{s,a}$, confidence level $\delta$, 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 $\theta$: $X_i \sim \theta$;
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 $\psi_{s,a}$;
```