Soft-Robust Algorithms for Handling Model Misspecification

Elita A. Lobo
University of Massachusetts, Amherst
elitalobo@cs.umass.edu

Mohammad Ghavamzadeh
Google Research
ghavamza@gmail.com

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

Abstract

In reinforcement learning, robust policies for high-stakes decision-making problems with limited data are usually computed by optimizing the percentile criterion, which minimizes the probability of a catastrophic failure. Unfortunately, such policies are typically overly conservative as the percentile criterion is non-convex, difficult to optimize, and ignores the mean performance. To overcome these shortcomings, we study the soft-robust criterion, which uses risk measures to balance the mean and percentile criteria better. In this paper, we establish the soft-robust criterion’s fundamental properties, show that it is NP-hard to optimize, and propose and analyze two algorithms to optimize it approximately. Our theoretical analyses and empirical evaluations demonstrate that our algorithms compute much less conservative solutions than the existing approximate methods for optimizing the percentile-criterion.

1 Introduction

Markov Decision Process (MDP) is an established model for optimizing returns in sequential decision-making problems [33, 43]. In Reinforcement Learning (RL), MDPs must be estimated from sampled data. Gathering sufficient data through the exploration to estimate transition models with high precision is impractical in many high-stakes domains, like healthcare or agriculture [6, 10, 38]. Such data limitation inevitably causes errors in the estimates of transition models and rewards, and results in policies that can fail catastrophically when deployed. We aim to compute robust policies from limited data in a way that accounts for the uncertainty in transition models.

In this work, we consider model-based approaches that rely on parametric Bayesian models to represent the uncertainty in transition models. Instead of computing with the most-likely estimate of the transition models, these methods use Bayesian inference to derive a posterior distribution over possible transition models [6, 10, 38, 49–51]. The most common robust objective, given a posterior distribution over uncertain transition probability parameters, is the percentile criterion. The percentile criterion optimizes a user-specified α-quantile of the expected return [6, 38, 44]. Despite their strong robustness guarantees, percentile criterion-based algorithms suffer from three major shortcomings. These limitations can cause the learned policies to be pessimistic and restrict their use in real-world applications.

First, it is well-known that the percentile criterion can have multiple randomized optimal policies [21]. Most existing approaches that only optimize the percentile criterion fail to learn optimal robust policies that also maximize the mean performance. As a result, such policies can be unnecessarily pessimistic. To see this, consider a portfolio optimization problem where an agent has to choose between 4 stocks A, B, C, and D to invest in. Assume that either uniformly randomizing investment between stocks A and B (Option 1) or between stocks B, C, and D (Option 2), provide the best robustness guarantees, i.e., the true expected return would not drop below Y for 95% of possible stock market scenarios. Further, assume that investing in stock D results in a very large profit in a subset of the scenarios and as a result, the average expected return computed over all possible market scenarios for Option 2 is larger than that for Option 1. The percentile-criterion is blind to mean performance and it may randomly select Option 1, thereby potentially resulting in large losses.

Second, the percentile criterion is not convex which complicates its optimization. A widely used approach to optimize this criterion is to construct a compact and convex uncertainty set that accounts for α% of the model parameter values, and solve a min-max game wherein the agent maximizes the expected return under the worst parameter realization in the set [30, 38, 47]. This setup typically represents a Robust Markov De-
cision Process (RMDP) [22, 47]. In this setting, the uncertainty sets are termed as ambiguity sets and are usually constructed using statistical confidence intervals. Recent work shows that the ambiguity sets (that are confidence regions) are often very large and lead to overly-conservative policies [14, 38]. Although several works have proposed methods to optimize the size of the uncertainty sets [38, 47], they do not scale to large problems and can be computationally expensive [8].

Third, the percentile criterion does not take into account the magnitude of the risk associated with models that fall below the α quantile. In problems with heavy tail risks, for example, the portfolio optimization problem, the percentile criterion learns policies that are too optimistic and may result in disastrous worst-case outcomes [36].

To overcome the above limitations, we adopt the soft-robust criterion [1, 7]. Contrary to the percentile criterion, the soft-robust criterion optimizes a convex combination of mean and robust performances and is itself convex. We measure the robust performance in soft-robust criterion using the Conditional Value at Risk (CVaR) measure [36], which represents the mean of the expected return of the worst α% of the models. The CVaR measure also bounds from below the percentile criterion and takes into account the tail risks. We note that although the soft-robust criterion has been widely studied in finance and risk-averse RL, (see e.g., [5, 32, 45]), it has not been well-understood in the context of RMDPs and static model uncertainties. We discuss related work in greater depth in Section 6.

Unfortunately, just like the percentile criterion, optimization of the soft-robust criterion is an NP-hard problem. In this paper, we introduce a Mixed Integer Linear Program (MILP), the solution to which is the exact deterministic optimal policy for the soft-robust criterion. This formulation differs from most prior work [7, 26, 51] in a subtle but important way by assuming that the uncertain parameters are realized once for every episode instead of at every time-step. The assumption that the uncertain parameters can vary at every time step (dynamic model of uncertainty) is pessimistic because it does not allow the agent to exploit any information about the uncertain parameters that can be inferred from the state features to make more informed decisions.

Compared with prior work that assumes a dynamic model of the uncertainty, our MILP formulation is more natural and generates less conservative policies. However, MILPs do not scale well and only work well for small to medium size problems. To overcome the issue of scalability, we present an alternative method that builds on prior work to draw connections to Robust MDPs and tractably solve an approximation of the soft-robust criterion in large problems.

Contributions. As our main contribution, we define and analyze the soft-robust criterion in the context of RL. By assuming that the uncertain parameters are realized only once during the execution, instead of at every time step, we derive a more natural formulation that learns less-conservative policies. We show that our formulation allows us to compute exact deterministic optimal policies using a MILP. Inspired by prior work, we also propose a scalable algorithm that approximately optimizes the soft-robust objective using rectangular RMDPs with a specific type of ambiguity set. We also derive new explicit error bounds and convergence guarantees for the soft-robust objective. Finally, the experimental results demonstrate the promise of our methods.

2 Preliminaries

We assume that the agent’s interaction with the environment is modeled by an MDP [33, 43]. An MDP is a tuple \((S, A, P, r, p_0, \gamma)\) that consists of a set of states \(S = \{1, 2, \ldots, S\}\), set of actions \(A = \{1, 2, \ldots, A\}\), a reward function \(r : S \times A \times S \rightarrow \mathbb{R}\), a transition probability function \(P : S \times A \rightarrow \Delta^S\), an initial state distribution \(p_0 \in \Delta^S\), and a discount factor \(\gamma \in [0, 1]\). The symbol \(\Delta^S\) denotes the \(S\)-dimensional probability simplex. Our objective is to maximize the infinite-horizon discounted return and assume that \(|r(s, a)| \leq r_{\text{max}} \in \mathbb{R}\) for each \(s \in S\) and \(a \in A\).

We consider a batch RL setting [25], where the reward function \(r(s, a, s')\) is known but the true transition models \(P^*(s, a, s')\) are unknown and must be estimated from the batch of data \(D = \{(s_i, a_i, s'_i)\}_{i=1}^M\), where \(s'_i\) is distributed according to \(P_{s_i, a_i}\). We take a parametric Bayesian approach to model the uncertainty over the true transition models \(P^*\) [6, 8, 38, 51]. In the Bayesian approach, the transition model \(P^*\) is a random variable. Using the batch of data \(D\), one can derive a posterior distribution over \(P^*\) conditional on the data \(D\), which we denote by \(\tilde{P} = P^* \mid D\). Let \(f\) be the probability measure function of the random variable \(\tilde{P} : \Omega \rightarrow (\Delta^S)^{S \times A}\) and let \(\tilde{P}^\omega\), \(\omega \in \Omega\) denote a sample from the posterior distribution of \(\tilde{P}\) with weight \(f_\omega\). We use \(N = |\Omega|\) to denote the size of the probability space and assume that \(\tilde{P} \sim f\) unless specified otherwise.

A policy \(\pi : S \rightarrow \Delta^A\) prescribes the probability of taking an action \(a \in A\) when the agent is in a state \(s \in S\). We denote the set of all randomized policies by \(\Pi = (\Delta^A)^S\) and the set of all deterministic policies by \(\Pi_D = A^S\). For a given realization of transition models \(P\), the initial state distribution \(p_0\), and a policy \(\pi\), the
expected discounted return is defined as
\[ \rho(\pi, P) = E\left[ \sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t, s_{t+1}) \right], \]
where \( s_0 \sim p_0, a_t \sim \pi(s_t), \) and \( s_{t+1} \sim P(s_t, a_t). \)

**Percentile Criterion.** Percentile optimization has been commonly used to derive robust policies and risk-adjusted discounted returns for an MDP under uncertainty \([6, 36, 38]\). The chance-constrained objective that this criterion optimizes is of the form
\[
\max_{\pi \in \Pi, y \in \mathbb{R}} \left\{ y \mid \mathbb{P}_{\pi, \gamma}[\rho(\pi, \hat{P}) \geq y] \geq \alpha \right\}, \tag{1}
\]
where \( y \) lower-bounds the true expected discounted return with confidence \( \alpha \in [0, 1] \). Optimizing the percentile criterion is equivalent to optimizing the Value at Risk (VaR) of the expected discounted return with respect to the uncertainty in transition models \( \hat{P} \). Thus, we may write the optimization problem in (1) as \( \max_{\pi \in \Pi} \text{VaR}^{\alpha}_g[\rho(\pi, \hat{P})] \), where VaR of a bounded random variable \( Z \) with PDF and CDF functions \( g \) and \( G \) is defined as \( \text{VaR}^{\alpha}_g[Z] = \inf\{z \in \mathbb{R} \mid G(z) \geq 1 - \alpha\} \). The subscript of VaR indicates the random variable when it is not clear from the context. Increasing the value of \( \alpha \) in (1) increases the confidence that the expected return is at least \( y \).

A common alternative to VaR that overcomes its shortcomings is the Conditional Value at Risk measure (CVaR). The CVaR of a random variable \( Z \sim g \) is defined as \([36]\)
\[
\text{CVaR}_g^\alpha[Z] = \max_{\epsilon \in \mathbb{R}}\left( z - \frac{1}{1-\alpha} E[(z - [z]_\epsilon)] \right), \tag{2}
\]
where \( [z - Z]_\epsilon = \max\{z - Z, 0\} \). The CVaR measure represents the average of the expected returns of the worst \( 1 - \alpha \) fraction of the models. Since CVaR lower bounds VaR, we will adopt this measure to derive a tight bound on the \( \alpha \)-quantile of the expected returns.

**Robust MDPs.** A Robust MDPs (RMDP) \([22, 30, 47]\) generalizes an MDP by allowing uncertain transition models. It consists of the same components as an MDP, except the fixed transition model \( P \) is replaced by an ambiguity set \( \mathcal{P} \subseteq \{P : S \times A \rightarrow \Delta^S\} \) of plausible transition models. The objective in an RMDP \([22, 30, 47]\) is to compute a policy that achieves the highest return for the worst-case realization of the transition models in the ambiguity set \( \mathcal{P} \),
\[
\max_{\pi \in \Pi} \min_{P \in \mathcal{P}} \rho(\pi, P). \tag{3}
\]

General RMDPs are NP-hard to solve, but they are tractable for broad classes of ambiguity sets \( \mathcal{P} \). The simplest such class is the SA-rectangular ambiguity set that is defined as \([16, 23, 30, 47]\)
\[
\mathcal{P} = \{p \in (\Delta^S)^{S \times A} \mid p_{s,a} \in \mathcal{P}_{s,a}, \forall s \in S, \forall a \in A\},
\]
for some \( \mathcal{P}_{s,a} \subseteq \Delta^S, s \in S, a \in A \). The SA-rectangular ambiguity sets assume that the transition models corresponding to each state-action pair are independent. S-rectangular ambiguity sets are also tractable and generalize SA-rectangular sets to allow correlation between transition models of actions in the same state and are defined as \([47]\)
\[
\mathcal{P} = \{p \in (\Delta^S)^{S \times A} \mid (p_{s,a})_{a \in A} \in \mathcal{P}_s, \forall s \in S\}, \tag{4}
\]
for some \( \mathcal{P}_s \subseteq (\Delta^S)^A, s \in S \). Each element of \( \mathcal{P}_s \) maps each action \( a \in A \) to a probability distribution in \( \Delta^S \) over the next states.

Ambiguity sets for RMDPs are generally constructed as some form of confidence regions around the mean of transition models \( \bar{p}_{s,a} \), \( s \in S, a \in A \) \([6, 31, 38, 49]\). For instance, an SA-rectangular set may be constructed for each state \( s \in S \) and action \( a \in A \) as
\[
\mathcal{P}_{s,a} = \{p \in \Delta^S \mid \|p - \bar{p}_{s,a}\|_1 \leq \epsilon_{s,a}\},
\]
where the budget \( \epsilon_{s,a} \in \mathbb{R}_+ \) determines the size of the confidence intervals around \( \bar{p}_{s,a} \). Other popular alternatives to the \( L_1 \)-norm (related to the total variation distance) are the Wasserstein metric \([13, 52]\), and Kullback-Liebler divergence \([22]\). In Section 4.2, we introduce a new way of constructing S-rectangular ambiguity sets that are not confidence regions and result in less conservative solutions.

In S-rectangular and SA-rectangular RMDPs, the optimal value function exists, is unique, and is the fixed-point of the S-rectangular robust Bellman operator \( T_{\Psi} : \mathbb{R}^S \rightarrow \mathbb{R}^S \) that is defined for each \( v \in \mathbb{R}^S \) and \( s \in S \) as \([22, 47]\)
\[
(\Psi v)(s) = \max_{d \in \Delta^S} \min_{P_\epsilon \in \mathcal{P}_s} \sum_{a \in A} d_a \left( r_{s,a} + \gamma \cdot P(s_{a,v}^T v) \right), \tag{5}
\]
where the optimal policies may need to be randomized.

**Robust Projected Value Iteration.** The Robust Projected Value Iteration (RPVI) is a type of value iteration introduced for scaling RMDPs to problems with large state spaces. Approximating the value function of a policy in such problems with limited data is particularly challenging. This method represents value function using a linear value function approximation \( v = \Phi w \), where \( \Phi \in \mathbb{R}^{S \times l} \) is the sample feature matrix and \( w \in \mathbb{R}^l \) is the weight vector. To find the optimal weight vector \( w \), RPVI iteratively solves the robust projected Bellman equation given by \( v = \Psi T_{\Psi} v \), where \( \Psi \) is the projection matrix onto the subspace of \( \Phi \) with respect to a Euclidean norm. We refer the readers to Tamar et al. \([44]\) for more details.
3 Soft-Robust Framework

As discussed in Section 1, the percentile criterion has several shortcomings that render it unsuitable for high-risk decision problems. To overcome these limitations, we propose to optimize the soft-robust objective \( \rho^s : \Pi \rightarrow \mathbb{R} \), which we define for a policy \( \pi \in \Pi \) as

\[
\rho^s(\pi) = (1 - \lambda) \mathbb{E} \left[ \rho(\pi, \hat{P}) \right] + \lambda \text{CVaR}^\alpha \left[ \rho(\pi, \hat{P}) \right],
\]

where the two summands in (6) represent the mean performance and robust performance, respectively. The parameters \( \alpha \in [0, 1] \) and \( \lambda \in [0, 1] \) allow a fine-grained control over the robustness of the computed policy.

The parameter \( \lambda \in [0, 1] \) balances the importance of mean and robust performance. When \( \lambda = 0 \), any given policy \( \pi \) is evaluated only based on the mean of its expected discounted return (mean performance). This ignores the impact of any high-risk model that may occur with a small probability. Setting \( \lambda = 1 \), on the other hand, can result in the agent learning a policy that optimizes the robust performance, but may perform poorly with respect to the mean.

The risk-aversion parameter \( \alpha \in [0, 1] \) gives us control over the quality of the robust policy. A higher value of \( \alpha \) makes the CVaR term to place a larger weight on the worst model parameters, which makes the learned policies more robust, but also more conservative.

Next, we describe a more natural interpretation of the soft-robust criterion and show that analogous to the percentile-criterion, the soft-robust criterion is NP-hard to solve and its optimal policies may be randomized.

**Lemma 1.** The soft-robust optimization in (6) equals to the saddle-point problem

\[
\max_{\pi \in \Pi} \min_{\xi \in \Xi} \mathbb{E}_{\rho \sim \xi} \left[ \rho(\pi, \hat{P}) \right],
\]

where \( \Xi \) is defined as

\[
\Xi = \left\{ \xi \in \Delta^{|\Pi|} \mid (1 - \lambda) \cdot f \leq \xi \leq \frac{1 + \lambda \alpha - \alpha}{1 - \alpha} \cdot f \right\}.
\]

Lemma 1 follows directly by substituting the robust representation of CVaR from (2) into (6). We report the proof in Appendix B. This lemma provides a game-theoretic interpretation of the soft-robust algorithm, where the agent plays an adversarial game against nature. At the beginning of each episode, nature samples a transition model from the worst distribution of known transition models in the ambiguity set \( \Xi \), while the agent chooses the best action at every step to maximize the weighted sum of expected returns for the selected model. Note that in this formulation, the transition model is sampled only once after the agent specifies the policy. Since the model remains fixed (static) throughout each episode, we will alternatively refer to this formulation as the static soft-robust criterion.

**Proposition 1.** There may be no stationary deterministic policy \( \pi \in \Pi_D \) that attains the optimal objective of the soft-robust optimization problem (7).

Proposition 1 can be proved by setting \( \alpha = 1 \) and using Theorem 2 in [4] that shows the optimal policy for a weighted sum of expected returns in an infinite horizon setting may be randomized. Note that we state the soft-robust criterion in the context of stationary policies. However, history-dependent randomized policies may outperform stationary ones [12].

**Proposition 2.** Computing the optimal policy of the soft-robust problem (7) is NP-hard.

The proof follows by setting \( \alpha = 0 \) and using Theorem 1 in [5] that shows optimizing the policy for the weighted sum of expected returns in an infinite-horizon setting is NP hard.

4 Soft-Robust Optimization

As Proposition 2 suggests, computing optimal randomized policies for the soft-robust criterion is intractable. In this section, we present two tractable methods that approximately solve the soft-robust objective. We first approximate the soft-robust criterion using a MILP and show that its solution corresponds to the deterministic optimal policy of the soft-robust criterion. Because MILPs cannot scale beyond medium-sized problems, we propose a scalable method that builds on the connection between the soft-robust criterion and RMDPs, and show that the soft-robust objective can be approximately solved using variants of the existing scalable RMDP techniques.

4.1 Soft-Robust Optimization using MILP

Before formulating the soft-robust criterion as a MILP and discussing its implications, we posit the underlying assumptions and notation. We assume a tabular MDP setting where the transition model \( \hat{P} \) can be represented by a \( S \times A \times S \) array. We also assume that \( \hat{P} \) can be approximately represented by the posterior samples \( \hat{P}^\omega \) defined in Section 2, as is common in methods like sample average approximation (SAA) [30]. In simple settings, like with Dirichlet priors, the samples can be derived by sampling from the analytical posterior. In more complex and realistic settings, the discrete approximation comes from MCMC samples (e.g., [2, 9]). We denote by \( u(s, a, \omega) \in \mathbb{R}_{+}^{S \times A \times |\Omega|} \), the state-action occupancy frequency of the deterministic optimal soft-
The MILP formulation in Figure 1, which we call SR-MILP, optimizes the soft-robust objective only over the set of deterministic policies, $\Pi_D$, whereas the global optimal policy for (6) or (7) may be history-dependent and randomized. Although this may seem like a major drawback, to the best of our knowledge, there is no method for computing history-dependent policies in an infinite-horizon setting. Moreover, in practice, deterministic policies are often preferred over randomized ones, especially in high-risk medical decision-making problems, where randomizing between different treatment regimes outside clinical trials may be considered unethical.

The SR-MILP formulation in Figure 1 has two important properties that make it an attractive method for optimizing the soft-robust criterion. First, as shown above, the deterministic optimal policies computed by the MILP are exact. Second, recall that the soft-robust formulation assumes that the transition model for each episode is fixed, as shown in Lemma 1. The MILP formulation in Figure 1 retains this assumption. As a result, it allows the agent to use any information about the model that can be inferred from the current state features to predict the model for that episode and take more informed decisions. In contrast, methods that assume dynamic model uncertainty, cannot exploit any leaked model information and instead require the agent to consider all possible models at every step, which can be too conservative. For example, consider a cancer treatment problem, where the agent has to decide on the amount of chemotherapy to be administered to a patient. The current state of the cancer patient may reveal some information about how the patient has been reacting to the chemotherapy treatment. This in turn may enable the agent to predict the patient model and take more informed decisions instead of optimizing the policy for the worst possible patient model at every stage of the treatment. Thus, we conclude that our MILP formulation has the potential for generating much less conservative policies than the existing methods.

However, it is important to note that while MILPs work well for problems of small to medium size (hundreds of states), they do not scale to large or infinite state problems. In such cases, the dynamic model uncertainty assumption is useful for tractably optimizing an approximation of the soft-robust objective. In the next section, we adopt this assumption to approximately solve the soft-robust criterion in large problems and bound the errors introduced by making this assumption.

### 4.2 Soft-Robust Optimization using RMDPs

To derive scalable approximations to the soft-robust criterion, we draw connections to RMDPs, for which scalable algorithms are available. Robust representation of any coherent risk measure can be reduced to S-rectangular distributionally RMDPs [51]. We adopt similar assumptions to derive the S-rectangular RMDP corresponding to the soft-robust criterion.

Our work differs from Xu and Mannor [51] in two main aspects. First, we explicitly derive the ambiguity sets of the S-rectangular RMDP corresponding to the soft-robust criterion, instead of assuming an abstract representation of ambiguity sets for coherent risk measures. Second, we present a scalable algorithm for optimizing the S-rectangular RMDP approximation of the soft-robust criterion. Note that, in contrast to the percentile criterion, the ambiguity sets corresponding...
Step 2: We approximate the soft-robust optimization criterion in (7) as

$$\max_{\pi \in \Pi} \rho^D(\pi) = \max_{\pi \in \Pi} \min_{\xi \in \Xi} \rho\left(\pi, E_{\hat{P} \cdot \xi}(\hat{P})\right).$$  \hspace{1cm} (9)$$

Note that, in contrast to (7), the expectation in (9) is inside of the return function $\rho$. This approximate formulation is helpful because, as can be shown readily, it equals to a non-rectangular RMDP in (3), with the ambiguity set $P_D \subseteq (\Delta^S)^{S \times A}$ constructed as

$$P_D = \left\{ \sum_{\omega \in \Omega} \xi_\omega \cdot \hat{P}_s^\omega \mid \xi \in \Xi \right\}. \hspace{1cm} (10)$$

The superscript $D$ indicates that this is the non-rectangular ambiguity set corresponding to the dynamic formulation.

Before proceeding to the second step, it is important to discuss the interpretation of the dynamic formulation in (9). It is tempting to interpret this formulation as a version of (7) in which a new realization of the model is sampled from the same distribution in every time step. This is, however, not true in general.

Step 2: Although solving a non-rectangular RMDP is NP-hard [47], it can be turned into a tractable rectangular one by a process called rectangularization in the context of dynamic risk measures [20, 37]. Rectangularization constructs the smallest rectangular set that comprises the entire non-rectangular set. Since the rectangular set is a superset of the non-rectangular one and the robust objective optimizes for the worst-case, the rectangular robust objective bounds the non-rectangular one from below.

To formalize the rectangularization step, assume that the soft-robust ambiguity sets in (8), which we denote by $\mathcal{P}_s^R = \bigotimes_{s \in S} P^R_s$, is S-rectangular. Consider the ambiguity set decomposition as in (4) with

$$\mathcal{P}_s^R = \left\{ \sum_{\omega \in \Omega} \xi_\omega \cdot \hat{P}_s^\omega \mid \xi \in \Xi \right\}, \hspace{1cm} (11)$$

where $\Xi$ is defined in (8) and $\hat{P}^\omega_s \in (\Delta^S)^A$ is a value of the posterior sample $\hat{P}^\omega_s$ in state $s$. As noted in the introduction, such S-rectangular RMDP can be solved efficiently [16]. The mnemonic superscript $R$ in $\mathcal{P}_s^R$ indicates that this is the S-rectangular ambiguity set.

Finally, the following optimization problem defines the S-rectangular objective $\rho^R: \Pi \rightarrow \mathbb{R}$:

$$\rho^R(\pi) = \min_{P \in \mathcal{P}_s^R} \rho(\pi, P). \hspace{1cm} (12)$$

The following theorem shows that the rectangularized return is a lower-bound on the return of the intractable non-rectangular RMDP.

**Theorem 1.** The ambiguity sets $P_D$ and $P_s^R$ defined in (10) and (11) respectively satisfy that $P_D \subseteq P_s^R$, and, therefore,

$$\rho^R(\pi) \leq \rho^D(\pi), \hspace{1cm} \forall \pi \in \Pi.$$
and \( s \in \mathcal{S} \), can be computed by solving:
\[
\max_{d \in \Delta^A, z \in \mathbb{R}} \left( 1 - \lambda \right) \sum_{a \in \mathcal{A}} \sum_{w \in \Omega} d_a f_w (\tilde{P}_{s,a}^w)^\top (r_{s,a} + \gamma \cdot v_s) + \lambda \cdot \left( z - \frac{1}{1 - \alpha} \sum_{w \in \Omega} f_w \cdot y_w \right)
\]
subject to \( y_w \geq z - \sum_{a \in \mathcal{A}} d_a (\tilde{P}_{s,a}^w)^\top (r_{s,a} + \gamma \cdot v) \), \( \forall \omega \in \Omega \).

**Proposition 4.** The optimal objective in (13) equals to the optimal S-rectangular update in (5) with the ambiguity set defined as in (11).

**Proof.** The correctness of the linear program in (13) follows from (11) by algebraic manipulation. \( \square \)

### 4.3 Soft-Robust Value Iteration Algorithm

In this section, we describe the soft-robust value iteration algorithm (SRVI) outlined in Algorithm 1. This algorithm uses a variant of the Robust Projected Value Iteration (RPVI) [44] to efficiently learn the optimal S-rectangular soft-robust value function using linear function approximation. The original RPVI algorithm [44] is designed to only learn deterministic optimal policies and value functions for SA-rectangular RMDPs; we extend it to computing randomized policies for S-rectangular soft-robust MDPs.

Let \( v \in \mathbb{R}^S \) denote the soft-robust value function approximated using a linear weighted sum of features:
\[
v(s) = \phi(s)^\top w,
\]
where \( \phi(s) \in \mathbb{R}^l \) with \( l \ll S \) represents the \( l \)-dimensional features of state \( s \) and \( w \in \mathbb{R}^l \) is the weight vector. Let \( \Phi \in \mathbb{R}^{M \times l} \) denote the sample feature matrix of \( \phi \) and \( q \in \Delta^S \) denote the steady state distribution over states \( s \in \mathcal{S} \) induced by policy \( \pi \). We further denote by \( \Psi \), the projection operator onto the subspace \( \Phi \) with respect to a weighted Euclidean norm.

Let \( \pi_S \in \Pi \) represent the optimal soft-robust policy for the S-rectangular RMDP with soft-robust ambiguity set given in (11). We compute the soft-robust value function \( v \in \mathbb{R}^S \) corresponding to \( \pi_S \) by iteratively solving the projected soft-robust Bellman equation \( v = \Psi \tilde{P} v \) using the RPVI algorithm [44], as outlined in Algorithm 1.

SRVI in Algorithm 1 proceeds as follows. The algorithm takes the posterior distribution of transition probability models \( f \), confidence \( \alpha \), and risk factor \( \lambda \) as input. The posterior distribution can be learned from the batch of data \( D \) using Bayesian regression models [48] or variational inference techniques [18]. We then sample \( N \) parametric transition models \( \{P_{\theta}^1, \ldots, P_{\theta}^N\} \) from the posterior distribution \( f \) and constructs the mean transition model \( \tilde{P}_\theta \).

In each iteration \( k \) of the algorithm, we simulate episodes using the mean transition probability model \( \tilde{P}_\theta \) and reconstruct data \( D \) to approximately reflect the stationary state distribution induced by the current policy. Let \( \sigma_s : \mathbb{R}^S \rightarrow \mathbb{R}, s \in \mathcal{S} \) be the value of the S-rectangular soft-robust Bellman operator for value function \( v = \Phi w \) with the weight vector \( w \in \mathbb{R}^l \) defined as
\[
\sigma_s(w) = \max_{d \in \Delta^A} \min_{\xi \in \mathbb{R}} \sum_{a \in \mathcal{A}} \sum_{w \in \Omega} d_a \cdot \xi_w \cdot \mathbb{E}_{(s',s) \sim \tilde{P}_{\theta}(s,a)} \left[ r(s,a,s') + \gamma \cdot \Phi (s')^\top w \right],
\]
where \( \Xi \) is defined in (8). Further, let \( \Sigma(w) = (\sigma_s(w))_{t=1,\ldots,M} \) be the sample matrix of the soft-robust Bellman values. In each iteration of the algorithm, we update the weight vector \( w_k \) using the reconstructed data as
\[
w_{k+1} = (\Phi^\top Q \Phi)^{-1} \Phi^\top Q P \Sigma(w_k),
\]
where \( Q = \text{diag}(q) \). Since it is not possible to exactly compute the terms in (15), we approximate them using Sample Average Approximate (SAA) [41] as
\[
\Phi^\top Q \Phi \sim \frac{1}{M} \sum_{t=1}^M \phi(s_t)\phi(s_t)^\top,
\]
\[
\Phi^\top Q P \Sigma(w) \sim \frac{1}{M} \sum_{t=1}^M \sigma_s(w).
\]

---

**Algorithm 1:** Soft-Robust Value Iteration (SRVI) for Problems with Large State Spaces

**Input:** Confidence \( \alpha \), risk factor \( \lambda \), posterior distribution \( f \)

**Output:** Soft-Robust value function \( v \), target Bellman residual \( \epsilon \)

Initialize: weight vector \( w_0 \), counter \( k \leftarrow 1 \);
Sample \( N \) parametric \( P_{\theta}^1, \ldots, P_{\theta}^N \) from \( f \);
Compute the mean model \( \tilde{P}_\theta = \mathbb{E}[P_{\theta}] \) from samples \( P_{\theta}^1, \ldots, P_{\theta}^N \);
repeat

\[
\begin{align*}
&\text{Using mean model } \tilde{P}_\theta \text{ and policy obtained by solving (14), simulate episodes and construct } D_k \leftarrow \{(s_t, a_t, s_{t+1})\}_{t=1}^M \text{ and } \Phi_k ;
&\text{Set } w_k \leftarrow (\Phi_k^\top Q \Phi_k)^{-1} (\Phi_k^\top Q P \Sigma(w_k)) \text{ using (14) and (16)} ;
&k \leftarrow k + 1 ;
\end{align*}
\]

until \( \|\Phi_k^\top w_k - \Phi_k^\top w_{k-1}\|_\infty \leq \epsilon ;\)
return \( w_k \)
The optimization problem in (14) can be solved by formulating it as a stochastic linear program and then using the SAA method [44]. The expected features $E_{s_{t+1}, a_{t+1}}[\Phi(s_{t+1})]$ of state $s_{t+1}$ can be approximated by collecting i.i.d. samples of $s_{t+1}$ from the posterior distribution $P^\omega(s_{t+1})$ and computing the empirical mean of the state features.

We repeatedly update the weight vector $w$ using (15) until the soft-robust value function $\Phi^T w$ converges to the unique projected fixed-point of $T_{pR}$. Given the optimal weight vector $w^*$, the optimal policy for any state $s_t \in S$ can be computed online by solving the inner optimization problem in (14).

5 RMDP Approximation Error

In this section, we derive new bounds on the approximation errors incurred when solving the RMDP formulation described in Sections 4.2 and 4.3. The error bounds provide insight into when the proposed approximations may be useful to lay out the possible directions for further improvements. Our results show, in particular, that the error introduced by resorting to the RMDP formulation depends on two main factors. The first factor depends on how the model uncertainty impacts the occupancy frequency. The second factor depends on whether the worst-case model is consistent across states. Finally, we assume throughout the section that $\Omega$ is finite to avoid technical issues that may distract from the main ideas.

We first show that the error introduced by maximizing the dynamic criterion instead of the static one is bounded by the difference between occupancy frequencies among the uncertain models. We use the symbol $d^\omega_\pi \in \mathbb{R}^S$ to denote the state occupancy frequency for each policy $\pi \in \Pi$ and model $\omega \in \Omega$ defines as

$$d^\omega_\pi = \left( I - \gamma \cdot (P^\pi)^T \right)^{-1} p_0. \quad (17)$$

Equipped with this notation, we can prove the following bound on the performance loss of the dynamically robust policy.

**Theorem 2.** The difference between static and dynamic returns is bounded for each $\pi \in \Pi$ as

$$|\rho^S(\pi) - \rho^S(\pi)| \leq \frac{\gamma \cdot r_{\max}}{1 - \gamma} \cdot \epsilon_1(\pi),$$

where

$$\epsilon_1(\pi) = \max_{\omega_1, \omega_2 \in \Omega} \|d^\omega_\pi - d^{\omega_2}_\pi\|_1.$$

We provide the proof of the theorem in Appendix E.

**Corollary 1.** The soft-robust return $\rho^S(\pi)_R \in \arg \max_{\pi \in \Pi} \rho^R(\pi)$ computed by Algorithm 1 satisfies that

$$\rho^S(\pi^*_R) - \rho^S(\pi^*_R) \leq \frac{1}{1 - \gamma} \left( 2 \cdot \gamma \cdot \epsilon_1 \cdot r_{\max} + \epsilon_2 \right),$$

where $\epsilon_1 = \max_{\pi \in \Pi} \epsilon_1(\pi)$, and $\epsilon_1(\pi)$ and $\epsilon_2$ are defined as in Theorems 2 and 3 respectively, and $\pi^*_R \in \arg \max_{\pi \in \Pi} \rho^S(\pi)$.

The proof of the corollary is reported in Appendix E.

6 Related Work

Numerous robust objectives for mitigating model uncertainty have been proposed in the literature. We discuss a number of them in more details in this section.
Dynamic robust objectives. A vast majority of work in Robust RL has studied objectives that assume a dynamic model uncertainty for achieve tractability. Mankowitz et al. [26] proposed algorithms that optimize entropy-regularized policies against the worst model in the uncertainty set. While these algorithms scale to continuous state and action spaces, they do not provide any kind of probabilistic guarantees on the expected returns like our framework and compute overly conservative policies. Actor-critic algorithms that optimize the mean of the expected returns computed for a fixed distribution over models in the uncertainty set [7].

In contrast prior work, our dynamic soft-robust algorithm dynamically computes the distribution over uncertain models that provide guarantees on the user-specified quantile of the expected returns for the optimal policy. Derman et al. [8] introduced scalable algorithms that optimize an RMDP objective while accounting for changing dynamics. This framework also suffers from the shortcomings of the percentile criterion. Other related work constructs a plausible framework to incorporate any probabilistic information about the uncertain models in RMDPs [51]. Although their work shows a connection between coherent risk measures and distributionally-robust MDPs, they do not aim to address the shortcomings of the percentile criterion. In the same vein as our work, policy gradient methods for optimizing CVaR of expected returns have been studied [15]. Nonetheless, their methods do not exploit the coherent properties of this measure and only tend to find local optimal policies.

Static robust objectives. Few works have focused on optimizing robust objectives while retaining the static model uncertainty assumption [3, 4, 29, 42]. However, we note that the robust objectives used in these works are quite different than ours. Steimle et al. [42] proposed a mixed integer linear program and a fast heuristic algorithm to optimize the weighted expected returns across different models in a finite-horizon setting, whereas our objective optimizes the policy for the worst distribution over models in the ambiguity set. Buchholz and Scheftelowitsch [4] used the same objective as in Steimle et al. [42], but considered both finite and infinite-horizon settings. They proposed a MILP for calculating the exact deterministic policy in the finite-horizon setting, and other approximation algorithms that optimize a finite class of randomized Markovian policies for the infinite-horizon case. In another similar work, Meraklı and Küçükyavuz [29] proposed an approximate MILP for optimizing the percentile-criterion. However, since the original objective is non-convex, the approximation may not generate optimal deterministic solutions.

Ambiguity set optimization. Other related work has focused on considering partial correlations between uncertain model parameters to mitigate the conservativeness of learned policies [12, 27, 28, 38]. Some examples are k-rectangular [27, 28] and r-rectangular [12] ambiguity sets. These approaches mitigate the conservativeness of S- and SA-rectangular ambiguity sets by capturing correlations between the uncertainty and by limiting number of times the uncertain parameters deviate from the mean parameters. Despite this progress, most of this works still relies on weak concentration bounds for the construction of ambiguity sets, which can make the ambiguity sets unnecessarily large and result in conservative policies.

7 Experimental Evaluation

In this section, we present two case studies to demonstrate the performance of the soft-robust criterion in discrete and continuous domains. For each experiment, we simulate episodes to generate a batch of data \( D \) from the true model \( P^* \). Using the data, we first construct a posterior distribution over transition models and then generate a set of models each for training and testing purposes. To compare the soft-robust criterion with other baseline robust objectives, we compare the mean and robust performance of the learned policies given by the average and CVaR measure of expected returns computed over the test models at 95% confidence level.

7.1 Integrated Pest Control Problem

The domain represents a simplified integrated pest control problem. The decision-maker must decide which, if any, pesticide to use during the growing season. The state represents the pest population, determined by trapings, and action represents the pesticide. Exponential pest growth dynamics drive the transition models and the rewards measure the net profit of the yields after accounting for the pesticides costs. The pest population reduces marketable yields which impacts the rewards.

The domain’s MDP consists of 51 states, each represents the current pest population as determined by trapping (0 means no pest population). There are 5 actions available, with each action representing the use of an increasingly potent pesticide. The true transition probabilities are based on a logistic model of population growth as described in [46]. The discount factor is \( \gamma = 0.9 \).

To compute the posterior distribution over \( \hat{P} \), we gather 300 state-action transition samples from a single episode. Using these transition samples, we fit an exponential population model [24] using the JAGS
modeling language and sample 100 posterior samples using MCMC sampling. We then use these samples to formulate and solve the MILP in Figure 1 and to run Algorithm 1. We use confidence $\alpha = 0.7$ for both the percentile criterion and soft-robust objective for the evaluation. We also use $\lambda = 0.5$ for the soft-robust objective.

Figure 2 compares the return distribution of the soft-robust MILP policy with the robust BCR solution [38], and the nominal policy, which solves the expected transition model $E[P]$. Although the nominal policy achieves the highest mean return, it has a significant probability of incurring a high loss (over $5,000$). The policy that targets the percentile criterion improves robustness, but still ignores the fat tail and degrades the mean return. The soft-robust objective balances the mean return with robustness.

Figure 3 illustrates the trade-off between mean and worst-case performance for several robust methods for different choices of $\lambda \in [0, 1]$ as indicated by the floating labels. We compare the optimal MILP policy in Figure 1 (SR-MILP) and Algorithm 1 (SRVI) with RSVF [38] and BCR [38]. Note that RSVF and BCR optimize the percentile criterion, which has no inherent notion of the trade-off between robust and mean performance. We simulate the effect of $\lambda$ by simply shrinking the ambiguity sets in the RMDP formulations (multiplying the budget by $\lambda$). The results show that our soft-robust algorithms outperform the earlier methods both in terms of the average and robust performance. The methods also provide a good trade-off between the mean and robust performance with an appropriate choice of $\lambda$.

### 7.2 Cancer Growth Simulator

The cancer simulator models the growth of tumors in cancer patients. The state is a 4-dimensional vector that captures the dynamics of the tumor’s growth. This problem requires a binary decision to be made every month on whether chemotherapy should be administered to the patient. We refer the readers to [34] for more details on the dynamics of this problem.

We model $P_{s_t,a_t}$ as a Multivariate Normal Random variable with a linear function of state features weighted by weight vector $w \in \mathbb{R}^d$ as mean and a fixed diagonal covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. We sample a batch of 600 samples using $P_{s_t,a_t}$ and the optimal policy to construct the dataset $P$. Using this data, we train a multivariate Bayesian linear regression model to predict the posterior distribution of the weight vector $w$ and the covariance matrix $\Sigma$. We assume a Normal prior $N(0, 1)$ for each element of the weight vector $w$ and a HalfNormal(0.001) prior for the elements of the covariance matrix $\Sigma$. We construct the uncertainty set as shown in Algorithm 1 using 100 weight vectors and covariance matrices sampled from the posterior distribution using MCMC [19].

In each experiment, we expand the set of features using the polynomial basis of degree 2. We construct the mean model by taking the mean of the model parameters sampled from the posterior and use it for simulating episodes in all our experiments. We train 3 agents using SRVI, soft-robust actor-critic, and worst-case actor-critic algorithms. We represent the value functions of each algorithm using linear function approximators and the policies for soft-robust and worst-case actor-critic using a neural network with two hidden layers of size 200 and ReLU activation. We compute the optimal policy for SRVI online by solving the LP given in (14) using the learned value function. We set the discount factor to $\gamma = 0.8$. For each algorithm, we use 50 samples from the posterior model distribution for training and 50 samples for testing. We evaluate the policy learned by each algorithm by comparing the mean and CVaR of its expected return over all test
We proposed a new generic framework based on the soft-robust criterion to a non-rectangular RMDP and then making it tractable by rectangularization. The experimental results demonstrate that the proposed methods show promise. Future work should focus on broader and more rigorous empirical evaluation and approaches that can further tighten the approximation bounds while preserving tractability.

### 8 Conclusion

We proposed a new generic framework based on the soft-robust criterion that can balance expected and robust performance in reinforcement learning and handle heavy tail risks. We showed that a Mixed Integer Linear Program can be formulated to derive exact deterministic optimal solutions for the soft-robust objective. Since MILPs can only solve small-medium size problems, we propose a specific type of RMDPs as a scalable alternative to optimize an approximation of the soft-robust criterion. The approach is based on reducing the

**Table 1: Test results for cancer simulator**

| Method (λ)       | Mean Return | Robust Return |
|------------------|-------------|---------------|
| S-SRVI (1.0)     | 13.35 ± 0.58 | 0.19 ± 0.25   |
| S-SRVI (0.75)    | 14.69 ± 0.89 | -1.58 ± 1.60  |
| S-SRVI (0.0)     | 9.16 ± 0.80  | -21.13 ± 5.33 |
| SA-SRVI (1.0)    | 13.62 ± 1.30 | -1.64 ± 1.89  |
| SA-SRVI (0.75)   | 14.69 ± 0.72 | -0.14 ± 1.21  |
| SA-SRVI (0.0)    | 12.70 ± 0.40 | -17.18 ± 2.43 |
| Soft-Robust AC   | 8.33 ± 0.66  | -3.44 ± 0.63  |
| Worst-case AC    | 10.86 ± 0.72 | -0.73 ± 0.45  |

models.

Table 1 shows the comparison between mean and robust performance of S-Rectangular Soft-Robust Value Iteration (S-SRVI), SA-Rectangular Soft-Robust Value Iteration (SA-SRVI), Soft-Robust Actor-Critic (Soft-Robust AC) [7], and Worst-Case Actor-Critic (Worst-Case AC) [20] on the test models. The robust performance is computed as the CVaR of expected returns computed over the set of test models at 90% confidence interval. Notice that the S-SRVI algorithm has the highest robust performance when $\lambda = 1.0$ and the robustness of performance diminishes with increasing $\lambda$. Further, we observe that the SA-SRVI algorithm outperforms S-SRVI in mean performance at $\lambda = 0.75$ and $\lambda = 0.0$, but has lower robust performance than S-SRVI. This behavior is expected since SA-SRVI can only learn deterministic policies and hence it compensates loss in robust performance by learning robust policies with higher mean performance. Both, SA-SRVI and S-SRVI outperform Soft-Robust AC in mean and robust performance. This behavior as well is expected since Soft-Robust AC does not explicitly handle tail risks and is likely to find local optimal policies. Finally, note that while Worst-Case AC has comparable robust performance to SRVI, its mean performance is still low as it learns policies that only account for worst-case models and does not optimize for the overall mean performance.

### References

[1] Ben-Tal, A., Bertsimas, D., and Brown, D. B. (2010). A Soft Robust Model for Optimization Under Ambiguity. *Operations Research*, 58(4):1220–1234.

[2] Brown, D. S. and Niekum, S. (2018). Efficient Probabilistic Performance Bounds for Inverse Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*, pages 2754–2762.

[3] Buchholz, P. and Scheftelowitsch, D. (2019). Computation of weighted sums of rewards for concurrent MDPs. *Mathematical Methods of Operations Research*, 89(1):1–42.

[4] Buchholz, P. and Scheftelowitsch, D. (2020). Concurrent MDPs with finite Markovian policies. pages 37–53.

[5] Chow, Y. and Ghavamzadeh, M. (2014). Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, pages 3509–3517, Cambridge, MA, USA. MIT Press.

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

[7] Derman, E., Mankowitz, D., Mann, T. A., and Mannor, S. (2018). Soft-Robust Actor-Critic Policy-Gradient. In *Uncertainty in Artificial Intelligence (UAI)*.

[8] Derman, E., Mankowitz, D. J., Mann, T. A., and Mannor, S. (2019). A Bayesian approach to robust reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

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

[10] Ghavamzadeh, M., Engel, Y., and Valko, M. (2016). Bayesian policy gradient and actor-critic algorithms. *Journal of Machine Learning Research*, 17:1–53.

[11] Gottesman, O., Liu, Y., Sussex, S., Brunskill, E., and Doshi-Velez, F. (2019). Combining parametric and nonparametric models for off-policy evaluation. In *International Conference of Machine Learning (ICML)*, pages 4236–4245.

[12] Goyal, V. and Grand-Clément, J. (2018). Robust Markov decision process: Beyond rectangularity.
[13] Grand-Clement, J. and Kroer, C. (2020). First-Order Methods for Wasserstein Distributionally Robust MDP.

[14] Gupta, V. (2019). Near-optimal Bayesian ambiguity sets for distributionally robust optimization. Management Science.

[15] Hiraoka, T., Imagawa, T., Mori, T., Onishi, T., and Tsuruoka, Y. (2019). Learning robust options by conditional value at risk optimization. In NeurIPS.

[16] Ho, C. P., Petrik, M., and Wiesemann, W. (2018). Fast Bellman updates for robust MDPs. In International Conference on Machine Learning (ICML), pages 1979–1988.

[17] Ho, C. P., Petrik, M., and Wiesemann, W. (2020). Partial Policy Iteration for L1-Robust Markov Decision Processes. Technical report.

[18] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. Journal of Machine Learning Research, 14(1):1303–1347.

[19] Hoffman, M. D. and Gelman, A. (2011). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo.

[20] Iancu, D., Petrik, M., and Subramanian, D. (2011). Tight approximations of dynamic risk measures. Mathematics of Operations Research, 40.

[21] Iancu, D. A. and Trichakis, N. (2014). Pareto Efficiency in Robust Optimization. Management Science, 60(1):130–147.

[22] Iyengar, G. (2005). Robust dynamic programming. Mathematics Operations Research, 30:257–280.

[23] Kaufman, D. L. and Schaefer, A. J. (2013). Robust modified policy iteration. INFORMS Journal on Computing, 25(3):396–410.

[24] Kery, M. and Schaub, M. (2012). Bayesian Population Analysis Using WinBUGS. Elsevier.

[25] Lange, S., Gabel, T., and Riedmiller, M. (2012). Batch Reinforcement Learning. In Reinforcement Learning, pages 45–73.

[26] Mankowitz, D. J., Levine, N., Jeong, R. U., Abdolmaleki, A., Springenberg, J. T., Mann, T., Hester, T., and Riedmiller, M. A. (2020). Robust reinforcement learning for continuous control with model misspecification. ArXiv, abs/1906.07516.

[27] Mannor, S., Mebel, O., and Xu, H. (2012). Lightning does not strike twice: Robust MDPs with coupled uncertainty. ArXiv, abs/1206.4643.

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

[29] Merahi, M. and Küçükyavuz, S. (2019). Risk aversion to parameter uncertainty in Markov decision processes with an application to slow-onset disaster relief. ISE Transactions, pages 1–52.

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

[31] Petrik, M., Ghavamzadeh, M., and Chow, Y. (2016). Safe policy improvement by minimizing robust baseline regret. In International Conference on Neural Information Processing Systems, pages 2306–2314, USA.

[32] Prashanth, L. (2014). Policy gradients for CVaR-constrained MDPs.

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

[34] Ribba, B., Kaloshi, G., Peyre, M., Ricard, D., Calvez, V., Tod, M., Cajavec-Bernard, B., Idbah, A., Pismaras, D., Dainese, L., Pallud, J., Cartalat-Carel, S., Delatte, J., Honorat, J., Grenier, E., and Ducray, F. (2012). A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. Clinical Cancer Research, 18:5071–5080.

[35] Rockafellar, R. T. (1970). Convex Analysis.

[36] Rockafellar, R. T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. Journal of Risk, 2:21–41.

[37] Roorda, B., Schumacher, J., and Engwerda, J. (2005). Coherent acceptability measures in multi-period models. Mathematical Finance, 15:589–612.

[38] Russel, R. H. and Petrik, M. (2019). Beyond confidence regions: Tight Bayesian ambiguity sets for robust MDPs. In International Conference on Neural Information Processing Systems.

[39] Schied, A. (2004). Risk measures and robust optimization problems. In Symposium on Probability and Stochastic Processes.

[40] Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). Lectures on stochastic programming: Modeling and theory.

[41] Shapiro, A., Tekaya, W., da Costa, J. P., and Soares, M. P. (2013). Risk-neutral and risk-averse stochastic dual dynamic programming method. European Journal of Operational Research, (2):375–391.

[42] Steinle, L. N., Kaufman, D. L., and Denton, B. T. (2018). Multi-model markov decision processes.

[43] Sutton, R. S. and Barto, A. G. (2018). Reinforcement Learning: An Introduction. 2nd edition.

[44] Tamar, A., Mannor, S., and Xu, H. (2014). Scaling up robust MDPs using function approximation. In...
[45] Tang, Y., Yu Zhang, J., and Salakhutdinov, R. (2019). Worst cases policy gradients. In CoRL.

[46] Tirinzoni, A., Chen, X., Petrik, M., and Ziebart, B. D. (2018). Policy-conditioned uncertainty sets for robust Markov decision processes. In International Conference on Neural Information Processing Systems (NIPS), page 8953–8963.

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

[48] Wundervald, B. (2019). Bayesian linear regression.

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

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

[51] Xu, H. and Mannor, S. (2012). Distributionally robust Markov decision processes. Mathematics of Operations Research, 37(2):288–300.

[52] Yang, I. (2017). A Convex Optimization Approach to Distributionally Robust Markov Decision Processes With Wasserstein Distance. IEEE Control Systems Letters, 1(1):164–169.
A Auxiliary Results

The following lemma will be useful when bounding the RMDP approximation of the soft-robust objective.

**Lemma 2.** The vector-induced norms for a stochastic matrix \( P \in \mathbb{R}^{S \times S} \) satisfy that
\[
\|P\|_\infty = \|P^T\|_1 = 1.
\]

**Proof.** Let \( \mathcal{L}_1 = \{ x \in \mathbb{R}^S \mid \|x\|_1 = 1 \} \) be the \( L_1 \) ball and let \( \mathcal{L}_\infty = \{ x \in \mathbb{R}^S \mid \|x\|_\infty = 1 \} \) be the \( L_\infty \) ball. Then, using basic linear algebra, definition of induced matrix norms in steps (a), and the duality of the vector \( L_1 \) and \( L_\infty \) norm in step (b), we can establish the desired result as follows:
\[
\|P^T\|_1 = \max_{x \in \mathcal{L}_1} \|P^T x\|_1 = \max_{x \in \mathcal{L}_1} \max_{y \in \mathcal{L}_\infty} \|P y\|_\infty = \|P\|_\infty .
\]

The result follows because, as it is well-known, \( \|P\|_\infty = 1 \) for any stochastic matrix \( P \).

The following generic lemma establishes the bounds on the error between a maximizer of a function and a maximizer of an approximation of that function.

**Lemma 3.** Let \( x^* \in \arg\max_{x \in \mathcal{X}} f(x) \) and \( \tilde{x}^* \in \arg\max_{x \in \mathcal{X}} \tilde{f}(x) \) be the maximizers of some function \( f : \mathcal{X} \to \mathbb{R} \) and its approximation \( \tilde{f} : \mathcal{X} \to \mathbb{R} \) respectively. Then the optimality gap of \( \tilde{x}^* \) in non-negative and bounded by:
\[
f(x^*) - f(\tilde{x}^*) \leq |f(x^*) - \tilde{f}(x^*)| + |f(\tilde{x}^*) - \tilde{f}(\tilde{x}^*)| \leq 2 \cdot \max_{x \in \mathcal{X}} |f(x) - \tilde{f}(x)| .
\]

Moreover, when \( \tilde{f}(x) \leq f(x) \) for all \( x \in \mathcal{X} \), then the optimality gap of \( \tilde{x}^* \) reduces to:
\[
f(x^*) - f(\tilde{x}^*) \leq f(x^*) - \tilde{f}(x^*) .
\]

**Proof.** First, the following basic inequality follows by algebraic manipulation as:
\[
f(x^*) - f(\tilde{x}^*) = f(x^*) - f(\tilde{x}^*) + \tilde{f}(\tilde{x}^*) - \tilde{f}(\tilde{x}^*)
\]
\[
= f(x^*) - \tilde{f}(\tilde{x}^*) + \{ \tilde{f}(\tilde{x}^*) - f(\tilde{x}^*) \}
\]
\[
\leq f(x^*) - \tilde{f}(\tilde{x}^*) + \{ \tilde{f}(\tilde{x}^*) - f(\tilde{x}^*) \}
\]
\[
\leq \max_{x \in \mathcal{X}} |f(x) - \tilde{f}(x)|.
\]

Then, the inequality (18) follows from (20b) because \( x^* \in \mathcal{X} \) and \( \tilde{x}^* \in \mathcal{X} \) and therefore
\[
|f(x^*) - \tilde{f}(x^*)| \leq \max_{x \in \mathcal{X}} |f(x) - \tilde{f}(x)|
\]
\[
|\tilde{f}(\tilde{x}^*) - f(\tilde{x}^*)| \leq \max_{x \in \mathcal{X}} |f(x) - \tilde{f}(x)| .
\]

The inequality (19) follows from (20a) because \( \tilde{f}(x) \leq f(x) \) for all \( x \in \mathcal{X} \) and therefore
\[
f(x^*) - f(\tilde{x}^*) \leq f(x^*) - \tilde{f}(x^*) + \{ \tilde{f}(\tilde{x}^*) - f(\tilde{x}^*) \}
\]
\[
\leq f(x^*) - \tilde{f}(x^*)
\]
\[
\text{Because } \tilde{f}(\tilde{x}^*) - f(\tilde{x}^*) \leq 0 .
\]

B Proofs: Section 3

**Proof of Lemma 1.** First, we show that the negations of the terms in the soft-robust objective (6) are support functions [35] of convex sets. For any random variable \( X : \Omega \to \mathbb{R} \) with probability measure function \( f \), the robust representation of CVaR takes the following form (e.g., [36, 39]):
\[
\text{CVaR}^\alpha[X] = \min_{\xi \in \Delta^\Omega} \left\{ \sum_{\omega \in \Omega} \xi_\omega \cdot X(\omega) \mid \xi_\omega \leq \frac{1}{1 - \alpha} f_\omega, \forall \omega \in \Omega \right\} ,
\]

(21)
and, therefore, the CVaR term in (6) becomes

$$\text{CVaR}^\alpha_{\tilde{\rho} \sim f} \left[ \rho(\pi, \tilde{P}) \right] = \min_{\xi \in Q^{\text{CVaR}}} \sum_{\omega=1}^{D} \xi_\omega \cdot \rho(\pi, \tilde{P}^{\omega})$$  \hspace{1cm} (22)$$

where the set $Q^{\text{CVaR}}$ is defined as

$$Q^{\text{CVaR}} = \left\{ \xi \in \Delta^D \mid \xi_\omega \leq \frac{1}{1-\alpha} f_\omega, \ \omega \in \Omega \right\}.$$ \hspace{1cm} \hspace{1cm} (23)

As a result of (22), the function $X \mapsto -\text{CVaR}^\alpha_{\tilde{\rho} \sim f}[-X]$ for $X : \Omega \to \mathbb{R}$ is the support function of set $Q^{\text{CVaR}}$. Note that we are interpreting the random variable $X$ as a vector over $\mathbb{R}^\Omega$. Similarly, the mean term in (6) trivially equals to

$$\mathbb{E}_{\rho \sim f} \left[ \rho(\pi, \tilde{P}) \right] = \min_{\xi \in \mathbb{R}^k} \sum_{\omega \in \Omega} \xi_\omega \cdot \rho(\pi, \tilde{P}^{\omega}),$$

where $Q^\mathbb{E} = \{ f \}$ is a singleton set. As with the CVaR above, it can be seen readily that the function $X \mapsto -\mathbb{E}_{\rho \sim f}[-X]$ for $X : \Omega \to \mathbb{R}$ is the support function of $Q^\mathbb{E}$.

Next, any two support functions $f_1(z) = \max_{q \in Q_1} z^\top q$ and $f_2(z) = \max_{q \in Q_2} z^\top q$ over convex sets $Q_1, Q_2$ satisfy for $\lambda \in [0, 1]$ (see for example Chapter 13 of [35]),

$$\lambda \cdot f_1(z) + (1 - \lambda) \cdot f_2(z) = \max_q \left\{ q^\top z \mid q \in (\lambda \cdot Q_1 + (1 - \lambda) \cdot Q_2) \right\}.$$ \hspace{1cm} (24)

Multiplying the equality in (24) by $-1$, and using $-z$ as the parameter, we get:

$$-\lambda \cdot f_1(-z) - (1 - \lambda) \cdot f_2(-z) = -\max_q \left\{ -q^\top z \mid q \in (\lambda \cdot Q_1 + (1 - \lambda) \cdot Q_2) \right\}$$

$$= \min_q \left\{ q^\top z \mid q \in (\lambda \cdot Q_1 + (1 - \lambda) \cdot Q_2) \right\}.$$ \hspace{1cm} (25)

Consider the sets $Q_1 = Q^{\text{CVaR}}, Q_2 = Q^\mathbb{E}$ and support functions $f_1(X) = -\text{CVaR}^\alpha[-X]$ and $f_2(X) = -\mathbb{E}[-X]$ in (25). Then, we can reformulate (6) as:

$$\tilde{\rho}^S(\pi) = (1 - \lambda) \cdot \mathbb{E} \left[ \rho(\pi, \tilde{P}) \right] + \lambda \cdot \text{CVaR}^\alpha \left[ \rho(\pi, \tilde{P}) \right]$$

$$= \min_{\xi \in \Delta^N} \left\{ \sum_{\omega \in \Omega} \xi_\omega \cdot \rho(\pi, \tilde{P}^{\omega}) \mid \xi = \lambda \cdot \xi_1 + (1 - \lambda) \cdot \xi_2, \ \xi_1 \in Q^{\text{CVaR}}, \ \xi_2 \in Q^\mathbb{E} \right\}.$$ \hspace{1cm} \hspace{1cm} (26)

The the feasible set in the equation above in terms in $\xi, \xi_1 \in \mathbb{R}^N$ (note that $\xi_2 = f$) is represented by these inequalities,

$$\xi = \lambda \cdot \xi_1 + (1 - \lambda) \cdot f$$

$$\xi_1 \leq \frac{1}{1-\alpha} \cdot f$$

$$\xi \geq 0 \hspace{1cm} \xi_1 \geq 0 \hspace{1cm} 1^\top \xi = 1 \hspace{1cm} 1^\top \xi_1 = 1.$$ \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} (27)

Now, substituting $\xi_1 = \frac{1}{\lambda} \cdot (\xi - (1 - \lambda) \cdot f)$ to the inequalities above, we get

$$0 = 0 \hspace{1cm} \xi \geq 0 \hspace{1cm} 1^\top \xi = 1$$

$$\frac{1}{\lambda} \cdot (\xi - (1 - \lambda) \cdot f) \leq \frac{1}{1-\alpha} \cdot f$$

$$\frac{1}{\lambda} \cdot (\xi - (1 - \lambda) \cdot f) \geq 0 \hspace{1cm} 1^\top \left( \frac{1}{\lambda} \cdot (\xi - (1 - \lambda) \cdot f) \right) = 1,$$

which, using $\lambda \in [0, 1]$ and $f \in \Delta^N$, reduces to:

$$0 = 0 \hspace{1cm} \xi \geq 0 \hspace{1cm} 1^\top \xi = 1$$

$$\xi \leq \frac{\lambda}{1-\alpha} \cdot f + (1 - \lambda) \cdot f$$

$$\xi \geq (1 - \lambda) \cdot f \hspace{1cm} 0 = 0,$$

The result then follows by simple algebraic manipulation.
C Proofs: Section 4.1

Proof of Proposition 3. For a given MDP with transition probabilities $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta^\mathcal{S}$, and any fixed deterministic policy $\pi \in \Pi_D$ can be evaluated by solving the following linear program (often referred to as the dual formulation) [33]:

$$\begin{align*}
\text{maximize} & \quad \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} d(s, a) \sum_{s \in \mathcal{S}} P(s, a, s') r(s, a, s') \\
\text{subject to} & \quad d(s, a) = \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \gamma \cdot d(s', a') P(s', a', s) + p_0(s), \quad \forall s \in \mathcal{S}, a \in \mathcal{A} \\
& \quad d(s, a) \leq \frac{\pi(s, a)}{1 - \gamma}, \quad \forall s \in \mathcal{S}, a \in \mathcal{A} \\
& \quad \sum_{a \in \mathcal{A}} \pi(s, a) = 1, \quad \forall s \in \mathcal{S}.
\end{align*}$$

The optimal solution $d^*$ to (26) can be shown to be the occupancy frequency for the policy $\pi$. In addition, the optimal objective of (26) can be shown to be equal to the return $\rho(\pi, P)$. □

D Proofs: Section 4.2

Proof of Theorem 1. From the construction of $\mathcal{P}^R$, one can readily show that $\mathcal{P}^R \supseteq \mathcal{P}^D$, which trivially implies the desired lower bound inequality. □

E Proofs: Section 5

In this section, we describe the technical results that underlie the proof of Corollary 1.

The following lemma bounds the difference between a convex combination of occupancy frequencies and the occupancy frequency of the convex combination of transition functions. It serves as the main technical tool when bounding the difference between dynamic and static objectives.

Lemma 4. Consider stochastic matrices $P_i \in (\Delta^\mathcal{S})^\mathcal{S}$, $i = 1, \ldots, N$ with occupancy frequencies $d_i = (I - \gamma P_i^\top)^{-1} p_0$. Let $P_\beta = \sum_{i=1}^N \beta_i \cdot P_i$ be the convex combination of $P_i$ for a given $\beta \in \Delta^N$ and let $d_\beta = (I - \gamma P_\beta^\top)^{-1} p_0$ be its occupancy frequency. The convex combination of individual occupancy frequencies is denoted by $e_\beta = \sum_{i=1}^N \beta_i \cdot d_i$. Then:

$$\|d_\beta - e_\beta\|_1 \leq \frac{\gamma}{1-\gamma} \cdot \epsilon_1,$$

when $\|d_i - d_j\|_1 \leq \epsilon_1$ for each $i = 1, \ldots, N$ and $j = 1, \ldots, N$.

Proof. Recall that the following identities hold for the occupancy frequencies [33]:

$$\begin{align*}
d_i &= \gamma \cdot P_i^\top d_i + p_0, \quad i = 1, \ldots, N, \\
d_\beta &= \gamma \cdot P_\beta^\top d_\beta + p_0.
\end{align*}$$

Using the identity above and the fact that $\beta \in \Delta^\mathcal{S}$, we obtain a similar expression for $e_\beta$:

$$e_\beta = \gamma \sum_{i=1}^N \beta_i \cdot P_i^\top d_i + p_0.$$

Because $d_\beta$ need not be a convex combination of $d_i$ we use the following representation of the difference between
Applying the Lemma 2 combined with the Neumann series representation of matrix inverse implies that expressed in terms of their occupancy frequencies using (31) as

\[ d_\beta - e_\beta = \gamma \cdot P_\beta^T d_\beta - \gamma \sum_{i=1}^{N} \beta_i \cdot P_i^T d_i \]

from (27) and (28)

\[ = \gamma \cdot P_\beta^T d_\beta - \gamma \cdot P_\beta^T e_\beta + \gamma \cdot P_i^T e_\beta - \gamma \sum_{i=1}^{N} \beta_i \cdot P_i^T d_i \]

add 0

\[ = \gamma \cdot P_\beta^T d_\beta - \gamma \cdot P_\beta^T e_\beta + \gamma \sum_{i=1}^{N} \beta_i \cdot P_i^T e_\beta - \gamma \sum_{i=1}^{N} \beta_i \cdot P_i^T d_i \]

definition of \( P_\beta \)

\[ = \gamma \cdot P_\beta^T (d_\beta - e_\beta) + \gamma \sum_{i=1}^{N} \beta_i \cdot P_i^T (e_\beta - d_i) . \]

simplify

Next, subtracting \( \gamma \cdot P_\beta^T (d_\beta - e_\beta) \) from both sides of the equality above, and multiplying by the appropriate matrix inverse leads to:

\[ d_\beta - e_\beta = \gamma \sum_{i=1}^{N} \beta_i \cdot (I - \gamma \cdot P_\beta^T)^{-1} P_i^T (e_\beta - d_i) . \]  

(29)

Applying the \( L_1 \) norm to both sides of (29) we get that:

\[
\|d_\beta - e_\beta\|_1 = \gamma \sum_{i=1}^{N} \beta_i \cdot \| (I - \gamma \cdot P_\beta^T)^{-1} P_i^T (e_\beta - d_i) \|_1 \\
\leq \gamma \sum_{i=1}^{N} \beta_i \cdot \|(I - \gamma \cdot P_\beta^T)^{-1}\|_1 \cdot \|P_i^T\|_1 \cdot \|e_\beta - d_i\|_1 \\
\leq \gamma \sum_{i=1}^{N} \beta_i \cdot \|(I - \gamma \cdot P_\beta^T)^{-1}\|_1 \cdot \|e_\beta - d_i\|_1 . 
\]

Lemma 2

Then Lemma 2 combined with the Neumann series representation of matrix inverse implies that \( \|(I - \gamma \cdot P_\beta^T)^{-1}\|_1 \leq 1/(1 - \gamma) \). It can also be shown readily by basic algebra that \( \|e_\beta - d_i\|_1 \leq \epsilon_1 \). The desired result then follows because \( \beta \in \Delta^S \).

**Proof of Theorem 2.** Before proving the result, we recall several necessary definitions and identities. The static and dynamic returns are defined as

\[ \rho^S(\pi) = \min_{\xi \in \Xi} \mathbb{E}_{\rho \sim \xi} \left[ \rho(\pi, \hat{P}) \right] \quad (30a) \]

\[ \rho^D(\pi) = \min_{\xi \in \Xi} \rho \left( \pi, \mathbb{E}_{\rho \sim \xi} \left[ \hat{P} \right] \right) . \quad (30b) \]

Recall also that the return of a policy \( \pi \) in an MDP with the transition matrix \( P_\pi \in (\Delta^S)^S \) can be expressed in terms of the occupancy frequency \( d_\pi \), defined in (17), as

\[ \rho(\pi, P) = p_0^T v_\pi = p_0^T (I - \gamma \cdot P_\pi)^{-1} r_\pi = d_\pi^T r_\pi . \]

(31)

Now, let \( \xi^S \in \Delta^{[S]} \) and \( \xi^D \in \Delta^{[S]} \) be optimal in (30a) and (30b). Then the soft-robust returns in (30) can be expressed in terms of their occupancy frequencies using (31) as

\[ \rho^S(\pi) = \mathbb{E}_{\rho \sim \xi^S} \left[ \rho(\pi, \hat{P}) \right] = \sum_{\omega \in \Omega} \xi^S_\omega \cdot p_0^T (I - \gamma \cdot \hat{P}_\omega)^{-1} r_\pi = \sum_{\omega \in \Omega} \xi^S_\omega \cdot (d^\omega_\pi)^T r_\pi \]

\[ \rho^D(\pi) = \rho \left( \pi, \mathbb{E}_{\rho \sim \xi^D} \left[ \hat{P} \right] \right) = p_0^T \left( I - \gamma \sum_{\omega \in \Omega} \xi^D_\omega \cdot \hat{P}_\omega \right)^{-1} r_\pi = (d^D_\pi)^T r_\pi . \]

(32)
Next, we get for each $\pi \in \Pi$ that
\[
\rho^D(\pi) - \rho^S(\pi) \leq \mathbb{E}_{\rho \sim \xi^D} \left[ \rho(\pi, \hat{P}) \right] - \rho \left( \pi, \mathbb{E}_{\rho \sim \xi^S} \left[ \hat{P} \right] \right) = (d^S_\pi)^\top r_\pi - \sum_{\omega \in \Omega} \xi^S_\omega \cdot (d^S_\pi)^\top r_\pi
\]
From (32) and $\xi^S \in \Xi$
\[
\leq \left\| d^S_\pi - \sum_{\omega \in \Omega} \xi^S_\omega \cdot d^S_\pi \right\|_1 \cdot \| r_\pi \|_\infty \quad \text{Holder's inequality}
\]
\[
\leq \frac{\gamma \cdot \epsilon_1}{1 - \gamma} \cdot r_{\max}
\]
From Lemma 4.

Similarly, the reverse inequality follows as
\[
\rho^S(\pi) - \rho^D(\pi) \leq \mathbb{E}_{\rho \sim \xi^D} \left[ \rho(\pi, \hat{P}) \right] - \rho \left( \pi, \mathbb{E}_{\rho \sim \xi^D} \left[ \hat{P} \right] \right) = \sum_{\omega \in \Omega} \xi^D_\omega \cdot (d^D_\pi)^\top r_\pi - (d^D_\pi)^\top r_\pi
\]
From (32) and $\xi^D \in \Xi$
\[
\leq \left\| \sum_{\omega \in \Omega} \xi^D_\omega \cdot d^D_\pi - d^D_\pi \right\|_1 \cdot \| r_\pi \|_\infty \quad \text{Holder's inequality}
\]
\[
\leq \frac{\gamma \cdot \epsilon_1(\pi)}{1 - \gamma} \cdot r_{\max}
\]
From Lemma 4.

Combining the two inequalities above, we obtain that
\[
|\rho^S(\pi) - \rho^D(\pi)| \leq \frac{\gamma \cdot \epsilon_1(\pi)}{1 - \gamma} \cdot r_{\max}
\]
which proves the result.

Proof of Theorem 3. To establish this bound, define a robust Bellman value operator $\mathfrak{T}^{\pi, \xi} : \mathbb{R}^S \to \mathbb{R}^S$ for any policy $\pi \in \Pi$, nature’s response $\xi \in \Xi$, value function $v \in \mathbb{R}^S$, and state $s \in S$ as
\[
(\mathfrak{T}^{\pi, \xi} v)_s = \sum_{a \in \mathcal{A}} \sum_{\omega \in \Omega} \xi_\omega \cdot s, a \cdot (\hat{P}^\omega_{s,a})^\top (r_{s,a} + \gamma \cdot v).
\]
The operator $\mathfrak{T}^{\pi, \xi}$ is linear and has a unique fixed point $v^{\pi, \xi} \in \mathbb{R}^S$ which satisfies $\mathfrak{T}^{\pi, \xi} v^{\pi, \xi} = v^{\pi, \xi}$ [17]. Similarly, we define a robust S-rectangular Bellman value operator $\mathfrak{T}^\pi : \mathbb{R}^S \to \mathbb{R}^S$ defined for any policy $\pi \in \Pi$, value function $v \in \mathbb{R}^S$, and state $s \in S$ as
\[
(\mathfrak{T}^\pi v)_s = \min_{\xi \in \Xi} \left( \mathfrak{T}^{\pi, \xi} v \right)_s.
\]
Note that for a fixed policy $\pi \in \Pi$, the operator $\mathfrak{T}^\pi$ is equivalent to the Bellman operator in MDPs and satisfies the same properties.

Equipped with the definitions above, we proceed to bound the error $\rho^D(\pi^*_D) - \rho^R(\pi^*_D)$. Let $\xi^*_D$ be the minimizer for $\rho^D(\pi^*_D)$ in (7) and therefore
\[
\rho^D(\pi^*_D) = p_0 v^{\pi^*_D, \xi^*_D}.
\]
Similarly, let $\xi^*_R$ be the minimizer to $\rho^R(\pi^*_D)$ in (12) and therefore
\[
\rho^R(\pi^*_D) = p_0 v^{\pi^*_D, \xi^*_R}.
\]
Exploiting the fact that $\mathfrak{T}^\pi$ is an MDP Bellman operator and using standard arguments for MDP value functions (for example, Corollary 4 in [17]) we get that:
\[
\rho^D(\pi^*_D) - \rho^R(\pi^*_D) = p_0 v^{\pi^*_D, \xi^*_D} - p_0 v^{\pi^*_D, \xi^*_R} \leq \left\| p_0 \right\|_1 \cdot \left\| v^{\pi^*_D, \xi^*_D} - v^{\pi^*_D, \xi^*_R} \right\|_\infty = \left\| v^{\pi^*_D, \xi^*_D} - v^{\pi^*_D, \xi^*_R} \right\|_\infty
\]
\[
\leq \frac{1}{1 - \gamma} \cdot \left\| \mathfrak{T}^{\pi^*_D} v^{\pi^*_D, \xi^*_D} - v^{\pi^*_D, \xi^*_R} \right\|_\infty \leq \frac{1}{1 - \gamma} \cdot \epsilon_2,
\]
for the $\epsilon_2$ stated in the theorem. Finally, we employ Lemma 3 combined with Theorem 1 to show that

$$0 \leq \rho^D(\pi_D) - \rho^R(\pi_R) \leq \rho^D(\pi_D^*) - \rho^R(\pi_R^*) \leq \frac{1}{1 - \gamma} \cdot \epsilon_2,$$

which shows the desired result.

**Proof of Corollary 1.** The result follows by algebraic manipulation as

$$\rho^S(\pi_S^*) - \rho^S(\pi_R^*) = \rho^S(\pi_S) - \rho^D(\pi_D) + \rho^D(\pi_D) - \rho^D(\pi_R^*) + \rho^D(\pi_R^*) - \rho^S(\pi_S) = 0$$

Lemma 3 & Theorem 2

$$\leq \frac{2\gamma \cdot r_{max} \cdot \epsilon_1}{1 - \gamma} + \rho^D(\pi_D) - \rho^D(\pi_R^*)$$

Theorem 2

$$= \frac{2\gamma \cdot r_{max} \cdot \epsilon_1}{1 - \gamma} + \rho^D(\pi_D) - \rho^D(\pi_R^*)$$

Theorem 3

$$\leq \frac{2\gamma \cdot r_{max} \cdot \epsilon_1}{1 - \gamma} + \frac{\epsilon_2}{1 - \gamma}.$$