Beyond Confidence Regions: Tight Bayesian Ambiguity Sets for Robust MDPs

Reazul Hasan Russel
Department of Computer Science
University of New Hampshire
rrussel@cs.unh.edu

Marek Petrik
Department of Computer Science
University of New Hampshire
mpetrik@cs.unh.edu

Abstract

Robust MDPs (RMDPs) can be used to compute policies with provable worst-case guarantees in reinforcement learning. The quality and robustness of an RMDP solution are determined by the ambiguity set—the set of plausible transition probabilities—which is usually constructed as a multi-dimensional confidence region. Existing methods construct ambiguity sets as confidence regions using concentration inequalities which leads to overly conservative solutions. This paper proposes a new paradigm that can achieve better solutions with the same robustness guarantees without using confidence regions as ambiguity sets. To incorporate prior knowledge, our algorithms optimize the size and position of ambiguity sets using Bayesian inference. Our theoretical analysis shows the safety of the proposed method, and the empirical results demonstrate its practical promise.

1 Introduction

Markov decision processes (MDPs) provide a versatile framework for modeling reinforcement learning problems [4, 33, 38]. However, they assume that transition probabilities and rewards are known exactly which is rarely the case. Limited data sets, modeling errors, value function approximation, and noisy data are common reasons for errors in transition probabilities [16, 30, 45]. This results in policies that are brittle and fail when implemented. This is particularly true in the case of *batch* reinforcement learning [18, 20, 23, 32, 42].

A promising framework for computing *robust policies* is based on Robust MDPs (RMDPs). RMDPs relax the need for precisely known transition probabilities. Instead, transition probabilities can take on any value from a so-called *ambiguity set* which represents a set of plausible transition probabilities [9, 14, 24, 29, 32, 40, 46, 47]. RMDPs are also reminiscent of dynamic zero-sum games: the decision maker chooses the best actions, while the adversarial nature chooses the worst transition probabilities from the ambiguity set.

The practical utility of using RMDPs has been hindered by the lack of good ways of constructing ambiguity sets that lead to solutions that are robust without being too conservative. The standard approach to constructing ambiguity sets from concentration inequalities [1, 32, 42, 44] leads to theoretical guarantees but provides solutions that hopelessly conservative. Many problem-specific methods have been proposed too, but they are hard to use and typically lack finite-sample guarantees [3, 5, 16, 28].

The main contribution of this work is to introduce a new method for constructing ambiguity sets that are both significantly *less conservative* than existing ones [21, 32, 42] and also provide strong finite-sample guarantees. Similarly to some prior work on robust reinforcement learning and optimization, we use Bayesian assumptions to take advantage of domain knowledge which is often available [7, 8, 13, 47]. Our main innovation is to realize that the natural approach to building ambiguity sets as confidence intervals is unnecessarily conservative. Surprisingly, in the Bayesian setting, using a

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

95% confidence region for the transition probabilities is unnecessarily conservative to achieve 95% confidence in the robustness of the solution. We also derive new L_1 concentration inequalities of possible independent interest.

The remainder of the paper is organized as follows. Section 2 formally describes the framework and goals of the paper. Section 3 describes our main contribution, RSVF, a new method for constructing tight ambiguity sets from Bayesian models that are adapted to the optimal policy. We provide theoretical justification for the robustness of RSVF, but detailed theoretical analysis of its performance guarantees is beyond the scope of this work. Then, Section 4 overviews related work and outlines methods that build ambiguity sets as frequentist confidence regions or Bayesian credible sets. Finally, Section 5 presents empirical results on several problem domains.

2 Problem Statement: Data-driven RMDPs

This section formalizes our goals and reviews relevant results for robust Markov decision processes (RMDPs). Throughout the paper, we use the symbol Δ^S to denote the probability simplex in $\mathbb{R}^S_+$. The symbols $\mathbf{1}$ and $\mathbf{0}$ denote vectors of all ones and zeros, respectively, of an appropriate size. The symbol $\mathbf{I}$ represents the identity matrix.

2.1 Safe Return Estimate: VaR

The underlying reinforcement learning problem is a Markov decision process with states $\mathcal{S} = \{1, \dots, S\}$ and actions $\mathcal{A} = \{1, \dots, A\}$. The rewards $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are known but the true transition probabilities $P^\star : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$ are unknown. The transition probability vector for a state s and an action a is denoted by $p^\star_{s,a}$. As this is a *batch* reinforcement learning setting, a fixed dataset $\mathcal{D}$ of transition samples is provided: $\mathcal{D} = (s_i \in \mathcal{S}, a_i \in \mathcal{A}, s'_i \in \mathcal{S})_{i=1,\dots,m}$. The only assumption about $\mathcal{D}$ is that the state s' in $(s, a, s') \in \mathcal{S}$ is distributed according to the *true* transition probabilities $s' \sim P^\star(s, a, \cdot)$, no assumptions are made on the sampling policy. Note that in the Bayesian approach, $P^\star$ is a random variable and we assume to have a prior distribution available.

The objective is to maximize the standard γ-discounted infinite horizon return [33]. Because this paper analyzes the impact of using different transition probabilities, we use a subscript to indicate which ones are used. The optimal value function for some transition probabilities P is, therefore, denoted as $v^\star_P : \mathcal{S} \to \mathbb{R}$, and the value function for a *deterministic policy* $\pi : \mathcal{S} \to \mathcal{A}$ is denoted as v^π_P. The set of all deterministic stationary policies is denoted by Π. The total return $\rho(\pi, P)$ of a policy π under transition probabilities P is:

$$\rho(\pi, P) = p_0^\mathsf{T} v^\pi_P,$$

where p_0 is the initial distribution.

Ideally, we could compute a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the return $\rho(\pi, P^\star)$, but $P^\star$ is unknown. Ignoring the uncertainty in $P^\star$ completely leads to brittle policies. Instead, a common objective in robust reinforcement learning is to maximize a plausible *lower-bound* on the return. Having a safe return estimate is very important since it can inform the stakeholder that the policy may not be good enough when deployed. The objective of computing a policy π that maximizes a *high-confidence* lower bound on the return can be expressed as [8, 21, 31, 42]:

$$\max_{\pi \in \Pi} \mathrm{V@R}^\delta_{P^\star}[\rho(\pi, P^\star)], \tag{1}$$

where $\mathrm{V@R}^\delta$ is the popular value-at-risk measure at a risk level δ [35]. This objective is also sometimes known as *percentile optimization* [8]. It is important to note that the risk metric is applied over possible values of the uncertain parameter and not over the distribution of returns. For example, if $\mathrm{V@R}^{0.05}_{P^\star}[\rho(\pi, P^\star)] = -1$ then for 5% of uncertain transition probabilities $P^\star$, the return is -1 or smaller.

Because solving the optimization problem in (1) is NP-hard [8], we instead maximize a lower bound $\tilde{\rho}(\pi)$. We call this lower bound a *safe return estimate* and it is defined as follows.

Definition 2.1 (Safe Return Estimate)**.** The estimate $\tilde{\rho} : \Pi \to \mathbb{R}$ of return is called *safe* for a policy π with probability $1 - \delta$ if $\tilde{\rho}(\pi) \leq \mathrm{V@R}^\delta_{P^\star}[\rho(\pi, P^\star)]$, or in other words if it satisfies:

$$\mathbb{P}_{P^\star}\left[\tilde{\rho}(\pi) \leq \rho(\pi, P^\star) \,\middle|\, \mathcal{D}\right] \geq 1 - \delta\,.$$

Recall that under Bayesian assumptions, $P^\star$ is a random variable and the guarantees are conditional on the dataset $\mathcal{D}$. This is different from the frequentist approach, in which the random variable is $\mathcal{D}$ and the guarantees are conditional on $P^\star$. The relative merits of Bayesian versus frequentist approaches to robust optimization have been discussed in earlier work [8, 47], but we emphasize that each approach presents a different set of advantages. An insightful discussion of the differences between the two approaches can be found, for example, in Sections 5.2.2 and 6.1.1 of Murphy (2012).

The following example will be used throughout the paper to demonstrate the proposed methods and visualize simple ambiguity sets.

Example 2.1. Consider an MDP with 3 states: s_1, s_2, s_3 and a single action a_1. Assume that the true, but unknown, transition probability is $P^\star(s_1, a_1, \cdot) = [0.3, 0.2, 0.5]$. The known prior distribution over $p^\star_{s_1,a_1}$ is Dirichlet with concentration parameters $\alpha = (1, 1, 1)$. The dataset $\mathcal{D}$ is comprised of 3 occurrences of transitions (s_1, a_1, s_1), 2 of transitions (s_1, a_1, s_2), and 5 of transitions (s_1, a_1, s_3). The posterior distribution over $p^\star_{s_a,a_1}$ is also Dirichlet with $\alpha = (4, 3, 6)$. Note that this is a probability distribution over transition probability distributions. Fig. 1 depicts the posterior distribution projected onto the probability simplex along with a 90% confidence region centered on the posterior mean.

2.2 Robust MDPs

Robust Markov Decision Processes (RMDPs) are a convenient model and tractable model that generalizes MDPs. We will use RMDPs to maximize a tractable lower bound on V@R objective in (1) and compute a *safe* return estimate. Our RMDP model has the same states $\mathcal{S}$, actions $\mathcal{A}$, rewards $r_{s,a}$ as the MDP. The transition probabilities for each state s and action a, denoted as $p_{s,a} \in \Delta^S$, are assumed chosen adversarialy from an *ambiguity set* $\mathcal{P}_{s,a}$. We use $\mathcal{P}$ to refer cumulatively to $\mathcal{P}_{s,a}$ for all states s and actions a.

We restrict our attention to sa-rectangular ambiguity sets, which allow the adversarial nature to choose the worst transition probability independently for each state and action [22, 45]. Limitations of rectangular ambiguity sets are known well [12, 25, 43] but they represent a simple, tractable, and practical model. A convenient way of defining ambiguity sets is to use a norm-distance from a given *nominal transition probability* $\bar{p}_{s,a}$:

$$\mathcal{P}_{s,a} = \left\{ p \in \Delta^S \,:\, \|p - \bar{p}_{s,a}\|_1 \le \psi_{s,a} \right\} \tag{2}$$

for a given $\psi_{s,a} \ge 0$ and a nominal point $\bar{p}_{s,a}$. We focus on ambiguity sets defined by the L_1 norm because they give rise to RMDPs that can be solved very efficiently [15].

RMDPs have properties that are similar to regular MDPs (see, for example, [2, 19, 22, 28, 45]). The robust Bellman operator $\widehat{T}_\mathcal{P}$ for an ambiguity set $\mathcal{P}$ for a state s computes the best action with respect to the worst-case realization of the transition probabilities:

$$(\widehat{T}_\mathcal{P} v)(s) := \max_{a \in \mathcal{A}} \min_{p \in \mathcal{P}_{s,a}} \left(r_{s,a} + \gamma \cdot p^\mathsf{T} v \right) \tag{3}$$

The symbol $\widehat{T}^\pi_\mathcal{P}$ denotes a robust Bellman update for a given *stationary* policy π. The optimal robust value function $\hat{v}^\star$, and the robust value function $\hat{v}^\pi$ for a policy π are unique and must, similarly to MDPs, satisfy $\hat{v}^\star = \widehat{T}_\mathcal{P} \hat{v}^\star$ and $\hat{v}^\pi = \widehat{T}^\pi_\mathcal{P} \hat{v}^\pi$. In general, we use a hat to denote quantities in the RMDP and omit it for the MDP. When the ambiguity set $\mathcal{P}$ is not obvious from the context, we use it as a subscript $\hat{v}^\star_\mathcal{P}$. The robust return $\hat{\rho}$ is defined as [16]:

$$\hat{\rho}(\pi, \mathcal{P}) = \min_{P \in \mathcal{P}} \rho(\pi, P) = p_0^\mathsf{T} \hat{v}^\pi_\mathcal{P} \,,$$

where $p_0 \in \Delta^S$ is the initial distribution. In the remainder of the paper, we describe methods that construct $\mathcal{P}$ from $\mathcal{D}$ in order to guarantee that $\hat{\rho}$ is a tight lower bound on V@R of the returns.

3 Optimized Bayesian Ambiguity Sets

In this section, we describe the new algorithm for constructing Bayesian ambiguity sets that can compute less-conservative lower bounds on the return. RSVF (robustification with sensible value functions) is a Bayesian method that uses samples from the posterior distribution over $P^\star$ to construct tight ambiguity sets.

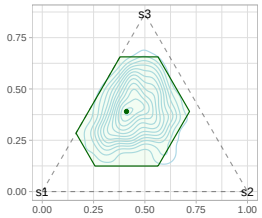

Figure 1: Contours of the posterior distribution and the 90%-confidence region.

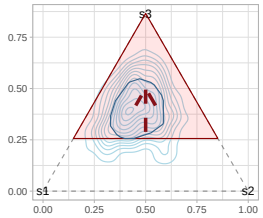

Figure 2: Optimal Bayesian ambiguity set (red) for a value function $v = (0, 0, 1)$.

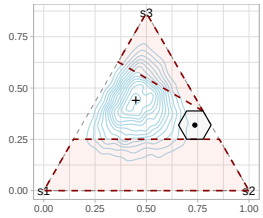

Figure 3: Sets $\mathcal{K}_{s_1,a_1}(v_i)$ (dashed red) for $i = 1, 2$ and $\mathcal{L}_{s_1,a_1}(\{v_1, v_2\})$ (black).

Before describing the algorithm, we use the setting of Example 2.1 to motivate our approach. To minimize distractions by technicalities, assume that the goal is to compute the return for a *single* time step starting from state s_1. Assume also that the value function $v = (1, 0, 0)$ is known, all rewards from s_1 are 0, and $\gamma = 1$. Recall that our goal is to construct a safe return estimate $\tilde{\rho}(\pi)$ of $\mathrm{V@R}^{0.1}_{P^\star}[\rho(\pi, P^\star)]$ at the 90% level. When the value function is known, it is possible to construct the *optimal* ambiguity set $\mathcal{P}^\star$ such that $\hat{\rho}(\pi) = \min_{p\in\mathcal{P}^\star} p^\mathsf{T} v = \mathrm{V@R}^{0.1}_{P^\star}[\rho(\pi, P^\star)]$ as:

$$\mathcal{P}^\star = \left\{ p \in \Delta^3 \,:\, p^\mathsf{T} v \geq \mathrm{V@R}^{0.1}_{P^\star}[\rho(\pi, P^\star)] \right\}.$$

It can be shown readily that this ambiguity set is optimal in the sense that any set for which $\tilde{\rho}(\pi)$ is exact must be a subset of $\mathcal{P}^\star$ [13]. Fig. 2 depicts the optimal ambiguity set along with the arrow that indicates the direction along which v increases.

The optimal ambiguity set described above cannot be used directly, unfortunately, because the value function is unknown. It would be tempting to construct the ambiguity set as the *intersection* of optimal sets for all possible value functions; a polyhedral approximation of this set is shown in Fig. 2 using a blue color. Unfortunately, this approach is not (usually) correct and will not lead to a safe return estimate. This can be shown from the fact that support functions to convex sets are convex and V@R is not a convex (concave) function [6, 34]; see Gupta (2015) for a more detailed discussion.

Since it is not possible, in general, to simply consider the intersection of optimal ambiguity sets for all possible value functions, we approximate the optimal ambiguity set for a few reasonable value functions. For this purpose, we use a set $\mathcal{K}_{s,a}(v)$ defined as follows:

$$\mathcal{K}_{s,a}(v) = \left\{ p \in \Delta^S \,:\, p^\mathsf{T} v \leq \mathrm{V@R}^{\zeta}_{P^\star}\left[(p^\star_{s,a})^\mathsf{T} v\right] \right\},$$

where $\zeta = 1 - \delta/(SA)$. The bottom dashed set in Fig. 3 depicts this set $\mathcal{K}$ for $v = (0, 0, 1)$ in Example 2.1. The intuition behind this construction is as follows. If any ambiguity set $\mathcal{P}_{s,a}$ intersects $\mathcal{K}_{s,a}(\hat{v}^\pi_\mathcal{P})$ for each state s, a then the value function $\hat{v}^\pi_\mathcal{P}$ is safe: $\max_{p\in\mathcal{K}_{s,a}(v)} p^\mathsf{T} v \leq \mathrm{V@R}^{\zeta}_{P^\star}\left[(p^\star_{s,a})^\mathsf{T} v\right]$. See Lemma B.2 for the formal statement.

The set $\mathcal{K}_{s,a}(v)$ is sufficient, when the value function is known, but we need to generalize the approach to unknown value functions. The set $\mathcal{L}_{s,a}(\mathcal{V})$ provides such a guarantee for a set of possible value functions (POV) $\mathcal{V}$. Its center is chosen to minimize its size while intersecting $\mathcal{K}_{s,a}(v)$ for each v in $\mathcal{V}$ and is constructed as follows.

$$\begin{aligned} &\mathcal{L}_{s,a}(\mathcal{V}) = \left\{ p \in \Delta^S \,:\, \|p - \theta_{s,a}(\mathcal{V})\|_1 \leq \psi_{s,a}(\mathcal{V}) \right\} \\ &\psi_{s,a}(\mathcal{V}) = \min_{p\in\Delta^S} f(p), \quad \theta_{s,a}(\mathcal{V}) \in \arg\min_{p\in\Delta^S} f(p), \quad f(p) = \max_{v\in\mathcal{V}} \min_{q\in\mathcal{K}_{s,a}(v)} \|q - p\|_1 \end{aligned} \tag{4}$$

The optimization in (4) can be represented and solved as a linear program. Fig. 3 shows the set $\mathcal{L}$ in black solid color. It is the smallest L_1-constrained set that intersects the two $\mathcal{K}$ sets for value functions $v_1 = (0, 0, 1)$ and $v_2 = (2, 1, 0)$ in Example 2.1.

We are now ready to describe RSVF, which is outlined in Algorithm 1. RSVF takes an optimistic approach to approximating the optimal ambiguity set. It starts with a small set of potential optimal value functions (POV) and constructs an ambiguity set that is safe for these value functions. It keeps increasing the POV set until $\hat{v}^\star$ is in the set and the policy is safe. To simplify presentation,

Algorithm 1: RSVF: Adapted Ambiguity Sets

Input: Confidence $1-\delta$ and posterior $\mathbb{P}_{P^\star}[\cdot \mid \mathcal{D}]$
Output: Policy π and lower bound $\tilde{\rho}(\pi)$

$k \leftarrow 0$;
Pick some initial value function $\hat{v}_0$;
Initialize POV: $\mathcal{V}_0 \leftarrow \emptyset$;
repeat
  Augment POV: $\mathcal{V}_{k+1} \leftarrow \mathcal{V}_k \cup \{v_k\}$;
  For all s, a update $\mathcal{P}_{s,a}^{k+1} \leftarrow \mathcal{L}_{s,a}(\mathcal{V}_{k+1})$;
  Solve $\hat{v}_{k+1} \leftarrow \hat{v}^\star_{\mathcal{P}_{k+1}}$ and $\hat{\pi}_{k+1} \leftarrow \hat{\pi}^\star_{\mathcal{P}_{k+1}}$;
  $k \leftarrow k+1$;
until *safe for all* s, a*:* $\mathcal{K}_{s,a}(\hat{v}_k) \cap \mathcal{P}_{s,a}^k \neq \emptyset$;
return $(\hat{\pi}_k, p_0^\mathsf{T}\hat{v}_k)$;

Algorithm 1 is not guaranteed to terminate in finite time; the actual implementation switches to BCI described in Section 4.2 after 100 iterations, which guarantees its termination.

The following theorem states that Algorithm 1 produces a safe estimate of the true return.

Theorem 3.1. *Suppose that Algorithm 1 terminates with a policy $\hat{\pi}_k$ and a value function $\hat{v}_k$ in the iteration k. Then, the return estimate $\tilde{\rho}(\hat{\pi}) = p_0^\mathsf{T}\hat{v}_k$ is safe:* $\mathbb{P}_{P^\star}\left[p_0^\mathsf{T}\hat{v}_k \leq p_0^\mathsf{T} v_{P^\star}^{\hat{\pi}_k} \;\middle|\; \mathcal{D}\right] \geq 1-\delta$.

Before discussing the proof of Theorem 3.1, it is important to mention its limitations. This result shows only that the return estimate $\hat{\rho}$ is safe; it does not show that it is good. There are, of course, naive safe estimates such as $\tilde{\rho}(\pi) = (1-\gamma)^{-1}\min_{s,a} r_{s,a}$. Since RSVF tightly approximates the optimal ambiguity sets, we expect it to perform significantly better. The theoretical analysis of this of the approximation error of $\hat{\rho}$ is beyond the scope of this work and we present empirical evidence in Section 5 instead.

All proofs can be found in Appendix B. The proof is technical but conceptually simple. It is based on two main properties. The first one is the construction of optimal ambiguity sets for the known value function as outlined above. The second is the fact that the ambiguity set needs to be robust with only with respect to the robust value function $\hat{v}$ and *not* the optimal value function $v^\star$. This is subtle, but *crucial* since $\hat{v}$ is a constant while $v^\star$ is a random variable in the Bayesian setting. The RSVF approach, therefore, does not work when frequentist guarantees are required. Confidence regions, described in Section 4, are designed for situations when robustness is required with respect to a random variable, and are therefore overly conservative in our setting. See Appendix E for more in-depth discussion.

4 Ambiguity Sets as Confidence Regions

In this section, we describe the standard approach to constructing ambiguity sets as multidimensional confidence regions and propose its extension to the Bayesian setting. Confidence regions derived from concentration inequalities have been used previously to compute bounds on the true return in off-policy policy evaluation [41, 42]. These methods, unfortunately, do not readily generalize to the policy optimization setting, which we target. Other work has focused on reducing variance rather than on high-probability bounds [18, 23, 26]. Methods for exploration in reinforcement learning, such as MBIE or UCRL2, also construct ambiguity sets using concentration inequalities [10, 17, 37, 37, 39] and compute optimistic (upper) bounds to guide exploration.

4.1 Distribution-free (Frequentist) Confidence Interval

Distribution-free confidence regions are used widely in reinforcement learning to achieve robustness [32, 42] and to guide exploration [36, 39]. The confidence region is constructed around the mean transition probability by combining the Hoeffding inequality with the union bound [32, 44].

We refer to this set as a *Hoeffding confidence region* and define it as follows for each s and a:

$$\mathcal{P}^H_{s,a} = \left\{ p \in \Delta^S : \|p - \bar{p}_{s,a}\|_1 \leq \sqrt{\frac{2}{n_{s,a}} \log \frac{SA2^S}{\delta}} \right\},$$

where $\bar{p}_{s,a}$ is the mean transition probability computed from $\mathcal{D}$ and $n_{s,a}$ is the number of transitions in $\mathcal{D}$ originating from state s and an action a.

Theorem 4.1. *The robust value function $\hat{v}_{\mathcal{P}^H}$ for the ambiguity set $\mathcal{P}^H$ satisfies:*

$$\mathbb{P}_{\mathcal{D}} \left[\hat{v}^{\pi}_{\mathcal{P}^H} \leq v^{\pi}_{P^\star},\ \forall \pi \in \Pi \mid P^\star \right] \geq 1 - \delta\ . \tag{5}$$

In addition, if $\hat{\pi}^\star_{\mathcal{P}^H}$ is the optimal solution to the RMDP, then $p_0^\intercal \hat{v}^\star_{\mathcal{P}^H}$ is a safe *return estimate of $\hat{\pi}^\star_{\mathcal{P}^H}$.*

To better understand the limitations of using concentration inequalities, we compare with new, and significantly tighter, frequentist ambiguity sets. The size of $\mathcal{P}^H$ grows as a square root of the number of states because of the 2^S term. This means that the size of $\mathcal{D}$ must scale about quadratically with the number of states to achieve the same confidence. Under some restrictive assumptions, the ambiguity set can be shown to be:

$$\mathcal{P}^M_{s,a} = \left\{ p \in \Delta^S\ :\ \|p - \bar{p}_{s,a}\|_1 \leq \sqrt{\frac{2}{n_{s,a}} \log \frac{S^2 A}{\delta}} \right\}\ .$$

This auxiliary result is proved in Appendix C.1. We emphasize that the aim of this bound is to understand the limitations of distribution free bounds, and we use it even when the necessary assumptions are violated.

4.2 Bayesian Credible Region (BCI)

We now describe how to construct ambiguity sets from Bayesian credible (or confidence) regions. To the best of our knowledge, this approach has not been studied explicitly. The construction starts with a (hierarchical) Bayesian model that can be used to sample from the posterior probability of $P^\star$ given data $\mathcal{D}$. The implementation of the Bayesian model is irrelevant as long as it generates posterior samples efficiently. For example, one may use a Dirichlet posterior, or use MCMC sampling libraries like JAGS, Stan, or others [11].

The posterior distribution is used to optimize for the *smallest* ambiguity set around the mean transition probability. Smaller sets, for a fixed nominal point, are likely to result in less conservative robust estimates. The BCI ambiguity set is defined as follows:

$$\mathcal{P}^B_{s,a} = \left\{ p \in \Delta^S\ :\ \|p - \bar{p}_{s,a}\|_1 \leq \psi^B_{s,a} \right\}, \quad \bar{p}_{s,a} = \mathbb{E}_{P^\star}[p^\star_{s,a} \mid \mathcal{D}]\ .$$

There is no closed-form expression for the Bayesian ambiguity set size. It must be computed by solving the following optimization problem for each state s and action a:

$$\psi^B_{s,a} = \min_{\psi \in \mathbb{R}_+} \left\{ \psi\ :\ \mathbb{P}\left[\|p^\star_{s,a} - \bar{p}_{s,a}\|_1 > \psi \mid \mathcal{D} \right] < \frac{\delta}{SA} \right\}\ .$$

The nominal point $\bar{p}_{s,a}$ is fixed (not optimized) to preserve tractability. This optimization problem can be solved by the Sample Average Approximation (SAA) algorithm [35]. Algorithm 2, in the appendix, summarizes the sort-based method. The main idea is to sample from the posterior distribution and then choose the minimal size $\psi_{s,a}$ that satisfies the constraint. We assume that it is possible to draw enough samples from $P^\star$ that the sampling error becomes negligible. Because the finite-sample analysis of SAA is simple but tedious, we omit it.

Theorem 4.2. *The robust value function $\hat{v}_{\mathcal{P}^B}$ for the ambiguity set $\mathcal{P}^B$ satisfies:*

$$\mathbb{P}_{P^\star} \left[\hat{v}^{\pi}_{\mathcal{P}^B} \leq v^{\pi}_{P^\star},\ \forall \pi \in \Pi \mid \mathcal{D} \right] \geq 1 - \delta\ .$$

In addition, if $\hat{\pi}^\star_{\mathcal{P}^B}$ is the optimal solution to the RMDP, then $p_0^\intercal \hat{v}^\star_{\mathcal{P}^B}$ is a safe *return estimate of $\hat{\pi}^\star_{\mathcal{P}^B}$.*

The proof is provided in Appendix B. Similar to other results, this theorem only proves that the constructed lower bound on the return is safe. It does not address the tightness of the bound.

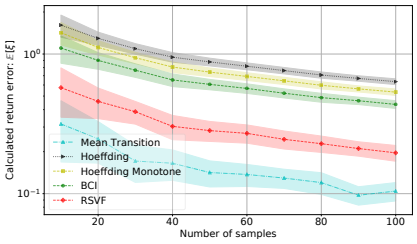

Figure 4: Expected regret of safe estimates with 95% confidence regions for the Bellman update with an uninformative prior.

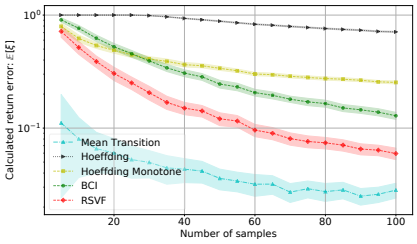

Figure 5: Expected regret of safe estimates with 95% confidence regions for the Bellman update with an informative prior.

5 Empirical Evaluation

In this section, we empirically evaluate the safe estimates computed using Hoeffding, BCI, and RSVF ambiguity sets. We start by assuming a true model and generate simulated datasets from it. Each dataset is then used to construct an ambiguity set and a safe estimate of policy return. The performance of the methods is measured using the average of the absolute errors of the estimates compared with the true returns of the *optimal* policies. All of our experiments use a 95% confidence for the safety of the estimates.

We compare ambiguity sets constructed using BCI, RSVF, with the Hoeffding sets. To reduce the conservativeness of Hoeffding sets when transition probabilities are sparse, we use a modification inspired by the Good-Turing bounds [39]. That is that any transitions from s, a to s' are impossible if they are not in $\mathcal{D}$. We also compare with the "Hoeffding Monotone" formulation $\mathcal{P}^M$ even when there is no guarantee that the value function is really monotone. Finally, we compare the results with the "Mean Transition" which solves the expected model $\bar{p}_{s,a}$ with no safety guarantees.

Next in 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, Section 5.2, evaluates the approach on an MDP with an informative prior.

We do not evaluate the computational complexity of the methods since they target problems constrained by data and not computation. The Bayesian methods are generally more computationally demanding but the scale depends significantly on the type of the prior model used. All Bayesian methods draw $1,000$ samples from the posterior for each state and action.

5.1 Bellman Update

In this section, we consider a transition from a single state s_0 and action a_0 to 5 states $s_1, \ldots, s_5$. The value function for the states $s_1, \ldots, s_5$ is fixed to be $[1, 2, 3, 4, 5]$. RSVF is run for a single iteration with the given value function. The single iteration of RSVF in this simplistic setting helps to quantify the possible benefit of using RSVF-style methods over BCI. The ground truth is generated from the corresponding prior for each one of the problems.

Uninformative Dirichlet Priors This setting considers a uniform Dirichlet distribution with $\alpha = [1, 1, 1, 1, 1]$ as the prior. This prior provides little information. Figure 4 compares the computed robust return errors. The value ξ represents the regret of predicted returns, which is the absolute difference between the *true* optimal value and the robust estimate: $\xi = |\rho(\pi^\star_{P^\star}, P^\star) - \tilde{\rho}(\hat{\pi}^\star)|$. Here, $\tilde{\rho}$ is the robust estimate and $\hat{\pi}^\star$ is the optimal robust solution. The smaller the value, the tighter and less conservative the safe estimate is. The number of samples is the size of dataset $\mathcal{D}$. All results are computed by averaging over 200 simulated datasets of the given size generated from the true $P^\star$. The results show that BCI improves on both types of Hoeffding bounds and RSVF further improves on BCI. The mean estimate provides the tightest bounds, but it does not provide any meaningful safety guarantees.

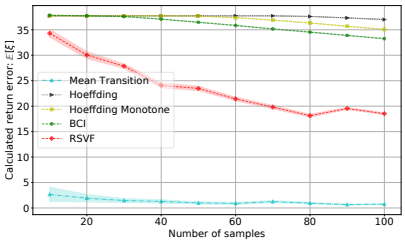

Figure 6: Expected regret of safe estimates with 95% confidence regions for the RiverSwim: an MDP with an uninformative prior.

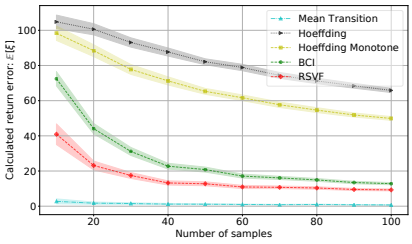

Figure 7: Expected regret of safe estimates with 90% confidence regions for the ExpPopulation: an MDP with an informative prior.

Informative Gaussian Priors To evaluate the effect of using an informative prior, we use a problem inspired by inventory optimization. The states $s_1, \ldots, s_5$ represent inventory levels. The inventory level corresponds to the state index (1 in the state s_1) except that the inventory in the current state s_0 is 5. The demand is assumed to be Normally distributed with an unknown mean μ and a *known* standard deviation $\sigma = 1$. The prior over μ is Normal with the mean $\mu_0 = 3$ and, therefore, the posterior over μ is also Normal. The current action assumes that no product is ordered and, therefore, only the demand is subtracted from s_0.

5.2 Full MDP

In this section, we evaluate the methods using MDPs with relatively small state-spaces. They can be used with certain types of value function approximation, like aggregation [30], but we evaluate them only on tabular problems to prevent approximation errors from skewing the results. To prevent the sampling policy from influencing the results, each dataset $\mathcal{D}$ has the same number of samples from each state.

Uninformative Prior We first use the standard RiverSwim domain for the evaluation [36]. The methods are evaluated identically to the Bellman update above. That is, we generate synthetic datasets from the ground truth and then compare the expected regret of the robust estimate with respect to the true return of the *optimal* policy for the ground truth. As the prior distribution, we use the uniform Dirichlet distribution over all states. Figure 6 shows the expected robust regret over 100 repetitions. The x-axis represents the number of samples in $\mathcal{D}$ for each state. It is apparent that BCI improves only slightly on the Hoeffding sets since the prior is not informative. RSVF, on the other hand, shows a significant improvement over BCI. All robust methods have safety violations of 0% indicating that even RSVF is unnecessarily conservative here.

Informative Prior Next, we evaluate RSVF on the MDP model of a simple exponential population model [43]. Robustness plays an important role in ecological models because they are often complex, stochastic, and data collection is expensive. Yet, it is important that the decisions are robust due to their long term impacts. Figure 7 shows the average regret of safe predictions. BCI can leverage the prior information to compute tighter bounds, but RSVF further improves on BCI. The rate of safety violations is again 0% for all robust methods.

6 Summary and Conclusion

This paper proposes new Bayesian algorithms for constructing ambiguity sets in RMDPs, improving over standard distribution-free methods. BCI makes it possible to flexibly incorporate prior domains knowledge and is easy to generalize to other shapes of ambiguity sets (like L_2) without having to prove new concentration inequalities. Finally, RSVF improves on BCI by constructing tighter ambiguity sets that are not confidence regions. Our experimental results and theoretical analysis indicate that the new ambiguity sets provide much tighter safe return estimates. The only drawbacks of the Bayesian methods are that they need priors and may increase the computational complexity.

Acknowledgments

We would like to thank Vishal Gupta and the anonymous referees for their insightful comments and suggestions. This work was supported by NSF under grants number 1815275 and 1717368.

References

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

[2] Bagnell, J. A., Ng, A. Y., and Schneider, J. G. Solving Uncertain Markov Decision Processes. *Carnegie Mellon Research Showcase*, pp. 948–957, 2001.

[3] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust Optimization*. Princeton University Press, 2009.

[4] Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-dynamic programming*. 1996.

[5] Bertsimas, D., Kallus, N., and Gupta, V. *Data-driven robust optimization*. Springer Berlin Heidelberg, 2017.

[6] Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[7] Castro, P. S. and Precup, D. Smarter Sampling in Model-Based Bayesian Reinforcement Learning. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. LNAI*, 6321:200–2014, 2010.

[8] Delage, E. and Mannor, S. Percentile Optimization for Markov Decision Processes with Parameter Uncertainty. *Operations Research*, 58(1):203–213, 2010.

[9] Delgado, K. V., De Barros, L. N., Dias, D. B., and Sanner, S. Real-time dynamic programming for Markov decision processes with imprecise probabilities. *Artificial Intelligence*, 230:192–223, 2016.

[10] Dietterich, T., Taleghan, M., and Crowley, M. PAC optimal planning for invasive species management: Improved exploration for reinforcement learning from simulator-defined MDPs. *AAAI*, 2013.

[11] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2014.

[12] Goyal, V. and Grand-Clement, J. Robust Markov Decision Process: Beyond Rectangularity. Technical report, 2018.

[13] Gupta, V. Near-Optimal Bayesian Ambiguity Sets for Distributionally Robust Optimization. 2015.

[14] Hanasusanto, G. and Kuhn, D. Robust Data-Driven Dynamic Programming. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[15] Ho, C. P., Petrik, M., and Wiesemann, W. Fast Bellman Updates for Robust MDPs. In *International Conference on Machine Learning (ICML)*, volume 80, pp. 1979–1988, 2018.

[16] Iyengar, G. N. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

[17] Jaksch, T., Ortner, R., and Auer, P. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(1):1563–1600, 2010.

[18] Jiang, N. and Li, L. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 2015.

[19] Kalyanasundaram, S., Chong, E. K. P., and Shroff, N. B. Markov decision processes with uncertain transition rates: Sensitivity and robust control. In *IEEE Conference on Decision and Control*, pp. 3799–3804, 2002.

[20] Lange, S., Gabel, T., and Riedmiller, M. Batch Reinforcement Learning. In *Reinforcement Learning*, pp. 45–73. 2012.

[21] Laroche, R. and Trichelair, P. Safe Policy Improvement with Baseline Bootstrapping, 2018.

[22] Le Tallec, Y. *Robust, Risk-Sensitive, and Data-driven Control of Markov Decision Processes*. PhD thesis, MIT, 2007.

[23] Li, L., Munos, R., and Szepesvári, C. Toward Minimax Off-policy Value Estimation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[24] Mannor, S., Mebel, O., and Xu, H. Lightning does not strike twice: Robust MDPs with coupled uncertainty. In *International Conference on Machine Learning (ICML)*, 2012.

[25] Mannor, S., Mebel, O., and Xu, H. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

[26] Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. G. Safe and Efficient Off-Policy Reinforcement Learning. In *Conference on Neural Information Processing Systems (NIPS)*, 2016.

[27] Murphy, K. *Machine Learning: A Probabilistic Perspective*. 2012.

[28] Nilim, A. and El Ghaoui, L. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[29] Petrik, M. Approximate dynamic programming by minimizing distributionally robust bounds. In *International Conference of Machine Learning (ICML)*, 2012.

[30] Petrik, M. and Subramanian, D. RAAM : The benefits of robustness in approximating aggregated MDPs in reinforcement learning. In *Neural Information Processing Systems (NIPS)*, 2014.

[31] Petrik, M., Chow, Y., and Ghavamzadeh, M. Safe Policy Improvement by Minimizing Robust Baseline Regret. In *ICML Workshop on Reliable Machine Learning in the Wild*, pp. 1–25, 2016.

[32] Petrik, M., Mohammad Ghavamzadeh, and Chow, Y. Safe Policy Improvement by Minimizing Robust Baseline Regret. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[33] Puterman, M. L. *Markov decision processes: Discrete stochastic dynamic programming*. 2005.

[34] Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on Stochastic Programming*. SIAM, 2009.

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

[36] Strehl, A. and Littman, M. An analysis of model-based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74:1309–1331, 2008.

[37] Strehl, A. L. *Probably Approximately Correct (PAC) Exploration in Reinforcement Learning*. PhD thesis, Rutgers University, 2007.

[38] Sutton, R. S. and Barto, A. *Reinforcement learning*. 1998.

[39] Taleghan, M. A., Dietterich, T. G., Crowley, M., Hall, K., and Albers, H. J. PAC Optimal MDP Planning with Application to Invasive Species Management. *Journal of Machine Learning Research*, 16:3877–3903, 2015.

[40] Tamar, A., Mannor, S., and Xu, H. Scaling up Robust MDPs Using Function Approximation. In *International Conference of Machine Learning (ICML)*, 2014.

[41] Thomas, P. S. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference of Machine Learning (ICML)*, 2016.

[42] Thomas, P. S., Teocharous, G., and Ghavamzadeh, M. High Confidence Off-Policy Evaluation. In *Annual Conference of the AAAI*, 2015.

[43] Tirinzoni, A., Milano, P., Chen, X., and Ziebart, B. D. Policy-Conditioned Uncertainty Sets for Robust Markov Decision Processes. In *Neural Information Processing Systems (NIPS)*, 2018.

[44] Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the L1 deviation of the empirical distribution. 2003.

[45] Wiesemann, W., Kuhn, D., and Rustem, B. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[46] Xu, H. and Mannor, S. The robustness-performance tradeoff in Markov decision processes. *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[47] Xu, H. and Mannor, S. Parametric regret in uncertain Markov decision processes. In *IEEE Conference on Decision and Control (CDC)*, pp. 3606–3613, 2009.

A Technical Results

The following proposition shows that the guarantee of a safe estimate on the return is achieved when the true transition model is contained in the ambiguity set.

Lemma A.1. *Suppose that an ambiguity set $\mathcal{P}$ satisfies $\mathbb{P}_{\mathcal{D}}\left[p^\star_{s,a} \in \mathcal{P}_{s,a} \mid P^\star\right] \geq 1 - \delta/(SA)$ for each state s and action a. Then:*

$$\mathbb{P}_{\mathcal{D}}\left[\hat{v}^\pi_{\mathcal{P}} \leq v^\pi_{P^\star},\ \forall \pi \in \Pi \mid P^\star\right] \geq 1 - \delta\ .$$

Proof. We omit $\mathcal{P}$ and $P^\star$ from the notation in the proof since they are fixed. From Lemma B.1, we have that $\hat{v}^\pi \leq v^\pi$ if

$$\widehat{T}^\pi \hat{v}^\pi \leq T^\pi \hat{v}^\pi\ .$$

That is, for each state s and action a:

$$\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} \hat{v}^\pi \leq (p^\star_{s,a})^\mathsf{T} \hat{v}^\pi.$$

Using the identity above, the probability that the robust value function is a lower bound can be bounded as follows:

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}}\left[\hat{v}^\pi_{\mathcal{P}} \leq v^\pi_P,\ \forall \pi \in \Pi \mid P^\star\right] &= \mathbb{P}_{\mathcal{D}}\left[\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} \hat{v}^\pi \leq (p^\star_{s,a})^\mathsf{T} \hat{v}^\pi,\ \forall \pi \in \Pi, s \in \mathcal{S}, a \in \mathcal{A} \mid P^\star\right] \geq \\
&\geq \mathbb{P}_{\mathcal{D}}\left[(p^\star_{s,a})^\mathsf{T} \hat{v}^\pi \leq (p^\star_{s,a})^\mathsf{T} \hat{v}^\pi,\ \forall \pi \in \Pi, s \in \mathcal{S}, a \in \mathcal{A} \mid P^\star \in \mathcal{P}, P^\star\right] \mathbb{P}_{\mathcal{D}}\left[P^\star \in \mathcal{P} \mid P^\star\right] + \\
&\quad + \mathbb{P}_{\mathcal{D}}\left[P^\star \notin \mathcal{P} \mid P^\star\right] \geq 1\, \mathbb{P}_{\mathcal{D}}\left[P^\star \in \mathcal{P} \mid P^\star\right] + 0\, \mathbb{P}_{\mathcal{D}}\left[P^\star \notin \mathcal{P} \mid P^\star\right] \geq \\
&\geq \mathbb{P}_{\mathcal{D}}\left[P^\star \in \mathcal{P} \mid P^\star\right]\ .
\end{aligned}$$

Now, from the union bound over all states and actions, we get:

$$\mathbb{P}_{\mathcal{D}}\left[\hat{v}^\pi > v^\pi | P^\star\right] \leq \mathbb{P}_{\mathcal{D}}\left[P^\star \notin \mathcal{P} \mid P^\star\right] \leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathbb{P}_{\mathcal{D}}\left[p^\star_{s,a} \notin \mathcal{P}_{s,a} \mid P^\star\right] \leq \delta\ ,$$

which completes the proof. □

The next proposition is the Bayesian equivalent of Lemma A.1.

Lemma A.2. *Suppose that an ambiguity set $\mathcal{P}$ satisfies $\mathbb{P}_{P^\star}\left[p^\star_{s,a} \in \mathcal{P}_{s,a} \mid \mathcal{D}\right] \geq 1 - \delta/(SA)$ for each state s and action a. Then:*

$$\mathbb{P}_{P^\star}\left[\hat{v}^\pi_{\mathcal{P}} \leq v^\pi_{P^\star},\ \forall \pi \in \Pi \mid \mathcal{D}\right] \geq 1 - \delta\ .$$

Proof. We omit $\mathcal{P}$ and $P^\star$ from the notation in the proof since they are fixed. From Lemma B.1, we have that $\hat{v}^\pi \leq v^\pi$ if

$$\widehat{T}^\pi \hat{v}^\pi \leq T^\pi \hat{v}^\pi\ .$$

That is, for each state s and action a:

$$\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} \hat{v}^\pi \leq (p^\star_{s,a})^\mathsf{T} \hat{v}^\pi.$$

Using the identity above, the probability that the robust value function is a lower bound can be bounded as follows:

$$\begin{aligned}
\mathbb{P}_{P^\star}\left[\hat{v}^\pi_{\mathcal{P}} \leq v^\pi_P,\ \forall \pi \in \Pi \mid \mathcal{D}\right] &= \mathbb{P}_{P^\star}\left[\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} \hat{v}^\pi \leq (p^\star_{s,a})^\mathsf{T} \hat{v}^\pi,\ \forall \pi \in \Pi, s \in \mathcal{S}, a \in \mathcal{A} \mid \mathcal{D}\right] \geq \\
&\geq \mathbb{P}_{P^\star}\left[(p^\star_{s,a})^\mathsf{T} \hat{v}^\pi \leq (p^\star_{s,a})^\mathsf{T} \hat{v}^\pi,\ \forall \pi \in \Pi, s \in \mathcal{S}, a \in \mathcal{A} \mid P^\star \in \mathcal{P}, \mathcal{D}\right] \mathbb{P}_{P^\star}\left[P^\star \in \mathcal{P} \mid \mathcal{D}\right] + \\
&\quad + \mathbb{P}_{P^\star}\left[P^\star \notin \mathcal{P} \mid \mathcal{D}\right] \geq 1\, \mathbb{P}_{P^\star}\left[P^\star \in \mathcal{P} \mid \mathcal{D}\right] + 0\, \mathbb{P}_{P^\star}\left[P^\star \notin \mathcal{P} \mid \mathcal{D}\right] \geq \\
&\geq \mathbb{P}_{P^\star}\left[P^\star \in \mathcal{P} \mid \mathcal{D}\right]\ .
\end{aligned}$$

Now, from the union bound over all states and actions, we get:

$$\mathbb{P}_{P^\star}\left[\hat{v}^\pi > v^\pi | \mathcal{D}\right] \leq \mathbb{P}_{P^\star}\left[P^\star \notin \mathcal{P} \mid \mathcal{D}\right] \leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathbb{P}_{P^\star}\left[p^\star_{s,a} \notin \mathcal{P}_{s,a} \mid \mathcal{D}\right] \leq \delta\ ,$$

which completes the proof. □

B Technical Proofs

B.1 Proof of Theorem 3.1

Before presenting the proof of the theorem, we need to show some auxiliary results.

The following lemma shows that when the robust Bellman update lower-bounds the true Bellman update then the value function estimate is safe.

Lemma B.1. *Consider a policy π, its robust value function $\hat{v}^\pi$, and true value function v^π such that $\hat{v}^\pi = \widehat{T}^\pi \hat{v}^\pi$ and $v^\pi = T^\pi v^\pi$. Then, $\hat{v}^\pi \leq v^\pi$ element-wise whenever $\widehat{T}^\pi \hat{v}^\pi \leq T^\pi \hat{v}^\pi$. That is, if $\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} \hat{v}^\pi \leq p_{s,a}^\mathsf{T} \hat{v}^\pi$ for each state s and action $a = \pi(s)$ then $\hat{v}^\pi \leq v^\pi$.*

Lemma B.1 implies readily that the inequality above is satisfied when $p^\star_{s,a} \in \mathcal{P}_{s,a}$.

Proof. Using the assumption $\widehat{T}^\pi \hat{v}^\pi \leq T^\pi \hat{v}^\pi$, and from $\hat{v}^\pi = \widehat{T}^\pi \hat{v}^\pi$ and $v^\pi = T^\pi v^\pi$, we get by algebraic manipulation:

$$\hat{v}^\pi - v^\pi = \widehat{T}^\pi \hat{v}^\pi - T_P^\pi v^\pi \leq T^\pi \hat{v}^\pi - T^\pi v^\pi = \gamma P_\pi (\hat{v}^\pi - v^\pi) \,.$$

Here, P_π is the transition probability matrix for the policy π. Subtracting $\gamma P_\pi(\hat{v}^\pi - v^\pi)$ from the above inequality gives:

$$(\mathbf{I} - \gamma P_\pi)(\hat{v}^\pi - v^\pi) \leq \mathbf{0} \,,$$

where $\mathbf{I}$ is the identity matrix. Because the matrix $(\mathbf{I} - \gamma P_{\pi^\star})^{-1}$ is monotone, as can be seen from its Neumann series, we get:

$$\hat{v}^\pi - v^\pi \leq (\mathbf{I} - \gamma P_\pi)^{-1} \mathbf{0} = \mathbf{0} \,,$$

which proves the result. □

The next lemma formalizes the safety-sufficiency of $\mathcal{K}$. Note that the rewards $r_{s,a}$ are not a factor in this lemma because they are certain and cancel out.

Lemma B.2. *Consider any ambiguity set $\mathcal{P}_{s,a}$ and a value function v. Then $\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} v \leq (p^\star_{s,a})^\mathsf{T} v$ with probability $1 - \delta/(SA)$ if and only if $\mathcal{P}_{s,a} \cap \mathcal{K}_{s,a}(v) \neq \emptyset$.*

Proof. To show the "if" direction, let $\hat{p} \in \mathcal{P}_{s,a} \cap \mathcal{K}_{s,a}(v)$. Such $\hat{p}$ exists because the intersection is nonempty. Then, $\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} v \leq \hat{p}^\mathsf{T} v \leq \mathrm{V@R}^\zeta_{P^\star}\left[(p^\star_{s,a})^\mathsf{T} v\right]$. By definition, $\mathrm{V@R}^\zeta_{P^\star}\left[(p^\star_{s,a})^\mathsf{T} v\right] \leq (p^\star_{s,a})^\mathsf{T} v$ with probability $1 - \delta/(SA)$.

To show the "only if" direction, suppose that $\hat{p}$ is a minimizer in $\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} v$. The premise translates to $\mathbb{P}_{P^\star}[\hat{p}^\mathsf{T} v \leq (p^\star_{s,a})^\mathsf{T} v \mid \mathcal{D}] \geq 1 - \delta/(SA)$. Therefore, $\mathrm{V@R}^\zeta_{P^\star}\left[(p^\star_{s,a})^\mathsf{T} v\right] \geq \hat{p}^\mathsf{T} v$ and $\hat{p} \in \mathcal{P}_{s,a} \cap \mathcal{K}_{s,a}$ and the intersection is non-empty. □

The following lemma formalizes the properties of $\mathcal{L}_{s,a}$.

Lemma B.3. *For any finite set $\mathcal{V}$ of value functions, the following inequality holds for all $v \in \mathcal{V}$ simultaneously:*

$$\mathbb{P}_{P^\star}\left[\min_{p \in \mathcal{L}_{s,a}(\mathcal{V})} p^\mathsf{T} v \leq (p^\star_{s,a})^\mathsf{T} v \;\middle|\; \mathcal{D}\right] \geq 1 - \frac{\delta}{SA} \,.$$

Proof. Assume an arbitrary $v \in \mathcal{V}$ and let $q^\star_v \in \arg\min_{q \in \mathcal{K}_{s,a}(v)} \|q - \theta_{s,a}(\mathcal{V})\|_1$ using the notation of (4). From the definition of $\theta_{s,a}(\mathcal{V})$ in (4), the value q_v is in the ambiguity set $\mathcal{L}_{s,a}(\mathcal{V})$. Given that also $q_v \in \mathcal{K}_{s,a}(v)$, Lemma B.2 shows that:

$$\mathbb{P}_{P^\star}\left[\min_{p \in \mathcal{L}_{s,a}(\mathcal{V})} p^\mathsf{T} v \leq (p^\star_{s,a})^\mathsf{T} v \;\middle|\; \mathcal{D}\right] \geq 1 - \frac{\delta}{SA} \,,$$

because $q_v \in \mathcal{L}_{s,a}(v) \cup \mathcal{K}_{s,a}(v) \neq \emptyset$. This completes the proof since v is any from $\mathcal{V}$. □

We are now ready to prove the theorem.

Proof. Recall that Algorithm 1 terminates only if $\mathcal{K}_{s,a}(\hat{v}_k) \cap \mathcal{P}^k_{s,a} \neq \emptyset$ for each state s and action a. Then, according to Lemma B.2, we get with probability $1 - \delta/(SA)$:

$$\min_{p \in \mathcal{P}^k_{s,a}} p^\mathsf{T} \hat{v}_k \leq (p^\star_{s,a})^\mathsf{T} \hat{v}_k$$

for any fixed state s and action a. By the union bound, the inequality holds simultaneously for all states and actions with probability $1 - \delta$. That means that with probability $1 - \delta$ we can derive the following using basic algebra:

$$\begin{aligned} \min_{p \in \mathcal{P}^k_{s,a}} p^\mathsf{T} \hat{v}_k &\leq (p^\star_{s,a})^\mathsf{T} \hat{v}_k && \forall s \in \mathcal{S}, a \in \mathcal{A} \\ r_{s,a} + \min_{p \in \mathcal{P}^k_{s,a}} p^\mathsf{T} \hat{v}_k &\leq r_{s,a} + (p^\star_{s,a})^\mathsf{T} \hat{v}_k && \forall s \in \mathcal{S}, a \in \mathcal{A} \\ \widehat{T}^{\hat{\pi}_k}_{\mathcal{P}^k} \hat{v}_k &\leq T^{\hat{\pi}_k}_{P^\star} \hat{v}_k \end{aligned}$$

Note that $\hat{v}_k$ is the robust value function for the policy $\hat{\pi}_k$ since $\hat{v}_k = \hat{v}^\star_{\mathcal{P}_k}$ and $\hat{\pi}_k = \hat{\pi}^\star_{\mathcal{P}_k}$. Lemma B.1 finally implies that $\hat{v}_k \leq v^{\hat{\pi}_k}_{P^\star}$ with probability $1 - \delta$. □

B.2 Proof of Theorem 4.1

Proof. The first part of the statement follows directly from Lemma A.1 and Lemma C.1. The second part of the statement follows from the fact that the lower bound property holds uniformly across all policies. □

B.3 Proof of Theorem 4.2

Proof. The first part of the statement follows directly from Lemma A.2 and the definition of $\psi^B_{s,a}$. The second part of the statement follows from the fact that the lower bound property holds uniformly across all policies. □

C L_1 Concentration Inequality Bounds

In this section, we describe a new elementary proof of a bound on the L_1 distance between the estimated transition probability distribution and the true one. It simplifies the proofs of Weissman et al. (2003) but also leads to coarser bounds. We include the proof here in order to derive the tighter bound in Appendix C.1. Note that in the frequentist setting the ambiguity set $\mathcal{P}$ is a random variable that is a function of the dataset $\mathcal{D}$.

Recall that our ambiguity sets are defined as L_1 balls around the expected transition probabilities $\bar{p}_{s,a}$:

$$\mathcal{P}_{s,a} = \{p \in \Delta^S \,:\, \|p - \bar{p}_{s,a}\|_1 \leq \psi_{s,a}\} \,. \tag{6}$$

Lemma A.1 implies that the size of the L_1 balls must be chosen as follows:

$$\mathbb{P}\left[\|\bar{p}(s,a) - p^\star(s,a)\|_1 \leq \psi_{s,a}\right] \geq 1 - \delta/(SA) \,. \tag{7}$$

We can now express the necessary size $\psi_{s,a}$ of the ambiguity sets in terms of $n_{s,a}$, which denotes the number of samples in $\mathcal{D}$ that originate with a state s and an action a.

Lemma C.1 (L_1 Error bound)**.** *Suppose that $\bar{p}_{s,a}$ is the empirical estimate of the transition probability obtained from $n_{s,a}$ samples for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Then:*

$$\mathbb{P}\left[\|\bar{p}_{s,a} - p^\star_{s,a}\|_1 \geq \psi_{s,a}\right] \leq (2^S - 2) \exp\left(-\frac{\psi^2_{s,a} n_{s,a}}{2}\right) \,.$$

Therefore, for any $\delta \in [0,1]$:

$$\mathbb{P}\left[\|\bar{p}_{s,a} - p^\star_{s,a}\|_1 \leq \sqrt{\frac{2}{n_{s,a}} \log \frac{SA(2^S - 2)}{\delta}}\right] \leq 1 - \delta/(SA) \,.$$

Proof. To shorten the notation, we omit the indexes s, a throughout the proof; for example $\bar{p}$ is used instead of the full $\bar{p}_{s,a}$. First, express the L_1 distance between two distributions $\bar{p}$ and $p^\star$ in terms of an optimization problem. Let $\mathbf{1}_\mathcal{Q} \in \mathbb{R}^\mathcal{S}$ be the indicator vector for some subset $\mathcal{Q} \subset \mathcal{S}$. Then:

$$\begin{aligned}\|\bar{p} - p^\star\|_1 &= \max_z \left\{ z^\mathsf{T}(\bar{p} - p^\star) \,:\, \|z\|_\infty \leq 1 \right\} = \\ &= \max_{\mathcal{Q} \in 2^\mathcal{S}} \left\{ \mathbf{1}_\mathcal{Q}^\mathsf{T}(\bar{p} - p^\star) - (\mathbf{1} - \mathbf{1}_\mathcal{Q})^\mathsf{T}(\bar{p} - p^\star) \,:\, 0 < |\mathcal{Q}| < m \right\} \\ &\overset{\text{(a)}}{=} 2 \max_{\mathcal{Q} \in 2^\mathcal{S}} \left\{ \mathbf{1}_\mathcal{Q}^\mathsf{T}(\bar{p} - p^\star) \,:\, 0 < |\mathcal{Q}| < m \right\} .\end{aligned}$$

Here, (a) holds because $\mathbf{1}^\mathsf{T}(\bar{p} - p^\star) = 0$. Using the expression above, the target probability can be bounded as follows:

$$\begin{aligned}\mathbb{P}\left[\|\bar{p} - p^\star\|_1 > \psi\right] &= \mathbb{P}\left[2 \max_{\mathcal{Q} \in 2^\mathcal{S}} \left\{ \mathbf{1}_\mathcal{Q}^\mathsf{T}(\bar{p} - p^\star) \,:\, 0 < |\mathcal{Q}| < m \right\} > \psi \right] \\ &\overset{\text{(a)}}{\leq} (|\mathcal{Q}| - 2) \max_{\mathcal{Q} \in 2^\mathcal{S}} \left\{ \mathbb{P}\left[\mathbf{1}_\mathcal{Q}^\mathsf{T}(\bar{p} - p^\star) > \frac{\psi}{2} \right] \,:\, 0 < |\mathcal{Q}| < m \right\} \\ &\overset{\text{(b)}}{\leq} (|\mathcal{Q}| - 2) \exp\left(-\frac{\psi^2 n}{2} \right) = (2^S - 2) \exp\left(-\frac{\psi^2 n}{2} \right) .\end{aligned}$$

The inequality (a) follows from union bound and the inequality (b) follows from the Hoeffding's inequality since $\mathbf{1}_\mathcal{Q}^\mathsf{T}\bar{p} \in [0, 1]$ for any $\mathcal{Q}$ with the mean of $\mathbf{1}_\mathcal{Q}^\mathsf{T}\bar{p}^\star$. □

C.1 Ambiguity Sets for Monotone Value Functions

A significant limitation of the result in Lemma C.1 is that the ψ depends linearly on the number of states. We now explore an assumption that can alleviate this important drawback when the value functions are guaranteed to be monotone. In particular, the monotonicity assumption states that the value functions v of the optimal robust policy must be non-decreasing in some arbitrary order which must be known ahead of time. Assume, therefore, without loss of generality that:

$$v_1 \geq v_2 \geq \ldots \geq v_n \,, \tag{8}$$

where v_i is the value of state i.

Admittedly, monotonicity is a restrictive assumption, but we explore it in order to understand the greatest possible gains from tightening the known concentration inequalities. Yet, monotonicity of this type occurs in some problems, such as inventory management in which the value does not decrease with increasing inventory levels or medical problems in which the value does not increase with a deteriorating health state.

It is important to note that any MDP algorithm that relies on the assumption (8) needs to also enforce it. That means, the algorithm must prevent generating value functions that violate the monotonicity assumption. Practically, this could be achieved by representing the value function as a linear combination of monotone features.

The bound Lemma C.1 is large because of the term 2^S which derives from the use of a union bound. The union bound is used because the L_1 norm can be represented as a maximum over an exponentially many linear functions:

$$\|x\|_1 = \max_{\mathcal{Q} \subseteq \mathcal{J}} \left(\mathbf{1}_\mathcal{Q} - \mathbf{1}_{\mathcal{J} \setminus \mathcal{Q}} \right)^\mathsf{T} x \,.$$

Here, the set $\mathcal{J} = 2^S$ represents all indexes of x and $\mathbf{1}_\mathcal{Q}$ is a vector that is one for all elements of $\mathcal{Q}$ and zero otherwise. We now show that under the monotonicity property (8), the L_1 norm can be represented as a maximum over a *linear* (in states) number of linear functions. In particular, the worst-case optimization problem of the nature:

$$\begin{array}{ll} \min\limits_{p} & v^\mathsf{T} p \\ \text{s.t.} & \left(\mathbf{1}_\mathcal{Q} - \mathbf{1}_{\mathcal{J} \setminus \mathcal{Q}} \right)^\mathsf{T} (p - \bar{p}) \leq \psi, \quad \forall \mathcal{Q} \subseteq \mathcal{J} \\ & \mathbf{1}^\mathsf{T} p = 1, \\ & p \geq 0 \end{array} \tag{9}$$

can be replaced by the following optimization problem:

$$\begin{aligned} \min_{p} \quad & v^{\mathsf{T}} p \\ \text{s.t.} \quad & (\mathbf{1}_{k\ldots n} - \mathbf{1}_{1\ldots(k-1)})^{\mathsf{T}}(p - \bar{p}) \leq \psi, \quad \forall k = 0, \ldots, (n+1) \\ & \mathbf{1}^{\mathsf{T}} p = 1, \\ & p \geq 0 \end{aligned} \tag{10}$$

Lemma C.2. *Suppose that* (8) *is satisfied. Then the optimal objective values of* (9) *and* (10) *coincide.*

Proof. Let f^a be the optimal objective of (9) and let f^b be the optimal objective of (10). The inequality $f^a \geq f^b$ can be shown readily since (10) only relaxes some of the constraints of (9).

It remains to show that $f^a \leq f^b$. To show the inequality by contradiction, assume that each optimal solution p^b to (10) is infeasible in (9) (otherwise $f^a \leq f^b$). Let the constraint violated by p^b be:

$$\left(\mathbf{1}_{\mathcal{C}} - \mathbf{1}_{2^S \setminus \mathcal{C}}\right)^{\mathsf{T}} (p - \bar{p}) \leq \psi,$$

for some set $\mathcal{C}$. Since this constraint is not present in (10), that means that there exist i and j such that $i < j$, $i \in \mathcal{C}$, $j \notin \mathcal{C}$, and because the constraint is violated:

$$p^b_i = \bar{p}_i - \epsilon, \quad \text{or} \quad p^b_j = \bar{p}_j + \epsilon$$

for some $\epsilon > 0$. Assume now that $p^b_i = \bar{p}_i - \epsilon$, the case when $p^b_j = \bar{p}_j + \epsilon$ follows similarly.

Now, choose the largest $k > i$ possible, and let $p^a = p^b$, with the exception of:

$$p^a_i = p^b_i + \epsilon, \quad \text{and} \quad p^a_k = p^b_k - \epsilon\,.$$

This does not increase the violation of the constraint by p^a over p^b:

$$\left(\mathbf{1}_{\mathcal{C}} - \mathbf{1}_{2^S \setminus \mathcal{C}}\right)^{\mathsf{T}} (p^a - \bar{p}) \leq \left(\mathbf{1}_{\mathcal{C}} - \mathbf{1}_{2^S \setminus \mathcal{C}}\right)^{\mathsf{T}} (p^b - \bar{p}),$$

And it does not increase the objective function:

$$v^{\mathsf{T}} p^a = v^{\mathsf{T}} p^b - \epsilon(v_i - v_j) \leq v^{\mathsf{T}} p^b,$$

and thus remains optimal in (10). Repeating these steps until no constraints are violated leads to a contradiction with the lack of an optimal solution to (10) that is not optimal in (9). □

Lemma C.2 shows that we can replace the L_1 ambiguity set in (6) by the following set without affecting the solution.

$$\mathcal{P}_{s,a} = \{p \in \Delta^S \,:\, (\mathbf{1}_{k\ldots n} - \mathbf{1}_{1\ldots(k-1)})^{\mathsf{T}}(p - \bar{p}_{s,a}) \leq \psi_{s,a}, \quad \forall k = 0, \ldots, (n+1)\} \tag{11}$$

Now, following the same steps as the proof of Lemma C.1 but using (11) in place of (6) gives us the following result.

Lemma C.3 (L_1 Error bound)**.** *Suppose that $\bar{p}_{s,a}$ is the empirical estimate of the transition probability obtained from $n_{s,a}$ samples for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Then:*

$$\mathbb{P}\left[\|\bar{p}_{s,a} - p^\star_{s,a}\|_1 \geq \psi_{s,a}\right] \leq S \exp\left(-\frac{\psi^2_{s,a} n_{s,a}}{2}\right)\,.$$

Therefore, for any $\delta \in [0,1]$:

$$\mathbb{P}\left[\|\bar{p}_{s,a} - p^\star_{s,a}\|_1 \leq \sqrt{\frac{2}{n_{s,a}} \log \frac{S^2 A}{\delta}}\right] \leq 1 - \delta/(SA)\,.$$

D Detailed Description of Selected Algorithms

D.1 Computing Bayesian Confidence Interval

Algorithm 2: 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* $\psi_{s,a}$;

E Why Not Credible Regions

Constructing ambiguity sets from confidence regions seems intuitive and natural. It may be surprising that RSVF abandons this intuitive approach. In this section, we describe two reasons why confidence regions are unnecessarily conservative compared to RSVF sets.

The first reason why confidence regions are too conservative is because they assume that the value function depends on the true model $P^\star$. To see this, consider the setting of Example 2.1 with $r_{s_1,a_1} = 0$. When an ambiguity set $\mathcal{P}_{s_1,a_1}$ is built as a confidence region such that $\mathbb{P}[p^\star_{s_1,a_1} \in \mathcal{P}_{s_1,a_1}] \geq 1-\delta$, it satisfies:

$$\mathbb{P}_{P^\star}\left[\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} v \leq (p^\star_{s,a})^\mathsf{T} v,\ \forall v \in \mathbb{R}^S \,\middle|\, \mathcal{D}\right] \geq 1-\delta.$$

Notice the value function inside of the probability operator. Lemma B.1 shows that this guarantee is needlessly strong. It is, instead, sufficient that the inequality in Lemma B.1 holds just for $\hat{v}^\pi$ which is independent of $P^\star$ in the Bayesian setting. The following weaker condition is sufficient to guarantee safety:

$$\mathbb{P}_{P^\star}\left[\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} v \leq (p^\star_{s,a})^\mathsf{T} v \,\middle|\, \mathcal{D}\right] \geq 1-\delta,\ \forall v \in \mathbb{R}^S \tag{12}$$

Notice that v is outside of the probability operator. This set is smaller and provides the same guarantees, but may be more difficult to construct [13].

The second reason why confidence regions are too conservative is because they construct a uniform lower bound for all policies π as is apparent in Theorem 4.2. This is unnecessary, again, as Lemma B.1 shows. The robust Bellman update only needs to lower bound the Bellman update for the computed value function $\hat{v}^\pi$, not for all value functions. As a result, (12), can be further relaxed to:

$$\mathbb{P}_{P^\star}\left[\min_{p \in \mathcal{P}_{s,a}} p^\mathsf{T} \hat{v}^{\pi_R} \leq (p^\star_{s,a})^\mathsf{T} \hat{v}^{\pi_R} \,\middle|\, \mathcal{D}\right] \geq 1-\delta, \tag{13}$$

where π_R is the optimal solution to the robust MDP. RSVF is less conservative because it constructs ambiguity sets that satisfy the weaker requirement of (13) rather than confidence regions. Deeper theoretical analysis of the benefits of using RSVF sets is very important but is beyond the scope of this work. Examples that show the benefits to be arbitrarily large or small can be constructed readily by properly choosing the priors over probability distributions.