Sample Complexity of Robust Reinforcement Learning with a Generative Model

Kishan Panaganti
Texas A&M University

Dileep Kalathil
Texas A&M University

Abstract

The Robust Markov Decision Process (RMDP) framework focuses on designing control policies that are robust against the parameter uncertainties due to the mismatches between the simulator model and real-world settings. An RMDP problem is typically formulated as a max-min problem, where the objective is to find the policy that maximizes the value function for the worst possible model that lies in an uncertainty set around a nominal model. The standard robust dynamic programming approach requires the knowledge of the nominal model for computing the optimal robust policy. In this work, we propose a model-based reinforcement learning (RL) algorithm for learning an ε-optimal robust policy when the nominal model is unknown. We consider three different forms of uncertainty sets, characterized by the total variation distance, chi-square divergence, and KL divergence. For each of these uncertainty sets, we give a precise characterization of the sample complexity of our proposed algorithm. In addition to the sample complexity results, we also present a formal analytical argument on the benefit of using robust policies. Finally, we demonstrate the performance of our algorithm on two benchmark problems.

1 Introduction

Reinforcement Learning (RL) algorithms typically require a large number of data samples to learn a control policy, which makes the training of RL algorithms directly on the real-world systems expensive and potentially dangerous. This problem is typically avoided by training the RL algorithm on a simulator and transferring the trained policy to the real-world system. However, due to multiple reasons such as the approximation errors incurred while modeling, changes in the real-world parameters over time and possible adversarial disturbances in the real-world, there will be inevitable mismatches between the simulator model and the real-world system. For example, the standard simulator settings of the sensor noise, action delays, friction, and mass of a mobile robot can be different from that of the actual real-world robot. This mismatch between the simulator and real-world model parameters, often called 'simulation-to-reality gap', can significantly degrade the real-world performance of the RL algorithms trained on a simulator model.

Robust Markov Decision Process (RMDP) addresses the *planning* problem of computing the optimal policy that is robust against the parameter mismatch between the simulator and real-world system. The RMDP framework was first introduced in (Iyengar, 2005; Nilim and El Ghaoui, 2005). The RMDP problem has been analyzed extensively in the literature (Xu and Mannor, 2010; Wiesemann et al., 2013; Yu and Xu, 2015; Mannor et al., 2016; Russel and Petrik, 2019), considering different types of uncertainty set and computationally efficient algorithms. However, these works are limited to the planning problem, which assumes the knowledge of the system. Robust RL algorithms for learning the optimal robust policy have also been proposed (Roy et al., 2017; Panaganti and Kalathil, 2021), but they only provide asymptotic convergence guarantees. Robust RL problem has also been addressed using deep RL methods (Pinto et al., 2017; Derman et al., 2018, 2020; Mankowitz et al., 2020; Zhang et al., 2020a). However, these works are empirical in nature and do not provide any theoretical guarantees for the learned policies. In particular, there are few works that provide *robust RL algorithms with provable (non-asymptotic) finite-sample performance guarantees.*

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

In this work, we address the problem of developing a model-based robust RL algorithm with provable finite-sample guarantees on its performance, characterized by the metric of sample complexity in a PAC (probably approximately correct) sense. The RMDP framework assumes that the real-world model lies within some uncertainty set $\mathcal{P}$ around a nominal (simulator) model P^o. The goal is to learn a policy that performs the best under the worst possible model in this uncertainty set. We do not assume that the algorithm knows the exact simulator model (and hence the exact uncertainty set). Instead, similar to the standard (non-robust) RL setting (Singh and Yee, 1994; Azar et al., 2013; Haskell et al., 2016; Sidford et al., 2018; Agarwal et al., 2020; Li et al., 2020; Kalathil et al., 2021), we assume that the algorithm has access to a generative sampling model that can generate next-state samples for all state-action pairs according to the nominal simulator model. In this context, we answer the following important question: *How many samples from the nominal simulator model do we need to learn an ε-optimal robust policy with high probability?*

Our contributions: The main contributions of our work are as follows:

(1) We propose a model-based robust RL algorithm, which we call the robust empirical value iteration algorithm (REVI), for learning an approximately optimal robust policy. We consider three different classes of RMDPs with three different uncertainty sets: (i) Total Variation (TV) uncertainty set, (ii) Chi-square uncertainty set, and (iii) Kullback-Leibler (KL) uncertainty set, each characterized by its namesake distance measure. Robust RL problem is much more challenging than the standard (non-robust) RL problems due to the inherent nonlinearity associated with the robust dynamic programming and the resulting unbiasedness of the empirical estimates. We overcome this challenge analytically by developing a series of upperbounds that are amenable to using concentration inequality results (which are typically useful only in the unbiased setting), where we exploit a uniform concentration bound with a covering number argument. We rigorously characterize the sample complexity of the proposed algorithm for each of these uncertainty sets. We also make a precise comparison with the sample complexity of non-robust RL.

(2) We give a formal argument for the need for using a robust policy when the simulator model is different from the real-world model. More precisely, we analytically address the question *'why do we need robust policies?'*, by showing that the worst case performance of a non-robust policy can be arbitrarily bad (as bad as a random policy) when compared to that of a robust policy. While the need for robust policies have been discussed in the literature qualitatively, to the best of our knowledge, this is the first work that gives an analytical answer to the above question.

(3) Finally, we demonstrate the performance of our REVI algorithm in two experiment settings and for two different uncertainty sets. In each setting, we show that the policy learned by our proposed REVI algorithm is indeed robust against the changes in the model parameters. We also illustrate the convergence of our algorithm with respect to the number of samples and the number of iterations.

1.1 Related Work

Robust RL: An RMDP setting where some state-action pairs are adversarial and the others are stationary was considered by (Lim et al., 2013), who proposed an online algorithm to address this problem. An approximate robust dynamic programming approach with linear function approximation was proposed in (Tamar et al., 2014). State aggregation and kernel-based function approximation for robust RL were studied in (Petrik and Subramanian, 2014; Lim and Autef, 2019). (Roy et al., 2017) proposed a robust version of the Q-learning algorithm. (Panaganti and Kalathil, 2021) developed a least squares policy iteration approach to learn the optimal robust policy using linear function approximation with provable guarantees. A soft robust RL algorithm was proposed in (Derman et al., 2018) and a maximum a posteriori policy optimization approach was used in (Mankowitz et al., 2020). While the above mentioned works make interesting contributions to the area of robust RL, they focus either on giving asymptotic performance guarantees or on the empirical performance without giving provable guarantees. In particular, they do not provide provable guarantees on the finite-sample performance of the robust RL algorithms.

The closest to our work is (Zhou et al., 2021), which analyzed the sample complexity with a KL uncertainty set. Our work is different in two significant aspects: Firstly, we consider the total variation uncertainty set and chi-square uncertainty set, in addition to the KL uncertainty set. The analysis for these uncertainty sets are very challenging and significantly different from that of the KL uncertainty set. Secondly, we give a more precise characterization of the sample complexity bound for the KL uncertainty set by clearly specifying the exponential dependence on $(1-\gamma)^{-1}$, where γ is the discount factor, which was left unspecified in (Zhou et al., 2021).

While this paper was under review, we were notified of a concurrent work (Yang et al., 2021), which also provides similar sample complexity bounds for robust

Table 1: Comparison of the sample complexities of different uncertainty sets and the best known result in the non-robust setting (Li et al., 2020). Here $|\mathcal{S}|$ and $|\mathcal{A}|$ are the cardinality of the state and action spaces, c_r is the robust RL problem parameter, and γ is the discount factor. We consider the optimality gap $\varepsilon \in (0, c/(1-\gamma))$, where $c > 0$ is a constant. We refer to Section 3.2 for further details.

Uncertainty set	TV	Chi-square	KL	Non-robust
Sample Complexity	$\mathcal{O}(\frac{\|\mathcal{S}\|^2\|\mathcal{A}\|}{(1-\gamma)^4\varepsilon^2})$	$\mathcal{O}(\frac{c_r\|\mathcal{S}\|^2\|\mathcal{A}\|}{(1-\gamma)^4\varepsilon^2})$	$\mathcal{O}(\frac{\|\mathcal{S}\|^2\|\mathcal{A}\|\exp(1/(1-\gamma))}{c_r^2(1-\gamma)^4\varepsilon^2})$	$\mathcal{O}(\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^3\varepsilon^2})$

RL. Our proof technique is significantly different from their work. Moreover, we also provide open-source experimental results that illustrate the performance of our robust RL algorithm.

Other related works: Robust control is a well-studied area (Zhou et al., 1996; Dullerud and Paganini, 2013) in the classical control theory. Recently, there are some interesting works that address the robust RL problem using this framework, especially focusing on the linear quadratic regulator setting (Zhang et al., 2020b). Our framework of robust MDP is significantly different from this line of work. Risk sensitive RL algorithms (Borkar, 2002) and adversarial RL algorithms (Pinto et al., 2017) also address the robustness problem implicitly. Our approach is different from these works also.

Notations: For any set $\mathcal{X}$, $|\mathcal{X}|$ denotes its cardinality. For any vector x, $\|x\|$ denotes its infinity norm $\|x\|_\infty$.

2 Preliminaries: Robust Markov Decision Process

A Markov Decision Process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, r, P, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. The transition probability function $P_{s,a}(s')$ represents the probability of transitioning to state s' when action a is taken at state s. P is also called the the model of the system. We consider a finite MDP setting where $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite. We will also assume that $r(s, a) \in [0, 1]$, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, without loss of generality.

A (deterministic) policy π maps each state to an action. The value of a policy π for an MDP with model P, evaluated at state s is given by

$$V_{\pi,P}(s) = \mathbb{E}_{\pi,P}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s], \qquad (1)$$

where $a_t = \pi(s_t), s_{t+1} \sim P_{s_t,a_t}(\cdot)$. The optimal value function V_P^* and the optimal policy π_P^* of an MDP with the model P are defined as

$$V_P^* = \max_\pi V_{\pi,P}, \quad \pi_P^* = \arg\max_\pi V_{\pi,P}. \qquad (2)$$

Uncertainty set: Unlike the standard MDP which considers a single model (transition probability function), the RMDP formulation considers a set of models. We call this set as the *uncertainty set* and denote it as $\mathcal{P}$. We assume that the set $\mathcal{P}$ satisfies the standard *rectangularity condition* (Iyengar, 2005). We note that a similar uncertainty set can be considered for the reward function at the expense of additional notations. However, since the analysis will be similar and the samples complexity guarantee will be identical upto a constant, without loss of generality, we assume that the reward function is known and deterministic. We specify a robust MDP as a tuple $M = (\mathcal{S}, \mathcal{A}, r, \mathcal{P}, \gamma)$.

The uncertainty set $\mathcal{P}$ is typically defined as

$$\mathcal{P} = \otimes \mathcal{P}_{s,a}, \text{ where, } \mathcal{P}_{s,a} = \{P_{s,a} \in [0,1]^{|\mathcal{S}|} : D(P_{s,a}, P^o_{s,a}) \le c_r, \sum_{s' \in \mathcal{S}} P_{s,a}(s') = 1\}, \qquad (3)$$

where $P^o = (P^o_{s,a}, (s, a) \in \mathcal{S} \times \mathcal{A})$ is the nominal transition probability function, $c_r > 0$ indicates the level of robustness, and $D(\cdot, \cdot)$ is a distance metric between two probability distributions. In the following, we call P^o as the nominal model. In other words, $\mathcal{P}$ is the set of all valid transition probability functions in the neighborhood of the nominal model P^o, where the neighborhood is defined using the distance metric $D(\cdot, \cdot)$. We note that the radius c_r can depend on the state-action pair (s, a). We omit this to reduce the notation complexity. We also note that for $c_r \downarrow 0$, we recover the non-robust regime.

We consider three different uncertainty sets corresponding to three different distance metrics $D(\cdot, \cdot)$.

1. *Total Variation (TV) uncertainty set ($\mathcal{P}^{\text{tv}}$)*: We define $\mathcal{P}^{\text{tv}} = \otimes \mathcal{P}^{\text{tv}}_{s,a}$, where $\mathcal{P}^{\text{tv}}_{s,a}$ is defined as in (3) using the total variation distance

$$D_{\text{tv}}(P_{s,a}, P^o_{s,a}) = (1/2)\|P_{s,a} - P^o_{s,a}\|_1. \qquad (4)$$

2. *Chi-square uncertainty set ($\mathcal{P}^{\text{c}}$)*: We define $\mathcal{P}^{\text{c}} = \otimes \mathcal{P}^{\text{c}}_{s,a}$, where $\mathcal{P}^{\text{c}}_{s,a}$ is defined as in (3) using the Chi-square distance

$$D_{\text{c}}(P_{s,a}, P^o_{s,a}) = \sum_{s' \in \mathcal{S}} \frac{(P_{s,a}(s') - P^o_{s,a}(s'))^2}{P^o_{s,a}(s')}. \qquad (5)$$

3. *Kullback-Leibler (KL) uncertainty set ($\mathcal{P}^{\text{kl}}$)*: We define $\mathcal{P}^{\text{kl}} = \otimes \mathcal{P}^{\text{kl}}_{s,a}$, where $\mathcal{P}^{\text{kl}}_{s,a}$ is defined as in (3) using the Kullback-Leibler (KL) distance

$$D_{\text{kl}}(P_{s,a}, P^o_{s,a}) = \sum_{s'} P_{s,a}(s') \log \frac{P_{s,a}(s')}{P^o_{s,a}(s')}. \quad (6)$$

We note that the sample complexity and its analysis will depend on the specific form of the uncertainty set.

Robust value iteration: The goal of the RMDP problem is to compute the optimal robust policy which maximizes the value even under the worst model in the uncertainty set. Formally, the *robust value function* V_π corresponding to a policy π and the *optimal robust value function* V^* are defined as (Iyengar, 2005; Nilim and El Ghaoui, 2005)

$$V^\pi = \inf_{P \in \mathcal{P}} V_{\pi,P}, \quad V^* = \sup_\pi \inf_{P \in \mathcal{P}} V_{\pi,P}. \quad (7)$$

The *optimal robust policy* π^* is such that the robust value function corresponding to it matches the optimal robust value function, i.e., $V^{\pi^*} = V^*$. It is known that there exists a deterministic optimal policy (Iyengar, 2005) for the RMDP problem. So, we will restrict our attention to the class of deterministic policies.

For any set $\mathcal{B}$ and a vector v, let

$$\sigma_{\mathcal{B}}(v) = \inf\{u^\top v : u \in \mathcal{B}\}.$$

Using this notation, we can define the *robust Bellman operator* (Iyengar, 2005) as $T(V)(s) = \max_a (r(s,a) + \gamma \sigma_{\mathcal{P}_{s,a}}(V))$. It is known that T is a contraction mapping in infinity norm and the V^* is the unique fixed point of T (Iyengar, 2005). Since T is a contraction, *robust value iteration* can be used to compute V^*, similar to the non-robust MDP setting (Iyengar, 2005). More precisely, the robust value iteration, defined as $V_{k+1} = TV_k$, converges to V^*, i.e., $V_k \to V^*$. Similar to the optimal robust value function, we can also define the optimal robust action-value function as $Q^*(s,a) = r(s,a) + \gamma \sigma_{\mathcal{P}_{s,a}}(V^*)$. Similar to the non-robust setting, it is straight forward to show that $\pi^*(s) = \arg\max_a Q^*(s,a)$ and $V^*(s) = \max_a Q^*(s,a)$.

3 Algorithm and Sample Complexity

The robust value iteration requires the knowledge of the nominal model P^o and the radius of the uncertainty set c_r to compute V^* and π^*. While c_r may be available as design parameter, the form of the nominal model may not be available in most practical problems. So, we do not assume the knowledge of the nominal model P^o. Instead, similar to the non-robust RL setting, we assume only to have access to the samples from a generative model, which can generate samples of the next state s' according to $P^o_{s,a}(\cdot)$, given the state-action pair (s,a) as the input. We propose a model-based robust RL algorithm that uses these samples to estimate the nominal model and uncertainty set.

Algorithm 1 Robust Empirical Value Iteration (REVI) Algorithm

Input: Loop termination number K
Initialize: $Q_0 = 0$
Compute the empirical uncertainty set $\widehat{\mathcal{P}}$ according to (8)
for $k = 0, \cdots, K-1$ **do**
  $V_k(s) = \max_a Q_k(s,a), \ \forall s$
  $Q_{k+1}(s,a) = r(s,a) + \gamma \sigma_{\widehat{\mathcal{P}}_{s,a}}(V_k), \ \forall (s,a)$
end for
Output: $\pi_K(s) = \arg\max_a Q_K(s,a), \forall s \in \mathcal{S}$

3.1 Robust Empirical Value Iteration (REVI) Algorithm

We first get a maximum likelihood estimate $\widehat{P}^o$ of the nominal model P^o by following the standard approach (Azar et al., 2013, Algorithm 3). More precisely, we generate N next-state samples corresponding to each state-action pairs. Then, the maximum likelihood estimate $\widehat{P}^o$ is given by $\widehat{P}^o_{s,a}(s') = N(s,a,s')/N$, where $N(s,a,s')$ is the number of times the state s' is realized out of the total N transitions from the state-action pair (s,a). Given $\widehat{P}^o$, we can get an empirical estimate $\widehat{\mathcal{P}}$ of the uncertainty set $\mathcal{P}$ as,

$$\widehat{\mathcal{P}} = \otimes \widehat{\mathcal{P}}_{s,a}, \text{ where, } \widehat{\mathcal{P}}_{s,a} = \{P \in [0,1]^{\mathcal{S}} : D(P_{s,a}, \widehat{P}_{s,a}) \le c_r, \sum_{s' \in \mathcal{S}} P_{s,a}(s') = 1\}, \quad (8)$$

where D is one of the metrics specified in (4) - (6).

For finding an approximately optimal robust policy, we now consider the empirical RMDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, r, \widehat{\mathcal{P}}, \gamma)$ and perform robust value iteration using $\widehat{\mathcal{P}}$. This is indeed our approach, which we call the Robust Empirical Value Iteration (REVI) Algorithm. The optimal robust policy and value function of $\widehat{M}$ are denoted as $\widehat{\pi}^\star, \widehat{V}^\star$, respectively.

3.2 Sample Complexity

In this section we give the sample complexity guarantee of the REVI algorithm for the three uncertainty sets. We first consider the TV uncertainty set.

Theorem 1 (TV Uncertainty Set)**.** *Consider an RMDP with a total variation uncertainty set $\mathcal{P}^{\text{tv}}$. Fix $\delta \in (0,1)$ and $\varepsilon \in (0, 24\gamma/(1-\gamma))$. Consider the REVI*

algorithm with $K \geq K_0$ *and* $N \geq N^{\text{tv}}$, *where*

$$K_0 = \frac{1}{\log(1/\gamma)} \log(\frac{8\gamma}{\varepsilon(1-\gamma)^2}) \text{ and} \tag{9}$$

$$N^{\text{tv}} = \frac{72\gamma^2|\mathcal{S}|}{(1-\gamma)^4\varepsilon^2} \log(\frac{144\gamma|\mathcal{S}||\mathcal{A}|}{(\delta\varepsilon(1-\gamma)^2)}). \tag{10}$$

Then, $\|V^* - V^{\pi_K}\| \leq \varepsilon$ *with probability at least* $1-2\delta$.

Remark 1. The total number of samples needed in the REVI algorithm is $N_{\text{total}} = N|\mathcal{S}||\mathcal{A}|$. So the sample complexity of the REVI algorithm with the TV uncertainty set is $\mathcal{O}(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^4\varepsilon^2})$.

Remark 2 (Comparison with the sample complexity of the non-robust RL). For the non-robust setting, the lowerbound for the total number of samples from the generative sampling device is $\Omega(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta})$ (Azar et al., 2013, Theorem 3). The variance reduced value iteration algorithm proposed in (Sidford et al., 2018) achieves a sample complexity of $\mathcal{O}(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta\varepsilon})$, matching the lower bound. However, this work is restricted to $\varepsilon \in (0,1)$, whereas ε can be considered upto the value $1/(1-\gamma)$ for the MDP problems. Recently, this result has been further improved recently by (Agarwal et al., 2020) and (Li et al., 2020), which considered $\varepsilon \in (0, 1/\sqrt{(1-\gamma)})$ and $\varepsilon \in (0, 1/(1-\gamma))$, respectively.

Theorem 1 for the robust RL setting also considers ε upto $\mathcal{O}(1/(1-\gamma))$. However, the sample complexity obtained is worse by a factor of $|\mathcal{S}|$ and $1/(1-\gamma)$ when compared to the non-robust setting. These additional terms are appearing in our result due to a covering number argument we used in the proof, which seems necessary for getting a tractable bound. However, it is not clear if this is fundamental to the robust RL problem with TV uncertainty set. We leave this investigation for our future work.

We next consider the chi-square uncertainty set.

Theorem 2 (Chi-square Uncertainty Set). *Consider an RMDP with a Chi-square uncertainty set* $\mathcal{P}^{\text{c}}$. *Fix* $\delta \in (0,1)$ *and* $\varepsilon \in (0, 16\gamma/(1-\gamma))$, *for an absolute constant* $c_1 > 1$. *Consider the REVI algorithm with* $K \geq K_0$ *and* $N \geq N^{\text{c}}$, *where* K_0 *is as given in* (9) *and*

$$N^{\text{c}} = \frac{64\gamma^2(2c_r+1)|\mathcal{S}|}{(1-\gamma)^4\varepsilon^2} \log(\frac{192|\mathcal{S}||\mathcal{A}|\gamma}{(\delta\varepsilon(1-\gamma)^2)}). \tag{11}$$

Then, $\|V^* - V^{\pi_K}\| \leq \varepsilon$ *with probability at least* $1-2\delta$.

Remark 3. The sample complexity of the algorithm with the chi-square uncertainty set is $\mathcal{O}(\frac{|\mathcal{S}|^2|\mathcal{A}|c_r}{(1-\gamma)^4\varepsilon^2})$. The order of sample complexity remains the same compared to that of the TV uncertainty set given in Theorem 1.

Finally, we consider the KL uncertainty set.

Theorem 3 (KL Uncertainty Set). *Consider an RMDP with a KL uncertainty set* $\mathcal{P}^{\text{kl}}$. *Fix* $\delta \in (0,1)$ *and* $\varepsilon \in (0, 1/(1-\gamma))$. *Consider the REVI algorithm with* $K \geq K_0$ *and* $N \geq N^{\text{kl}}$, *where* K_0 *is as in* (9) *and*

$$N^{\text{kl}} = \frac{8\gamma^2|\mathcal{S}|}{c_r^2(1-\gamma)^4\varepsilon^2} \exp(\frac{2\lambda_{\text{kl}}+4}{\lambda_{\text{kl}}(1-\gamma)}) \log(\frac{9|\mathcal{S}||\mathcal{A}|}{\delta\lambda_{\text{kl}}(1-\gamma)}), \tag{12}$$

and λ_{kl} *is a problem dependent parameter but independent of* N^{kl}. *Then,* $\|V^* - V^{\pi_K}\| \leq \varepsilon$ *with probability at least* $1-2\delta$.

Remark 4. The sample complexity with the KL uncertainty set is $\mathcal{O}(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^4\varepsilon^2 c_r^2} \exp(\frac{1}{(1-\gamma)}))$. We note that (Zhou et al., 2021) also considered the robust RL problem with KL uncertainty set. They provided a sample complexity bound of the form $\mathcal{O}(\frac{C|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^4\varepsilon^2 c_r^2})$, However the exponential dependence on $1/(1-\gamma)$ was hidden inside the constant C. In this work, we clearly specify the depends on the factor $1/(1-\gamma)$.

4 Why Do We Need Robust Policies?

In the introduction, we have given a qualitative description about the need for finding a robust policy. In this section, we give a formal argument to show that the worst case performance of a non-robust policy can be arbitrarily bad (as bad as a random policy) when compared to that of a robust policy.

We consider a simple setting with an uncertainty set that contains only two models, i.e., $\mathcal{P} = \{P^o, P'\}$. Let π^* be the optimal robust policy. Following the notation in (2), let $\pi^o = \pi_{P^o}$ and $\pi' = \pi_{P'}$ be the non-robust optimal policies when the model is P^o and P', respectively. Assume that nominal model is P^o and we decide to employ the non-robust policy π^o. The worst case performance of π^o is characterized by its robust value function V^{π^o} which is $\min\{V_{\pi^o,P^o}, V_{\pi^o,P'}\}$.

We now state the following result.

Theorem 4 (Robustness Gap). *There exists a robust MDP* M *with uncertainty set* $\mathcal{P} = \{P^o, P'\}$, *discount factor* $\gamma \in (\gamma_o, 1]$, *and state* $s_1 \in \mathcal{S}$ *such that*

$$V^{\pi^o}(s_1) \leq V^{\pi^*}(s_1) - c/(1-\gamma),$$

where c *is a positive constant,* π^* *is the optimal robust policy, and* $\pi^o = \pi_{P^o}$ *is the non-robust optimal policy when the model is* P^o.

Theorem 4 states that the worst case performance of the non-robust policy π^o is lower than that of the optimal robust policy π^*, and this performance gap is

$\Omega(1/(1-\gamma))$. Since $|r(s,a)| \le 1, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$ by assumption, $\|V_{\pi,P}\| \le 1/(1-\gamma)$ for any policy π and any model P. Therefore, the difference between the optimal (robust) value function and the (robust) value function of an arbitrary policy cannot be greater than $\mathcal{O}(1/(1-\gamma))$. Thus the worst-case performance of the non-robust policy π^o can be as bad as an arbitrary policy in an order sense.

5 Sample Complexity Analysis

In this section we explain the key ideas used in the analysis of the REVI algorithm for obtaining the sample complexity bound for each of the uncertainty sets. Recall that we consider an RMDP M and its empirical estimate version as $\widehat{M}$.

To bound $\|V^* - V^{\pi_K}\|$, we split it into three terms as $\|V^* - V^{\pi_K}\| \le \|V^* - \widehat{V}^*\| + \|\widehat{V}^* - \widehat{V}^{\pi_K}\| + \|\widehat{V}^{\pi_K} - V^{\pi_K}\|$, and analyze each term separately.

Analyzing the second term, $\|\widehat{V}^* - \widehat{V}^{\pi_K}\|$, is similar to that of non-robust algorithms. Due to the contraction property of the robust Bellman operator, it is straight forward to show that $\|\widehat{V}^* - \widehat{V}^{\pi_{k+1}}\| \le \gamma \|\widehat{V}^* - \widehat{V}^{\pi_k}\|$ for any k. This exponential convergence, with some additional results from the MDP theory, enables us to get a bound $\|\widehat{V}^* - \widehat{V}^{\pi_K}\| \le 2\gamma^{K+1}/(1-\gamma)^2$.

The analysis of terms $\|V^* - \widehat{V}^*\|$ and $\|\widehat{V}^{\pi_K} - V^{\pi_K}\|$ are however non-trivial and significantly more challenging compared to the non-robust setting. We will focus on the latter, and the analysis of the former is similar.

For any policy π and for any state s, and denoting $a = \pi(s)$, we have

$$V^\pi(s) - \widehat{V}^\pi(s) = \gamma \sigma_{\mathcal{P}_{s,a}}(V^\pi) - \gamma \sigma_{\widehat{\mathcal{P}}_{s,a}}(\widehat{V}^\pi)$$
$$= \gamma(\sigma_{\mathcal{P}_{s,a}}(V^\pi) - \sigma_{\mathcal{P}_{s,a}}(\widehat{V}^\pi)) + \gamma(\sigma_{\mathcal{P}_{s,a}}(\widehat{V}^\pi) - \sigma_{\widehat{\mathcal{P}}_{s,a}}(\widehat{V}^\pi)) \tag{13}$$

To bound the first term in (13), we present a result that shows that $\sigma_{\mathcal{P}_{s,a}}$ is 1-Lipschitz in the sup-norm.

Lemma 1. *For any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and for any $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$, we have $|\sigma_{\mathcal{P}_{s,a}}(V_1) - \sigma_{\mathcal{P}_{s,a}}(V_2)| \le \|V_1 - V_2\|$ and $|\sigma_{\widehat{\mathcal{P}}_{s,a}}(V_1) - \sigma_{\widehat{\mathcal{P}}_{s,a}}(V_2)| \le \|V_1 - V_2\|$.*

Using the above lemma, the first term in (13) will be bounded by $\gamma\|V^\pi - \widehat{V}^\pi\|$ and the discount factor makes this term amenable to getting a closed form bound.

Obtaining a bound for $\sigma_{\mathcal{P}_{s,a}}(\widehat{V}^\pi) - \sigma_{\widehat{\mathcal{P}}_{s,a}}(\widehat{V}^\pi)$ is the most challenging part of our analysis. In the non-robust setting, this will be equivalent to the error term $P^o_{s,a}V - \widehat{P}_{s,a}V$, which is unbiased and can be easily bounded using concentration inequalities. In the robust setting, however, because of the nonlinear nature of the function $\sigma(\cdot)$, $\mathbb{E}[\sigma_{\widehat{\mathcal{P}}_{s,a}}(\widehat{V}^\pi)] \ne \sigma_{\mathcal{P}_{s,a}}(\widehat{V}^\pi)$. So, using concentration inequalities to get a bound is not immediate. Our strategy is to find appropriate upper-bound for this term that is amenable to using concentration inequalities. To that end, we will analyze this term separately for each of the three uncertainty set.

5.1 Total variation uncertainty set

We will first get following upperbound:

Lemma 2 (TV uncertainty set)**.** *Let $\mathcal{V} = \{V \in \mathbb{R}^{|\mathcal{S}|} : \|V\| \le 1/(1-\gamma)\}$. For any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and for any $V \in \mathcal{V}$,*

$$|\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)| \le 2 \max_{\mu\in\mathcal{V}} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu|. \tag{14}$$

While the term $|\widehat{P}_{s,a}\mu - P^o_{s,a}\mu|$ in (14) can be upperbounded using the standard Hoeffding's inequality, bounding $\max_{\mu\in\mathcal{V}} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu|$ is more challenging as it requires a uniform bound. Since μ can take a continuum of values, a simple union bound argument will also not work. We overcome this issue by using a covering number argument and obtain the following bound.

Lemma 3. *Let $V \in \mathbb{R}^{|\mathcal{S}|}$ with $\|V\| \le 1/(1-\gamma)$. For any $\eta, \delta \in (0,1)$,*

$$\max_{\mu:0\le\mu\le V} \max_{s,a} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu| \le \frac{1}{1-\gamma}\sqrt{\frac{|\mathcal{S}|}{2N}\log\Big(\frac{12|\mathcal{S}||\mathcal{A}|}{(\delta\eta(1-\gamma))}\Big)} + 2\eta,$$

with probability at least $1-\delta/2$.

We note that this uniform bound adds an additional $\sqrt{|\mathcal{S}|}$ factor compared to the non-robust setting, which results in an additional $|\mathcal{S}|$ in the sample complexity. Combining these, we finally get the following result.

Proposition 1. *Let $\mathcal{V} = \{V \in \mathbb{R}^{|\mathcal{S}|} : \|V\| \le 1/(1-\gamma)\}$. For any $\eta, \delta \in (0,1)$, with probability at least $1-\delta$,*

$$\max_{V\in\mathcal{V}} \max_{s,a} |\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)| \le C^{\text{tv}}_u(N,\eta,\delta), \text{ where,}$$
$$C^{\text{tv}}_u(N,\eta,\delta) = 4\eta + \frac{2}{1-\gamma}\sqrt{\frac{|\mathcal{S}|\log(6|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{2N}}. \tag{15}$$

Tracing back the steps to (13), we can get an arbitrary small bound for $\|V^\pi - \widehat{V}^\pi\|$ by selecting N appropriately, as specified in Theorem 1.

5.2 Chi-square uncertainty set

We will first get the following upperbound:

Lemma 4 (Chi-square uncertainty set). *For any* $(s,a) \in \mathcal{S} \times \mathcal{A}$ *and for any* $V \in \mathbb{R}^{|\mathcal{S}|}, \|V\| \leq 1/(1-\gamma)$,

$$|\sigma_{\widehat{\mathcal{P}}^{\mathrm{c}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\mathrm{c}}_{s,a}}(V)| \leq \max_{\mu:0\leq\mu\leq V} |\sqrt{c_r \mathrm{Var}_{\widehat{P}_{s,a}}(V-\mu)} - \sqrt{c_r \mathrm{Var}_{P^o_{s,a}}(V-\mu)}| + \max_{\mu:0\leq\mu\leq V} |\widehat{P}_{s,a}(V-\mu) - P^o_{s,a}(V-\mu)|. \quad (16)$$

The second term of (16) can be bounded using Lemma 3. However, the first term, which involves the square-root of the variance is more challenging. We use a concentration inequality that is applicable for variance to overcome this challenge. Finally, we get the following result.

Proposition 2. *Let* $\mathcal{V} = \{V \in \mathbb{R}^{|\mathcal{S}|} : \|V\| \leq 1/(1-\gamma)\}$. *For any* $\eta, \delta \in (0,1)$, *with probability at least* $(1-\delta)$,

$$\max_{V\in\mathcal{V}} \max_{s,a} |\sigma_{\widehat{\mathcal{P}}^{\mathrm{c}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\mathrm{c}}_{s,a}}(V)| \leq C^{\mathrm{c}}_u(N,\eta,\delta), \text{ where,}$$

$$C^{\mathrm{c}}_u(N,\eta,\delta) \leq \sqrt{\frac{32\eta c_r}{1-\gamma}} + 2\eta + \frac{1}{1-\gamma}\sqrt{\frac{(2c_r+1)|\mathcal{S}|\log(12|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{N}}, \quad (17)$$

Now, by selecting appropriate N as specified in Theorem 2, we can show the ε-optimality of π_K.

The details on the KL uncertainty set analysis is included in the appendix.

6 Experiments

In this section we demonstrate the convergence behavior and robust performance of our REVI algorithm using numerical experiments. We consider two different settings, namely, the *Gambler's Problem* environment (Sutton and Barto, 2018, Example 4.3) and *FrozenLake8x8* environment in OpenAI Gym (Brockman et al., 2016). We also consider the TV uncertainty set and chi-square uncertainty set. We solve the optimization problem $\sigma_{\widehat{\mathcal{P}}}$ and $\sigma_{\mathcal{P}}$ using the Scipy (Virtanen et al., 2020) optimization library.

We illustrate the following important characteristics of the REVI algorithm:

(1) Rate of convergence with respect to the number of iterations: To demonstrate this, we plot $\|V_k - V^*\|$ against the iteration number k, where V_k is the value at the kth step of the REVI algorithm with $N = 5000$. We compute V^* using the full knowledge of the uncertainty set for benchmarking the performance of the REVI algorithm.

(2) Rate of convergence with respect to the number of samples: To show this, we plot $\|V_K(N) - V^*\|$ against the number of samples N, where $V_K(N)$ is final value obtained from the REVI algorithm using N samples.

(3) Robustness of the learned policy: To demonstrate this, we plot the number of times the robust policy π_K (obtained from the REVI algorithm) successfully completed the task as a function of the change in an environment parameter. We perform 1000 trials for each environment and each uncertainty set, and plot the fraction of the success.

Gambler's Problem: In gambler's problem, a gambler starts with a random balance in her account and makes bets on a sequence of coin flips, winning her stake with heads and losing with tails, until she wins \$100 or loses all money. This problem can be formulated as a chain MDP with states in $\{1, \cdots, 99\}$ and when in state s the available actions are in $\{0, 1, \cdots, \min(s, 100-s)\}$. The agent is rewarded 1 after reaching a goal and rewarded 0 in every other timestep. The biased coin probability is fixed throughout the game. We denote its heads-up probability as p_h and use 0.6 as a nominal model for training our algorithm. We also fix $c_r = 0.2$ for the chi-square uncertainty set experiments and $c_r = 0.4$ for the TV uncertainty set experiments.

The red curves with square markers in the first two plots in Fig. 1 show the rate of convergence with respect to the number of iterations for TV and chi-square uncertainty sets respectively. As expected, convergence is fast due to the contraction property of the robust Bellman operator.

The blue curves with triangle markers in the first two plots in Fig. 1 show the rate of convergence with respect to the number of samples for TV and chi-square uncertainty sets. We generated these curves for 10 different seed runs. The bold line depicts the mean of these runs and the error bar is the standard deviation. As expected, the plots show that $V_K(N)$ converges to V^* as N increases.

We then demonstrate the robustness of the approximate robust policy π_K (obtained with $N = 100, 500, 3000$) by evaluating its performance on environments with different values of p_h. We plot the fraction of the wins out of 1000 trails. We also plot the performance the optimal robust policy π^* as a benchmark. The third and fourth plot in Fig. 1 show the results with TV and chi-square uncertainty sets respectively. We note that the performance of the non-robust policy decays drastically as we decrease the parameter p_h from its nominal value 0.6. On the other hand, the optimal robust policy performs consistently better under this change in the environment. We also note that

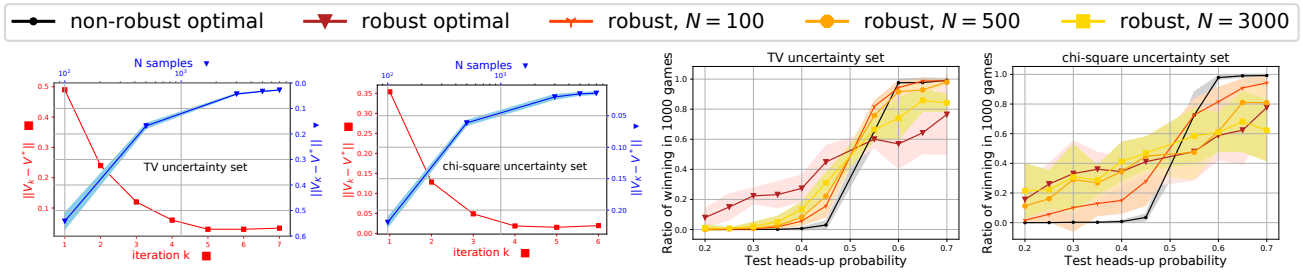

Figure 1: *Experiment results for the Gambler's problem.* The first two plots shows the rate of convergence with respect to the number of iterations (k) and the rate of convergence with respect to the number of samples (N) for the TV and chi-square uncertainty set, respectively. The third and fourth plots shows the robustness of the learned policy against changes in the model parameter (heads-up probability).

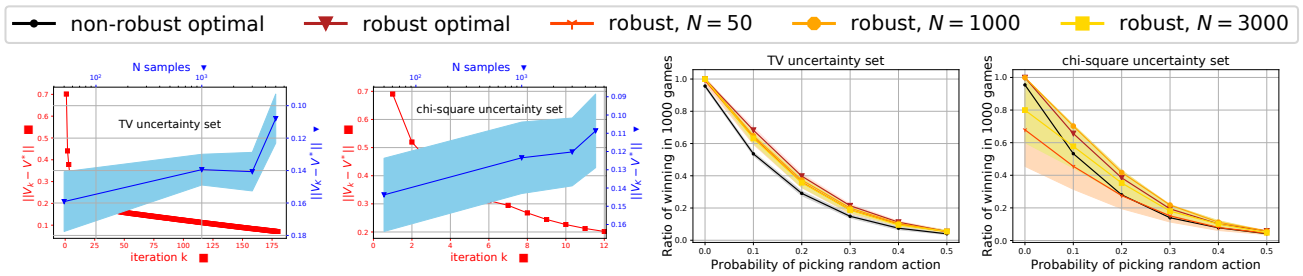

Figure 2: *Experiment results for the FrozenLake8x8 environment.* The first two plots shows the rate of convergence with respect to the number of iterations (k) and the rate of convergence with respect to the number of samples (N) for the TV and chi-square uncertainty set, respectively. The third and fourth plots shows the robustness of the learned policy against changes in the model parameter (probability of picking a random action).

$\pi_K(N)$ closely follows the performance of π^* for large enough N.

Frozen Lake environment: *FrozenLake8x8* is a gridworld environment of size 8×8. It consists of some flimsy tiles which makes the agent fall into the water. The agent is rewarded 1 after reaching a goal tile without falling and rewarded 0 in every other timestep. We use the *FrozenLake8x8* environment with default design as our nominal model except that we make the probability of transitioning to a state in the intended direction to be 0.4 (the default value is 1/3). We also set $c_r = 0.35$ for the chi-square uncertainty set experiments and $c_r = 0.7$ for the TV uncertainty set experiments.

The red curves in the first two plots in Fig. 2 show the rate of convergence with respect to the number of iterations for TV and chi-square uncertainty sets respectively. The blue curves in the first two plots in Fig. 2 show the rate of convergence with respect to the number of samples for TV and chi-square uncertainty sets respectively. The behavior is similar to the one observed in the case of gambler's problem.

We demonstrate the robustness of the learned policy by evaluating it on FrozenLake test environments with action perturbations. In the real-world settings, due to model mismatch or noise in the environments, the resulting action can be different from the intended action. We model this by picking a random action with some probability at each time step. In addition, we change the probability of transitioning to a state in the intended direction to be 0.2 for these test environments. We observe that the performance of the robust RL policy is consistently better than the non-robust policy as we introduce model mismatch in terms of the probability of picking random actions. We also note that $\pi_K(N)$ closely follows the performance of π^* for large enough N.

We note that we have included our code for experiments in this GitHub page. We note that we can employ a hyperparameter learning strategy to find the best value of c_r. We demonstrate this on the FrozenLake environment for the TV uncertainty set. We computed the optimal robust policy for each $c_r \in \{0.1, 0.2, \ldots, 1.6\}$. We tested these policies across 1000 games with random action probabilities $\{0.1, 0.2, \ldots, 0.5\}$ on the test environment. We found that the policy for $c_r = 1.2$ has the best winning ratio across all the random action probabilities. We do not exhibit exhaustive experiments on this hyperparameter learning strategy as it is out of scope of the intent of this manuscript.

7 Conclusion and Future Work

We presented a model-based robust reinforcement learning algorithm called Robust Empirical Value Iteration algorithm, where we used an approximate robust Bellman updates in the vanilla robust value iteration algorithm. We provided a finite sample performance characterization of the learned policy with respect to the optimal robust policy for three different uncertainty sets, namely, the total variation uncertainty set, the chi-square uncertainty set, and the Kullback-Leibler uncertainty set. We also demonstrated the performance of REVI algorithm on two different environments showcasing its theoretical properties of convergence. We also showcased the REVI algorithm based policy being robust to the changes in the environment as opposed to the non-robust policies.

The goal of this work was to develop the fundamental theoretical results for the finite state space and action space regime. As mentioned earlier, the sub-optimality of the sample complexity of our REVI algorithm in factors $|\mathcal{S}|$ and $1/(1-\gamma)$ needs more investigation and refinements in the analyses. In the future, we will extend this idea to robust RL with linear and nonlinear function approximation architectures and for more general models in deep RL.

Acknowledgments

Dileep Kalathil gratefully acknowledges funding from the U.S. National Science Foundation (NSF) grants NSF-EAGER-1839616, NSF-CRII-CPS-1850206 and NSF-CAREER-EPCN-2045783.

References

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. 2, 5

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.*, 91(3):325–349. 2, 4, 5

Borkar, V. S. (2002). Q-learning for risk-sensitive control. *Mathematics of operations research*, 27(2):294–311. 3

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*. 7

Derman, E., Mankowitz, D., Mann, T., and Mannor, S. (2020). A bayesian approach to robust reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 648–658. 1

Derman, E., Mankowitz, D. J., Mann, T. A., and Mannor, S. (2018). Soft-robust actor-critic policy-gradient. In *AUAI press for Association for Uncertainty in Artificial Intelligence*, pages 208–218. 1, 2

Dullerud, G. E. and Paganini, F. (2013). *A course in robust control theory: a convex approach*, volume 36. Springer Science & Business Media. 3

Haskell, W. B., Jain, R., and Kalathil, D. (2016). Empirical dynamic programming. *Mathematics of Operations Research*, 41(2):402–429. 2

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280. 1, 3, 4, 13, 14, 15, 18

Kalathil, D., Borkar, V. S., and Jain, R. (2021). Empirical Q-Value Iteration. *Stochastic Systems*, 11(1):1–18. 2

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872. 2, 3, 5

Lim, S. H. and Autef, A. (2019). Kernel-based reinforcement learning in robust Markov decision processes. In *International Conference on Machine Learning*, pages 3973–3981. 2

Lim, S. H., Xu, H., and Mannor, S. (2013). Reinforcement learning in robust Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 701–709. 2

Mankowitz, D. J., Levine, N., Jeong, R., Abdolmaleki, A., Springenberg, J. T., Shi, Y., Kay, J., Hester, T., Mann, T., and Riedmiller, M. (2020). Robust reinforcement learning for continuous control with model misspecification. In *International Conference on Learning Representations*. 1, 2

Mannor, S., Mebel, O., and Xu, H. (2016). Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509. 1

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample-variance penalization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*. 11

Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798. 1, 4, 18, 21

Panaganti, K. and Kalathil, D. (2021). Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *Proceedings of the 38th International Conference on Machine Learning*, pages 511–520. 1, 2

Petrik, M. and Subramanian, D. (2014). Raam: The benefits of robustness in approximating aggregated mdps in reinforcement learning. In *NIPS*, pages 1979–1987. 2

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. (2017). Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. 1, 3

Roy, A., Xu, H., and Pokutta, S. (2017). Reinforcement learning under model mismatch. In *Advances in Neural Information Processing Systems*, pages 3043–3052. 1, 2

Russel, R. H. and Petrik, M. (2019). Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps. *Advances in Neural Information Processing Systems*. 1

Sidford, A., Wang, M., Wu, X., Yang, L. F., and Ye, Y. (2018). Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5192–5202. 2, 5

Singh, S. P. and Yee, R. C. (1994). An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233. 2, 14

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press. 7

Tamar, A., Mannor, S., and Xu, H. (2014). Scaling up robust mdps using function approximation. In *International Conference on Machine Learning*, pages 181–189. 2

van Handel, R. (2014). Probability in High Dimension. Technical report, Princeton University NJ. 11

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University press. 11

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272. 7

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

Xu, H. and Mannor, S. (2010). Distributionally robust Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 2505–2513. 1

Yang, W., Zhang, L., and Zhang, Z. (2021). Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*. 2

Yu, P. and Xu, H. (2015). Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543. 1

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. (2020a). Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037. 1

Zhang, K., Hu, B., and Basar, T. (2020b). Policy optimization for H_2 linear control with H_∞ robustness guarantee: Implicit regularization and global convergence. In *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control, L4DC 2020, Online Event, Berkeley, CA, USA, 11-12 June 2020*, volume 120 of *Proceedings of Machine Learning Research*, pages 179–190. 3

Zhou, K., Doyle, J. C., Glover, K., et al. (1996). *Robust and optimal control*, volume 40. Prentice hall New Jersey. 3

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. 2, 5, 17

Appendix

A Useful Technical Results

In this section we state some existing results that are useful in our analysis.

Lemma 5 (Hoeffding's inequality (Vershynin, 2018, Theorem 2.2.6)). *Let $X_1, \cdots, X_T$ be independent random variables. Assume that $X_t \in [m_t, M_t]$ for every t with $M_t > m_t$. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\sum_{t=1}^{T}(X_t - \mathbb{E}[X_t]) \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^{T}(M_t - m_t)^2}\right).$$

Lemma 6 (Self-bounding variance inequality (Maurer and Pontil, 2009, Theorem 10)). *Let $X_1, \cdots, X_T$ be independent and identically distributed random variables with finite variance, that is, $\mathrm{Var}(X_1) < \infty$. Assume that $X_t \in [0, M]$ for every t with $M > 0$, and let $S_T^2 = \frac{1}{T}\sum_{t=1}^{T} X_t^2 - (\frac{1}{T}\sum_{t=1}^{T} X_t)^2$. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(|S_T - \sqrt{\mathrm{Var}(X_1)}| \geq \varepsilon\right) \leq 2\exp\left(-\frac{T\varepsilon^2}{2M^2}\right).$$

Proof. The proof of this lemma directly follows from (Maurer and Pontil, 2009, Theorem 10) by noting that we can rewrite S_T^2 as follows

$$\frac{T}{T-1}S_T^2 = \frac{1}{T(T-1)}\sum_{i,j=1}^{T}(X_i - X_j)^2.$$

Also, note that we apply (Maurer and Pontil, 2009, Theorem 10) for the scaled random variables $X_t/M \in [0, 1]$. □

We now provide a covering number result that is useful to get high probability concentration bounds for value function classes. We first define minimal η-cover of a set.

Definition 1 (Minimal η-cover; (van Handel, 2014, Definition 5.5)). *A set $\mathcal{N}_{\mathcal{V}}(\eta)$ is called an η-cover for a metric space $(\mathcal{V}, d)$ if for every $V \in \mathcal{V}$, there exists a $V' \in \mathcal{N}$ such that $d(V, V') \leq \eta$. Furthermore, $\mathcal{N}_{\mathcal{V}}(\eta)$ with the minimal cardinality ($|\mathcal{N}_{\mathcal{V}}(\eta)|$) is called a minimal η-cover.*

From (van Handel, 2014, Exercise 5.5 and Lemma 5.13) we have the following result.

Lemma 7 (Covering Number). *Let $\mathcal{V} = \{V \in \mathbb{R}^{|\mathcal{S}|} \ : \ \|V\| \leq V_{\max}\}$. Let $\mathcal{N}_{\mathcal{V}}(\eta)$ be a minimal η-cover of $\mathcal{V}$ with respect to the distance metric $d(V, V') = \|V - V'\|$ for some fixed $\eta \in (0, 1)$. Then we have*

$$\log|\mathcal{N}_{\mathcal{V}}(\eta)| \leq |\mathcal{S}| \cdot \log\left(\frac{3\,V_{\max}}{\eta}\right).$$

Proof. We will consider the normalized metric space $(\mathcal{V}_n, d_n)$, where

$$\mathcal{V}_n := \mathcal{V}/V_{\max} = \{V \in \mathbb{R}^{|\mathcal{S}|} \ : \ \|V\| \leq 1\}$$

and $d_n := d/V_{\max}$ to make use of the fact that the covering number is invariant to the scaling of a metric space. Let $\eta_n := \eta/V_{\max}$. Then, it follows from (van Handel, 2014, Exercise 5.5 and Lemma 5.13) that

$$\log|\mathcal{N}_{\mathcal{V}}(\eta)| = \log|\mathcal{N}_{\mathcal{V}_n}(\eta_n)| \leq |\mathcal{S}| \cdot \log\left(\frac{3}{\eta_n}\right) = |\mathcal{S}| \cdot \log\left(\frac{3\,V_{\max}}{\eta}\right).$$

This completes the proof. □

Here we present another covering number result, with a similar proof as Lemma 7, that is useful to get our upperbound for the KL uncertainty set.

Lemma 8 (Covering Number of a bounded real line). *Let $\Theta \subset \mathbb{R}$ with $\Theta = [l, u]$ for some real numbers $u > l$. Let $\mathcal{N}_\Theta(\eta)$ be a minimal η-cover of Θ with respect to the distance metric $d(\theta, \theta') = |\theta - \theta'|$ for some fixed $\eta \in (0, 1)$. Then we have $|\mathcal{N}_\Theta(\eta)| \leq 3(u - l)/\eta$.*

B Proof of the Theorems

B.1 Concentration Results

Here, we prove Lemma 3. We state the following result first.

Lemma 9. *For any $V \in \mathbb{R}^{|\mathcal{S}|}$ with $\|V\| \le V_{\max}$, with probability at least $1-\delta$,*

$$\max_{(s,a)} |P^o_{s,a}V - \widehat{P}_{s,a}V| \le V_{\max}\sqrt{\frac{\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{2N}}$$

Proof. Fix any (s,a) pair. Consider a discrete random variable X taking value $V(j)$ with probability $P^o_{s,a}(j)$ for all $j \in \{1,2,\cdots,|\mathcal{S}|\}$. From Hoeffding's inequality (Lemma 5), we have

$$\mathbb{P}(P^o_{s,a}V - \widehat{P}_{s,a}V \ge \varepsilon) \le \exp(-2N\varepsilon^2/V^2_{\max}), \quad \mathbb{P}(\widehat{P}_{s,a}V - P^o_{s,a}V \ge \varepsilon) \le \exp(-2N\varepsilon^2/V^2_{\max}).$$

Choosing $\varepsilon = V_{\max}\sqrt{\frac{\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{2N}}$, we get $\mathbb{P}(|P^o_{s,a}V - \widehat{P}_{s,a}V| \ge V_{\max}\sqrt{\frac{\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{2N}}) \le \frac{\delta}{|\mathcal{S}||\mathcal{A}|}$. Now, using union bound, we get

$$\mathbb{P}(\max_{(s,a)} |P^o_{s,a}V - \widehat{P}_{s,a}V| \ge V_{\max}\sqrt{\frac{\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{2N}}) \le \sum_{s,a} \mathbb{P}(|P^o_{s,a}V - \widehat{P}_{s,a}V| \ge V_{\max}\sqrt{\frac{\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{2N}}) \le \delta.$$

This completes the proof. □

Proof of Lemma 3: Let $\mathcal{V} = \{V \in \mathbb{R}^{|\mathcal{S}|} : \|V\|_\infty \le 1/(1-\gamma)\}$. Let $\mathcal{N}_{\mathcal{V}}(\eta)$ be a minimal η-cover of $\mathcal{V}$. Fix a $\mu \le V$. By the definition of $\mathcal{N}_{\mathcal{V}}(\eta)$, there exists a $\mu' \in \mathcal{N}_{\mathcal{V}}(\eta)$ such that $\|\mu - \mu'\| \le \eta$. Now, for these particular μ and μ', we get

$$\begin{aligned}
|\widehat{P}_{s,a}\mu - P^o_{s,a}\mu| &\le |\widehat{P}_{s,a}\mu - \widehat{P}_{s,a}\mu'| + |\widehat{P}_{s,a}\mu' - P^o_{s,a}\mu'| + |P^o_{s,a}\mu' - P^o_{s,a}\mu| \\
&\overset{(a)}{\le} \|\widehat{P}_{s,a}\|_1\|\mu - \mu'\|_\infty + |\widehat{P}_{s,a}\mu' - P^o_{s,a}\mu'| + \|P^o_{s,a}\|_1\|\mu' - \mu\|_\infty \\
&\le \sup_{\mu' \in \mathcal{N}_{\mathcal{V}}(\eta)} \max_{s,a} |\widehat{P}_{s,a}\mu' - P^o_{s,a}\mu'| + 2\eta
\end{aligned}$$

where (a) follows from Hölder's inequality. Now, taking max on both sides with respect to μ and (s,a) we get

$$\begin{aligned}
\sup_{\mu \in \mathcal{V}} \max_{s,a} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu| &\le \sup_{\mu' \in \mathcal{N}_{\mathcal{V}}(\eta)} \max_{s,a} |\widehat{P}_{s,a}\mu' - P^o_{s,a}\mu'| + 2\eta \\
&\overset{(b)}{\le} \frac{1}{1-\gamma}\sqrt{\frac{\log(4|\mathcal{S}||\mathcal{A}||\mathcal{N}_{\mathcal{V}}(\eta)|/\delta)}{2N}} + 2\eta \\
&\overset{(c)}{\le} \frac{1}{1-\gamma}\sqrt{\frac{|\mathcal{S}|\log(12|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{2N}} + 2\eta,
\end{aligned}$$

with probability at least $1-\delta/2$. Here, (b) follows from Lemma 9 and the union bound and (c) from Lemma 7. □

B.2 Proof of Theorem 1

Proof of Lemma 1. We only prove the first inequality since the proof is analogous for the other inequality. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$ we have

$$\begin{aligned}
\sigma_{\mathcal{P}_{s,a}}(V_2) - \sigma_{\mathcal{P}_{s,a}}(V_1) &= \inf_{q \in \mathcal{P}_{s,a}} q^\top V_2 - \inf_{\tilde{q} \in \mathcal{P}_{s,a}} \tilde{q}^\top V_1 = \inf_{q \in \mathcal{P}_{s,a}} \sup_{\tilde{q} \in \mathcal{P}_{s,a}} q^\top V_2 - \tilde{q}^\top V_1 \\
&\ge \inf_{q \in \mathcal{P}_{s,a}} q^\top (V_2 - V_1) = \sigma_{\mathcal{P}_{s,a}}(V_2 - V_1). \qquad (18)
\end{aligned}$$

By definition, for any arbitrary $\varepsilon > 0$, there exists a $P_{s,a} \in \mathcal{P}_{s,a}$ such that

$$P_{s,a}^{\top}(V_2 - V_1) - \varepsilon \leq \sigma_{\mathcal{P}_{s,a}}(V_2 - V_1). \tag{19}$$

Using (19) and (18),

$$\sigma_{\mathcal{P}_{s,a}}(V_1) - \sigma_{\mathcal{P}_{s,a}}(V_2) \leq P_{s,a}^{\top}(V_1 - V_2) + \varepsilon \overset{(a)}{\leq} \|P_{s,a}\|_1 \|V_1 - V_2\| + \varepsilon = \|V_1 - V_2\| + \varepsilon$$

where (a) follows from Holder's inequality. Since ε is arbitrary, we get, $\sigma_{\mathcal{P}_{s,a}}(V_1) - \sigma_{\mathcal{P}_{s,a}}(V_2) \leq \|V_1 - V_2\|$. Exchanging the roles of V_1 and V_2 completes the proof. □

Proof of Lemma 2. Fix any (s,a) pair. From (Iyengar, 2005, Lemma 4.3) we have that

$$\sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V) = P^o_{s,a}V + \max_{\mu:0\leq\mu\leq V} \Big(- P^o_{s,a}\mu - c_r \max_s(V_\mu(s)) + c_r \min_s(V_\mu(s)) \Big) \tag{20}$$

$$\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) = \widehat{P}_{s,a}V + \max_{\mu:0\leq\mu\leq V} \Big(- \widehat{P}_{s,a}\mu - c_r \max_s(V_\mu(s)) + c_r \min_s(V_\mu(s)) \Big), \tag{21}$$

where $0 \leq \mu \leq V$ is an elementwise inequality and $V_\mu(s) = V(s) - \mu(s)$ for all $s \in \mathcal{S}$.

Using the fact that $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$, it directly follows that

$$|\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)| \leq |\widehat{P}_{s,a}V - P^o_{s,a}V| + \max_{\mu:0\leq\mu\leq V} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu|.$$

Further simplifying we get

$$\begin{aligned}|\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)| &\leq |\widehat{P}_{s,a}V - P^o_{s,a}V| + \max_{\mu:0\leq\mu\leq V} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu| \\ &\leq \max_{\mu\in\mathcal{V}} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu| + \max_{\mu:0\leq\mu\leq V} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu| \leq 2\max_{\mu\in\mathcal{V}} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu|.\end{aligned}$$

This completes the proof. □

We are now ready to prove Proposition 1.

Proof of Proposition 1. For any given $V \in \mathcal{V}$ and (s,a), from Lemma 2, we have

$$|\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)| \leq 2\max_{\mu\in\mathcal{V}} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu| \leq 2\max_{\mu\in\mathcal{V}} \max_{s,a} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu|.$$

Taking the maximum over V and (s,a) on both sides, we get

$$\max_{V\in\mathcal{V}} \max_{s,a} |\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)| \leq 2\max_{\mu\in\mathcal{V}} \max_{s,a} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu|. \tag{22}$$

Now, from the proof of Lemma 3, for any $\eta, \delta \in (0,1)$, we get

$$\max_{\mu\in\mathcal{V}} \max_{s,a} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu| \leq \frac{1}{1-\gamma}\sqrt{\frac{|\mathcal{S}|\log(6|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{2N}} + 2\eta, \tag{23}$$

with probability greater than $1-\delta$. Using (23) in (22), we get the desired result. □

We also need the following result that specifies the amplification when replacing the algorithm iterate value function with the value function of the policy towards approximating the optimal value.

Lemma 10. *Let V_k and Q_k be as given in the REVI algorithm for $k \geq 1$. Also, let $\pi_k = \arg\max_a Q_k(s,a)$. Then,*

$$\|\widehat{V}^* - \widehat{V}^{\pi_k}\| \leq \frac{2\gamma}{1-\gamma}\|V_k - \widehat{V}^*\|.$$

Furthermore,

$$\|V^* - V^{\pi_k}\| \leq \frac{2}{1-\gamma}\|Q_k - Q^*\|.$$

Proof. The proof is similar to the proof in (Singh and Yee, 1994, Main Theorem, Corollary 2). A straight forward modification to this proof, using the fact that $\sigma_{\widehat{\mathcal{P}}_{s,a}}$ and $\sigma_{\mathcal{P}_{s,a}}$ are 1-Lipschitz functions as shown in Lemma 1, will give the desired result. □

Proof of Theorem 1. Recall the empirical RMDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, r, \widehat{\mathcal{P}}^{\text{tv}}, \gamma)$. For any policy π, let $\widehat{V}^{\pi}$ be robust value function of policy π with respect to the RMDP $\widehat{M}$. The optimal robust policy, value function, and state-action value function of $\widehat{M}$ are denoted as $\widehat{\pi}^{\star}, \widehat{V}^{\star}$ and $\widehat{Q}^{\star}$, respectively. Also, for any policy π, we have $\widehat{Q}^{\pi}(s,a) = r(s,a) + \gamma\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(\widehat{V}^{\pi})$ and $Q^{\pi}(s,a) = r(s,a) + \gamma\sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V^{\pi})$.

Let V_k and Q_k be as given in the REVI algorithm for $k \geq 1$. Also, let $\pi_k(s) = \arg\max_a Q_k(s,a)$. Now,

$$\|V^* - V^{\pi_k}\| \leq \|V^* - \widehat{V}^*\| + \|\widehat{V}^* - \widehat{V}^{\pi_k}\| + \|\widehat{V}^{\pi_k} - V^{\pi_k}\|. \tag{24}$$

1) Bounding the first term in (24): Let $\mathcal{V} = \{V \in \mathbb{R}^{|\mathcal{S}|} : \|V\| \leq 1/(1-\gamma)\}$. For any $s \in \mathcal{S}$,

$$\begin{aligned}
V^*(s) - \widehat{V}^*(s) = Q^*(s,\pi^*(s)) - \widehat{Q}^*(s,\hat{\pi}^*(s)) &\overset{(a)}{\leq} Q^*(s,\pi^*(s)) - \widehat{Q}^*(s,\pi^*(s)) \\
&\overset{(b)}{=} \gamma\sigma_{\mathcal{P}^{\text{tv}}_{s,\pi^*(s)}}(V^*) - \gamma\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,\pi^*(s)}}(\widehat{V}^*) \\
&= \gamma(\sigma_{\mathcal{P}^{\text{tv}}_{s,\pi^*(s)}}(V^*) - \sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,\pi^*(s)}}(V^*)) + \gamma(\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,\pi^*(s)}}(V^*) - \sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,\pi^*(s)}}(\widehat{V}^*)) \\
&\overset{(c)}{\leq} \gamma(\sigma_{\mathcal{P}^{\text{tv}}_{s,\pi^*(s)}}(V^*) - \sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,\pi^*(s)}}(V^*)) + \gamma\|V^* - \widehat{V}^*\| \\
&\leq \gamma \max_{V\in\mathcal{V}} \max_{s,a} |\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)| + \gamma\|V^* - \widehat{V}^*\|
\end{aligned}$$

where (a) follows since $\hat{\pi}^*$ is the robust optimal policy for $\widehat{M}$, (b) follows from the definitions of Q^* and $\widehat{Q}^*$, (c) follows from Lemma 1. Similarly analyzing for $\widehat{V}^*(s) - V^*(s)$, we get

$$\|V^* - \widehat{V}^*\| \leq \frac{\gamma}{(1-\gamma)} \max_{V\in\mathcal{V}} \max_{s,a} |\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)|. \tag{25}$$

Now, using Proposition 1, with probability greater than $1-\delta$, we get

$$\|V^* - \widehat{V}^*\| \leq \frac{\gamma}{(1-\gamma)} C_u^{\text{tv}}(N,\eta,\delta), \tag{26}$$

where $C_u^{\text{tv}}(N,\eta,\delta)$ is given in equation (15) in the statement of Proposition 1.

2) Bounding the second term in (24): Let $\widehat{T}$ be the robust Bellman operator corresponding to $\widehat{M}$. So, $\widehat{T}$ is a γ-contraction mapping and $\widehat{V}^*$ is its unique fixed point (Iyengar, 2005). The REVI iterates $V_k, k \geq 0$, with $V_0 = 0$, can now be expressed as $V_{k+1} = \widehat{T}V_k$. Using the properties of $\widehat{T}$, we get

$$\|V_k - \widehat{V}^*\| = \|\widehat{T}V_{k-1} - \widehat{T}\widehat{V}^*\| \leq \gamma\|V_{k-1} - \widehat{V}^*\| \leq \cdots \leq \gamma^k\|V_0 - \widehat{V}^*\| \leq \gamma^k/(1-\gamma). \tag{27}$$

Now, using Lemma 10, we get

$$\|\widehat{V}^{\pi_k} - \widehat{V}^*\| \leq \frac{2\gamma^{k+1}}{(1-\gamma)^2}. \tag{28}$$

3) Bounding the third term in (24): This is similar to bounding the first term. For any $s \in \mathcal{S}$,

$$\begin{aligned}
V^{\pi_k}(s) - \widehat{V}^{\pi_k}(s) = Q^{\pi_k}(s,\pi_k(s)) - \widehat{Q}^{\pi_k}(s,\pi_k(s)) &= \gamma\sigma_{\mathcal{P}_{s,\pi_k(s)}}(V^{\pi_k}) - \gamma\sigma_{\widehat{\mathcal{P}}_{s,\pi_k(s)}}(\widehat{V}^{\pi_k}) \\
&= \gamma(\sigma_{\mathcal{P}_{s,\pi_k(s)}}(V^{\pi_k}) - \sigma_{\mathcal{P}_{s,\pi_k(s)}}(\widehat{V}^{\pi_k})) + \gamma(\sigma_{\mathcal{P}_{s,\pi_k(s)}}(\widehat{V}^{\pi_k}) - \sigma_{\widehat{\mathcal{P}}_{s,\pi_k(s)}}(\widehat{V}^{\pi_k})) \\
&\overset{(d)}{\leq} \gamma\|V^{\pi_k} - \widehat{V}^{\pi_k}\| + \gamma(\sigma_{\mathcal{P}_{s,\pi_k(s)}}(\widehat{V}^{\pi_k}) - \sigma_{\widehat{\mathcal{P}}_{s,\pi_k(s)}}(\widehat{V}^{\pi_k})) \\
&\leq \gamma\|V^{\pi_k} - \widehat{V}^{\pi_k}\| + \gamma \max_{V\in\mathcal{V}} \max_{s,a} |\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)|
\end{aligned}$$

where (d) follows from Lemma 1. Similarly analyzing for $\widehat{V}^{\pi_k}(s) - V^{\pi_k}(s)$, we get,

$$\|V^{\pi_k} - \widehat{V}^{\pi_k}\| \leq \frac{\gamma}{(1-\gamma)} \max_{V \in \mathcal{V}} \max_{s,a} |\sigma_{\widehat{\mathcal{P}}^{\text{tv}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{tv}}_{s,a}}(V)|. \tag{29}$$

Now, using Proposition 1, with probability greater than $1-\delta$, we get

$$\|V^{\pi_k} - \widehat{V}^{\pi_k}\| \leq \frac{\gamma}{(1-\gamma)} C_u^{\text{tv}}(N, \eta, \delta). \tag{30}$$

Using (26) - (30) in (24), we get, with probability at least $1-2\delta$,

$$\|V^* - V^{\pi_k}\| \leq \frac{2\gamma^{k+1}}{(1-\gamma)^2} + \frac{2\gamma}{(1-\gamma)} C_u^{\text{tv}}(N, \eta, \delta). \tag{31}$$

Using the value of $C_u^{\text{tv}}(N, \eta, \delta)$ as given in Proposition 1, we get

$$\|V^* - V^{\pi_k}\| \leq \frac{2\gamma^{k+1}}{(1-\gamma)^2} + \frac{4\gamma}{(1-\gamma)^2} \sqrt{\frac{|\mathcal{S}| \log(6|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{2N}} + \frac{8\gamma\eta}{(1-\gamma)} \tag{32}$$

with probability at least $1-2\delta$.

Now, choose $\eta = \varepsilon(1-\gamma)/(24\gamma)$. Since $\varepsilon \in (0, 24\gamma/(1-\gamma))$, this particular η is in $(0,1)$. Now, choosing

$$k \geq K_0 = \frac{1}{\log(1/\gamma)} \log(\frac{6\gamma}{\varepsilon(1-\gamma)^2}), \tag{33}$$

$$N \geq N^{\text{tv}} = \frac{72\gamma^2}{(1-\gamma)^4} \frac{|\mathcal{S}| \log(144\gamma|\mathcal{S}||\mathcal{A}|/(\delta\varepsilon(1-\gamma)^2))}{\varepsilon^2}, \tag{34}$$

we get $\|V^* - V^{\pi_k}\| \leq \varepsilon$ with probability at least $1-2\delta$. □

B.3 Proof of Theorem 2

Proof of Lemma 4. Fix an (s,a) pair. From (Iyengar, 2005, Lemma 4.2), we have

$$\sigma_{\mathcal{P}^{\text{c}}_{s,a}}(V) = \max_{\mu: 0 \leq \mu \leq V} \left(P^o_{s,a}(V-\mu) - \sqrt{c_r \text{Var}_{P^o_{s,a}}(V-\mu)} \right), \tag{35}$$

where $\text{Var}_{P^o_{s,a}}(V-\mu) = P^o_{s,a}(V-\mu)^2 - (P^o_{s,a}(V-\mu))^2$. We get a similar expression for $\sigma_{\widehat{\mathcal{P}}^{\text{c}}_{s,a}}(V)$. Using these expressions, with the additional facts that $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$ and $\max_x (f(x)+g(x)) \leq \max_x f(x) + \max_x g(x)$, we get the desired result. □

We state the following concentration result that is useful for the proof of Proposition 2.

Lemma 11. *For any $V \in \mathbb{R}_+^{|\mathcal{S}|}$ with $\|V\| \leq V_{\max}$, with probability at least $1-\delta$,*

$$\max_{(s,a)} |\sqrt{Var_{P^o_{s,a}} V} - \sqrt{Var_{\widehat{P}_{s,a}} V}| \leq V_{\max} \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}$$

Proof. Fix any (s,a) pair. Consider a discrete random variable X taking value $V(j)$ with probability $P^o_{s,a}(j)$ for all $j \in \{1, 2, \cdots, |\mathcal{S}|\}$. From the Self-bounding variance inequality (Lemma 6), we have

$$\mathbb{P}(|\sqrt{\text{Var}_{P^o_{s,a}} V} - \sqrt{\text{Var}_{\widehat{P}_{s,a}} V}| \geq \varepsilon) \leq 2\exp(-N\varepsilon^2/(2V_{\max}^2)).$$

Choosing $\varepsilon = V_{\max}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}$, we get $\mathbb{P}(|P^o_{s,a}V - \widehat{P}_{s,a}V| \geq V_{\max}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}) \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|}$. Now, using union bound, we get

$$\mathbb{P}(\max_{(s,a)} |\sqrt{\text{Var}_{P^o_{s,a}} V} - \sqrt{\text{Var}_{\widehat{P}_{s,a}} V}| \geq V_{\max}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}})$$
$$\leq \sum_{s,a} \mathbb{P}(|\sqrt{\text{Var}_{P^o_{s,a}} V} - \sqrt{\text{Var}_{\widehat{P}_{s,a}} V}| \geq V_{\max}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}) \leq \delta.$$

This completes the proof. □

We are now ready to prove Proposition 2.

Proof of Proposition 2. Fix an (s,a) pair. From Lemma 4, for any given $V \in \mathcal{V}$, we have

$$|\sigma_{\widehat{\mathcal{P}}^{\mathrm{c}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\mathrm{c}}_{s,a}}(V)| \leq \max_{\mu:0\leq\mu\leq V} |\sqrt{c_r \mathrm{Var}_{\widehat{P}_{s,a}}(V-\mu)} - \sqrt{c_r \mathrm{Var}_{P^o_{s,a}}(V-\mu)}| + \max_{\mu:0\leq\mu\leq V} |\widehat{P}_{s,a}(V-\mu) - P^o_{s,a}(V-\mu)|.$$

By a simple variable substitution, we get

$$|\sigma_{\widehat{\mathcal{P}}^{\mathrm{c}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\mathrm{c}}_{s,a}}(V)| \leq \max_{\mu\in\mathcal{V}_+} \max_{s,a} |\sqrt{c_r \mathrm{Var}_{\widehat{P}_{s,a}}\mu} - \sqrt{c_r \mathrm{Var}_{P^o_{s,a}}\mu}| + \max_{\mu\in\mathcal{V}} \max_{s,a} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu|,$$

which will give

$$\max_{V\in\mathcal{V}} \max_{s,a} |\sigma_{\widehat{\mathcal{P}}^{\mathrm{c}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\mathrm{c}}_{s,a}}(V)| \leq \max_{\mu\in\mathcal{V}_+} \max_{s,a} |\sqrt{c_r \mathrm{Var}_{\widehat{P}_{s,a}}\mu} - \sqrt{c_r \mathrm{Var}_{P^o_{s,a}}\mu}| + \max_{\mu\in\mathcal{V}} \max_{s,a} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu|, \tag{36}$$

where $\mathcal{V}_+ = \{V \in \mathbb{R}_+^{|\mathcal{S}|} : \|V\| \leq 1/(1-\gamma)\}$.

We will first bound the second term on the RHS of (36). From the proof of Lemma 3, for any $\eta, \delta \in (0,1)$, we get

$$\max_{\mu\in\mathcal{V}} \max_{s,a} |\widehat{P}_{s,a}\mu - P^o_{s,a}\mu| \leq \frac{1}{1-\gamma}\sqrt{\frac{|\mathcal{S}|\log(12|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{2N}} + 2\eta, \tag{37}$$

with probability greater than $1-\delta/2$.

Now, we will focus on the first term on the RHS of (36). Fix a $\mu \in \mathcal{V}_+$. Consider a minimal η-cover $\mathcal{N}_{\mathcal{V}_+}(\eta)$ of the set $\mathcal{V}_+$. By definition, there exists $\mu' \in \mathcal{N}_{\mathcal{V}_+}(\eta)$ such that $\|\mu - \mu'\| \leq \eta$. Now, following the same step as in the proof of Lemma 3, we get

$$\begin{aligned}
|\sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu} - \sqrt{\mathrm{Var}_{P^o_{s,a}}\mu}| &\leq |\sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu} - \sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu'}| + |\sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu'} - \sqrt{\mathrm{Var}_{P^o_{s,a}}\mu'}| + |\sqrt{\mathrm{Var}_{P^o_{s,a}}\mu} - \sqrt{\mathrm{Var}_{P^o_{s,a}}\mu'}| \\
&\overset{(a)}{\leq} |\sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu'} - \sqrt{\mathrm{Var}_{P^o_{s,a}}\mu'}| + \sqrt{|\mathrm{Var}_{\widehat{P}_{s,a}}\mu - \mathrm{Var}_{\widehat{P}_{s,a}}\mu'|} + \sqrt{|\mathrm{Var}_{P^o_{s,a}}\mu - \mathrm{Var}_{P^o_{s,a}}\mu'|} \\
&\overset{(b)}{\leq} |\sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu'} - \sqrt{\mathrm{Var}_{P^o_{s,a}}\mu'}| + \sqrt{|\widehat{P}_{s,a}(\mu^2 - \mu'^2)|} + \sqrt{|(\widehat{P}_{s,a}\mu)^2 - (\widehat{P}_{s,a}\mu')^2)|} + \\
&\qquad \sqrt{|P^o_{s,a}(\mu^2 - \mu'^2)|} + \sqrt{|(P^o_{s,a}\mu)^2 - (P^o_{s,a}\mu')^2)|} \\
&\overset{(c)}{\leq} |\sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu'} - \sqrt{\mathrm{Var}_{P^o_{s,a}}\mu'}| + \sqrt{\frac{32\eta}{1-\gamma}} \\
&\leq \sup_{\mu'\in\mathcal{N}_{\mathcal{V}_+}(\eta)} \max_{s,a} |\sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu'} - \sqrt{\mathrm{Var}_{P^o_{s,a}}\mu'}| + \sqrt{\frac{32\eta}{1-\gamma}}
\end{aligned}$$

where (a) follows from the fact $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$ for all $x, y \in \mathbb{R}_+$, (b) follows from the fact $|\sqrt{x+y}| \leq \sqrt{x} + \sqrt{y}$ for all $x, y \in \mathbb{R}_+$, and (c) follows by using the fact $x^2 - y^2 = (x+y)(x-y)$, $\|\mu\| \leq 1/(1-\gamma)$, and $\|\mu'\| \leq 1/(1-\gamma)$ with Hölder's inequality. Now, taking max on both sides with respect to μ and (s,a) we get

$$\begin{aligned}
\sup_{\mu\in\mathcal{V}_+} \max_{s,a} |\sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu} - \sqrt{\mathrm{Var}_{P^o_{s,a}}\mu}| &\leq \sup_{\mu'\in\mathcal{N}_{\mathcal{V}_+}(\eta)} \max_{s,a} |\sqrt{\mathrm{Var}_{\widehat{P}_{s,a}}\mu'} - \sqrt{\mathrm{Var}_{P^o_{s,a}}\mu'}| + \sqrt{\frac{32\eta}{1-\gamma}} \\
&\overset{(d)}{\leq} \frac{1}{1-\gamma}\sqrt{\frac{2\log(4|\mathcal{S}||\mathcal{A}||\mathcal{N}_{\mathcal{V}_+}(\eta)|/\delta)}{N}} + \sqrt{\frac{32\eta}{1-\gamma}} \\
&\overset{(e)}{\leq} \frac{1}{1-\gamma}\sqrt{\frac{2|\mathcal{S}|\log(12|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{N}} + \sqrt{\frac{32\eta}{1-\gamma}},
\end{aligned} \tag{38}$$

with probability at least $1-\delta/2$. Here, (d) follows from Lemma 11 and the union bound and (e) from Lemma 7.

Applying (37) and (38) in (36), we get

$$\max_{V\in\mathcal{V}}\max_{s,a}|\sigma_{\widehat{\mathcal{P}}^{\mathrm{c}}_{s,a}}(V)-\sigma_{\mathcal{P}^{\mathrm{c}}_{s,a}}(V)|\leq\frac{1}{1-\gamma}\sqrt{\frac{2c_r|\mathcal{S}|\log(12|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{N}}+\sqrt{\frac{32\eta c_r}{1-\gamma}}+\frac{1}{1-\gamma}\sqrt{\frac{|\mathcal{S}|\log(12|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{2N}}+2\eta,$$

with probability greater than $1-\delta$. This completes the proof. □

Proof of Theorem 2. The basic steps of the proof is similar to that of Theorem 1. So, we present only the important steps.

Following the same steps as given before (26) and using Proposition 2, we get, with probability greater than $1-\delta$,

$$\|V^*-\widehat{V}^*\|\leq\frac{\gamma}{(1-\gamma)}C^{\mathrm{c}}_u(N,\eta,\delta)\tag{39}$$

Similarly, following the steps as given before (28), we get

$$\|\widehat{V}^{\pi_k}-\widehat{V}^*\|\leq\frac{2\gamma^{k+1}}{(1-\gamma)^2}.\tag{40}$$

In the same vein, following the steps as given before (30) and using Proposition 2, we get, with probability greater than $1-\delta$,

$$\|V^{\pi_k}-\widehat{V}^{\pi_k}\|\leq\frac{\gamma}{(1-\gamma)}C^{\mathrm{c}}_u(N,\eta,\delta).\tag{41}$$

Using (39) - (41), similar to (31), we get, with probability greater than $1-2\delta$,

$$\|V^*-V^{\pi_k}\|\leq\frac{2\gamma^{k+1}}{(1-\gamma)^2}+\frac{2\gamma}{(1-\gamma)}C^{\mathrm{c}}_u(N,\eta,\delta).\tag{42}$$

Using the value of $C^{\mathrm{c}}_u(N,\eta,\delta)$ as given in Proposition 2, we get, with probability greater than $1-2\delta$,

$$\|V^*-V^{\pi_k}\|\leq\frac{2\gamma^{k+1}}{(1-\gamma)^2}+\frac{8\gamma\sqrt{2\eta c_r}}{(1-\gamma)^{3/2}}+\frac{4\gamma\eta}{1-\gamma}+\frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{(2c_r+1)|\mathcal{S}|\log(12|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{N}}.$$

We can now choose k,ε,η to make each of the term on the RHS of the above inequality small. In particular, we select $\varepsilon\in(0,\min\{16\gamma/(1-\gamma),32\gamma\sqrt{2c_r}/(1-\gamma)^{3/2}\})$ and $\eta=\min\{\varepsilon(1-\gamma)/(16\gamma),\varepsilon^2(1-\gamma)^3/(2048c_r\gamma^2)\}$. Note that this choice also ensure $\eta\in(0,1)$. Now, by choosing

$$k\geq K_0=\frac{1}{\log(1/\gamma)}\cdot\log(\frac{8\gamma}{\varepsilon(1-\gamma)^2}),\tag{43}$$

$$N\geq N^{\mathrm{c}}=\frac{64\gamma^2}{(1-\gamma)^4}\cdot\frac{(2c_r+1)|\mathcal{S}|\log(12|\mathcal{S}||\mathcal{A}|/(\delta\eta(1-\gamma)))}{\varepsilon^2},\tag{44}$$

we will get $\|V^*-V^{\pi_k}\|\leq\varepsilon$ with probability at least $1-2\delta$. □

B.4 Proof of Theorem 3

We state a result from (Zhou et al., 2021) that will be useful in the proof of Theorem 3.

Lemma 12 ((Zhou et al., 2021, Lemma 4)). *Fix any $\delta\in(0,1)$. Let $X\sim P$ be a bounded random variable with $X\in[0,M]$ and let P_N denote the empirical distribution of P with N samples. For $t>0$, for any*

$$\lambda^*\in\arg\max_{\lambda\geq0}\left\{-\lambda\log(\mathbb{E}_P[\exp(-X/\lambda)])-\lambda t\right\},$$

(1) $\lambda^* = 0$. Furthermore, let the support of X be finite. Then there exists a problem dependent constant

$$N'(\delta, t, P) := \max\{\log(2/\delta)/\log(1/(1 - \min_{x \in \text{supp}(X)} P(X = x))) \ , \ 2M^2 \log(4/\delta)/(P(X = \operatorname{ess\,inf} X) - \exp(-t))^2\},$$

such that for $N \geq N'(\delta, t, P)$ we have, with probability at least $1 - \delta$,

$$0 \in \arg\max_{\lambda \geq 0} \{-\lambda \log(\mathbb{E}_{P_N}[\exp(-X/\lambda)]) - \lambda t\}.$$

(2) $\lambda^* > 0$. Then there exists a problem dependent constant

$$N''(\delta, t, P) := \max_{\lambda \in \{\underline{\lambda}, \lambda^*, M/t\}} \frac{8M^2 \exp(2M/\lambda)}{\tau^2} \log(6/\delta),$$

where $\underline{\lambda} = \lambda^*/2 > 0$ (independent of N) and

$$\tau = \min\{\underline{\lambda} \log(\mathbb{E}_P[\exp(-X/\underline{\lambda})]) + \underline{\lambda} t, (M/t) \log(\mathbb{E}_P[\exp(-tX/M)]) + M\} \\ - (\lambda^* \log(\mathbb{E}_P[\exp(-X/\lambda^*)]) + \lambda^* t) > 0,$$

such that for $N \geq N''(\delta, t, P)$, with probability at least $1 - \delta$, there exists a

$$\widehat{\lambda}^* \in \arg\max_{\lambda \geq 0} \{-\lambda \log(\mathbb{E}_{P_N}[\exp(-X/\lambda)]) - \lambda t\},$$

such that $\lambda^*, \widehat{\lambda}^* \in [\underline{\lambda}, M/t]$.

We now prove the following result.

Lemma 13. *For any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and for any $V \in \mathbb{R}^{|\mathcal{S}|}$ with $\|V\| \leq 1/(1-\gamma)$,*

$$|\sigma_{\widehat{\mathcal{P}}^{\text{kl}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\text{kl}}_{s,a}}(V)| \leq \frac{\exp(1/\lambda_{\text{kl}}(1-\gamma))}{c_r(1-\gamma)} \max_{\lambda \in [\lambda_{\text{kl}}, \frac{1}{c_r(1-\gamma)}]} |(P^o_{s,a} - \widehat{P}_{s,a}) \exp(-V/\lambda)| \tag{45}$$

holds with probability at least $1 - \delta/(2|\mathcal{S}||\mathcal{A}|)$ for $N \geq \max\{N'(\delta/(4|\mathcal{S}||\mathcal{A}|), c_r, P^o_{s,a}), N''(\delta/(4|\mathcal{S}||\mathcal{A}|), c_r, P^o_{s,a})\}$, where both N', N'' are defined as in Lemma 12.

Proof. Fix any (s, a) pair. From (Iyengar, 2005, Lemma 4.1), we have

$$\sigma_{\mathcal{P}^{\text{kl}}_{s,a}}(V) = \max_{\lambda \geq 0} \ (-c_r\lambda - \lambda \log(P^o_{s,a} \exp(-V/\lambda))), \quad \sigma_{\widehat{\mathcal{P}}^{\text{kl}}_{s,a}}(V) = \max_{\lambda \geq 0} \ (-c_r\lambda - \lambda \log(\widehat{P}_{s,a} \exp(-V/\lambda))), \tag{46}$$

where $\exp(-V/\lambda)$ is an element-wise exponential function. It is straight forward to show that $(-c_r\lambda - \lambda \log(P^o_{s,a} \exp(-V/\lambda)))$ is a concave function in λ. So, there exists an optimal solution λ^*. Similarly, let $\widehat{\lambda}^*$ be the optimal solution of the second problem above.

We can now give an upperbound for $\lambda^*, \widehat{\lambda}^*$ as follows: Since $\sigma_{\mathcal{P}^{\text{kl}}_{s,a}}(V) \geq 0$, we have

$$0 \leq -c_r\lambda^* - \lambda^* \log(P^o_{s,a} \exp(-V/\lambda^*)) \overset{(a)}{\leq} -c_r\lambda^* - \lambda^* \log(\exp(-1/(\lambda^*(1-\gamma)))) \leq -c_r\lambda^* + 1/(1-\gamma),$$

from which we can conclude that $\lambda^* \leq 1/(c_r(1-\gamma))$. Same argument applies for the case of $\widehat{\lambda}^*$.

From (Nilim and El Ghaoui, 2005, Appendix C) it follows that whenever the maximizer λ^* is 0 ($\widehat{\lambda}^*$ is 0), we have

$\sigma_{\mathcal{P}^{\mathrm{kl}}_{s,a}}(V) = V_{\min}$ ($\sigma_{\widehat{\mathcal{P}}^{\mathrm{kl}}_{s,a}}(V) = V_{\min}$) where $V_{\min} = \min_{j\in\mathcal{S}} V(j)$. We include this part in detail for completeness.

$$\begin{aligned}
\lim_{\lambda\downarrow 0} -c_r\lambda - \lambda\log(P^o_{s,a}\exp(-V/\lambda)) &= \lim_{\lambda\downarrow 0} -c_r\lambda - \lambda\log(\exp(-V_{\min}/\lambda)\sum_{s'} P^o_{s,a}(s')\exp((V_{\min}-V(s'))/\lambda)) \\
&= \lim_{\lambda\downarrow 0} V_{\min} - c_r\lambda - \lambda\log(\sum_{s'} P^o_{s,a}(s')\exp((V_{\min}-V(s'))/\lambda)) \\
&= \lim_{\lambda\downarrow 0} V_{\min} - c_r\lambda - \lambda\log(\sum_{s':V(s')=V_{\min}} P^o_{s,a}(s') + \sum_{s':V(s')>V_{\min}} P^o_{s,a}(s')\exp((V_{\min}-V(s'))/\lambda)) \\
&\overset{(a)}{=} \lim_{\lambda\downarrow 0} V_{\min} - c_r\lambda - \lambda\log(\sum_{s':V(s')=V_{\min}} P^o_{s,a}(s') + \mathcal{O}(\exp(-t/\lambda))) \\
&\overset{(b)}{=} \lim_{\lambda\downarrow 0} V_{\min} - c_r\lambda - \lambda\log(\sum_{s':V(s')=V_{\min}} P^o_{s,a}(s')) - \lambda\log(1+\mathcal{O}(\exp(-t/\lambda))) \\
&\overset{(c)}{=} \lim_{\lambda\downarrow 0} V_{\min} - \lambda(c_r + \log(\sum_{s':V(s')=V_{\min}} P^o_{s,a}(s'))) - \mathcal{O}(\lambda\exp(-t/\lambda)) = V_{\min},
\end{aligned}$$

where (a) follows by taking $t = \min_{s':V(s')>V_{\min}} V(s') - V_{\min} > 0$, and (b) and (c) follows from the Taylor series expansion. Thus when λ^* is 0, we have $\sigma_{\mathcal{P}^{\mathrm{kl}}_{s,a}}(V) = V_{\min}$. A similar argument applies for $\sigma_{\widehat{\mathcal{P}}^{\mathrm{kl}}_{s,a}}(V)$.

Now consider the case when $\lambda^* = 0$. From Lemma 12, it follows that, with probability at least $1-\delta/(4|\mathcal{S}||\mathcal{A}|)$, $\widehat{\lambda}^* = 0$ for $N \geq N'(\delta/(4|\mathcal{S}||\mathcal{A}|), c_r, P^o_{s,a})$, where N' is defined in Lemma 12. Thus, whenever $\lambda^* = 0$, we have $|\sigma_{\widehat{\mathcal{P}}_{s,a}}(V) - \sigma_{\mathcal{P}_{s,a}}(V)| = |V_{\min} - V_{\min}| = 0$, with probability at least $1-\delta/(4|\mathcal{S}||\mathcal{A}|)$. Thus having resolving this trivial case, we now focus on the case when $\lambda^* > 0$.

Consider the case when $\lambda^* > 0$. Let $\lambda_{\mathrm{kl}} := \lambda^*/2 > 0$ (dependent on $P^o_{s,a}, V$, and c_r but independent of N). Again from Lemma 12, if $\lambda^* \in [\lambda_{\mathrm{kl}}, 1/(c_r(1-\gamma))]$, then with probability at least $1-\delta/(4|\mathcal{S}||\mathcal{A}|)$ we have $\widehat{\lambda}^* \in [\lambda_{\mathrm{kl}}, 1/(c_r(1-\gamma))]$ for $N \geq N''(\delta/(4|\mathcal{S}||\mathcal{A}|), c_r, P^o_{s,a})$, where N'' is defined in Lemma 12.

From these arguments, it is clear that we can restrict the optimization problem (46) to the set $\lambda \in [\lambda_{\mathrm{kl}}, 1/(c_r(1-\gamma))$. Using this, with the additional fact that $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$, we get

$$|\sigma_{\widehat{\mathcal{P}}^{\mathrm{kl}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\mathrm{kl}}_{s,a}}(V)| \leq \max_{\lambda\in[\lambda_{\mathrm{kl}}, \frac{1}{c_r(1-\gamma)}]} |\lambda\log(\frac{\widehat{P}_{s,a}\exp(-V/\lambda)}{P^o_{s,a}\exp(-V/\lambda)})|. \tag{47}$$

Now,

$$\begin{aligned}
\left|\log(\frac{\widehat{P}_{s,a}\exp(-V/\lambda)}{P^o_{s,a}\exp(-V/\lambda)})\right| = \left|\log(1+\frac{(\widehat{P}_{s,a}-P^o_{s,a})\exp(-V/\lambda)}{P^o_{s,a}\exp(-V/\lambda)})\right| &\leq \frac{|(P^o_{s,a}-\widehat{P}_{s,a})\exp(-V/\lambda)|}{|P^o_{s,a}\exp(-V/\lambda)|} \\
&\overset{(d)}{\leq} \frac{|(P^o_{s,a}-\widehat{P}_{s,a})\exp(-V/\lambda)|}{\exp(\frac{-1}{\lambda_{\mathrm{kl}}(1-\gamma)})},
\end{aligned} \tag{48}$$

where (d) follows since $\lambda \geq \lambda_{\mathrm{kl}}$ and $\|V\| \leq 1/(1-\gamma)$. Using (48) in (47) along with the fact that $\lambda \leq 1/(c_r(1-\gamma))$, we get the desired result. □

Proof of Theorem 3. The basic steps of the proof is similar to that of Theorem 1. So, we present only the important steps.

Following the same steps as given before (25) and (29) , we get

$$\|V^* - \widehat{V}^*\| + \|V^{\pi_k} - \widehat{V}^{\pi_k}\| \leq \frac{2\gamma}{(1-\gamma)} \max_{V\in\mathcal{V}} \max_{s,a} |\sigma_{\widehat{\mathcal{P}}^{\mathrm{kl}}_{s,a}}(V) - \sigma_{\mathcal{P}^{\mathrm{kl}}_{s,a}}(V)|. \tag{49}$$

Similarly, following the steps as given before (28), we get

$$\|\widehat{V}^{\pi_k} - \widehat{V}^*\| \leq \frac{2\gamma^{k+1}}{(1-\gamma)^2}. \tag{50}$$

Using Lemma 13 in (49), we get

$$\|V^* - \widehat{V}^*\| + \|V^{\pi_k} - \widehat{V}^{\pi_k}\| \leq \frac{2\gamma}{(1-\gamma)} \frac{\exp(1/(\lambda_{\text{kl}}(1-\gamma)))}{c_r(1-\gamma)} \max_{s,a} \max_{V \in \mathcal{V}} \max_{\lambda \in [\lambda_{\text{kl}}, \frac{1}{c_r(1-\gamma)}]} |(P^o_{s,a} - \widehat{P}_{s,a}) \exp(-V/\lambda)|. \tag{51}$$

We now bound the max term in (51). We reparameterize $1/\lambda$ as θ and consider the set $\Theta = [c_r(1-\gamma), \frac{1}{\lambda_{\text{kl}}}]$. Also, consider the minimal η-cover $\mathcal{N}_\Theta(\eta)$ of Θ and fix a $V \in \mathcal{V}$. Then, for any given $\theta \in \Theta$, there exits a $\theta' \in \mathcal{N}_\Theta(\eta)$ such that $|\theta - \theta'| \leq \eta$. Now, for this particular θ, θ',

$$\begin{aligned}
&|(P^o_{s,a} - \widehat{P}_{s,a}) \exp(-V\theta)| = |(\widehat{P}_{s,a} - P^o_{s,a})(\exp(-V\theta') \circ \exp(-V(\theta - \theta'))| \\
&\overset{(c)}{\leq} |(\widehat{P}_{s,a} - P^o_{s,a}) \exp(-V\theta')| \exp(\eta/(1-\gamma)) \leq \max_{s,a} \max_{\theta' \in \mathcal{N}_\Theta(\eta)} |(\widehat{P}_{s,a} - P^o_{s,a}) \exp(-V\theta')| \exp(\eta/(1-\gamma)),
\end{aligned}$$

where (c) follows because V is non-negative and $\|V\| \leq 1/(1-\gamma)$. Now consider a minimal η-cover $\mathcal{N}_\mathcal{V}(\eta)$ of the set $\mathcal{V}$. By definition, there exists $V' \in \mathcal{N}_\mathcal{V}(\eta)$ such that $\|V - V'\| \leq \eta$. So, we get

$$\begin{aligned}
|(P^o_{s,a} - \widehat{P}_{s,a}) \exp(-V\theta)| &\leq |(\widehat{P}_{s,a} - P^o_{s,a}) \exp(-V\theta')| \exp(\eta/(1-\gamma)) \\
&= |(\widehat{P}_{s,a} - P^o_{s,a})(\exp(-V'\theta') \circ \exp(\theta'(V' - V)))| \exp(\eta/(1-\gamma)) \\
&\overset{(d)}{\leq} |(\widehat{P}_{s,a} - P^o_{s,a})(\exp(-V'\theta'))| \exp(\eta/(1-\gamma)) \exp(\eta/\lambda_{\text{kl}}) \\
&\leq \max_{s,a} \max_{V' \in \mathcal{V}} \max_{\theta' \in \mathcal{N}_\Theta(\eta)} |(\widehat{P}_{s,a} - P^o_{s,a})(\exp(-V'\theta'))| \exp(\eta/(1-\gamma)) \exp(\eta/\lambda_{\text{kl}})
\end{aligned}$$

where (d) follows because $\theta' \in \mathcal{N}_\Theta(\eta) \subseteq \Theta$. Now, taking maximum on both sides with respect to (s,a), θ, and V, we get

$$\begin{aligned}
\max_{s,a} \max_{\theta \in \Theta} \max_{V \in \mathcal{V}} |(\widehat{P}_{s,a} - P^o_{s,a}) \exp(-V\theta)| &\leq \exp(\eta/(1-\gamma)) \exp(\eta/\lambda_{\text{kl}}) \max_{s,a} \max_{V' \in \mathcal{V}} \max_{\theta' \in \mathcal{N}_\Theta(\eta)} |(\widehat{P}_{s,a} - P^o_{s,a}) \exp(-V'\theta')| \\
&\overset{(e)}{\leq} \exp(\eta/(1-\gamma)) \exp(\eta/\lambda_{\text{kl}}) \sqrt{\frac{\log(2|\mathcal{S}||\mathcal{A}||\mathcal{N}_\Theta(\eta)||\mathcal{N}_\mathcal{V}(\eta)|/\delta)}{2N}} \\
&\overset{(f)}{\leq} \exp(\eta/(1-\gamma)) \exp(\eta/\lambda_{\text{kl}}) \sqrt{\frac{|\mathcal{S}| \log(18|\mathcal{S}||\mathcal{A}|/(\delta\eta^2(1-\gamma)\lambda_{\text{kl}}))}{2N}}
\end{aligned} \tag{52}$$

with probability greater than $1-\delta$. Here, (e) follows from Lemma 9 with a union bound accounting for $|\mathcal{N}_\Theta(\eta)|$, $|\mathcal{N}_\mathcal{V}(\eta)|$ and the fact that $\|\exp(-V'\theta')\| \leq 1$, and (f) follows from Lemmas 7 and 8.

Using (49) - (52), we get, with probability greater than $1-\delta$,

$$\begin{aligned}
\|V^* - V^{\pi_k}\| \leq &\frac{2\gamma^{k+1}}{(1-\gamma)^2} + \\
&\frac{2\gamma}{(1-\gamma)} \frac{\exp(1/(\lambda_{\text{kl}}(1-\gamma)))}{c_r(1-\gamma)} \exp(\eta/(1-\gamma)) \exp(\eta/\lambda_{\text{kl}}) \sqrt{\frac{|\mathcal{S}| \log(18|\mathcal{S}||\mathcal{A}|/(\delta\eta^2(1-\gamma)\lambda_{\text{kl}}))}{2N}}.
\end{aligned}$$

We can now choose k, ε, η to make each of the term on the RHS of the above inequality small. In particular, choosing $\eta = 1$, $\varepsilon \in (0, 1/(1-\gamma))$, and k, N satisfying the conditions

$$\begin{aligned}
&k \geq K_0 = \frac{1}{\log(1/\gamma)} \cdot \log\left(\frac{4}{\varepsilon(1-\gamma)^2}\right) \qquad \text{and} \\
&N \geq N^{\text{kl}} = \max \Bigg\{ \max_{s,a} N'(\delta/(4|\mathcal{S}||\mathcal{A}|), c_r, P^o_{s,a}), \ \max_{s,a} N''(\delta/(4|\mathcal{S}||\mathcal{A}|), c_r, P^o_{s,a}), \\
&\qquad\qquad \frac{8\gamma^2|\mathcal{S}|}{c_r^2(1-\gamma)^4\varepsilon^2} \exp\left(\frac{4+2\lambda_{\text{kl}}}{\lambda_{\text{kl}}(1-\gamma)}\right) \log\left(\frac{18|\mathcal{S}||\mathcal{A}|}{\delta\lambda_{\text{kl}}(1-\gamma)}\right) \Bigg\},
\end{aligned}$$

we get $\|V^* - V^{\pi_k}\| \leq \varepsilon$ with probability greater than $1-\delta$. □

B.5 Proof of Theorem 4

Proof. We consider the deterministic MDP $(\mathcal{S}, \mathcal{A}, r, P^o, \gamma)$ shown in Fig.3 to be the nominal model. We fix $\gamma \in (0.01, 1]$ and $s_1 = 0$. The state space is $\mathcal{S} = \{0, 1\}$ and action space is $\mathcal{A} = \{a_l, a_r\}$, where a_l denotes 'move left' and a_r denotes 'move right' action. Reward for state 1 and action a_r pair is $r(1, a_r) = 1$, for state 0 and action a_r pair is $r(0, a_r) = -100\gamma/99$, and the reward is 0 for all other (s, a). Transition function P^o is deterministic, as indicated by the arrows.

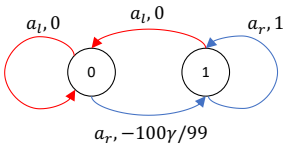

Figure 3: Transitions and rewards corresponding to the nominal model P^o. The states $\{0, 1\}$ are given inside the circles, and the actions $\{a_l, a_r\}$ and associated rewards are given on the corresponding transitions.

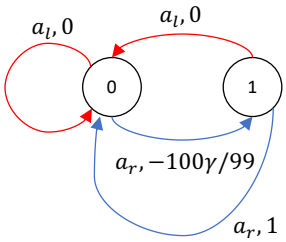

Figure 4: Transitions and rewards corresponding to the model P'.

Similarly, we consider another deterministic model P', as shown in Fig.4. We consider the set $\mathcal{P} = \{P^o, P'\}$.

It is straight forward to show that taking action a_r in any state is the optimal non-robust policy π^o corresponding to the nominal model P^o. This is obvious if for state $s = 1$. For $s = 0$, notice that taking action a_l will give a value zero and taking action a_r will give a value $\frac{\gamma}{1-\gamma} - \frac{100\gamma}{99}$. Since $\gamma > 0.01$, taking action a_r will give a positive value and hence is optimal. So, we get

$$V_{\pi^o, P^o}(0) = \frac{\gamma}{1-\gamma} - \frac{100\gamma}{99}.$$

We can now compute $V_{\pi^o, P'}(0)$ using the recursive equation

$$V_{\pi^o, P'}(0) = -\frac{100\gamma}{99} + \gamma + \gamma^2 V_{\pi^o, P'}(0).$$

Solving this, we get $V_{\pi^o, P'}(0) = -\gamma/(99(1-\gamma^2))$.

Now the robust value of π^o is given by

$$V^{\pi^o}(0) = \min\{V_{\pi^o, P^o}(0), V_{\pi^o, P'}(0)\} = -\gamma/(99(1-\gamma^2)).$$

We will now compute the optimal non-robust value from state 0 of model P'.

$$\begin{aligned}\max_{\pi} V_{\pi, P'}(0) = \max\{&V_{(\pi(0)=a_r, \pi(1)=a_r), P'}(0),\ V_{(\pi(0)=a_l, \pi(1)=a_l), P'}(0),\\ &V_{(\pi(0)=a_r, \pi(1)=a_l), P'}(0),\ V_{(\pi(0)=a_l, \pi(1)=a_r), P'}(0)\}\\ = \max\{&-\frac{\gamma}{99(1-\gamma^2)},\ 0,\ -\frac{100\gamma}{99(1+\gamma^2)},\ 0\} = 0.\end{aligned}$$

Now, we find the optimal robust value $V^*(0)$. From the perfect duality result of robust MDP (Nilim and El Ghaoui, 2005, Theorem 1), we have

$$V^*(0) = \min\{\max_{\pi} V_{\pi, P^o}(0), \max_{\pi} V_{\pi, P'}(0)\} = \min\{V_{\pi^o, P^o}(0), \max_{\pi} V_{\pi, P'}(0)\} = 0.$$

We finally have

$$V^*(0) - V^{\pi^o}(0) = \frac{\gamma}{99(1-\gamma^2)} \geq \frac{\gamma}{198(1-\gamma)},$$

where the inequality follows since $1 + \gamma \leq 2$. Thus, setting $c = \gamma/198$ and $\gamma_o = 0.01$, completes the proof of this theorem. □