Tight Bayesian Ambiguity Sets for Robust MDPs

Reazul Hasan Russel **Marek Petrik**
Department of Computer Science
University of New Hampshire
Durham, NH 03824 USA
rrussel, mpetrik at cs.unh.edu

Abstract

Robustness is important for sequential decision making in a stochastic dynamic environment with uncertain probabilistic parameters. We address the problem of using robust MDPs (RMDPs) to compute policies with provable worst-case guarantees in reinforcement learning. The quality and robustness of an RMDP solution is determined by its ambiguity set. Existing methods construct ambiguity sets that lead to impractically conservative solutions. In this paper, we propose RSVF, which achieves less conservative solutions with the same worst-case guarantees by 1) leveraging a Bayesian prior, 2) optimizing the size and location of the ambiguity set, and, most importantly, 3) relaxing the requirement that the set is a confidence interval. Our theoretical analysis shows the safety of RSVF, 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]. MDPs assume that transition probabilities are known precisely, but this is rarely the case in reinforcement learning. Errors in transition probabilities often results in policies that are brittle and fail in real-world deployments. A promising framework for robust reinforcement learning are robust MDPs (RMDPs) which assume that the transition probabilities and/or rewards are not known precisely. Instead, they can take on any value from a so-called *ambiguity set* which represents a set of plausible values [Xu and Mannor, 2006, 2009; Mannor *et al.*, 2012; Petrik, 2012; Hanasusanto and Kuhn, 2013; Tamar *et al.*, 2014; Delgado *et al.*, 2016; Petrik *et al.*, 2016]. The choice of an ambiguity set determines the trade-off between robustness and average performance of an RMDP.

The main contribution of this paper is RSVF, a new *data-driven* Bayesian approach to constructing *ambiguity* sets for RMDPs. The method computes policies with tighter safe estimates (Definition 2.1) by introducing two new ideas. First, it is based on Bayesian posterior distributions rather than distribution-free bounds. Second, RSVF does not construct ambiguity sets as simple confidence intervals. Confidence intervals as ambiguity sets are a sufficient but not a necessary condition. RSVF uses the structure of the value function to optimize the *location* and *shape* of the ambiguity set to guarantee lower bounds directly without necessarily enforcing the requirement for the set to be a confidence interval.

2 Problem Statement: Data-driven RMDPs

We propose to use Robust Markov Decision Processes (RMDPs) with states $\mathscr{S} = \{1, \ldots, S\}$ and actions $\mathscr{A} = \{1, \ldots, A\}$ to compute a policy with the maximal *safe* estimate of return.

Infer to Control: NIPS 2018 Workshop on Probabilistic Reinforcement Learning and Structured Control.

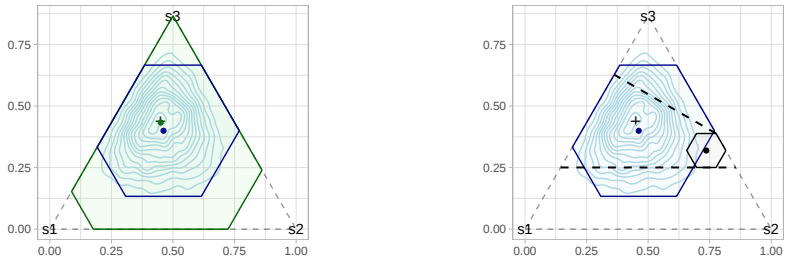

Figure 1: Comparison of 90% L_1 confidence intervals, Left: Hoeffding (green) and Bayesian (blue), Right: RSVF (green) and BCI (blue).

Definition 2.1 (Safe Estimate of Return)**.** We say that an estimate of policy return $\tilde{\rho} : \Pi \to \mathbb{R}$ is *safe* with probability δ for a given dataset $\mathscr{D} \subseteq \{(s, a, s') \ : \ s, s' \in \mathscr{S}, a \in \mathscr{A}\}$, if it satisfies: $\mathbb{P}_{P^\star}\big[\tilde{\rho}(\pi) \leqslant \rho(\pi, P^\star) \big| \mathscr{D}\big] \geqslant 1-\delta$, for each stationary deterministic policy π. Here $P^\star : \mathscr{S} \times \mathscr{A} \to \Delta^{\mathscr{S}}$ is the true, but unknown transition probabilities, and $\rho(\pi)$ is the return for a policy π.

In standard *batch* RL setting, $\mathscr{D}$ can be used to estimate the transition probabilities, but is assumed to be not known precisely for the RMDP and is constrained to be in the *ambiguity set* $\mathscr{P}_{s,a}$, defined for each state and action (s,a-rectangular). The most common method for defining ambiguity sets is to use norm-bounded distance from a *nominal* probability distribution $\bar{p}$: $\mathscr{P}_{s,a} = \{p \in \Delta^S \ : \ \|p - \bar{p}_{s,a}\|_1 \leqslant \psi_{s,a}\}$ for a given $\psi_{s,a} \geqslant 0$ and a nominal point $\bar{p}_{s,a}$. We assume that the rewards $r : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ are known. The *objective* is to maximize the γ-discounted infinite horizon return.

RMDPs satisfy similar properties as regular MDPs [Iyengar, 2005; Tamar *et al.*, 2014]. The robust Bellman operator $\widehat{T}_{\mathscr{P}}$ is defined for a state s as: $(\widehat{T}v)(s) := \max_{a \in \mathscr{A}} \min_{p \in \mathscr{P}_{s,a}} (r_{s,a} + \gamma \cdot p^\mathsf{T} v)$. The robust return is defined as [Iyengar, 2005]: $\hat{\rho}(\pi) = \min_{P \in \mathscr{P}} \rho(\pi, P) = p_0^\mathsf{T} \hat{v}^\pi$, where $p_0 \in \Delta^S$ is the initial distribution. In general, we use hat $\hat{(\cdot)}$ to denote quantities in RMDP.

3 Ambiguity Sets as Confidence Intervals

In this section, we describe the standard approach to constructing ambiguity sets as distribution-free confidence intervals and propose its extension to the Bayesian setting.

Distribution-free Confidence Interval The use of distribution-free error bounds on the L_1 norm is common in reinforcement learning [Petrik *et al.*, 2016; Taleghan *et al.*, 2015; Strehl and Littman, 2004]. The confidence interval is constructed around the mean transition probability by combining the Hoeffding inequality with the union bound [Weissman *et al.*, 2003; Petrik *et al.*, 2016]. The Hoeffding ambiguity set is defined as: $\mathscr{P}^H_{s,a} = \left\{ \|p^\star_{s,a} - \bar{p}_{s,a}\|_1 \leqslant \sqrt{\frac{2}{n_{s,a}} \log \frac{SA2^S}{\delta}} \right\}$ where $\bar{p}_{s,a}$ is the mean transition probability computed from $\mathscr{D}$ and $n_{s,a}$ is the number of transitions observed originating from state s and an action a. An important limitation of $\mathscr{P}^H$ is that the size of the ambiguity set grows linearly with the number of states S.

Bayesian Confidence Interval (BCI) Here we assume that data $\mathscr{D}$ is available and a hierarchical Bayesian model can be used to infer a probability distribution over $P^\star$ analytically or using MCMC methods like Stan [Gelman *et al.*, 2014]. To construct the ambiguity set $\mathscr{P}^B$, we optimize for the *smallest* ambiguity set around the mean transition probability with the assumption that a smaller ambiguity set will lead to a tighter lower bound estimate. Formally, the optimization problem to compute $\psi_{s,a}$ for each state s and action a is: $\min_{\psi \in \mathbb{R}_+} \left\{ \psi \ : \ \mathbb{P}\left[\|p^\star_{s,a} - \bar{p}_{s,a}\|_1 > \psi \mid \mathscr{D} \right] < \frac{\delta}{SA} \right\}$,where nominal point is $\bar{p}_{s,a} = \mathbb{E}_{P^\star}[p^\star_{s,a} \mid \mathscr{D}]$.

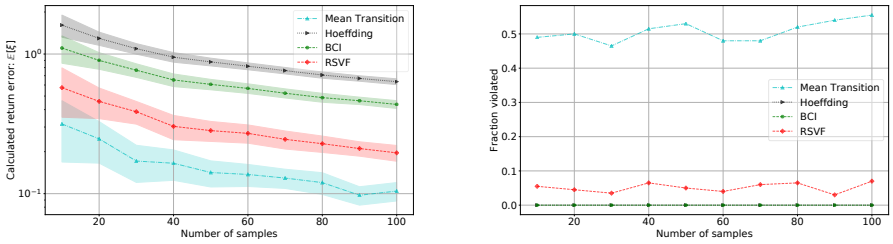

Figure 2: Single state with Dirichlet prior, return error with 95% Confidence & violations.

4 RSVF: Robustification With Sensible Value Functions

RSVF 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:

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

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 $\mathscr{P}$.

Algorithm 1: RSVF: Robustification with 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

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 $\mathscr{V}_0$;
5 Initialize counter $k \leftarrow 0$;
6 **while** *Eq.* (1) *is violated with* $\mathscr{V} = \{v_k\}$ **do**
7 Include v_k that violates Eq. (1): $\mathscr{V}_{k+1} \leftarrow \mathscr{V}_k \cup \{v_k\}$;
8 Construct $\mathscr{P}_{k+1}$ optimized for $\mathscr{V}_{k+1}$;
9 Compute robust value function v_{k+1} and policy π_{k+1} for $\mathscr{P}_{k+1}$;
10 $k \leftarrow k + 1$;
11 **return** $(\pi_k, p_0^\mathsf{T} v_k)$;

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

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

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:

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

In other words, we construct the set to minimize its radius while still intersecting the hyperplane for each v in $\mathscr{V}$. Algorithm 1, 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.

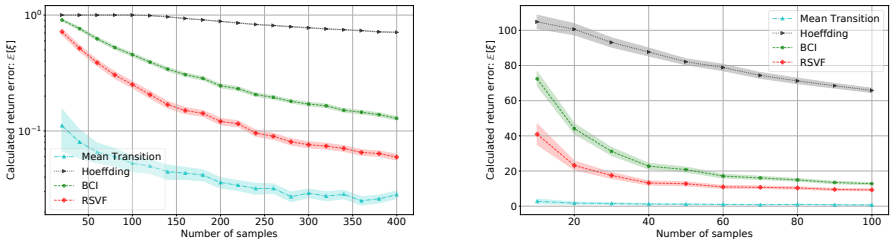

Figure 3: Return error with a Gaussian prior with 95% confidence, Left: Single state, Right: Full MDP, X-axis is the number of samples per state-action.

5 Empirical Evaluation

In this section, we evaluate the safe estimates computed by BCI and RSVF empirically. We assume a true model of each problem and generate a number of simulated data sets for the known distribution. We compute the largest safe estimate for the optimal return and compare it with the optimal return for the true model. We compare our results with "Hoeffding Inequality" based distance $\mathscr{P}^H$ and "Mean Transition" which simply solves the expected model $\bar{p}_{s,a}$ and provides no safety guarantees. 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.

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 Fig. 2 (left) and Fig. 3 (left) shows. As expected, the mean estimate provides the tightest bound, but Fig. 2 (right) illustrates that it does not provide any meaningful safety guarantees.

Full MDP with Informative Prior Next, we evaluate RSVF on a full MDP problem. Standard RL benchmarks, like cart-pole or arcade games, lack meaningful Bayesian priors. We instead use a simple exponential population model, based on the management of an invasive species [Taleghan *et al.*, 2015]. The population N_t of the invasive species at time t evolves according to the exponential dynamics $N_{t+1} = \min(\lambda_t N_t, K)$. Here, λ is the growth rate and K is the carrying capacity of the environment. A land manager needs to decide, at each time t, whether to take a control action which influences the growth rate λ. If z_t is the indicator of whether the control action was taken, the growth rate λ_t is defined as: $\lambda_t = \bar{\lambda} - z_t N_t \beta_1 - z_t \max(0, N_t - \bar{N})^2 \beta_2 + \mathcal{N}(0, \sigma_y^2)$, where β_1 and β_2 are the coefficients of control effectiveness. We also assume that we only observe y_t, a noisy estimate of population N_t: $y_t \sim N_t + \mathcal{N}(0, \sigma_y^2)$. In the MDP model, the population observation defines the state. There are two actions: to apply or not to apply the control measure. Transition probabilities are given by the population evolution function. The reward for the MDP captures the costs of high invasive population and the application of the treatment.

Fig. 3 (right) depicts the average predicted regret over the different datasets. The distribution-free methods are very conservative, BCI improves on this behavior somewhat, but RSVF provides bounds that are even tighter than BCI by almost a factor of 2. The rate of violations is 0 for all robust methods. This indicates that RSVF is overly conservative in this case too since its rate of violations is also close to 0. This is due its reliance on the union bound across multiple states, the approximate construction of the individual ambiguity sets, and the inherent rectangularity assumption.

6 Conclusion

We propose, in this paper, a new Bayesian approach to the construction of ambiguity sets in robust reinforcement learning. This approach has several important advantages over standard distribution-free methods used in the past. Our experimental results and theoretical analysis indicate that the Bayesian ambiguity sets can lead to much tighter safe return estimates.

References

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

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.

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

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.

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.

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

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.