Improved Sample Complexity of Robust Reinforcement Learning with KL Uncertainty Sets

Yudan Wang
University of Central Florida
yudan.wang@ucf.edu

Zilong Deng
Arizona State University
zdeng43@asu.edu

Nathaniel D. Bastian
Johns Hopkins University
ndbastian@jhu.edu

Shaofeng Zou
Arizona State University
zou@asu.edu

Abstract

In this paper, we study distributionally robust reinforcement learning with Kullback–Leibler (KL) divergence-defined uncertainty sets. The goal is to find a policy that maximizes the robust value function, defined as the worst-case value over all transition kernels in the uncertainty set. Assuming access to a generative model, we aim to understand the sample complexity of finding an ϵ-optimal robust policy. The best-known sample complexity results in the literature show a non-trivial gap of $\mathcal{O}\{\max\{p_\wedge^{-1}(1-\gamma)^{-1}, (1-\gamma)^{-2}\}\}$ between the upper and lower bounds, where $p_\wedge$ denotes the minimum positive transition probability of the nominal transition kernel, and γ is the discount factor. More importantly, existing results on the lower bound only cover a limited range of the uncertainty level. In this paper, we develop tighter and complete upper and lower bounds for robust RL with KL-defined uncertainty sets. Our upper bound is the tightest among all existing studies, and it improves upon the best-known bound by at least the order of $\mathcal{O}(\min\{p_\wedge^{-1}, (1-\gamma)^{-1}\})$ for uncertainty level $\delta \in (0, 1)$. Furthermore, our lower bound holds for any uncertainty level. Our upper and lower bounds (nearly) match in various regimes, providing improved results for this problem.

1 Introduction

Reinforcement learning (RL) [26] provides a powerful framework for agents to learn optimal policies through interaction with an environment, and it has achieved remarkable successes in recent years such as beating human champions in the game of Go [24, 25], achieving grandmaster level in video games [27, 17, 7], autonomous driving [11], robotics [12], and large language models [6]. However, existing RL approaches usually assume that a learned policy will be deployed in the same environment as the one in which it was trained. Such an assumption is often violated in practice due to, e.g., adversarial perturbations [21], modeling errors between the simulator and real-world applications [37], non-stationary environments [19], and a limited amount of training data [32]. The discrepancy between the training and test environments leads to a model mismatch, resulting in a notable decline in performance and limiting the suitability of RL in crucial domains.

Noteworthy efforts have been made to develop distributionally robust RL approaches to address the above challenge. The basic idea is to take a conservative approach in the face of model mismatch/uncertainty. An uncertainty set of models is constructed, e.g., one consisting of models close to a simulator, and the robust policy is obtained by optimizing the worst-case performance over the uncertainty set.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The robust planning setting where the uncertainty set is fully known has been well studied in the literature [9, 18, 31]. However, the statistical sample complexity of learning a robust policy in the learning setting, where only samples from the nominal transition kernel are available, has not been well understood until recently. We note that robust RL problems with different uncertainty sets are fundamentally different, and in this paper we focus on the uncertainty set defined by Kullback–Leibler (KL) divergence:

$$\mathcal{P}(\delta) = \{P : \mathrm{KL}(P\|P^0) \leq \delta\}, \tag{1}$$

where KL denotes the KL divergence, δ denotes the uncertainty level and P^0 is the nominal transition kernel. In the learning setting, P^0 is not known, and samples from P^0 are available. The goal is to optimize the worst-case value function over the uncertainty set $\mathcal{P}(\delta)$. Other divergence measures to construct the uncertainty set have also been studied in the literature, e.g., total variation, χ^2, R-contamination, Wasserstein distance, and we will discuss these results in related work.

There have been several studies on the sample complexity of robust RL with KL-defined uncertainty set. So far, the best-known upper bound in the literature is $\widetilde{\mathcal{O}}\left(p_\wedge^{-1}(1-\gamma)^{-4}\epsilon^{-2}|\mathcal{S}||\mathcal{A}|\right)$ [15], where $p_\wedge$ denotes the minimum positive transition probability of the nominal transition kernel P^0 (see definition in Section 2), γ denotes the discount factor, $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, and ϵ denotes the optimality gap. However, when the uncertainty level $\delta \to 0$, the problem shall reduce to a standard RL problem, whose minimax optimal sample complexity shall be $\widetilde{\Theta}\left((1-\gamma)^{-3}\epsilon^{-2}|\mathcal{S}||\mathcal{A}|\right)$ [28, 13, 1]. Therefore, the best-known bound may not be tight when δ is small (also see Figure 1). Moreover, as we summarized in Table 5, existing results on the lower bound only cover a limited range of the uncertainty level δ, and there exists a non-trivial gap of at least $\mathcal{O}(\max\{p_\wedge^{-1}(1-\gamma)^{-1}, (1-\gamma)^{-2}\})$ between the existing upper and lower bounds.

In this paper, we develop tighter and complete upper and lower bounds for robust RL with KL-defined uncertainty sets.

- We first establish an upper bound of $\widetilde{\mathcal{O}}\big((1-\gamma)^{-3}\epsilon^{-2}|\mathcal{S}||\mathcal{A}|\big(1+\min\{\exp(\delta/p_\wedge), \sqrt{\delta}/p_\wedge, 1/p_\wedge\}+(1-\gamma)^{-1}\min\{\sqrt{\delta}, 1\}\big)\big)$ for $\delta \in (0,1)$, which is the state-of-the-art (SOTA) in the literature. This result improves upon the best-known bound by at least an order of $\mathcal{O}(\min\{p_\wedge^{-1}, (1-\gamma)^{-1}\})$. Moreover, when the uncertainty level is small, i.e., $\delta \in (0, \min\{(1-\gamma)^2, p_\wedge\})$, our complexity order reduces to the minimax optimal sample complexity for standard MDP [1].
- We further provide a lower bound of $\widetilde{\Omega}((1-\gamma)^{-3}\epsilon^{-2}|\mathcal{S}||\mathcal{A}|\max\left\{1, \delta e^\delta/(1-\gamma), \delta^2(1-\gamma)/p_\wedge\right\})$ for any $\delta \in [1-\gamma, -\log(p_\wedge))$. Moreover, for $\delta \in [-\log(p_\wedge), \infty)$, we prove that this KL-defined robust RL problem reduces to a worst-path RL problem [5, 4], and the lower bound is then $\Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{p_\wedge}\right)$. Together with the lower bound results in [22] for $\delta < 1-\gamma$, these results provide the first complete lower bound characterization that applies to any uncertainty level δ. For any uncertainty level, the gap between our upper and lower bounds is at most $\mathcal{O}(\max\{(1-\gamma)^{-1}, (1-\gamma)/p_\wedge\})$.
- For a large uncertainty level $\delta \in [-\log(p_\wedge), \infty)$, we further characterize the minimax optimal sample complexity $\widetilde{\Theta}\left(\frac{|\mathcal{S}||\mathcal{A}|}{p_\wedge}\right)$ by proving that the problem is equivalent to a worst-path RL problem [5, 4].
- We comprehensively discuss the tightness of our upper and lower bounds, and show that in various regimes the upper and lower bounds match (up to logarithmic terms), and further quantify the gap for regimes where the bounds do not match. See Figure 1 for a complete schematic illustration. Overall, our results provide a near-minimax-optimal characterization of the sample complexity for this problem.
- We develop a novel analysis of the upper bounds. Specifically, we bound the variance induced by the distributionally robust optimization problem using the variance of the value function under worst-case transitions. This approach can be extended to other distributionally robust problems defined with similar uncertainty sets.

1.1 Related Work

In this section, we review related work on the sample complexity of robust RL.

REFERENCE	UPPER BOUND	LOWER BOUND	COMMENTS
[20]	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert e^{\frac{1}{1-\gamma}}}{(1-\gamma)^4\epsilon^2}\right)$	-	
[35]	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert e^{\frac{1}{1-\gamma}}}{(1-\gamma)^4\delta^2\epsilon^2}\right)$	-	EPISODIC MDPS
[36]	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^4 p_\wedge^2\epsilon^2}\right)$	-	
[22]	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge\delta(1-\gamma)^4\epsilon^2}\right)$	$\widetilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right), \delta\in(0,1-\gamma)$ $\widetilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge(1-\gamma)^2\epsilon^2}\right), \delta\sim\log(\frac{1}{p_\wedge})$	
[29]	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge^6(1-\gamma)^5\epsilon^2}\right)$	-	MODEL-FREE
[30]	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge^3(1-\gamma)^5\epsilon^2}\right)$	-	
[34]	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge(1-\gamma)^5\epsilon^2}\right)$	-	
[30] (VR)	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge^3(1-\gamma)^4\epsilon^2}\right)$	-	
[15]	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge(1-\gamma)^4\epsilon^2}\right)$	-	s-REC
OUR WORK	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\frac{\min\{p_\wedge e^{\delta/p_\wedge},1\}}{p_\wedge}$ $+\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\frac{\min\{\sqrt{\delta},1\}e^{\delta}}{1-\gamma})$ $\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\right), \delta\geq\log(\frac{1}{p_\wedge})$[1]	$\widetilde{\Omega}\left(\frac{\delta^2\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge(1-\gamma)^2\epsilon^2}\right), \delta\in(1-\gamma,\log(\frac{1}{p_\wedge}))$ $\widetilde{\Omega}\left(\frac{\delta e^{\delta}\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right), \delta\in(1-\gamma,\log(\frac{1}{p_\wedge}))$ $\Omega\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\right), \delta\geq\log(\frac{1}{p_\wedge})$[1]	NEAR-MINIMAX

Table 1: Upper and Lower Bounds under KL Uncertainty Set. VR: Variance reduction. s-rec: s-rectangular. [1]: 1) When $\delta\geq\log(1/p_\wedge)$, we prove that the KL-defined robust RL reduces to a worst-path RL problem and admits minimax-optimal sample complexity of $\widetilde{\Theta}(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge})$ as shown in [4]. 2) When $\delta\in(0,1-\gamma)$, we apply the lower bound result of $\widetilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$ established in [22].

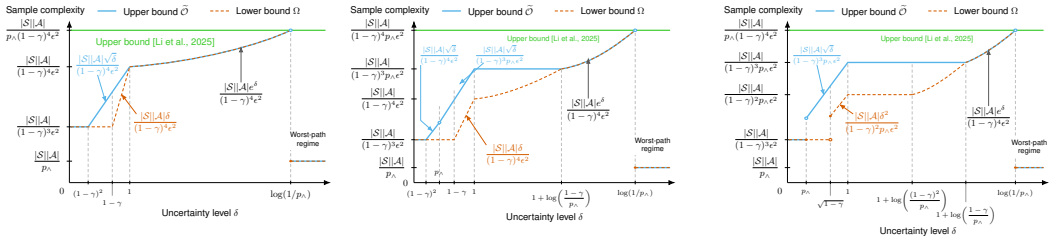
Standard RL. For standard RL, the minimax optimal sample complexity has been derived, $\widetilde{\mathcal{O}}\left((1-\gamma)^{-3}\epsilon^{-2}\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\right)$, for model-based approaches in [1] and for Q-learning-type algorithms [13, 28].

Robust RL with KL-defined Uncertainty Set. Sample complexity upper bounds for model-based algorithms have been developed in [20, 36, 15, 22, 38, 35]. The detailed comparisons are summarized in Table 1. Early results have an $\exp((1-\gamma)^{-1})$ factor in the upper bound. With a deeper understanding of the KL-defined robust RL problem, the bounds have been progressively improved. The current best-known upper bound is [15] $\widetilde{\mathcal{O}}(p_\wedge^{-1}(1-\gamma)^{-4}\epsilon^{-2}\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert)$. However, when compared with the minimax lower bound [22], there remains at least a gap of $\mathcal{O}(\max\{p_\wedge^{-1}(1-\gamma)^{-1},(1-\gamma)^{-2}\})$. Compared to these existing results, our upper and lower bounds are much tighter, and they (nearly) match each other in various regimes.

There are also several studies focusing on model-free algorithms [29, 30, 34]. These model-free methods generally require larger sample sizes than their model-based counterparts (see Table 1). Although our result is presented in the model-based setting, our analysis can be adapted to model-free algorithms, potentially leading to improved sample complexity.

Other Uncertainty Sets. (1) For robust RL with uncertainty sets defined by total variation (TV), several works have developed sample complexity bounds, e.g., [20, 36, 2, 23, 34]. In particular, [23] established the minimax optimal sample complexity of $\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2\epsilon^2}\min\left\{\frac{1}{\delta},\frac{1}{1-\gamma}\right\}\right)$. (2) For robust RL with χ^2-defined uncertainty sets, there are a few studies, e.g., [20, 36, 23, 34]. Notably, [23] provided a near-optimal result, where the upper bound of $\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\left(1+\frac{\max\{\delta,\sqrt{\delta}\}}{1-\gamma}\right)\right)$ and lower bound of $\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\left(1+\frac{\delta}{1-\gamma}\right)\right)$ match within a certain range of the uncertainty level. (3) For robust RL with uncertainty sets defined by L_p distance, [3] presented a minimax optimal complexity of $\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2\epsilon^2}\min\left\{\frac{1}{\delta},\frac{1}{1-\gamma}\right\}\right)$. (4) For robust RL with R-contamination uncertainty

sets, [33] presented an upper bound of $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-5}\epsilon^{-2})$ on the sample complexity. (5) There are also studies on robust RL with Wasserstein distance-defined uncertainty sets, e.g., [16, 35]. We emphasize that robust RL problems with different uncertainty sets are fundamentally different, and therefore, these results for other uncertainty sets may not be used to solve robust RL with KL-defined uncertainty sets studied in this paper.



(a) $p_\wedge \geq 1-\gamma$. (b) $(1-\gamma)^2 \leq p_\wedge < 1-\gamma$. (c) $0 < p_\wedge < (1-\gamma)^2$.

Figure 1: Illustrations of the obtained sample complexity upper and lower bounds under KL-defined uncertainty sets

2 Problem Formulation and Preliminaries

Distributionally robust RL is typically modeled mathematically as a distributionally robust Markov decision process (DR-MDP). A DR-MDP can be represented by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, \mathcal{P}, \gamma)$. Here, $\mathcal{S}, \mathcal{A}$ are the state and action spaces; $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function; and γ is the discount factor. Different from standard MDPs, $\mathcal{P}$ denotes the uncertainty set of transition kernels. Specifically, in this paper, we consider the following (s,a)-rectangular uncertainty set defined by KL divergence: $\mathcal{P}(\delta) = \prod_{s,a} \mathcal{P}_{s,a}(\delta) = \prod_{s,a}\{P_{s,a} : \mathrm{KL}(P_{s,a}\|P^0_{s,a}) \leq \delta\}$, where KL is the KL divergence: $\mathrm{KL}(\mu\|\mu_0) = \mathbb{E}_\mu[\log(\mu/\mu_0)]$ for any two probability distributions μ and μ_0. Here, the uncertainty set is centered at a nominal transition kernel P^0 and $P^0_{s,a}$ denotes the nominal transition probability distribution conditioned on s, a. Define the minimal non-zero probability of the nominal transition kernel as

$$p_\wedge = \min_{s',s,a:P^0(s'|s,a)>0} P^0(s'|s,a). \tag{2}$$

A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ denotes the probability of taking actions under different states and represents the strategy of the agent. Here, $\Delta(\mathcal{A})$ denotes the probability simplex over $\mathcal{A}$. Given the transition kernel P, the value function of a policy π is defined as the expected cumulative reward the agent receives by following the policy starting from s: $V^\pi_P(s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s, P, \pi\right]$. The Q-function is defined as the expected cumulative reward starting from s and action a: $Q^\pi_P(s,a) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s, a_0 = a, P, \pi\right]$.

The goal in a DR-MDP is to optimize the worst-case performance over the uncertainty set, i.e., to optimize the robust value function:

$$\pi^* = \arg\max_\pi V^\pi = \arg\max_\pi \inf_{P\in\mathcal{P}(\delta)} V^\pi_P, \tag{3}$$

where $V^\pi = \inf_{P\in\mathcal{P}(\delta)} V^\pi_P$ denotes the worst-case value function of policy π. Let $V^* = V^{\pi^*}$ be the optimal robust value function. Define the robust state-action value function: $Q^\pi(s,a) = \inf_{P\in\mathcal{P}(\delta)} Q^\pi_P(s,a)$. The optimal robust state-action value function is hence denoted by

$$Q^*(s,a) = \sup_\pi Q^\pi(s,a), \tag{4}$$

and the optimal robust policy can be directly obtained: $\pi^*(s) = \arg\max_a Q^*(s,a)$. It was shown in [9] that the optimal robust Q-function satisfies the following robust Bellman equation:

$$Q^*(s,a) = r(s,a) + \gamma \inf_{P\in\mathcal{P}(\delta)} \mathbb{E}_P\left[\max_{a'\in\mathcal{A}} Q^*(s',a')\right]. \tag{5}$$

Besides, the robust Bellman optimality operator is defined as

$$\mathcal{T}(Q(s,a)) = r(s,a) + \gamma \inf_{P \in \mathcal{P}(\delta)} \mathbb{E}_{s' \sim P_{s,a}}[V(s')], \tag{6}$$

where $V(s') = \max_{a'} Q(s', a')$.

2.1 Strong Duality

For a general uncertainty set $\mathcal{P}$, directly computing $\inf_{p \in \mathcal{P}} p^\top V$ for any vector V is usually computationally expensive. However, this optimization problem can be equivalently solved using its dual, which is a convex optimization problem [9, 8]. These results play a crucial role in our algorithm design; therefore, we introduce the dual form corresponding to the KL-defined uncertainty set as follows.

Lemma 2.1 ([9, 8])**.** *The optimization problem*

$$\min \quad \mathbb{E}_{s \sim q}[V(s)] \qquad \textit{subject to} \quad q \in \{KL\,(q \| p) \leq \delta, q \in \Delta(\mathcal{S})\}\,, \tag{7}$$

is equivalent to

$$\max_{\lambda \geq 0} \left\{ -\lambda \log \left(\mathbb{E}_{s \sim p} \left[\exp\left(-\frac{V(s)}{\lambda} \right) \right] \right) - \lambda \delta \right\}. \tag{8}$$

2.2 Generative Model and Empirical Transition Kernel

In this work, we study the sample complexity of robust RL with a KL-defined uncertainty set in the generative model framework introduced in [10]. This is a widely used sampling mechanism for understanding the fundamental sample complexity of RL problems [20, 36, 22, 1]. In the generative model setting, we assume sampling access to a generative model of the nominal transition kernel P^0, and we can query the next state s' sampled from $P^0_{s,a}$, given any state-action pair (s, a). Specifically, we sample N i.i.d. samples from $P^0_{s,a}$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$s_{n,s,a} \overset{i.i.d.}{\sim} P^0(\cdot|s,a), n = 1, 2, ..., N, \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Then, the total number of samples is $N|\mathcal{S}||\mathcal{A}|$. Using samples from the generative model, we compute the empirical estimate $\widehat{P}^0 = \prod_{s,a} \widehat{P}^0(\cdot|s,a)$ of the nominal transition kernel using a maximum likelihood estimate as follows:

$$\widehat{P}^0(s'|s,a) := \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left\{ s_{n,s,a} = s' \right\}. \tag{9}$$

Here, $\mathbb{I}$ denotes the indicator function. In the next section, we will use the empirical estimate of the nominal transition kernel in the robust value iteration algorithm [9, 18] and understand its sample complexity.

Notation. Throughout this paper, $\mathcal{O}(\cdot)$ denotes an upper bound up to universal constant factors. Specifically, given functions f and g defined on a domain D, we write $f(x) = O(g(x))$ if there exists a constant $M > 0$ such that $|f(x)| \leq Mg(x)$ for all $x \in D$. Besides, $\widetilde{\mathcal{O}}(\cdot)$ denotes an upper bound up to universal constant and logarithmic factors. Similarly, $\Omega(\cdot)$ denotes a lower bound up to universal constant factors, and $\widetilde{\Omega}(\cdot)$ denotes a lower bound up to universal constant and logarithmic factors. Moreover, $\Theta(\cdot)$ denotes a tight (minimax-optimal) bound up to universal constant factors, that is, $f(x) = \Theta(g(x))$ if $f(x) = \mathcal{O}(g(x))$ and $f(x) = \Omega(g(x))$. Similarly, $\widetilde{\Theta}(\cdot)$ denotes a tight bound up to universal constant and logarithmic factors.

3 Main Results

We use the standard robust value iteration algorithm with the empirically estimated uncertainty set. Firstly, from the generative model, we draw N samples from every state-action pair, and we can get the empirical nominal transition kernel $\widehat{P}^0(\cdot|s,a)$ for $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$. Based on the empirical

Algorithm 1 Robust Value Iteration Algorithm

Initialization: step size β, iteration number T,sample count N, and $Q = 0$
Sample N i.i.d. samples from $P^0(\cdot|s,a), \forall(s,a)$
Compute the empirical transition kernel $\widehat{P}^0(\cdot|s,a)$ according to Eq. 9
for $t = 0, ..., T-1$ **do**
  Update the robust Q-function: $Q(s,a) \leftarrow \hat{\mathcal{T}}(Q(s,a))$ according to Eq. 12, $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$
end for
Output: $\pi_T(\cdot|s) = \arg\max_a Q(s,a), \forall s \in \mathcal{S}$

nominal transition kernel, we further define the empirical uncertainty set, where we replace the unknown true nominal transition kernel P^0 with its empirical estimate:

$$\widehat{\mathcal{P}}_{s,a} = \{q : \mathrm{KL}(q \mid \widehat{P}^0_{s,a}) \leq \delta\}, \qquad \text{and } \widehat{\mathcal{P}} = \prod_{s,a} \widehat{\mathcal{P}}_{s,a}. \tag{10}$$

Then, we further define the empirical robust Bellman optimality operator as follows:

$$\widehat{\mathcal{T}}(Q(s,a)) = r(s,a) + \gamma \inf_{q \in \widehat{\mathcal{P}}_{s,a}(\delta)} \mathbb{E}_q[V(s')], \tag{11}$$

where $V(s') = \max_{a'} Q(s',a')$. By Lemma 2.1, the robust Bellman optimality operator can be computed as follows:

$$\widehat{\mathcal{T}}(Q(s,a)) = r(s,a) + \gamma \max_{\lambda \geq 0} \left\{ -\lambda \log \left(\mathbb{E}_{\widehat{P}^0(\cdot|s,a)} \left[\exp\left(-\frac{V}{\lambda} \right) \right] \right) - \lambda\delta \right\}. \tag{12}$$

In Algorithm 1, we present the robust value iteration algorithm with an empirically estimated uncertainty set. We note that the convergence of the robust value iteration algorithm has been shown in prior works [9, 18] for any given uncertainty set. In the following, our focus is on understanding how many samples are needed to guarantee that the solution to Algorithm 1 using the empirical uncertainty set is close to the optimal robust policy defined w.r.t. the original uncertainty set in eq. (3).

3.1 Sample Complexity Upper Bound

We first present the upper bound on the sample complexity of Algorithm 1.

Theorem 3.1. *Let π_T be the output of Algorithm 1 with uncertainty level $0 < \delta \leq 1$, $T = \log(\frac{N}{1-\gamma})$ and $V_T = V^{\pi_T}$. Then, with probability at least $\eta \in (0,1)$, the gap between the optimal robust policy and output policy can be bounded by*

$$\|V^* - V_T\|_\infty \leq \mathcal{O}\left((1-\gamma)^{-3/2} \frac{1}{\sqrt{N}} \sqrt{1 + \min\left\{ \exp\left(\frac{\delta}{p_\wedge}\right), \frac{\sqrt{\delta}}{p_\wedge}, \frac{1}{p_\wedge} \right\} + \frac{\min\{\sqrt{\delta}, 1\}}{1-\gamma}} \right).$$

To obtain an ϵ-optimal policy, i.e. $\|V^* - V_T\|_\infty \leq \epsilon$, the total sample complexity is

$$\mathcal{O}\left((1-\gamma)^{-3} \epsilon^{-2} |\mathcal{S}||\mathcal{A}| \log\left(\frac{|\mathcal{S}||\mathcal{A}|N}{(1-\gamma)\eta} \right) \left(1 + \min\left\{ \exp\left(\frac{\delta}{p_\wedge}\right), \frac{\sqrt{\delta}}{p_\wedge}, \frac{1}{p_\wedge} \right\} + \frac{\min\{\sqrt{\delta}, 1\}}{1-\gamma} \right) \right).$$

As we discussed in the introduction, existing upper bound results have a gap of $\mathcal{O}(p_\wedge^{-1}(1-\gamma)^{-1})$ compared to the minimax lower bound of standard RL when the uncertainty level is small. The intuition is that the problem shall reduce to the standard RL setting when the uncertainty level is small (see Figs. 1). In contrast, our sample complexity reduces to the minimax-optimal sample complexity of standard RL: $\mathcal{O}((1-\gamma)^{-3}\epsilon^{-2}|\mathcal{S}||\mathcal{A}|)$ [1, 28, 13], when the uncertainty level is small.

Corollary 3.2. *When the uncertainty level δ is small: $\delta \in (0, \min\{(1-\gamma)^2, p_\wedge\})$, the sample complexity of Algorithm 1 reduces to the minimax-optimal sample complexity of standard RL:*

$$\mathcal{O}((1-\gamma)^{-3}\epsilon^{-2}|\mathcal{S}||\mathcal{A}|). \tag{13}$$

Compared with the previous KL-defined robust RL works, our results establish a state-of-the-art (SOTA) upper bound which improves upon the best known bound by at least an order of $\mathcal{O}(\min\{(1-\gamma)^{-1}, p_\wedge^{-1}\})$ [20, 36, 15] when uncertainty level $\delta \leq 1$. Moreover, our work is the first to achieve the same sample complexity order as standard (non-robust) RL when the uncertainty level is sufficiently small.

Theorem 3.3 (Large uncertainty level $\delta > 1$)**.** *Let π_T be the output of Algorithm 1 with uncertainty level $1 \leq \delta \leq \log(\frac{1}{p_\wedge})$, $T = \log(\frac{N}{1-\gamma})$ and $V_T = V^{\pi_T}$. Then, with probability at least $\eta \in (0,1)$, the gap between the optimal robust policy and output policy can be bounded by*

$$\|V^* - V_T\|_\infty \leq \mathcal{O}\left((1-\gamma)^{-3/2}\frac{1}{\sqrt{N}}\sqrt{\frac{1}{p_\wedge} + \frac{e^\delta}{1-\gamma}}\right).$$

To obtain an ϵ-optimal policy, i.e. $\|V^* - V_T\|_\infty \leq \epsilon$, the total sample complexity is

$$\mathcal{O}\left((1-\gamma)^{-3}\epsilon^{-2}|\mathcal{S}||\mathcal{A}|\log\left(\frac{|\mathcal{S}||\mathcal{A}|N}{(1-\gamma)\eta}\right)\left(\frac{1}{p_\wedge} + \frac{e^\delta}{1-\gamma}\right)\right).$$

Here, we consider the regime $e^\delta < p_\wedge^{-1}$. When the uncertainty level satisfies $\delta \geq \log\left(p_\wedge^{-1}\right)$, the robust RL problem with KL uncertainty sets reduces to a worst-path problem; see the discussion in Section 3.3.

3.2 Sample Complexity Lower Bounds

To fully understand the improvement and optimality of the upper bound in Theorem 3.1, we derive minimax lower bounds. In this section, we focus on the case when the uncertainty level $\delta < \log(p_\wedge^{-1})$. The case when $\delta \geq \log(p_\wedge^{-1})$ will be studied in section 3.3.

Theorem 3.4. *Consider any $(\mathcal{S}, \mathcal{A}, \gamma, \delta, \epsilon)$ such that $\gamma \in (3/4, 1), \delta \in [1-\gamma, \log(1/p_\wedge))$, and $\epsilon < \frac{1}{1000(1-\gamma)(\delta+\log(2/(1-\gamma)))}$. Then there exist two DR-MDPs $\mathcal{M}_0, \mathcal{M}_1$ defined by the KL-defined uncertainty sets, and a dataset with N independent samples per (s,a) pair from the nominal transition kernel (for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively), such that*

$$\inf_{\widehat{\pi}} \max\left\{\mathbb{P}_0\left(||V_0^* - V_0^{\widehat{\pi}}||_\infty > \epsilon\right), \mathbb{P}_1\left(||V_1^* - V_1^{\widehat{\pi}}||_\infty > \epsilon\right)\right\} > \frac{1}{8}, \tag{14}$$

provided that the total number of samples satisfies

$$N|\mathcal{S}||\mathcal{A}| \leq \widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|\delta e^\delta}{(1-\gamma)^4\epsilon^2}\right).$$

Here, V_0^* (resp. V_1^*) denotes the optimal robust value function under $\mathcal{M}_0$ (resp. $\mathcal{M}_1$), $V_0^{\hat{\pi}}$ (resp. $V_1^{\hat{\pi}}$) denotes the robust value function of the policy returned by any learning algorithm $\hat{\pi}$ using samples from $\mathcal{M}_0$ (resp. $\mathcal{M}_1$). Moreover, $\mathbb{P}_0\left(||V_0^* - V_0^{\widehat{\pi}}||_\infty > \epsilon\right)$ (resp. $\mathbb{P}_1$) denotes the probability that the output policy $\widehat{\pi}$ fails to converge to the optimal robust policy using samples from the DR-MDP $\mathcal{M}_0$. Specifically, this event occurs when the gap between the optimal value function V_0^* and the value function induced by the algorithm's output policy $V_0^{\widehat{\pi}}$ exceeds ϵ in the ℓ_∞ norm. When eq. (14) holds, there exists at least one MDP in $\mathcal{M}_0, \mathcal{M}_1$ under which no algorithm can guarantee that the output policy converges to the optimal robust policy.

Theorem 3.4 provides a lower bound for KL-defined robust RL when uncertainty level $\delta \in [1-\gamma, -\log(p_\wedge))$. For $\delta \in (0, 1-\gamma)$, we borrow the lower bound of $\Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$ from [22]. Combining our results with the one in [22], we obtain a lower bound of $\Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\max\{\frac{\delta e^\delta}{1-\gamma}, 1\}\right)$ that holds for any $\delta \in (0, -\log(p_\wedge))$. We further observe that when the uncertainty level is large, i.e., $\delta = \mathcal{O}(1)$, a sample complexity of $\Omega\left(\frac{e^\delta|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\right)$ is necessary. This lower bound is strictly larger than the minimax-optimal bound for standard RL, implying that KL-defined robust RL is intrinsically more challenging than standard RL.

Theorem 3.5. *Under the same conditions as in Theorem 3.4 and $p_\wedge \leq \frac{1}{8}$, eq.* (14) *holds provided that the total number of samples satisfies*

$$N|\mathcal{S}||\mathcal{A}| \leq \widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|\delta^2}{(1-\gamma)^2 p_\wedge \epsilon^2}\right). \tag{15}$$

Theorem 3.4 and Theorem 3.5 are derived using two different constructions of hard MDPs, and therefore, they hold simultaneously. Combining Theorem 3.4, Theorem 3.5 and results from [22], we obtain the following minimax lower bound.

Corollary 3.6. *For any $\delta \in (1-\gamma, \log(1/p_\wedge))$, the overall minimax lower bound is given by*

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\max\left\{1, \frac{\delta e^\delta}{1-\gamma}, \frac{\delta^2(1-\gamma)}{p_\wedge}\right\}\right). \tag{16}$$

The lower bound above implies that when the uncertainty level remains on the order of a constant, even a sufficiently small minimum positive transition probability can trigger a large variance, substantially impeding the convergence of learning algorithms. This observation not only sharpens the theoretical picture but also offers a principled explanation for the performance degradation often seen in practice.

3.3 Worst-Path RL

In this section, we analyze the case when the uncertainty level $\delta \geq \log\left(p_\wedge^{-1}\right)$. Results here will provide insight about the fundamental difficulty of the problem when the uncertainty set is overly pessimistic, and the problem reduces to a worst-path RL problem [5, 4], in which case the agent aims to maximize the minimum possible cumulative reward. Moreover, together with the lower bound results in Section 3.2, our lower bound covers any value of δ.

Lemma 3.7. *For any $(s,a) \in \mathcal{S}\times\mathcal{A}$, if $\delta \geq \log\left(\frac{1}{p_\wedge}\right)$, then the worst-case transition kernel concentrates all probability mass on the worst-case next state:*

$$\inf_{P\in\mathcal{P}(\delta)} \mathbb{E}_{s'\sim P_{s,a}}[V(s')] = \min_{s':P(s'|s,a)>0} V(s').$$

Lemma 3.7 implies that when δ is large, the robust Bellman equation in eq. (5) reduces to $Q^*(s,a) = r(s,a) + \gamma \min_{s':P^0(s'|s,a)>0}\left(\max_{a'\in\mathcal{A}} Q^*(s',a')\right)$, which coincides with the worst-path RL Bellman equation (Equation (31)) in [4]. Therefore, when $\delta \geq \log\left(\frac{1}{p_\wedge}\right)$, the KL-defined DR-RL problem reduces to a worst-path RL problem.

In the following, we quote the minimax sample complexity results derived in [4] for the worst-path RL problem.

Theorem 3.8 ([4], Theorems 3 and 4)**.** *For a worst-path RL problem with $|\mathcal{S}|$ states, $|\mathcal{A}|$ actions, and minimum positive transition probability $p_\wedge$, the minimax-optimal sample complexity order is $\widetilde{\Theta}\left(\frac{|\mathcal{S}||\mathcal{A}|}{p_\wedge}\right)$.*

Combining Lemma 3.7 and Theorem 3.8, we have the following minimax sample complexity results for our KL-defined robust RL problem.

Corollary 3.9. *For any $\delta \geq \log\left(\frac{1}{p_\wedge}\right)$, the minimax optimal sample complexity for the KL-defined robust RL problem is $\widetilde{\Theta}\left(\frac{|\mathcal{S}||\mathcal{A}|}{p_\wedge}\right)$.*

Here, we note that the sample complexity lower bound established in [22] applies to uncertainty levels such that $\delta \in \left(\log(1/p_\wedge) - 6,\ \log(1/p_\wedge) - 5\right)$, and this set does not overlap with the set $[\log\left(\frac{1}{p_\wedge}\right), \infty)$ we consider here. Therefore, their results do not contradict ours. We also notice that the sample complexity in Theorem 3.9 does not depend on ϵ. As discussed in [5], this is because the worst-path RL problem only concerns the minimum value of successor states, which is independent of specific transition probabilities. Corollary 3.9 provides both upper and lower bounds for any $\delta \geq \log\left(\frac{1}{p_\wedge}\right)$. Combined with the lower-bound results in Section 3.2, we obtain the first complete characterization of the lower bound for any uncertainty level in the literature.

3.4 Tightness of Upper and Lower Bounds

In this section, we analyze the tightness of our upper and lower bounds. As will be shown below, for a wide range of uncertainty levels, our upper and lower bounds match with each other (up to logarithmic terms), and thus are minimax optimal. Moreover, for regimes where the bounds do not match, we also quantify the gap between the upper and lower bounds.

Case 1: $p_\wedge \geq 1-\gamma$**.** The upper bound is controlled by the standard-RL term and the term involving $\min\{\sqrt{\delta}, 1\}/(1-\gamma)$. Among the two new lower bounds, the one proportional to $|\mathcal{S}||\mathcal{A}|\delta/((1-\gamma)^4\epsilon^2)$ dominates. The rates in Table 2 match up to logarithmic factors for $\delta \leq (1-\gamma)^2$, for $1 \leq \delta < \log(1/p_\wedge)$, and in the worst-path regime. For $(1-\gamma)^2 < \delta < 1-\gamma$, their ratio is $\sqrt{\delta}/(1-\gamma)$ up to logarithmic factors; for $1-\gamma \leq \delta \leq 1$, it is $1/\sqrt{\delta}$. Thus the largest polynomial gap in this case is $(1-\gamma)^{-1/2}$.

Uncertainty level	Upper bound	Lower bound	Gap
$0 < \delta \leq (1-\gamma)^2$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}(1)$
$(1-\gamma)^2 < \delta < 1-\gamma$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\sqrt{\delta}}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}\Big(\frac{\sqrt{\delta}}{1-\gamma}\Big)$
$1-\gamma \leq \delta < 1$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\sqrt{\delta}}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\delta}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{\delta}}\Big)$
$1 \leq \delta < \log(\frac{1}{p_\wedge})$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert e^\delta}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert e^\delta}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}(1)$
$\delta \geq \log(\frac{1}{p_\wedge})$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\Big)$	$\widetilde{\mathcal{O}}(1)$

Table 2: Upper bounds, lower bounds, and gaps for $p_\wedge \geq 1-\gamma$. The factor δ is retained in the lower bound even when $1 \leq \delta < \log(1/p_\wedge)$; its size is then at most logarithmic in $1/p_\wedge$.

The analysis above shows that our upper and lower bounds are near minimax optimal when the minimum positive transition probability satisfies $p_\wedge \geq 1-\gamma$ (see Figure 1 (a)).

Case 2: $(1-\gamma)^2 \leq p_\wedge < 1-\gamma$**.** The dominant lower bound is the same as in Case 1. When $\delta \leq p_\wedge$, the exponential branch of the upper bound is bounded by an absolute constant. For $p_\wedge < \delta < 1$, the full minimum is retained in Table 3; the exponential branch need not become inactive immediately above $p_\wedge$. The rates match for $\delta \leq (1-\gamma)^2$ and in the worst-path regime. When $1 \leq \delta < \log(1/p_\wedge)$, the ratio of the displayed rate expressions is $(1-\gamma)/(\delta p_\wedge)$ before the logarithmic factors in the upper bound are restored.

Uncertainty level	Upper bound	Lower bound	Gap
$0 < \delta \leq (1-\gamma)^2$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}(1)$
$(1-\gamma)^2 < \delta \leq p_\wedge$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\sqrt{\delta}}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}\Big(\frac{\sqrt{\delta}}{1-\gamma}\Big)$
$p_\wedge < \delta < 1-\gamma$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\,\frac{\sqrt{\delta}}{p_\wedge}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\mathcal{O}\Big(\frac{\sqrt{\delta}}{p_\wedge}\Big)$
$1-\gamma \leq \delta < 1$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\,\frac{\sqrt{\delta}}{p_\wedge}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\delta}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}\Big(\frac{1-\gamma}{\sqrt{\delta}p_\wedge}\Big)$
$1 \leq \delta < 1+\log(\frac{1-\gamma}{p_\wedge})$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3 p_\wedge\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{e^\delta\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}\Big(\frac{1-\gamma}{p_\wedge e^\delta}\Big)$
$1+\log(\frac{1-\gamma}{p_\wedge}) \leq \delta < \log(1/p_\wedge)$	$\widetilde{\mathcal{O}}\Big(\frac{e^\delta\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{e^\delta\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}(1)$
$\delta \geq \log(1/p_\wedge)$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\Big)$	$\widetilde{\mathcal{O}}(1)$

Table 3: Upper bounds, lower bounds, and gaps for $(1-\gamma)^2 \leq p_\wedge < 1-\gamma$. Intermediate uncertainty levels retain all three branches of the upper bound.

Case 3: $p_\wedge < (1-\gamma)^2$**.** The minimum involving $p_\wedge$ controls the upper bound throughout $0 < \delta \leq 1$. For $\delta \geq 1-\gamma$, the dominant lower bound is now the one proportional to $|\mathcal{S}||\mathcal{A}|\delta/((1-\gamma)^2 p_\wedge\epsilon^2)$. Table 4 uses this support-dependent lower bound only on the range established by Theorem 3.5.

When $1 \le \delta < \log(1/p_\wedge)$, the ratio of the displayed rate expressions is $1/(\delta(1-\gamma))$ before restoring logarithmic factors.

Uncertainty level	Upper bound	Lower bound	Gap
$0 < \delta \le p_\wedge$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}(1)$
$p_\wedge < \delta < \sqrt{1-\gamma}$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\frac{\sqrt{\delta}}{p_\wedge}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}\Big(\frac{\sqrt{\delta}}{p_\wedge}\Big)$
$\sqrt{1-\gamma} \le \delta < 1$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\frac{\sqrt{\delta}}{p_\wedge}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\delta^2}{(1-\gamma)^2 p_\wedge\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}\Big(\frac{1}{\delta\sqrt{\delta}(1-\gamma)}\Big)$
$1 \le \delta < 1 + \log(\frac{(1-\gamma)^2}{p_\wedge})$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3 p_\wedge\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2 p_\wedge\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}\Big(\frac{1}{1-\gamma}\Big)$
$1 + \log(\frac{(1-\gamma)^2}{p_\wedge}) \le \delta < 1 + \log(\frac{1-\gamma}{p_\wedge})$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3 p_\wedge\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{e^\delta\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}\Big(\frac{1-\gamma}{p_\wedge e^\delta}\Big)$
$1 + \log(\frac{1-\gamma}{p_\wedge}) \le \delta < \log(1/p_\wedge)$	$\widetilde{\mathcal{O}}\Big(\frac{e^\delta\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\Omega}\Big(\frac{e^\delta\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\Big)$	$\widetilde{\mathcal{O}}(1)$
$\delta \ge \log(1/p_\wedge)$	$\widetilde{\mathcal{O}}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\Big)$	$\widetilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\Big)$	$\widetilde{\mathcal{O}}(1)$

Table 4: Upper bounds, lower bounds, and gaps for $p_\wedge < (1-\gamma)^2$. For $\delta < 1-\gamma$, the lower-bound entries use only the cited standard-RL result.

The cases above are illustrated in Figure 1 (b) and (c).

4 Conclusion

In this paper, we study the sample complexity of distributionally robust reinforcement learning under KL-defined uncertainty sets. We develop a comprehensive understanding of this problem with tighter upper bounds and, for the first time, establish minimax lower bounds that hold across the entire range of uncertainty levels. Our results close the gap between upper and lower bounds and recover the minimax-optimal rates of standard reinforcement learning in the vanishing-uncertainty regime. A key insight of our analysis is a precise characterization of how uncertainty level, discount factor, and minimum positive transition probability jointly govern the intrinsic difficulty of KL-defined robust RL. In particular, our lower bounds reveal that even under constant uncertainty, small support probabilities can induce large variance, fundamentally limiting learning efficiency. This phenomenon highlights a structural source of hardness that does not arise in standard RL and explains the degradation observed in robust settings. Beyond KL-defined uncertainty sets, our analytical techniques, especially the transfer of robustness-induced variance into value-function variance and the hard-MDP constructions used for lower bounds, may be of independent interest and applicable to other divergence-based robust optimization and robust RL frameworks.

Acknowledgements

The work of Y. Wang, Z. Deng and S. Zou was supported in part by DARPA under Agreement No. D25AP00191, and by the National Science Foundation under Grants ECCS-2438392 (CAREER), ECCS-2622695, and CCF-2438429. The work of Y. Wang was additionally supported by UCF ECE department.

References

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

[2] P. Clavier, E. Le Pennec, and M. Geist. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

[3] P. Clavier, L. Shi, E. Le Pennec, E. Mazumdar, A. Wierman, and M. Geist. Near-optimal distributionally robust reinforcement learning with general l_p norms. *Advances in Neural Information Processing Systems*, 37:1750–1810, 2024.

[4] Z. Deng, S. Khan, and S. Zou. Near-optimal sample complexity for iterated cvar reinforcement learning with a generative model. *arXiv preprint arXiv:2503.08934*, 2025.

[5] Y. Du, S. Wang, and L. Huang. Provably efficient risk-sensitive reinforcement learning: Iterated cvar and worst path. *arXiv preprint arXiv:2206.02678*, 2022.

[6] Y. Du, O. Watkins, Z. Wang, C. Colas, T. Darrell, P. Abbeel, A. Gupta, and J. Andreas. Guiding pretraining in reinforcement learning with large language models. In *International Conference on Machine Learning*, pages 8657–8677. PMLR, 2023.

[7] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[8] Z. Hu and L. J. Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

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

[10] M. J. Kearns. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

[11] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE transactions on intelligent transportation systems*, 23(6):4909–4926, 2021.

[12] J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[13] G. Li, C. Cai, Y. Chen, Y. Wei, and Y. Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236, 2024.

[14] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *arXiv preprint arXiv:2006.03041*, 2020.

[15] Z. Li, S. Wang, and N. Si. Near-optimal sample complexities of divergence-based s-rectangular distributionally robust reinforcement learning. *arXiv preprint arXiv:2505.12202*, 2025.

[16] S. Mandal, Y. Murthy, and R. Srikant. Finite-time bounds for distributionally robust td learning with linear function approximation. *arXiv preprint arXiv:2510.01721*, 2025.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing Atari with deep reinforcement learning. *NIPS Deep Learning Workshop*, 2013.

[18] A. Nilim and L. El Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 839–846, 2004.

[19] S. Padakandla, P. KJ, and S. Bhatnagar. Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, 50(11):3590–3606, 2020.

[20] K. Panaganti and D. Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR, 2022.

[21] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826, 2017.

[22] L. Shi and Y. Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

[23] L. Shi, G. Li, Y. Wei, Y. Chen, M. Geist, and Y. Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

[24] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

[25] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[26] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[27] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[28] M. J. Wainwright. Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

[29] S. Wang, N. Si, J. Blanchet, and Z. Zhou. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3370–3398. PMLR, 2023.

[30] S. Wang, N. Si, J. Blanchet, and Z. Zhou. Sample complexity of variance-reduced distributionally robust q-learning. *arXiv preprint arXiv:2305.18420*, 2023.

[31] Y. Wang, A. Velasquez, G. Atia, A. Prater-Bennette, and S. Zou. Robust average-reward markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15215–15223, 2023.

[32] Y. Wang, J. Xiong, and S. Zou. Achieving minimax optimal sample complexity of offline reinforcement learning: A dro-based approach. *Transactions on Machine Learning Research*, 2024.

[33] Y. Wang and S. Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

[34] Y. Wang, S. Zou, and Y. Wang. Model-free robust reinforcement learning with sample complexity analysis. In *Uncertainty in Artificial Intelligence*, pages 3470–3513. PMLR, 2024.

[35] Z. Xu, K. Panaganti, and D. Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9728–9754. PMLR, 2023.

[36] W. Yang, L. Zhang, and Z. Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

[37] W. Zhao, J. P. Queralta, and T. Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 737–744. IEEE, 2020.

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

Contents

1 Introduction **1**

1.1 Related Work 2

2 Problem Formulation and Preliminaries **4**

2.1 Strong Duality 5

2.2 Generative Model and Empirical Transition Kernel 5

3 Main Results **5**

3.1 Sample Complexity Upper Bound 6

3.2 Sample Complexity Lower Bounds 7

3.3 Worst-Path RL 8

3.4 Tightness of Upper and Lower Bounds 9

4 Conclusion **10**

A Robust RL with Other Uncertainty Sets **13**

B Limitation **13**

C Proof of Theorem 3.1: Upper Bound Analysis **14**

C.1 Dual representation and worst-case transitions 14

C.2 Error decomposition: statistical error and planning error 16

C.3 Support retention and a stable dual comparison 17

C.4 Concentration results 19

C.5 Variance transfer for small uncertainty 26

C.6 A resolvent bound for value variance 30

C.7 Sample-complexity analysis for small uncertainty level: $\delta \leq 1$. 31

C.8 Sample-complexity analysis for large uncertainty level: $\delta > 1$ 35

D Lower Bound Proof **43**

D.1 Hard MDP design and lower bound proof for Theorem 3.4 43

D.2 Proof of Theorem 3.5 48

E Proof of Lemma 3.7 **52**

A Robust RL with Other Uncertainty Sets

In this section, we further compare our results with other sample-complexity results for robust RL across various uncertainty sets.

B Limitation

In this section, we discuss the limitations of our work. This paper is primarily theoretical, and our results provide a comprehensive understanding of KL-based DR-RL. However, further effort is

REFERENCE	SETS	UPPER BOUND	LOWER BOUND	-
[28] [14] [1]	-	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	SAME	MINIMAX
[23]	TV	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2\epsilon^2}\min\left\{\frac{1}{\delta},\frac{1}{1-\gamma}\right\}\right)$	SAME	MINIMAX
[23]	χ^2	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\left(1+\frac{\max\{\delta,\sqrt{\delta}\}}{1-\gamma}\right)\right)$	$\widetilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\left(1+\frac{\delta}{1-\gamma}\right)\right)$	NEAR-MINIMAX
[3]	L_p	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2\epsilon^2\max\{1-\gamma,\delta\}}\right)$	SAME	MINIMAX
[22]	KL	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\delta p_\wedge(1-\gamma)^4\epsilon^2}\right)$	$\widetilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right),\delta\in(0,1-\gamma)$ $\widetilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge(1-\gamma)^2\epsilon^2}\right),\delta\sim\log(\frac{1}{p_\wedge})$	
[15]	KL	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge(1-\gamma)^4\epsilon^2}\right)$	-	s-REC
OUR WORK	KL	$\widetilde{\mathcal{O}}\big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\frac{\min\{p_\wedge e^{\delta/p_\wedge},1\}}{p_\wedge}$ $+\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\frac{e^\delta\min\{\sqrt{\delta},1\}}{1-\gamma}\big)$ $\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\right),\delta\geq\log(\frac{1}{p_\wedge})$[1]	$\widetilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge(1-\gamma)^2\epsilon^2}\right),\delta\in(1-\gamma,\log(\frac{1}{p_\wedge}))$ $\widetilde{\Omega}\left(\frac{\delta e^\delta\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right),\delta\in(1-\gamma,\log(\frac{1}{p_\wedge}))$ $\Omega\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{p_\wedge}\right),\delta\geq\log(1/p_\wedge)$[1]	NEAR-MINIMAX

Table 5: Upper and Lower Bounds: standard RL, TV, χ^2, l_p and KL defined uncertainty sets. [1]: 1) When $\delta\geq\log(1/p_\wedge)$, we prove that the KL-defined robust RL reduces to a worst-path RL problem and admits minimax-optimal sample complexity of $\widetilde{\Theta}(\frac{|\mathcal{S}||\mathcal{A}|}{p_\wedge})$ as shown in [4]. 2) When $\delta\in(0,1-\gamma)$, we apply the lower bound result of $\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$ established in [22].

required to develop exact minimax-optimal upper and lower bounds and to conduct simulations that validate these findings.

C Proof of Theorem 3.1: Upper Bound Analysis

In this section, we prove Theorem 3.1.

We first introduce the notation and supporting lemmas in Sections C.1 to C.6, which also provide insight into the structure of DR-RL with KL uncertainty sets. We then use these lemmas to complete the proof in Section C.7.

C.1 Dual representation and worst-case transitions

For a value vector V, recall the dual objective of strong duality in Lemma 2.1:

$$f(P^0_{s,a},V,\lambda)=-\lambda\log\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda}]-\lambda\delta,\qquad\lambda>0. \tag{17}$$

At $\lambda=0$, $f(P^0_{s,a},V,\lambda)$ is defined by its limit: $\lim_{\lambda\to0^+}f(P^0_{s,a},V,\lambda)=\min_{s':P^0(s'|s,a)>0}V(s')$.

For $\delta>0$, denote population dual maximizers λ^* by $\lambda^*=\arg\max_{\lambda\geq0}f(P^0_{s,a},V,\lambda)$. Denote by empirical dual maximizers $\widehat{\lambda}^*$ by $\widehat{\lambda}^*=\arg\max_{\lambda\geq0}f(\widehat{P}^0_{s,a},V,\lambda)$, respectively. Here, we suppress the dependence of λ^* and $\widehat{\lambda}^*$ on s,a,V without causing any confusion.

If $\lambda^*>0$, define the weight function $h^{\lambda^*,V}_{s,a}(s')$ as the ratio of the worst-case distribution $\underline{P}^V(s'|s,a)\in\arg\min_{P_{s,a}\in\mathcal{P}_{s,a}(\delta)}\mathbb{E}_{P_{s,a}}[V]$ to the nominal distribution $P^0(s'|s,a)$. The closed-form of the population worst-case distribution $\underline{P}^V(s'|s,a)$ can then be written as follows from [8]:

$$h^{\lambda^*,V}_{s,a}(s')=\frac{e^{-V(s')/\lambda^*}}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]},\quad\underline{P}^V(s'|s,a)=P^0(s'|s,a)h^{\lambda^*,V}_{s,a}(s'). \tag{18}$$

If $\widehat{\lambda}^* > 0$, the empirical worst-case distribution $\underline{\widehat{P}}^V(s'|s,a) \in \arg\min_{P_{s,a}\in\widehat{\mathcal{P}}_{s,a}(\delta)} \mathbb{E}_{P_{s,a}}[V]$ is defined in the same way as:

$$\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}(s') = \frac{e^{-V(s')/\widehat{\lambda}^*}}{\mathbb{E}_{\widehat{P}^0_{s,a}}[e^{-V/\widehat{\lambda}^*}]}, \quad \underline{\widehat{P}}^V(s'|s,a) = \widehat{P}^0(s'|s,a)\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}(s'), \tag{19}$$

where the uncertainty set $\widehat{\mathcal{P}}_{s,a}$ is centered at empirical distribution $\widehat{P}^0_{s,a}$.

For given policy π, define the transitions $\underline{P}^{V,\pi}$ and $\underline{\widehat{P}}^{V,\pi}$ by averaging over $a \sim \pi(\cdot|s)$ as

$$\underline{P}^{V,\pi}(s'|s) = \sum_{a\in\mathcal{A}} \pi(a|s)\underline{P}^V(s'|s,a); \quad \underline{\widehat{P}}^{V,\pi}(s'|s) = \sum_{a\in\mathcal{A}} \pi(a|s)\underline{\widehat{P}}^V(s'|s,a) \tag{20}$$

Next, we introduce a lemma that characterizes the properties of the weight function and the worst-case distributions when $\lambda^* > 0$.

Lemma C.1. *For $\lambda^* > 0$, the weight function and the worst-case distribution satisfy*

$$\begin{aligned} &(a). \quad \mathbb{E}_{P^0_{s,a}}[h^{\lambda^*,V}_{s,a}] = 1, && (b). \quad \max_{s'} h^{\lambda^*,V}_{s,a}(s') \le p_\wedge^{-1}, \\ &(c). \quad \mathrm{KL}(\underline{P}^V_{s,a}\|P^0_{s,a}) = \delta, && (d). \quad \frac{e^{-\overline{V}/\lambda^*}}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]} = e^\delta, \end{aligned} \tag{21}$$

where $\overline{V} = \mathbb{E}_{\underline{P}^V_{s,a}}[V]$.

Proof. First, 21.(a) follows directly from the definition in equation 18. Specifically, we have

$$\mathbb{E}_{P^0_{s,a}}[h^{\lambda^*,V}_{s,a}] = \sum_{s':P^0(s'|s,a)>0} P^0(s'|s,a)h^{\lambda^*,V}_{s,a}(s') = \frac{\sum_{s'} P^0(s'|s,a)e^{-V(s')/\lambda^*}}{\mathbb{E}_{P^0_{s,a}}\left[e^{-V/\lambda^*}\right]} = 1. \tag{22}$$

Second, since $P^0(s'|s,a)h^{\lambda^*,V}_{s,a}(s') \le 1$, therefore $h^{\lambda^*,V}_{s,a}(s') \le 1/P^0(s'|s,a) \le p_\wedge^{-1}$, which completes the proof of 21.(b).

Third, we provide the proof of 21.(c). To differentiate the dual objective $f(P^0_{s,a}, V, \lambda)$, first note that, for $\lambda > 0$,

$$\frac{d}{d\lambda}\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda}] = \frac{1}{\lambda^2}\mathbb{E}_{P^0_{s,a}}[Ve^{-V/\lambda}]. \tag{23}$$

The product rule then gives

$$\partial_\lambda f(P^0_{s,a}, V, \lambda) = -\log\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda}] - \frac{\mathbb{E}_{P^0_{s,a}}[Ve^{-V/\lambda}]}{\lambda\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda}]} - \delta. \tag{24}$$

Consequently, by Karush–Kuhn–Tucker (KKT) conditions, differentiating the dual objective gives

$$\partial_\lambda f(P^0_{s,a}, V, \lambda^*) = 0, \tag{25}$$

at an interior maximizer $\lambda^* > 0$. Therefore, we can conclude that

$$\delta = -\log\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}] - \frac{\mathbb{E}_{P^0_{s,a}}[Ve^{-V/\lambda^*}]}{\lambda^*\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]} \tag{26}$$

Moreover, using $\log\left(h^{\lambda^*,V}_{s,a}\right) = -V/\lambda^* - \log\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]$, the 21.(c) is equal to

$$\begin{aligned} \mathrm{KL}(\underline{P}^V_{s,a}\|P^0_{s,a}) &= \mathbb{E}_{\underline{P}^V_{s,a}}\left[\log\left(h^{\lambda^*,V}_{s,a}\right)\right] && (27) \\ &= \mathbb{E}_{\underline{P}^V_{s,a}}\left[-\frac{V}{\lambda^*}\right] - \log\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}] && (28) \\ &= \frac{\mathbb{E}_{P^0_{s,a}}[-V/\lambda^*\cdot e^{-V/\lambda^*}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]} - \log\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}] = \delta. && (29) \end{aligned}$$

This completes the proof of 21.(c).

Last, by equation 18, we have

$$\overline{V} = \mathbb{E}_{\underline{P}^V_{s,a}}[V] = \frac{\mathbb{E}_{P^0_{s,a}}[Ve^{-V/\lambda^*}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]}. \tag{30}$$

Substituting this identity into equation 26 and rearranging gives

$$-\log \mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}] = \delta + \frac{\overline{V}}{\lambda^*}. \tag{31}$$

Exponentiating both sides and multiplying by $e^{-\overline{V}/\lambda^*}$ yields

$$\frac{e^{-\overline{V}/\lambda^*}}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]} = e^{\delta}, \tag{32}$$

which proves 21.(d). □

C.2 Error decomposition: statistical error and planning error

Let $\widehat{\pi}^*$ be the optimal policy for the empirical robust MDP defined by $\widehat{\mathcal{P}}$, and let π_T be the policy returned by robust value iteration. Denote by $\widehat{V}_\pi$ the robust value function of policy π under the empirical uncertainty set $\widehat{\mathcal{P}}$. Write $\widehat{V}^* = \widehat{V}_{\widehat{\pi}^*}$, and let $\widehat{Q}^*$ denote the optimal robust Q-function defined by $\widehat{\mathcal{P}}$. We then decompose the error as follows and discuss each term one by one:

$$\begin{aligned} V_{\pi^*} - V_{\pi_T} = \quad & \underbrace{\left(V_{\pi^*} - \widehat{V}_{\pi^*}\right)}_{\text{Term I: statistical error under } \pi^*} + \underbrace{\left(\widehat{V}_{\pi^*} - \widehat{V}_{\widehat{\pi}^*}\right)}_{\text{Term II: optimality gap}} \\ & + \underbrace{\left(\widehat{V}_{\widehat{\pi}^*} - \widehat{V}_{\pi_T}\right)}_{\text{Term III: planning error}} + \underbrace{\left(\widehat{V}_{\pi_T} - V_{\pi_T}\right)}_{\text{Term IV:statistical error under } \pi_T}. \end{aligned} \tag{33}$$

Term II is componentwise nonpositive by the optimality of $\widehat{\pi}^*$ in the empirical robust MDP. Term III is the planning error and can be bounded using the contraction property of the empirical Bellman operator. Terms I and IV compare population and empirical robust values under π^* and π_T, respectively, which are referred to as the statistical error.

First, we bound term III. By the γ-contraction property of the robust Bellman optimality operator and the initialization $Q_0 = 0$, we have $\|Q_T - \widehat{Q}^*\|_\infty \le \gamma^T/(1-\gamma)$. Recall that $\widehat{\mathcal{T}}$ is the empirical robust Bellman optimality operator:

$$(\widehat{\mathcal{T}}V)(s) = \max_{a\in\mathcal{A}} \left[r(s,a) + \gamma \inf_{q\in\widehat{\mathcal{P}}_{s,a}(\delta)} \mathbb{E}_{s'\sim q}[V(s')] \right]. \tag{34}$$

Denote by $\widehat{\mathcal{T}}^{\pi_T}$ the empirical robust Bellman operator for a policy π_T:

$$(\widehat{\mathcal{T}}^{\pi_T}V)(s) = \sum_{a\in\mathcal{A}} \pi_T(a \mid s) \left[r(s,a) + \gamma \inf_{q\in\widehat{\mathcal{P}}_{s,a}(\delta)} \mathbb{E}_{s'\sim q}[V(s')] \right], \qquad s\in\mathcal{S}. \tag{35}$$

Since π_T is greedy with respect to Q_T, we have that

$$0 \le \widehat{\mathcal{T}}\widehat{V}^* - \widehat{\mathcal{T}}^{\pi_T}\widehat{V}^* \le 2\|Q_T - \widehat{Q}^*\|_\infty \mathbf{1}. \tag{36}$$

Using the Bellman fixed-point equations and the γ-contraction property gives

$$\|\widehat{V}^* - \widehat{V}_{\pi_T}\|_\infty \le 2\|Q_T - \widehat{Q}^*\|_\infty + \gamma\|\widehat{V}^* - \widehat{V}_{\pi_T}\|_\infty. \tag{37}$$

Rearranging and substituting the value-iteration error bound, we obtain that

$$\|\widehat{V}_{\widehat{\pi}^*} - \widehat{V}_{\pi_T}\|_\infty \le \frac{2\gamma^T}{(1-\gamma)^2}. \tag{38}$$

Therefore, using equation 38 and the fact that term II is nonpositive, the error decomposition in equation 33 can be bounded as follows:

$$\|V_{\pi^*} - V_{\pi_T}\|_\infty \le \|V_{\pi^*} - \widehat{V}_{\pi^*}\|_\infty + \|\widehat{V}_{\pi_T} - V_{\pi_T}\|_\infty + \|\widehat{V}_{\widehat{\pi}^*} - \widehat{V}_{\pi_T}\|_\infty \tag{39}$$

$$\le \underbrace{\|V_{\pi^*} - \widehat{V}_{\pi^*}\|_\infty}_{\text{term I}} + \underbrace{\|\widehat{V}_{\pi_T} - V_{\pi_T}\|_\infty}_{\text{term IV}} + \frac{2\gamma^T}{(1-\gamma)^2}. \tag{40}$$

For a fixed policy π, the robust Bellman equations yield the following bound on the value-function statistical error:

$$\|V_\pi - \widehat{V}_\pi\|_\infty \le \gamma \max_{W \in \{V_\pi, \widehat{V}_\pi\}} \left\| (I - \gamma \underline{\widehat{P}}^{W,\pi})^{-1} \left| \underline{\widehat{P}}^{V_\pi,\pi} V_\pi - \underline{P}^{V_\pi,\pi} V_\pi \right| \right\|_\infty. \tag{41}$$

Absolute value above is taken componentwise.

Recall the definition of empirical robust Bellman operator for a policy π in equation 35. Similarly, we can also define the robust Bellman operator for a policy π using the true uncertainty set $\mathcal{P}$, denoted by $\mathcal{T}^\pi$:

$$(\mathcal{T}^\pi V)(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left[r(s,a) + \gamma \inf_{q \in \mathcal{P}_{s,a}(\delta)} \mathbb{E}_{s' \sim q}[V(s')] \right], \qquad s \in \mathcal{S}. \tag{42}$$

To show equation 41, we first note that $\widehat{V}_\pi$ and V_π are fixed-points of robust Bellman operators $\widehat{\mathcal{T}}^\pi$ and $\mathcal{T}^\pi$: $\widehat{V}_\pi = \widehat{\mathcal{T}}^\pi \widehat{V}_\pi$ and $V_\pi = \mathcal{T}^\pi V_\pi$. Applying these, we obtain that

$$\widehat{V}_\pi - V_\pi = \gamma\big(\underline{\widehat{P}}^{\widehat{V}_\pi,\pi} \widehat{V}_\pi - \underline{\widehat{P}}^{V_\pi,\pi} V_\pi\big) + \gamma\big(\underline{\widehat{P}}^{V_\pi,\pi} V_\pi - \underline{P}^{V_\pi,\pi} V_\pi\big). \tag{43}$$

For each state-action, since $\underline{\widehat{P}}^{\widehat{V}_\pi,\pi}$ corresponds to the worst-case transition kernel with respect to $\widehat{V}_\pi$, we have that

$$\underline{\widehat{P}}^{\widehat{V}_\pi,\pi} \widehat{V}_\pi \le \underline{\widehat{P}}^{V_\pi,\pi} \widehat{V}_\pi. \tag{44}$$

Similarly, for V_π, we have that

$$\underline{\widehat{P}}^{V_\pi,\pi} V_\pi \le \underline{\widehat{P}}^{\widehat{V}_\pi,\pi} V_\pi. \tag{45}$$

Substituting equation 45 and equation 44 into equation 43, respectively and rearranging yields

$$(I - \gamma \underline{\widehat{P}}^{\widehat{V}_\pi,\pi})(\widehat{V}_\pi - V_\pi) \ge \gamma\big(\underline{\widehat{P}}^{V_\pi,\pi} V_\pi - \underline{P}^{V_\pi,\pi} V_\pi\big), \tag{46}$$

$$(I - \gamma \underline{\widehat{P}}^{V_\pi,\pi})(\widehat{V}_\pi - V_\pi) \le \gamma\big(\underline{\widehat{P}}^{V_\pi,\pi} V_\pi - \underline{P}^{V_\pi,\pi} V_\pi\big). \tag{47}$$

For every stochastic matrix P, $(I - \gamma P)^{-1} = \sum_{t=0}^\infty \gamma^t P^t$ has nonnegative entries. Multiplying the two inequalities by their respective inverses preserves inequality directions. Finally, $|(I-\gamma P)^{-1} x| \le (I - \gamma P)^{-1}|x|$ componentwise. The positive and negative parts of $\widehat{V}_\pi - V_\pi$ are thus bounded by the two quantities in equation 41.

C.3 Support retention and a stable dual comparison

We first establish a high-probability lower bound on the empirical transition probabilities. With sufficiently many samples, each positive nominal transition probability is estimated to be at least three quarters of its true value, simultaneously for all state-action pairs and successor states. Consequently, the empirical and nominal transition kernels have the same support. This lower bound will be used to compare the expectations appearing in the two dual objectives.

Lemma C.2 (Uniform lower bound on empirical transition probabilities)**.** *If* $N \ge \frac{32}{p_\wedge} \log \frac{|\mathcal{S}|^2|\mathcal{A}|}{\eta}$, *then, with probability at least* $1 - \eta$,

$$\widehat{P}^0(s'|s,a) \ge \tfrac{3}{4} P^0(s'|s,a) \quad \forall (s,a,s') \text{ with } P^0(s'|s,a) > 0. \tag{48}$$

Moreover, the empirical and nominal transition distributions share the same support for every state-action pair (s,a).

Proof. For any (s, a, s') such that $P^0(s'|s, a) > 0$, $N\widehat{P}^0(s'|s, a)$ is a binomial random variable with mean $NP^0(s'|s, a)$. The multiplicative Chernoff inequality gives

$$\Pr\Big(\widehat{P}^0(s'|s, a) < \tfrac{3}{4}P^0(s'|s, a)\Big) \le e^{-NP^0(s'|s,a)/32} \le e^{-Np_\wedge/32}. \tag{49}$$

A union bound over at most $|\mathcal{S}|^2|\mathcal{A}|$ entries proves equation 48.

Sampling cannot create successors outside the nominal support, and the event prevents any nominal successor from disappearing, which proves the last claim in the lemma. □

Conditioned on the high probability event in Lemma C.2, we next bound the statistical errors in equation 33. Denote by $h^{\lambda,V}_{s,a}(s') := \frac{e^{-V(s')/\lambda}}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda}]}, \forall\lambda > 0$.

Lemma C.3. *Conditioned on the event in Lemma C.2, for any value vector V, we have*

$$\begin{aligned}\left|\inf_{P_{s,a}\in\widehat{\mathcal{P}}_{s,a}} \mathbb{E}_{P_{s,a}}[V] - \inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{P_{s,a}}[V]\right| &\le \max_{\lambda\in\{\lambda^*,\widehat{\lambda}^*\}} \lambda\left|\log\frac{\mathbb{E}_{\widehat{P}^0_{s,a}}[e^{-V/\lambda}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda}]}\right| \\ &\le \frac{4}{3}\max_{\lambda\in\{\lambda^*,\widehat{\lambda}^*\}} \lambda\left|(\widehat{P}^0_{s,a} - P^0_{s,a})h^{\lambda,V}_{s,a}\right|. \end{aligned} \tag{50}$$

If either of the dual maximizers $\lambda^*, \widehat{\lambda}^*$ equals zero, the corresponding terms on the right-hand side are interpreted by continuous extension and equal zero.

Proof. We first consider $\lambda^*, \widehat{\lambda}^* > 0$. By the definition of $\lambda^*, \widehat{\lambda}^*$, evaluating the dual objective f in equation 17 at $\widehat{P}^0_{s,a}, P^0_{s,a}$ with $\lambda^*, \widehat{\lambda}^*$ for the same V gives

$$\begin{aligned} f(\widehat{P}^0_{s,a}, V, \lambda^*) - f(P^0_{s,a}, V, \lambda^*) &\le f(\widehat{P}^0_{s,a}, V, \widehat{\lambda}^*) - f(P^0_{s,a}, V, \widehat{\lambda}^*) \\ &= -\widehat{\lambda}^* \log\frac{\mathbb{E}_{\widehat{P}^0_{s,a}}[e^{-V/\widehat{\lambda}^*}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\widehat{\lambda}^*}]}. \end{aligned} \tag{51}$$

Similarly, we have that

$$\begin{aligned} f(P^0_{s,a}, V, \lambda^*) - f(\widehat{P}^0_{s,a}, V, \widehat{\lambda}^*) &\le f(P^0_{s,a}, V, \lambda^*) - f(\widehat{P}^0_{s,a}, V, \lambda^*) \\ &= \lambda^* \log\frac{\mathbb{E}_{\widehat{P}^0_{s,a}}[e^{-V/\lambda^*}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]}. \end{aligned} \tag{52}$$

Combining equation 51 and equation 52 proves the first inequality in equation 50.

Conditioned on the event in Lemma C.2, it follows that $\mathbb{E}_{\widehat{P}^0_{s,a}}[Z] \ge (3/4)\mathbb{E}_{P^0_{s,a}}[Z]$ for any non-negative Z. Then, if $x \ge \frac{3}{4}y > 0$, it holds that $\min\{x, y\} \ge \frac{3}{4}y$. We can conclude that $|\log x - \log y| \le |x - y|/\min\{x, y\} \le \frac{4}{3}|x - y|/|x|$. Applying this on the right hand side of the first inequality in equation 50 shows the second inequality.

If $\lambda^* = \widehat{\lambda}^* = 0$, it holds that

$$f(\widehat{P}^0_{s,a}, V, \widehat{\lambda}^*) - f(P^0_{s,a}, V, 0) = \min_{s'} V(s') - \min_{s'} V(s') = 0. \tag{53}$$

If $\lambda^* = 0 < \widehat{\lambda}^*$, then, it holds that

$$0 \le f(\widehat{P}^0_{s,a}, V, \widehat{\lambda}^*) - f(P^0_{s,a}, V, 0) \tag{54}$$

$$\le f(\widehat{P}^0_{s,a}, V, \widehat{\lambda}^*) - f(P^0_{s,a}, V, \widehat{\lambda}^*) \tag{55}$$

$$= -\widehat{\lambda}^* \log\frac{\mathbb{E}_{\widehat{P}^0_{s,a}}[e^{-V/\widehat{\lambda}^*}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\widehat{\lambda}^*}]} \tag{56}$$

$$\le \frac{4}{3}\widehat{\lambda}^*\left|(\widehat{P}^0_{s,a} - P^0_{s,a})h^{\widehat{\lambda}^*,V}_{s,a}\right|. \tag{57}$$

If $\widehat{\lambda}^* = 0 < \lambda^*$, the same argument gives

$$0 \le f(P^0_{s,a}, V, \lambda^*) - f(\widehat{P}^0_{s,a}, V, 0) \tag{58}$$

$$\le f(P^0_{s,a}, V, \lambda^*) - f(\widehat{P}^0_{s,a}, V, \lambda^*) \tag{59}$$

$$= \lambda^* \log \frac{\mathbb{E}_{\widehat{P}^0_{s,a}}[e^{-V/\lambda^*}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]} \tag{60}$$

$$\le \frac{4}{3}\lambda^* \left| (\widehat{P}^0_{s,a} - P^0_{s,a}) h^{\lambda^*, V}_{s,a} \right|. \tag{61}$$

This completes the proof. □

C.4 Concentration results

To bound the statistical error terms in equation 33, Lemma C.3 reduces the comparison of empirical and nominal robust expectations to controlling the difference $(\widehat{P}^0_{s,a} - P^0_{s,a})h$, where h is the corresponding weight vector. In this subsection, we use Bernstein's inequality to derive concentration bounds for this difference. We first establish absolute and relative bounds for a vector h fixed independently of the transition samples.

Lemma C.4 (Absolute and relative Bernstein bounds)**.** *Let $X_1, \dots, X_n$ be n independent samples from a discrete distribution $P \in \Delta(\mathcal{S})$, and let $\widehat{P}$ be their empirical distribution. Fix a function $g : \mathcal{S} \to \mathbb{R}$ independently of these samples, with $0 \le g \le B_g$, and suppose $\mu_g = \mathbb{E}_P[g] > 0$. Then, with probability at least $1 - \eta$,*

$$|(\widehat{P} - P)g| \le \sqrt{\frac{2\operatorname{Var}_P(g)\log(2/\eta)}{n}} + \frac{2B_g \log(2/\eta)}{3n}, \tag{62}$$

$$\frac{|(\widehat{P} - P)g|}{\mathbb{E}_P[g]} \le \sqrt{\frac{2\operatorname{Var}_P(g)\log(2/\eta)}{n(\mathbb{E}_P[g])^2}} + \frac{2B_g \log(2/\eta)}{3n\mathbb{E}_P[g]}. \tag{63}$$

Conditioned on the above high probability event, if the right-hand side of equation 62 is strictly smaller than $\mathbb{E}_P[g]$, then

$$\frac{|(\widehat{P} - P)g|}{\mathbb{E}_{\widehat{P}}[g]} \le \frac{\sqrt{\frac{2\operatorname{Var}_P(g)\log(2/\eta)}{n}} + \frac{2B_g \log(2/\eta)}{3n}}{\mathbb{E}_P[g] - \sqrt{\frac{2\operatorname{Var}_P(g)\log(2/\eta)}{n}} - \frac{2B_g \log(2/\eta)}{3n}}. \tag{64}$$

In particular, if the right-hand side of equation 63 is at most $1/2$, the empirical-relative error is at most twice that right-hand side of equation 63.

Proof. Set $Y_i = g(X_i) - \mu_g$, and therefore Y_i's are also independent and

$$\mathbb{E}[Y_i] = 0, \qquad \mathbb{E}[Y_i^2] = \operatorname{Var}_P(g), \qquad |Y_i| \le B_g. \tag{65}$$

The two-sided Bernstein inequality therefore states that, for $t > 0$,

$$\Pr\left(\left| \frac{1}{n} \sum_{i=1}^{n} Y_i \right| > t \right) \le 2\exp\left(-\frac{nt^2}{2\operatorname{Var}_P(g) + (2/3)B_g t} \right). \tag{66}$$

For $x > 0$, the positive solution of $nt^2 - (2/3)B_g x t - 2\operatorname{Var}_P(g)x = 0$ is

$$\frac{B_g x}{3n} + \sqrt{\frac{2\operatorname{Var}_P(g)x}{n} + \frac{B_g^2 x^2}{9n^2}} \le \sqrt{\frac{2\operatorname{Var}_P(g)x}{n}} + \frac{2B_g x}{3n}. \tag{67}$$

Taking $x = \log(2/\eta)$ proves equation 62. Further dividing by the positive μ_g proves equation 63.

For the empirical denominator, the same event gives

$$\mathbb{E}_{\widehat{P}}[g] \ge \mathbb{E}_P[g] - |(\widehat{P} - P)g| \tag{68}$$

$$\ge \mathbb{E}_P[g] - \sqrt{\frac{2\operatorname{Var}_P(g)\log(2/\eta)}{n}} - \frac{2B_g \log(2/\eta)}{3n} > 0. \tag{69}$$

We then divide equation 62 by this lower bound to obtain equation 64.

If the nominal-relative upper bound is at most $1/2$, the empirical mean is at least $\mu_g/2$, so replacing the denominator by $\mu_g/2$ proves the final assertion. All three bounds follow from the same concentration event, which holds with probability at least $1-\eta$. □

We next apply Bernstein's inequality to the weight function $h_{s,a}^{\lambda,V}$ for fixed V and $\lambda > 0$. Since $\mathbb{E}_{P_{s,a}^0}[h_{s,a}^{\lambda,V}] = 1$, we consider the centered random variable $\lambda(h_{s,a}^{\lambda,V} - 1)$, whose empirical mean can be written as $\lambda(\widehat{P}_{s,a}^0 - P_{s,a}^0)h_{s,a}^{\lambda,V}$. We show that its absolute value is bounded by $\operatorname{span}(V)/p_\wedge$ on the support of $P_{s,a}^0$. This yields a Bernstein bound whose $1/N$ remainder has no additional factor of λ.

Lemma C.5. *Fix a value vector V and a dual parameter $\lambda > 0$, both chosen independently of the samples drawn from $P_{s,a}^0$. With probability at least $1-\eta$, we have*

$$\lambda\left|(\widehat{P}_{s,a}^0 - P_{s,a}^0)h_{s,a}^{\lambda,V}\right| \le \lambda\sqrt{\frac{2\operatorname{Var}_{P_{s,a}^0}(h_{s,a}^{\lambda,V})\log(2/\eta)}{N}} + \frac{2\operatorname{span}(V)\log(2/\eta)}{3p_\wedge N}, \tag{70}$$

where $\operatorname{span}(V) = \max_s V(s) - \min_s V(s)$.

Proof. Throughout this proof, $\lambda > 0$, V, and (s,a) are fixed. Suppressing the dependence on these quantities, we define the scalar function $h : \mathbb{R} \to \mathbb{R}$ by

$$h(v) = \frac{e^{-v/\lambda}}{\mathbb{E}_{P_{s,a}^0}\left[e^{-V/\lambda}\right]}, v \in \mathbb{R}. \tag{71}$$

Let $V_c = -\lambda \log \mathbb{E}_{P_{s,a}^0}[e^{-V/\lambda}]$. Then $\min V \le V_c \le \max V$ and $h(V_c) = 1$.

Since $h'(v) = -h(v)/\lambda$, the function h is decreasing. For any s' with $P^0(s' \mid s,a) > 0$, the mean-value theorem gives a point ξ between $V(s')$ and V_c such that

$$h(V(s')) - 1 = h(V(s')) - h(V_c) = h'(\xi)(V(s') - V_c). \tag{72}$$

Consequently,

$$\lambda|h(V(s')) - 1| = h(\xi)|V(s') - V_c|. \tag{73}$$

Because h is decreasing and $h(V_c) = 1$,

$$h(\xi) \le \max\{h(V(s')), h(V_c)\} = \max\{h(V(s')), 1\}. \tag{74}$$

Together with $|V(s') - V_c| \le \operatorname{span}(V)$, this gives

$$\lambda|h(V(s')) - 1| \le \operatorname{span}(V)\max\{h(V(s')), 1\}. \tag{75}$$

Moreover, the denominator in the definition of h satisfies

$$\mathbb{E}_{P_{s,a}^0}[e^{-V/\lambda}] \ge P^0(s' \mid s,a)e^{-V(s')/\lambda}. \tag{76}$$

Hence,

$$h(V(s')) \le \frac{1}{P^0(s' \mid s,a)} \le \frac{1}{p_\wedge}. \tag{77}$$

Since $1 \le p_\wedge^{-1}$, we conclude that

$$\lambda|h(V(s')) - 1| \le \frac{\operatorname{span}(V)}{p_\wedge} \tag{78}$$

for every s' in the support of $P_{s,a}^0$.

Let $s'_1, \dots, s'_N$ be the independent samples drawn from $P_{s,a}^0$. Using $h(V(s')) = h_{s,a}^{\lambda,V}(s')$, define

$$Y_i = \lambda\big(h(V(s'_i)) - 1\big) = \lambda\big(h_{s,a}^{\lambda,V}(s'_i) - 1\big). \tag{79}$$

These variables are independent and satisfy

$$\begin{aligned}\mathbb{E}[Y_i] &= 0,\\ \mathrm{Var}(Y_i) &= \lambda^2 \,\mathrm{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a}),\\ |Y_i| &\le \frac{\mathrm{span}(V)}{p_\wedge}.\end{aligned} \tag{80}$$

Furthermore, since $\mathbb{E}_{P^0_{s,a}}[h^{\lambda,V}_{s,a}] = 1$,

$$\frac{1}{N}\sum_{i=1}^{N} Y_i = \lambda(\widehat{P}^0_{s,a} - P^0_{s,a})h^{\lambda,V}_{s,a}. \tag{81}$$

Applying the two-sided Bernstein inequality to this average with failure probability η proves equation 70. □

Lemma C.5 requires both the value vector V and the dual parameter $\lambda > 0$ to be chosen independently of the samples drawn from $P^0_{s,a}$. It therefore does not directly apply when either quantity depends on these samples. To address this dependence, we construct auxiliary models in which a given state is absorbing, so that their value functions, for each fixed auxiliary parameter, are independent of the samples from that state. We then apply Bernstein's inequality conditionally on the samples from the other states and take a union bound over finite grids of the auxiliary parameter and λ. Controlling the discretization error extends these bounds to the required sample-dependent value functions and their nominal and empirical dual optimizers. The following lemma thus provides a bound on the one-step estimation error that can be used in the subsequent analysis.

Recall that λ^* and $\widehat{\lambda}^*$ are the nominal and empirical dual optimizers for the given (s, a, V), respectively.

Lemma C.6. *Let $\delta, \eta \in (0, 1)$. Suppose that N samples are drawn from each nominal distribution $P^0_{s,a}$, with all samples mutually independent, and that*

$$N \ge \frac{192}{p_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}. \tag{82}$$

Then, with probability at least $1-\eta$, the following inequalities hold simultaneously for all (s, a, s') and $V \in \{V_{\pi^*}, \widehat{V}_{\pi^*}, \widehat{V}_{\widehat{\pi}^*}\}$:

$$\left|\widehat{P}^0(s'|s,a) - P^0(s'|s,a)\right| \le P^0(s'|s,a)\sqrt{\frac{12}{Np_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}} \le \frac{1}{4}P^0(s'|s,a), \tag{83}$$

and

$$\begin{aligned}\left|\underline{\widehat{P}}^V(\cdot \mid s,a)V - \underline{P}^V(\cdot \mid s,a)V\right| &\le 4\sqrt{\frac{1}{N}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\max_{\lambda\in\{\lambda^*,\widehat{\lambda}^*\}}\lambda\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a})}\\ &\quad + \frac{64}{p_\wedge N(1-\gamma)}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}.\end{aligned} \tag{84}$$

If either dual optimizer λ^* or $\widehat{\lambda}^*$ equals zero, its contribution to the above maximum is defined by

$$\lambda\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a})}\Big|_{\lambda=0} := \lim_{\lambda\downarrow 0}\lambda\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a})} = 0.$$

In particular, equation 83 implies

$$\mathrm{supp}(\widehat{P}^0_{s,a}) = \mathrm{supp}(P^0_{s,a}) \qquad \textit{for every } (s,a).$$

Proof. Set $L = \log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}$, $C = 3 + 2\log(1/p_\wedge)$. Fix a population-optimal policy π^*, and it is independent of the samples.

We first establish the relative probability bound. Equation (82) implies $\sqrt{12L/(Np_\wedge)} \le 1/4$. For every (s,a,s') with $P^0(s'|s,a) > 0$, the multiplicative Chernoff inequality gives

$$\begin{aligned}\Pr\left(\left|\widehat{P}^0(s'|s,a) - P^0(s'|s,a)\right| > P^0(s'|s,a)\sqrt{\frac{12L}{Np_\wedge}}\right) \\ \le 2\exp\left(-\frac{4LP^0(s'|s,a)}{p_\wedge}\right) \le 2e^{-4L},\end{aligned} \tag{85}$$

where the last inequality uses $P^0(s'|s,a) \ge p_\wedge$. If $P^0(s'|s,a) = 0$, then $\widehat{P}^0(s'|s,a) = 0$ almost surely. A union bound over all (s,a,s') therefore shows that equation 83 holds on an event $\mathcal{E}_{\mathrm{rel}}$ with

$$\Pr(\mathcal{E}_{\mathrm{rel}}^c) \le 2|\mathcal{S}|^2|\mathcal{A}|e^{-4L} \le \frac{\eta}{2}, \tag{86}$$

where

$$\mathcal{E}_{\mathrm{rel}} := \left\{\left|\widehat{P}^0(s'|s,a) - P^0(s'|s,a)\right| \le P^0(s'|s,a)\sqrt{\frac{12L}{Np_\wedge}} \quad \text{for all } (s,a,s')\right\}. \tag{87}$$

On this event, the nominal and empirical supports agree, and

$$\frac{3}{4}P^0(s'|s,a) \le \widehat{P}^0(s'|s,a) \le \frac{5}{4}P^0(s'|s,a) \qquad \text{for every } (s,a,s'). \tag{88}$$

Moreover,

$$\begin{aligned}d_{\mathrm{TV}}(P^0_{s,a}, \widehat{P}^0_{s,a}) &= \frac{1}{2}\sum_{s'\in\mathcal{S}}\left|\widehat{P}^0(s'|s,a) - P^0(s'|s,a)\right| \\ &\le \sqrt{\frac{3L}{Np_\wedge}} \qquad \text{for every } (s,a).\end{aligned} \tag{89}$$

Step 1: Bernstein's inequality for fixed functions. For $V \in [0,(1-\gamma)^{-1}]^{|\mathcal{S}|}$ and $\lambda > 0$, define the centered function $\phi^{\lambda,V}_{s,a} : \mathrm{supp}(P^0_{s,a}) \to \mathbb{R}$ by

$$\phi^{\lambda,V}_{s,a}(s') := \lambda\big(h^{\lambda,V}_{s,a}(s') - 1\big) = \lambda\left(\frac{e^{-V(s')/\lambda}}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda}]} - 1\right). \tag{90}$$

The definition of $\phi^{\lambda,V}_{s,a}$ and the normalization $\mathbb{E}_{P^0_{s,a}}[h^{\lambda,V}_{s,a}] = 1$ give the mean and variance identities below

$$\begin{aligned}\mathbb{E}_{P^0_{s,a}}[\phi^{\lambda,V}_{s,a}] &= 0, \\ \left|\phi^{\lambda,V}_{s,a}(s')\right| &\le \frac{1}{p_\wedge(1-\gamma)} \qquad \text{for every } s' \in \mathrm{supp}(P^0_{s,a}), \\ \mathrm{Var}_{P^0_{s,a}}(\phi^{\lambda,V}_{s,a}) &= \lambda^2\,\mathrm{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a}).\end{aligned} \tag{91}$$

The absolute bound follows from equation 78, together with $\mathrm{span}(V) \le (1-\gamma)^{-1}$.

If V and λ are fixed independently of the samples from (s,a), Bernstein's inequality gives

$$\left|(\widehat{P}^0_{s,a} - P^0_{s,a})\phi^{\lambda,V}_{s,a}\right| \le \sqrt{\frac{8L}{N}\,\mathrm{Var}_{P^0_{s,a}}(\phi^{\lambda,V}_{s,a})} + \frac{16L}{3p_\wedge N(1-\gamma)}, \tag{92}$$

which holds with probability at least $1 - 2e^{-4L}$.

The discretization argument in Step 2 also uses the limiting cases $\lambda = 0$ and $\lambda = \infty$. We therefore define

$$\begin{aligned}\phi^{0,V}_{s,a}(s') &:= \lim_{\lambda\downarrow 0}\phi^{\lambda,V}_{s,a}(s') = 0, \\ \phi^{\infty,V}_{s,a}(s') &:= \lim_{\lambda\to\infty}\phi^{\lambda,V}_{s,a}(s') = \mathbb{E}_{P^0_{s,a}}[V] - V(s').\end{aligned} \tag{93}$$

The bound equation 92 is immediate for $\phi^{0,V}_{s,a} = 0$. For $\phi^{\infty,V}_{s,a}$, it follows from Bernstein's inequality because this function has mean zero under $P^0_{s,a}$ and absolute value at most $(1-\gamma)^{-1}$.

Step 2: Deriving uniform concentration. We first establish regularity estimates for the centered function $\phi^{\lambda,V}_{s,a}$ to control the approximation error of the finite grids constructed below. For a fixed state–action pair (s,a), any $V, W \in [0,(1-\gamma)^{-1}]^{|\mathcal{S}|}$, and finite $\lambda, \mu \geq 0$, we have

$$\|\phi^{\lambda,V}_{s,a} - \phi^{\lambda,W}_{s,a}\|_\infty \leq \frac{2}{p_\wedge}\|V - W\|_\infty, \tag{94}$$

$$\|\phi^{\lambda,V}_{s,a} - \phi^{\mu,V}_{s,a}\|_\infty \leq \frac{C}{p_\wedge}|\lambda - \mu|. \tag{95}$$

For positive dual parameters, these inequalities follow by differentiating equation 90. The calculation uses $h^{\lambda,V}_{s,a}(s') \leq p_\wedge^{-1}$ whenever $P^0(s'|s,a) > 0$, the normalization $\mathbb{E}_{P^0_{s,a}}[h^{\lambda,V}_{s,a}] = 1$, and the bounds

$$0 \leq \mathbb{E}_{P^0_{s,a}}\left[h^{\lambda,V}_{s,a} \log h^{\lambda,V}_{s,a}\right] \leq \log(1/p_\wedge),$$
$$|x \log x| \leq p_\wedge^{-1}\log(1/p_\wedge) + 1/e, \qquad 0 < x \leq p_\wedge^{-1}. \tag{96}$$

The continuous extension at $\lambda = 0$ gives the same Lipschitz bounds when either parameter is zero.

For large λ, we approximate $\phi^{\lambda,V}_{s,a}$ by its limiting function $\phi^{\infty,V}_{s,a}$ defined in equation 93. A Taylor expansion in $1/\lambda$ around zero yields

$$\|\phi^{\lambda,V}_{s,a} - \phi^{\infty,V}_{s,a}\|_\infty \leq \frac{e}{2\lambda(1-\gamma)^2} \leq \frac{2}{\lambda(1-\gamma)^2}, \qquad \lambda \geq \frac{1}{1-\gamma}. \tag{97}$$

Indeed, for $0 \leq t \leq 1-\gamma$, the second derivative of $e^{-tV(s')}/\mathbb{E}_{P^0_{s,a}}[e^{-tV}]$ with respect to t has absolute value at most $e/(1-\gamma)^2$. Thus, sufficiently large values of λ can be covered using the single limiting function $\phi^{\infty,V}_{s,a}$.

To handle the dependence of the empirical value functions on the samples, we construct an auxiliary model for each state s. For a fixed parameter $\theta \in [0,(1-\gamma)^{-1}]$, replace the nominal transition distribution at (s,a) by the point mass e_s and the reward by $(1-\gamma)\theta$, for every action a. Keep the empirical model at all other states unchanged. Since the forward-KL uncertainty set centered at e_s contains only e_s, state s is absorbing and its value is θ.

Let $W^{\pi^*}_s(\theta)$ denote the robust value of the fixed policy π^* in this auxiliary model, and let $W^*_s(\theta)$ denote its optimal robust value. For each fixed θ, both vectors depend only on samples collected from state–action pairs (x,a) with $x \neq s$. Conditional on these samples, the two value vectors are fixed, while the samples from (s,a) remain independent samples from $P^0_{s,a}$. Thus, the Bernstein bound from Step 1 applies to the centered functions associated with either auxiliary value and any fixed λ.

These auxiliary families recover the empirical value functions at suitable choices of θ:

$$W^{\pi^*}_s\left(\widehat{V}_{\pi^*}(s)\right) = \widehat{V}_{\pi^*},$$
$$W^*_s\left(\widehat{V}_{\widehat{\pi}^*}(s)\right) = \widehat{V}_{\widehat{\pi}^*}. \tag{98}$$

Indeed, the original empirical value satisfies the unchanged Bellman equations at every state other than s. At state s, setting θ equal to that value gives $(1-\gamma)\theta + \gamma\theta = \theta$. The identities in equation 98 therefore follow from uniqueness of the corresponding Bellman fixed points.

Moreover, changing θ to θ' changes the auxiliary reward vector by at most $(1-\gamma)|\theta - \theta'|$. The contraction properties of the policy and optimal Bellman operators imply

$$\left\|W^{\pi^*}_s(\theta) - W^{\pi^*}_s(\theta')\right\|_\infty \leq |\theta - \theta'|,$$
$$\|W^*_s(\theta) - W^*_s(\theta')\|_\infty \leq |\theta - \theta'|. \tag{99}$$

Thus, discretizing the scalar parameter θ also controls the approximation error of the corresponding value vectors.

We now choose deterministic grids for $\theta \in [0,(1-\gamma)^{-1}]$ and $\lambda \in [0, N/(1-\gamma)]$, including their endpoints, with respective maximum spacings

$$\frac{1}{N(1-\gamma)} \qquad \text{and} \qquad \frac{1}{NC(1-\gamma)}. \tag{100}$$

We also include the limiting function at $\lambda = \infty$. For each (s, a) and each pair of grid points, form the centered functions associated with $W_s^{\pi^*}(\theta)$ and $W_s^*(\theta)$. Include also the centered functions associated with the fixed value V_{π^*} at every λ-grid point.

The θ-grid contains at most $N + 1$ points, and the λ-grid, including ∞, contains at most $\lceil N^2 C \rceil + 2$ points. Consequently, the total number of centered grid functions over all (s, a) is at most

$$|\mathcal{S}||\mathcal{A}|(2N + 3)(\lceil N^2 C \rceil + 2) \le 12|\mathcal{S}||\mathcal{A}|N^3 C. \tag{101}$$

We next bound the approximation error of these grids. Fix (s, a) and $V \in \{V_{\pi^*}, \widehat{V}_{\pi^*}, \widehat{V}_{\widehat{\pi}^*}\}$. By equation 98 and equation 99, there is a value vector V_g included in the grid construction such that

$$\|V - V_g\|_\infty \le \frac{1}{N(1-\gamma)}. \tag{102}$$

For $V = V_{\pi^*}$, we simply take $V_g = V_{\pi^*}$.

If $0 \le \lambda \le N/(1-\gamma)$, choose a grid point λ_g with $|\lambda - \lambda_g| \le 1/[NC(1-\gamma)]$. The preceding Lipschitz bounds give

$$\begin{aligned}\|\phi_{s,a}^{\lambda,V} - \phi_{s,a}^{\lambda_g,V_g}\|_\infty &\le \frac{2}{p_\wedge}\|V - V_g\|_\infty + \frac{C}{p_\wedge}|\lambda - \lambda_g| \\ &\le \frac{3}{p_\wedge N(1-\gamma)}.\end{aligned} \tag{103}$$

If $\lambda > N/(1-\gamma)$, use the limiting grid function $\phi_{s,a}^{\infty,V_g}$ instead. The Taylor bound above and the 2-Lipschitz property of $V \mapsto \mathbb{E}_{P_{s,a}^0}[V] - V$ yield

$$\begin{aligned}\|\phi_{s,a}^{\lambda,V} - \phi_{s,a}^{\infty,V_g}\|_\infty &\le \frac{2}{\lambda(1-\gamma)^2} + 2\|V - V_g\|_\infty \\ &\le \frac{4}{N(1-\gamma)} \le \frac{4}{p_\wedge N(1-\gamma)}.\end{aligned} \tag{104}$$

Therefore, for every required V and every $\lambda \ge 0$, there is a centered grid function ϕ_{grid} satisfying

$$\|\phi_{s,a}^{\lambda,V} - \phi_{\mathrm{grid}}\|_\infty \le \frac{4}{p_\wedge N(1-\gamma)}. \tag{105}$$

Applying equation 92 conditional on the samples collected from the other states and taking the union bound gives simultaneous Bernstein bounds on these grids with failure probability at most $\eta/2$. Together with equation 83, with probability at least $1 - \eta$, for every required V, λ,

$$\begin{aligned}&\lambda\left|(\widehat{P}_{s,a}^0 - P_{s,a}^0)h_{s,a}^{\lambda,V}\right| \\ &\overset{\text{(i)}}{\le} \left|(\widehat{P}_{s,a}^0 - P_{s,a}^0)\phi_{\mathrm{grid}}\right| + \frac{8}{p_\wedge N(1-\gamma)} \\ &\overset{\text{(ii)}}{\le} \sqrt{\frac{8L}{N}\operatorname{Var}_{P_{s,a}^0}(\phi_{\mathrm{grid}})} + \frac{16L}{3p_\wedge N(1-\gamma)} + \frac{8}{p_\wedge N(1-\gamma)} \\ &\overset{\text{(iii)}}{\le} \lambda\sqrt{\frac{8L}{N}\operatorname{Var}_{P_{s,a}^0}(h_{s,a}^{\lambda,V})} + \frac{16L}{3p_\wedge N(1-\gamma)} \\ &\qquad + \frac{4}{p_\wedge N(1-\gamma)}\left(2 + \sqrt{\frac{8L}{N}}\right) \\ &\overset{\text{(iv)}}{\le} \lambda\sqrt{\frac{8L}{N}\operatorname{Var}_{P_{s,a}^0}(h_{s,a}^{\lambda,V})} + \frac{40L}{p_\wedge N(1-\gamma)}.\end{aligned} \tag{106}$$

Here, (i) follows from equation 105, the identity $(\widehat{P}_{s,a}^0 - P_{s,a}^0)\mathbf{1} = 0$, and $\|\widehat{P}_{s,a}^0 - P_{s,a}^0\|_1 \le 2$; (ii) uses the simultaneous Bernstein bound for the grid functions; (iii) combines equation 105 and equation 91 with the fact that standard deviation is 1-Lipschitz under the supremum norm; and (iv) uses equation 82 and $L \ge 1$.

Step 3: Comparing the empirical and nominal transitions. On the event established in Step 2, which has probability at least $1-\eta$, equation 83 gives, simultaneously for every (s,a), value vector V, and $\lambda > 0$,

$$\frac{3}{4} \le \frac{\mathbb{E}_{\widehat{P}^0_{s,a}}[e^{-V/\lambda}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda}]} \le \frac{5}{4}. \tag{107}$$

Using $|\log x| \le \frac{4}{3}|x-1|$ on $[3/4, 5/4]$ and evaluating each dual objective at the other objective's maximizer yields

$$\begin{aligned} &\left|\underline{\widehat{P}}^V(\cdot \mid s,a)V - \underline{P}^V(\cdot \mid s,a)V\right| \\ &\qquad \le \frac{4}{3} \max_{\lambda \in \{\lambda^*, \widehat{\lambda}^*\}} \lambda \left|(\widehat{P}^0_{s,a} - P^0_{s,a}) h^{\lambda,V}_{s,a}\right|. \end{aligned} \tag{108}$$

For each value vector V listed in the lemma, applying equation 106 and using $4\sqrt{8}/3 < 4$ and $160/3 < 64$ gives equation 84. If either optimizer equals zero, its corresponding term in the dual comparison is defined by continuous extension and equals zero, since the two dual objectives coincide at $\lambda = 0$ when the nominal and empirical supports agree. □

Let $\widehat{\pi} := \pi_T$ denote the deterministic policy returned by robust value iteration, with ties resolved by a fixed rule. Choose π^* to be a deterministic population-optimal policy.

Lemma C.7 (Comparison of value variances)**.** *Suppose* $0 < \delta \le 1$ *and*

$$N \ge \frac{192}{p_\wedge} \log \frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}. \tag{109}$$

On the event of Lemma C.6, the following inequalities hold at every state:

$$\begin{aligned} &\max\left\{\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{V}_{\pi^*},\pi^*}}(\widehat{V}_{\pi^*})}, \sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\widehat{V}_{\pi^*},\pi^*}}(\widehat{V}_{\pi^*})}\right\} \\ &\le \frac{2}{\sqrt{3}}\sqrt{\mathrm{Var}_{\underline{P}^{V_{\pi^*},\pi^*}}(V_{\pi^*})} + \left(\frac{2}{\sqrt{3}}\|\widehat{V}_{\pi^*} - V_{\pi^*}\|_\infty + \frac{2\delta^{1/4}}{1-\gamma}\right)\mathbf{1}, \end{aligned} \tag{110}$$

$$\begin{aligned} &\max\left\{\sqrt{\mathrm{Var}_{\underline{P}^{V_{\pi^*},\pi^*}}(V_{\pi^*})}, \sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{V_{\pi^*},\pi^*}}(V_{\pi^*})}\right\} \\ &\le \frac{2}{\sqrt{3}}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\widehat{V}_{\pi^*},\pi^*}}(\widehat{V}_{\pi^*})} + \left(\frac{2}{\sqrt{3}}\|V_{\pi^*} - \widehat{V}_{\pi^*}\|_\infty + \frac{2\delta^{1/4}}{1-\gamma}\right)\mathbf{1}, \end{aligned} \tag{111}$$

$$\begin{aligned} &\max\left\{\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{V}_{\widehat{\pi}^*},\widehat{\pi}}}(\widehat{V}_{\widehat{\pi}^*})}, \sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\widehat{V}_{\widehat{\pi}^*},\widehat{\pi}}}(\widehat{V}_{\widehat{\pi}^*})}\right\} \\ &\le \frac{2}{\sqrt{3}}\sqrt{\mathrm{Var}_{\underline{P}^{V_{\widehat{\pi}},\widehat{\pi}}}(V_{\widehat{\pi}})} + \left(\frac{2}{\sqrt{3}}\|\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}}\|_\infty + \frac{2\delta^{1/4}}{1-\gamma}\right)\mathbf{1}. \end{aligned} \tag{112}$$

These conclusions also hold when a nominal or empirical dual optimizer equals zero.

Proof. Fix a state-action pair (s,a). By equation 83, for every value function V,

$$\mathrm{Var}_{\widehat{P}^0_{s,a}}(V) \le \frac{5}{4}\mathrm{Var}_{P^0_{s,a}}(V), \qquad \mathrm{Var}_{P^0_{s,a}}(V) \le \frac{4}{3}\mathrm{Var}_{\widehat{P}^0_{s,a}}(V).$$

To see this, center under the distribution on the right and use $\frac{3}{4}P^0_{s,a} \le \widehat{P}^0_{s,a} \le \frac{5}{4}P^0_{s,a}$. Pinsker's inequality also shows that replacing either base row by any distribution in its KL uncertainty set changes the standard deviation of a value function by at most $\delta^{1/4}/[2^{1/4}(1-\gamma)]$. Indeed, for a center in $[0,(1-\gamma)^{-1}]$, the squared deviation lies in $[0,(1-\gamma)^{-2}]$, so its expectations differ by at most $\sqrt{\delta/2}/(1-\gamma)^2$; minimizing over the center and taking square roots gives the assertion.

Consequently,

$$
\begin{aligned}
&\sqrt{\mathrm{Var}_{\underline{\widehat{P}}_{s,a}^{\widehat{V}_{\pi^*}}}(\widehat{V}_{\pi^*})} \\
&\le \sqrt{\mathrm{Var}_{\widehat{P}^0_{s,a}}(\widehat{V}_{\pi^*})} + \frac{\delta^{1/4}}{2^{1/4}(1-\gamma)} \\
&\le \frac{2}{\sqrt{3}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}_{\pi^*})} + \frac{\delta^{1/4}}{2^{1/4}(1-\gamma)} \\
&\le \frac{2}{\sqrt{3}}\sqrt{\mathrm{Var}_{\underline{P}^{V_{\pi^*}}_{s,a}}(\widehat{V}_{\pi^*})} + \frac{(1+2/\sqrt{3})\delta^{1/4}}{2^{1/4}(1-\gamma)} \\
&\le \frac{2}{\sqrt{3}}\sqrt{\mathrm{Var}_{\underline{P}^{V_{\pi^*}}_{s,a}}(V_{\pi^*})} + \frac{2}{\sqrt{3}}\|\widehat{V}_{\pi^*} - V_{\pi^*}\|_\infty + \frac{2\delta^{1/4}}{1-\gamma}.
\end{aligned}
$$

The last step uses the triangle inequality for standard deviations and $(1+2/\sqrt{3})2^{-1/4} < 2$. The same argument applies to the nominal worst-case row at $\widehat{V}_{\pi^*}$, with both distributions compared to $P^0_{s,a}$. Choosing $a = \pi^*(s)$ proves equation 110. Interchanging the nominal and empirical rows and the two value functions proves equation 111. Replacing the two value functions by $\widehat{V}_{\widehat{\pi}^*}, V_{\widehat{\pi}}$ and choosing $a = \widehat{\pi}(s)$ proves equation 112. The argument also covers zero dual optimizers because their worst-case distributions belong to the same KL uncertainty sets. □

C.5 Variance transfer for small uncertainty

In the later part, we will provide a tighter bound when uncertainty level δ is small. Here, we first control the density ratio $\underline{P}^V(s'|s,a)/P^0(s'|s,a)$. The next two lemmas give tighter bounds from the KL constraint; combining them with the minimum nominal probability will produce the three branches used in the variance estimate. We begin with the exponential bound.

Lemma C.8 (Probability ratio bound under a KL constraint)**.** *Let p_1, p_2 be probability distributions on a finite set $\mathcal{X}$ with $\mathrm{KL}(p_1\|p_2) \le \delta$. Then, for every x with $p_2(x) > 0$,*

$$
p_1(x) \le p_2(x)\exp\left(1 + \frac{\delta}{p_2(x)}\right). \tag{113}
$$

Proof. Fix x with $p_2(x) > 0$. Since $\delta \ge 0$, the claim is immediate when $p_1(x) \le p_2(x)$. We therefore assume $p_1(x) > p_2(x) > 0$.

We first note that, for $u, v > 0$, the inequality $\log t \le t - 1$ gives

$$
u\log(u/v) = -u\log(v/u) \ge -u\left(\frac{v}{u} - 1\right) = u - v. \tag{114}
$$

This inequality also holds for $u = 0$ by continuity, with the convention $0\log(0/v) = 0$.

Let $\mathcal{X}_2 = \mathrm{supp}(p_2)$, where supp denotes the support. Finite KL divergence implies that $p_1(y) = 0$ for every $y \notin \mathcal{X}_2$. Applying the preceding scalar inequality to each $y \in \mathcal{X}_2 \setminus \{x\}$, we obtain

$$
\begin{aligned}
\delta &\ge \mathrm{KL}(p_1\|p_2) && (115)\\
&= p_1(x)\log\frac{p_1(x)}{p_2(x)} + \sum_{y\in\mathcal{X}_2\setminus\{x\}} p_1(y)\log\frac{p_1(y)}{p_2(y)} && (116)\\
&\ge p_1(x)\log\frac{p_1(x)}{p_2(x)} + \sum_{y\in\mathcal{X}_2\setminus\{x\}} \big(p_1(y) - p_2(y)\big) && (117)\\
&= p_1(x)\log\frac{p_1(x)}{p_2(x)} + p_2(x) - p_1(x). && (118)
\end{aligned}
$$

Dividing by $p_1(x) > 0$ and rearranging yields

$$
\begin{aligned}
\log\frac{p_1(x)}{p_2(x)} &\le \frac{\delta}{p_1(x)} + 1 - \frac{p_2(x)}{p_1(x)} && (119)\\
&\le \frac{\delta}{p_2(x)} + 1, && (120)
\end{aligned}
$$

where the last inequality uses $\delta \geq 0$ and $p_1(x) > p_2(x) > 0$. Exponentiating gives

$$p_1(x) \leq p_2(x) \exp\left(1 + \frac{\delta}{p_2(x)}\right), \tag{121}$$

which proves the claim. □

The preceding exponential bound is useful when δ is small relative to the nominal probability $p_2(x)$. We next derive an alternative bound using Pinsker's inequality, which controls $|p_1(x) - p_2(x)|$. Dividing by $p_2(x)$ yields a bound of $1 + \sqrt{\delta/2}/p_2(x)$ on the probability ratio. For nominal transition probabilities satisfying $p_2(x) \geq p_\wedge$, this gives the $\sqrt{\delta}/p_\wedge$ term in the subsequent density ratio estimate.

Lemma C.9 (Pinsker bound on probability differences)**.** *Let p_1 and p_2 be probability distributions on a finite set $\mathcal{X}$ satisfying $\mathrm{KL}(p_1 \| p_2) \leq \delta$. Then, for every $x \in \mathcal{X}$,*

$$|p_1(x) - p_2(x)| \leq \sqrt{\delta/2}. \tag{122}$$

Proof. For finite distributions, total variation has the equivalent form:

$$d_{\mathrm{TV}}(p_1, p_2) = \frac{1}{2} \sum_y |p_1(y) - p_2(y)| = \sup_B |p_1(B) - p_2(B)|. \tag{123}$$

Applying Pinsker's inequality yields

$$|p_1(x) - p_2(x)| \leq d_{\mathrm{TV}}(p_1, p_2) \leq \sqrt{\frac{\mathrm{KL}(p_1 \| p_2)}{2}} \leq \sqrt{\delta/2}. \tag{124}$$

This completes the proof. □

We now apply Lemmas C.8 and C.9 to the worst-case transition distribution $\underline{P}^V_{s,a}$ and the nominal distribution $P^0_{s,a}$. On the support of $P^0_{s,a}$, the inequalities $\underline{P}^V(s' \mid s, a) \leq 1$ and $P^0(s' \mid s, a) \geq p_\wedge$ give an additional upper bound of $1/p_\wedge$ on the density ratio. Combining these three bounds yields the following result.

Lemma C.10 (Density ratio bounds for worst-case transitions)**.** *Fix a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and a value vector V. Suppose that $\mathrm{KL}(\underline{P}^V_{s,a} \| P^0_{s,a}) \leq \delta$. Then the maximum density ratio on the support of $P^0_{s,a}$ satisfies*

$$\max_{\substack{s' \in \mathcal{S}: \\ P^0(s'|s,a) > 0}} \frac{\underline{P}^V(s' \mid s, a)}{P^0(s' \mid s, a)} \leq \min\left\{ \exp\left(1 + \frac{\delta}{p_\wedge}\right), 1 + \frac{\sqrt{\delta/2}}{p_\wedge}, \frac{1}{p_\wedge} \right\}. \tag{125}$$

Proof. Fix any $s' \in \mathcal{S}$ with $P^0(s' \mid s, a) > 0$. By the definition of $p_\wedge$, $P^0(s' \mid s, a) \geq p_\wedge > 0$.

Applying Lemma C.8 with $p_1 = \underline{P}^V_{s,a}$ and $p_2 = P^0_{s,a}$ gives

$$\frac{\underline{P}^V(s' \mid s, a)}{P^0(s' \mid s, a)} \leq \exp\left(1 + \frac{\delta}{P^0(s' \mid s, a)}\right) \tag{126}$$

$$\leq \exp\left(1 + \frac{\delta}{p_\wedge}\right). \tag{127}$$

Next, Lemma C.9, applied to the same distributions, yields

$$\underline{P}^V(s' \mid s, a) \leq P^0(s' \mid s, a) + \sqrt{\delta/2}. \tag{128}$$

Dividing by $P^0(s' \mid s, a)$, we obtain

$$\frac{\underline{P}^V(s' \mid s, a)}{P^0(s' \mid s, a)} \leq 1 + \frac{\sqrt{\delta/2}}{P^0(s' \mid s, a)} \tag{129}$$

$$\leq 1 + \frac{\sqrt{\delta/2}}{p_\wedge}. \tag{130}$$

Finally, since $\underline{P}^V(s' \mid s,a) \le 1$,

$$\frac{\underline{P}^V(s' \mid s,a)}{P^0(s' \mid s,a)} \le \frac{1}{P^0(s' \mid s,a)} \le \frac{1}{p_\wedge}. \tag{131}$$

These three bounds hold for every s' in the support of $P^0_{s,a}$, and their right-hand sides are independent of s'. Taking their minimum and then maximizing over s' in this support proves the claim. □

We next use Lemma C.10 to relate the variance of $h^{\lambda^*,V}_{s,a}$ under the nominal transition distribution to the variance of V under the worst-case transition distribution. The proof compares the exponential weight at $V(s')$ with its value at $\overline{V} = \mathbb{E}_{\underline{P}^V_{s,a}}[V]$. The mean-value theorem then cancels the explicit factor λ^* against the reciprocal factor from the derivative. A change of measure, together with the density ratio bound, controls the resulting weighted second moment in terms of $\mathrm{Var}_{\underline{P}^V_{s,a}}(V)$, while total variation bounds the remaining discrepancy between the nominal and worst-case second moments. The following lemma states this variance comparison and its simplified form for $0 < \delta \le 1$.

Lemma C.11 (Variance transfer)**.** *Let $0 < \delta \le 1$, and fix a state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and a value vector V. Suppose that $\lambda^* > 0$ is a maximizer of the population dual objective $f(P^0_{s,a}, V, \lambda)$ over $\lambda \ge 0$. Then*

$$\begin{aligned}&\lambda^* \sqrt{\mathrm{Var}_{P^0_{s,a}}\big(h^{\lambda^*,V}_{s,a}\big)} \\ &\le 8\sqrt{\left(1 + \min\left\{\exp\left(\frac{\delta}{p_\wedge}\right), \frac{\sqrt{\delta}}{p_\wedge}, \frac{1}{p_\wedge}\right\}\right)\mathrm{Var}_{\underline{P}^V_{s,a}}(V)} + 4\delta^{1/4}\,\mathrm{span}(V).\end{aligned} \tag{132}$$

Proof. As in the proof of Lemma C.5, suppress the dependence on the fixed quantities (s,a), V, and λ^*, and define

$$h(v) = \frac{e^{-v/\lambda^*}}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}]}, \qquad v \in \mathbb{R}. \tag{133}$$

Then $h(V(s')) = h^{\lambda^*,V}_{s,a}(s')$ and $h'(v) = -h(v)/\lambda^*$. Recall $\overline{V} = \mathbb{E}_{\underline{P}^V_{s,a}}[V]$. By Lemma C.1, $h(\overline{V}) = e^\delta$.

For each s' in the support of $P^0_{s,a}$, the mean-value theorem gives a point $\xi(s')$ between $V(s')$ and $\overline{V}$ such that

$$\lambda^*\big(h(V(s')) - h(\overline{V})\big) = -h(\xi(s'))\big(V(s') - \overline{V}\big). \tag{134}$$

Since h is positive and decreasing,

$$h(\xi(s'))^2 \le \max\{h(V(s'))^2, e^{2\delta}\} \le h(V(s'))^2 + e^{2\delta}. \tag{135}$$

Since variance is bounded above by the mean squared deviation from any constant, using $h(V(s')) = h^{\lambda^*,V}_{s,a}(s')$ gives

$$\begin{aligned}(\lambda^*)^2\,\mathrm{Var}_{P^0_{s,a}}\big(h^{\lambda^*,V}_{s,a}\big) &\le (\lambda^*)^2 \mathbb{E}_{s'\sim P^0_{s,a}}\big[(h(V(s')) - h(\overline{V}))^2\big] &(136)\\ &\le \mathbb{E}_{s'\sim P^0_{s,a}}\big[h(V(s'))^2 (V(s') - \overline{V})^2\big] \\ &\quad + e^{2\delta}\mathbb{E}_{s'\sim P^0_{s,a}}\big[(V(s') - \overline{V})^2\big]. &(137)\end{aligned}$$

Equation (137) follows from squaring equation 134, applying equation 135, and taking expectations with respect to $s' \sim P^0_{s,a}$.

To bound the first term in equation 137, let

$$H = \max_{s' : P^0(s'|s,a) > 0} h(V(s')). \tag{138}$$

Using the identity $\underline{P}^V(s' \mid s,a) = P^0(s' \mid s,a)h(V(s'))$, we have

$$\mathbb{E}_{s'\sim P^0_{s,a}}\left[h(V(s'))^2(V(s')-\overline{V})^2\right] = \mathbb{E}_{s'\sim \underline{P}^V_{s,a}}\left[h(V(s'))(V(s')-\overline{V})^2\right] \tag{139}$$

$$\leq H \operatorname{Var}_{\underline{P}^V_{s,a}}(V). \tag{140}$$

For the second term in equation 137, note that $\min_s V(s) \leq \overline{V} \leq \max_s V(s)$. Hence, for every $s' \in \mathcal{S}$, $0 \leq (V(s')-\overline{V})^2 \leq \operatorname{span}(V)^2$.

The bounded-function inequality for total variation and Pinsker's inequality therefore give

$$\mathbb{E}_{s'\sim P^0_{s,a}}\left[(V(s')-\overline{V})^2\right] \leq \operatorname{Var}_{\underline{P}^V_{s,a}}(V) + \operatorname{span}(V)^2 d_{\mathrm{TV}}(P^0_{s,a}, \underline{P}^V_{s,a})$$

$$\leq \operatorname{Var}_{\underline{P}^V_{s,a}}(V) + \operatorname{span}(V)^2\sqrt{\delta/2}, \tag{141}$$

where we used $\mathrm{KL}(\underline{P}^V_{s,a} \| P^0_{s,a}) = \delta$.

Substituting equation 140 into equation 137 and taking square roots yields the first inequality below. Applying equation 141 then gives the second. In both steps, we use $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$:

$$\lambda^* \sqrt{\operatorname{Var}_{P^0_{s,a}}\left(h^{\lambda^*,V}_{s,a}\right)} \leq \sqrt{H}\sqrt{\operatorname{Var}_{\underline{P}^V_{s,a}}(V)} + e^{\delta}\sqrt{\mathbb{E}_{s'\sim P^0_{s,a}}\left[(V(s')-\overline{V})^2\right]} \tag{142}$$

$$\leq (\sqrt{H}+e^{\delta})\sqrt{\operatorname{Var}_{\underline{P}^V_{s,a}}(V)} + e^{\delta}(\delta/2)^{1/4}\operatorname{span}(V). \tag{143}$$

Finally, set

$$K_\delta = 1 + \min\left\{\exp\left(\frac{\delta}{p_\wedge}\right), \frac{\sqrt{\delta}}{p_\wedge}, \frac{1}{p_\wedge}\right\}. \tag{144}$$

Lemma C.10 implies

$$H \leq \min\left\{\exp\left(1+\frac{\delta}{p_\wedge}\right), 1+\frac{\sqrt{\delta/2}}{p_\wedge}, \frac{1}{p_\wedge}\right\} \leq eK_\delta. \tag{145}$$

The last inequality follows by comparing the three terms with e times one plus their respective counterparts in the definition of K_δ. Since $0 < \delta \leq 1$ and $K_\delta \geq 1$,

$$\sqrt{H} + e^{\delta} \leq \sqrt{eK_\delta} + e \leq (\sqrt{e}+e)\sqrt{K_\delta} \leq 8\sqrt{K_\delta}. \tag{146}$$

Also,

$$e^{\delta}(\delta/2)^{1/4} \leq e\,2^{-1/4}\delta^{1/4} \leq 4\delta^{1/4}. \tag{147}$$

Substituting these bounds and recalling $h(V(s')) = h^{\lambda^*,V}_{s,a}(s')$ proves equation 132. □

Empirical version of the variance bound in Lemma C.11. The preceding proof also yields a variance bound for the empirical transition kernel.

Let $0 < \delta \leq 1$, and fix a state-action pair (s,a) and a value vector V. Suppose that $\widehat{\lambda}^* > 0$ maximizes the empirical dual objective $f(\widehat{P}^0_{s,a}, V, \lambda)$ over $\lambda \geq 0$. Recall that the corresponding weight function and worst-case transition distribution are given by

$$\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}(s') = \frac{e^{-V(s')/\widehat{\lambda}^*}}{\mathbb{E}_{s''\sim\widehat{P}^0_{s,a}}\left[e^{-V(s'')/\widehat{\lambda}^*}\right]}, \tag{148}$$

$$\widehat{\underline{P}}^V(s' \mid s,a) = \widehat{P}^0(s' \mid s,a)\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}(s'). \tag{149}$$

Define the minimum positive empirical transition probability as

$$\widehat{p}_\wedge = \min_{\substack{s,a,s':\\ \widehat{P}^0(s'|s,a)>0}} \widehat{P}^0(s' \mid s,a). \tag{150}$$

The normalization and stationarity arguments in Lemma C.1 give

$$\mathbb{E}_{s'\sim\widehat{P}^0_{s,a}}\left[\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}(s')\right] = 1, \tag{151}$$

$$\mathrm{KL}\big(\widehat{\underline{P}}^V_{s,a}\|\widehat{P}^0_{s,a}\big) = \delta. \tag{152}$$

Thus, for each realization of the transition samples, Lemma C.11 applies with $P^0_{s,a}$, λ^*, and $p_\wedge$ replaced by $\widehat{P}^0_{s,a}$, $\widehat{\lambda}^*$, and $\widehat{p}_\wedge$, respectively.

On the event in equation 48, we further have $\widehat{p}_\wedge \geq (3/4)p_\wedge$. Consequently, on this event,

$$\begin{aligned}
&\widehat{\lambda}^*\sqrt{\mathrm{Var}_{\widehat{P}^0_{s,a}}\big(\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}\big)} \\
&\leq 8\sqrt{\left(1+\min\left\{\exp\left(\frac{\delta}{\widehat{p}_\wedge}\right),\frac{\sqrt{\delta}}{\widehat{p}_\wedge},\frac{1}{\widehat{p}_\wedge}\right\}\right)\mathrm{Var}_{\widehat{\underline{P}}^V_{s,a}}(V)} + 4\delta^{1/4}\,\mathrm{span}(V) \qquad (153)\\
&\leq 8\sqrt{\left(1+\min\left\{\exp\left(\frac{4\delta}{3p_\wedge}\right),\frac{4\sqrt{\delta}}{3p_\wedge},\frac{4}{3p_\wedge}\right\}\right)\mathrm{Var}_{\widehat{\underline{P}}^V_{s,a}}(V)} + 4\delta^{1/4}\,\mathrm{span}(V). \qquad (154)
\end{aligned}$$

The second inequality follows from $1/\widehat{p}_\wedge \leq 4/(3p_\wedge)$ and the monotonicity of each term inside the minimum.

This variance bound holds deterministically for each realization of the empirical model. Here, $\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}$ is normalized under $\widehat{P}^0_{s,a}$, whereas $h^{\widehat{\lambda}^*,V}_{s,a}$ in equation 50 is normalized under $P^0_{s,a}$. Applying the present bound in the statistical error analysis therefore requires a comparison of these normalizations and the corresponding variances, together with a valid concentration argument for quantities selected using the samples.

C.6 A resolvent bound for value variance

Lemma C.12. *Let $0<\gamma<1$, let P be a transition matrix on $\mathcal{S}$, and let $r\in[0,1]^{|\mathcal{S}|}$ be a reward vector. Let V be the unique solution to the Bellman equation: $V = r+\gamma PV$, and define*

$$\mathrm{Var}_P(V) = P(V^2) - (PV)^2.$$

Then

$$\left\|(I-\gamma P)^{-1}\sqrt{\mathrm{Var}_P(V)}\right\|_\infty \leq \frac{\sqrt{1+\gamma}}{2\gamma(1-\gamma)^{3/2}}, \tag{155}$$

where the squares and square root are taken componentwise.

Proof. Let $(s_t)_{t\geq 0}$ be a Markov chain with transition matrix P, and define the discounted return

$$G = \sum_{t=0}^{\infty}\gamma^t r(s_t).$$

The Bellman equation implies that $\mathbb{E}[G \mid s_0 = s] = V(s)$. Write

$$W(s) = \mathrm{Var}(G \mid s_0 = s), \qquad u = \mathrm{Var}_P(V).$$

Since $r(s_0)$ is deterministic conditional on s_0, the Markov property and the law of total variance give

$$W(s) = \gamma^2\mathbb{E}[W(s_1)\mid s_0=s] + \gamma^2\,\mathrm{Var}(V(s_1)\mid s_0=s). \tag{156}$$

Equivalently, $W = \gamma^2 PW + \gamma^2 u$, and hence

$$W = \gamma^2(I-\gamma^2P)^{-1}u.$$

Because $0\leq r\leq 1$, the return satisfies $0\leq G\leq (1-\gamma)^{-1}$. Therefore, for every state s,

$$W(s) \leq \mathbb{E}\left[\left(G-\frac{1}{2(1-\gamma)}\right)^2\middle| s_0=s\right] \leq \frac{1}{4(1-\gamma)^2}.$$

It follows that

$$(I-\gamma^2 P)^{-1}u \le \frac{\mathbf{1}}{4\gamma^2(1-\gamma)^2}, \tag{157}$$

where the inequality is componentwise.

Next, the matrix inverse identity gives

$$\begin{aligned}(I-\gamma P)^{-1} &= (I-\gamma^2 P)^{-1} \\ &\quad + \gamma(1-\gamma)(I-\gamma P)^{-1}P(I-\gamma^2 P)^{-1}.\end{aligned} \tag{158}$$

The inverse matrices have nonnegative entries, and $P\mathbf{1} = \mathbf{1}$ implies

$$(I-\gamma P)^{-1}P\mathbf{1} = \frac{\mathbf{1}}{1-\gamma}.$$

Applying the matrix identity to u and using equation 157, we obtain

$$\begin{aligned}(I-\gamma P)^{-1}u &\le \frac{\mathbf{1} + \gamma(1-\gamma)(I-\gamma P)^{-1}P\mathbf{1}}{4\gamma^2(1-\gamma)^2} \\ &= \frac{1+\gamma}{4\gamma^2(1-\gamma)^2}\mathbf{1}.\end{aligned} \tag{159}$$

Finally, write $R = (I-\gamma P)^{-1}$. Its entries are nonnegative, and each row sums to $(1-\gamma)^{-1}$. Thus, for every state s, the Cauchy–Schwarz inequality gives

$$\begin{aligned}\left[(R\sqrt{u})(s)\right]^2 &= \left(\sum_{s'} R(s,s')\sqrt{u(s')}\right)^2 \\ &\le \left(\sum_{s'} R(s,s')\right)\left(\sum_{s'} R(s,s')u(s')\right) \\ &= \frac{(Ru)(s)}{1-\gamma} \\ &\le \frac{1+\gamma}{4\gamma^2(1-\gamma)^3},\end{aligned} \tag{160}$$

where the last inequality follows from equation 159. Taking square roots and maximizing over s proves equation 155. □

The bound requires $V = r + \gamma PV$ with $r \in [0,1]^{|\mathcal{S}|}$, where the same matrix P appears in both the resolvent and $\mathrm{Var}_P(V)$. When P is an empirical worst-case transition matrix and V is a population robust value function, this Bellman equation need not hold. Applying the bound to such a pair therefore requires an additional comparison argument.

C.7 Sample-complexity analysis for small uncertainty level: $\delta \le 1$.

Theorem C.13 (Policy guarantee for $0 < \delta \le 1$)**.** *Let $0 < \gamma < 1$ and $\delta, \eta \in (0,1)$. Suppose that rewards are known and belong to $[0,1]$, and N samples are drawn independently from the nominal transition distribution for every state–action pair, with independence between the samples from different pairs. Let $\widehat{\pi}$ be the output policy after T iterations of robust value iteration with $Q_0 = 0$, using the empirical transition kernel and a fixed rule to choose among actions attaining the same maximum. If*

$$\begin{aligned}N \ge \max\Bigg\{ &\frac{192}{p_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}, \\ &\frac{65536}{(1-\gamma)^2}\left(1+\min\left\{e^{2\delta/p_\wedge}, \frac{2\sqrt{\delta}}{p_\wedge}+1, \frac{2}{p_\wedge}\right\}\right)\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}\Bigg\},\end{aligned} \tag{161}$$

then, with probability at least $1-\eta$*,*

$$\|V_{\pi^*}-V_{\widehat{\pi}}\|_\infty \le 1024\sqrt{\frac{1}{N}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\left(\frac{\sqrt{\min\left\{e^{2\delta/p_\wedge},\,\frac{2\sqrt{\delta}}{p_\wedge}+1,\,\frac{2}{p_\wedge}\right\}}}{(1-\gamma)^{3/2}}+\frac{\delta^{1/4}}{(1-\gamma)^2}\right)$$
$$+\frac{512}{p_\wedge N(1-\gamma)^2}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}+\frac{8\gamma^T}{(1-\gamma)^2}. \tag{162}$$

Proof. Work on the event of Lemma C.6. Nominal and empirical optimal policies are denoted by π^* and $\widehat{\pi}^*$, respectively. Every value function below belongs to $[0,(1-\gamma)^{-1}]^{|\mathcal{S}|}$.

One-step estimation error.

Fix (s,a) and a value vector V listed in Lemma C.6. For $\widehat{\lambda}^*>0$, the relative probability bound equation 83 implies

$$\widehat{\lambda}^*\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\widehat{\lambda}^*,V}_{s,a})}\le\frac{5}{2\sqrt{3}}\widehat{\lambda}^*\sqrt{\mathrm{Var}_{\widehat{P}^0_{s,a}}(\widehat{h}^{\widehat{\lambda}^*,V}_{s,a})}.$$

Indeed, the relative probability bound gives $\mathrm{Var}_{P^0_{s,a}}(h^{\widehat{\lambda}^*,V}_{s,a})\le\frac{4}{3}\mathrm{Var}_{\widehat{P}^0_{s,a}}(h^{\widehat{\lambda}^*,V}_{s,a})$. Moreover, $h^{\widehat{\lambda}^*,V}_{s,a}$ equals $\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}$ multiplied by the ratio of the empirical to nominal normalizing expectations, which is at most $5/4$. Apply Lemma C.11 to the nominal and empirical distributions at their respective positive dual optimizers, using $\widehat{p}_\wedge\ge(3/4)p_\wedge$ for the empirical distribution. Combining these bounds with the preceding comparison and equation 84 gives, for every required V and (s,a), equations (83), (106), and (107), together with $|\log x|\le(4/3)|x-1|$ on $[3/4,5/4]$, give

$$|f(\widehat{P}^0_{s,a},V,\lambda)-f(P^0_{s,a},V,\lambda)|\le\sqrt{\frac{1}{N}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\min\left\{4\lambda\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a})},\right.$$
$$\left.\frac{11}{2}\lambda\sqrt{\mathrm{Var}_{\widehat{P}^0_{s,a}}(\widehat{h}^{\lambda,V}_{s,a})}\right\}+\frac{54}{p_\wedge N(1-\gamma)}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}. \tag{163}$$

The normalization comparison in (83) and (107) contributes $5/(2\sqrt{3})$ to the empirical coefficient. Terms at $\lambda=0$ use continuity; scalar weights follow (133).

For every $\lambda>0$,

$$\lambda\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a})}\le3\lambda^*\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a})}+\frac{3\delta^{1/4}}{1-\gamma}+\sqrt{\frac{2}{p_\wedge(1-\gamma)}\left[\underline{P}^V_{s,a}V-f(P^0_{s,a},V,\lambda)\right]}. \tag{164}$$

The bracket is nonnegative.

Since $x(\log x)^2\le(x-1)^2$, normalization and the exponential-tilt identity imply

$$\sqrt{\mathrm{Var}_{\underline{P}^V_{s,a}}(V)}\le\lambda^*\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a})}. \tag{165}$$

Both sides vanish when $\lambda^*=0$.

For $\lambda\ge\lambda^*>0$, Lemma C.1 gives $h^{\lambda,V}_{s,a}(\underline{P}^V_{s,a}V)\le e^\delta=h^{\lambda^*,V}_{s,a}(\underline{P}^V_{s,a}V)$. Below $\underline{P}^V_{s,a}V$, monotonicity of $\lambda(e^{x/\lambda}-1)$ bounds the centered weight difference, multiplied by λ, by $\lambda^*|h^{\lambda^*,V}_{s,a}-1|$; above it, the mean-value theorem bounds that difference by $e^\delta|V-\underline{P}^V_{s,a}V|$. Centering the variance at $h^{\lambda,V}_{s,a}(\underline{P}^V_{s,a}V)$ and applying (141) and equation 165 therefore give

$$\lambda\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a})}\le\sqrt{1+e^{2\delta}}\,\lambda^*\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a})}+\frac{e^\delta(\delta/2)^{1/4}}{1-\gamma}.$$

Since $\sqrt{1+e^2}<3$, this proves equation 164 in this case. For $\lambda^*=0$, the first region contributes zero and the same argument applies.

Next suppose that $0 < \lambda < \lambda^*$ and $\lambda \le (1-\gamma)^{-1}$. The exponential-tilt identities give

$$\begin{aligned}&\lambda \,\mathrm{KL}\left(\underline{P}^V_{s,a} \,\middle\|\, P^0_{s,a} h^{\lambda,V}_{s,a}\right)\\ &\quad = \underline{P}^V_{s,a} V - f(P^0_{s,a}, V, \lambda).\end{aligned}$$

The triangle inequality in $L^2(P^0_{s,a})$, followed by Pinsker's inequality and $P^0(s' \mid s,a) \ge p_\wedge$ on the support, yields

$$\begin{aligned}&\lambda \sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a})}\\ &\le \frac{\lambda}{\lambda^*}\, \lambda^* \sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a})}\\ &\quad + \sqrt{\frac{2\lambda}{p_\wedge}\left[\underline{P}^V_{s,a} V - f(P^0_{s,a}, V, \lambda)\right]}.\end{aligned}$$

This proves equation 164 in this case. For $\lambda > (1-\gamma)^{-1}$, the weight is at most e; the mean-value theorem, (141), and equation 165 bound the left side by the first two terms in equation 164. This completes the estimate, whose empirical version uses $\widehat{p}_\wedge \ge 3p_\wedge/4$.

Retain the dual optimization loss:

$$\begin{aligned}&\widehat{\underline{P}}^V_{s,a} V - \underline{P}^V_{s,a} V\\ &= f(\widehat{P}^0_{s,a}, V, \widehat{\lambda}^*) - f(P^0_{s,a}, V, \widehat{\lambda}^*)\\ &\quad - \left[\underline{P}^V_{s,a} V - f(P^0_{s,a}, V, \widehat{\lambda}^*)\right].\end{aligned}$$

Apply equation 163 and equation 164, and complete the square in the square root of the bracket. For the opposite sign, evaluate the empirical objective at λ^*. Interchanging the centers gives the empirical bound; its quadratic remainder has coefficient $121/6$, compared with 8 for the nominal bound. Since $54 + 121/6 < 75$, both yield

$$\begin{aligned}&\left|\widehat{\underline{P}}^V_{s,a} V - \underline{P}^V_{s,a} V\right|\\ &\le \frac{33}{2}\sqrt{\frac{1}{N}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\left[\min\left\{\lambda^*\sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a})},\right.\right.\\ &\qquad\qquad \left.\left.\widehat{\lambda}^*\sqrt{\mathrm{Var}_{\widehat{P}^0_{s,a}}(\widehat{h}^{\widehat{\lambda}^*,V}_{s,a})}\right\} + \frac{\delta^{1/4}}{1-\gamma}\right]\\ &\quad + \frac{75}{p_\wedge N(1-\gamma)}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}. \qquad (166)\end{aligned}$$

Comparison at different value vectors.

For every $V, W \in [0, (1-\gamma)^{-1}]^{|\mathcal{S}|}$,

$$\begin{aligned}&\lambda^* \sqrt{\mathrm{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a})}\\ &\quad \le 5\sqrt{1 + \min\left\{e^{2\delta/p_\wedge},\, 1 + \frac{2\sqrt{\delta}}{p_\wedge},\, \frac{2}{p_\wedge}\right\}}\left[\sqrt{\mathrm{Var}_{\underline{P}^W_{s,a}}(W)} + 2\|V - W\|_\infty\right] + \frac{7\delta^{1/4}}{1-\gamma}. \qquad (167)\end{aligned}$$

The same estimate holds with empirical centers, weights, and worst-case rows.

By Lemma C.10 and $\widehat{p}_\wedge \ge 3p_\wedge/4$ from (83), the maximum optimized density for either center is bounded by

$$e\left(1 + \min\left\{e^{2\delta/p_\wedge},\, 1 + \frac{2\sqrt{\delta}}{p_\wedge},\, \frac{2}{p_\wedge}\right\}\right). \qquad (168)$$

We use the refined variance-transfer estimate (143) from the proof of Lemma C.11, with $\mathrm{span}(V) \le (1-\gamma)^{-1}$. At a zero optimizer the optimized influence and value variance are zero, and any density is interpreted by its zero-temperature limit.

Consider $V + t(W - V)$, $0 \le t \le 1$, and let λ_t^* be its dual optimizer. Wherever this optimizer is positive and the value variance is positive, differentiating the active KL constraint gives

$$\frac{d\lambda_t^*}{dt} = \lambda_t^* \frac{\mathrm{Cov}_{\underline{P}_{s,a}^{V+t(W-V)}}(V + t(W-V), W - V)}{\mathrm{Var}_{\underline{P}_{s,a}^{V+t(W-V)}}(V + t(W-V))}.$$

Differentiate the $L^2(P_{s,a}^0)$ norm of $\lambda_t^*(h_{s,a}^{\lambda_t^*, V+t(W-V)} - 1)$. Cauchy–Schwarz and the preceding identity yield

$$\begin{aligned}
&\left| \frac{d}{dt} \left[\lambda_t^* \sqrt{\mathrm{Var}_{P_{s,a}^0}(h_{s,a}^{\lambda_t^*, V+t(W-V)})} \right] \right| \\
&\quad \le \left[\frac{\lambda_t^* \sqrt{\mathrm{Var}_{P_{s,a}^0}(h_{s,a}^{\lambda_t^*, V+t(W-V)})}}{\sqrt{\mathrm{Var}_{\underline{P}_{s,a}^{V+t(W-V)}}(V + t(W-V))}} + 2\sqrt{\max_{s'} h_{s,a}^{\lambda_t^*, V+t(W-V)}(s')} \right] \|W - V\|_\infty. \quad (169)
\end{aligned}$$

Here the weight derivative contributes one maximum-density term; the optimizer derivative contributes the ratio and the second maximum-density term.

Whenever the influence on the left of equation (143) exceeds $7\delta^{1/4}/(1-\gamma)$, its ratio to the worst-case standard deviation is at most

$$\frac{3}{2}\left(\sqrt{\max_{s'} h_{s,a}^{\lambda_t^*, V+t(W-V)}(s')} + e^{\delta} \right),$$

because $3e^{\delta} 2^{-1/4} < 7$. Equations 168 and equation 169, together with $(7/2)\sqrt{e} + (3/2)e < 10$, therefore bound the derivative in that region by

$$10\sqrt{1 + \min\left\{ e^{2\delta/p_\wedge},\, 1 + \frac{2\sqrt{\delta}}{p_\wedge},\, \frac{2}{p_\wedge} \right\}} \|W - V\|_\infty.$$

The optimizer is unique in $[0, 1/(\delta(1-\gamma))]$. The influence is continuous, including at zero optimizers, since the density is bounded by the reciprocal of the minimum positive base probability. Hence clipping it from below at $7\delta^{1/4}/(1-\gamma)$ gives the same Lipschitz bound on the entire segment. At W, (143), $\sqrt{e} + e < 5$, and $e2^{-1/4} < 7$ give the endpoint bound in equation 167, completing the estimate for both centers.

Bellman comparisons.

Monotonicity, contraction, and $Q_0 = 0$ give $0 \le Q_T \le \widehat{Q}^*$ and $\|\widehat{Q}^* - Q_T\|_\infty \le \gamma^T/(1-\gamma)$. Since $\widehat{\pi}$ is greedy for Q_T, its optimal empirical Bellman residual satisfies

$$0 \le \widehat{V}_{\widehat{\pi}^*} - r(\cdot, \widehat{\pi}(\cdot)) - \gamma \widehat{\underline{P}}^{\widehat{V}_{\widehat{\pi}^*}, \widehat{\pi}} \widehat{V}_{\widehat{\pi}^*} \le \frac{\gamma^T}{1-\gamma}\mathbf{1}.$$

Subtract the population Bellman equation for $V_{\widehat{\pi}}$ and use the nominal minimizing property to obtain

$$\begin{aligned}
\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}} &\le (I - \gamma \underline{P}^{V_{\widehat{\pi}}, \widehat{\pi}})^{-1} \left[\frac{\gamma^T}{1-\gamma}\mathbf{1} \right. \\
&\qquad \left. + \gamma \left(\widehat{\underline{P}}^{\widehat{V}_{\widehat{\pi}^*}, \widehat{\pi}} - \underline{P}^{\widehat{V}_{\widehat{\pi}^*}, \widehat{\pi}} \right) \widehat{V}_{\widehat{\pi}^*} \right]. \quad (170)
\end{aligned}$$

Propagation and absorption.

Apply equation 166 and equation 167 to these two comparisons and equation 170 with the respective pairs

$$(\widehat{V}_{\pi^*}, V_{\pi^*}), \qquad (V_{\pi^*}, \widehat{V}_{\pi^*}), \qquad (\widehat{V}_{\widehat{\pi}^*}, V_{\widehat{\pi}}).$$

Choose the nominal term in the first and third comparisons and the empirical term in the second. The resulting coefficients are bounded by 83, 165, 132, and 75, using $(33/2) \cdot 5 < 83$, $(33/2) \cdot 10 = 165$,

and $(33/2)(7+1) = 132$. Lemma C.12 bounds each matched variance after discounted propagation by $1/(\sqrt{2}(1-\gamma)^{3/2})$; the inverses map $\mathbf{1}$ to $\mathbf{1}/(1-\gamma)$. Population and empirical optimality give $V_{\widehat{\pi}} - \widehat{V}_{\widehat{\pi}^*} \le V_{\pi^*} - \widehat{V}_{\pi^*}$, so the two signed comparisons for π^* also control the negative part of the third comparison. Taking the maximum yields

$$\begin{aligned}
&\max\left\{\|V_{\pi^*} - \widehat{V}_{\pi^*}\|_\infty, \|\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}}\|_\infty\right\} \\
&\le \sqrt{\frac{1 + \min\left\{e^{2\delta/p_\wedge},\, 1 + \dfrac{2\sqrt{\delta}}{p_\wedge},\, \dfrac{2}{p_\wedge}\right\}}{N} \log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}} \\
&\quad \cdot \left[\frac{83}{\sqrt{2}(1-\gamma)^{3/2}} + \frac{165}{1-\gamma}\max\left\{\|V_{\pi^*} - \widehat{V}_{\pi^*}\|_\infty, \|\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}}\|_\infty\right\}\right] \\
&\quad + \frac{132\delta^{1/4}}{(1-\gamma)^2}\sqrt{\frac{1}{N}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}} + \frac{75}{p_\wedge N(1-\gamma)^2}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge} + \frac{\gamma^T}{(1-\gamma)^2}. \qquad (171)
\end{aligned}$$

The coefficient of the repeated error maximum in equation 171 is at most $165/256 < 1$. Moving it to the left multiplies the remaining terms by at most $256/91$.

Empirical optimality bounds the policy loss by twice this error maximum, giving a total multiplier $512/91$. Since

$$\frac{512\cdot 83}{91} < 1024, \qquad \frac{512\cdot 132}{91} < 1024, \qquad \frac{512\cdot 75}{91} < 512, \qquad \frac{512}{91} < 8.$$

the claimed bound follows on the event of probability at least $1-\eta$. □

C.8 Sample-complexity analysis for large uncertainty level: $\delta > 1$

Lemma C.14 (Variance bound for $\delta > 1$)**.** *Fix (s,a) and $\delta > 1$. Suppose every positive probability in $P^0_{s,a}$ is at least $p_\wedge$, and let λ^* maximize $f(P^0_{s,a}, V, \lambda)$ over $\lambda \ge 0$. Use the worst-case transition $\underline{P}^V_{s,a}$ and weight $h^{\lambda^*,V}_{s,a}$ defined in equation 18. Then*

$$(\lambda^*)^2 \operatorname{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a}) \le \frac{2}{p_\wedge}\operatorname{Var}_{\underline{P}^V_{s,a}}(V) + 8e^\delta \operatorname{span}(V)^2. \qquad (172)$$

At $\lambda^* = 0$, interpret the left-hand side as its limit as $\lambda \downarrow 0$. That limit is zero, so equation 172 remains valid.

Proof. Suppose first that $\lambda^* > 0$. The dual first-order condition and the definition of $h^{\lambda^*,V}_{s,a}$ imply

$$\log h^{\lambda^*,V}_{s,a} = \delta - \frac{V - \mathbb{E}_{\underline{P}^V_{s,a}}[V]}{\lambda^*},$$

$$\mathbb{E}_{\underline{P}^V_{s,a}}[\log h^{\lambda^*,V}_{s,a}] = \delta, \qquad (173)$$

$$\operatorname{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a}) = \mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}] - 1. \qquad (174)$$

The last identity uses $\mathbb{E}_{P^0_{s,a}}[h^{\lambda^*,V}_{s,a}] = 1$ and $P^0_{s,a}h^{\lambda^*,V}_{s,a} = \underline{P}^V_{s,a}$. Moreover, $h^{\lambda^*,V}_{s,a} \le p_\wedge^{-1}$ on the nominal support. Jensen's inequality applied to the dual normalizer gives

$$\begin{aligned}
\lambda^*\delta &= -\lambda^* \log \mathbb{E}_{P^0_{s,a}}[e^{-V/\lambda^*}] - \mathbb{E}_{\underline{P}^V_{s,a}}[V] \\
&\le \mathbb{E}_{P^0_{s,a}}[V] - \mathbb{E}_{\underline{P}^V_{s,a}}[V] \le \operatorname{span}(V). \qquad (175)
\end{aligned}$$

Define

$$\psi = \log \mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}] - \delta.$$

Jensen's inequality and equation 173 give $\psi \ge 0$. We consider the following two cases.

Case 1: $\psi > 2$**.** Then $\mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}] = e^{\delta+\psi} > e^{\delta+2}$. Splitting this expectation at half its value gives

$$\begin{aligned}&\mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}]\\ &\le \frac{1}{2}\mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}] + \frac{1}{p_\wedge}\Pr_{\underline{P}^V_{s,a}}\left(h^{\lambda^*,V}_{s,a} \ge \frac{1}{2}\mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}]\right).\end{aligned}$$

Hence

$$\Pr_{\underline{P}^V_{s,a}}\left(h^{\lambda^*,V}_{s,a} \ge \frac{1}{2}\mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}]\right) \ge \frac{p_\wedge}{2}\mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}]. \tag{176}$$

On this event, $\log h^{\lambda^*,V}_{s,a} - \delta \ge \psi - \log 2 > 1$. By equation 173,

$$\begin{aligned}\frac{\operatorname{Var}_{\underline{P}^V_{s,a}}(V)}{(\lambda^*)^2} &= \mathbb{E}_{\underline{P}^V_{s,a}}[(\log h^{\lambda^*,V}_{s,a} - \delta)^2]\\ &\ge \Pr_{\underline{P}^V_{s,a}}\left(h^{\lambda^*,V}_{s,a} \ge \frac{1}{2}\mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}]\right)\\ &\ge \frac{p_\wedge}{2}\mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}].\end{aligned}$$

Combining this with equation 174 bounds $(\lambda^*)^2 \operatorname{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a})$ by $2\operatorname{Var}_{\underline{P}^V_{s,a}}(V)/p_\wedge$.

Case 2: $0 \le \psi \le 2$**.** Since $\mathbb{E}_{\underline{P}^V_{s,a}}[h^{\lambda^*,V}_{s,a}] = e^{\delta+\psi} \le e^{\delta+2}$, we obtain

$$\begin{aligned}(\lambda^*)^2 \operatorname{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a}) &\le (\lambda^*)^2(e^{\delta+2}-1)\\ &\le \frac{e^{\delta+2}-1}{\delta^2}\operatorname{span}(V)^2 \le 8e^\delta \operatorname{span}(V)^2,\end{aligned}$$

where the second inequality follows from equation 175, and the last uses $\delta > 1$ and $e^2 < 8$. In either case, the sum of the two nonnegative terms on the right-hand side of equation 172 bounds the left-hand side.

For every $\lambda > 0$, the nominally normalized weight satisfies $h^{\lambda,V}_{s,a} \le p_\wedge^{-1}$, and therefore

$$0 \le \lambda^2 \operatorname{Var}_{P^0_{s,a}}(h^{\lambda,V}_{s,a}) \le \lambda^2(p_\wedge^{-1}-1) \longrightarrow 0 \qquad \text{as } \lambda \downarrow 0.$$

If $\lambda^* = 0$, strong duality gives

$$\mathbb{E}_{\underline{P}^V_{s,a}}[V] = f(P^0_{s,a}, V, 0) = \min_{s':P^0(s'|s,a)>0} V(s').$$

Finite KL divergence restricts the worst-case distribution to the nominal support. The displayed equality therefore forces it to assign mass only to minimizing states, so $\operatorname{Var}_{\underline{P}^V_{s,a}}(V) = 0$. This argument does not use stationarity at zero. □

Lemma C.15 (Uniform estimation for $\delta > 1$)**.** *Fix $\delta > 1$ and $0 < \gamma < 1$. Suppose rewards are known and in $[0,1]$, and N samples are drawn independently from each nominal transition $P^0_{s,a}$, independently across state–action pairs. For $0 < \eta < 1$, assume*

$$N \ge \frac{192}{p_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}. \tag{177}$$

With probability at least $1-\eta$, the supports agree, $3P^0/4 \le \widehat{P}^0 \le 5P^0/4$, and equation 84 holds simultaneously for every (s,a) and $V \in \{V_{\pi^*}, \widehat{V}_{\pi^*}, \widehat{V}_{\widehat{\pi}^*}\}$. On this event,

$$\begin{aligned}&\left|\underline{\widehat{P}}^V(\cdot \mid s,a)V - \underline{P}^V(\cdot \mid s,a)V\right|\\ &\le 16\sqrt{\frac{1}{Np_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\min\left\{\sqrt{\operatorname{Var}_{\underline{P}^V(\cdot|s,a)}(V)}, \sqrt{\operatorname{Var}_{\underline{\widehat{P}}^V(\cdot|s,a)}(V)}\right\}\\ &\quad + 21\sqrt{\frac{e^\delta}{N}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\operatorname{span}(V)\\ &\quad + \frac{652}{p_\wedge N(1-\gamma)}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}.\end{aligned} \tag{178}$$

Proof. The proof of Lemma C.6 applies to each fixed $\delta > 1$: its support estimate, Bernstein inequality, and covering grids do not use $\delta \le 1$. The absorbing-state values remain bounded by $(1-\gamma)^{-1}$ and satisfy the same Bellman contraction. In particular, the grid argument is uniform in $\lambda \ge 0$ and covers both data-dependent optimizers. Thus its stated support and concentration conclusions hold on a single event of probability at least $1-\eta$.

For $\widehat{\lambda}^* > 0$, the relative probability bounds give

$$\frac{3}{4} \le \frac{\mathbb{E}_{\widehat{P}^0_{s,a}}[e^{-V/\widehat{\lambda}^*}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\widehat{\lambda}^*}]} \le \frac{5}{4}. \tag{179}$$

The nominal and empirical weights at this optimizer satisfy

$$h^{\widehat{\lambda}^*,V}_{s,a} = \frac{\mathbb{E}_{\widehat{P}^0_{s,a}}[e^{-V/\widehat{\lambda}^*}]}{\mathbb{E}_{P^0_{s,a}}[e^{-V/\widehat{\lambda}^*}]} \widehat{h}^{\widehat{\lambda}^*,V}_{s,a}.$$

Since $\mathbb{E}_{\widehat{P}^0_{s,a}}[\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}] = 1$ and $P^0_{s,a} \le (4/3)\widehat{P}^0_{s,a}$,

$$\begin{aligned}
\operatorname{Var}_{P^0_{s,a}}(\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}) &\le \mathbb{E}_{P^0_{s,a}}[(\widehat{h}^{\widehat{\lambda}^*,V}_{s,a} - 1)^2] \\
&\le \frac{4}{3}\mathbb{E}_{\widehat{P}^0_{s,a}}[(\widehat{h}^{\widehat{\lambda}^*,V}_{s,a} - 1)^2] \\
&= \frac{4}{3}\operatorname{Var}_{\widehat{P}^0_{s,a}}(\widehat{h}^{\widehat{\lambda}^*,V}_{s,a}).
\end{aligned}$$

Together with equation 179, this yields

$$\begin{aligned}
&\widehat{\lambda}^* \sqrt{\operatorname{Var}_{P^0_{s,a}}(h^{\widehat{\lambda}^*,V}_{s,a})} \\
&\le \frac{5}{2\sqrt{3}} \widehat{\lambda}^* \sqrt{\operatorname{Var}_{\widehat{P}^0_{s,a}}(\widehat{h}^{\widehat{\lambda}^*,V}_{s,a})}.
\end{aligned} \tag{180}$$

Apply Lemma C.14 to the nominal and empirical optimizers, using the probability floors $p_\wedge$ and $3p_\wedge/4$, respectively. Taking square roots and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for $x, y \ge 0$ gives

$$\begin{aligned}
&\lambda^* \sqrt{\operatorname{Var}_{P^0_{s,a}}(h^{\lambda^*,V}_{s,a})} \\
&\le \sqrt{\frac{2}{p_\wedge}} \sqrt{\operatorname{Var}_{\underline{P}^V_{s,a}}(V)} + \sqrt{8e^\delta}\operatorname{span}(V), \\
&\widehat{\lambda}^* \sqrt{\operatorname{Var}_{P^0_{s,a}}(h^{\widehat{\lambda}^*,V}_{s,a})} \\
&\le \frac{5}{2\sqrt{3}} \left(\sqrt{\frac{8}{3p_\wedge}} \sqrt{\operatorname{Var}_{\underline{\widehat{P}}^V_{s,a}}(V)} + \sqrt{8e^\delta}\operatorname{span}(V) \right),
\end{aligned}$$

where the second estimate also uses equation 180. The empirical coefficients dominate the nominal ones. Substituting both bounds into equation 84 gives

$$\begin{aligned}
&\left| \underline{\widehat{P}}^V(\cdot \mid s,a)V - \underline{P}^V(\cdot \mid s,a)V \right| \\
&\le \frac{20\sqrt{2}}{3} \sqrt{\frac{1}{Np_\wedge} \log \frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}} \max\left\{ \sqrt{\operatorname{Var}_{\underline{P}^V(\cdot|s,a)}(V)}, \sqrt{\operatorname{Var}_{\underline{\widehat{P}}^V(\cdot|s,a)}(V)} \right\} \\
&\quad + 20\sqrt{\frac{2e^\delta}{3N} \log \frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}} \operatorname{span}(V) \\
&\quad + \frac{64}{p_\wedge N(1-\gamma)} \log \frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}.
\end{aligned}$$

We next compare the two value variances in this estimate. Suppose first that both optimizers are positive. The distribution $P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}$ uses the nominal base and the empirical optimizer. The relative probability bound in Lemma C.6 gives

$$\frac{3}{5} \le \frac{\underline{\widehat{P}}^V(s' \mid s,a)}{P^0(s' \mid s,a)h^{\widehat{\lambda}^*,V}_{s,a}(s')} \le \frac{5}{3}.$$

The ratio equals $\widehat{P}^0(s' \mid s,a)/P^0(s' \mid s,a)$ divided by its expectation under $P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}$. For a positive function, its variance divided by the square of its mean is at most the squared difference between its upper and lower bounds divided by four times their product. This follows by expanding the nonnegative product of its distances to these bounds and completing the square. Since the base ratio differs from one by at most the relative error in equation 83, we obtain

$$\begin{aligned}
&\mathrm{Var}_{P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}}\left(\frac{\underline{\widehat{P}}^V(\cdot \mid s,a)}{P^0(\cdot \mid s,a)h^{\widehat{\lambda}^*,V}_{s,a}(\cdot)}\right) \\
&\le \frac{\dfrac{12}{Np_\wedge}\log\dfrac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}{1-\dfrac{12}{Np_\wedge}\log\dfrac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}} \le \frac{16}{15}\frac{12}{Np_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}.
\end{aligned}$$

The last inequality uses equation 177. Centering under $P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}$ and using Cauchy–Schwarz therefore gives

$$\begin{aligned}
\mathrm{Var}_{\underline{\widehat{P}}^V_{s,a}}(V) &\le \frac{5}{3}\,\mathrm{Var}_{P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}}(V), \\
\left|\mathbb{E}_{\underline{\widehat{P}}^V_{s,a}}[V] - \mathbb{E}_{P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}}[V]\right| &\le \frac{4}{\sqrt{15}}\sqrt{\frac{12}{Np_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\sqrt{\mathrm{Var}_{P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}}(V)}.
\end{aligned}$$

The distributions $P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}$ and $\underline{P}^V_{s,a} = P^0_{s,a}h^{\lambda^*,V}_{s,a}$ are exponential tilts of the same base. Their density ratio is monotone in V, so their distributions of V are stochastically ordered. This follows directly from the covariance inequality for an increasing function and a decreasing function. The squared deviation from $\mathbb{E}_{\underline{P}^V_{s,a}}[V]$ is $2\,\mathrm{span}(V)$-Lipschitz on the range of V. Adding or subtracting this squared deviation from $2\,\mathrm{span}(V)V$ gives increasing functions. Applying the expectation order to these two functions yields

$$\begin{aligned}
&\mathrm{Var}_{P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}}(V) \\
&\le \mathrm{Var}_{\underline{P}^V_{s,a}}(V) + 2\,\mathrm{span}(V)\left|\mathbb{E}_{P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}}[V] - \mathbb{E}_{\underline{P}^V_{s,a}}[V]\right| \\
&\le \mathrm{Var}_{\underline{P}^V_{s,a}}(V) + 2\,\mathrm{span}(V)\left|(\underline{\widehat{P}}^V_{s,a} - \underline{P}^V_{s,a})V\right| \\
&\quad + \frac{8}{\sqrt{15}}\,\mathrm{span}(V)\sqrt{\frac{12}{Np_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\sqrt{\mathrm{Var}_{P^0_{s,a}h^{\widehat{\lambda}^*,V}_{s,a}}(V)}.
\end{aligned}$$

Solving this quadratic inequality for the last standard deviation, and then using the preceding factor $5/3$, gives

$$\begin{aligned}
\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^V_{s,a}}(V)} \le \sqrt{\frac{5}{3}}\sqrt{\mathrm{Var}_{\underline{P}^V_{s,a}}(V)} + \sqrt{\frac{10}{3}\,\mathrm{span}(V)\left|(\underline{\widehat{P}}^V_{s,a} - \underline{P}^V_{s,a})V\right|} \\
+ \frac{8}{3}\,\mathrm{span}(V)\sqrt{\frac{12}{Np_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}.
\end{aligned}$$

The same inequality with the two worst-case variances interchanged also holds. Use the empirical base with the nominal optimizer: the reciprocal base ratios, after normalization, obey the same pointwise and variance bounds, so the remaining argument is unchanged.

Substitute either comparison into the preceding transition estimate and apply Young's inequality:

$$\frac{40}{3}\sqrt{\frac{5\operatorname{span}(V)}{3Np_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}\left|(\widehat{\underline{P}}^V_{s,a}-\underline{P}^V_{s,a})V\right|}$$
$$\le\frac{1}{5}\left|(\widehat{\underline{P}}^V_{s,a}-\underline{P}^V_{s,a})V\right|+\frac{10000\operatorname{span}(V)}{27Np_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}.$$

Move one fifth of the absolute difference to the left. Since $\operatorname{span}(V)\le(1-\gamma)^{-1}$ and

$$\frac{25}{3}\sqrt{\frac{10}{3}}<16,\qquad 25\sqrt{\frac{2}{3}}<21,\qquad 80+\frac{400\sqrt{6}}{9}+\frac{12500}{27}<652,$$

we obtain

$$\left|(\widehat{\underline{P}}^V_{s,a}-\underline{P}^V_{s,a})V\right|$$
$$\le 16\sqrt{\frac{1}{Np_\wedge}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\min\left\{\sqrt{\operatorname{Var}_{\underline{P}^V_{s,a}}(V)},\sqrt{\operatorname{Var}_{\widehat{\underline{P}}^V_{s,a}}(V)}\right\}$$
$$+21\sqrt{\frac{e^\delta}{N}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\operatorname{span}(V)$$
$$+\frac{652}{Np_\wedge(1-\gamma)}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}.$$

This proves equation 178. When an optimizer equals zero, take the corresponding exponential tilt limit. Its value equals the minimum on the common support almost surely. Every worst-case distribution at that zero optimizer has this same distribution of V, so all the mean and variance inequalities above remain valid. □

Theorem C.16 (Policy guarantee for $\delta>1$)**.** *Suppose $0<\gamma<1$, $\delta>1$, $0<\eta<1$, rewards are known and in $[0,1]$, and the sampling assumptions and sample condition of Lemma C.15 hold. Let $\widehat{\pi}$ be obtained after T iterations of empirical robust value iteration from $Q_0=0$, with a fixed rule for choosing among maximizing actions. On the event of Lemma C.15, whose probability is at least $1-\eta$,*

$$\|V_{\pi^*}-V_{\widehat{\pi}}\|_\infty\le 34\sqrt{\frac{1}{Np_\wedge(1-\gamma)^3}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}$$
$$+63\sqrt{\frac{e^\delta}{N(1-\gamma)^4}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}$$
$$+\frac{2304\big(1+\log(4/(3p_\wedge))\big)}{Np_\wedge(1-\gamma)^3}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}+\frac{6\gamma^T}{(1-\gamma)^2}. \tag{181}$$

If $\delta\ge\log(4/(3p_\wedge))$, then on the same event

$$\|V_{\pi^*}-V_{\widehat{\pi}}\|_\infty\le\frac{2\gamma^T}{(1-\gamma)^2}. \tag{182}$$

Proof. Work on the event of Lemma C.15. All values belong to $[0,(1-\gamma)^{-1}]^{|\mathcal{S}|}$. The proof uses the value variances in equation 178 before bounding their discounted propagation.

For every fixed nominal row and $V,W\in[0,(1-\gamma)^{-1}]^{|\mathcal{S}|}$, we first prove

$$\left|\operatorname{Var}_{\underline{P}^V_{s,a}}(V)-\operatorname{Var}_{\underline{P}^W_{s,a}}(W)\right|$$
$$\le\frac{4\log(1/p_\wedge)}{1-\gamma}\|V-W\|_\infty. \tag{183}$$

Suppose first that the dual optimizer at V is positive, and let λ_t denote the dual optimizer at $V+t(W-V)$ for t near zero. The worst-case distribution is then unique and assigns positive

probability to every state in the nominal support. Moreover, $\mathrm{Var}_{\underline{P}^V_{s,a}}(V) > 0$: otherwise V would be constant on that support, and the dual objective would be strictly decreasing for $\lambda > 0$. Differentiating the active constraint

$$\mathrm{KL}\Big(\underline{P}^{V+t(W-V)}_{s,a} \,\Big\|\, P^0_{s,a}\Big) = \delta$$

and the exponential formula for the worst-case distribution gives

$$\left.\frac{d\lambda_t}{dt}\right|_{t=0} = \frac{\lambda^* \,\mathrm{Cov}_{\underline{P}^V_{s,a}}(V, W-V)}{\mathrm{Var}_{\underline{P}^V_{s,a}}(V)}.$$

The derivative exists by the implicit function theorem, since the derivative of the KL divergence with respect to λ is $-\mathrm{Var}_{\underline{P}^V_{s,a}}(V)/(\lambda^*)^3 < 0$. Differentiating the value variance therefore yields

$$\begin{aligned}
&\left.\frac{d}{dt}\,\mathrm{Var}_{\underline{P}^{V+t(W-V)}_{s,a}}\big(V + t(W-V)\big)\right|_{t=0} \\
&= 2\,\mathrm{Cov}_{\underline{P}^V_{s,a}}(V, W-V) \\
&\quad - \frac{1}{\lambda^*}\,\mathrm{Cov}_{\underline{P}^V_{s,a}}\Big(\big(V - \mathbb{E}_{\underline{P}^V_{s,a}}[V]\big)^2, W-V\Big) \\
&\quad + \frac{\mathbb{E}_{\underline{P}^V_{s,a}}[\big(V - \mathbb{E}_{\underline{P}^V_{s,a}}[V]\big)^3]}{\lambda^*\,\mathrm{Var}_{\underline{P}^V_{s,a}}(V)}\,\mathrm{Cov}_{\underline{P}^V_{s,a}}(V, W-V).
\end{aligned}$$

By equation 173 and $h^{\lambda^*,V}_{s,a} \le p_\wedge^{-1}$,

$$\begin{aligned}
&\mathbb{E}_{\underline{P}^V_{s,a}}\Big[\Big|V - \mathbb{E}_{\underline{P}^V_{s,a}}[V]\Big|\Big] \\
&= 2\mathbb{E}_{\underline{P}^V_{s,a}}\Big[\max\Big\{\mathbb{E}_{\underline{P}^V_{s,a}}[V] - V, 0\Big\}\Big] \le 2\lambda^*\big(\log(1/p_\wedge) - \delta\big).
\end{aligned}$$

The expected positive and negative deviations from the mean are equal. Bounding their squares by their respective endpoint distances times the absolute deviations therefore gives

$$\mathrm{Var}_{\underline{P}^V_{s,a}}(V) \le \frac{\mathrm{span}(V)}{2}\mathbb{E}_{\underline{P}^V_{s,a}}\Big[\Big|V - \mathbb{E}_{\underline{P}^V_{s,a}}[V]\Big|\Big].$$

Cauchy–Schwarz then bounds the displayed mean absolute deviation by $\mathrm{span}(V)/2$. Combining these facts with the preceding multiplier bound and $\mathrm{span}(V) \le (1-\gamma)^{-1}$ gives

$$\begin{aligned}
\mathrm{Var}_{\underline{P}^V_{s,a}}(V) &\le \frac{\lambda^*\big(\log(1/p_\wedge) - \delta\big)}{1-\gamma}, \\
\Big|\mathbb{E}_{\underline{P}^V_{s,a}}[\big(V - \mathbb{E}_{\underline{P}^V_{s,a}}[V]\big)^3]\Big| &\le \frac{\mathrm{Var}_{\underline{P}^V_{s,a}}(V)}{1-\gamma}.
\end{aligned}$$

These inequalities bound the absolute values of the three derivative terms, respectively, by

$$\frac{1}{1-\gamma}\|W-V\|_\infty, \qquad \frac{2\big(\log(1/p_\wedge) - \delta\big)}{1-\gamma}\|W-V\|_\infty, \qquad \frac{2\big(\log(1/p_\wedge) - \delta\big)}{1-\gamma}\|W-V\|_\infty.$$

For the middle bound, center $W - V$ in the covariance and use $\Big|W - V - \mathbb{E}_{\underline{P}^V_{s,a}}[W-V]\Big| \le 2\|W-V\|_\infty$. A positive optimizer implies $\delta \le \log(1/p_\wedge)$. Since $\delta > 1$, summing the three bounds gives a derivative of absolute value at most $4\log(1/p_\wedge)\|W-V\|_\infty/(1-\gamma)$.

This derivative bound extends to the entire segment $V + t(W-V)$, $0 \le t \le 1$, including zero optimizers. Indeed, the KL uncertainty set on the nominal support is compact. Whenever value vectors converge, every limit of their minimizing distributions minimizes the limiting objective. At a positive optimizer that distribution is unique; at a zero optimizer every minimizer is supported on states attaining the minimum value and has zero value variance, by Lemma C.14. Consequently, the value variance is continuous along the segment. It is differentiable with the preceding bound on every open interval where it is positive, and is zero elsewhere. Integrating on such intervals, and using continuity at their endpoints, proves equation 183.

The same argument applies to the empirical row on the event of Lemma C.15, with probability floor $3p_\wedge/4$. Hence

$$\begin{aligned}&\left|\mathrm{Var}_{\underline{\widehat{P}}^V_{s,a}}(V) - \mathrm{Var}_{\underline{\widehat{P}}^W_{s,a}}(W)\right| \\ &\le \frac{4\log(4/(3p_\wedge))}{1-\gamma}\|V-W\|_\infty. \end{aligned} \tag{184}$$

Taking square roots in these comparisons gives, in particular,

$$\sqrt{\mathrm{Var}_{\underline{P}^W_{s,a}}(W)} \le \sqrt{\mathrm{Var}_{\underline{P}^V_{s,a}}(V)} + \sqrt{\frac{4\log(1/p_\wedge)}{1-\gamma}\|V-W\|_\infty},$$

$$\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^W_{s,a}}(W)} \le \sqrt{\mathrm{Var}_{\underline{\widehat{P}}^V_{s,a}}(V)} + \sqrt{\frac{4\log(4/(3p_\wedge))}{1-\gamma}\|V-W\|_\infty}.$$

The nominal and empirical Bellman equations give the following three componentwise comparisons:

$$\widehat{V}_{\pi^*} - V_{\pi^*} \le \gamma(I-\gamma\underline{P}^{V_{\pi^*},\pi^*})^{-1}\big(\underline{\widehat{P}}^{\widehat{V}_{\pi^*},\pi^*} - \underline{P}^{\widehat{V}_{\pi^*},\pi^*}\big)\widehat{V}_{\pi^*}, \tag{185}$$

$$V_{\pi^*} - \widehat{V}_{\pi^*} \le \gamma(I-\gamma\underline{\widehat{P}}^{\widehat{V}_{\pi^*},\pi^*})^{-1}\big(\underline{P}^{V_{\pi^*},\pi^*} - \underline{\widehat{P}}^{V_{\pi^*},\pi^*}\big)V_{\pi^*}, \tag{186}$$

$$\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}} \le (I-\gamma\underline{P}^{V_{\widehat{\pi}},\widehat{\pi}})^{-1}\Big[\frac{2\gamma^T}{1-\gamma}\mathbf{1} + \gamma\big(\underline{\widehat{P}}^{\widehat{V}_{\widehat{\pi}^*},\widehat{\pi}} - \underline{P}^{\widehat{V}_{\widehat{\pi}^*},\widehat{\pi}}\big)\widehat{V}_{\widehat{\pi}^*}\Big]. \tag{187}$$

For the first inequality, split the difference of the Bellman equations at the nominal robust expectation of $\widehat{V}_{\pi^*}$. The nominal robust-expectation difference is at most $\underline{P}^{V_{\pi^*},\pi^*}(\widehat{V}_{\pi^*} - V_{\pi^*})$, because this transition minimizes the expectation of V_{π^*}. Rearranging and using the nonnegative inverse proves the inequality. Interchanging the two models gives the second inequality. For the third, contraction gives $\|Q_T - Q^*_{\widehat{\pi}^*}\|_\infty \le \gamma^T/(1-\gamma)$. Greediness of $\widehat{\pi}$ therefore implies

$$0 \le \widehat{V}_{\widehat{\pi}^*} - r(\cdot,\widehat{\pi}(\cdot)) - \gamma\underline{\widehat{P}}^{\widehat{V}_{\widehat{\pi}^*},\widehat{\pi}}\widehat{V}_{\widehat{\pi}^*} \le \frac{2\gamma^T}{1-\gamma}\mathbf{1}.$$

Subtracting the nominal Bellman equation for $V_{\widehat{\pi}}$ and using its minimizing transition gives equation 187.

Apply Lemma C.15 at $\widehat{V}_{\pi^*}$, V_{π^*}, and $\widehat{V}_{\widehat{\pi}^*}$, respectively. In the first and third comparisons, use the nominal variance in equation 178; in the second, use the empirical variance. The value-variance comparison proved above then bounds each required standard deviation by the standard deviation of the value function of the matrix in its inverse, plus

$$\sqrt{\frac{4\log(4/(3p_\wedge))}{1-\gamma}\max\Big\{\|V_{\pi^*} - \widehat{V}_{\pi^*}\|_\infty, \|\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}}\|_\infty\Big\}}.$$

Here the empirical probability floor is at least $3p_\wedge/4$, which explains the common logarithm. The three pairs of value functions are, respectively, $(\widehat{V}_{\pi^*}, V_{\pi^*})$ under the nominal base, $(V_{\pi^*}, \widehat{V}_{\pi^*})$ under the empirical base, and $(\widehat{V}_{\widehat{\pi}^*}, V_{\widehat{\pi}})$ under the nominal base. Thus the variance comparison is deterministic and requires no concentration at $V_{\widehat{\pi}}$.

For each resulting variance, the corresponding value function satisfies its own Bellman equation with rewards in $[0,1]$. Lemma C.12 consequently bounds its discounted propagation, including the factor γ, by $1/[\sqrt{2}(1-\gamma)^{3/2}]$. Each inverse maps $\mathbf{1}$ to $\mathbf{1}/(1-\gamma)$, and every span is at most $(1-\gamma)^{-1}$. Moreover, $V_{\widehat{\pi}} \le V_{\pi^*}$ and $\widehat{V}_{\pi^*} \le \widehat{V}_{\widehat{\pi}^*}$, so the negative part of $\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}}$ is bounded by the positive

part of $V_{\pi^*} - \widehat{V}_{\pi^*}$. The three comparisons therefore control both absolute errors and give

$$\begin{aligned}
&\max\left\{\|V_{\pi^*} - \widehat{V}_{\pi^*}\|_\infty, \|\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}}\|_\infty\right\} \\
&\le 8\sqrt{2}\sqrt{\frac{1}{Np_\wedge(1-\gamma)^3}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}} \\
&\quad + 16\sqrt{\frac{4\log(4/(3p_\wedge))}{Np_\wedge(1-\gamma)^3}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}\sqrt{\max\left\{\|V_{\pi^*} - \widehat{V}_{\pi^*}\|_\infty, \|\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}}\|_\infty\right\}} \\
&\quad + 21\sqrt{\frac{e^\delta}{N(1-\gamma)^4}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}} + \frac{652}{Np_\wedge(1-\gamma)^2}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge} + \frac{2\gamma^T}{(1-\gamma)^2}.
\end{aligned}$$

Applying Young's inequality to the term containing the square root of the repeated maximum absorbs one third of that maximum into the left-hand side. The additional term before division by $2/3$ is

$$\frac{768\log(4/(3p_\wedge))}{Np_\wedge(1-\gamma)^3}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}.$$

Finally,

$$0 \le V_{\pi^*} - V_{\widehat{\pi}} \le (V_{\pi^*} - \widehat{V}_{\pi^*}) + (\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}}),$$

so the policy loss is at most twice the maximum. Using $24\sqrt{2} < 34$ and

$$2304\log(4/(3p_\wedge)) + 1956 \le 2304\big(1 + \log(4/(3p_\wedge))\big)$$

proves equation 181.

Finally, if $\delta \ge \log(4/(3p_\wedge))$, a point mass at any state in the common support has KL divergence at most δ from either transition row. Hence, for every V and (s,a),

$$\underline{P}^V_{s,a}V = \underline{\widehat{P}}^V_{s,a}V = \min_{s':P^0(s'|s,a)>0} V(s').$$

The nominal and empirical robust Bellman operators therefore agree, so $V_{\pi^*} = \widehat{V}_{\widehat{\pi}^*}$. The empirical Bellman residual of $\widehat{\pi}$ proved above is also its nominal residual. Contraction for this policy consequently gives

$$\|V_{\pi^*} - V_{\widehat{\pi}}\|_\infty = \|\widehat{V}_{\widehat{\pi}^*} - V_{\widehat{\pi}}\|_\infty \le \frac{1}{1-\gamma}\frac{2\gamma^T}{1-\gamma},$$

which proves equation 182. □

Corollary C.17 (Sample requirements for $\delta > 1$)**.** *Under the assumptions of Theorem C.16, let $0 < \epsilon \le (1-\gamma)^{-1}$ and choose*

$$T \ge \left\lceil \frac{\log(24/((1-\gamma)^2\epsilon))}{-\log\gamma} \right\rceil. \tag{188}$$

For every $\delta > 1$, the sample condition

$$\begin{aligned}
N \ge \max\Bigg\{ &\frac{192}{p_\wedge}, \frac{9216}{\epsilon^2}\left(\frac{1}{p_\wedge(1-\gamma)^3} + \frac{e^\delta}{(1-\gamma)^4}\right) \\
&+ \frac{6144\big(1+\log(4/(3p_\wedge))\big)}{p_\wedge(1-\gamma)^3\epsilon}\Bigg\}\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}
\end{aligned} \tag{189}$$

ensures $\|V_{\pi^*} - V_{\widehat{\pi}}\|_\infty \le \epsilon$ with probability at least $1-\eta$. The sufficient total sample count is

$$\widetilde{\mathcal{O}}\left(|\mathcal{S}||\mathcal{A}|\left[\frac{1}{p_\wedge} + \frac{1}{p_\wedge(1-\gamma)^3\epsilon^2} + \frac{e^\delta}{(1-\gamma)^4\epsilon^2} + \frac{1}{p_\wedge(1-\gamma)^3\epsilon}\right]\right). \tag{190}$$

For $0 < \epsilon \le 1$, this simplifies to

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^2}\left[\frac{1}{p_\wedge(1-\gamma)^3} + \frac{e^\delta}{(1-\gamma)^4}\right]\right). \tag{191}$$

For $\delta \ge \log(4/(3p_\wedge))$, only equation 177 is required, giving total sample count $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|/p_\wedge)$.

Proof. For equation 181, Cauchy–Schwarz and $34^2 + 63^2 = 5125 < 72^2$ bound the sum of the two square-root terms by

$$72\sqrt{\frac{1}{N}\left(\frac{1}{p_\wedge(1-\gamma)^3} + \frac{e^\delta}{(1-\gamma)^4}\right)\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}}.$$

Young's inequality bounds this display by

$$\frac{3\epsilon}{8} + \frac{3456}{\epsilon N}\left(\frac{1}{p_\wedge(1-\gamma)^3} + \frac{e^\delta}{(1-\gamma)^4}\right)\log\frac{36|\mathcal{S}|^2|\mathcal{A}|N}{\eta p_\wedge}.$$

Adding the N^{-1} term in equation 181 and applying equation 189 bounds the sum by $3\epsilon/4$: the identities $(3/8)\cdot 9216 = 3456$ and $(3/8)\cdot 6144 = 2304$ account for its two sample coefficients. The choice of T bounds the planning term $6\gamma^T/(1-\gamma)^2$ by $\epsilon/4$.

Multiplication by $|\mathcal{S}||\mathcal{A}|$ gives the total sample requirements; the displayed logarithms contribute only logarithmic factors. If $\epsilon \le 1$, the $p_\wedge^{-1}$ and $[p_\wedge(1-\gamma)^3\epsilon]^{-1}$ terms are dominated by $[p_\wedge(1-\gamma)^3\epsilon^2]^{-1}$, which proves equation 191. The last assertion follows directly from equation 182. □

D Lower Bound Proof

In this section, we present our proof of the lower bounds.

D.1 Hard MDP design and lower bound proof for Theorem 3.4

Theorem D.1 (Restatement of Theorem 3.4)**.** *Suppose $|\mathcal{S}| \ge 2$, $|\mathcal{A}| \ge 2$, $\gamma \in (3/4, 1)$, $\delta \ge 1-\gamma$, $0 < \epsilon < 1$, and*

$$0 < \epsilon \le \frac{1}{754(1-\gamma)\left(\delta + \log\frac{2}{1-\gamma}\right)}.$$

There exist two DR-MDPs $\mathcal{M}_0, \mathcal{M}_1$ with the same known rewards in $[0,1]$ and KL-defined (s,a)-rectangular uncertainty sets of radius δ such that, given N independent samples per state-action pair, independently across pairs,

$$\inf_{\widehat{\pi}} \max_{\phi\in\{0,1\}} \mathbb{P}_\phi\Big(\|V_\phi^* - V_\phi^{\widehat{\pi}}\|_\infty > \epsilon\Big) > \frac{1}{8}, \tag{192}$$

provided that

$$N|\mathcal{S}||\mathcal{A}| \le \frac{1}{8457}\,\frac{\delta e^\delta |\mathcal{S}||\mathcal{A}|}{(1+\delta)(1-\gamma)^4\epsilon^2\left(\delta + \log\frac{2}{1-\gamma}\right)^2}.$$

The infimum is over all learning algorithms and their returned policies, including randomization. The failure probability is at least $3/8$. The minimal non-zero nominal transition probability is determined by the construction and satisfies $\delta < \log(1/p_\wedge)$. Consequently, the total sample complexity is lower bounded by

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|\delta e^\delta}{(1-\gamma)^4\epsilon^2}\right),$$

where the suppressed factors are logarithmic in $1/p_\wedge$ and $1/(1-\gamma)$ and are displayed explicitly in the sample condition above.

We construct a family of hard instances to establish the minimax lower bound for KL-defined robust RL. Consider a class of two-state MDPs $\{\mathcal{M}_\phi\}$ parameterized by p with an uncertainty level δ.

The state space includes a state s_0 and an absorbing state s_1, with two available actions $a \in \{0,1\}$:

$$s_0 : r(s_0) = 0, \qquad s_1 : r(s_1) = 1.$$

The nominal transition probabilities from s_0 are defined as follows:

$$P(s_0 \mid s_0, a=\phi) = p, \qquad P(s_1 \mid s_0, a=\phi) = 1-p,$$
$$P(s_0 \mid s_0, a=1-\phi) = q, \qquad P(s_1 \mid s_0, a=1-\phi) = 1-q.$$

State s_1 is absorbing for both actions:

$$P(s_1 \mid s_1, a) = 1, \qquad \forall a \in \{0, 1\}.$$

In the robust formulation, the adversary selects a worst-case transition kernel within the KL uncertainty set. The resulting worst-case transitions preserve the same structure:

$$\begin{aligned} P^{\mathrm{wc}}(s_1 \mid s_0, a = \phi) &= 1 - \tilde{p}, & P^{\mathrm{wc}}(s_0 \mid s_0, a = \phi) &= \tilde{p}, \\ P^{\mathrm{wc}}(s_1 \mid s_0, a = 1 - \phi) &= 1 - \tilde{q}, & P^{\mathrm{wc}}(s_0 \mid s_0, a = 1 - \phi) &= \tilde{q}. \end{aligned}$$

Since s_1 is absorbing and yields unit reward at every step, its value function is

$$V(s_1) = \frac{1}{1 - \gamma}.$$

For a policy that always selects action ϕ, the robust Bellman equation at s_0 is

$$\begin{aligned} V_\phi(s_0) &= \gamma \tilde{p} V_\phi(s_0) + \gamma(1 - \tilde{p}) V(s_1) \\ &= \gamma \tilde{p} V_\phi(s_0) + \frac{\gamma(1 - \tilde{p})}{1 - \gamma}, \end{aligned}$$

which yields the closed-form expression

$$V_\phi(s_0) = \frac{\gamma(1 - \tilde{p})}{(1 - \gamma)(1 - \gamma\tilde{p})}.$$

Similarly, for the policy that always selects action $1 - \phi$, we obtain

$$\begin{aligned} V_{1-\phi}(s_0) &= \gamma \tilde{q} V_{1-\phi}(s_0) + \gamma(1 - \tilde{q}) V(s_1) \\ &= \gamma \tilde{q} V_{1-\phi}(s_0) + \frac{\gamma(1 - \tilde{q})}{1 - \gamma}, \end{aligned}$$

and hence

$$V_{1-\phi}(s_0) = \frac{\gamma(1 - \tilde{q})}{(1 - \gamma)(1 - \gamma\tilde{q})}.$$

We now compare the values induced by the two deterministic policies $\pi(s_0) = \phi$ and $\pi(s_0) = 1 - \phi$. Define the value gap as

$$\Delta V = \left| V_\phi(s_0) - V_{1-\phi}(s_0) \right|.$$

Substituting the closed-form expressions for the value functions, we obtain

$$\begin{aligned} \Delta V &= \left| \frac{\gamma(1 - \tilde{q})}{(1 - \gamma)(1 - \gamma\tilde{q})} - \frac{\gamma(1 - \tilde{p})}{(1 - \gamma)(1 - \gamma\tilde{p})} \right| \\ &= \left| \frac{\gamma\big((1 - \tilde{q})(1 - \gamma\tilde{p}) - (1 - \tilde{p})(1 - \gamma\tilde{q})\big)}{(1 - \gamma)(1 - \gamma\tilde{q})(1 - \gamma\tilde{p})} \right| \\ &= \left| \frac{\gamma(1 - \gamma)(\tilde{p} - \tilde{q})}{(1 - \gamma)(1 - \gamma\tilde{q})(1 - \gamma\tilde{p})} \right| \\ &= \frac{\gamma|\tilde{p} - \tilde{q}|}{(1 - \gamma\tilde{q})(1 - \gamma\tilde{p})}. \end{aligned} \tag{193}$$

We set

$$\begin{aligned} &p = \gamma e^{-\delta/\gamma}, \qquad q = p e^{\varepsilon}, \\ &\varepsilon = \frac{8(1 + \gamma)^2(1 - \gamma)^2\epsilon}{\gamma^2} \left(\delta + \log \frac{2}{1 - \gamma} \right). \end{aligned} \tag{194}$$

Since $\delta \geq 1-\gamma$ and $e^{(1-\gamma)/\gamma} \geq 1 + (1-\gamma)/\gamma = 1/\gamma$, we have $p \leq \gamma^2$. To verify the perturbation range, first note that $\gamma > 3/4$ gives

$$128\frac{(1+\gamma)^2}{\gamma^2} \leq \frac{6272}{9} < 754.$$

Convexity of the exponential on $[-1/3, 0]$ also gives

$$\begin{aligned} p &\leq \gamma e^{-(1-\gamma)/\gamma} \\ &\leq \gamma - 3(1-e^{-1/3})(1-\gamma), \end{aligned}$$

since $0 < (1-\gamma)/\gamma < 1/3$. Consequently,

$$\frac{1-\gamma}{1-p} \leq \frac{1}{4-3e^{-1/3}} \leq \frac{113}{209},$$

where the last inequality follows from $e^{1/3} \geq 1 + 1/3 + 1/18 + 1/162 = 113/81$. Hence

$$256\frac{(1+\gamma)^2(1-\gamma)}{\gamma^2(1-p)} \leq \frac{1417472}{1881} < 754.$$

Substituting the theorem's accuracy condition into equation 194 now yields

$$0 < \varepsilon \leq \frac{1-\gamma}{16}, \qquad \varepsilon \leq \frac{1-p}{32}. \tag{195}$$

In particular,

$$\begin{aligned} 0 < q - p = p(e^{\varepsilon} - 1) \leq 2p\varepsilon \leq \frac{1-p}{2}, \qquad 1 - q \geq \frac{1-p}{2}, \\ q = \gamma e^{-\delta/\gamma + \varepsilon} < \gamma, \end{aligned} \tag{196}$$

since $\delta \geq 1-\gamma$. Moreover,

$$\log\frac{1}{p} - \delta = -\log\gamma + \frac{\delta(1-\gamma)}{\gamma} \geq (1-\gamma)\left(1 + \frac{\delta}{\gamma}\right), \tag{197}$$

$$\log\frac{1}{x} - \delta \geq \log\frac{1}{p} - \delta - \varepsilon \geq \frac{15}{16}\left(\log\frac{1}{p} - \delta\right) > 0, \qquad x \in [p, q]. \tag{198}$$

For each $x \in [p, q]$, the worst-case self-loop probability y is therefore the unique solution in $(x, 1)$ of

$$\mathrm{KL}(y\|x) = y\log\frac{y}{x} + (1-y)\log\frac{1-y}{1-x} = \delta.$$

Indeed, this divergence is zero at $y = x$, equals $\log(1/x) > \delta$ at $y = 1$, and is strictly increasing for $y > x$. At $x = p$ and $x = q$, the solutions are $\widetilde{p}$ and $\widetilde{q}$, respectively, so

$$\mathrm{KL}(\widetilde{p}\|p) = \mathrm{KL}(\widetilde{q}\|q) = \delta. \tag{199}$$

We next bound the difference between the worst-case probabilities. The partial derivatives of the binary KL divergence are

$$\frac{\partial}{\partial x}\mathrm{KL}(y\|x) = \frac{x-y}{x(1-x)}, \qquad \frac{\partial}{\partial y}\mathrm{KL}(y\|x) = \log\frac{y(1-x)}{x(1-y)} > 0.$$

The implicit function theorem gives

$$x\frac{dy}{dx} = \frac{y-x}{(1-x)\log\left(\dfrac{y(1-x)}{x(1-y)}\right)} > 0. \tag{200}$$

By convexity of $\mathrm{KL}(\cdot\|x)$ between x and 1,

$$\delta \leq \frac{y-x}{1-x}\log\frac{1}{x}.$$

Since $-\log\gamma \le (1-\gamma)/\gamma$ and $\delta \ge 1-\gamma$, it follows that

$$\frac{y-x}{1-x} \ge \frac{\delta}{\log(1/p)} \ge \frac{\gamma}{2}, \qquad \log\frac{1}{p} \le \frac{\delta+1-\gamma}{\gamma} \le \frac{2\delta}{\gamma}. \tag{201}$$

To bound the logarithm in equation 200, rearrange $\mathrm{KL}(y\|x) = \delta$ to obtain

$$\begin{aligned}\log\frac{1}{x} - \delta &= -y\log y - (1-y)\log(1-y) + (1-y)\log\frac{1-x}{x} \\ &\le \left(\log\frac{1}{p} + 2\right)\sqrt{1-y}.\end{aligned}$$

Here $-y\log y \le 1-y$ and $-(1-y)\log(1-y) \le 2(\sqrt{1-y} - (1-y))$, both following from $\log t \le t-1$, imply the stated bound on the entropy. Together with equation 198, this yields

$$\sqrt{1-y} \ge \frac{15[\log(1/p)-\delta]}{16(\log(1/p)+2)} \ge \frac{15(1-\gamma)}{32}.$$

For the last inequality, equation 197 and $\delta+\gamma \ge 1$ give

$$\frac{\log(1/p)+2}{\log(1/p)-\delta} \le \frac{\delta+1+\gamma}{(1-\gamma)(\delta+\gamma)} \le \frac{2}{1-\gamma}.$$

Consequently, throughout $x \in [p,q]$,

$$\begin{aligned}\log\frac{y(1-x)}{x(1-y)} &\le \log\frac{1}{p} + \log(1-p) + 2\log\frac{32}{15(1-\gamma)} \\ &\le 2\left(\delta + \log\frac{2}{1-\gamma}\right).\end{aligned} \tag{202}$$

For the last inequality, use $\log(1/p) \le (\delta+1-\gamma)/\gamma$ and $1-p \le \min\{1, 1-\gamma+\delta\}$. Moreover,

$$\left(2-\frac{1}{\gamma}\right)\delta - \frac{1-\gamma}{\gamma} - \log\min\{1,1-\gamma+\delta\} \ge 2\gamma - \frac{1}{\gamma} > \frac{1}{6} > \log\frac{256}{225}.$$

Indeed, the expression on the left decreases up to $\delta=\gamma$ and increases thereafter; its value there is $2\gamma - 1/\gamma$. Combining equation 200, equation 201, and equation 202, and integrating over $[p,q]$, gives

$$\begin{aligned}\widetilde{q} - \widetilde{p} &= \int_p^q \frac{dy}{dx}\,dx \\ &\ge \frac{\gamma\log(q/p)}{4\left(\delta + \log\frac{2}{1-\gamma}\right)} = \frac{2(1-\gamma^2)^2\epsilon}{\gamma}.\end{aligned} \tag{203}$$

The reference point γ satisfies

$$\mathrm{KL}(\gamma\|p) = \delta + (1-\gamma)\log\frac{1-\gamma}{1-p} < \delta. \tag{204}$$

Hence $\widetilde{q} > \widetilde{p} > \gamma$. In particular,

$$\Delta V > \frac{\gamma(\widetilde{q}-\widetilde{p})}{(1-\gamma^2)^2} \ge 2\epsilon. \tag{205}$$

For a stationary randomized policy $\widehat{\pi}$, its worst-case self-loop probability at s_0 is $\widetilde{p} + (1-\widehat{\pi}(\phi \mid s_0))(\widetilde{q}-\widetilde{p})$. Substituting this probability into its robust Bellman equation gives

$$\begin{aligned}\|V_\phi^* - V_\phi^{\widehat{\pi}}\|_\infty &= \frac{\gamma(\widetilde{q}-\widetilde{p})(1-\widehat{\pi}(\phi \mid s_0))}{(1-\gamma\widetilde{p})\left[1-\gamma\big(\widetilde{p} + (1-\widehat{\pi}(\phi \mid s_0))(\widetilde{q}-\widetilde{p})\big)\right]} \\ &\ge \frac{\gamma(\widetilde{q}-\widetilde{p})}{(1-\gamma\widetilde{p})^2}(1-\widehat{\pi}(\phi \mid s_0)).\end{aligned} \tag{206}$$

The equality holds because the value at s_1 is independent of the policy. Moreover, $\widetilde{p} > \gamma$ and equation 203 give

$$\frac{\gamma(\widetilde{q}-\widetilde{p})}{(1-\gamma\widetilde{p})^2} > \frac{\gamma(\widetilde{q}-\widetilde{p})}{(1-\gamma^2)^2} \geq 2\epsilon.$$

To distinguish the two models, we bound the KL divergence of their sample distributions. The two models interchange p and q between actions 0 and 1. Independence of the observations and the exact sum of the two Bernoulli KL divergences give

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) &= \mathbb{E}_{\mathbb{P}_0}\left[\log\frac{\mathbb{P}_0}{\mathbb{P}_1}\right] = N\big(\mathrm{KL}(p\|q) + \mathrm{KL}(q\|p)\big) \\
&= N(q-p)\left(\varepsilon + \log\frac{1-p}{1-q}\right) \\
&\leq \frac{N(q-p)\varepsilon}{1-q} \leq \frac{Np\varepsilon^2}{1-p-\varepsilon} \leq \frac{32Np\varepsilon^2}{31(1-p)}.
\end{aligned} \tag{207}$$

The ratio in the expectation is evaluated at the observed dataset; independent algorithmic randomization leaves this divergence unchanged. The first inequality uses $\log((1-p)/(1-q)) \leq (q-p)/(1-q)$ and $q - p = q(1-e^{-\varepsilon}) \leq q\varepsilon$. For the second, $e^{\varepsilon} \leq (1-\varepsilon)^{-1}$ gives

$$\frac{q-p}{1-q} = \frac{p(e^{\varepsilon}-1)}{1-pe^{\varepsilon}} \leq \frac{p\varepsilon}{1-p-\varepsilon}.$$

The last inequality follows from $\varepsilon \leq (1-p)/32$ in equation 195. If

$$N \leq \frac{31}{2^{14}} \frac{\gamma^4(1-p)}{p(1+\gamma)^4(1-\gamma)^4\epsilon^2\left(\delta + \log\frac{2}{1-\gamma}\right)^2}, \tag{208}$$

then substituting equation 194 into equation 207 gives

$$\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) \leq \frac{2048}{31}\frac{31}{2^{14}} = \frac{1}{8}.$$

For fixed $\delta > 0$, the following derivative is nonpositive on $\gamma \in [3/4, 1]$:

$$\begin{aligned}
&\frac{\partial}{\partial\gamma}\left[\frac{\gamma^4}{(1+\gamma)^4}\left(\frac{e^{\delta/\gamma}}{\gamma} - 1\right)\right] \\
&\quad = \frac{\gamma^2}{(1+\gamma)^5}\left[e^{\delta/\gamma}\left(3 - \gamma - \frac{(1+\gamma)\delta}{\gamma}\right) - 4\gamma\right] \leq 0.
\end{aligned}$$

Indeed, maximizing the exponential term in brackets over $\delta \geq 0$ gives $(1+\gamma)e^{2(1-\gamma)/(1+\gamma)} \leq 2e^{2/7} < 3 \leq 4\gamma$. Evaluating at $\gamma = 1$ and using $e^{\delta} \geq 1 + \delta$ therefore yields

$$\begin{aligned}
\frac{\gamma^4}{(1+\gamma)^4}\frac{1-p}{p} &= \frac{\gamma^4}{(1+\gamma)^4}\left(\frac{e^{\delta/\gamma}}{\gamma} - 1\right) \\
&\geq \frac{e^{\delta}-1}{16} \geq \frac{\delta e^{\delta}}{16(1+\delta)}.
\end{aligned} \tag{209}$$

Since $1/8457 < 31/2^{18}$, this inequality shows that the sample condition in Theorem 3.4 implies equation 208. By Pinsker's inequality, the total variation distance between $\mathbb{P}_0$ and $\mathbb{P}_1$ is at most $1/4$. Choose model 0 when $\widehat{\pi}(0 \mid s_0) \geq 1/2$ and model 1 otherwise. An incorrect decision implies a robust value loss strictly greater than ϵ by equation 206. Moreover,

$$\begin{aligned}
&\mathbb{P}_0\big(\widehat{\pi}(0 \mid s_0) < 1/2\big) + \mathbb{P}_1\big(\widehat{\pi}(0 \mid s_0) \geq 1/2\big) \\
&\quad = 1 - \mathbb{P}_0\big(\widehat{\pi}(0 \mid s_0) \geq 1/2\big) + \mathbb{P}_1\big(\widehat{\pi}(0 \mid s_0) \geq 1/2\big) \geq \frac{3}{4}.
\end{aligned}$$

Consequently, the maximum failure probability is at least $3/8$.

The same conclusion holds for policies depending on the past. Use the fixed worst-case rows with self-loop probabilities $\widetilde{p}$ and $\widetilde{q}$, and evaluate the policy on histories in which every transition returns

to s_0. Each actual action history that remains at s_0 for t steps has probability at least $\widetilde{p}^t$ times its probability under these fixed-state histories. At s_0, each action other than ϕ has Bellman value loss $\gamma(\widetilde{q}-\widetilde{p})/(1-\gamma\widetilde{p})$ relative to V_ϕ^*; action ϕ has zero loss. Summing the discounted losses, and using that the fixed rows are feasible in the uncertainty sets, yields

$$V_\phi^*(s_0) - V_\phi^{\widehat{\pi}}(s_0) \geq \frac{\gamma(\widetilde{q}-\widetilde{p})}{(1-\gamma\widetilde{p})^2}\left(1-(1-\gamma\widetilde{p})\mathbb{E}_{\widehat{\pi}}\left[\sum_{t=0}^{\infty}(\gamma\widetilde{p})^t\mathbf{1}\{a_t=\phi\}\right]\right).$$

Here the expectation is over the policy's randomization along these fixed-state histories. Choose model 0 if

$$(1-\gamma\widetilde{p})\mathbb{E}_{\widehat{\pi}}\left[\sum_{t=0}^{\infty}(\gamma\widetilde{p})^t\mathbf{1}\{a_t=0\}\right] \geq \frac{1}{2},$$

and model 1 otherwise. The normalized frequencies of actions 0 and 1 sum to one, so an incorrect decision again implies loss strictly greater than ϵ. The same total variation bound gives maximum failure probability at least $3/8$.

For $|\mathcal{S}| > 2$, add absorbing states with reward zero, unreachable from s_0. For $|\mathcal{A}| > 2$, give every additional action at s_0 the inferior nominal row q in both models. All added rows have identical data distributions in the two models. The value-loss bounds continue to hold, and the frequencies of actions 0 and 1 now sum to at most one, which suffices for the same decision rule. The minimal non-zero nominal probability is $p_\wedge = \min\{p, 1-q\}$, so $\delta < \log(1/p) \leq \log(1/p_\wedge)$. For $\delta \geq 1$, $q \leq e^{-1+1/64} < 1/2$ and therefore $p_\wedge = p = \gamma e^{-\delta/\gamma}$. There are $N|\mathcal{S}||\mathcal{A}|$ samples in total. Combining equation 208 and equation 209 with the theorem's sample condition and taking the infimum over learning algorithms proves equation 14. Finally, $\delta < \log(1/p_\wedge)$ gives

$$\begin{aligned}&(1+\delta)\left(\delta + \log\frac{2}{1-\gamma}\right)^2 \\ &\quad\leq \left(1+\log\frac{1}{p_\wedge}\right)\left(\log\frac{1}{p_\wedge}+\log\frac{2}{1-\gamma}\right)^2.\end{aligned}$$

Thus the factors suppressed in the stated $\widetilde{\Omega}$ bound are logarithmic in $1/p_\wedge$ and $1/(1-\gamma)$. This completes the proof of Theorem 3.4.

D.2 Proof of Theorem 3.5

We construct a family of hard instances to establish the minimax lower bound for KL-defined robust RL. Consider a class of three-state MDPs $\{\mathcal{M}_\phi : \phi \in \{0,1\}\}$ parameterized by p, q with an uncertainty level δ.

The state space includes a state s_0 and two absorbing states s_1, s_2, with two available actions $a \in \{0,1\}$:

$$s_0 : r(s_0) = 0, \qquad s_1 : r(s_1) = 1, \qquad s_2 : r(s_2) = 0.$$

The nominal transition probabilities from s_0 are defined as follows:

$$\begin{aligned} P(s_1 \mid s_0, a=\phi) &= p, & P(s_2 \mid s_0, a=\phi) &= 1-p, \\ P(s_1 \mid s_0, a=1-\phi) &= q, & P(s_2 \mid s_0, a=1-\phi) &= 1-q. \end{aligned}$$

States s_1 and s_2 are absorbing for both actions:

$$\begin{aligned} P(s_1 \mid s_1, a) &= 1, \qquad \forall a \in \{0,1\}; \\ P(s_2 \mid s_2, a) &= 1, \qquad \forall a \in \{0,1\}. \end{aligned}$$

Case 1: $0 < \delta < \log(1/p_\wedge)$

We set

$$0 < \epsilon \leq \frac{\gamma\delta}{8(1-\gamma)\log(1/p_\wedge)\log\left(\dfrac{4}{p_\wedge^3[\log(1/p_\wedge)-\delta]^2}\right)}\min\left\{1, \frac{\log(1/p_\wedge)-\delta}{2}\right\}.$$

Then, we have

$$0 < \varepsilon \le p_\wedge \min\left\{1, \frac{\log(1/p_\wedge) - \delta}{2}\right\}. \tag{210}$$

Consequently,

$$p_\wedge = 1 - p < 1 - q \le 2p_\wedge < \frac{1}{4}, \qquad \frac{3}{4} < 1 - 2p_\wedge \le q < p < 1.$$

The minimal non-zero nominal transition probability is exactly $p_\wedge$ in each model.

The worst-case transition from s_0 has support contained in $\{s_1, s_2\}$. For $x \in [p_\wedge, p_\wedge + \varepsilon]$, let y be the worst-case probability of reaching s_2 when its nominal probability is x. The corresponding KL divergence is

$$\mathrm{KL}(y\|x) = y \log \frac{y}{x} + (1-y) \log \frac{1-y}{1-x}.$$

For every x in this interval, equation 210 and $\log(1+t) \le t$ for $t \ge 0$ imply

$$\begin{aligned} \log \frac{1}{x} - \delta &= \log \frac{1}{p_\wedge} - \delta - \log \frac{x}{p_\wedge} \\ &\ge \log \frac{1}{p_\wedge} - \delta - \frac{\varepsilon}{p_\wedge} \ge \frac{\log(1/p_\wedge) - \delta}{2} > 0. \end{aligned} \tag{211}$$

The divergence is zero at $y = x$, equals $\log(1/x) > \delta$ at $y = 1$, and is strictly increasing for $y > x$, since

$$\frac{\partial}{\partial y} \mathrm{KL}(y\|x) = \log \frac{y(1-x)}{x(1-y)} > 0.$$

Hence there is a unique $y \in (x, 1)$ satisfying

$$\mathrm{KL}(y\|x) = \delta. \tag{212}$$

Since the divergence is strictly increasing for $y > x$, the worst-case probabilities of reaching s_2 and s_1 are y and $1 - y$, respectively.

By convexity of $y \mapsto \mathrm{KL}(y\|x)$ between $y = x$ and $y = 1$, we obtain

$$\delta = \mathrm{KL}(y\|x) \le \frac{y-x}{1-x} \log \frac{1}{x}, \qquad \frac{y-x}{1-x} \ge \frac{\delta}{\log(1/x)}. \tag{213}$$

Rearranging equation 212, we obtain

$$\begin{aligned} \log \frac{1}{x} - \delta &= -y \log y - (1-y) \log(1-y) + (1-y) \log \frac{1-x}{x} \\ &\le 2\sqrt{1-y} + (1-y) \log \frac{1}{p_\wedge} \\ &\le \big(\log(1/p_\wedge) + 2\big)\sqrt{1-y}. \end{aligned} \tag{214}$$

For the first inequality, $\log t \le t - 1$ gives

$$-y \log y \le 1 - y, \qquad -(1-y)\log(1-y) \le 2\big(\sqrt{1-y} - (1-y)\big),$$

and $x \ge p_\wedge$ gives $\log((1-x)/x) \le \log(1/p_\wedge)$. The last inequality in equation 214 uses $1 - y \le \sqrt{1-y}$. Combining equation 211 and equation 214, and using $\log(1/p_\wedge) + 2 \le 1/p_\wedge$ for $0 < p_\wedge < 1/8$, gives

$$1 - y \ge \frac{[\log(1/p_\wedge) - \delta]^2}{4[\log(1/p_\wedge) + 2]^2} \ge \frac{p_\wedge^2 [\log(1/p_\wedge) - \delta]^2}{4}. \tag{215}$$

It follows that

$$0 < \log \frac{y(1-x)}{x(1-y)} \le \log \frac{1}{p_\wedge(1-y)} \le \log\left(\frac{4}{p_\wedge^3 [\log(1/p_\wedge) - \delta]^2}\right). \tag{216}$$

The partial derivative of the KL divergence with respect to x is

$$\frac{\partial}{\partial x} \mathrm{KL}(y\|x) = \frac{x-y}{x(1-x)}.$$

Since $0 < x < y < 1$ and the partial derivative with respect to y is positive, the implicit function theorem applies throughout $[p_\wedge, p_\wedge + \varepsilon]$. Differentiating equation 212 and using equation 213, we obtain

$$\begin{aligned}\frac{dy}{dx} &= \frac{y - x}{x(1-x)\log\left(\dfrac{y(1-x)}{x(1-y)}\right)} \\ &\geq \frac{\delta}{x\log(1/x)\log\left(\dfrac{y(1-x)}{x(1-y)}\right)}.\end{aligned}$$

Now $x \leq 2p_\wedge$, $\log(1/x) \leq \log(1/p_\wedge)$, and equation 216 show that, throughout the interval,

$$\frac{dy}{dx} \geq \frac{\delta}{2p_\wedge \log(1/p_\wedge)\log\left(\dfrac{4}{p_\wedge^3[\log(1/p_\wedge) - \delta]^2}\right)}. \tag{217}$$

Let $\widetilde{p}$ and $\widetilde{q}$ be the worst-case probabilities of reaching s_1 under the actions with nominal probabilities p and q, respectively. They equal $1 - y$ at $x = p_\wedge$ and $x = p_\wedge + \varepsilon$, respectively. Integrating equation 217 gives

$$\begin{aligned}\widetilde{p} - \widetilde{q} &= \int_{p_\wedge}^{p_\wedge + \varepsilon} \frac{dy}{dx}\, dx \\ &\geq \frac{\delta\varepsilon}{2p_\wedge \log(1/p_\wedge)\log\left(\dfrac{4}{p_\wedge^3[\log(1/p_\wedge) - \delta]^2}\right)} = \frac{4(1-\gamma)\epsilon}{\gamma}.\end{aligned} \tag{218}$$

Since finite KL divergence preserves the supports of deterministic transitions, s_1 and s_2 remain absorbing, with robust values $1/(1-\gamma)$ and zero, respectively. Thus action ϕ is optimal in model $\mathcal{M}_\phi$, and

$$V_\phi^*(s_0) = \frac{\gamma\widetilde{p}}{1-\gamma}.$$

For any returned randomized policy $\widehat{\pi}$, the (s,a)-rectangular uncertainty set allows the transition for each action to be minimized separately. The robust Bellman equation gives

$$\begin{aligned}V_\phi^{\widehat{\pi}}(s_0) &= \gamma\mathbb{E}_{a\sim\widehat{\pi}(\cdot|s_0)}\left[\inf_{P_{s_0,a}\in\mathcal{P}_{s_0,a}(\delta)} \mathbb{E}_{s'\sim P_{s_0,a}}\left[V_\phi^{\widehat{\pi}}(s')\right]\right] \\ &= \frac{\gamma}{1-\gamma}\left[\widehat{\pi}(\phi \mid s_0)\widetilde{p} + \big(1 - \widehat{\pi}(\phi \mid s_0)\big)\widetilde{q}\right].\end{aligned}$$

Combining this with equation 218, we obtain

$$\begin{aligned}\|V_\phi^* - V_\phi^{\widehat{\pi}}\|_\infty &= \frac{\gamma}{1-\gamma}(\widetilde{p} - \widetilde{q})\big(1 - \widehat{\pi}(\phi \mid s_0)\big) \\ &\geq 4\epsilon\big(1 - \widehat{\pi}(\phi \mid s_0)\big).\end{aligned} \tag{219}$$

The equality holds because the values at the absorbing states do not depend on the policy. Since s_0 is visited only initially, the same calculation applies to policies depending on the past.

To distinguish the two MDPs, we bound the KL divergence of their data distributions. The models interchange p and q between actions 0 and 1, and all other transition distributions and rewards are identical. With N independent samples per state-action pair, independence gives

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) &= \mathbb{E}_{\mathbb{P}_0}\left[\log\frac{\mathbb{P}_0}{\mathbb{P}_1}\right] \\ &= N\big(\mathrm{KL}(p\|q) + \mathrm{KL}(q\|p)\big) \\ &= N\big(\mathrm{KL}(p_\wedge\|p_\wedge + \varepsilon) + \mathrm{KL}(p_\wedge + \varepsilon\|p_\wedge)\big).\end{aligned} \tag{220}$$

The ratio in the expectation is evaluated at the observed dataset. For any $x, y \in (0,1)$, the inequality $\log t \le t-1$ gives

$$\begin{aligned}\mathrm{KL}(x\|y) &\le x\left(\frac{x}{y}-1\right)+(1-x)\left(\frac{1-x}{1-y}-1\right)\\ &= \frac{(x-y)^2}{y(1-y)}.\end{aligned}$$

When $x, y \in [p_\wedge, 2p_\wedge]$, the denominator is at least $p_\wedge(1-2p_\wedge) \ge 3p_\wedge/4$. Therefore

$$\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) \le \frac{8N\varepsilon^2}{3p_\wedge} \le \frac{4N\varepsilon^2}{p_\wedge}. \tag{221}$$

In particular, if

$$N \le \frac{p_\wedge}{32\varepsilon^2} = \frac{\gamma^2\delta^2}{2048p_\wedge(1-\gamma)^2\epsilon^2\log^2(1/p_\wedge)\log^2\left(\dfrac{4}{p_\wedge^3[\log(1/p_\wedge)-\delta]^2}\right)}, \tag{222}$$

then the KL divergence of the two data distributions is at most $1/8$.

For any event E, grouping its probability and that of its complement in the KL divergence gives

$$\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) \ge \mathrm{KL}\big(\mathbb{P}_0(E)\|\mathbb{P}_1(E)\big) \ge 2\big(\mathbb{P}_0(E)-\mathbb{P}_1(E)\big)^2.$$

The first inequality follows from convexity of $t\log t$, applied separately on the event and its complement. The second follows because, for fixed y, the function $x \mapsto \mathrm{KL}(x\|y)$ is zero with zero derivative at $x=y$ and has second derivative $1/[x(1-x)] \ge 4$; boundary cases follow by continuity. It follows that

$$\sup_E |\mathbb{P}_0(E)-\mathbb{P}_1(E)| \le \frac{1}{4}. \tag{223}$$

Including independent algorithmic randomization with the observations leaves the KL divergence and equation 223 unchanged, since its law is identical in both models.

Given the returned policy, choose model 0 when $\widehat{\pi}(0 \mid s_0) \ge 1/2$, and choose model 1 otherwise. Under $\mathcal{M}_0$, choosing model 1 implies $\widehat{\pi}(0 \mid s_0) < 1/2$ and hence a value loss greater than 2ϵ by equation 219. Under $\mathcal{M}_1$, choosing model 0 implies $\widehat{\pi}(1 \mid s_0) \le 1/2$ and hence a value loss at least 2ϵ. Both losses are strictly greater than ϵ. Equation 223 gives

$$\begin{aligned}&\mathbb{P}_0\big(\widehat{\pi}(0 \mid s_0) < 1/2\big) + \mathbb{P}_1\big(\widehat{\pi}(0 \mid s_0) \ge 1/2\big)\\ &= 1 - \mathbb{P}_0\big(\widehat{\pi}(0 \mid s_0) \ge 1/2\big) + \mathbb{P}_1\big(\widehat{\pi}(0 \mid s_0) \ge 1/2\big) \ge \frac{3}{4}.\end{aligned}$$

Consequently, for every learning algorithm,

$$\max_{\phi\in\{0,1\}} \mathbb{P}_\phi\Big(\|V_\phi^* - V_\phi^{\widehat{\pi}}\|_\infty > \epsilon\Big) \ge \frac{3}{8} > \frac{1}{8}. \tag{224}$$

Taking the infimum over learning algorithms preserves this lower bound.

For $|\mathcal{S}| > 3$, add $|\mathcal{S}|-3$ absorbing states with reward zero, unreachable from s_0. For $|\mathcal{A}| > 2$, give each additional action at s_0 the probabilities q and $1-q$ in both models. At absorbing states, every action remains absorbing with the same reward. The minimal non-zero nominal probability remains $p_\wedge$, and the added rows contribute zero to equation 220. Since every action other than ϕ at s_0 has worst-case probability $\widetilde{q}$ of reaching s_1, equation 219 continues to hold. Moreover, $\widehat{\pi}(0 \mid s_0) \ge 1/2$ implies $\widehat{\pi}(1 \mid s_0) \le 1/2$, so the same error bound applies.

There are N samples per state-action pair, hence $N|\mathcal{S}||\mathcal{A}|$ samples in total. Multiplying equation 222 by $|\mathcal{S}||\mathcal{A}|$ gives

$$N|\mathcal{S}||\mathcal{A}| \le \frac{\gamma^2\delta^2|\mathcal{S}||\mathcal{A}|}{2048p_\wedge(1-\gamma)^2\epsilon^2\log^2(1/p_\wedge)\log^2\left(\dfrac{4}{p_\wedge^3[\log(1/p_\wedge)-\delta]^2}\right)}.$$

Combining this sample condition with equation 224 and taking the infimum over all learning algorithms gives equation 14.

Case 2: $\delta \geq \log(1/p_\wedge)$

With the same states and rewards, we set

$$p = 1, \qquad q = 1 - p_\wedge.$$

Thus action ϕ at s_0 reaches s_1 with probability one, whereas action $1-\phi$ reaches s_2 with nominal probability $p_\wedge$. A transition to s_2 with probability one has KL divergence $\log(1/p_\wedge) \leq \delta$ from the latter row. The deterministic row of action ϕ cannot change its support. Consequently, the worst-case probabilities of reaching s_1 are

$$\widetilde{p} = 1, \qquad \widetilde{q} = 0.$$

The same robust Bellman calculation as above gives

$$\|V_\phi^* - V_\phi^{\widehat{\pi}}\|_\infty = \frac{\gamma}{1-\gamma}\big(1 - \widehat{\pi}(\phi \mid s_0)\big). \tag{225}$$

Consider the event that all samples from both actions 0 and 1 at s_0 are s_1. In either model, one of these two rows is deterministic and the other reaches s_1 with probability $1 - p_\wedge$. The event therefore has the same probability $(1-p_\wedge)^N$ under both models. Conditional on this event, the entire observed dataset has the same distribution in both models: the samples from these two rows are fixed, and all remaining rows have identical distributions. The same is true after including the independent randomization of the algorithm.

Use the same decision rule as before: choose model 0 if $\widehat{\pi}(0 \mid s_0) \geq 1/2$ and model 1 otherwise. The sum of its conditional error probabilities on the event is one. Hence

$$\begin{aligned}&\mathbb{P}_0\big(\widehat{\pi}(0 \mid s_0) < 1/2\big) + \mathbb{P}_1\big(\widehat{\pi}(0 \mid s_0) \geq 1/2\big)\\ &\geq (1-p_\wedge)^N \geq 1 - Np_\wedge \geq \frac{3}{4},\end{aligned}$$

where the last inequality follows from $N \leq 1/(4p_\wedge)$. An incorrect choice of model implies $\widehat{\pi}(\phi \mid s_0) \leq 1/2$ and therefore a value loss at least $\gamma/[2(1-\gamma)]$, which is strictly greater than ϵ under the stated accuracy condition. It follows that the maximum failure probability is at least $3/8$. Additional states and actions are added exactly as above, with every additional action at s_0 having nominal probabilities $q = 1 - p_\wedge$ and $1 - q = p_\wedge$ in both models. They preserve both the conditional equality of the data distributions and the minimal non-zero nominal probability $p_\wedge$.

Taking the infimum over learning algorithms gives equation 14, which completes the proof of Theorem 3.5.

E Proof of Lemma 3.7

Proof of Lemma 3.7. From the duality Lagrange function, we can get that

$$\bar{V} = \max_{\lambda \geq 0} \left\{ -\lambda \log\left(\mathbb{E}_{P(\cdot|s,a)}\big[\exp(-V(s')/\lambda)\big] \right) - \lambda\delta \right\}. \tag{226}$$

Then, define

$$\lambda^\star = \arg\max_{\lambda \geq 0} \left\{ -\lambda \log\left(\mathbb{E}_{P(\cdot|s,a)}\big[\exp(-V(s')/\lambda)\big] \right) - \lambda\delta \right\}.$$

We could get the worst-case mean value function as

$$\bar{V} = -\lambda^\star \log\left(\mathbb{E}_{P(\cdot|s,a)}\big[\exp(-V(s')/\lambda^\star)\big] \right) - \lambda^\star\delta.$$

Next, we set worst-case state, worst-case value function and corresponding probability as:

$$s_{\min} = \arg\min_{s': P(s'|s,a)>0} V(s'), V(s_{\min}) = V_{\min}, P_{\min} = P(s_{\min} \mid s, a). \tag{227}$$

Then, from the fact $\mathbb{E}_{P(\cdot|s,a)}\big[\exp(-V(s')/\lambda^\star)\big] \geq P_{\min}\exp(-V_{\min}/\lambda^\star)$, we could get the following result:

$$\begin{aligned}\bar{V} &\leq -\lambda^\star \log\left(P_{\min}\exp(-V_{\min}/\lambda^\star) \right) - \lambda^\star\delta,\\ &\leq V_{\min} - \lambda^\star \log(P_{\min}) - \lambda^\star\delta\\ &\leq V_{\min} + \lambda^\star\delta - \lambda^\star\delta\\ &\leq V_{\min},\end{aligned} \tag{228}$$

where the third inequality follows from $\log(P_{\min}) > -\delta$.

Besides, combined with the fact $\bar{V} \geq V_{\min}$, we can conclude that

$$\bar{V} = V_{\min}.$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please check Section B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper provide the full set of assumptions and a complete and correct proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is theory work, which does not include any directly social impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not include experiments.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The LLM was used solely for grammar polishing and LaTeX formatting in this work; all technical results have been independently validated.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.