The Curious Price of Distributional Robustness in Reinforcement Learning with a Generative Model

Laixi Shi[*] Caltech Gen Li[†] CUHK Yuting Wei[‡] UPenn Yuxin Chen[‡] UPenn Matthieu Geist[§] Cohere Yuejie Chi[¶] CMU

May 2023; Revised April 12, 2024

Abstract

This paper investigates model robustness in reinforcement learning (RL) to reduce the sim-to-real gap in practice. We adopt the framework of distributionally robust Markov decision processes (RMDPs), aimed at learning a policy that optimizes the worst-case performance when the deployed environment falls within a prescribed uncertainty set around the nominal MDP. Despite recent efforts, the sample complexity of RMDPs remained mostly unsettled regardless of the uncertainty set in use. It was unclear if distributional robustness bears any statistical consequences when benchmarked against standard RL. Assuming access to a generative model that draws samples based on the nominal MDP, we characterize the sample complexity of RMDPs when the uncertainty set is specified via either the total variation (TV) distance or χ^2 divergence. The algorithm studied here is a model-based method called *distributionally robust value iteration*, which is shown to be near-optimal for the full range of uncertainty levels. Somewhat surprisingly, our results uncover that RMDPs are not necessarily easier or harder to learn than standard MDPs. The statistical consequence incurred by the robustness requirement depends heavily on the size and shape of the uncertainty set: in the case w.r.t. the TV distance, the minimax sample complexity of RMDPs is always smaller than that of standard MDPs; in the case w.r.t. the χ^2 divergence, the sample complexity of RMDPs can often far exceed the standard MDP counterpart.

Keywords: distributionally robust RL, robust Markov decision processes, sample complexity, distributionally robust value iteration, model-based RL

Contents

1 Introduction **2**
1.1 Prior art and open questions 3
1.2 Main contributions 4

2 Problem formulation **7**

3 Model-based algorithm: distributionally robust value iteration **9**

4 Theoretical guarantees: sample complexity analyses **11**
4.1 The case of TV distance: RMDPs are easier to learn than standard MDPs 11
4.2 The case of χ^2 divergence: RMDPs can be harder than standard MDPs 12

[*] Department of Computing Mathematical Sciences, California Institute of Technology, CA 91125, USA. Part of L. Shi's work was completed when she was at CMU.

[†] Department of Statistics, The Chinese University of Hong Kong, Hong Kong.

[‡] Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

[§] Cohere.

[¶] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

5 Analysis: the TV case 14
5.1 Preliminaries of the analysis 14
5.2 Proof of the upper bound with TV distance: Theorem 1 16
5.3 Proof of the lower bound with TV distance: Theorem 2 19

6 Offline distributionally robust RL with uniform coverage 23
6.1 The case of TV distance 23
6.2 The case of χ^2 divergence 24

7 Other related works 25

8 Discussions 26

A Proof of the preliminaries 32
A.1 Proof of Lemma 1 and Lemma 2 32
A.2 Proof of Lemma 5 33

B Proof of the auxiliary lemmas for Theorem 1 34
B.1 Proof of Lemma 6 34
B.2 Proof of Lemma 7 35
B.3 Proof of Lemma 8 35
B.4 Proof of Lemma 9 40
B.5 Proof of Lemma 11 43
B.6 Proof of Lemma 12 44
B.7 Proof of Lemma 14 45
B.8 Proof of Lemma 15 49

C Proof of the auxiliary facts for Theorem 2 50
C.1 Proof of Lemma 10 50
C.2 Proof of the claim (75) 52

D Proof of the upper bound with χ^2 divergence: Theorem 3 53
D.1 Proof of Theorem 3 54
D.2 Proof of the auxiliary lemmas 55

E Proof of the lower bound with χ^2 divergence: Theorem 4 59
E.1 Construction of the hard problem instances 59
E.2 Establishing the minimax lower bound 60
E.3 Proof of the auxiliary facts 63

F Proof for the offline setting 67
F.1 Proof of the upper bounds: Corollary 1 and Corollary 3 67
F.2 Proof of the lower bounds: Corollary 2 and Corollary 4 69

1 Introduction

Reinforcement learning (RL) strives to learn desirable sequential decisions based on trial-and-error interactions with an unknown environment. As a fast-growing subfield of artificial intelligence, it has achieved remarkable success in a variety of applications, such as networked systems (Qu et al., 2022), trading (Park and Van Roy, 2015), operations research (de Castro Silva et al., 2003; Pan et al., 2023; Zhao et al., 2021), large language model alignment (OpenAI, 2023; Ziegler et al., 2019), healthcare (Fatemi et al., 2021; Liu et al., 2019), robotics and control (Kober et al., 2013; Mnih et al., 2013). Due to the unprecedented dimensionality of the state-action space, the issue of data efficiency inevitably lies at the core of modern RL practice. A large portion of recent efforts in RL has been directed towards designing sample-efficient algorithms and understanding the fundamental statistical bottleneck for a diverse range of RL scenarios.

While standard RL has been heavily investigated recently, its use can be significantly hampered in practice due to the sim-to-real gap or uncertainty (Bertsimas et al., 2019); for instance, a policy learned in an ideal, nominal environment might fail catastrophically when the deployed environment is subject to small changes in task objectives or adversarial perturbations (Klopp et al., 2017; Mahmood et al., 2018; Zhang et al., 2020a). Consequently, in addition to maximizing the long-term cumulative reward, robustness emerges as another critical goal for RL, especially in high-stakes applications such as robotics, autonomous driving, clinical trials, financial investments, and so on. Towards achieving this, distributionally robust RL (Bäuerle and Glauner, 2022; Cai et al., 2016; Iyengar, 2005; Nilim and El Ghaoui, 2005; Xu and Mannor, 2012), which leverages insights from distributionally robust optimization and supervised learning (Bertsimas et al., 2018; Blanchet and Murthy, 2019; Chen et al., 2019; Duchi and Namkoong, 2018; Gao, 2020; Lam, 2019; Rahimian and Mehrotra, 2019), becomes a natural yet versatile framework; the aim is to learn a policy that performs well even when the deployed environment deviates from the nominal one in the face of environment uncertainty.

In this paper, we pursue fundamental understanding about whether, and how, the choice of distributional robustness bears statistical implications in learning a desirable policy, through the lens of sample complexity. More concretely, imagine that one has access to a generative model (also called a simulator) that draws samples from a Markov decision processes (MDP) with a nominal transition kernel (Kearns and Singh, 1999). Standard RL aims to learn the optimal policy tailored to the nominal kernel, for which the minimax sample complexity limit has been fully settled (Azar et al., 2013b; Li et al., 2023b). In contrast, distributionally robust RL seeks to learn a more *robust* policy using the same set of samples, with the aim of optimizing the worst-case performance when the transition kernel is arbitrarily chosen from some *prescribed* uncertainty set around the nominal kernel; this setting is frequently referred to as robust MDPs (RMDPs).[1] Clearly, the RMDP framework helps ensure that the performance of the learned policy does not fail catastrophically as long as the sim-to-real gap is not overly large. It is then natural to wonder how the robustness consideration impacts data efficiency: is there a statistical premium that one needs to pay in quest of additional robustness?

Compared with standard MDPs, the class of RMDPs encapsulates richer models, given that one is allowed to prescribe the shape and size of the uncertainty set. Oftentimes, the uncertainty set is hand-picked as a small ball surrounding the nominal kernel, with the size and shape of the ball specified by some distance-like metric ρ between probability distributions and some uncertainty level σ. To ensure tractability of solving RMDPs, the uncertainty set is often selected to obey certain structures. For instance, a number of prior works assumed that the uncertainty set can be decomposed as a product of independent uncertainty subsets over each state or state-action pair (Wiesemann et al., 2013; Zhou et al., 2021), dubbed as the s- and (s,a)-rectangularity, respectively. The current paper adopts the second choice by assuming (s,a)-rectangularity for the uncertainty set. An additional challenge with RMDPs arises from distribution shift, where the transition kernel drawn from the uncertainty set can be different from the nominal kernel. This challenge leads to complicated nonlinearity and nested optimization in the problem structure not present in standard MDPs.

1.1 Prior art and open questions

In this paper, we focus attention on RMDPs in the context of γ-discounted infinite-horizon setting, assuming access to a generative model. The uncertainty set considered herein is specified using one of the f-divergence metrics: the total variation (TV) distance and the χ^2 divergence. These two choices are motivated by their practical appeals: easy to implement, and already adopted by empirical RL (Lee et al., 2021; Pan et al., 2023).

A popular learning approach is model-based, which first estimates the nominal transition kernel using a plug-in estimator based on the collected samples, and then runs a planning algorithm (e.g., a robust variant of value iteration) on top of the estimated kernel. Despite the surge of recent activities, however, existing statistical guarantees for the above paradigm remained highly inadequate, as we shall elaborate on momentarily (see Table 1 and Table 2 respectively for a summary of existing results). For concreteness, let S be the size of the state space, A the size of the action space, γ the discount factor (so that the effective horizon is $\frac{1}{1-\gamma}$), and σ the uncertainty level. We are interested in how the sample complexity — the number of samples needed for an algorithm to output a policy whose robust value function (the worst-case value

[1] While it is straightforward to incorporate additional uncertainty of the reward in our framework, we do not consider it here for simplicity, since the key challenge is to deal with the uncertainty of the transition kernel.

Result type	Reference	Sample complexity	
		$0 < \sigma \lesssim 1-\gamma$	$1-\gamma \lesssim \sigma < 1$
Upper bound	Yang et al. (2022)	$\frac{S^2 A}{\sigma^2(1-\gamma)^4\varepsilon^2}$	
	Panaganti and Kalathil (2022)	$\frac{S^2 A}{(1-\gamma)^4\varepsilon^2}$	
	This paper	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$
Lower bound	Yang et al. (2022)	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA(1-\gamma)}{\sigma^4\varepsilon^2}$
	This paper	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$

Table 1: Comparisons between our results and prior arts for finding an ε-optimal robust policy in the infinite-horizon RMDPs with a generative model, where the uncertainty set is measured w.r.t. the TV distance. Here, S, A, γ, and $\sigma \in (0,1)$ are the state space size, the action space size, the discount factor, and the uncertainty level, respectively, and all logarithmic factors are omitted in the table. Our results provide the first matching upper and lower bounds (up to log factors), improving upon all prior results.

over all the transition kernels in the uncertainty set) is at most ε away from the optimal robust one — scales with all these salient problem parameters.

- *Large gaps between existing upper and lower bounds.* There remained large gaps between the sample complexity upper and lower bounds established in prior literature, regardless of the divergence metric in use. Specifically, considering the cases using either TV distance or χ^2 divergence, the state-of-the-art upper bounds (Panaganti and Kalathil, 2022) scales quadratically with the size S of the state space, while the lower bound (Yang et al., 2022) exhibits only linear scaling with S. Moreover, in the χ^2 divergence case, the state-of-the-art upper bound grows linearly with the uncertainty level σ when $\sigma \gtrsim 1$,[2] while the lower bound (Yang et al., 2022) is inversely proportional to σ. These lead to unbounded gaps between the upper and lower bounds as σ grows. *Can we hope to close these gaps for RMDPs?*

- *Benchmarking with standard MDPs.* Perhaps a more pressing issue is that, past works failed to provide an affirmative answer regarding how to benchmark the sample complexity of RMDPs with that of standard MDPs regardless of the chosen shape (determined by ρ) or size (determined by σ) of the uncertainty set, given the large unresolved gaps mentioned above. Specifically, existing sample complexity upper (resp. lower) bounds are all larger (resp. smaller) than the sample size requirement for standard MDPs. As a consequence, it remains mostly unclear *whether learning RMDPs is harder or easier than learning standard MDPs.*

1.2 Main contributions

To address the aforementioned questions, this paper develops strengthened sample complexity upper bounds on learning RMDPs with the TV distance and χ^2 divergence in the infinite-horizon setting, using a model-based approach called distributionally robust value iteration (DRVI). Improved minimax lower bounds are also developed to help gauge the tightness of our upper bounds and enable benchmarking with standard MDPs. The novel analysis framework developed herein leads to new insights into the interplay between the geometry of uncertainty sets and statistical hardness.

[2] Let $\mathcal{X} := \big(S, A, \frac{1}{1-\gamma}, \sigma, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$. The notation $f(\mathcal{X}) = O(g(\mathcal{X}))$ or $f(\mathcal{X}) \lesssim g(\mathcal{X})$ indicates that there exists a universal constant $C_1 > 0$ such that $f \le C_1 g$, the notation $f(\mathcal{X}) \gtrsim g(\mathcal{X})$ indicates that $g(\mathcal{X}) = O(f(\mathcal{X}))$, and the notation $f(\mathcal{X}) \asymp g(\mathcal{X})$ indicates that $f(\mathcal{X}) \lesssim g(\mathcal{X})$ and $f(\mathcal{X}) \gtrsim g(\mathcal{X})$ hold simultaneously. Additionally, the notation $\widetilde{O}(\cdot)$ is defined in the same way as $O(\cdot)$ except that it hides logarithmic factors.

Result type	Reference	Sample complexity		
		$0 < \sigma \lesssim 1-\gamma$	$1-\gamma \lesssim \sigma \lesssim \frac{1}{1-\gamma}$	$\sigma \gtrsim \frac{1}{1-\gamma}$
Upper bound	Panaganti and Kalathil (2022)	$\frac{S^2A(1+\sigma)}{(1-\gamma)^4\varepsilon^2}$		
	Yang et al. (2022)	$\frac{S^2A(1+\sigma)^2}{(\sqrt{1+\sigma}-1)^2(1-\gamma)^4\varepsilon^2}$		
	This paper	$\frac{SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}$		
Lower bound	Yang et al. (2022)	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$	
	This paper	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA\sigma}{(1-\gamma)^4(1+\sigma)^4\varepsilon^2}$	$\frac{SA\sigma}{\varepsilon^2}$

Table 2: Comparisons between our results and prior art on finding an ε-optimal robust policy in the infinite-horizon RMDPs with a generative model, where the uncertainty set is measured w.r.t. the χ^2 divergence. Here, S, A, γ, and $\sigma \in (0, \infty)$ are the state space size, the action space size, the discount factor, and the uncertainty level, respectively, and all logarithmic factors are omitted in the table. Improving upon all prior results, our theory is tight (up to log factors) when $\sigma \asymp 1$, and otherwise loose by no more than a polynomial factor in $1/(1-\gamma)$.

Sample complexity of RMDPs under the TV distance. We summarize our results and compare them with past works in Table 1; see Figure 1(a) for a graphical illustration.

- **Minimax-optimal sample complexity.** We prove that DRVI reaches ε accuracy as soon as the sample complexity is on the order of

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\varepsilon^2}\min\left\{\frac{1}{1-\gamma}, \frac{1}{\sigma}\right\}\right)$$

  for all $\sigma \in (0, 1)$, assuming that ε is small enough. In addition, a matching minimax lower bound (modulo some logarithmic factor) is established to guarantee the tightness of the upper bound. To the best of our knowledge, this is the *first* minimax-optimal sample complexity for RMDPs, which was previously unavailable regardless of the divergence metric and uncertainty level in use and is over the full range of the uncertainty level.

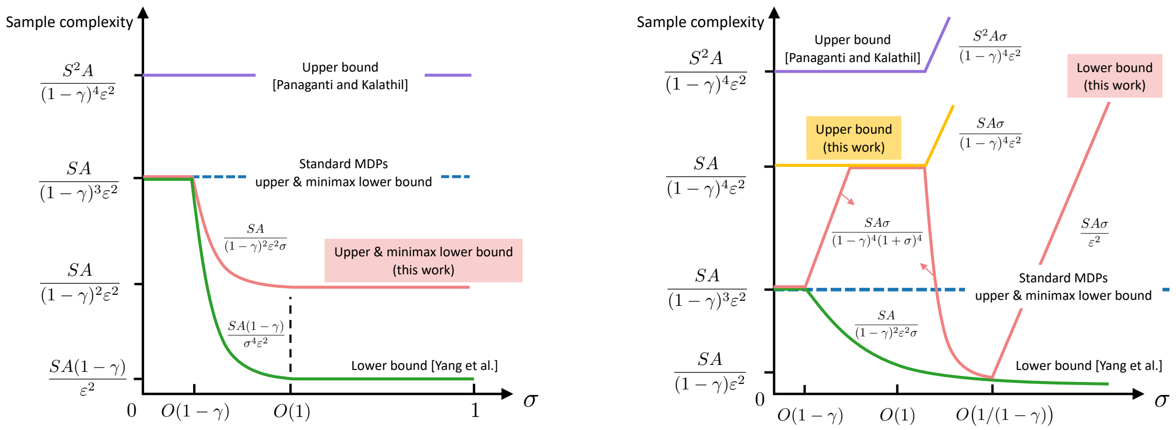
- **RMDPs are easier to learn than standard MDPs under the TV distance.** Given the sample complexity $\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right)$ of standard MDPs (Li et al., 2023b), it can be seen that learning RMDPs under the TV distance is never harder than learning standard MDPs; more concretely, the sample complexity for RMDPs matches that of standard MDPs when $\sigma \lesssim 1-\gamma$, and becomes smaller by a factor of $\sigma/(1-\gamma)$ when $1-\gamma \lesssim \sigma < 1$. Therefore, in this case, distributional robustness comes almost for free, given that we do not need to collect more samples.

Sample complexity of RMDPs under the χ^2 divergence. We summarize our results and provide comparisons with prior works in Table 2; see Figure 1(b) for an illustration.

- **Near-optimal sample complexity.** We demonstrate that DRVI yields ε accuracy as soon as the sample complexity is on the order of

$$\widetilde{O}\left(\frac{SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\right)$$

  for all $\sigma \in (0, \infty)$, which is the first sample complexity in this setting that scales linearly in the size S of the state space; in other words, our theory breaks the quadratic scaling bottleneck that was present in prior works (Panaganti and Kalathil, 2022; Yang et al., 2022). We have also developed a strengthened



(a) TV distance

(b) χ^2 divergence

Figure 1: Illustrations of the obtained sample complexity upper and lower bounds for learning RMDPs with comparisons to state-of-the-art and the sample complexity of standard MDPs, where the uncertainty set is specified using the TV distance (a) and the χ^2 divergence (b).

lower bound that is optimized by leveraging the geometry of the uncertainty set under different ranges of σ. Our theory is tight when $\sigma \asymp 1$, and is otherwise loose by at most a polynomial factor of the effective horizon $1/(1-\gamma)$ (regardless of the uncertainty level σ). This significantly improves upon prior results (as there exists an unbounded gap between prior upper and lower bounds as $\sigma \to \infty$).

- **RMDPs can be harder to learn than standard MDPs under the χ^2 divergence.** Somewhat surprisingly, our improved lower bound suggests that RMDPs in this case can be much harder to learn than standard MDPs, at least for a certain range of uncertainty levels. We single out two regimes of particular interest. Firstly, when $\sigma \asymp 1$, the sample size requirement of RMDPs is on the order of $\frac{SA}{(1-\gamma)^4\varepsilon^2}$ (up to log factor), which is provably larger than the one for standard MDPs by a factor of $\frac{1}{1-\gamma}$. Secondly, the lower bound continues to increase as σ grows and exceeds the sample complexity of standard MDPs when $\sigma \gtrsim \frac{1}{(1-\gamma)^3}$.

In sum, our sample complexity bounds not only strengthen the prior art in the development of both upper and lower bounds, but also unveil that the additional robustness consideration might affect the sample complexity in a somewhat surprising manner. As it turns out, RMDPs are not necessarily harder nor easier to learn than standard MDPs; the conclusion is far more nuanced and highly dependent on both the size and shape of the uncertainty set. This constitutes a curious phenomenon that has not been elucidated in prior analyses.

Technical novelty. Our upper bound analyses require careful treatments of the impact of the uncertainty set upon the value functions, and decouple the statistical dependency across the iterates of the robust value iteration using tailored leave-one-out arguments (Agarwal et al., 2020; Li et al., 2022b) that have not been introduced to the RMDP setting previously. Turning to the lower bound, we develop new hard instances that differ from those for standard MDPs (Azar et al., 2013a; Li et al., 2024). These new instances draw inspiration from the asymmetric structure of RMDPs induced by the additional infimum operator in the robust value function. In addition, we construct a series of hard instances depending on the uncertainty level σ to establish the tight lower bound as σ varies.

Extension: offline RL with uniform coverage. Last but not least, we extend our analysis framework to accommodate a widely studied offline setting with uniform data coverage (Yang et al., 2022; Zhou et al., 2021) in Section 6. In particular, given a historical dataset with minimal coverage probability $\mu_{\min}$ over the state-action space (see Assumption 1), we provide sample complexity results for both cases with TV distance or χ^2 divergence, where in effect the dependency with the size of the state-action space SA is replaced by $1/\mu_{\min}$. The sample complexity upper bounds significantly improve upon prior art (Yang et al., 2022) by a factor of $\frac{S}{(1-\gamma)^2}$ (resp. $S(1+\sigma)$) when the uncertainty set is measured by the TV distance (resp. the χ^2 divergence).

Notation and paper organization. Throughout this paper, we denote by $\Delta(\mathcal{S})$ the probability simplex over a set $\mathcal{S}$ and $x = \big[x(s,a)\big]_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{SA}$ (resp. $x = \big[x(s)\big]_{s\in\mathcal{S}} \in \mathbb{R}^{S}$) as any vector that constitutes certain values for each state-action pair (resp. state). In addition, we denote by $x \circ y = \big[x(s)\cdot y(s)\big]_{s\in\mathcal{S}}$ the Hadamard product of any two vectors $x, y \in \mathbb{R}^S$.

The remainder of this paper is structured as follows. Section 2 presents the background about discounted infinite-horizon standard MDPs and formulates distributionally robust MDPs. In Section 3, a model-based approach is introduced, tailored to both the TV distance and the χ^2 divergence. Both upper and lower bounds on the sample complexity are developed in Section 4, covering both divergence metrics. Section 5 provides an outline of our analysis. Section 6 further extends the findings to the offline RL setting with uniform data coverage. We then summarize several additional related works in Section 7 and conclude the main paper with further discussions in Section 8. The proof details are deferred to the appendix.

2 Problem formulation

In this section, we formulate distributionally robust Markov decision processes (RMDPs) in the discounted infinite-horizon setting, introduce the sampling mechanism, and describe our goal.

Standard MDPs. To begin, we first introduce the standard Markov decision processes (MDPs), which facilitate the understanding of RMDPs. A discounted infinite-horizon MDP is represented by $\mathcal{M} = \big(\mathcal{S}, \mathcal{A}, \gamma, P, r\big)$, where $\mathcal{S} = \{1, \cdots, S\}$ and $\mathcal{A} = \{1, \cdots, A\}$ are the finite state and action spaces, respectively, $\gamma \in [0,1)$ is the discounted factor, $P : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ denotes the probability transition kernel, and $r : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the immediate reward function which is assumed to be deterministic. A policy is denoted by $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which specifies the action selection probability over the action space in any state. When the policy is deterministic, we overload the notation and refer to $\pi(s)$ as the action selected by policy π in state s. To characterize the cumulative reward, the value function $V^{\pi,P}$ for any policy π under the transition kernel P is defined by

$$\forall s \in \mathcal{S}: \qquad V^{\pi,P}(s) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\Big|\, s_0 = s\right], \tag{1}$$

where the expectation is taken over the randomness of the trajectory $\{s_t, a_t\}_{t=0}^{\infty}$ generated by executing policy π under the transition kernel P, namely, $a_t \sim \pi(\cdot \,|\, s_t)$ and $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ for all $t \geq 0$. Similarly, the Q-function $Q^{\pi,P}$ associated with any policy π under the transition kernel P is defined as

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad Q^{\pi,P}(s,a) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\Big|\, s_0 = s, a_0 = a\right], \tag{2}$$

where the expectation is again taken over the randomness of the trajectory under policy π.

Distributionally robust MDPs. We now introduce the distributionally robust MDP (RMDP) tailored to the discounted infinite-horizon setting, denoted by $\mathcal{M}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}_\rho^\sigma(P^0), r\}$, where $\mathcal{S}, \mathcal{A}, \gamma, r$ are identical to those in the standard MDP. A key distinction from the standard MDP is that: rather than assuming a fixed transition kernel P, it allows the transition kernel to be chosen arbitrarily from a prescribed uncertainty set $\mathcal{U}_\rho^\sigma(P^0)$ centered around a *nominal* kernel $P^0 : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$, where the uncertainty set is specified using some distance metric ρ of radius $\sigma > 0$. In particular, given the nominal transition kernel P^0 and some uncertainty level σ, the uncertainty set—with the divergence metric $\rho : \Delta(\mathcal{S})\times\Delta(\mathcal{S}) \to \mathbb{R}^+$—is specified as

$$\mathcal{U}_\rho^\sigma(P^0) \coloneqq \otimes\, \mathcal{U}_\rho^\sigma(P^0_{s,a}) \qquad \text{with} \quad \mathcal{U}_\rho^\sigma(P^0_{s,a}) \coloneqq \big\{P_{s,a} \in \Delta(\mathcal{S}) : \rho\left(P_{s,a}, P^0_{s,a}\right) \leq \sigma\big\}, \tag{3}$$

where we denote a vector of the transition kernel P or P^0 at state-action pair (s,a) respectively as

$$P_{s,a} \coloneqq P(\cdot \,|\, s,a) \in \mathbb{R}^{1\times S}, \qquad P^0_{s,a} \coloneqq P^0(\cdot \,|\, s,a) \in \mathbb{R}^{1\times S}. \tag{4}$$

In other words, the uncertainty is imposed in a decoupled manner for each state-action pair, obeying the so-called (s,a)-rectangularity (Wiesemann et al., 2013; Zhou et al., 2021).

In RMDPs, we are interested in the worst-case performance of a policy π over all the possible transition kernels in the uncertainty set. This is measured by the *robust value function* $V^{\pi,\sigma}$ and the *robust Q-function* $Q^{\pi,\sigma}$ in $\mathcal{M}_{\mathsf{rob}}$, defined respectively as

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad V^{\pi,\sigma}(s) \coloneqq \inf_{P\in\mathcal{U}^\sigma_\rho(P^0)} V^{\pi,P}(s), \qquad Q^{\pi,\sigma}(s,a) \coloneqq \inf_{P\in\mathcal{U}^\sigma_\rho(P^0)} Q^{\pi,P}(s,a). \tag{5}$$

Optimal robust policy and robust Bellman operator. As a generalization of properties of standard MDPs, it is well-known that there exists at least one deterministic policy that maximizes the robust value function (resp. robust Q-function) simultaneously for all states (resp. state-action pairs) (Iyengar, 2005; Nilim and El Ghaoui, 2005). Therefore, we denote the *optimal robust value function* (resp. *optimal robust Q-function*) as $V^{\star,\sigma}$ (resp. $Q^{\star,\sigma}$), and the optimal robust policy as $\pi^\star$, which satisfy

$$\forall s\in\mathcal{S}: \quad V^{\star,\sigma}(s) \coloneqq V^{\pi^\star,\sigma}(s) = \max_\pi V^{\pi,\sigma}(s), \tag{6a}$$

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad Q^{\star,\sigma}(s,a) \coloneqq Q^{\pi^\star,\sigma}(s,a) = \max_\pi Q^{\pi,\sigma}(s,a). \tag{6b}$$

A key machinery in RMDPs is a generalization of Bellman's optimality principle, encapsulated in the following *robust Bellman consistency equation* (resp. *robust Bellman optimality equation*):

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad Q^{\pi,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma_\rho(P^0_{s,a})} \mathcal{P}V^{\pi,\sigma}, \tag{7a}$$

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma_\rho(P^0_{s,a})} \mathcal{P}V^{\star,\sigma}. \tag{7b}$$

The robust Bellman operator (Iyengar, 2005; Nilim and El Ghaoui, 2005) is denoted by $\mathcal{T}^\sigma(\cdot): \mathbb{R}^{SA}\to\mathbb{R}^{SA}$ and defined as follows:

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad \mathcal{T}^\sigma(Q)(s,a) \coloneqq r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma_\rho(P^0_{s,a})} \mathcal{P}V, \quad \text{with} \quad V(s) \coloneqq \max_a Q(s,a). \tag{8}$$

Given that $Q^{\star,\sigma}$ is the unique fixed point of $\mathcal{T}^\sigma$, one can recover the optimal robust value function and Q-function using a procedure termed *distributionally robust value iteration* (DRVI). Generalizing the standard value iteration, DRVI starts from some given initialization and recursively applies the robust Bellman operator until convergence. As has been shown previously, this procedure converges rapidly due to the γ-contraction property of $\mathcal{T}^\sigma$ w.r.t. the ℓ_∞ norm (Iyengar, 2005; Nilim and El Ghaoui, 2005).

Specification of the divergence ρ. We consider two popular choices of the uncertainty set measured in terms of two different f-divergence metric: the total variation distance and the χ^2 divergence, given respectively by (Tsybakov, 2009)

$$\rho_{\mathsf{TV}}\left(P_{s,a}, P^0_{s,a}\right) \coloneqq \frac{1}{2}\left\|P_{s,a} - P^0_{s,a}\right\|_1 = \frac{1}{2}\sum_{s'\in\mathcal{S}} P^0(s' \,|\, s,a)\left|1 - \frac{P(s' \,|\, s,a)}{P^0(s' \,|\, s,a)}\right|, \tag{9}$$

$$\rho_{\chi^2}\left(P_{s,a}, P^0_{s,a}\right) \coloneqq \sum_{s'\in\mathcal{S}} P^0(s' \,|\, s,a)\left(1 - \frac{P(s' \,|\, s,a)}{P^0(s' \,|\, s,a)}\right)^2. \tag{10}$$

Note that $\rho_{\mathsf{TV}}\left(P_{s,a}, P^0_{s,a}\right) \in [0,1]$ and $\rho_{\chi^2}\left(P_{s,a}, P^0_{s,a}\right) \in [0,\infty)$ in general. As we shall see shortly, these two choices of divergence metrics result in drastically different messages when it comes to sample complexities.

Sampling mechanism: a generative model. Following Panaganti and Kalathil (2022); Zhou et al. (2021), we assume access to a generative model or a simulator (Kearns and Singh, 1999), which allows us to collect N independent samples for each state-action pair generated based on the *nominal* kernel P^0:

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A},\qquad s_{i,s,a}\overset{i.i.d}{\sim}P^0(\cdot\,|\,s,a),\qquad i=1,2,\cdots,N. \tag{11}$$

The total sample size is, therefore, NSA.

Goal. Given the collected samples, the task is to learn the robust optimal policy for the RMDP — w.r.t. some prescribed uncertainty set $\mathcal{U}^\sigma(P^0)$ around the nominal kernel — using as few samples as possible. Specifically, given some target accuracy level $\varepsilon>0$, the goal is to seek an ε-optimal robust policy $\widehat{\pi}$ obeying

$$\forall s\in\mathcal{S}:\quad V^{\star,\sigma}(s)-V^{\widehat{\pi},\sigma}(s)\le\varepsilon. \tag{12}$$

3 Model-based algorithm: distributionally robust value iteration

We consider a model-based approach tailored to RMDPs, which first constructs an empirical nominal transition kernel based on the collected samples, and then applies distributionally robust value iteration (DRVI) to compute an optimal robust policy.

Empirical nominal kernel. The empirical nominal transition kernel $\widehat{P}^0\in\mathbb{R}^{SA\times S}$ can be constructed on the basis of the empirical frequency of state transitions, i.e.,

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}:\quad \widehat{P}^0(s'\,|\,s,a):=\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\big\{s_{i,s,a}=s'\big\}, \tag{13}$$

which leads to an empirical RMDP $\widehat{\mathcal{M}}_{\mathsf{rob}}=\{\mathcal{S},\mathcal{A},\gamma,\mathcal{U}^\sigma_\rho(\widehat{P}^0),r\}$. Analogously, we can define the corresponding robust value function (resp. robust Q-function) of policy π in $\widehat{\mathcal{M}}_{\mathsf{rob}}$ as $\widehat{V}^{\pi,\sigma}$ (resp. $\widehat{Q}^{\pi,\sigma}$) (cf. (6)). In addition, we denote the corresponding *optimal robust policy* as $\widehat{\pi}^\star$ and the *optimal robust value function* (resp. *optimal robust Q-function*) as $\widehat{V}^{\star,\sigma}$ (resp. $\widehat{Q}^{\star,\sigma}$) (cf. (7)), which satisfies the robust Bellman optimality equation:

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}:\quad \widehat{Q}^{\star,\sigma}(s,a)=r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^\sigma_\rho(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}. \tag{14}$$

Equipped with $\widehat{P}^0$, we can define the empirical robust Bellman operator $\widehat{\mathcal{T}}^\sigma$ as

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}:\quad \widehat{\mathcal{T}}^\sigma(Q)(s,a):=r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^\sigma_\rho(\widehat{P}^0_{s,a})}\mathcal{P}V,\quad\text{with}\quad V(s):=\max_a Q(s,a). \tag{15}$$

DRVI: distributionally robust value iteration. To compute the fixed point of $\widehat{\mathcal{T}}^\sigma$, we introduce distributionally robust value iteration (DRVI), which is summarized in Algorithm 1. From an initialization $\widehat{Q}_0=0$, the update rule at the t-th ($t\ge1$) iteration can be formulated as:

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}:\quad \widehat{Q}_t(s,a)=\widehat{\mathcal{T}}^\sigma\big(\widehat{Q}_{t-1}\big)(s,a)=r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^\sigma_\rho(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}_{t-1}, \tag{16}$$

where $\widehat{V}_{t-1}(s)=\max_a\widehat{Q}_{t-1}(s,a)$ for all $s\in\mathcal{S}$. However, directly solving (16) is computationally expensive since it involves optimization over an S-dimensional probability simplex at each iteration, especially when the dimension of the state space $\mathcal{S}$ is large. Fortunately, in view of strong duality (Iyengar, 2005), (16) can be equivalently solved using its dual problem, which concerns optimizing a *scalar* dual variable and thus can be solved efficiently. In what follows, we shall illustrate this for the two choices of the divergence ρ of interest (cf. (9) and (10)). Before continuing, for any $V\in\mathbb{R}^S$, we denote $[V]_\alpha$ as its clipped version by some non-negative value α, namely,

$$[V]_\alpha(s):=\begin{cases}\alpha, & \text{if } V(s)>\alpha,\\ V(s), & \text{otherwise.}\end{cases} \tag{17}$$

Algorithm 1: Distributionally robust value iteration (DRVI) for infinite-horizon RMDPs.

```
input: empirical nominal transition kernel $\widehat{P}^0$; reward function $r$; uncertainty level $\sigma$; number of
  iterations $T$.
initialization: $\widehat{Q}_0(s,a) = 0$, $\widehat{V}_0(s) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
for $t = 1, 2, \cdots, T$ do
    for $s \in \mathcal{S}, a \in \mathcal{A}$ do
        Set $\widehat{Q}_t(s,a)$ according to (16);
    for $s \in \mathcal{S}$ do
        Set $\widehat{V}_t(s) = \max_a \widehat{Q}_t(s,a)$;
output: $\widehat{Q}_T$, $\widehat{V}_T$ and $\widehat{\pi}$ obeying $\widehat{\pi}(s) \coloneqq \arg\max_a \widehat{Q}_T(s,a)$.
```

- TV distance, where the uncertainty set is $\mathcal{U}^\sigma_\rho(\widehat{P}^0_{s,a}) \coloneqq \mathcal{U}^\sigma_{\mathsf{TV}}(\widehat{P}^0_{s,a}) \coloneqq \mathcal{U}^\sigma_{\rho_{\mathsf{TV}}}(\widehat{P}^0_{s,a})$ w.r.t. the TV distance $\rho = \rho_{\mathsf{TV}}$ defined in (9). In particular, we have the following lemma due to strong duality, which is a direct consequence of Iyengar (2005, Lemma 4.3).

  Lemma 1 (Strong duality for TV)**.** *Consider any probability vector $P \in \Delta(\mathcal{S})$, any fixed uncertainty level σ and the uncertainty set $\mathcal{U}^\sigma(P) \coloneqq \mathcal{U}^\sigma_{\mathsf{TV}}(P)$. For any vector $V \in \mathbb{R}^S$ obeying $V \geq 0$, recalling the definition of $[V]_\alpha$ in (17), one has*

$$\inf_{\mathcal{P} \in \mathcal{U}^\sigma(P)} \mathcal{P}V = \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ P\left[V\right]_\alpha - \sigma\left(\alpha - \min_{s'} \left[V\right]_\alpha (s')\right)\right\}. \tag{18}$$

  In view of the above lemma, the following dual update rule is equivalent to (16) in DRVI:

$$\widehat{Q}_t(s,a) = r(s,a) + \gamma \max_{\alpha \in \left[\min_s \widehat{V}_{t-1}(s), \max_s \widehat{V}_{t-1}(s)\right]} \left\{ \widehat{P}^0_{s,a}\left[\widehat{V}_{t-1}\right]_\alpha - \sigma\left(\alpha - \min_{s'} \left[\widehat{V}_{t-1}\right]_\alpha (s')\right)\right\}. \tag{19}$$

- χ^2 divergence, where the uncertainty set is $\mathcal{U}^\sigma_\rho(\widehat{P}^0_{s,a}) \coloneqq \mathcal{U}^\sigma_{\chi^2}(\widehat{P}^0_{s,a}) \coloneqq \mathcal{U}^\sigma_{\rho_{\chi^2}}(\widehat{P}^0_{s,a})$ w.r.t. the χ^2 divergence $\rho = \rho_{\chi^2}$ defined in (10). We introduce the following lemma which directly follows from (Iyengar, 2005, Lemma 4.2).

  Lemma 2 (Strong duality for χ^2)**.** *Consider any probability vector $P \in \Delta(\mathcal{S})$, any fixed uncertainty level σ and the uncertainty set $\mathcal{U}^\sigma(P) \coloneqq \mathcal{U}^\sigma_{\chi^2}(P)$. For any vector $V \in \mathbb{R}^S$ obeying $V \geq 0$, one has*

$$\inf_{\mathcal{P} \in \mathcal{U}^\sigma(P)} \mathcal{P}V = \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ P[V]_\alpha - \sqrt{\sigma \mathsf{Var}_P\left([V]_\alpha\right)}\right\}, \tag{20}$$

  where $\mathsf{Var}_P(\cdot)$ is defined as (40).

  In view of the above lemma, the update rule (16) in DRVI can be equivalently written as:

$$\widehat{Q}_t(s,a) = r(s,a) + \gamma \max_{\alpha \in \left[\min_s \widehat{V}_{t-1}(s), \max_s \widehat{V}_{t-1}(s)\right]} \left\{ \widehat{P}^0_{s,a}\left[\widehat{V}_{t-1}\right]_\alpha - \sqrt{\sigma \mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}_{t-1}\right]_\alpha\right)}\right\}. \tag{21}$$

The proofs of Lemma 1 and Lemma 2 are provided in Appendix A. To complete the description, we output the greedy policy of the final Q-estimate $\widehat{Q}_T$ as the final policy $\widehat{\pi}$, namely,

$$\forall s \in \mathcal{S}: \quad \widehat{\pi}(s) = \arg\max_a \widehat{Q}_T(s,a). \tag{22}$$

Encouragingly, the iterates $\left\{\widehat{Q}_t\right\}_{t\geq 0}$ of DRVI converge linearly to the fixed point $\widehat{Q}^{\star,\sigma}$, owing to the appealing γ-contraction property of $\widehat{\mathcal{T}}^\sigma$.

4 Theoretical guarantees: sample complexity analyses

We now present our main results, which concern the sample complexities of learning RMDPs when the uncertainty set is specified using the TV distance or the χ^2 divergence. Somewhat surprisingly, different choices of the uncertainty set can lead to dramatically different consequences in the sample size requirement.

4.1 The case of TV distance: RMDPs are easier to learn than standard MDPs

We start with the case where the uncertainty set is measured via the TV distance. The following theorem, whose proof is deferred to Section 5.2, develops an upper bound on the sample complexity of DRVI in order to return an ε-optimal robust policy. The key challenge of the analysis lies in careful control of the robust value function $V^{\pi,\sigma}$ as a function of the uncertainty level σ.

Theorem 1 (Upper bound under TV distance)**.** *Let the uncertainty set be $\mathcal{U}_\rho^\sigma(\cdot) = \mathcal{U}_{\mathsf{TV}}^\sigma(\cdot)$, as specified by the TV distance (9). Consider any discount factor $\gamma \in \big[\frac{1}{4}, 1\big)$, uncertainty level $\sigma \in (0,1)$, and $\delta \in (0,1)$. Let $\widehat{\pi}$ be the output policy of Algorithm 1 after $T = C_1 \log\big(\frac{N}{1-\gamma}\big)$ iterations. Then with probability at least $1-\delta$, one has*

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \leq \varepsilon \tag{23}$$

for any $\varepsilon \in \Big(0, \sqrt{1/\max\{1-\gamma,\sigma\}}\Big]$, as long as the total number of samples obeys

$$NSA \geq \frac{C_2 SA}{(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2} \log\left(\frac{SAN}{(1-\gamma)\delta}\right). \tag{24}$$

Here, $C_1, C_2 > 0$ are some large enough universal constants.

Remark 1. Note that Theorem 1 is not only valid when invoking Algorithm 1. In fact, the theorem holds for any oracle planning algorithm (designed based on the empirical transitions $\widehat{P}^0$) whose output policy $\widehat{\pi}$ obeys

$$\big\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\big\|_\infty \leq O\left(\frac{(1-\gamma)^2}{N}\log\left(\frac{SAN}{(1-\gamma)\delta}\right)\right). \tag{25}$$

Before discussing the implications of Theorem 1, we present a matching minimax lower bound that confirms the tightness and optimality of the upper bound, which in turn pins down the sample complexity requirement for learning RMDPs with TV distance. The proof is based on constructing new hard instances inspired by the asymmetric structure of RMDPs, with the details postponed to Section 5.3.

Theorem 2 (Lower bound under TV distance)**.** *Consider any tuple $(S, A, \gamma, \sigma, \varepsilon)$ obeying $\sigma \in (0, 1-c_0]$ with $0 < c_0 \leq \frac{1}{8}$ being any small enough positive constant, $\gamma \in \big[\frac{1}{2}, 1\big)$, and $\varepsilon \in \big(0, \frac{c_0}{256(1-\gamma)}\big]$. We can construct a collection of infinite-horizon RMDPs $\mathcal{M}_0, \mathcal{M}_1$ defined by the uncertainty set $\mathcal{U}_\rho^\sigma(\cdot) = \mathcal{U}_{\mathsf{TV}}^\sigma(\cdot)$, an initial state distribution φ, and a dataset with N independent samples for each state-action pair over the nominal transition kernel (for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively), such that*

$$\inf_{\widehat{\pi}} \max\Big\{\mathbb{P}_0\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big),\, \mathbb{P}_1\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big)\Big\} \geq \frac{1}{8},$$

provided that

$$NSA \leq \frac{c_0 SA \log 2}{8192(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2}.$$

Here, the infimum is taken over all estimators $\widehat{\pi}$, and $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) denotes the probability when the RMDP is $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).

Below, we interpret the above theorems and highlight several key implications about the sample complexity requirements for learning RMDPs for the case w.r.t. the TV distance.

Near minimax-optimal sample complexity. Theorem 1 shows that the total number of samples required for DRVI (or any oracle planning algorithm claimed in Remark 1) to yield ε-accuracy is

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}\right). \tag{26}$$

Taken together with the minimax lower bound asserted by Theorem 2, this confirms the near optimality of the sample complexity (up to some logarithmic factor) almost over the full range of the uncertainty level σ. Importantly, this sample complexity scales linearly with the size of the state-action space, and is inversely proportional to σ in the regime where $\sigma \gtrsim 1-\gamma$.

RMDPs is easier than standard MDPs with TV distance. Recall that the sample complexity requirement for learning standard MDPs with a generative model is (Agarwal et al., 2020; Azar et al., 2013a; Li et al., 2023b)

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right) \tag{27}$$

in order to yield ε accuracy. Comparing this with the sample complexity requirement in (26) for RMDPs under the TV distance, we confirm that the latter is at least as easy as — if not easier than — standard MDPs. In particular, when $\sigma \lesssim 1-\gamma$ is small, the sample complexity of RMDPs is the same as that of standard MDPs as in (27), which is as anticipated since the RMDP reduces to the standard MDP when $\sigma = 0$. On the other hand, when $1-\gamma \lesssim \sigma < 1$, the sample complexity of RMDPs simplifies to

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}\right), \tag{28}$$

which is smaller than that of standard MDPs by a factor of $\sigma/(1-\gamma)$.

Comparison with state-of-the-art bounds. For the upper bound, our results (cf. Theorem 1) significantly improves over the prior art $\widetilde{O}\left(\frac{S^2A}{(1-\gamma)^4\varepsilon^2}\right)$ of Panaganti and Kalathil (2022) by at least a factor of $\frac{S}{1-\gamma}$ and even $\frac{S}{(1-\gamma)^2}$ when the uncertainty level $1-\gamma \lesssim \sigma < 1$ is large. Turning to the lower bound side, Yang et al. (2022) developed a lower bound for RMDPs under the TV distance, which scales as

$$\widetilde{O}\left(\frac{SA(1-\gamma)}{\varepsilon^2}\min\left\{\frac{1}{(1-\gamma)^4},\frac{1}{\sigma^4}\right\}\right).$$

Clearly, this is worse than ours by a factor of $\frac{\sigma^3}{(1-\gamma)^3} \in \left(1, \frac{1}{(1-\gamma)^3}\right)$ in the regime where $1-\gamma \lesssim \sigma < 1$.

4.2 The case of χ^2 divergence: RMDPs can be harder than standard MDPs

We now switch attention to the case when the uncertainty set is measured via the χ^2 divergence. The theorem below presents an upper bound on the sample complexity for this case, whose proof is deferred to Appendix D.

Theorem 3 (Upper bound under χ^2 divergence)**.** *Let the uncertainty set be $\mathcal{U}_\rho^\sigma(\cdot) = \mathcal{U}_{\chi^2}^\sigma(\cdot)$, as specified using the χ^2 divergence (10). Consider any uncertainty level $\sigma \in (0,\infty)$, $\gamma \in [1/4, 1)$ and $\delta \in (0,1)$. With probability at least $1-\delta$, the output policy $\widehat{\pi}$ from Algorithm 1 with at most $T = c_1 \log\left(\frac{N}{1-\gamma}\right)$ iterations yields*

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \le \varepsilon \tag{29}$$

for any $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$, as long as the total number of samples obeying

$$NSA \ge \frac{c_2 SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\log\left(\frac{SAN}{\delta}\right). \tag{30}$$

Here, $c_1, c_2 > 0$ are some large enough universal constants.

***Remark* 2.** Akin to Remark 1, the sample complexity derived in Theorem 3 continues to hold for any oracle planning algorithm that outputs a policy $\widehat{\pi}$ obeying $\big\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\big\|_\infty \leq O\Big(\frac{\log(\frac{SAN}{(1-\gamma)\delta})}{N^2}\Big)$.

In addition, in order to gauge the tightness of Theorem 3 and understand the minimal sample complexity requirement under the χ^2 divergence, we further develop a minimax lower bound as follows; the proof is deferred to Appendix E.

Theorem 4 (Lower bound under χ^2 divergence)**.** *Consider any $(S, A, \gamma, \sigma, \varepsilon)$ obeying $\gamma \in [\frac{3}{4}, 1)$, $\sigma \in (0, \infty)$, and*

$$\varepsilon \leq c_3 \begin{cases} \frac{1}{1-\gamma} & \text{if } \sigma \in \big(0, \frac{1-\gamma}{4}\big) \\ \max\Big\{\frac{1}{(1+\sigma)(1-\gamma)}, 1\Big\} & \text{if } \sigma \in \big[\frac{1-\gamma}{4}, \infty\big) \end{cases} \tag{31}$$

for some small universal constant $c_3 > 0$. Then we can construct two infinite-horizon RMDPs $\mathcal{M}_0, \mathcal{M}_1$ defined by the uncertainty set $\mathcal{U}_\rho^\sigma(\cdot) = \mathcal{U}_{\chi^2}^\sigma(\cdot)$, an initial state distribution φ, and a dataset with N independent samples per (s, a) pair over the nominal transition kernel (for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively), such that

$$\inf_{\widehat{\pi}} \max\Big\{\mathbb{P}_0\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big), \, \mathbb{P}_1\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big)\Big\} \geq \frac{1}{8}, \tag{32}$$

provided that the total number of samples

$$NSA \leq c_4 \begin{cases} \frac{SA}{(1-\gamma)^3\varepsilon^2} & \text{if } \sigma \in \big(0, \frac{1-\gamma}{4}\big) \\ \frac{\sigma SA}{\min\{1, (1-\gamma)^4(1+\sigma)^4\}\varepsilon^2} & \text{if } \sigma \in \big[\frac{1-\gamma}{4}, \infty\big) \end{cases} \tag{33}$$

for some universal constant $c_4 > 0$.

We are now positioned to single out several key implications of the above theorems.

Nearly tight sample complexity. In order to achieve ε-accuracy for RMDPs under the χ^2 divergence, Theorem 3 asserts that a total number of samples on the order of

$$\widetilde{O}\left(\frac{SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\right). \tag{34}$$

is sufficient for DRVI (or any other oracle planning algorithm as discussed in Remark 2). Taking this together with the minimax lower bound in Theorem 4 confirms that the sample complexity is near-optimal — up to a polynomial factor of the effective horizon $\frac{1}{1-\gamma}$ — over the entire range of the uncertainty level σ. In particular,

- when $\sigma \asymp 1$, our sample complexity $\widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right)$ is sharp and matches the minimax lower bound;
- when $\sigma \gtrsim \frac{1}{(1-\gamma)^3}$, our sample complexity correctly predicts the linear dependency with σ, suggesting that more samples are needed when one wishes to account for a larger χ^2-based uncertainty sets.

RMDPs can be much harder to learn than standard MDPs with χ^2 divergence. The minimax lower bound developed in Theorem 4 exhibits a curious non-monotonic behavior of the sample size requirement over the entire range of the uncertainty level $\sigma \in (0, \infty)$ when the uncertainty set is measured via the χ^2 divergence. When $\sigma \lesssim 1 - \gamma$, the lower bound reduces to

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right),$$

which matches with that of standard MDPs, as $\sigma = 0$ corresponds to standard MDP. However, two additional regimes are worth calling out:

$$1 - \gamma \lesssim \sigma \lesssim \frac{1}{(1-\gamma)^{1/3}}: \qquad \widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\min\left\{\sigma, \frac{1}{\sigma^3}\right\}\right),$$

$$\sigma \gtrsim \frac{1}{(1-\gamma)^3}: \qquad \widetilde{O}\left(\frac{SA\sigma}{\varepsilon^2}\right),$$

both of which are *greater* than that of standard MDPs, indicating learning RMDPs under the χ^2 divergence can be much harder.

Comparison with state-of-the-art bounds. Our upper bound significantly improves over the prior art $\widetilde{O}\left(\frac{S^2A(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\right)$ of Panaganti and Kalathil (2022) by a factor of S, and provides the *first* finite-sample complexity that scales *linearly* with respect to S for discounted infinite-horizon RMDPs, which typically exhibit more complicated statistical dependencies than the finite-horizon counterpart. On the other hand, Yang et al. (2022) established a lower bound on the order of $\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}\right)$ when $\sigma \gtrsim 1-\gamma$, which is always smaller than the requirement of standard MDPs, and diminishes when σ grows. Consequently, Yang et al. (2022) does not lead to the rigorous justification that RMDPs can be much harder than standard MDPs, nor the correct linear scaling of the sample size as σ grows.

5 Analysis: the TV case

This section presents the key technical steps for proving our main results of the TV case.

5.1 Preliminaries of the analysis

5.1.1 Additional notations and basic facts

For convenience, we introduce the notation $[T] \coloneqq \{1, \cdots, T\}$ for any positive integer $T > 0$. Moreover, for any two vectors $x = [x_i]_{1\le i\le n}$ and $y = [y_i]_{1\le i\le n}$, the notation $x \le y$ (resp. $x \ge y$) means $x_i \le y_i$ (resp. $x_i \ge y_i$) for all $1 \le i \le n$. And for any vecvor x, we overload the notation by letting $x \circ x = \big[x(s,a)^2\big]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ (resp. $x \circ x = \big[x(s)^2\big]_{s\in\mathcal{S}}$). With slight abuse of notation, we denote 0 (resp. 1) as the all-zero (resp. all-one) vector, and drop the subscript ρ to write $\mathcal{U}^\sigma(\cdot) = \mathcal{U}^\sigma_\rho(\cdot)$ whenever the argument holds for all divergence ρ.

Matrix notation. To continue, we recall or introduce some additional matrix notation that is useful throughout the analysis.

- $P^0 \in \mathbb{R}^{SA\times S}$: the matrix of the nominal transition kernel with $P^0_{s,a}$ as the (s,a)-th row.
- $\widehat{P}^0 \in \mathbb{R}^{SA\times S}$: the matrix of the estimated nomimal transition kernel with $\widehat{P}^0_{s,a}$ as the (s,a)-th row.
- $r \in \mathbb{R}^{SA}$: a vector representing the reward function r (so that $r_{(s,a)} = r(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$).
- $\Pi^\pi \in \{0,1\}^{S\times SA}$: a projection matrix associated with a given deterministic policy π taking the following form

$$\Pi^\pi = \begin{pmatrix} e_{\pi(1)}^\top & 0^\top & \cdots & 0^\top \\ 0^\top & e_{\pi(2)}^\top & \cdots & 0^\top \\ \vdots & \vdots & \ddots & \vdots \\ 0^\top & 0^\top & \cdots & e_{\pi(S)}^\top \end{pmatrix}, \tag{35}$$

  where $e_{\pi(1)}^\top, e_{\pi(2)}^\top, \dots, e_{\pi(S)}^\top \in \mathbb{R}^A$ are standard basis vectors.
- $r_\pi \in \mathbb{R}^S$: a reward vector restricted to the actions chosen by the policy π, namely, $r_\pi(s) = r(s,\pi(s))$ for all $s \in \mathcal{S}$ (or simply, $r_\pi = \Pi^\pi r$).
- $\mathrm{Var}_P(V) \in \mathbb{R}^{SA}$: for any transition kernel $P \in \mathbb{R}^{SA\times S}$ and vector $V \in \mathbb{R}^S$, we denote the (s,a)-th row of $\mathrm{Var}_P(V)$ as

$$\mathsf{Var}_P(s,a) \coloneqq \mathrm{Var}_{P_{s,a}}(V). \tag{36}$$

- $P^V \in \mathbb{R}^{SA\times S}$, $\widehat{P}^V \in \mathbb{R}^{SA\times S}$: the matrices representing the probability transition kernel in the uncertainty set that leads to the worst-case value for any vector $V \in \mathbb{R}^S$. We denote $P^V_{s,a}$ (resp. $\widehat{P}^V_{s,a}$) as the (s,a)-th row of the transition matrix P^V (resp. $\widehat{P}^V$). In truth, the (s,a)-th rows of these transition matrices are defined as

$$P^V_{s,a} = \mathrm{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}V, \qquad \text{and} \qquad \widehat{P}^V_{s,a} = \mathrm{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}V. \tag{37a}$$

Furthermore, we make use of the following short-hand notation:

$$P^{\pi,V}_{s,a} := P^{V^{\pi,\sigma}}_{s,a} = \mathrm{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}V^{\pi,\sigma}, \qquad P^{\pi,\widehat{V}}_{s,a} := P^{\widehat{V}^{\pi,\sigma}}_{s,a} = \mathrm{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widehat{V}^{\pi,\sigma}, \tag{37b}$$

$$\widehat{P}^{\pi,V}_{s,a} := \widehat{P}^{V^{\pi,\sigma}}_{s,a} = \mathrm{argmin}_{P\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}PV^{\pi,\sigma}, \qquad \widehat{P}^{\pi,\widehat{V}}_{s,a} := \widehat{P}^{\widehat{V}^{\pi,\sigma}}_{s,a} = \mathrm{argmin}_{P\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}P\widehat{V}^{\pi,\sigma}. \tag{37c}$$

The corresponding probability transition matrices are denoted by $P^{\pi,V} \in \mathbb{R}^{SA\times S}$, $P^{\pi,\widehat{V}} \in \mathbb{R}^{SA\times S}$, $\widehat{P}^{\pi,V} \in \mathbb{R}^{SA\times S}$ and $\widehat{P}^{\pi,\widehat{V}} \in \mathbb{R}^{SA\times S}$, respectively.

- $P^\pi \in \mathbb{R}^{S\times S}$, $\widehat{P}^\pi \in \mathbb{R}^{S\times S}$, $\underline{P}^{\pi,V} \in \mathbb{R}^{S\times S}$, $\underline{P}^{\pi,\widehat{V}} \in \mathbb{R}^{S\times S}$, $\underline{\widehat{P}}^{\pi,V} \in \mathbb{R}^{S\times S}$ and $\underline{\widehat{P}}^{\pi,\widehat{V}} \in \mathbb{R}^{S\times S}$: six *square* probability transition matrices w.r.t. policy π over the states, namely

$$\begin{aligned} &P^\pi := \Pi^\pi P^0, \qquad \widehat{P}^\pi := \Pi^\pi \widehat{P}^0, \qquad \underline{P}^{\pi,V} := \Pi^\pi P^{\pi,V}, \qquad \underline{P}^{\pi,\widehat{V}} := \Pi^\pi P^{\pi,\widehat{V}}, \\ &\underline{\widehat{P}}^{\pi,V} := \Pi^\pi \widehat{P}^{\pi,V}, \qquad \text{and} \qquad \underline{\widehat{P}}^{\pi,\widehat{V}} := \Pi^\pi \widehat{P}^{\pi,\widehat{V}}. \end{aligned} \tag{38}$$

We denote P^π_s as the s-th row of the transition matrix P^π; similar quantities can be defined for the other matrices as well.

Kullback-Leibler (KL) divergence. First, for any two distributions P and Q, we denote by $\mathsf{KL}(P \parallel Q)$ the Kullback-Leibler (KL) divergence of P and Q. Letting $\mathsf{Ber}(p)$ be the Bernoulli distribution with mean p, we also introduce

$$\mathsf{KL}(p \parallel q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \quad \text{and} \quad \chi^2(p \parallel q) := \frac{(p-q)^2}{q} + \frac{(p-q)^2}{1-q} = \frac{(p-q)^2}{q(1-q)}, \tag{39}$$

which represent respectively the KL divergence and the χ^2 divergence of $\mathsf{Ber}(p)$ from $\mathsf{Ber}(q)$ (Tsybakov, 2009).

Variance. For any probability vector $P \in \mathbb{R}^{1\times S}$ and vector $V \in \mathbb{R}^S$, we denote the variance

$$\mathrm{Var}_P(V) := P(V \circ V) - (PV) \circ (PV). \tag{40}$$

The following lemma bounds the Lipschitz constant of the variance function.

Lemma 3. *Consider any* $0 \le V_1, V_2 \le \frac{1}{1-\gamma}$ *obeying* $\|V_1 - V_2\|_\infty \le x$ *and any probability vector* $P \in \Delta(S)$*, one has*

$$|\mathrm{Var}_P(V_1) - \mathrm{Var}_P(V_2)| \le \frac{2x}{(1-\gamma)}. \tag{41}$$

Proof of Lemma 3: It is immediate to check that

$$\begin{aligned} |\mathrm{Var}_P(V_1) - \mathrm{Var}_P(V_2)| &= |P(V_1 \circ V_1) - (PV_1) \circ (PV_1) - P(V_2 \circ V_2) + (PV_2) \circ (PV_2)| \\ &\le \big|P\big(V_1 \circ V_1 - V_2 \circ V_2\big)\big| + |(PV_1 + PV_2)P(V_1 - V_2)| \\ &\le 2\|V_1 + V_2\|_\infty \|V_1 - V_2\|_\infty \le \frac{2x}{(1-\gamma)}. \end{aligned} \tag{42}$$

where the penultimate inequality holds by the triangle inequality.

5.1.2 Facts of the robust Bellman operator and the empirical robust MDP

γ-contraction of the robust Bellman operator. It is worth noting that the robust Bellman operator (cf. (8)) shares the nice γ-contraction property of the standard Bellman operator, stated as below.

Lemma 4 (γ-Contraction)**.** *(Iyengar, 2005, Theorem 3.2) For any $\gamma \in [0,1)$, the robust Bellman operator $\mathcal{T}^\sigma(\cdot)$ (cf. (8)) is a γ-contraction w.r.t. $\|\cdot\|_\infty$. Namely, for any $Q_1, Q_2 \in \mathbb{R}^{SA}$ s.t. $Q_1(s,a), Q_2(s,a) \in \big[0, \frac{1}{1-\gamma}\big]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, one has*

$$\|\mathcal{T}^\sigma(Q_1) - \mathcal{T}^\sigma(Q_2)\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty . \tag{43}$$

Additionally, $Q^{\star,\sigma}$ is the unique fixed point of $\mathcal{T}^\sigma(\cdot)$ obeying $0 \le Q^{\star,\sigma}(s,a) \le \frac{1}{1-\gamma}$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

Bellman equations of the empirical robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}$. To begin with, recall that the empirical robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^\sigma(\widehat{P}^0), r\}$ based on the estimated nominal distribution $\widehat{P}^0$ constructed in (13) and its corresponding robust value function (resp. robust Q-function) $\widehat{V}^{\pi,\sigma}$ (resp. $\widehat{Q}^{\pi,\sigma}$).

Note that $\widehat{Q}^{\star,\sigma}$ is the unique fixed point of $\widehat{\mathcal{T}}^\sigma(\cdot)$ (see Lemma 4), the empirical robust Bellman operator constructed using $\widehat{P}^0$. Moreover, similar to (7), for $\widehat{\mathcal{M}}_{\mathsf{rob}}$, the Bellman's optimality principle gives the following *robust Bellman consistency equation* (resp. *robust Bellman optimality equation*):

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad \widehat{Q}^{\pi,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\pi,\sigma}, \tag{44a}$$

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \quad \widehat{Q}^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}. \tag{44b}$$

With these in mind, combined with the matrix notation (introduced at the beginning of Section 5), for any policy π, we can write the robust Bellman consistency equations as

$$Q^{\pi,\sigma} = r + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0)} \mathcal{P}V^{\pi,\sigma} \quad \text{and} \quad \widehat{Q}^{\pi,\sigma} = r + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0)} \mathcal{P}\widehat{V}^{\pi,\sigma}, \tag{45}$$

which leads to

$$\begin{aligned} V^{\pi,\sigma} &= r_\pi + \gamma\Pi^\pi \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0)} \mathcal{P}V^{\pi,\sigma} \overset{\text{(i)}}{=} r_\pi + \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}, \\ \widehat{V}^{\pi,\sigma} &= r_\pi + \gamma\Pi^\pi \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0)} \mathcal{P}\widehat{V}^{\pi,\sigma} \overset{\text{(ii)}}{=} r_\pi + \gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}, \end{aligned} \tag{46}$$

where (i) and (ii) holds by the definitions in (35), (37) and (38).

Encouragingly, the above property of the robust Bellman operator ensures the fast convergence of DRVI. We collect this consequence in the following lemma, whose proof is postponed to Appendix A.2.

Lemma 5. *Let $\widehat{Q}_0 = 0$. The iterates $\{\widehat{Q}_t\}, \{\widehat{V}_t\}$ of* DRVI *(cf. Algorithm 1) obey*

$$\forall t \ge 0: \quad \big\|\widehat{Q}_t - \widehat{Q}^{\star,\sigma}\big\|_\infty \le \frac{\gamma^t}{1-\gamma} \quad \text{and} \quad \big\|\widehat{V}_t - \widehat{V}^{\star,\sigma}\big\|_\infty \le \frac{\gamma^t}{1-\gamma}. \tag{47}$$

Furthermore, the output policy $\widehat{\pi}$ obeys

$$\big\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\big\|_\infty \le \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}, \qquad \text{where} \quad \big\|\widehat{V}^{\star,\sigma} - \widehat{V}_{T-1}\big\|_\infty =: \varepsilon_{\mathsf{opt}}. \tag{48}$$

5.2 Proof of the upper bound with TV distance: Theorem 1

Throughout this section, for any transition kernel P, the uncertainty set is taken as (see (9))

$$\mathcal{U}^\sigma(P) \coloneqq \mathcal{U}^\sigma_{\mathsf{TV}}(P) = \otimes\, \mathcal{U}^\sigma_{\mathsf{TV}}(P_{s,a}), \qquad \mathcal{U}^\sigma_{\mathsf{TV}}(P_{s,a}) \coloneqq \Big\{P'_{s,a} \in \Delta(\mathcal{S}) : \frac{1}{2}\big\|P'_{s,a} - P_{s,a}\big\|_1 \le \sigma\Big\}. \tag{49}$$

5.2.1 Technical lemmas

We begin with a key lemma that is new and distinguishes robust MDPs with TV distance from standard MDPs , which plays a critical role in obtaining the sample complexity upper bound in Theorem 1. This lemma concerns the dynamic range of the robust value function $V^{\pi,\sigma}$ (cf. (5)) for any fixed policy π, which produces tighter control than that in standard MDP (cf. $\frac{1}{1-\gamma}$) when σ is large. This lemma The proof is deferred to Appendix B.1.

Lemma 6. *For any nominal transition kernel $P \in \mathbb{R}^{SA\times S}$, any fixed uncertainty level σ, and any policy π, its corresponding robust value function $V^{\pi,\sigma}$ (cf. (5)) satisfies*

$$\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) - \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \le \frac{1}{\gamma \max\{1-\gamma,\sigma\}}.$$

With the above lemma in hand, we introduce the following lemma that is useful throughout this section, whose proof is postponed in Appendix B.2.

Lemma 7. *Consider an MDP with transition kernel matrix P and reward function $0 \le r \le 1$. For any policy π and its associated state transition matrix $P_\pi := \Pi^\pi P$ and value function $0 \le V^{\pi,P} \le \frac{1}{1-\gamma}$ (cf. (1)), one has*

$$(I-\gamma P_\pi)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \le \sqrt{\frac{8(\max_s V^{\pi,P}(s) - \min_s V^{\pi,P}(s))}{\gamma^2(1-\gamma)^2}}1.$$

5.2.2 Proof of Theorem 1

Recall that the proof for standard RL (Agarwal et al., 2020; Li et al., 2023b) deals with the upper and lower bound of the value function estimate gap identically. In contrast, the proof of Theorem 1 needs tailored argument for the robust RL setting — controlling the upper and lower bound of the value function estimate gap in an *asymmetric way* — motivated by the varying worst-case transition kernels associated with different value functions. Before proceeding, applying Lemma 5 yields that for any $\varepsilon_{\mathsf{opt}} > 0$, as long as $T \ge \log(\frac{1}{(1-\gamma)\varepsilon_{\mathsf{opt}}})$, one has

$$\big\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\big\|_\infty \le \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}, \tag{50}$$

allowing us to justify the more general statement in Remark 1. To control the performance gap $\big\|V^{\star,\sigma} - V^{\widehat{\pi},\sigma}\big\|_\infty$, the proof is divided into several key steps.

Step 1: decomposing the error. Recall the optimal robust policy $\pi^\star$ w.r.t. $\mathcal{M}_{\mathsf{rob}}$ and the optimal robust policy $\widehat{\pi}^\star$, the optimal robust value function $\widehat{V}^{\star,\sigma}$ (resp. robust value function $\widehat{Q}^{\pi,\sigma}$) w.r.t. $\widehat{\mathcal{M}}_{\mathsf{rob}}$. The term of interest $V^{\star,\sigma} - V^{\widehat{\pi},\sigma}$ can be decomposed as

$$\begin{aligned}
V^{\star,\sigma} - V^{\widehat{\pi},\sigma} &= \Big(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\Big) + \Big(\widehat{V}^{\pi^\star,\sigma} - \widehat{V}^{\widehat{\pi}^\star,\sigma}\Big) + \Big(\widehat{V}^{\widehat{\pi}^\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\Big) + \Big(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\Big) \\
&\overset{\text{(i)}}{\le} \Big(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\Big) + \Big(\widehat{V}^{\widehat{\pi}^\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\Big) + \Big(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\Big) \\
&\overset{\text{(ii)}}{\le} \Big(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\Big) + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1 + \Big(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\Big)
\end{aligned} \tag{51}$$

where (i) holds by $\widehat{V}^{\pi^\star,\sigma} - \widehat{V}^{\widehat{\pi}^\star,\sigma} \le 0$ since $\widehat{\pi}^\star$ is the robust optimal policy for $\widehat{\mathcal{M}}_{\mathsf{rob}}$, and (ii) comes from the fact in (50).

To control the two important terms in (51), we first consider a more general term $\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma}$ for any policy π. Towards this, plugging in (46) yields

$$\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma} = r_\pi + \gamma \underline{\widehat{P}}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma} - \big(r_\pi + \gamma \underline{P}^{\pi,V}V^{\pi,\sigma}\big)$$

$$
\begin{aligned}
&= \left(\gamma \widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) + \left(\gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \\
&\overset{\text{(i)}}{\leq} \gamma\left(\underline{P}^{\pi,V} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right) + \left(\gamma \widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right),
\end{aligned}
$$

where (i) holds by observing

$$
\underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} \leq \underline{P}^{\pi,V} \widehat{V}^{\pi,\sigma}
$$

due to the optimality of $\underline{P}^{\pi,\widehat{V}}$ (cf. (37)). Rearranging terms leads to

$$
\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma} \leq \gamma \left(I - \gamma \underline{P}^{\pi,V}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right). \tag{52}
$$

Similarly, we can also deduce

$$
\begin{aligned}
\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma} &= r_\pi + \gamma \widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \left(r_\pi + \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \\
&= \left(\gamma \widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) + \left(\gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \\
&\geq \gamma \left(\underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} V^{\pi,\sigma}\right) + \left(\gamma \widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) \\
&\geq \gamma \left(I - \gamma \underline{P}^{\pi,\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right).
\end{aligned} \tag{53}
$$

Combining (52) and (53), we arrive at

$$
\begin{aligned}
\left\|\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma}\right\|_\infty \leq \gamma \max\Big\{ &\left\| \left(I - \gamma \underline{P}^{\pi,V}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right)\right\|_\infty, \\
&\left\| \left(I - \gamma \underline{P}^{\pi,\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right)\right\|_\infty \Big\}.
\end{aligned} \tag{54}
$$

By decomposing the error in a symmetric way, we can similarly obtain

$$
\begin{aligned}
\left\|\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma}\right\|_\infty \leq \gamma \max\Big\{ &\left\| \left(I - \gamma \widehat{\underline{P}}^{\pi,V}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right)\right\|_\infty, \\
&\left\| \left(I - \gamma \widehat{\underline{P}}^{\pi,\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right)\right\|_\infty \Big\}.
\end{aligned} \tag{55}
$$

With the above facts in mind, we are ready to control the two terms $\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty$ and $\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty$ in (51) separately. More specifically, taking $\pi = \pi^\star$, applying (55) leads to

$$
\begin{aligned}
\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty \leq \gamma \max\Big\{ &\left\| \left(I - \gamma \widehat{\underline{P}}^{\pi^\star,V}\right)^{-1} \left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right)\right\|_\infty, \\
&\left\| \left(I - \gamma \widehat{\underline{P}}^{\pi^\star,\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right)\right\|_\infty \Big\}.
\end{aligned} \tag{56}
$$

Similarly, taking $\pi = \widehat{\pi}$, applying (54) leads to

$$
\begin{aligned}
\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty \leq \gamma \max\Big\{ &\left\| \left(I - \gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}\right)\right\|_\infty, \\
&\left\| \left(I - \gamma \underline{P}^{\widehat{\pi},V}\right)^{-1} \left(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}\right)\right\|_\infty \Big\}.
\end{aligned} \tag{57}
$$

Step 2: controlling $\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty$ and $\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty$ separately and summing up. First, we introduce the following two lemmas that control the two main terms in (51), respectively. The first lemma controls the value function estimation error associated with the optimal policy $\pi^\star$ induced by the randomness of the generated dataset. The proof are postponed to Appendix B.3 and B.4.

Lemma 8. *Consider any $\delta \in (0,1)$. With probability at least $1-\delta$, taking $N \geq \frac{16\log(\frac{SAN}{\delta})}{(1-\gamma)^2}$, one has*

$$\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty \leq 160\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} + \frac{8\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}. \tag{58}$$

Unlike the term $\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\|_\infty$ associated with the fixed policy $\pi^\star$ (independent from the dataset), to control $\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty$, we need to deal with the additional complicated statistical dependency between the learned policy $\widehat{\pi}$ and the empirical RMDP constructed by the dataset.

Lemma 9. *Taking $\varepsilon_{\mathsf{opt}} \leq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}$ and $N \geq \frac{16\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2}$, with probability at least $1-\delta$, one has*

$$\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty \leq 24\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} + \frac{28\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}. \tag{59}$$

Summing up the results in (58) and (59) and inserting back to (51) complete the proof as follows: taking $\varepsilon_{\mathsf{opt}} \leq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}$ and $N \geq \frac{16\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2}$, with probability at least $1-\delta$,

$$\begin{aligned}
\left\|V^{\star,\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty &\leq \left\|V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\right\|_\infty + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} + \left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty \\
&\leq \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} + 160\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} + \frac{8\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N} \\
&\quad + 24\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} + \frac{28\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2} \\
&\leq 184\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} + \frac{36\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2} \\
&\leq 1508\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}},
\end{aligned} \tag{60}$$

where the last inequality holds by $\gamma \geq \frac{1}{4}$ and $N \geq \frac{16\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2}$.

5.3 Proof of the lower bound with TV distance: Theorem 2

To achieve a tight lower bound for robust MDPs, we construct new hard instances that are different from those for standard MDPs (Azar et al., 2013a), addressing two new challenges: 1) Due to robustness requirement, the recursive step (or bootstrapping) in robust MDPs has asymmetric structures over all states, since the worst-case transition probability depends on the value function and puts more weights on the states with lower values. Inspired by such an asymmetric structure, we develop new hard instances by setting larger rewards on the states with action-invariant transition kernels to achieve a tighter lower bound. Note that standard MDPs do not have such reward allocation challenges, where the bootstrapping step is determined by a fixed transition probability independent from the value function. 2) As the uncertainty level can vary within $0 < \sigma \leq 1$, for given any fixed uncertainty level σ, a tailored σ-dependent hard instance is required to achieve a tight lower bound, leading to the construction of a series of different instances as σ varies. Instead, standard RL only needs to construct one hard instance (i.e., $\sigma = 0$). By constructing a new class of hard instances addressing the above challenges, we develop a new lower bound in Theorem 2 that is tighter than prior art (Yang et al., 2022), which used an identical hard instance for all uncertainty levels $0 < \sigma \leq 1$.

5.3.1 Construction of the hard problem instances

Construction of two hard MDPs. Suppose there are two standard MDPs defined as below:

$$\left\{\mathcal{M}_\phi = \left(\mathcal{S}, \mathcal{A}, P^\phi, r, \gamma\right) \mid \phi = \{0,1\}\right\}.$$

Here, γ is the discount parameter, $\mathcal{S} = \{0, 1, \ldots, S-1\}$ is the state space. Given any state $s \in \{2, 3, \cdots, S-1\}$, the corresponding action space are $\mathcal{A} = \{0, 1, 2, \cdots, A-1\}$. While for states $s = 0$ or $s = 1$, the action space is only $\mathcal{A}' = \{0, 1\}$. For any $\phi \in \{0, 1\}$, the transition kernel P^ϕ of the constructed MDP $\mathcal{M}_\phi$ is defined as

$$P^\phi(s' \,|\, s, a) = \begin{cases} p\mathbb{1}(s' = 1) + (1-p)\mathbb{1}(s' = 0) & \text{if} \quad (s, a) = (0, \phi) \\ q\mathbb{1}(s' = 1) + (1-q)\mathbb{1}(s' = 0) & \text{if} \quad (s, a) = (0, 1-\phi) \\ \mathbb{1}(s' = 1) & \text{if} \quad s \geq 1 \end{cases}, \tag{61}$$

where p and q are set to satisfy

$$0 \leq p \leq 1 \quad \text{and} \quad 0 \leq q = p - \Delta \tag{62}$$

for some p and $\Delta > 0$ that shall be introduced later. The above transition kernel P^ϕ implies that state 1 is an absorbing state, namely, the MDP will always stay after it arrives at 1.

Then, we define the reward function as

$$r(s, a) = \begin{cases} 1 & \text{if } s = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{63}$$

Additionally, we choose the following initial state distribution:

$$\varphi(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{otherwise} \end{cases}. \tag{64}$$

Here, the constructed two instances are set with different probability transition from state 0 with reward 0 but not state 1 with reward 1 (which were used in standard MDPs (Li et al., 2022b)), yielding a larger gap between the value functions of the two instances.

Uncertainty set of the transition kernels. Recalling the uncertainty set assumed throughout this section is defined as $\mathcal{U}^\sigma(P^\phi)$ with TV distance:

$$\mathcal{U}^\sigma(P^\phi) := \mathcal{U}^\sigma_{\mathsf{TV}}(P^\phi) = \otimes\, \mathcal{U}^\sigma_{\mathsf{TV}}(P^\phi_{s,a}), \qquad \mathcal{U}^\sigma_{\mathsf{TV}}(P^\phi_{s,a}) := \left\{ P'_{s,a} \in \Delta(\mathcal{S}) : \frac{1}{2}\left\| P'_{s,a} - P^\phi_{s,a} \right\|_1 \leq \sigma \right\}, \tag{65}$$

where $P^\phi_{s,a} := P^\phi(\cdot \,|\, s, a)$ is defined similar to (4). In addition, without loss of generality, we recall the radius $\sigma \in (0, 1 - c_0]$ with $0 < c_0 < 1$. With the uncertainty level in hand, taking $c_1 := \frac{c_0}{2}$, p and Δ which determines the instances obey

$$p = (1 + c_1) \max\{1 - \gamma, \sigma\} \qquad \text{and} \qquad \Delta \leq c_1 \max\{1 - \gamma, \sigma\}, \tag{66}$$

which ensure $0 \leq p \leq 1$ as follows:

$$(1 + c_1)\,\sigma \leq 1 - c_0 + c_1\sigma \leq 1 - \frac{c_0}{2} < 1, \qquad (1 + c_1)\,(1 - \gamma) \leq \frac{3}{2}(1 - \gamma) \leq \frac{3}{4} < 1. \tag{67}$$

Consequently, applying (62) directly leads to

$$p \geq q \geq \max\{1 - \gamma, \sigma\}. \tag{68}$$

To continue, for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we denote the infimum probability of moving to the next state s' associated with any perturbed transition kernel $P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})$ as

$$\underline{P}^\phi(s' \,|\, s, a) := \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})} P(s' \,|\, s, a) = \max\{P(s' \,|\, s, a) - \sigma, 0\}, \tag{69}$$

where the last equation can be easily verified by the definition of $\mathcal{U}^\sigma(P^\phi)$ in (65). As shall be seen, the transition from state 0 to state 1 plays an important role in the analysis, for convenience, we denote

$$\underline{p} := \underline{P}^\phi(1 \,|\, 0, \phi) = p - \sigma, \qquad \underline{q} := \underline{P}^\phi(1 \,|\, 0, 1 - \phi) = q - \sigma, \tag{70}$$

which follows from the fact that $p \geq q \geq \sigma$ in (68).

Robust value functions and robust optimal policies. To proceed, we are ready to derive the corresponding robust value functions, identify the optimal policies, and characterize the optimal values. For any MDP $\mathcal{M}_\phi$ with the above uncertainty set, we denote $\pi_\phi^\star$ as the optimal policy, and the robust value function of any policy π (resp. the optimal policy $\pi_\phi^\star$) as $V_\phi^{\pi,\sigma}$ (resp. $V_\phi^{\star,\sigma}$). Then, we introduce the following lemma which describes some important properties of the robust (optimal) value functions and optimal policies. The proof is postponed to Appendix C.1.

Lemma 10. *For any $\phi = \{0,1\}$ and any policy π, the robust value function obeys*

$$V_\phi^{\pi,\sigma}(0) = \frac{\gamma\big(z_\phi^\pi - \sigma\big)}{(1-\gamma)\left(1 + \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma(1-\sigma)}\right)\left(1 - \gamma\left(1-\sigma\right)\right)}, \tag{71}$$

where z_ϕ^π is defined as

$$z_\phi^\pi := p\pi(\phi \,|\, 0) + q\pi(1-\phi \,|\, 0). \tag{72}$$

In addition, the robust optimal value functions and the robust optimal policies satisfy

$$V_\phi^{\star,\sigma}(0) = \frac{\gamma\left(p-\sigma\right)}{(1-\gamma)\left(1 + \frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}\right)\left(1-\gamma\left(1-\sigma\right)\right)}, \tag{73a}$$

$$\pi_\phi^\star(\phi \,|\, s) = 1, \qquad \textit{for } s \in \mathcal{S}. \tag{73b}$$

5.3.2 Establishing the minimax lower bound

Note that our goal is to control the quantity w.r.t. any policy estimator $\widehat{\pi}$ based on the chosen initial distribution φ in (64) and the dataset consisting of N samples over each state-action pair generated from the nominal transition kernel P^ϕ, which gives

$$\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \big\rangle = V_\phi^{\star,\sigma}(0) - V_\phi^{\widehat{\pi},\sigma}(0).$$

Step 1: converting the goal to estimate ϕ. We make the following useful claim which shall be verified in Appendix C.2: With $\varepsilon \le \frac{c_1}{32(1-\gamma)}$, letting

$$\Delta = 32(1-\gamma)\max\{1-\gamma,\sigma\}\varepsilon \le c_1 \max\{1-\gamma,\sigma\} \tag{74}$$

which satisfies (66), it leads to that for any policy $\widehat{\pi}$,

$$\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \big\rangle \ge 2\varepsilon\big(1 - \widehat{\pi}(\phi \,|\, 0)\big). \tag{75}$$

With this connection established between the policy $\widehat{\pi}$ and its sub-optimality gap as depicted in (75), we can now proceed to build an estimate for ϕ. Here, we denote $\mathbb{P}_\phi$ as the probability distribution when the MDP is $\mathcal{M}_\phi$, where ϕ can take on values in the set $\{0,1\}$.

Let's assume momentarily that an estimated policy $\widehat{\pi}$ achieves

$$\mathbb{P}_\phi\big\{\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \big\rangle \le \varepsilon\big\} \ge \frac{7}{8}, \tag{76}$$

then in view of (75), we necessarily have $\widehat{\pi}(\phi \,|\, 0) \ge \frac{1}{2}$ with probability at least $\frac{7}{8}$. With this in mind, we are motivated to construct the following estimate $\widehat{\phi}$ for $\phi \in \{0,1\}$:

$$\widehat{\phi} = \arg\max_{a \in \{0,1\}} \widehat{\pi}(a \,|\, 0), \tag{77}$$

which obeys

$$\mathbb{P}_\phi\big\{\widehat{\phi} = \phi\big\} \ge \mathbb{P}_\phi\big\{\widehat{\pi}(\phi \,|\, 0) > 1/2\big\} \ge \frac{7}{8}. \tag{78}$$

Subsequently, our aim is to demonstrate that (78) cannot occur without an adequate number of samples, which would in turn contradict (75).

Step 2: probability of error in testing two hypotheses. Equipped with the aforementioned groundwork, we can now delve into differentiating between the two hypotheses $\phi \in \{0, 1\}$. To achieve this, we consider the concept of minimax probability of error, defined as follows:

$$p_{\mathrm{e}} := \inf_{\psi} \max \big\{ \mathbb{P}_0(\psi \neq 0), \, \mathbb{P}_1(\psi \neq 1) \big\}. \tag{79}$$

Here, the infimum is taken over all possible tests ψ constructed from the samples generated from the nominal transition kernel P^{ϕ}.

Moving forward, let us denote μ_{ϕ} (resp. $\mu_{\phi}(s)$) as the distribution of a sample tuple (s_i, a_i, s_i') under the nominal transition kernel P^{ϕ} associated with $\mathcal{M}_{\phi}$ and the samples are generated independently. Applying standard results from Tsybakov (2009, Theorem 2.2) and the additivity of the KL divergence (cf. Tsybakov (2009, Page 85)), we obtain

$$\begin{aligned} p_{\mathrm{e}} &\geq \frac{1}{4} \exp\Big(- NSA \cdot \mathsf{KL}\big(\mu_0 \parallel \mu_1\big) \Big) \\ &= \frac{1}{4} \exp\Big\{ - N\Big(\mathsf{KL}\big(P^0(\cdot \,|\, 0, 0) \parallel P^1(\cdot \,|\, 0, 0)\big) + \mathsf{KL}\big(P^0(\cdot \,|\, 0, 1) \parallel P^1(\cdot \,|\, 0, 1)\big) \Big) \Big\}, \end{aligned} \tag{80}$$

where the last inequality holds by observing that

$$\begin{aligned} \mathsf{KL}\big(\mu_0 \parallel \mu_1\big) &= \frac{1}{SA} \sum_{s, a, s'} \mathsf{KL}\big(P^0(s' \,|\, s, a) \parallel P^1(s' \,|\, s, a)\big) \\ &= \frac{1}{SA} \sum_{a \in \{0, 1\}} \mathsf{KL}\big(P^0(\cdot \,|\, 0, a) \parallel P^1(\cdot \,|\, 0, a)\big), \end{aligned}$$

Here, the last equality holds by the fact that $P^0(\cdot \,|\, s, a)$ and $P^1(\cdot \,|\, s, a)$ only differ when $s = 0$.

Now, our focus shifts towards bounding the terms involving the KL divergence in (80). Given $p \geq q \geq \max\{1 - \gamma, \sigma\}$ (cf. (68)), applying Tsybakov (2009, Lemma 2.7) gives

$$\begin{aligned} \mathsf{KL}\big(P^0(\cdot \,|\, 0, 1) \parallel P^1(\cdot \,|\, 0, 1)\big) = \mathsf{KL}\left(p \parallel q\right) &\leq \frac{(p - q)^2}{(1 - p)p} \overset{\text{(i)}}{=} \frac{\Delta^2}{p(1 - p)} \\ &\overset{\text{(ii)}}{=} \frac{1024(1 - \gamma)^2 \max\{1 - \gamma, \sigma\}^2 \varepsilon^2}{p(1 - p)} \\ &\leq \frac{1024(1 - \gamma)^2 \max\{1 - \gamma, \sigma\} \varepsilon^2}{1 - p} \leq \frac{4096}{c_1} (1 - \gamma)^2 \max\{1 - \gamma, \sigma\} \varepsilon^2, \end{aligned} \tag{81}$$

where (i) stems from the definition in (62), (ii) follows by the expression of Δ in (74), and the last inequality arises from $1 - q \geq 1 - p \geq \frac{c_0}{4}$ (see (67)).

Note that it can be shown that $\mathsf{KL}\big(P^0(\cdot \,|\, 0, 0) \parallel P^1(\cdot \,|\, 0, 0)\big)$ can be upper bounded in a same manner. Substituting (81) back into (80) demonstrates that: if the sample size is selected as

$$N \leq \frac{c_1 \log 2}{8192(1 - \gamma)^2 \max\{1 - \gamma, \sigma\} \varepsilon^2}, \tag{82}$$

then one necessarily has

$$p_{\mathrm{e}} \geq \frac{1}{4} \exp\left\{ - N \frac{8192}{c_1} (1 - \gamma)^2 \max\{1 - \gamma, \sigma\} \varepsilon^2 \right\} \geq \frac{1}{8}, \tag{83}$$

Step 3: putting the results together. Lastly, suppose that there exists an estimator $\widehat{\pi}$ such that

$$\mathbb{P}_0 \big\{ \big\langle \varphi, V_0^{\star, \sigma} - V_0^{\widehat{\pi}, \sigma} \big\rangle > \varepsilon \big\} < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1 \big\{ \big\langle \varphi, V_1^{\star, \sigma} - V_1^{\widehat{\pi}, \sigma} \big\rangle > \varepsilon \big\} < \frac{1}{8}.$$

According to Step 1, the estimator $\widehat{\phi}$ defined in (77) must satisfy

$$\mathbb{P}_0 \big(\widehat{\phi} \neq 0 \big) < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1 \big(\widehat{\phi} \neq 1 \big) < \frac{1}{8}.$$

However, this cannot occur under the sample size condition (82) to avoid contradiction with (83). Thus, we have completed the proof.

6 Offline distributionally robust RL with uniform coverage

In this section, we extend our theoretical analysis to broader sampling mechanism scenarios with offline datasets. We first specify the offline settings as below.

Offline/batch dataset. Suppose that we observe a batch/historical dataset $\mathcal{D}^{\mathsf{b}} = \{(s_i, a_i, r_i, s_i')\}_{1\leq i\leq N_{\mathsf{b}}}$ consisting of N_{b} sample transitions generated independently. Specifically, the state-action pair (s_i, a_i) is drawn from some behavior distribution $\mu^{\mathsf{b}} \in \Delta(\mathcal{S}\times\mathcal{A})$, followed by a next state s_i' drawn over the nominal transition kernel P^0, i.e.,

$$(s_i, a_i) \overset{\text{i.i.d.}}{\sim} \mu^{\mathsf{b}} \quad \text{and} \quad s_i' \overset{\text{i.i.d.}}{\sim} P^0(\cdot \,|\, s_i, a_i), \qquad 1 \leq i \leq N_{\mathsf{b}}. \tag{84}$$

We consider uniform coverage historical dataset that is widely studied in offline settings for both standard RL and robust RL (Chen and Jiang, 2019; Jin et al., 2020b; Liao et al., 2022; Yang et al., 2022; Zhou et al., 2021), specified in the following assumption.

Assumption 1. *Suppose the historical dataset $\mathcal{D}^{\mathsf{b}}$ obeys*

$$\mu_{\min} \coloneqq \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu^{\mathsf{b}}(s,a) > 0. \tag{85}$$

Armed with the above dataset $\mathcal{D}^{\mathsf{b}}$, the empirical nominal transition kernel $\widehat{P}^0 \in \mathbb{R}^{SA\times S}$ can be constructed through (13) analogously. Then in such offline setting, we introduce the sample complexity upper bounds for DRVI and information-theoretical lower bounds in the cases of TV or χ^2 divergence respectively. The proof of the following corollaries are postponed to Appendix F.

6.1 The case of TV distance

With above historical dataset $\mathcal{D}^{\mathsf{b}}$ in hand, we achieve the following corollary implied by Theorem 1.

Corollary 1 (Upper bound under TV distance). *Let the uncertainty set be $\mathcal{U}_\rho^\sigma(\cdot) = \mathcal{U}_{\mathsf{TV}}^\sigma(\cdot)$ defined in (9), and $C_3, C_4 > 0$ be some large enough universal constants. Consider any discount factor $\gamma \in \left[\frac{1}{4}, 1\right)$, uncertainty level $\sigma \in (0,1)$, and $\delta \in (0,1)$. Let $\widehat{\pi}$ be the output policy of Algorithm 1 after $T = C_3 \log\left(\frac{N_{\mathsf{b}}}{1-\gamma}\right)$ iterations, based on a dataset $\mathcal{D}^{\mathsf{b}}$ satisfying Assumption 1. Then with probability at least $1-\delta$, one has*

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \leq \varepsilon \tag{86}$$

for any $\varepsilon \in \left(0, \sqrt{1/\max\{1-\gamma,\sigma\}}\right]$, as long as the total number of samples obeys

$$N_{\mathsf{b}} \geq \frac{C_4}{\mu_{\min}(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2} \log\left(\frac{N_{\mathsf{b}}SA}{(1-\gamma)\delta}\right). \tag{87}$$

We also derive a lower bound in the offline setting by adapting Theorem 2.

Corollary 2 (Lower bound under TV distance). *Let the uncertainty set be $\mathcal{U}_\rho^\sigma(\cdot) = \mathcal{U}_{\mathsf{TV}}^\sigma(\cdot)$ defined in (9). Consider any tuple $(S, \gamma, \sigma, \varepsilon, \mu_{\min})$ that obeys $\mu_{\min} > 0$, $\sigma \in (0, 1-c_0]$ with $0 < c_0 \leq \frac{1}{8}$ being any small enough positive constant, $\gamma \in \left[\frac{1}{2}, 1\right)$, and $\varepsilon \in \left(0, \frac{c_0}{256(1-\gamma)}\right]$. We can construct two infinite-horizon RMDPs $\mathcal{M}_0, \mathcal{M}_1$, an initial state distribution φ, and a dataset with N_{b} samples satisfying Assumption 1 (for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively) such that*

$$\inf_{\widehat{\pi}} \max\left\{\mathbb{P}_0\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big), \, \mathbb{P}_1\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big)\right\} \geq \frac{1}{8},$$

provided that

$$N_{\mathsf{b}} \leq \frac{c_0 \log 2}{8192\mu_{\min}(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2}.$$

Here, the infimum is taken over all estimators $\widehat{\pi}$, and $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) denotes the probability when the RMDP is $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).

Discussions. In the offline setting with uniform coverage dataset (cf. Assumption 1), Corollary 1 shows that DRVI algorithm can find an ε-optimal policy with the following sample complexity

$$\widetilde{O}\left(\frac{1}{\mu_{\min}(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}\right), \tag{88}$$

which is near minimax optimal with respect to all salient parameters (up to logarithmic factors) almost over the full range of the uncertainty level σ, verified by the lower bound in Corollary 2. Our sample complexity upper bound (Corollary 1) significantly improves over the prior art $\widetilde{O}\left(\frac{S(2+\sigma)^2}{\mu_{\min}\sigma^2(1-\gamma)^4\varepsilon^2}\right)$ (Yang et al., 2022) by at least a factor of $\frac{S}{(1-\gamma)^2}$, and even more than $\frac{S}{(1-\gamma)^3}$ when the uncertainty level $0<\sigma\lesssim 1-\gamma$ is small.

6.2 The case of χ^2 divergence

With uncertainty sets measured by the χ^2 divergence, we obtain the following upper bounds for DRVI and information-theoretical lower bounds, adapted from Theorem 3 and Theorem 4 respectively.

Corollary 3 (Upper bound under χ^2 divergence)**.** *Let the uncertainty set be $\mathcal{U}^{\sigma}_{\rho}(\cdot)=\mathcal{U}^{\sigma}_{\chi^2}(\cdot)$ specified by the χ^2 divergence (cf. (10)), and $c_1,c_2>0$ be some large enough universal constants. Consider any uncertainty level $\sigma\in(0,\infty)$, $\gamma\in[1/4,1)$ and $\delta\in(0,1)$. Given a dataset $\mathcal{D}^{\mathsf{b}}$ satisfying Assumption 1, with probability at least $1-\delta$, the output policy $\widehat{\pi}$ from Algorithm 1 with at most $T=c_1\log\left(\frac{N_{\mathsf{b}}}{1-\gamma}\right)$ iterations yields*

$$\forall s\in\mathcal{S}:\quad V^{\star,\sigma}(s)-V^{\widehat{\pi},\sigma}(s)\le\varepsilon \tag{89}$$

for any $\varepsilon\in\left(0,\frac{1}{1-\gamma}\right]$, as long as the total number of samples obeying

$$N_{\mathsf{b}}\ge\frac{c_2(1+\sigma)}{\mu_{\min}(1-\gamma)^4\varepsilon^2}\log\left(\frac{N_{\mathsf{b}}}{\mu_{\min}\delta}\right). \tag{90}$$

Corollary 4 (Lower bound under χ^2 divergence)**.** *Let the uncertainty set be $\mathcal{U}^{\sigma}_{\rho}(\cdot)=\mathcal{U}^{\sigma}_{\chi^2}(\cdot)$, and $c_3,c_4>0$ be some universal constants. Consider any tuple $(S,\gamma,\sigma,\varepsilon,\mu_{\min})$ obeying $\mu_{\min}>0$, $\gamma\in[\frac{3}{4},1)$, $\sigma\in(0,\infty)$, and*

$$\varepsilon\le c_3\begin{cases}\frac{1}{1-\gamma} & \text{if } \sigma\in\left(0,\frac{1-\gamma}{4}\right)\\ \max\left\{\frac{1}{(1+\sigma)(1-\gamma)},1\right\} & \text{if } \sigma\in\left[\frac{1-\gamma}{4},\infty\right).\end{cases} \tag{91}$$

Then we can construct two infinite-horizon RMDPs $\mathcal{M}_0,\mathcal{M}_1$, an initial state distribution φ, and a dataset with N_{b} independent samples satisfying Assumption 1 over the nominal transition kernel (for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively), such that

$$\inf_{\widehat{\pi}}\max\left\{\mathbb{P}_0\big(V^{\star,\sigma}(\varphi)-V^{\widehat{\pi},\sigma}(\varphi)>\varepsilon\big),\,\mathbb{P}_1\big(V^{\star,\sigma}(\varphi)-V^{\widehat{\pi},\sigma}(\varphi)>\varepsilon\big)\right\}\ge\frac{1}{8}, \tag{92}$$

provided that the total number of samples

$$N_{\mathsf{b}}\le c_4\begin{cases}\frac{1}{\mu_{\min}(1-\gamma)^3\varepsilon^2} & \text{if } \sigma\in\left(0,\frac{1-\gamma}{4}\right)\\ \frac{\sigma}{\min\{1,(1-\gamma)^4(1+\sigma)^4\}\mu_{\min}\varepsilon^2} & \text{if } \sigma\in\left[\frac{1-\gamma}{4}.\infty\right)\end{cases} \tag{93}$$

Discussions. Corollary 3 indicates that in the offline setting with uniform coverage dataset (cf. Assumption 1), DRVI can achieve ε-accuracy for RMDPs under the χ^2 divergence with a total number of samples on the order of

$$\widetilde{O}\left(\frac{(1+\sigma)}{\mu_{\min}(1-\gamma)^4\varepsilon^2}\right). \tag{94}$$

The above upper bound is relatively tight, since it matches the lower bound derived in Corollary 4 when the uncertainty level $\sigma\asymp 1$ and correctly captures the linear dependency with σ when the uncertainty level $\sigma\gtrsim\frac{1}{(1-\gamma)^3}$ is large. In addition, it significantly improves upon the prior art $\widetilde{O}\left(\frac{S(1+\sigma)^2}{\mu_{\min}(\sqrt{1+\sigma}-1)^2(1-\gamma)^4\varepsilon^2}\right)$ (Yang et al., 2022) by at least a factor of $S(1+\sigma)$.

7 Other related works

This section briefly discusses a small sample of other related works. We limit our discussions primarily to provable RL algorithms in the tabular setting with finite state and action spaces, which are most related to the current paper.

Finite-sample guarantees for standard RL. A surge of recent research has utilized the toolkit from high-dimensional probability/statistics to investigate the performance of standard RL algorithms in non-asymptotic settings. There has been a considerable amount of research into non-asymptotic sample analysis of standard RL for a variety of settings; partial examples include, but are not limited to, the works via probably approximately correct (PAC) bounds for the generative model setting (Agarwal et al., 2020; Azar et al., 2013b; Beck and Srikant, 2012; Chen et al., 2020; Kearns and Singh, 1999; Li et al., 2023a, 2022a, 2023b; Sidford et al., 2018; Wainwright, 2019) and the offline setting (Chen and Jiang, 2019; Jin et al., 2021; Li et al., 2024; Liao et al., 2022; Rashidinejad et al., 2021; Shi et al., 2022; Uehara et al., 2022; Woo et al., 2024; Xie et al., 2021; Yan et al., 2022; Yin et al., 2021), as well as the online setting via both regret-based and PAC-base analyses (Bai et al., 2019; Dong et al., 2019; Jafarnia-Jahromi et al., 2020; Jin et al., 2018, 2020a; Li et al., 2021, 2023c; Woo et al., 2023; Yang et al., 2021; Zhang et al., 2020b).

Robustness in RL. While standard RL has achieved remarkable success, current RL algorithms still have significant drawbacks in that the learned policy could be completely off if the deployed environment is subject to perturbation, model mismatch, or other structural changes. To address these challenges, an emerging line of works begin to address robustness of RL algorithms with respect to the uncertainty or perturbation over different components of MDPs — state, action, reward, and the transition kernel; see Moos et al. (2022) for a recent review. Besides the framework of distributionally robust MDPs (RMDPs) (Iyengar, 2005) adopted by this work, to promote robustness in RL, there exist various other works including but not limited to Han et al. (2022); Qiaoben et al. (2021); Sun et al. (2021); Xiong et al. (2022); Zhang et al. (2021, 2020a) investigating the robustness w.r.t. state uncertainty, where the agent's policy is chosen based on a perturbed observation generated from the state by adding restricted noise or adversarial attack. Besides, Tan et al. (2020); Tessler et al. (2019) considered the robustness w.r.t. the uncertainty of the action, namely, the action is possibly distorted by an adversarial agent abruptly or smoothly, and Ding et al. (2023) tackles robustness against spurious correlations..

Distributionally robust RL. Rooted in the literature of distributionally robust optimization, which has primarily been investigated in the context of supervised learning (Bertsimas et al., 2018; Blanchet and Murthy, 2019; Duchi and Namkoong, 2018; Gao, 2020; Rahimian and Mehrotra, 2019), distributionally robust dynamic programming and RMDPs have attracted considerable attention recently (Badrinath and Kalathil, 2021; Derman and Mannor, 2020; Goyal and Grand-Clement, 2022; Ho et al., 2018, 2021; Iyengar, 2005; Kaufman and Schaefer, 2013; Smirnova et al., 2019; Tamar et al., 2014; Wolff et al., 2012; Xu and Mannor, 2012). In the context of RMDPs, both empirical and theoretical studies have been widely conducted, although most prior theoretical analyses focus on planning with an exact knowledge of the uncertainty set (Iyengar, 2005; Tamar et al., 2014; Xu and Mannor, 2012), or are asymptotic in nature (Roy et al., 2017).

Resorting to the tools of high-dimensional statistics, various recent works begin to shift attention to understand the finite-sample performance of provable robust RL algorithms, under diverse data generating mechanisms and forms of the uncertainty set over the transition kernel. Besides the infinite-horizon setting, finite-sample complexity bounds for RMDPs with the TV distance and the χ^2 divergence are also developed for the finite-horizon setting in Dong et al. (2022); Lu et al. (2024); Xu et al. (2023). In addition, many other forms of uncertainty sets have been considered. For example, Wang and Zou (2021) considered a R-contamination uncertain set and proposed a provable robust Q-learning algorithm for the online setting with similar guarantees as standard MDPs. The KL divergence is another popular choice widely considered, where Blanchet et al. (2023); Liang et al. (2023); Liu et al. (2022); Panaganti and Kalathil (2022); Shi and Chi (2022); Wang et al. (2023a,b,d); Xu et al. (2023); Yang et al. (2022); Zhou et al. (2021) investigated the sample complexity of both model-based and model-free algorithms under the simulator, offline settings, or single-trajectory setting. Xu et al. (2023) considered a variety of uncertainty sets including one associated with Wasserstein distance. Badrinath and Kalathil (2021); Liu and Xu (2024a,b); Ma et al. (2022); Panaganti et al. (2022); Ramesh et al. (2023); Wang et al. (2024) considered function approximation settings. Moreover, various other related issues have been explored such as the difference of various uncertainty types (Wang et al., 2023c), the iteration complexity of the policy-based methods (Kumar et al., 2023; Li and Lan, 2023; Li et al., 2022c), the case when the uncertainty level is instance-dependent small enough (Clavier et al., 2023), regularization-based robust RL (Yang et al., 2023; Zhang et al., 2023), and distributionally robust optimization for offline RL (Panaganti et al., 2023).

8 Discussions

This work has developed improved sample complexity bounds for learning RMDPs when the uncertainty set is measured via the TV distance or the χ^2 divergence, assuming availability of a generative model. Our results have not only strengthened the prior art in both the upper and lower bounds, but have also unlocked curious insights into how the quest for distributional robustness impacts the sample complexity. As a key takeaway of this paper, RMDPs are not necessarily harder nor easier to learn than standard MDPs, as the answer depends — in a rather subtle manner — on the specific choice of the uncertainty set. For the case w.r.t. the TV distance, we have settled the minimax sample complexity for RMDPs, which is never larger than that required to learn standard MDPs. Regarding the case w.r.t. the χ^2 divergence, we have uncovered that learning RMDPs can oftentimes be provably harder than the standard MDP counterpart. All in all, our findings help raise awareness that the choice of the uncertainty set not only represents a preference in robustness, but also exerts fundamental influences upon the intrinsic statistical complexity.

Moving forward, our work opens up numerous avenues for future studies, and we point out a few below.

- *Extensions to the finite-horizon setting.* It is likely that our current analysis framework can be extended to tackle finite-horizon RMDPs, which would help complete our understanding for the tabular cases.

- *Improved analysis for the case of χ^2 divergence.* While we have settled the sample complexity of RMDPs with the TV distance, the upper and lower bounds we have developed for RMDPs w.r.t. the χ^2 divergence still differ by some polynomial factor in the effective horizon. It would be of great interest to see how to close this gap.

- *A unified theory for other families of uncertainty sets.* Our work raises an interesting question concerning how the geometry of the uncertainty sets intervenes the sample complexity. Characterizing the tight sample complexity for RMDPs under a more general family of uncertainty sets — such as using ℓ_p distance or f-divergence, as well as s-rectangular sets — would be highly desirable.

- *Instance-dependent sample complexity analyses.* We note that we focus on understanding the minimax-optimal sample complexity of RMDPs, which might be rather pessimistic. When consider a given MDP, the feasible and reasonable magnitude of the uncertainty level σ is limited by a certain *instance-dependent finite threshold*. It will be desirable to study instance-dependent sample complexity of RMDPs, which might shed better light on guiding the practice.

Acknowledgement

The work of L. Shi and Y. Chi is supported in part by the grants ONR N00014-19-1-2404, NSF CCF-2106778, DMS-2134080, and CNS-2148212. L. Shi is also gratefully supported by the Leo Finzi Memorial Fellowship, Wei Shen and Xuehong Zhang Presidential Fellowship, and Liang Ji-Dian Graduate Fellowship at Carnegie Mellon University. The work of Y. Wei is supported in part by the NSF grants DMS-2147546/2015447, CAREER award DMS-2143215, CCF-2106778, and the Google Research Scholar Award. The work of Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009 and CCF-1907661. The authors also acknowledge Mengdi Xu, Zuxin Liu and He Wang for valuable discussions.

References

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

Azar, M., Munos, R., and Kappen, H. J. (2013a). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349.

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

Badrinath, K. P. and Kalathil, D. (2021). Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient Q-learning with low switching cost. *arXiv preprint arXiv:1905.12849*.

Bäuerle, N. and Glauner, A. (2022). Distributionally robust markov decision processes and their connection to risk measures. *Mathematics of Operations Research*, 47(3):1757–1780.

Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208.

Bertsimas, D., Gupta, V., and Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292.

Bertsimas, D., Sim, M., and Zhang, M. (2019). Adaptive distributionally robust optimization. *Management Science*, 65(2):604–618.

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. (2023). Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659*.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Cai, J.-F., Qu, X., Xu, W., and Ye, G.-B. (2016). Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction. *Applied and computational harmonic analysis*, 41(2):470–490.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2020). Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874*.

Chen, Z., Sim, M., and Xu, H. (2019). Distributionally robust optimization with infinitely constrained ambiguity sets. *Operations Research*, 67(5):1328–1344.

Clavier, P., Pennec, E. L., and Geist, M. (2023). Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372v1*.

de Castro Silva, J., Soma, N., and Maculan, N. (2003). A greedy search for the three-dimensional bin packing problem: the packing static stability case. *International Transactions in Operational Research*, 10(2):141–153.

Derman, E. and Mannor, S. (2020). Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*.

Ding, W., Shi, L., Chi, Y., and Zhao, D. (2023). Seeing is not believing: Robust reinforcement learning against spurious correlation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dong, J., Li, J., Wang, B., and Zhang, J. (2022). Online policy optimization for robust MDP. *arXiv preprint arXiv:2209.13841*.

Dong, K., Wang, Y., Chen, X., and Wang, L. (2019). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*.

Duchi, J. and Namkoong, H. (2018). Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*.

Fatemi, M., Killian, T. W., Subramanian, J., and Ghassemi, M. (2021). Medical dead-ends and learning to identify high-risk states and treatments. *Advances in Neural Information Processing Systems*, 34:4856–4870.

Gao, R. (2020). Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*.

Goyal, V. and Grand-Clement, J. (2022). Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*.

Han, S., Su, S., He, S., Han, S., Yang, H., and Miao, F. (2022). What is the solution for state adversarial multi-agent reinforcement learning? *arXiv preprint arXiv:2212.02705*.

Ho, C. P., Petrik, M., and Wiesemann, W. (2018). Fast bellman updates for robust MDPs. In *International Conference on Machine Learning*, pages 1979–1988. PMLR.

Ho, C. P., Petrik, M., and Wiesemann, W. (2021). Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46.

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

Jafarnia-Jahromi, M., Wei, C.-Y., Jain, R., and Luo, H. (2020). A model-free learning algorithm for infinite-horizon average-reward MDPs with near-optimal regret. *arXiv preprint arXiv:2006.04354*.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020a). Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020b). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096.

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

Kearns, M. J. and Singh, S. P. (1999). Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in neural information processing systems*, pages 996–1002.

Klopp, O., Lounici, K., and Tsybakov, A. B. (2017). Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564.

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

Kumar, N., Derman, E., Geist, M., Levy, K., and Mannor, S. (2023). Policy gradient for s-rectangular robust markov decision processes. *arXiv preprint arXiv:2301.13589*.

Lam, H. (2019). Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105.

Lee, J., Jeon, W., Lee, B., Pineau, J., and Kim, K.-E. (2021). Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR.

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2023a). Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research.*

Li, G., Chi, Y., Wei, Y., and Chen, Y. (2022a). Minimax-optimal multi-agent RL in Markov games with a generative model. *Neural Information Processing Systems.*

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022b). Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275.*

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2024). Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260.

Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. (2021). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2023b). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *accepted to Operations Research.*

Li, G., Yan, Y., Chen, Y., and Fan, J. (2023c). Minimax-optimal reward-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2304.07278.*

Li, Y. and Lan, G. (2023). First-order policy optimization for robust policy evaluation. *arXiv preprint arXiv:2307.15890.*

Li, Y., Zhao, T., and Lan, G. (2022c). First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579.*

Liang, Z., Ma, X., Blanchet, J., Zhang, J., and Zhou, Z. (2023). Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721.*

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. A. (2022). Batch policy learning in average reward markov decision processes. *Annals of statistics*, 50(6):3364.

Liu, S., Ngiam, K. Y., and Feng, M. (2019). Deep reinforcement learning for clinical decision support: a brief survey. *arXiv preprint arXiv:1907.09475.*

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. (2022). Distributionally robust Q-learning. In *International Conference on Machine Learning*, pages 13623–13643. PMLR.

Liu, Z. and Xu, P. (2024a). Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. *arXiv preprint arXiv:2402.15399.*

Liu, Z. and Xu, P. (2024b). Minimax optimal and computationally efficient algorithms for distributionally robust offline reinforcement learning. *arXiv preprint arXiv:2403.09621.*

Lu, M., Zhong, H., Zhang, T., and Blanchet, J. (2024). Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithm. *arXiv preprint arXiv:2404.03578.*

Ma, X., Liang, Z., Blanchet, J., Liu, M., Xia, L., Zhang, J., Zhao, Q., and Zhou, Z. (2022). Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620.*

Mahmood, A. R., Korenkevych, D., Vasan, G., Ma, W., and Bergstra, J. (2018). Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pages 561–591. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Moos, J., Hansel, K., Abdulsamad, H., Stark, S., Clever, D., and Peters, J. (2022). Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1):276–315.

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

OpenAI (2023). Gpt-4 technical report.

Pan, Y., Chen, Y., and Lin, F. (2023). Adjustable robust reinforcement learning for online 3d bin packing. *arXiv preprint arXiv:2310.04323*.

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

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. (2022). Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. (2023). Bridging distributionally robust learning and offline rl: An approach to mitigate distribution shift and partial data coverage. *arXiv preprint arXiv:2310.18434*.

Park, B. and Van Roy, B. (2015). Adaptive execution: Exploration and learning of price impact. *Operations Research*, 63(5):1058–1076.

Qiaoben, Y., Zhou, X., Ying, C., and Zhu, J. (2021). Strategically-timed state-observation attacks on deep reinforcement learning agents. In *ICML 2021 Workshop on Adversarial Machine Learning*.

Qu, G., Wierman, A., and Li, N. (2022). Scalable reinforcement learning for multiagent networked systems. *Operations Research*, 70(6):3601–3628.

Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

Ramesh, S. S., Sessa, P. G., Hu, Y., Krause, A., and Bogunovic, I. (2023). Distributionally robust model-based reinforcement learning with large state spaces. *arXiv preprint arXiv:2309.02236*.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Neural Information Processing Systems (NeurIPS)*.

Roy, A., Xu, H., and Pokutta, S. (2017). Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30.

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

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 19967–20025. PMLR.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018). Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196.

Smirnova, E., Dohmatob, E., and Mary, J. (2019). Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*.

Sun, K., Liu, Y., Zhao, Y., Yao, H., Jui, S., and Kong, L. (2021). Exploring the training robustness of distributional reinforcement learning against noisy state observations. *arXiv preprint arXiv:2109.08776*.

Tamar, A., Mannor, S., and Xu, H. (2014). Scaling up robust MDPs using function approximation. In *International conference on machine learning*, pages 181–189. PMLR.

Tan, K. L., Esfandiari, Y., Lee, X. Y., and Sarkar, S. (2020). Robustifying reinforcement learning agents via action space adversarial training. In *2020 American control conference (ACC)*, pages 3959–3964. IEEE.

Tessler, C., Efroni, Y., and Mannor, S. (2019). Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pages 6215–6224. PMLR.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*, volume 11. Springer.

Uehara, M., Shi, C., and Kallus, N. (2022). A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. J. (2019). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*.

Wang, H., Shi, L., and Chi, Y. (2024). Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*.

Wang, K., Gadot, U., Kumar, N., Levy, K., and Mannor, S. (2023a). Robust reinforcement learning via adversarial kernel approximation. *arXiv preprint arXiv:2306.05859*.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023b). A finite sample complexity bound for distributionally robust Q-learning. *arXiv preprint arXiv:2302.13203*.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023c). On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023d). Sample complexity of variance-reduced distributionally robust Q-learning. *arXiv preprint arXiv:2305.18420*.

Wang, Y. and Zou, S. (2021). Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34.

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

Wolff, E. M., Topcu, U., and Murray, R. M. (2012). Robust control of uncertain Markov decision processes with temporal logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3372–3379. IEEE.

Woo, J., Joshi, G., and Chi, Y. (2023). The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. *arXiv preprint arXiv:2305.10697*.

Woo, J., Shi, L., Joshi, G., and Chi, Y. (2024). Federated offline reinforcement learning: Collaborative single-policy coverage suffices. *arXiv preprint arXiv:2402.05876*.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34.

Xiong, Z., Eappen, J., Zhu, H., and Jagannathan, S. (2022). Defending observation attacks in deep reinforcement learning via detection and denoising. *arXiv preprint arXiv:2206.07188*.

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

Xu, Z., Panaganti, K., and Kalathil, D. (2023). Improved sample complexity bounds for distributionally robust reinforcement learning. *arXiv preprint arXiv:2303.02783*.

Yan, Y., Li, G., Chen, Y., and Fan, J. (2022). The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*.

Yang, K., Yang, L., and Du, S. (2021). Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR.

Yang, W., Wang, H., Kozuno, T., Jordan, S. M., and Zhang, Z. (2023). Avoiding model estimation in robust markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*.

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

Yin, M., Bai, Y., and Wang, Y.-X. (2021). Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*.

Zhang, H., Chen, H., Boning, D., and Hsieh, C.-J. (2021). Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*.

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

Zhang, R., Hu, Y., and Li, N. (2023). Regularized robust mdps and risk-sensitive mdps: Equivalence, policy gradient, and sample complexity. *arXiv preprint arXiv:2306.11626*.

Zhang, Z., Zhou, Y., and Ji, X. (2020b). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33.

Zhao, H., Yu, Y., and Xu, K. (2021). Learning efficient online 3d bin packing on packing configuration trees. In *International conference on learning representations*.

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

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Proof of the preliminaries

A.1 Proof of Lemma 1 and Lemma 2

Proof of Lemma 1. To begin with, applying (Iyengar, 2005, Lemma 4.3), the term of interest obeys

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P)} \mathcal{P}V = \max_{\mu\in\mathbb{R}^S,\mu\geq 0} \left\{ P\left(V-\mu\right) - \sigma\left(\max_{s'}\left\{V(s')-\mu(s')\right\} - \min_{s'}\left\{V(s')-\mu(s')\right\}\right)\right\}, \tag{95}$$

where $\mu(s')$ represents the s'-th entry of $\mu\in\mathbb{R}^S$. Denoting $\mu^\star$ as the optimal dual solution, taking $\alpha = \max_{s'}\left\{V(s')-\mu^\star(s')\right\}$, it is easily verified that $\mu^\star$ obeys

$$\mu^\star(s) = \begin{cases} V(s)-\alpha, & \text{if } V(s)>\alpha \\ 0, & \text{otherwise.}\end{cases} \tag{96}$$

Therefore, (95) can be solved by optimizing α as below (Iyengar, 2005, Lemma 4.3):

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P)} \mathcal{P}V = \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left\{ P\left[V\right]_\alpha - \sigma\left(\alpha - \min_{s'}\left[V\right]_\alpha(s')\right)\right\}. \tag{97}$$

Proof of Lemma 2. Due to strong duality (Iyengar, 2005, Lemma 4.2), it holds that

$$\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P)} \mathcal{P}V = \max_{\mu\in\mathbb{R}^{S},\mu\geq 0} \left\{ P\left(V-\mu\right) - \sqrt{\sigma \mathsf{Var}_{P}\left(V-\mu\right)} \right\}, \tag{98}$$

and the optimal $\mu^{\star}$ obeys

$$\mu^{\star}(s) = \begin{cases} V(s)-\alpha, & \text{if } V(s) > \alpha \\ 0, & \text{otherwise.} \end{cases} \tag{99}$$

for some $\alpha \in [\min_s V(s), \max_s V(s)]$. As a result, solving (98) is equivalent to optimizing the scalar α as below:

$$\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P)} \mathcal{P}V = \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left\{ P[V]_{\alpha} - \sqrt{\sigma \mathsf{Var}_{P}\left([V]_{\alpha}\right)} \right\}. \tag{100}$$

A.2 Proof of Lemma 5

Applying the γ-contraction property in Lemma 4 directly yields that for any $t \geq 0$,

$$\begin{aligned} \|\widehat{Q}_t - \widehat{Q}^{\star,\sigma}\|_{\infty} = \big\|\widehat{\mathcal{T}}^{\sigma}(\widehat{Q}_{t-1}) - \widehat{\mathcal{T}}^{\sigma}(\widehat{Q}^{\star,\sigma})\big\|_{\infty} &\leq \gamma \big\|\widehat{Q}_{t-1} - \widehat{Q}^{\star,\sigma}\big\|_{\infty} \\ &\leq \cdots \leq \gamma^{t}\big\|\widehat{Q}_0 - \widehat{Q}^{\star,\sigma}\big\|_{\infty} = \gamma^{t}\big\|\widehat{Q}^{\star,\sigma}\big\|_{\infty} \leq \frac{\gamma^{t}}{1-\gamma}, \end{aligned}$$

where the last inequality holds by the fact $\|\widehat{Q}^{\star,\sigma}\|_{\infty} \leq \frac{1}{1-\gamma}$ (see Lemma 4). In addition,

$$\big\|\widehat{V}_t - \widehat{V}^{\star,\sigma}\big\|_{\infty} = \max_{s\in\mathcal{S}} \left\| \max_{a\in\mathcal{A}} \widehat{Q}_t(s,a) - \max_{a\in\mathcal{A}} \widehat{Q}^{\star,\sigma}(s,a) \right\|_{\infty} \leq \big\|\widehat{Q}_t - \widehat{Q}^{\star,\sigma}\big\|_{\infty} \leq \frac{\gamma^{t}}{1-\gamma},$$

where the penultimate inequality holds by the maximum operator is 1-Lipschitz. This completes the proof of (47).

We now move to establish (48). Note that there exists at least one state $s_0 \in \mathcal{S}$ that is associated with the maximum of the value gap, i.e.,

$$\big\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\big\|_{\infty} = \widehat{V}^{\star,\sigma}(s_0) - \widehat{V}^{\widehat{\pi},\sigma}(s_0) \geq \widehat{V}^{\star,\sigma}(s) - \widehat{V}^{\widehat{\pi},\sigma}(s), \qquad \forall s \in \mathcal{S}.$$

Recall $\widehat{\pi}^{\star}$ is the optimal robust policy for the empirical RMDP $\widehat{\mathcal{M}}_{\mathsf{rob}}$. For convenience, we denote $a_1 = \widehat{\pi}^{\star}(s_0)$ and $a_2 = \widehat{\pi}(s_0)$. Then, since $\widehat{\pi}$ is the greedy policy w.r.t. $\widehat{Q}_T$, one has

$$r(s_0,a_1) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^{0}_{s_0,a_1})} \mathcal{P}\widehat{V}_{T-1} = \widehat{Q}_T(s_0,a_1) \leq \widehat{Q}_T(s_0,a_2) = r(s_0,a_2) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^{0}_{s_0,a_2})} \mathcal{P}\widehat{V}_{T-1}. \tag{101}$$

Recalling the notation in (37), the above fact and (48) altogether yield

$$\begin{aligned} r(s_0,a_1) + \gamma \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1}\left(\widehat{V}^{\star,\sigma} - \varepsilon_{\mathsf{opt}}1\right) &\leq r(s_0,a_1) + \gamma \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1}\widehat{V}_{T-1} \\ &\leq r(s_0,a_2) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^{0}_{s_0,a_2})} \mathcal{P}\widehat{V}_{T-1} \\ &\overset{\text{(i)}}{\leq} r(s_0,a_2) + \gamma \widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2}\widehat{V}_{T-1} \\ &\leq r(s_0,a_2) + \gamma \widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2}\left(\widehat{V}^{\star,\sigma} + \varepsilon_{\mathsf{opt}}1\right), \end{aligned} \tag{102}$$

where (i) follows from the optimality criteria. The term of interest can be controlled as

$$\begin{aligned} &\big\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\big\|_{\infty} \\ &= \widehat{V}^{\star,\sigma}(s_0) - \widehat{V}^{\widehat{\pi},\sigma}(s_0) \end{aligned}$$

$$
\begin{aligned}
&= r(s_0,a_1) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}^{\star,\sigma} - \Big(r(s_0,a_2) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}^{\widehat{\pi},\sigma} \Big) \\
&= r(s_0,a_1) - r(s_0,a_2) + \gamma \Big(\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}^{\star,\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}^{\widehat{\pi},\sigma} \Big) \\
&\overset{\text{(i)}}{\leq} 2\gamma\varepsilon_{\mathsf{opt}} + \gamma \Big(\widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2} \widehat{V}^{\star,\sigma} - \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1} \widehat{V}^{\star,\sigma} + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}^{\star,\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}^{\widehat{\pi},\sigma} \Big) \\
&= 2\gamma\varepsilon_{\mathsf{opt}} + \gamma \Big(\widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2} \widehat{V}^{\star,\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}^{\widehat{\pi},\sigma} \Big) + \gamma \Big(\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}^{\star,\sigma} - \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1} \widehat{V}^{\star,\sigma} \Big) \\
&\overset{\text{(ii)}}{\leq} 2\gamma\varepsilon_{\mathsf{opt}} + \gamma \widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2} \big(\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma} \big) + \gamma \Big(\widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1} \widehat{V}^{\star,\sigma} - \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1} \widehat{V}^{\star,\sigma} \Big) \\
&\leq 2\gamma\varepsilon_{\mathsf{opt}} + \gamma \big\| \widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma} \big\|_\infty, \qquad (103)
\end{aligned}
$$

where (i) holds by plugging in (102), and (ii) follows from $\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}^{\star,\sigma} \leq \mathcal{P}\widehat{V}^{\star,\sigma}$ for any $\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})$. Rearranging (103) leads to

$$
\big\| \widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma} \big\|_\infty \leq \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}.
$$

B Proof of the auxiliary lemmas for Theorem 1

B.1 Proof of Lemma 6

To begin, note that there at leasts exist one state s_0 for any $V^{\pi,\sigma}$ such that $V^{\pi,\sigma}(s_0) = \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s)$. With this in mind, for any policy π, one has by the definition in (5) and the Bellman's equation (7a),

$$
\begin{aligned}
\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) &= \max_{s\in\mathcal{S}} \mathbb{E}_{a\sim\pi(\cdot\,|\,s)} \Big[r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{s,a})} \mathcal{P}V^{\pi,\sigma} \Big] \\
&\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big(1 + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{s,a})} \mathcal{P}V^{\pi,\sigma} \Big),
\end{aligned}
$$

where the second line holds since the reward function $r(s,a) \in [0,1]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. To continue, note that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, there exists some $\widetilde{P}_{s,a} \in \mathbb{R}^S$ constructed by reducing the values of some elements of $P_{s,a}$ to obey $P_{s,a} \geq \widetilde{P}_{s,a} \geq 0$ and $\sum_{s'} (P_{s,a}(s') - \widetilde{P}_{s,a}(s')) = \sigma$. This implies $\widetilde{P}_{s,a} + \sigma e_{s_0}^\top \in \mathcal{U}^\sigma(P_{s,a})$, where e_{s_0} is the standard basis vector supported on s_0, since $\frac{1}{2}\|\widetilde{P}_{s,a} + \sigma e_{s_0}^\top - P_{s,a}\|_1 \leq \frac{1}{2}\|\widetilde{P}_{s,a} - P_{s,a}\|_1 + \frac{\sigma}{2} = \sigma$. Consequently,

$$
\begin{aligned}
\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{s,a})} \mathcal{P}V^{\pi,\sigma} \leq \Big(\widetilde{P}_{s,a} + \sigma e_{s_0}^\top \Big) V^{\pi,\sigma} &\leq \big\| \widetilde{P}_{s,a} \big\|_1 \big\| V^{\pi,\sigma} \big\|_\infty + \sigma V^{\pi,\sigma}(s_0) \\
&\leq (1-\sigma) \max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) + \sigma \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s), \qquad (104)
\end{aligned}
$$

where the second inequality holds by $\big\|\widetilde{P}_{s,a}\big\|_1 = \sum_{s'} \widetilde{P}_{s,a}(s') = -\sum_{s'} \Big(P_{s,a}(s') - \widetilde{P}_{s,a}(s') \Big) + \sum_{s'} P_{s,a}(s') = 1-\sigma$. Plugging this back to the previous relation gives

$$
\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \leq 1 + \gamma\,(1-\sigma) \max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) + \gamma\sigma \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s),
$$

which, by rearranging terms, immediately yields

$$
\begin{aligned}
\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) &\leq \frac{1 + \gamma\sigma \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s)}{1 - \gamma\,(1-\sigma)} \\
&\leq \frac{1}{(1-\gamma) + \gamma\sigma} + \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \leq \frac{1}{\gamma \max\{1-\gamma, \sigma\}} + \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s).
\end{aligned}
$$

B.2 Proof of Lemma 7

Observing that each row of P_π belongs to $\Delta(S)$, it can be directly verified that each row of $(1-\gamma)\left(I-\gamma P_\pi\right)^{-1}$ falls into $\Delta(S)$. As a result,

$$
\begin{aligned}
\left(I-\gamma P_\pi\right)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} &= \frac{1}{1-\gamma}(1-\gamma)\left(I-\gamma P_\pi\right)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \\
&\overset{\text{(i)}}{\leq} \frac{1}{1-\gamma}\sqrt{(1-\gamma)\left(I-\gamma P_\pi\right)^{-1}\mathrm{Var}_{P_\pi}(V^{\pi,P})} \\
&= \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^t\mathrm{Var}_{P_\pi}(V^{\pi,P})},
\end{aligned}
\tag{105}
$$

where (i) holds by Jensen's inequality.

To continue, denoting the minimum value of V as $V_{\min} = \min_{s\in\mathcal{S}} V^{\pi,P}(s)$ and $V' := V^{\pi,P} - V_{\min}1$. We control $\mathrm{Var}_{P_\pi}(V^{\pi,P})$ as follows:

$$
\begin{aligned}
&\mathrm{Var}_{P_\pi}(V^{\pi,P}) \\
&\overset{\text{(i)}}{=} \mathrm{Var}_{P_\pi}(V') = P_\pi\left(V'\circ V'\right) - \big(P_\pi V'\big)\circ\big(P_\pi V'\big) \\
&\overset{\text{(ii)}}{=} P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma^2}\left(V' - r_\pi + (1-\gamma)V_{\min}1\right)\circ\left(V' - r_\pi + (1-\gamma)V_{\min}1\right) \\
&= P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma^2}V'\circ V' + \frac{2}{\gamma^2}V'\circ\left(r_\pi - (1-\gamma)V_{\min}1\right) \\
&\quad - \frac{1}{\gamma^2}\left(r_\pi - (1-\gamma)V_{\min}1\right)\circ\left(r_\pi - (1-\gamma)V_{\min}1\right) \\
&\leq P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1,
\end{aligned}
\tag{106}
$$

where (i) holds by the fact that $\mathrm{Var}_{P_\pi}(V^{\pi,P} - b1) = \mathrm{Var}_{P_\pi}(V^{\pi,P})$ for any scalar b and $V^{\pi,P}\in\mathbb{R}^S$, (ii) follows from $V' = r_\pi + \gamma P_\pi V^{\pi,P} - V_{\min}1 = r_\pi - (1-\gamma)V_{\min}1 + \gamma P_\pi V'$, and the last line arises from $\frac{1}{\gamma^2}V'\circ V' \geq \frac{1}{\gamma}V'\circ V'$ and $\|r_\pi - (1-\gamma)V_{\min}1\|_\infty \leq 1$. Plugging (106) back to (105) leads to

$$
\begin{aligned}
&\left(I-\gamma P_\pi\right)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \leq \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^t\left(P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1\right)} \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^t\left(P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V'\right)\right|} + \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^t\frac{2}{\gamma^2}\|V'\|_\infty 1} \\
&\leq \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\left(\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^{t+1} - \sum_{t=0}^{\infty}\gamma^{t-1}\left(P_\pi\right)^t\right)\left(V'\circ V'\right)\right|} + \sqrt{\frac{2\|V'\|_\infty 1}{\gamma^2(1-\gamma)^2}} \\
&\overset{\text{(ii)}}{\leq} \sqrt{\frac{\|V'\|_\infty^2 1}{\gamma(1-\gamma)}} + \sqrt{\frac{2\|V'\|_\infty 1}{\gamma^2(1-\gamma)^2}} \\
&\leq \sqrt{\frac{8\|V'\|_\infty 1}{\gamma^2(1-\gamma)^2}},
\end{aligned}
\tag{107}
$$

where (i) holds by the triangle inequality, (ii) holds by following recursion, and the last inequality holds by $\|V'\|_\infty \leq \frac{1}{1-\gamma}$.

B.3 Proof of Lemma 8

Step 1: controlling $\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\|_\infty$: bounding the first term in (56). To control the two terms in (56), we first introduce the following lemma whose proof is postponed to Appendix B.5.

Lemma 11. *Consider any $\delta \in (0,1)$. Setting $N \geq \log(\frac{18SAN}{\delta})$, with probability at least $1-\delta$, one has*

$$\begin{aligned}\left|\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right| &\leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)}1 \\ &\leq 3\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}1, \end{aligned} \tag{108}$$

where $\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})$ *is defined in* (36).

Armed with the above lemma, now we control the first term on the right hand side of (56) as follows:

$$\begin{aligned}
&\left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right) \\
&\overset{\text{(i)}}{\leq} \left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left\|\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right\|_\infty \\
&\overset{\text{(ii)}}{\leq} \left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left(2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)}1\right) \\
&\leq \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)}\left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}1 + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})}}_{=:\mathcal{C}_1} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})\right|}}_{=:\mathcal{C}_2} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)}_{=:\mathcal{C}_3},
\end{aligned} \tag{109}$$

where (i) holds by $\left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1} \geq 0$, (ii) follows from Lemma 11, and the last inequality arise from

$$\begin{aligned}
\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} &= \left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right) + \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})} \\
&\leq \left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right) + \sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})\right|} + \sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})}
\end{aligned}$$

by applying the triangle inequality.

To continue, observing that each row of $\widehat{\underline{P}}^{\pi^\star,V}$ is a probability distribution obeying that the sum is 1, we arrive at

$$\left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}1 = \left(I + \sum_{t=1}^{\infty}\gamma^t\left(\widehat{\underline{P}}^{\pi^\star,V}\right)^t\right)1 = \frac{1}{1-\gamma}1. \tag{110}$$

Armed with this fact, we shall control the other three terms $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ in (109) separately.

- Consider $\mathcal{C}_1$. We first introduce the following lemma, whose proof is postponed to Appendix B.6.

  Lemma 12. *Consider any $\delta \in (0,1)$. With probability at least $1-\delta$, one has*

$$\left(I-\gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})} \leq 4\sqrt{\frac{\left(1+\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}}}1 \leq 4\sqrt{\frac{\left(1+\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\right)}{\gamma^3(1-\gamma)^3}}1.$$

Applying Lemma 12 and inserting back to (109) leads to

$$\begin{aligned}
\mathcal{C}_1 &= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})} \\
&\leq 8\sqrt{\frac{\log(\frac{18SAN}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}\left(1+\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\right)}1. \qquad (111)
\end{aligned}$$

- Consider $\mathcal{C}_2$. First, denote $V' := V^{\star,\sigma} - \min_{s'\in\mathcal{S}} V^{\star,\sigma}(s')1$, by Lemma 6, it follows that

$$0 \leq V' \leq \frac{1}{\gamma\max\{1-\gamma,\sigma\}}1. \qquad (112)$$

Then, we have for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, and $P_{s,a}\in\Delta(\mathcal{S})$, and $\widetilde{P}_{s,a}\in\mathcal{U}^\sigma(P_{s,a})$:

$$\begin{aligned}
\left|\mathsf{Var}_{\widetilde{P}_{s,a}}(V^{\star,\sigma}) - \mathsf{Var}_{P_{s,a}}(V^{\star,\sigma})\right| &= \left|\mathsf{Var}_{\widetilde{P}_{s,a}}(V') - \mathsf{Var}_{P_{s,a}}(V')\right| \\
&\leq \left\|\widetilde{P}_{s,a} - P_{s,a}\right\|_1 \left\|V'\right\|_\infty^2 \\
&\leq \frac{2\sigma}{\gamma^2(\max\{1-\gamma,\sigma\})^2} \leq \frac{2}{\gamma^2\max\{1-\gamma,\sigma\}}. \qquad (113)
\end{aligned}$$

Applying the above relation we obtain

$$\begin{aligned}
\mathcal{C}_2 &= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})\right|} \\
&= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left|\Pi^{\pi^\star}\left(\mathrm{Var}_{\widehat{P}^0}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{P}^{\pi^\star,V}}(V^{\star,\sigma})\right)\right|} \\
&\leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left\|\mathrm{Var}_{\widehat{P}^0}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{P}^{\pi^\star,V}}(V^{\star,\sigma})\right\|_\infty 1} \\
&\leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\frac{2}{\gamma^2\max\{1-\gamma,\sigma\}}}1 = 2\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1, \qquad (114)
\end{aligned}$$

where the last equality uses $\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}1 = \frac{1}{1-\gamma}$ (cf. (110)).

- Consider $\mathcal{C}_3$. The following lemma plays an important role.

Lemma 13. *(Panaganti and Kalathil, 2022, Lemma 6) Consider any $\delta\in(0,1)$. For any fixed policy π and fixed value vector $V\in\mathbb{R}^S$, one has with probability at least $1-\delta$,*

$$\left|\sqrt{\mathrm{Var}_{\widehat{P}^\pi}(V)} - \sqrt{\mathrm{Var}_{P^\pi}(V)}\right| \leq \sqrt{\frac{2\|V\|_\infty^2\log(\frac{2SA}{\delta})}{N}}1.$$

Applying Lemma 13 with $\pi=\pi^\star$ and $V=V^{\star,\sigma}$ leads to

$$\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})} \leq \sqrt{\frac{2\|V^{\star,\sigma}\|_\infty^2\log(\frac{2SA}{\delta})}{N}}1,$$

which can be plugged in (109) to verify

$$\mathcal{C}_3 = 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)$$

$$\leq \frac{4}{(1-\gamma)} \frac{\log(\frac{SAN}{\delta})\|V^{\star,\sigma}\|_\infty}{N} 1 \leq \frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N} 1, \tag{115}$$

where the last line uses $\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,V}\right)^{-1} 1 = \frac{1}{1-\gamma}$ (cf. (110)).

Finally, inserting the results of $\mathcal{C}_1$ in (111), $\mathcal{C}_2$ in (114), $\mathcal{C}_3$ in (115), and (110) back into (109) gives

$$\begin{aligned}
&\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(\underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right) \leq 8\sqrt{\frac{\log(\frac{18SAN}{\delta})}{\gamma^3(1-\gamma)^2 \max\{1-\gamma,\sigma\}N}\left(1+\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}\right)} 1 \\
&\quad + 2\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} 1 + \frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N} 1 + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2} 1 \\
&\leq 10\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}\left(1+\sqrt{\frac{\log(\frac{SAN}{\delta})}{(1-\gamma)^2 N}}\right)} 1 + \frac{5\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N} 1 \\
&\leq 160\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} 1 + \frac{5\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N} 1,
\end{aligned} \tag{116}$$

where the last inequality holds by the fact $\gamma \geq \frac{1}{4}$ and letting $N \geq \frac{\log(\frac{SAN}{\delta})}{(1-\gamma)^2}$.

Step 2: controlling $\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\|_\infty$: bounding the second term in (56). To proceed, applying Lemma 11 on the second term of the right hand side of (56) leads to

$$\begin{aligned}
&\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right) \\
&\leq 2\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{P_{\pi^\star}}(V^{\star,\sigma})} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)} 1\right) \\
&\leq \frac{2\log(\frac{18SAN}{\delta})}{N(1-\gamma)}\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1} 1 + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(\widehat{V}^{\pi^\star,\sigma})}}_{=:\mathcal{C}_4} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma})}\right)}_{=:\mathcal{C}_5} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\sqrt{\left|\mathrm{Var}_{\widehat{P}_{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(V^{\star,\sigma})\right|}\right)}_{=:\mathcal{C}_6} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\sqrt{\mathrm{Var}_{P_{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}_{\pi^\star}}(V^{\star,\sigma})}\right)}_{=:\mathcal{C}_7},
\end{aligned} \tag{117}$$

where the last term $\widetilde{\mathcal{C}}_3$ can be controlled the same as $\mathcal{C}_3$ in (115). We now bound the above terms separately.

- Applying Lemma 7 with $P = \widehat{P}^{\pi^\star,\widehat{V}}$, $\pi = \pi^\star$ and taking $V = \widehat{V}^{\pi^\star,\sigma}$ which obeys $\widehat{V}^{\pi^\star,\sigma} = r_{\pi^\star} + \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}} \widehat{V}^{\pi^\star,\sigma}$, and in view of (110), the term $\mathcal{C}_4$ in (117) can be controlled as follows:

$$\mathcal{C}_4 = 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(\widehat{V}^{\pi^\star,\sigma})}$$

$$
\begin{aligned}
&\le 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\frac{8(\max_s \widehat{V}^{\pi^\star,\sigma}(s)-\min_s \widehat{V}^{\pi^\star,\sigma}(s))}{\gamma^2(1-\gamma)^2}}1\\
&\le 8\sqrt{\frac{\log(\frac{18SAN}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1, \qquad (118)
\end{aligned}
$$

where the last inequality holds by applying Lemma 6.

- To continue, considering $\mathcal{C}_5$, we directly observe that (in view of (110))

$$
\begin{aligned}
\mathcal{C}_5 &= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(V^{\pi^\star,\sigma}-\widehat{V}^{\pi^\star,\sigma})}\\
&\le 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1. \qquad (119)
\end{aligned}
$$

- Then, it is easily verified that $\mathcal{C}_6$ can be controlled similarly as (114) as follows:

$$
\mathcal{C}_6 \le 2\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1. \qquad (120)
$$

- Similarly, $\mathcal{C}_7$ can be controlled the same as (115) shown below:

$$
\mathcal{C}_7 \le \frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1. \qquad (121)
$$

Combining the results in (118), (119), (120), and (121) and inserting back to (117) leads to

$$
\begin{aligned}
&\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\right) \le 8\sqrt{\frac{\log(\frac{18SAN}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1\\
&\qquad + 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1 + 2\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1 + \frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1\\
&\le 80\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1 + 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1 + \frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1, \qquad (122)
\end{aligned}
$$

where the last inequality follows from the assumption $\gamma \ge \frac{1}{4}$.

Finally, inserting (116) and (122) back to (56) yields

$$
\begin{aligned}
&\left\|\widehat{V}^{\pi^\star,\sigma}-V^{\pi^\star,\sigma}\right\|_\infty\\
&\le \max\Bigg\{160\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{5\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N},\\
&\qquad 80\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty+\frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}\Bigg\}\\
&\le 160\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{8\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}, \qquad (123)
\end{aligned}
$$

where the last inequality holds by taking $N \ge \frac{16\log(\frac{SAN}{\delta})}{(1-\gamma)^2}$.

B.4 Proof of Lemma 9

Step 1: controlling $\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\|_\infty$: bounding the first term in (57). To begin with, we introduce the following lemma which controls the main term on the right hand side of (57), which is proved in Appendix B.7.

Lemma 14. *Consider any $\delta \in (0,1)$. Taking $N \geq \log\left(\frac{54SAN^2}{(1-\gamma)\delta}\right)$, with probability at least $1-\delta$, one has*

$$\begin{aligned}
\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right| &\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})}1 + \frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)}1 + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1 \\
&\leq 10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}}1 + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1. \qquad (124)
\end{aligned}$$

With Lemma 14 in hand, we have

$$\begin{aligned}
&\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right) \\
&\overset{\text{(i)}}{\leq} \left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right| \\
&\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})} + \left(I - \gamma P_Q^{\widehat{\pi},V^{\widehat{\pi}}}\right)^{-1}\left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)1 \\
&\overset{\text{(ii)}}{\leq} \left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}\right)1 + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})}}_{=:\mathcal{D}_1} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})\right|}}_{=:\mathcal{D}_2} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma})\right|}}_{=:\mathcal{D}_3}, \qquad (125)
\end{aligned}$$

where (i) and (ii) hold by the fact that each row of $(1-\gamma)\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}$ is a probability vector that falls into $\Delta(\mathcal{S})$.

The remainder of the proof will focus on controlling the three terms in (125) separately.

- For $\mathcal{D}_1$, we introduce the following lemma, whose proof is postponed to B.8.

  Lemma 15. *Consider any $\delta \in (0,1)$. Taking $N \geq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2}$ and $\varepsilon_{\mathsf{opt}} \leq \frac{1-\gamma}{\gamma}$, one has with probability at least $1-\delta$,*

  $$\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \leq 6\sqrt{\frac{1}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}}}1 \leq 6\sqrt{\frac{1}{(1-\gamma)^3\gamma^2}}1.$$

  Applying Lemma 15 and (110) to (125) leads to

  $$\begin{aligned}
  \mathcal{D}_1 &= 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \\
  &\leq 12\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1. \qquad (126)
  \end{aligned}$$

- Applying Lemma 3 with $\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\|_\infty \leq \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}$ and (110), $\mathcal{D}_2$ can be controlled as

$$\begin{aligned}\mathcal{D}_2 &= 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\Big)^{-1}\sqrt{\Big|\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})\Big|} \\ &\leq 4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\Big)^{-1}\frac{\sqrt{\gamma\varepsilon_{\mathsf{opt}}}}{1-\gamma} \leq 4\sqrt{\frac{\gamma\varepsilon_{\mathsf{opt}}\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^4N}}1. \end{aligned} \tag{127}$$

- $\mathcal{D}_3$ can be controlled similar to $\mathcal{C}_2$ in (114) as follows:

$$\begin{aligned}\mathcal{D}_3 &= 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\Big)^{-1}\sqrt{\Big|\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma})\Big|} \\ &\leq 4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\Big)^{-1}\sqrt{\frac{1}{\gamma^2\max\{1-\gamma,\sigma\}}}1 \leq 4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1 \end{aligned} \tag{128}$$

Finally, summing up the results in (126), (127), and (128) and inserting them back to (125) yields: taking $N \geq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2}$ and $\varepsilon_{\mathsf{opt}} \leq \frac{1-\gamma}{\gamma}$, with probability at least $1-\delta$,

$$\begin{aligned}&\Big(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\Big)^{-1}\Big(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\Big) \leq \left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}\right)1 \\ &\quad + 12\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1 + 4\sqrt{\frac{\gamma\varepsilon_{\mathsf{opt}}\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^4N}}1 + 4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1 \\ &\leq 16\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1 + \frac{14\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}1, \end{aligned} \tag{129}$$

where the last inequality holds by taking $\varepsilon_{\mathsf{opt}} \leq \min\left\{\frac{1-\gamma}{\gamma}, \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}\right\} = \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}$.

Step 2: controlling $\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\|_\infty$: bounding the second term in (57). Towards this, applying Lemma 14 leads to

$$\begin{aligned}&\Big(I - \gamma\underline{P}^{\widehat{\pi},V}\Big)^{-1}\Big(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\Big) \leq \Big(I - \gamma\underline{P}^{\widehat{\pi},V}\Big)^{-1}\Big|\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\Big| \\ &\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I - \gamma\underline{P}^{\widehat{\pi},V}\Big)^{-1}\sqrt{\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})} + \Big(I - \gamma\underline{P}^{\widehat{\pi},V}\Big)^{-1}\left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)1 \\ &\leq \left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}\right)1 + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I - \gamma\underline{P}^{\widehat{\pi},V}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(V^{\widehat{\pi},\sigma})}}_{=:\mathcal{D}_4} \\ &\quad + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I - \gamma\underline{P}^{\widehat{\pi},V}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma})}}_{=:\mathcal{D}_5} \\ &\quad + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\Big)^{-1}\sqrt{\Big|\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\widehat{\pi},\sigma})\Big|}}_{=:\mathcal{D}_6} \end{aligned}$$

$$+ \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\Big)^{-1}\sqrt{\Big|\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})-\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\star,\sigma})\Big|}}_{=:\mathcal{D}_7}. \tag{130}$$

We shall bound each of the terms separately.

- Applying Lemma 7 with $P=\underline{P}^{\widehat{\pi},V}$, $\pi=\widehat{\pi}$, and taking $V=V^{\widehat{\pi},\sigma}$ which obeys $V^{\widehat{\pi},\sigma}=r_{\widehat{\pi}}+\gamma\underline{P}^{\widehat{\pi},V}V^{\widehat{\pi},\sigma}$, the term $\mathcal{D}_4$ can be controlled similar to (118) as follows:

$$\mathcal{D}_4 \le 8\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1. \tag{131}$$

- For $\mathcal{D}_5$, it is observed that

$$\begin{aligned}\mathcal{D}_5 &= 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\Big(I-\gamma\underline{P}^{\widehat{\pi},V}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma})}\\ &\le 2\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2N}}\Big\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\Big\|_\infty 1. \end{aligned}\tag{132}$$

- Next, observing that $\mathcal{D}_6$ and $\mathcal{D}_7$ are almost the same as the terms $\mathcal{D}_2$ (controlled in (127)) and $\mathcal{D}_3$ (controlled in (128)) in (125), it is easily verified that they can be controlled as follows

$$\mathcal{D}_6 \le 4\sqrt{\frac{\gamma\varepsilon_{\mathsf{opt}}\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^4N}}1, \qquad \mathcal{D}_7 \le 4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1. \tag{133}$$

Then inserting the results in (131), (132), and (133) back to (130) leads to

$$\begin{aligned}&\Big(I-\gamma\underline{P}^{\widehat{\pi},V}\Big)^{-1}\Big(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\Big)\\ &\le\left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}\right)1+8\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1\\ &\quad+2\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2N}}\Big\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\Big\|_\infty 1+4\sqrt{\frac{\gamma\varepsilon_{\mathsf{opt}}\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^4N}}1+4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1\\ &\le 12\sqrt{\frac{2\log(\frac{8SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+\frac{14\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}1+2\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2N}}\Big\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\Big\|_\infty 1, \end{aligned}\tag{134}$$

where the last inequality holds by letting $\varepsilon_{\mathsf{opt}} \le \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}$, which directly satisfies $\varepsilon_{\mathsf{opt}} \le \frac{1-\gamma}{\gamma}$ by letting $N \ge \frac{\log(\frac{54SAN^2}{\delta})}{1-\gamma}$.

Finally, inserting (129) and (134) back to (57) yields: taking $\varepsilon_{\mathsf{opt}} \le \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}$ and $N \ge \frac{16\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2}$, with probability at least $1-\delta$, one has

$$\begin{aligned}&\Big\|\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma}\Big\|_\infty\\ &\le\max\Bigg\{16\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{14\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2},\\ &\quad 12\sqrt{\frac{2\log(\frac{8SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{14\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}+2\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2N}}\Big\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\Big\|_\infty\Bigg\}\\ &\le 24\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{28\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}. \end{aligned}\tag{135}$$

B.5 Proof of Lemma 11

Step 1: controlling the point-wise concentration. We first consider a more general term w.r.t. any fixed (independent from $\widehat{P}^0$) value vector V obeying $0 \leq V \leq \frac{1}{1-\gamma}1$ and any policy π. Invoking Lemma 1 leads to that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}\left|\widehat{P}^{\pi,V}_{s,a} V - P^{\pi,V}_{s,a} V\right| &\leq \Big| \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ \widehat{P}^0_{s,a} [V]_\alpha - \sigma \left(\alpha - \min_{s'} [V]_\alpha (s')\right)\right\} \\ &\qquad - \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ P^0_{s,a} [V]_\alpha - w\sigma \left(\alpha - \min_{s'} [V]_\alpha (s')\right)\right\} \Big| \\ &\leq \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \underbrace{\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right|}_{=: g_{s,a}(\alpha, V)}, \qquad (136)\end{aligned}$$

where the last inequality holds by that the maximum operator is 1-Lipschitz.

Then for a fixed α and any vector V that is independent with $\widehat{P}^0$, using the Bernstein's inequality, one has with probability at least $1-\delta$,

$$\begin{aligned} g_{s,a}(\alpha, V) = \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right| &\leq \sqrt{\frac{2\log(\frac{2}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}([V]_\alpha)} + \frac{2\log(\frac{2}{\delta})}{3N(1-\gamma)} \\ &\leq \sqrt{\frac{2\log(\frac{2}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{2\log(\frac{2}{\delta})}{3N(1-\gamma)}. \qquad (137)\end{aligned}$$

Step 2: deriving the uniform concentration. To obtain the union bound, we first notice that $g_{s,a}(\alpha, V)$ is 1-Lipschitz w.r.t. α for any V obeying $\|V\|_\infty \leq \frac{1}{1-\gamma}$. In addition, we can construct an ε_1-net N_{ε_1} over $[0, \frac{1}{1-\gamma}]$ whose size satisfies $|N_{\varepsilon_1}| \leq \frac{3}{\varepsilon_1(1-\gamma)}$ (Vershynin, 2018). By the union bound and (137), it holds with probability at least $1 - \frac{\delta}{SA}$ that for all $\alpha \in N_{\varepsilon_1}$,

$$g_{s,a}(\alpha, V) \leq \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{3N(1-\gamma)}. \qquad (138)$$

Combined with (136), it yields that,

$$\begin{aligned}\left|\widehat{P}^{\pi,V}_{s,a} V - P^{\pi,V}_{s,a} V\right| &\leq \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right| \\ &\overset{\text{(i)}}{\leq} \varepsilon_1 + \sup_{\alpha \in N_{\varepsilon_1}} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right| \\ &\overset{\text{(ii)}}{\leq} \varepsilon_1 + \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{3N(1-\gamma)} \qquad (139) \\ &\overset{\text{(iii)}}{\leq} \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N(1-\gamma)} \\ &\overset{\text{(iv)}}{\leq} 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)} \qquad (140) \\ &\leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\|V\|_\infty + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)} \\ &\leq 3\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}} \qquad (141)\end{aligned}$$

where (i) follows from that the optimal $\alpha^\star$ falls into the ε_1-ball centered around some point inside N_{ε_1} and $g_{s,a}(\alpha, V)$ is 1-Lipschitz, (ii) holds by (138), (iii) arises from taking $\varepsilon_1 = \frac{\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{3N(1-\gamma)}$, (iv) is verified by $|N_{\varepsilon_1}| \le \frac{3}{\varepsilon_1(1-\gamma)} \le 9N$, and the last inequality is due to the fact $\|V^{\star,\sigma}\|_\infty \le \frac{1}{1-\gamma}$ and letting $N \ge \log(\frac{18SAN}{\delta})$.

To continue, applying (140) and (141) with $\pi = \pi^\star$ and $V = V^{\star,\sigma}$ (independent with $\widehat{P}^0$) and taking the union bound over $(s,a) \in \mathcal{S} \times \mathcal{A}$ gives that with probability at least $1-\delta$, it holds simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$\begin{aligned}\left|\widehat{P}^{\pi^\star,V}_{s,a} V^{\star,\sigma} - P^{\pi^\star,V}_{s,a} V^{\star,\sigma}\right| &\le 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(V^{\star,\sigma})} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)} \\ &\le 3\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}. \end{aligned} \tag{142}$$

By converting (142) to the matrix form, one has with probability at least $1-\delta$,

$$\begin{aligned}\left|\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right| &\le 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)}1 \\ &\le 3\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}1. \end{aligned} \tag{143}$$

B.6 Proof of Lemma 12

Following the same argument as (105), it follows

$$\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})} = \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(\widehat{\underline{P}}^{\pi^\star,V}\right)^t \mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})}. \tag{144}$$

To continue, we first focus on controlling $\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})$. Towards this, denoting the minimum value of $V^{\star,\sigma}$ as $V_{\min} := \min_{s\in\mathcal{S}} V^{\star,\sigma}(s)$ and $V' := V^{\star,\sigma} - V_{\min}1$, we arrive at (see the robust Bellman's consistency equation in (46))

$$\begin{aligned} V' = V^{\star,\sigma} - V_{\min}1 &= r_{\pi^\star} + \gamma\underline{P}^{\pi^\star,V}V^{\star,\sigma} - V_{\min}1 \\ &= r_{\pi^\star} + \gamma\widehat{\underline{P}}^{\pi^\star,V}V^{\star,\sigma} + \gamma\Big(\underline{P}^{\pi^\star,V} - \widehat{\underline{P}}^{\pi^\star,V}\Big)V^{\star,\sigma} - V_{\min}1 \\ &= r_{\pi^\star} - (1-\gamma)V_{\min}1 + \gamma\widehat{\underline{P}}^{\pi^\star,V}V' + \gamma\Big(\underline{P}^{\pi^\star,V} - \widehat{\underline{P}}^{\pi^\star,V}\Big)V^{\star,\sigma} \\ &= r'_{\pi^\star} + \gamma\widehat{\underline{P}}^{\pi^\star,V}V' + \gamma\Big(\underline{P}^{\pi^\star,V} - \widehat{\underline{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}, \end{aligned} \tag{145}$$

where the last line holds by letting $r'_{\pi^\star} := r_{\pi^\star} - (1-\gamma)V_{\min}1 \le r_{\pi^\star}$. With the above fact in hand, we control $\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})$ as follows:

$$\begin{aligned} \mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma}) &\overset{\text{(i)}}{=} \mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V') = \widehat{\underline{P}}^{\pi^\star,V}(V'\circ V') - \big(\widehat{\underline{P}}^{\pi^\star,V}V'\big)\circ\big(\widehat{\underline{P}}^{\pi^\star,V}V'\big) \\ &\overset{\text{(ii)}}{=} \widehat{\underline{P}}^{\pi^\star,V}(V'\circ V') - \frac{1}{\gamma^2}\Big(V' - r'_{\pi^\star} - \gamma\Big(\underline{P}^{\pi^\star,V} - \widehat{\underline{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}\Big)^{\circ 2} \\ &= \widehat{\underline{P}}^{\pi^\star,V}(V'\circ V') - \frac{1}{\gamma^2}V'\circ V' + \frac{2}{\gamma^2}V'\circ\Big(r'_{\pi^\star} + \gamma\Big(\underline{P}^{\pi^\star,V} - \widehat{\underline{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}\Big) \\ &\quad - \frac{1}{\gamma^2}\Big(r'_{\pi^\star} + \gamma\Big(\underline{P}^{\pi^\star,V} - \widehat{\underline{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}\Big)^{\circ 2} \\ &\overset{\text{(iii)}}{\le} \widehat{\underline{P}}^{\pi^\star,V}(V'\circ V') - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1 + \frac{2}{\gamma}\|V'\|_\infty\Big|\Big(\underline{P}^{\pi^\star,V} - \widehat{\underline{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}\Big| \end{aligned} \tag{146}$$

$$\le \underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1 + \frac{6}{\gamma}\|V'\|_\infty\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}1, \tag{147}$$

where (i) holds by the fact that $\mathrm{Var}_{P_\pi}(V-b1)=\mathrm{Var}_{P_\pi}(V)$ for any scalar b and $V\in\mathbb{R}^S$, (ii) follows from (145), (iii) arises from $\frac{1}{\gamma^2}V'\circ V'\ge\frac{1}{\gamma}V'\circ V'$ and $-1\le r_{\pi^\star}-(1-\gamma)V_{\min}1=r'_{\pi^\star}\le r_{\pi^\star}\le 1$, and the last inequality holds by Lemma 11.

Plugging (147) into (144) leads to

$$\begin{aligned}
&\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})}\\
&\le\sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\left(\underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'+\frac{2}{\gamma^2}\|V'\|_\infty 1+\frac{6}{\gamma}\|V'\|_\infty\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}1\right)}\\
&\overset{\text{(i)}}{\le}\sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\left(\underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'\right)\right|}\\
&\quad+\sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\left(\frac{2}{\gamma^2}\|V'\|_\infty 1+\frac{6}{\gamma}\|V'\|_\infty\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}1\right)}\\
&\le\sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\left[\underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'\right]\right|}+\sqrt{\frac{\left(2+6\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\right)\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1,
\end{aligned}\tag{148}$$

where (i) holds by the triangle inequality. Therefore, the remainder of the proof shall focus on the first term, which follows

$$\begin{aligned}
&\left|\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\left(\underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'\right)\right|\\
&=\left|\left(\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^{t+1}-\sum_{t=0}^{\infty}\gamma^{t-1}\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\right)\left(V'\circ V'\right)\right|\le\frac{1}{\gamma}\|V'\|_\infty^2 1
\end{aligned}\tag{149}$$

by recursion. Inserting (149) back to (148) leads to

$$\begin{aligned}
&\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})}\\
&\le\sqrt{\frac{\|V'\|_\infty^2}{\gamma(1-\gamma)}}1+3\sqrt{\frac{\left(1+\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\right)\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1\\
&\le 4\sqrt{\frac{\left(1+\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\right)\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1\le 4\sqrt{\frac{\left(1+\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}}}1\le 4\sqrt{\frac{\left(1+\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\right)}{\gamma^3(1-\gamma)^3}}1,
\end{aligned}\tag{150}$$

where the penultimate inequality follows from applying Lemma 6 with $P=P^0$ and $\pi=\pi^\star$:

$$\|V'\|_\infty=\max_{s\in\mathcal{S}}V^{\star,\sigma}(s)-\min_{s\in\mathcal{S}}V^{\star,\sigma}(s)\le\frac{1}{\gamma\max\{1-\gamma,\sigma\}}.$$

B.7 Proof of Lemma 14

To begin with, for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, invoking the results in (136), we have

$$\left|\widehat{P}^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma}-P^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma}\right|\le\max_{\alpha\in[\min_s\widehat{V}^{\widehat{\pi},\sigma}(s),\max_s\widehat{V}^{\widehat{\pi},\sigma}(s)]}\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right|$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left(\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right| + \left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha - \left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)\right|\right)\\
&\leq \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left(\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right| + \left\|P^0_{s,a}-\widehat{P}^0_{s,a}\right\|_1 \left\|\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha - \left[\widehat{V}^{\star,\sigma}\right]_\alpha\right\|_\infty\right)\\
&\overset{\text{(ii)}}{\leq} \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right| + 2\left\|\widehat{V}^{\widehat{\pi},\sigma} - \widehat{V}^{\star,\sigma}\right\|_\infty\\
&\overset{\text{(iii)}}{\leq} \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right| + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}, \qquad (151)
\end{aligned}
$$

where (i) holds by the triangle inequality, and (ii) follows from $\left\|P^0_{s,a}-\widehat{P}^0_{s,a}\right\|_1 \leq 2$ and $\left\|\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha - \left[\widehat{V}^{\star,\sigma}\right]_\alpha\right\|_\infty \leq \left\|\widehat{V}^{\widehat{\pi},\sigma} - \widehat{V}^{\star,\sigma}\right\|_\infty$, and (iii) follows from (50).

To control $\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right|$ in (151) for any given $\alpha \in \left[0, \frac{1}{1-\gamma}\right]$, and tame the dependency between $\widehat{V}^{\star,\sigma}$ and $\widehat{P}^0$, we resort to the following leave-one-out argument motivated by (Agarwal et al., 2020; Li et al., 2022b; Shi and Chi, 2022). Specifically, we first construct a set of auxiliary RMDPs which simultaneously have the desired statistical independence between robust value functions and the estimated nominal transition kernel, and are minimally different from the original RMDPs under consideration. Then we control the term of interest associated with these auxiliary RMDPs and show the value is close to the target quantity for the desired RMDP. The process is divided into several steps as below.

Step 1: construction of auxiliary RMDPs with deterministic empirical nominal transitions. Recall that we target the empirical infinite-horizon robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}$ with the nominal transition kernel $\widehat{P}^0$. Towards this, we can construct an auxiliary robust MDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ for each state s and any non-negative scalar $u \geq 0$, so that it is the same as $\widehat{\mathcal{M}}_{\mathsf{rob}}$ except for the transition properties in state s. In particular, we define the nominal transition kernel and reward function of $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ as $P^{s,u}$ and $r^{s,u}$, which are expressed as follows

$$
\begin{cases}
P^{s,u}(s' \,|\, s,a) = \mathbb{1}(s'=s) & \text{for all } (s',a)\in\mathcal{S}\times\mathcal{A},\\
P^{s,u}(\cdot \,|\, \widetilde{s},a) = \widehat{P}^0(\cdot \,|\, \widetilde{s},a) & \text{for all } (\widetilde{s},a)\in\mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s}\neq s,
\end{cases}
\qquad (152)
$$

and

$$
\begin{cases}
r^{s,u}(s,a) = u & \text{for all } a\in\mathcal{A},\\
r^{s,u}(\widetilde{s},a) = r(\widetilde{s},a) & \text{for all } (\widetilde{s},a)\in\mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s}\neq s.
\end{cases}
\qquad (153)
$$

It is evident that the nominal transition probability at state s of the auxiliary $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$, i.e. it never leaves state s once entered. This useful property removes the randomness of $\widehat{P}^0_{s,a}$ for all $a\in\mathcal{A}$ in state s, which will be leveraged later.

Correspondingly, the robust Bellman operator $\widehat{\mathcal{T}}^{\sigma}_{s,u}(\cdot)$ associated with the RMDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is defined as

$$
\forall(\tilde{s},a)\in\mathcal{S}\times\mathcal{A}: \quad \widehat{\mathcal{T}}^{\sigma}_{s,u}(Q)(\tilde{s},a) = r^{s,u}(\tilde{s},a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{s,u}_{\tilde{s},a})} \mathcal{P}V, \qquad \text{with } V(\tilde{s}) = \max_a Q(\tilde{s},a). \qquad (154)
$$

Step 2: fixed-point equivalence between $\widehat{\mathcal{M}}_{\mathsf{rob}}$ and the auxiliary RMDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$. Recall that $\widehat{Q}^{\star,\sigma}$ is the unique fixed point of $\widehat{\mathcal{T}}^{\sigma}(\cdot)$ with the corresponding robust value $\widehat{V}^{\star,\sigma}$. We assert that the corresponding robust value function $\widehat{V}^{\star,\sigma}_{s,u^\star}$ obtained from the fixed point of $\widehat{\mathcal{T}}^{\sigma}_{s,u}(\cdot)$ aligns with the robust value function $\widehat{V}^{\star,\sigma}$ derived from $\widehat{\mathcal{T}}^{\sigma}(\cdot)$, as long as we choose u in the following manner:

$$
u^\star := u^\star(s) = \widehat{V}^{\star,\sigma}(s) - \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma}. \qquad (155)
$$

where e_s is the s-th standard basis vector in $\mathbb{R}^S$. Towards verifying this, we shall break our arguments in two different cases.

- **For state** s: One has for any $a \in \mathcal{A}$:

$$
\begin{aligned}
r^{s,u^\star}(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^{s,u^\star}_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma} &= u^\star + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma} \\
&= \widehat{V}^{\star,\sigma}(s) - \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma} + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma} = \widehat{V}^{\star,\sigma}(s), \qquad (156)
\end{aligned}
$$

  where the first equality follows from the definition of $P^{s,u^\star}_{s,a}$ in (152), and the second equality follows from plugging in the definition of $u^\star$ in (155).

- **For state** $s' \neq s$: It is easily verified that for all $a \in \mathcal{A}$,

$$
\begin{aligned}
r^{s,u^\star}(s',a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^{s,u^\star}_{s',a})} \mathcal{P}\widehat{V}^{\star,\sigma} &= r(s',a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s',a})} \mathcal{P}\widehat{V}^{\star,\sigma} \\
&= \widehat{\mathcal{T}}^\sigma(\widehat{Q}^{\star,\sigma})(s',a) = \widehat{Q}^{\star,\sigma}(s',a), \qquad (157)
\end{aligned}
$$

  where the first equality follows from the definitions in (153) and (152), and the last line arises from the definition of the robust Bellman operator in (15), and that $\widehat{Q}^{\star,\sigma}$ is the fixed point of $\widehat{\mathcal{T}}^\sigma(\cdot)$ (see Lemma 4).

Combining the facts in the above two cases, we establish that there exists a fixed point $\widehat{Q}^{\star,\sigma}_{s,u^\star}$ of the operator $\widehat{\mathcal{T}}^\sigma_{s,u^\star}(\cdot)$ by taking

$$
\begin{cases}
\widehat{Q}^{\star,\sigma}_{s,u^\star}(s,a) = \widehat{V}^{\star,\sigma}(s) & \text{for all } a \in \mathcal{A}, \\
\widehat{Q}^{\star,\sigma}_{s,u^\star}(s',a) = \widehat{Q}^{\star,\sigma}(s',a) & \text{for all } s' \neq s \text{ and } a \in \mathcal{A}.
\end{cases} \qquad (158)
$$

Consequently, we confirm the existence of a fixed point of the operator $\widehat{\mathcal{T}}^\sigma_{s,u^\star}(\cdot)$. In addition, its corresponding value function $\widehat{V}^{\star,\sigma}_{s,u^\star}$ also coincides with $\widehat{V}^{\star,\sigma}$. Note that the corresponding facts between $\widehat{\mathcal{M}}_{\mathsf{rob}}$ and $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ in Step 1 and step 2 holds in fact for any uncertainty set.

Step 3: building an ε-net for all reward values u. It is easily verified that

$$
0 \leq u^\star \leq \widehat{V}^{\star,\sigma}(s) \leq \frac{1}{1-\gamma}. \qquad (159)
$$

We can construct a N_{ε_2}-net over the interval $\left[0, \frac{1}{1-\gamma}\right]$, where the size is bounded by $|N_{\varepsilon_2}| \leq \frac{3}{\varepsilon_2(1-\gamma)}$ (Vershynin, 2018). Following the same arguments in the proof of Lemma 4, we can demonstrate that for each $u \in N_{\varepsilon_2}$, there exists a unique fixed point $\widehat{Q}^{\star,\sigma}_{s,u}$ of the operator $\widehat{\mathcal{T}}^\sigma_{s,u}(\cdot)$, which satisfies $0 \leq \widehat{Q}^{\star,\sigma}_{s,u} \leq \frac{1}{1-\gamma} \cdot 1$. Consequently, the corresponding robust value function also satisfies $\left\|\widehat{V}^{\star,\sigma}_{s,u}\right\|_\infty \leq \frac{1}{1-\gamma}$.

By the definitions in (152) and (153), we observe that for all $u \in N_{\varepsilon_2}$, $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is statistically independent from $\widehat{P}^0_{s,a}$. This independence indicates that $[\widehat{V}^{\star,\sigma}_{s,u}]_\alpha$ and $\widehat{P}^0_{s,a}$ are independent for a fixed α. With this in mind, invoking the fact in (140) and (141) and taking the union bound over all $(s,a,\alpha) \in \mathcal{S} \times \mathcal{A} \times N_{\varepsilon_1}$, $u \in N_{\varepsilon_2}$ yields that, with probability at least $1-\delta$, it holds for all $(s,a,u) \in \mathcal{S} \times \mathcal{A} \times N_{\varepsilon_2}$ that

$$
\begin{aligned}
\max_{\alpha\in[0,1/(1-\gamma)]} \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) [\widehat{V}^{\star,\sigma}_{s,u}]_\alpha \right| &\leq \varepsilon_2 + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,u})} + \frac{2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} \\
&\leq \varepsilon_2 + 3\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2 N}}, \qquad (160)
\end{aligned}
$$

where the last inequality holds by the fact $\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,u}) \leq \|\widehat{V}^{\star,\sigma}_{s,u}\|_\infty \leq \frac{1}{1-\gamma}$ and letting $N \geq \log\left(\frac{18SAN|N_{\varepsilon_2}|}{\delta}\right)$.

Step 4: uniform concentration. Recalling that $u^\star \in \big[0, \frac{1}{1-\gamma}\big]$ (see (159)), we can always find some $\overline{u} \in N_{\varepsilon_2}$ such that $|\overline{u} - u^\star| \le \varepsilon_2$. Consequently, plugging in the operator $\widehat{\mathcal{T}}^{\sigma}_{s,u}(\cdot)$ in (154) yields

$$\forall Q \in \mathbb{R}^{SA}: \quad \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\overline{u}}(Q) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(Q)\right\|_\infty = |\overline{u} - u^\star| \le \varepsilon_2$$

With this in mind, we observe that the fixed points of $\widehat{\mathcal{T}}^{\sigma}_{s,\overline{u}}(\cdot)$ and $\widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\cdot)$ obey

$$\begin{aligned}
\left\|\widehat{Q}^{\star,\sigma}_{s,\overline{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty &= \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\overline{u}}(\widehat{Q}^{\star,\sigma}_{s,\overline{u}}) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty \\
&\le \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\overline{u}}(\widehat{Q}^{\star,\sigma}_{s,\overline{u}}) - \widehat{\mathcal{T}}^{\sigma}_{s,\overline{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty + \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\overline{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star}) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty \\
&\le \gamma \left\|\widehat{Q}^{\star,\sigma}_{s,\overline{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty + \varepsilon_2,
\end{aligned}$$

where the last inequality holds by the fact that $\widehat{\mathcal{T}}^{\sigma}_{s,u}(\cdot)$ is a γ-contraction. It directly indicates that

$$\left\|\widehat{Q}^{\star,\sigma}_{s,\overline{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \frac{\varepsilon_2}{(1-\gamma)} \quad \text{and} \quad \left\|\widehat{V}^{\star,\sigma}_{s,\overline{u}} - \widehat{V}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \left\|\widehat{Q}^{\star,\sigma}_{s,\overline{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \frac{\varepsilon_2}{(1-\gamma)}. \tag{161}$$

Armed with the above facts, to control the first term in (151), invoking the identity $\widehat{V}^{\star,\sigma} = \widehat{V}^{\star,\sigma}_{s,u^\star}$ established in Step 2 gives that: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
&\max_{\alpha \in [\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}]_\alpha\right| \\
&\le \max_{\alpha \in [0, 1/(1-\gamma)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}]_\alpha\right| = \max_{\alpha \in [0, 1/(1-\gamma)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha\right| \\
&\overset{\text{(i)}}{\le} \max_{\alpha \in [0, 1/(1-\gamma)]} \left\{\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha\right| + \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left([\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha - [\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha\right)\right|\right\} \\
&\overset{\text{(ii)}}{\le} \max_{\alpha \in [0, 1/(1-\gamma)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha\right| + \frac{2\varepsilon_2}{(1-\gamma)} \\
&\overset{\text{(iii)}}{\le} \frac{2\varepsilon_2}{(1-\gamma)} + \varepsilon_2 + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,\overline{u}})} + \frac{2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} \\
&\le \frac{3\varepsilon_2}{(1-\gamma)} + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + \frac{2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} \\
&\qquad + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}}\sqrt{\left|\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,\overline{u}})\right|} \\
&\overset{\text{(iv)}}{\le} \frac{3\varepsilon_2}{(1-\gamma)} + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + \frac{2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} + 2\sqrt{\frac{2\varepsilon_2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N(1-\gamma)^2}} \\
&\le 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + \frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)} \qquad (162) \\
&\le 10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}}, \qquad (163)
\end{aligned}$$

where (i) holds by the triangle inequality, (ii) arises from (the last inequality holds by (161))

$$\begin{aligned}
\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left([\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha - [\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha\right)\right| &\le \left\|P^0_{s,a} - \widehat{P}^0_{s,a}\right\|_1 \left\|[\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha - [\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha\right\|_\infty \\
&\le 2\left\|\widehat{V}^{\star,\sigma}_{s,\overline{u}} - \widehat{V}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \frac{2\varepsilon_2}{(1-\gamma)},
\end{aligned} \tag{164}$$

(iii) follows from (160), (iv) can be verified by applying Lemma 3 with (161). Here, the penultimate inequality holds by letting $\varepsilon_2 = \frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}$, which leads to $|N_{\varepsilon_2}| \le \frac{3}{\varepsilon_2(1-\gamma)} \le \frac{3N}{1-\gamma}$, and the last inequality holds by the fact $\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}) \le \|\widehat{V}^{\star,\sigma}\|_\infty \le \frac{1}{1-\gamma}$ and letting $N \ge \log\left(\frac{54SAN^2}{(1-\gamma)\delta}\right)$.

Step 5: finishing up. Inserting (162) and (163) back into (151) and combining with (163) give that with probability at least $1-\delta$,

$$\begin{aligned}
\left|\widehat{P}^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma} - P^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma}\right| &\le \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}]_\alpha\right| + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} \\
&\le \max_{\alpha\in[0,1/(1-\gamma)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}]_\alpha\right| + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} \\
&\le 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + \frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} \\
&\le 10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}
\end{aligned} \tag{165}$$

holds for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

Finally, we complete the proof by compiling everything into the matrix form as follows:

$$\begin{aligned}
\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right| &\le 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})}1 + \frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)}1 + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1 \\
&\le 10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}}1 + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1.
\end{aligned} \tag{166}$$

B.8 Proof of Lemma 15

The proof can be achieved by directly applying the same routine as Appendix B.6. Towards this, similar to (144), we arrive at

$$\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \le \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(\underline{P}^{\widehat{\pi},\widehat{V}}\right)^t\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})}. \tag{167}$$

To control $\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})$, we denote the minimum value of $\widehat{V}^{\widehat{\pi},\sigma}$ as $V_{\min} = \min_{s\in\mathcal{S}}\widehat{V}^{\widehat{\pi},\sigma}(s)$ and $V' := \widehat{V}^{\widehat{\pi},\sigma} - V_{\min}1$. By the same argument as (146), we arrive at

$$\begin{aligned}
&\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma}) \\
&\le \underline{P}^{\widehat{\pi},\widehat{V}}\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1 + \frac{2}{\gamma}\|V'\|_\infty\left|\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}} - \underline{P}^{\widehat{\pi},\widehat{V}}\right)\widehat{V}^{\widehat{\pi},\sigma}\right| \\
&\le \underline{P}^{\widehat{\pi},\widehat{V}}\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1 + \frac{2}{\gamma}\|V'\|_\infty\left(10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)1,
\end{aligned} \tag{168}$$

where the last inequality makes use of Lemma 14. Plugging (168) back into (167) leads to

$$\begin{aligned}
\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} &\overset{\text{(i)}}{\le} \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t\left(\underline{P}^{\widehat{\pi},\widehat{V}}\right)^t\left(\underline{P}^{\widehat{\pi},\widehat{V}}\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V'\right)\right|} \\
&\quad + \sqrt{\frac{1}{(1-\gamma)^2\gamma^2}\left(2 + 20\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)\|V'\|_\infty 1}
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{\leq} \sqrt{\frac{\|V'\|_\infty^2}{\gamma(1-\gamma)}}1 + \sqrt{\frac{\left(2 + 20\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1 \\
&\overset{\text{(iii)}}{\leq} \sqrt{\frac{\|V'\|_\infty^2}{\gamma(1-\gamma)}}1 + \sqrt{\frac{24\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1 \leq 6\sqrt{\frac{\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1, \qquad (169)
\end{aligned}
$$

where (i) arises from following the routine of (148), (ii) holds by repeating the argument of (149), (iii) follows by taking $N \geq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2}$ and $\varepsilon_{\mathsf{opt}} \leq \frac{1-\gamma}{\gamma}$, and the last inequality holds by $\|V'\|_\infty \leq \|V^{\star,\sigma}\|_\infty \leq \frac{1}{1-\gamma}$.

Finally, applying Lemma 6 with $P = \widehat{P}^0$ and $\pi = \widehat{\pi}$ yields

$$
\|V'\|_\infty \leq \max_{s\in\mathcal{S}} \widehat{V}^{\widehat{\pi},\sigma}(s) - \min_{s\in\mathcal{S}} \widehat{V}^{\widehat{\pi},\sigma}(s) \leq \frac{1}{\gamma\max\{1-\gamma,\sigma\}},
$$

which can be inserted into (169) and gives

$$
\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \leq 6\sqrt{\frac{1}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}}}1 \leq 6\sqrt{\frac{1}{(1-\gamma)^3\gamma^2}}1.
$$

C Proof of the auxiliary facts for Theorem 2

C.1 Proof of Lemma 10

Deriving the robust value function over different states. For any $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$, we first characterize the robust value function of any policy π over different states. Before proceeding, we denote the minimum of the robust value function over states as below:

$$
V_{\phi,\min}^{\pi,\sigma} \coloneqq \min_{s\in\mathcal{S}} V_\phi^{\pi,\sigma}(s). \qquad (170)
$$

Clearly, there exists at least one state $s_{\phi,\min}^\pi$ that satisfies $V_\phi^{\pi,\sigma}(s_{\phi,\min}^\pi) = V_{\phi,\min}^{\pi,\sigma}$.

With this in mind, it is easily observed that for any policy π, the robust value function at state $s = 1$ obeys

$$
\begin{aligned}
V_\phi^{\pi,\sigma}(1) &= \mathbb{E}_{a\sim\pi(\cdot\,|\,1)}\left[r(1,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{1,a}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma}\right] \\
&\overset{\text{(i)}}{=} 1 + \gamma\mathbb{E}_{a\sim\pi(\cdot\,|\,1)}\left[\underline{P}^\phi(1\,|\,1,a)V_\phi^{\pi,\sigma}(1)\right] + \gamma\sigma V_{\phi,\min}^{\pi,\sigma} \overset{\text{(ii)}}{=} 1 + \gamma(1-\sigma)V_\phi^{\pi,\sigma}(1) + \gamma\sigma V_{\phi,\min}^{\pi,\sigma}, \qquad (171)
\end{aligned}
$$

where (i) holds by $r(1,a) = 1$ for all $a \in \mathcal{A}'$ and (69), and (ii) follows from $P^\phi(1\,|\,1,a) = 1$ for all $a \in \mathcal{A}'$.

Similarly, for any $s \in \{2,3,\cdots,S-1\}$, we have

$$
\begin{aligned}
V_\phi^{\pi,\sigma}(s) &= 0 + \gamma\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[\underline{P}^\phi(1\,|\,s,a)V_\phi^{\pi,\sigma}(1)\right] + \gamma\sigma V_{\phi,\min}^{\pi,\sigma} \\
&= \gamma\left(1-\sigma\right)V_\phi^{\pi,\sigma}(1) + \gamma\sigma V_{\phi,\min}^{\pi,\sigma}, \qquad (172)
\end{aligned}
$$

since $r(s,a) = 0$ for all $s \in \{2,3,\cdots,S-1\}$ and the definition in (69).

Finally, we move onto compute $V_\phi^{\pi,\sigma}(0)$, the robust value function at state 0 associated with any policy π. First, it obeys

$$
\begin{aligned}
V_\phi^{\pi,\sigma}(0) &= \mathbb{E}_{a\sim\pi(\cdot\,|\,0)}\left[r(0,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,a}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma}\right] \\
&= 0 + \gamma\pi(\phi\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,\phi}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma} + \gamma\pi(1-\phi\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,1-\phi}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma}. \qquad (173)
\end{aligned}
$$

Recall the transition kernel defined in (61) and the fact about the uncertainty set over state 0 in (70), it is easily verified that the following probability vector $P_1 \in \Delta(\mathcal{S})$ obeys $P_1 \in \mathcal{U}^\sigma(P^\phi_{0,\phi})$, which is defined as

$$\begin{aligned} P_1(0) &= 1 - p + \sigma \mathbb{1}\left(0 = s^\pi_{\phi,\min}\right), \qquad P_1(1) = \underline{p} = p - \sigma, \\ P_1(s) &= \sigma \mathbb{1}\left(s = s^\pi_{\phi,\min}\right), \qquad \forall s \in \{2, 3, \cdots, S-1\}, \end{aligned} \tag{174}$$

where $\underline{p} = p - \sigma$ due to (70). Similarly, the following probability vector $P_2 \in \Delta(\mathcal{S})$ also falls into the uncertainty set $\mathcal{U}^\sigma(P^\phi_{0,1-\phi})$:

$$\begin{aligned} P_2(0) &= 1 - q + \sigma \mathbb{1}\left(0 = s^\pi_{\phi,\min}\right), \qquad P_2(1) = \underline{q} = q - \sigma, \\ P_2(s) &= \sigma \mathbb{1}\left(0 = s^\pi_{\phi,\min}\right) \qquad \forall s \in \{2, 3, \cdots, S-1\}. \end{aligned} \tag{175}$$

It is noticed that P_0 and P_1 defined above are the worst-case perturbations, since the probability mass at state 1 will be moved to the state with the least value. Plugging the above facts about $P_1 \in \mathcal{U}^\sigma(P^\phi_{0,\phi})$ and $P_2 \in \mathcal{U}^\sigma(P^\phi_{0,1-\phi})$ into (173), we arrive at

$$\begin{aligned} V^{\pi,\sigma}_\phi(0) &\leq \gamma\pi(\phi \,|\, 0) P_1 V^{\pi,\sigma}_\phi + \gamma\pi(1-\phi \,|\, 0) P_2 V^{\pi,\sigma}_\phi \\ &= \gamma\pi(\phi \,|\, 0)\Big[(p-\sigma) V^{\pi,\sigma}_\phi(1) + (1-p) V^{\pi,\sigma}_\phi(0) + \sigma V^{\pi,\sigma}_{\phi,\min}\Big] \\ &\quad + \gamma\pi(1-\phi \,|\, 0)\Big[(q-\sigma) V^{\pi,\sigma}_\phi(1) + (1-q) V^{\pi,\sigma}_\phi(0) + \sigma V^{\pi,\sigma}_{\phi,\min}\Big] \\ &\overset{\text{(i)}}{=} \gamma\left(z^\pi_\phi - \sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1 - z^\pi_\phi) V^{\pi,\sigma}_\phi(0), \end{aligned} \tag{176}$$

where the last equality holds by the definition of z^π_ϕ in (72). To continue, recursively applying (176) yields

$$\begin{aligned} V^{\pi,\sigma}_\phi(0) &\leq \gamma\left(z^\pi_\phi - \sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1-z^\pi_\phi)\Big[\gamma\left(z^\pi_\phi - \sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1-z^\pi_\phi) V^{\pi,\sigma}_\phi(0)\Big] \\ &\overset{\text{(i)}}{\leq} \gamma\left(z^\pi_\phi - \sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1-z^\pi_\phi)\Big[\gamma z^\pi_\phi V^{\pi,\sigma}_\phi(1) + \gamma(1-z^\pi_\phi) V^{\pi,\sigma}_\phi(0)\Big] \\ &\leq \ldots \\ &\leq \gamma\left(z^\pi_\phi - \sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma z^\pi_\phi \sum_{t=1}^{\infty} \gamma^t (1-z^\pi_\phi)^t V^{\pi,\sigma}_\phi(1) + \lim_{t\to\infty} \gamma^t (1-z^\pi_\phi)^t V^{\pi,\sigma}_\phi(0) \\ &\overset{\text{(ii)}}{\leq} \gamma\left(z^\pi_\phi - \sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1-z^\pi_\phi)\frac{\gamma z^\pi_\phi}{1-\gamma(1-z^\pi_\phi)} V^{\pi,\sigma}_\phi(1) + 0 \\ &< \gamma\left(z^\pi_\phi - \sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1-z^\pi_\phi) V^{\pi,\sigma}_\phi(1) \\ &= \gamma\left(1 - \sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min}, \end{aligned} \tag{177}$$

where (i) uses $V^{\pi,\sigma}_{\phi,\min} \leq V^{\pi,\sigma}_\phi(1)$, (ii) follows from $\gamma(1 - z^\pi_\phi) < 1$, and the penultimate line follows from the trivial fact that $\frac{\gamma z^\pi_\phi}{1-\gamma(1-z^\pi_\phi)} < 1$.

Combining (171), (172), and (177), we have that for any policy π,

$$V^{\pi,\sigma}_\phi(0) = V^{\pi,\sigma}_{\phi,\min}, \tag{178}$$

which directly leads to

$$V^{\pi,\sigma}_\phi(1) = 1 + \gamma\left(1-\sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} = \frac{1 + \gamma\sigma V^{\pi,\sigma}_\phi(0)}{1 - \gamma\left(1-\sigma\right)}. \tag{179}$$

Let's now return to the characterization of $V^{\pi,\sigma}_\phi(0)$. In view of (178), the equality in (176) holds, and we have

$$V^{\pi,\sigma}_\phi(0) = \gamma\left(z^\pi_\phi - \sigma\right) V^{\pi,\sigma}_\phi(1) + \gamma\left(1 - z^\pi_\phi + \sigma\right) V^{\pi,\sigma}_\phi(0)$$

$$
\begin{aligned}
&\overset{\text{(i)}}{=} \gamma\left(z_\phi^\pi - \sigma\right)\frac{1+\gamma\sigma V_\phi^{\pi,\sigma}(0)}{1-\gamma\left(1-\sigma\right)} + \gamma\left(1 - z_\phi^\pi + \sigma\right)V_\phi^{\pi,\sigma}(0)\\
&= \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma\left(1-\sigma\right)} + \gamma\left(1 + \left(z_\phi^\pi - \sigma\right)\frac{\gamma\sigma - \left(1-\gamma\left(1-\sigma\right)\right)}{1-\gamma\left(1-\sigma\right)}\right)V_\phi^{\pi,\sigma}(0)\\
&= \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma\left(1-\sigma\right)} + \gamma\left(1 - \frac{(1-\gamma)\left(z_\phi^\pi - \sigma\right)}{1-\gamma\left(1-\sigma\right)}\right)V_\phi^{\pi,\sigma}(0),
\end{aligned}
$$

where (i) arises from (179). Solving this relation gives

$$
V_\phi^{\pi,\sigma}(0) = \frac{\frac{\gamma\left(z_\phi^\pi-\sigma\right)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1+\frac{\gamma\left(z_\phi^\pi-\sigma\right)}{1-\gamma(1-\sigma)}\right)}. \tag{180}
$$

The optimal robust policy and optimal robust value function. We move on to characterize the robust optimal policy and its corresponding robust value function. To begin with, denoting

$$
z \coloneqq \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma\left(1-\sigma\right)}, \tag{181}
$$

we rewrite (180) as

$$
V_\phi^{\pi,\sigma}(0) = \frac{z}{(1-\gamma)(1+z)} \eqqcolon f(z).
$$

Plugging in the fact that $z_\phi^\pi \geq q \geq \sigma > 0$ in (68), it follows that $z > 0$. So for any $z > 0$, the derivative of $f(z)$ w.r.t. z obeys

$$
\frac{(1-\gamma)(1+z) - (1-\gamma)z}{(1-\gamma)^2(1+z)^2} = \frac{1}{(1-\gamma)(1+z)^2} > 0. \tag{182}
$$

Observing that $f(z)$ is increasing in z, z is increasing in z_ϕ^π, and z_ϕ^π is also increasing in $\pi(\phi\,|\,0)$ (see the fact $p \geq q$ in (68)), the optimal policy in state 0 thus obeys

$$
\pi_\phi^\star(\phi\,|\,0) = 1. \tag{183}
$$

Considering that the action does not influence the state transition for all states $s > 0$, without loss of generality, we choose the robust optimal policy to obey

$$
\forall s > 0: \quad \pi_\phi^\star(\phi\,|\,s) = 1. \tag{184}
$$

Taking $\pi = \pi_\phi^\star$, we complete the proof by showing that the corresponding robust optimal robust value function at state 0 as follows:

$$
V_\phi^{\star,\sigma}(0) = \frac{\frac{\gamma\left(z_\phi^{\pi^\star}-\sigma\right)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1+\frac{\gamma\left(z_\phi^{\pi^\star}-\sigma\right)}{1-\gamma(1-\sigma)}\right)} = \frac{\frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1+\frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}\right)}. \tag{185}
$$

C.2 Proof of the claim (75)

Plugging in the definition of φ, we arrive at that for any policy π,

$$
\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma}\right\rangle = V_\phi^{\star,\sigma}(0) - V_\phi^{\pi,\sigma}(0) = \frac{\frac{\gamma\left(p-z_\phi^\pi\right)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1+\frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}\right)\left(1+\frac{\gamma\left(z_\phi^\pi-\sigma\right)}{1-\gamma(1-\sigma)}\right)}, \tag{186}
$$

which follows from applying (71) and basic calculus. Then, we proceed to control the above term in two cases separately in terms of the uncertainty level σ.

- When $\sigma \in (0, 1-\gamma]$. Then regarding the important terms in (186), we observe that

$$1-\gamma < 1-\gamma\,(1-\sigma) \leq 1-\gamma\,(1-(1-\gamma)) = (1-\gamma)(1+\gamma) \leq 2(1-\gamma), \tag{187}$$

which directly leads to

$$\frac{\gamma\big(z_\phi^\pi - \sigma\big)}{1-\gamma\,(1-\sigma)} \overset{\text{(i)}}{\leq} \frac{\gamma\,(p-\sigma)}{1-\gamma\,(1-\sigma)} \leq \frac{\gamma c_1(1-\gamma)}{1-\gamma\,(1-\sigma)} \overset{\text{(ii)}}{<} c_1\gamma, \tag{188}$$

where (i) holds by $z_\phi^\pi < p$, and (ii) is due to (187). Inserting (187) and (188) back into (186), we arrive at

$$\begin{aligned}\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma}\big\rangle &\geq \frac{\frac{\gamma\left(p-z_\phi^\pi\right)}{2(1-\gamma)}}{(1-\gamma)(1+c_1\gamma)^2} \geq \frac{\gamma\big(p-z_\phi^\pi\big)}{8(1-\gamma)^2} \\ &= \frac{\gamma\,(p-q)\,\big(1-\pi(\phi\,|\,0)\big)}{8(1-\gamma)^2} = \frac{\gamma\Delta\big(1-\pi(\phi\,|\,0)\big)}{8(1-\gamma)^2} \geq 2\varepsilon\big(1-\pi(\phi\,|\,0)\big),\end{aligned} \tag{189}$$

where the last inequality holds by setting ($\gamma \geq 1/2$)

$$\Delta = 32(1-\gamma)^2\varepsilon. \tag{190}$$

Finally, it is easily verified that

$$\varepsilon \leq \frac{c_1}{32(1-\gamma)} \qquad \Longrightarrow \qquad \Delta \leq c_1(1-\gamma).$$

- When $\sigma \in (1-\gamma, 1-c_1]$. Regarding (186), we observe that

$$\gamma\sigma < 1-\gamma\,(1-\sigma) = 1-\gamma+\gamma\sigma \leq (1+\gamma)\sigma \leq 2\sigma, \tag{191}$$

which directly leads to

$$\frac{\gamma\big(z_\phi^\pi - \sigma\big)}{1-\gamma\,(1-\sigma)} \leq \frac{\gamma\,(p-\sigma)}{1-\gamma\,(1-\sigma)} \leq \frac{\gamma c_1\sigma}{1-\gamma\,(1-\sigma)} \overset{\text{(i)}}{<} c_1, \tag{192}$$

where (i) holds by (191). Inserting (191) and (192) back into (186), we arrive at

$$\begin{aligned}\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma}\big\rangle &\geq \frac{\frac{\gamma\left(p-z_\phi^\pi\right)}{2\sigma}}{(1-\gamma)(1+c_1)^2} \geq \frac{\gamma\left(p-z_\phi^\pi\right)}{8(1-\gamma)\sigma} = \frac{\gamma\,(p-q)\,\big(1-\pi(\phi\,|\,0)\big)}{8(1-\gamma)\sigma} \\ &= \frac{\gamma\Delta\big(1-\pi(\phi\,|\,0)\big)}{8(1-\gamma)\sigma} \geq 2\varepsilon\big(1-\pi(\phi\,|\,0)\big),\end{aligned} \tag{193}$$

where the last inequality holds by letting ($\gamma \geq 1/2$)

$$\Delta = 32(1-\gamma)\sigma\varepsilon. \tag{194}$$

Finally, it is easily verified that

$$\varepsilon \leq \frac{c_1}{32(1-\gamma)} \qquad \Longrightarrow \qquad \Delta \leq c_1\sigma. \tag{195}$$

D Proof of the upper bound with χ^2 divergence: Theorem 3

The proof of Theorem 3 mainly follows the structure of the proof of Theorem 1 in Appendix 5.2. Throughout this section, for any nominal transition kernel P, the uncertainty set is taken as (see (10))

$$\mathcal{U}^\sigma(P) = \mathcal{U}^\sigma_{\chi^2}(P) \coloneqq \otimes\, \mathcal{U}^\sigma_{\chi^2}(P_{s,a}), \quad \mathcal{U}^\sigma_{\chi^2}(P_{s,a}) \coloneqq \Big\{P'_{s,a} \in \Delta(\mathcal{S}) : \sum_{s'\in\mathcal{S}} \frac{\big(P'(s'\,|\,s,a) - P(s'\,|\,s,a)\big)^2}{P(s'\,|\,s,a)} \leq \sigma\Big\}. \tag{196}$$

D.1 Proof of Theorem 3

In order to control the performance gap $\left\|V^{\star,\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty$, recall the error decomposition in (51):

$$V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \le \left(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\right) + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1 + \left(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right), \tag{197}$$

where $\varepsilon_{\mathsf{opt}}$ (cf. (50)) shall be specified later (which justifies Remark 2). To further control (197), we bound the remaining two terms separately.

Step 1: controlling $\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty$. Towards this, recall the bound in (56) which holds for any uncertainty set:

$$\begin{aligned} \left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty \le \gamma \max\Big\{ & \left\|\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right)\right\|_\infty, \\ & \left\|\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right)\right\|_\infty \Big\}. \end{aligned} \tag{198}$$

To control the main term $\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}$ in (198), we first introduce an important lemma whose proof is postponed to Appendix D.2.1.

Lemma 16. *Consider any $\sigma > 0$ and the uncertainty set $\mathcal{U}^\sigma(\cdot) \coloneqq \mathcal{U}^\sigma_{\chi^2}(\cdot)$. For any $\delta \in (0,1)$ and any fixed policy π, one has with probability at least $1-\delta$,*

$$\left\|\widehat{\underline{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right\|_\infty \le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}}.$$

Applying Lemma 16 by taking $\pi = \pi^\star$ gives

$$\left\|\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right\|_\infty \le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}}, \tag{199}$$

which directly leads to

$$\begin{aligned} & \left\|\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right)\right\|_\infty \\ & \le \left\|\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right\|_\infty \cdot \left\|\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,\widehat{V}}\right)^{-1} 1\right\|_\infty \le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^4 N}}. \end{aligned} \tag{200}$$

Similarly, we have

$$\left\|\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right)\right\|_\infty \le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^4 N}}. \tag{201}$$

Inserting (200) and (201) back to (198) yields

$$\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty \le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^4 N}}. \tag{202}$$

Step 2: controlling $\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty$. Recall the bound in (57) which holds for any uncertainty set:

$$\begin{aligned} \left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty \le \gamma \max\Big\{ & \left\|\left(I - \gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\left(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}\right)\right\|_\infty, \\ & \left\|\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}\right)\right\|_\infty \Big\}. \end{aligned} \tag{203}$$

We introduce the following lemma which controls $\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}$ in (203); the proof is deferred to Appendix D.2.2.

Lemma 17. *Consider the uncertainty set $\mathcal{U}^\sigma(\cdot) \coloneqq \mathcal{U}^\sigma_{\chi^2}(\cdot)$ and any $\delta \in (0,1)$. With probability at least $1-\delta$, one has*

$$\left\|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty \le 12\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^2N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} + 4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}}. \tag{204}$$

Repeating the arguments from (199) to (202) yields

$$\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty \le 12\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^4N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2} + 4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^4}}. \tag{205}$$

Finally, inserting (202) and (205) back to (197) complete the proof

$$\begin{aligned}
&\left\|V^{\star,\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty \\
&\le \left\|V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\right\|_\infty + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} + \left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty \\
&\le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^4N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} + 12\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^4N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2} + 4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^4}} \\
&\le 24\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^4N}},
\end{aligned} \tag{206}$$

where the last line holds by taking $\varepsilon_{\mathsf{opt}} \le \min\left\{\sqrt{\frac{32(1+\sigma)\log(\frac{36SAN^2}{\delta})}{N}}, \frac{4\log(\frac{36SAN^2}{\delta})}{N}\right\}$.

D.2 Proof of the auxiliary lemmas

D.2.1 Proof of Lemma 16

Step 1: controlling the point-wise concentration. Consider any fixed policy π and the corresponding robust value vector $V \coloneqq V^{\pi,\sigma}$ (independent from $\widehat{P}^0$). Invoking Lemma 2 leads to that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
&\left|\widehat{P}^{\pi,V}_{s,a}V^{\pi,\sigma} - P^{\pi,V}_{s,a}V^{\pi,\sigma}\right| \\
&= \Bigg|\max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left\{P^0_{s,a}[V]_\alpha - \sqrt{\sigma\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right\} \\
&\qquad - \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left\{\widehat{P}^0_{s,a}[V]_\alpha - \sqrt{\sigma\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)}\right\}\Bigg| \\
&\le \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha + \sqrt{\sigma\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\sigma\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\le \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right| + \\
&\qquad + \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\sqrt{\sigma}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right|,
\end{aligned} \tag{207}$$

where the first inequality follows by that the maximum operator is 1-Lipschitz, and the second inequality follows from the triangle inequality. Observing that the first term in (207) is exactly the same as (136), recalling the fact in (141) directly leads to: with probability at least $1-\delta$,

$$\max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right| \le 2\sqrt{\frac{\log(\frac{2SAN}{\delta})}{(1-\gamma)^2N}} \tag{208}$$

holds for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Then the remainder of the proof focuses on controlling the second term in (207).

Step 2: controlling the second term in (207). For any given $(s,a) \in \mathcal{S} \times \mathcal{A}$ and fixed $\alpha \in [0, \frac{1}{1-\gamma}]$, applying the concentration inequality (Panaganti and Kalathil, 2022, Lemma 6) with $\|[V]_\alpha\|_\infty \leq \frac{1}{1-\gamma}$, we arrive at

$$\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \leq \sqrt{\frac{2\log(\frac{2}{\delta})}{(1-\gamma)^2 N}} \tag{209}$$

holds with probability at least $1-\delta$. To obtain a uniform bound, we first observe the follow lemma proven in Appendix D.2.3.

Lemma 18. *For any V obeying $\|V\|_\infty \leq \frac{1}{1-\gamma}$, the function $J_{s,a}(\alpha, V) := \left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right|$ w.r.t. α obeys*

$$|J_{s,a}(\alpha_1, V) - J_{s,a}(\alpha_2, V)| \leq 4\sqrt{\frac{|\alpha_1 - \alpha_2|}{1-\gamma}}.$$

In addition, we can construct an ε_3-net N_{ε_3} over $[0, \frac{1}{1-\gamma}]$ whose size is $|N_{\varepsilon_3}| \leq \frac{3}{\varepsilon_3(1-\gamma)}$ (Vershynin, 2018). Armed with the above, we can derive the uniform bound over $\alpha \in [\min_s V(s), \max_s V(s)] \subset [0, 1/(1-\gamma)]$: with probability at least $1 - \frac{\delta}{SA}$, it holds that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
&\max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\leq \max_{\alpha \in [0, 1/(1-\gamma)]} \left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\overset{\text{(i)}}{\leq} 4\sqrt{\frac{\varepsilon_3}{1-\gamma}} + \sup_{\alpha \in N_{\varepsilon_3}} \left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\overset{\text{(ii)}}{\leq} 4\sqrt{\frac{\varepsilon_3}{1-\gamma}} + \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_3}|}{\delta})}{(1-\gamma)^2 N}} \\
&\overset{\text{(iii)}}{\leq} 2\sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_3}|}{\delta})}{(1-\gamma)^2 N}} \leq 2\sqrt{\frac{2\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}},
\end{aligned} \tag{210}$$

where (i) holds by the property of N_{ε_3}, (ii) follows from (209), (iii) arises from taking $\varepsilon_3 = \frac{\log(\frac{2SA|N_{\varepsilon_3}|}{\delta})}{8N(1-\gamma)}$, and the last inequality is verified by $|N_{\varepsilon_3}| \leq \frac{3}{\varepsilon_3(1-\gamma)} \leq 24N$.

Inserting (208) and (210) back to (207) and taking the union bound over $(s,a) \in \mathcal{S} \times \mathcal{A}$, we arrive at that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1-\delta$,

$$\begin{aligned}
\left|\widehat{P}^{\pi,V}_{s,a} V - P^{\pi,V}_{s,a} V\right| &\leq \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right| + \\
&\quad + \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left|\sqrt{\sigma \mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\sigma \mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\leq \sqrt{\frac{2\log(\frac{2SAN}{\delta})}{(1-\gamma)^2 N}} + 2\sqrt{\frac{2\sigma\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}} \leq 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}}.
\end{aligned}$$

Finally, we complete the proof by recalling the matrix form as below:

$$\left\|\underline{\widehat{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right\|_\infty \leq \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left|\widehat{P}^{\pi,V}_{s,a} V - P^{\pi,V}_{s,a} V\right| \leq 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}}.$$

D.2.2 Proof of Lemma 17

Step 1: decomposing the term of interest. The proof follows the routine of the proof of Lemma 14 in Appendix B.7. To begin with, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, following the same arguments of (207) yields

$$\left|\widehat{P}^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma} - P^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma}\right| \le \max_{\alpha\in\left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]} \left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right| +$$

$$+ \max_{\alpha\in\left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]} \sqrt{\sigma}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}\right|. \tag{211}$$

Invoking the fact in (165) (for proving Lemma 14), the first term in (211) obeys

$$\begin{aligned}\max_{\alpha\in\left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]} \left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right| &\le \max_{\alpha\in[0,1/(1-\gamma)]} \left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right| \\ &\le 4\sqrt{\frac{\log(\frac{3SAN^{3/2}}{(1-\gamma)\delta})}{(1-\gamma)^2N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}. \end{aligned} \tag{212}$$

The remainder of the proof will focus on controlling the second term of (211).

Step 2: controlling the second term of (211). Towards this, we recall the auxiliary robust MDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ defined in Appendix B.7. Taking the uncertainty set $\mathcal{U}^\sigma(\cdot) \coloneqq \mathcal{U}^\sigma_{\chi^2}(\cdot)$ for both $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ and $\widehat{\mathcal{M}}_{\mathsf{rob}}$, we recall the corresponding robust Bellman operator $\widehat{\mathcal{T}}^\sigma_{s,u}(\cdot)$ in (154) and the following definition in (155)

$$u^\star \coloneqq \widehat{V}^{\star,\sigma}(s) - \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma}. \tag{213}$$

Following the arguments in Appendix B.7, it can be verified that there exists a unique fixed point $\widehat{Q}^{\star,\sigma}_{s,u}$ of the operator $\widehat{\mathcal{T}}^\sigma_{s,u}(\cdot)$, which satisfies $0 \le \widehat{Q}^{\star,\sigma}_{s,u} \le \frac{1}{1-\gamma}1$. In addition, the corresponding robust value function coincides with that of the operator $\widehat{\mathcal{T}}^\sigma(\cdot)$, i.e., $\widehat{V}^{\star,\sigma}_{s,u} = \widehat{V}^{\star,\sigma}$.

We recall the N_{ε_2}-net over $\left[0,\frac{1}{1-\gamma}\right]$ whose size obeying $|N_{\varepsilon_2}| \le \frac{3}{\varepsilon_2(1-\gamma)}$ (Vershynin, 2018). Then for all $u \in N_{\varepsilon_2}$ and a fixed α, $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is statistically independent from $\widehat{P}^0_{s,a}$, which indicates the independence between $[\widehat{V}^{\star,\sigma}_{s,u}]_\alpha$ and $\widehat{P}^0_{s,a}$. With this in mind, invoking the fact in (210) and taking the union bound over all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $u \in N_{\varepsilon_2}$ yields that, with probability at least $1-\delta$,

$$\max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left([\widehat{V}^{\star,\sigma}_{s,u}]_\alpha\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left([\widehat{V}^{\star,\sigma}_{s,u}]_\alpha\right)}\right| \le 2\sqrt{\frac{2\log(\frac{24SAN|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2N}} \tag{214}$$

holds for all $(s,a,u)\in\mathcal{S}\times\mathcal{A}\times N_{\varepsilon_2}$.

To continue, we decompose the term of interest in (211) as follows:

$$\begin{aligned}&\max_{\alpha\in\left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}\right| \\
&\le \max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}\right| \\
&\overset{\text{(i)}}{\le} \max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}\right| \\
&\quad + \max_{\alpha\in[0,1/(1-\gamma)]}\left[\sqrt{\left|\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right) - \mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)\right|}\right. \\
&\quad \left. + \sqrt{\left|\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right) - \mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)\right|}\right] \\
&\overset{\text{(ii)}}{\le} \max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}\right|
\end{aligned}$$

$$
\begin{aligned}
&+ \max_{\alpha\in[0,1/(1-\gamma)]} 2\sqrt{\frac{2}{(1-\gamma)}\left\|\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha} - \left[\widehat{V}^{\star,\sigma}\right]_{\alpha}\right\|_{\infty}} \\
&\leq \max_{\alpha\in[0,1/(1-\gamma)]} \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_{\alpha}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_{\alpha}\right)} \right| + 4\sqrt{\frac{\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}}, \qquad (215)
\end{aligned}
$$

where (i) holds by the triangle inequality, (ii) arises from applying Lemma 3, and the last inequality holds by (50).

Armed with the above facts, invoking the identity $\widehat{V}^{\star,\sigma} = \widehat{V}^{\star,\sigma}_{s,u^\star}$ leads to that for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$,

$$
\begin{aligned}
&\max_{\alpha\in[0,1/(1-\gamma)]} \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_{\alpha}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_{\alpha}\right)} \right| \\
&= \max_{\alpha\in[0,1/(1-\gamma)]} \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,u^\star}\right]_{\alpha}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,u^\star}\right]_{\alpha}\right)} \right| \\
&\overset{\text{(i)}}{\leq} \max_{\alpha\in[0,1/(1-\gamma)]} \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_{\alpha}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_{\alpha}\right)} \right| \\
&\quad + \max_{\alpha\in[0,1/(1-\gamma)]} \left[\sqrt{\left|\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,u^\star}\right]_{\alpha}\right) - \mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_{\alpha}\right)\right|} \right. \\
&\quad \left. + \sqrt{\left|\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,u^\star}\right]_{\alpha}\right) - \mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_{\alpha}\right)\right|} \right] \\
&\overset{\text{(ii)}}{\leq} \max_{\alpha\in[0,1/(1-\gamma)]} \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_{\alpha}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_{\alpha}\right)} \right| + 4\sqrt{\frac{\varepsilon_2}{(1-\gamma)}} \\
&\overset{\text{(iii)}}{\leq} 2\sqrt{\frac{2\log(\frac{24SAN|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2 N}} + 4\sqrt{\frac{\varepsilon_2}{(1-\gamma)}} \\
&\leq 6\sqrt{\frac{2\log(\frac{36SAN^2|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2 N}}, \qquad (216)
\end{aligned}
$$

where (i) holds by the triangle inequality, (ii) arises from applying Lemma 3 and the fact $\left\|\widehat{V}^{\star,\sigma}_{s,\overline{u}} - \widehat{V}^{\star,\sigma}_{s,u^\star}\right\|_{\infty} \leq \frac{\varepsilon_2}{(1-\gamma)}$ (see (161)), (iii) follows from (214), and the last inequality holds by letting $\varepsilon_2 = \frac{2\log(\frac{24SAN|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)N}$, which leads to $|N_{\varepsilon_2}| \leq \frac{3}{\varepsilon_2(1-\gamma)} \leq \frac{3N}{2}$.

In summary, inserting (216) back to (215) and (215) leads to with probability at least $1-\delta$,

$$
\begin{aligned}
&\max_{\alpha\in\left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]} \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha}\right)} \right| \\
&\leq 6\sqrt{\frac{2\sigma\log(\frac{36SAN^2|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2 N}} + 4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}} \qquad (217)
\end{aligned}
$$

holds for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

Step 4: finishing up. Inserting (217) and (212) back to (211), we complete the proof: with probability at least $1-\delta$,

$$
\begin{aligned}
\left\|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right\|_{\infty} &\leq 4\sqrt{\frac{\log(\frac{3SAN^{3/2}}{(1-\gamma)\delta})}{(1-\gamma)^2 N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} + 6\sqrt{\frac{2\sigma\log(\frac{36SAN^2|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2 N}} + 4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}} \\
&\leq 12\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^2 N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} + 4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}}. \qquad (218)
\end{aligned}
$$

D.2.3 Proof of Lemma 18

For any $0 \leq \alpha_1, \alpha_2 \leq 1/(1-\gamma)$, one has

$$\begin{aligned}
&\left|J_{s,a}(\alpha_1, V) - J_{s,a}(\alpha_2, V)\right| \\
&= \left| \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left([V]_{\alpha_1}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left([V]_{\alpha_1}\right)} \right| - \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left([V]_{\alpha_2}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left([V]_{\alpha_2}\right)} \right| \right| \\
&\overset{\text{(i)}}{\leq} \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left([V]_{\alpha_1}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left([V]_{\alpha_1}\right)} - \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left([V]_{\alpha_2}\right)} + \sqrt{\mathsf{Var}_{P^0_{s,a}}\left([V]_{\alpha_2}\right)} \right| \\
&\leq \left| \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left([V]_{\alpha_1}\right)} - \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left([V]_{\alpha_2}\right)} \right| + \left| \sqrt{\mathsf{Var}_{P^0_{s,a}}\left([V]_{\alpha_1}\right)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}\left([V]_{\alpha_2}\right)} \right| \\
&\overset{\text{(ii)}}{\leq} \sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left([V]_{\alpha_2}\right) - \mathsf{Var}_{\widehat{P}^0_{s,a}}\left([V]_{\alpha_1}\right)} + \sqrt{\mathsf{Var}_{P^0_{s,a}}\left([V]_{\alpha_2}\right) - \mathsf{Var}_{P^0_{s,a}}\left([V]_{\alpha_1}\right)} \\
&\overset{\text{(iii)}}{\leq} \sqrt{\left|\widehat{P}^0_{s,a}\left[\left([V]_{\alpha_1}\right) \circ \left([V]_{\alpha_1}\right) - \left([V]_{\alpha_2}\right) \circ \left([V]_{\alpha_2}\right)\right]\right| + \left|\widehat{P}^0_{s,a}\left([V]_{\alpha_1} + [V]_{\alpha_2}\right) \cdot \widehat{P}^0_{s,a}\left([V]_{\alpha_1} - [V]_{\alpha_2}\right)\right|} \\
&\quad + \sqrt{\left|P^0_{s,a}\left[\left([V]_{\alpha_1}\right) \circ \left([V]_{\alpha_1}\right) - \left([V]_{\alpha_2}\right) \circ \left([V]_{\alpha_2}\right)\right]\right| + \left|P^0_{s,a}\left([V]_{\alpha_1} + [V]_{\alpha_2}\right) \cdot P^0_{s,a}\left([V]_{\alpha_1} - [V]_{\alpha_2}\right)\right|} \\
&\leq 2\sqrt{2(\alpha_1 + \alpha_2)|\alpha_1 - \alpha_2|} \leq 4\sqrt{\frac{|\alpha_1 - \alpha_2|}{1-\gamma}}.
\end{aligned} \tag{219}$$

where (i) holds by the fact $||x| - |y|| \leq |x - y|$ for all $x, y \in \mathbb{R}$, (ii) follows from the fact that $\sqrt{x} - \sqrt{y} \leq \sqrt{x-y}$ for any $x \geq y \geq 0$ and $\mathsf{Var}_P\left([V]_{\alpha_2}\right) \geq \mathsf{Var}_P\left([V]_{\alpha_1}\right)$ for any transition kernel $P \in \Delta(\mathcal{S})$, (iii) holds by the definition of $\mathrm{Var}_P(\cdot)$ defined in (40), and the last inequality arises from $0 \leq \alpha_1, \alpha_2 \leq 1/(1-\gamma)$.

E Proof of the lower bound with χ^2 divergence: Theorem 4

To prove Theorem 4, we shall first construct some hard instances and then characterize the sample complexity requirements over these instances. The structure of the hard instances are the same as the ones used in the proof of Theorem 2.

E.1 Construction of the hard problem instances

First, note that we shall use the same MDPs defined in Appendix 5.3.1 as follows

$$\left\{\mathcal{M}_\phi = \left(\mathcal{S}, \mathcal{A}, P^\phi, r, \gamma\right) \mid \phi = \{0, 1\}\right\}.$$

In particular, we shall keep the structure of the transition kernel in (61), reward function in (63) and initial state distribution in (64), while p and Δ shall be specified differently later.

Uncertainty set of the transition kernels. Recalling the uncertainty set associated with χ^2 divergence in (196), for any uncertainty level σ, the uncertainty set throughout this section is defined as $\mathcal{U}^\sigma(P^\phi)$:

$$\mathcal{U}^\sigma(P^\phi) \coloneqq \otimes\, \mathcal{U}^\sigma_{\chi^2}(P^\phi_{s,a}), \qquad \mathcal{U}^\sigma_{\chi^2}(P^\phi_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \sum_{s' \in \mathcal{S}} \frac{\left(P(s' \,|\, s, a) - P^\phi(s' \,|\, s, a)\right)^2}{P^\phi(s' \,|\, s, a)} \leq \sigma\right\}. \tag{220}$$

Clearly, $\mathcal{U}^\sigma(P^\phi_{s,a}) = P^\phi_{s,a}$ whenever the state transition is deterministic for χ^2 divergence. Here, q and Δ (whose choice will be specified later in more detail) which determine the instances are specified as

$$0 \leq q = \begin{cases} 1-\gamma & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right) \\ \frac{\sigma}{1+\sigma} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \infty\right) \end{cases}, \qquad p = q + \Delta, \tag{221}$$

and

$$0 < \Delta \leq \begin{cases} \frac{1}{4}(1-\gamma) & \text{if } \sigma \in \big(0, \frac{1-\gamma}{4}\big) \\ \min\left\{\frac{1}{4}(1-\gamma),\ \frac{1}{2(1+\sigma)}\right\} & \text{if } \sigma \in \big[\frac{1-\gamma}{4}, \infty\big) \end{cases}. \tag{222}$$

This directly ensures that

$$p = \Delta + q \leq \max\left\{\frac{\frac{1}{2}+\sigma}{1+\sigma}, \frac{5}{4}(1-\gamma)\right\} \leq 1$$

since $\gamma \in \big[\frac{3}{4}, 1\big)$.

To continue, for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we denote the infimum probability of moving to the next state s' associated with any perturbed transition kernel $P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})$ as

$$\underline{P}^\phi(s' \,|\, s, a) := \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})} P(s' \,|\, s, a). \tag{223}$$

In addition, we denote the transition from state 0 to state 1 as follows, which plays an important role in the analysis,

$$\underline{p} := \underline{P}^\phi(1 \,|\, 0, \phi), \qquad \underline{q} := \underline{P}^\phi(1 \,|\, 0, 1-\phi). \tag{224}$$

Before continuing, we introduce some facts about $\underline{p}$ and $\underline{q}$ which are summarized as the following lemma; the proof is postponed to Appendix E.3.1.

Lemma 19. *Consider any $\sigma \in (0, \infty)$ and any p, q, Δ obeying (221) and (222), the following properties hold*

$$\begin{cases} \frac{1-\gamma}{2} < \underline{q} < 1-\gamma, \quad \underline{q} + \frac{3}{4}\Delta \leq \underline{p} \leq \underline{q} + \Delta \leq \frac{5(1-\gamma)}{4} & \text{if } \sigma \in \big(0, \frac{1-\gamma}{4}\big), \\ \underline{q} = 0, \quad \frac{\sigma+1}{2}\Delta \leq \underline{p} \leq (3+\sigma)\Delta & \text{if } \sigma \in \big[\frac{1-\gamma}{4}, \infty\big). \end{cases} \tag{225}$$

Value functions and optimal policies. Armed with above facts, we are positioned to derive the corresponding robust value functions, the optimal policies, and its corresponding optimal robust value functions. For any RMDP $\mathcal{M}_\phi$ with the uncertainty set defined in (220), we denote the robust optimal policy as $\pi^\star_\phi$, the robust value function of any policy π (resp. the optimal policy $\pi^\star_\phi$) as $V^{\pi,\sigma}_\phi$ (resp. $V^{\star,\sigma}_\phi$). The following lemma describes some key properties of the robust (optimal) value functions and optimal policies whose proof is postponed to Appendix E.3.2.

Lemma 20. *For any $\phi = \{0, 1\}$ and any policy π, one has*

$$V^{\pi,\sigma}_\phi(0) = \frac{\gamma z^\pi_\phi}{(1-\gamma)\Big(1-\gamma\big(1-z^\pi_\phi\big)\Big)}, \tag{226}$$

where z^π_ϕ is defined as

$$z^\pi_\phi := \underline{p}\pi(\phi \,|\, 0) + \underline{q}\pi(1-\phi \,|\, 0). \tag{227}$$

In addition, the optimal value functions and the optimal policies obey

$$V^{\star,\sigma}_\phi(0) = \frac{\gamma \underline{p}}{(1-\gamma)\left(1-\gamma\left(1-\underline{p}\right)\right)}, \tag{228a}$$

$$\pi^\star_\phi(\phi \,|\, s) = 1, \qquad \textit{for } s \in \mathcal{S}. \tag{228b}$$

E.2 Establishing the minimax lower bound

Our goal is to control the performance gap w.r.t. any policy estimator $\widehat{\pi}$ based on the generated dataset and the chosen initial distribution φ in (64), which gives

$$\big\langle \varphi, V^{\star,\sigma}_\phi - V^{\widehat{\pi},\sigma}_\phi \big\rangle = V^{\star,\sigma}_\phi(0) - V^{\widehat{\pi},\sigma}_\phi(0). \tag{229}$$

Step 1: converting the goal to estimate ϕ. To achieve the goal, we first introduce the following fact which shall be verified in Appendix E.3.3: given

$$\varepsilon \leq \begin{cases} \frac{1}{72(1-\gamma)} & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right), \\ \frac{1}{256(1+\sigma)(1-\gamma)} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\right), \\ \frac{3}{32} & \text{if } \sigma > \frac{1}{3(1-\gamma)}. \end{cases} \tag{230}$$

choosing

$$\Delta = \begin{cases} 18(1-\gamma)^2\varepsilon & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right), \\ 64(1+\sigma)(1-\gamma)^2\varepsilon & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\right), \\ \frac{16}{3(1+\sigma)}\varepsilon & \text{if } \sigma > \frac{1}{3(1-\gamma)}. \end{cases} \tag{231}$$

which satisfies the requirement of Δ in (221), it holds that for any policy $\widehat{\pi}$,

$$\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \right\rangle \geq 2\varepsilon\big(1 - \widehat{\pi}(\phi \,|\, 0)\big). \tag{232}$$

Step 2: arriving at the final results. To continue, following the same definitions and argument in Appendix 5.3.2, we recall the minimax probability of the error and its property as follows:

$$p_{\mathrm{e}} \geq \frac{1}{4}\exp\left\{ -N\Big(\mathsf{KL}\big(P^0(\cdot \,|\, 0,0) \parallel P^1(\cdot \,|\, 0,0)\big) + \mathsf{KL}\big(P^0(\cdot \,|\, 0,1) \parallel P^1(\cdot \,|\, 0,1)\big)\Big)\right\}, \tag{233}$$

then we can complete the proof by showing $p_{\mathrm{e}} \geq \frac{1}{8}$ given the bound for the sample size N. In the following, we shall control the KL divergence terms in (233) in three different cases.

- Case 1: $\sigma \in \left(0, \frac{1-\gamma}{4}\right)$. In this case, applying $\gamma \in [\frac{3}{4}, 1)$ yields

$$\begin{aligned} & 1 - q > 1 - p = 1 - q - \Delta > \gamma - \frac{1-\gamma}{4} > \frac{3}{4} - \frac{1}{16} > \frac{1}{2}, \\ & p \geq q = 1 - \gamma. \end{aligned} \tag{234}$$

  Armed with the above facts, applying Tsybakov (2009, Lemma 2.7) yields

$$\begin{aligned} \mathsf{KL}\big(P^0(\cdot \,|\, 0,1) \parallel P^1(\cdot \,|\, 0,1)\big) = \mathsf{KL}\left(p \parallel q\right) \leq \frac{(p-q)^2}{(1-p)p} & \overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\ & \overset{\text{(ii)}}{=} \frac{324(1-\gamma)^4\varepsilon^2}{p(1-p)} \\ & \overset{\text{(iii)}}{\leq} 648(1-\gamma)^3\varepsilon^2, \end{aligned} \tag{235}$$

  where (i) follows from the definition in (221), (ii) holds by plugging in the expression of Δ in (231), and (iii) arises from (234). The same bound can be established for $\mathsf{KL}\big(P_1^0(\cdot \,|\, 0,0) \parallel P_1^1(\cdot \,|\, 0,0)\big)$. Substituting (235) back into (233) demonstrates that: if the sample size is chosen as

$$N \leq \frac{\log 2}{1296(1-\gamma)^3\varepsilon^2}, \tag{236}$$

  then one necessarily has

$$p_{\mathrm{e}} \geq \frac{1}{4}\exp\left\{ -N \cdot 1296(1-\gamma)^3\varepsilon^2 \right\} \geq \frac{1}{8}. \tag{237}$$

- Case 2: $\sigma \in \left[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\right)$. Applying the facts of Δ in (222), one has

$$1 - q > 1 - p = 1 - q - \Delta \geq \frac{1}{1+\sigma} - \frac{1}{2(1+\sigma)} = \frac{1}{2(1+\sigma)},$$

$$p \geq q = \frac{\sigma}{1+\sigma}. \tag{238}$$

Given (238), applying Tsybakov (2009, Lemma 2.7) yields

$$\begin{aligned} \mathsf{KL}\big(P^0(\cdot\,|\,0,1) \parallel P^1(\cdot\,|\,0,1)\big) = \mathsf{KL}\left(p \parallel q\right) \leq \frac{(p-q)^2}{(1-p)p} &\overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\ &\overset{\text{(ii)}}{=} \frac{4096(1+\sigma)^2(1-\gamma)^4\varepsilon^2}{p(1-p)} \\ &\overset{\text{(iii)}}{\leq} \frac{4096(1+\sigma)^2(1-\gamma)^4\varepsilon^2}{\frac{\sigma}{2(1+\sigma)^2}} \leq \frac{8192(1-\gamma)^4(1+\sigma)^4\varepsilon^2}{\sigma}, \end{aligned} \tag{239}$$

where (i) follows from the definition in (221), (ii) holds by plugging in the expression of Δ in (231), and (iii) arises from (238). The same bound can be established for $\mathsf{KL}\big(P_1^0(\cdot\,|\,0,0) \parallel P_1^1(\cdot\,|\,0,0)\big)$. Substituting (239) back into (80) demonstrates that: if the sample size is chosen as

$$N \leq \frac{\sigma \log 2}{16384(1-\gamma)^4(1+\sigma)^4\varepsilon^2}, \tag{240}$$

then one necessarily has

$$p_{\mathrm{e}} \geq \frac{1}{4}\exp\left\{-N\frac{16384(1-\gamma)^4(1+\sigma)^4\varepsilon^2}{\sigma}\right\} \geq \frac{1}{8}. \tag{241}$$

- Case 3: $\sigma > \frac{1}{3(1-\gamma)} \geq \frac{1}{3}$. Regarding this, one gives

$$\begin{aligned} &1-q > 1-p = 1-q-\Delta \geq \frac{1}{1+\sigma} - \frac{1}{4(1+\sigma)} \geq \frac{1}{2(1+\sigma)}, \\ &p \geq q \geq \frac{1}{4}. \end{aligned} \tag{242}$$

Given $p \geq q \geq 1/2$ and (242), applying Tsybakov (2009, Lemma 2.7) yields

$$\begin{aligned} \mathsf{KL}\big(P^0(\cdot\,|\,0,1) \parallel P^1(\cdot\,|\,0,1)\big) = \mathsf{KL}\left(p \parallel q\right) \leq \frac{(p-q)^2}{(1-p)p} &\overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\ &\overset{\text{(ii)}}{\leq} \frac{\frac{64}{(1+\sigma)^2}\varepsilon^2}{p(1-p)} \\ &\overset{\text{(iii)}}{\leq} \frac{492\varepsilon^2}{\sigma}, \end{aligned} \tag{243}$$

where (i) follows from the definition in (221), (ii) holds by plugging in the expression of Δ in (231), and (iii) arises from (242). The same bound can be established for $\mathsf{KL}\big(P_1^0(\cdot\,|\,0,0) \parallel P_1^1(\cdot\,|\,0,0)\big)$. Substituting (243) back into (80) demonstrates that: if the sample size is chosen as

$$N \leq \frac{\sigma \log 2}{984\varepsilon^2}, \tag{244}$$

then one necessarily has

$$p_{\mathrm{e}} \geq \frac{1}{4}\exp\left\{-N\frac{984\varepsilon^2}{\sigma}\right\} \geq \frac{1}{8}. \tag{245}$$

Step 3: putting things together. Finally, summing up the results in (236), (240), and (244), combined with the requirement in (230), one has when

$$\varepsilon \leq c_1 \begin{cases} \frac{1}{1-\gamma} & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right) \\ \max\left\{\frac{1}{(1+\sigma)(1-\gamma)}, 1\right\} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \infty\right) \end{cases}, \tag{246}$$

taking

$$N \leq c_2 \begin{cases} \frac{1}{(1-\gamma)^3\varepsilon^2} & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right) \\ \frac{\sigma}{\min\{1,(1-\gamma)^4(1+\sigma)^4\}\varepsilon^2} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \infty\right) \end{cases} \tag{247}$$

leads to $p_e \geq \frac{1}{8}$, for some universal constants $c_1, c_2 > 0$.

E.3 Proof of the auxiliary facts

We begin with some basic facts about the χ^2 divergence defined in (39) for any two Bernoulli distributions $\mathsf{Ber}(w)$ and $\mathsf{Ber}(x)$, denoted as

$$f(w,x) \coloneqq \chi^2(x \parallel w) = \frac{(w-x)^2}{w} + \frac{(1-w-(1-x))^2}{1-w} = \frac{(w-x)^2}{w(1-w)}. \tag{248}$$

For $x \in [0, w)$, it is easily verified that the partial derivative w.r.t. x obeys $\frac{\partial f(w,x)}{\partial x} = \frac{2(x-w)}{w(1-w)} < 0$, implying that

$$\forall\ x_1 < x_2 \in [0, w), \qquad f(w, x_1) > f(w, x_2). \tag{249}$$

In other words, the χ^2 divergence $f(w,x)$ increases as x decreases from w to 0.

Next, we introduce the following function for any fixed $\sigma \in (0, \infty)$ and any $x \in \left[\frac{\sigma}{1+\sigma}, 1\right)$:

$$f_\sigma(x) \coloneqq \inf_{\{y: \chi^2(y \| x) \leq \sigma, y \in [0,x]\}} y \overset{\text{(i)}}{=} \max\left\{0, x - \sqrt{\sigma x(1-x)}\right\} = x - \sqrt{\sigma x(1-x)}, \tag{250}$$

where (i) has been verified in Yang et al. (2022, Corollary B.2), and the last equality holds since $x \geq \frac{\sigma}{1+\sigma}$. The next lemma summarizes some useful facts about $f_\sigma(\cdot)$, which again has been verified in Yang et al. (2022, Lemma B.12 and Corollary B.2).

Lemma 21. *Consider any $\sigma \in (0, \infty)$. For $x \in [\frac{\sigma}{1+\sigma}, 1)$, $f_\sigma(x)$ is convex and differentiable, which obeys*

$$f'_\sigma(x) = 1 + \frac{\sqrt{\sigma}(2x-1)}{2\sqrt{x(1-x)}}.$$

E.3.1 Proof of Lemma 19

Let us control $\underline{q}$ and $\underline{p}$ respectively.

Step 1: controlling $\underline{q}$. We shall control $\underline{q}$ in different cases w.r.t. the uncertainty level σ.

- Case 1: $\sigma \in \left(0, \frac{1-\gamma}{4}\right)$. In this case, recall that $q = 1-\gamma$ defined in (221), applying (250) with $x = q$ leads to

$$1-\gamma = q > \underline{q} = f_\sigma(q) = 1-\gamma - \sqrt{\sigma\gamma(1-\gamma)} \geq 1-\gamma - \sqrt{\frac{1-\gamma}{4}\gamma(1-\gamma)} > \frac{1-\gamma}{2}. \tag{251}$$

- Case 2: $\sigma \in \left[\frac{1-\gamma}{4}, \infty\right)$. Note that it suffices to treat $P^{\phi}_{0,1-\phi}$ as a Bernoulli distribution $\mathsf{Ber}(q)$ over states 1 and 0, since we do not allow transition to other states. Recalling $q = \frac{\sigma}{1+\sigma}$ in (221) and noticing the fact that

$$f(q, 0) = \frac{q^2}{q} + \frac{(1-(1-q))^2}{1-q} = \frac{q}{(1-q)} = \sigma, \tag{252}$$

one has the probability $\mathsf{Ber}(0)$ falls into the uncertainty set of $\mathsf{Ber}(q)$) of size σ. As a result, recalling the definition (224) leads to

$$\underline{q} = \underline{P}^{\phi}(1 \,|\, 0, 1-\phi) = 0, \tag{253}$$

since $\underline{q} \geq 0$.

Step 2: controlling $\underline{p}$. To characterize the value of $\underline{p}$, we also divide into several cases separately.

- Case 1: $\sigma \in \left(0, \frac{1-\gamma}{4}\right)$. In this case, note that $p > q = 1-\gamma \geq \frac{\sigma}{1+\sigma}$. Therefore, applying that $f_\sigma(\cdot)$ is convex and the form of its derivative in Lemma 21, one has

$$\begin{aligned}\underline{p} &= f_\sigma(p) \geq f_\sigma(q) + f'_\sigma(q)(p-q)\\ &= \underline{q} + \left(1 + \frac{\sqrt{\sigma}(2q-1)}{2\sqrt{q(1-q)}}\right)\Delta \geq \underline{q} + \left(1 - \frac{\sqrt{\frac{1-\gamma}{4}}(1-2(1-\gamma))}{2\sqrt{(1-\gamma)\gamma}}\right)\Delta \geq \underline{q} + \frac{3\Delta}{4}.\end{aligned} \tag{254}$$

Similarly, applying Lemma 21 leads to

$$\begin{aligned}\underline{p} &= f_\sigma(p) \leq f_\sigma(q) + f'_\sigma(p)(p-q)\\ &= \underline{q} + \left(1 - \frac{\sqrt{\sigma}(1-2p)}{2\sqrt{p(1-p)}}\right)\Delta \leq \underline{q} + \Delta,\end{aligned} \tag{255}$$

where the last inequality holds by $1-2p > 0$ due to the fact $p = q + \Delta \leq \frac{5}{4}(1-\gamma) \leq \frac{5}{16} < \frac{1}{2}$ (cf. (222) and $\gamma \in [\frac{3}{4}, 1)$). To sum up, given $\sigma \in \left(0, \frac{1-\gamma}{4}\right)$, combined with (251), we arrive at

$$\underline{q} + \frac{3}{4}\Delta \leq \underline{p} \leq \underline{q} + \Delta \leq \frac{5(1-\gamma)}{4}, \tag{256}$$

where the last inequality holds by $\Delta \leq \frac{1}{4}(1-\gamma)$ (see (221)).

- Case 2: $\sigma \in \left[\frac{1-\gamma}{4}, \infty\right)$. We recall that $p = q + \Delta > q = \frac{\sigma}{1+\sigma}$ in (221). To derive the lower bound for $\underline{p}$ in (224), similar to (254), one has

$$\begin{aligned}\underline{p} &= f_\sigma(p) \geq f_\sigma(q) + f'_\sigma(q)(p-q)\\ &= \underline{q} + \left(1 + \frac{\sqrt{\sigma}(2q-1)}{2\sqrt{q(1-q)}}\right)\Delta\\ &\overset{\text{(i)}}{=} 0 + \left(1 + \frac{\sqrt{\sigma}\frac{\sigma-1}{1+\sigma}}{2\sqrt{\frac{\sigma}{1+\sigma}\frac{1}{1+\sigma}}}\right)\Delta = \left(1 + \frac{\sigma-1}{2}\right)\Delta = \left(\frac{\sigma+1}{2}\right)\Delta,\end{aligned} \tag{257}$$

where (i) follows from $q = \frac{\sigma}{1+\sigma}$ and $\underline{q} = 0$ (see (253)). For the other direction, similar to (255), we have

$$\begin{aligned}\underline{p} = f_\sigma(p) \leq f_\sigma(q) + f'_\sigma(p)(p-q) = \underline{q} + \left(1 + \frac{\sqrt{\sigma}(2p-1)}{2\sqrt{p(1-p)}}\right)\Delta\\ \overset{\text{(i)}}{=} \left(1 + \frac{\sqrt{\sigma}(2p-1)}{2\sqrt{p(1-p)}}\right)\Delta \overset{\text{(ii)}}{=} \left(1 + \frac{\sqrt{\sigma}\left(\frac{\sigma-1}{1+\sigma} + 2\Delta\right)}{2\sqrt{\left(\frac{\sigma}{1+\sigma} + \Delta\right)\left(\frac{1}{1+\sigma} - \Delta\right)}}\right)\Delta\end{aligned}$$

$$\overset{\text{(iii)}}{\le} \left(1+\frac{\sqrt{\sigma}(1+2\Delta)}{2\sqrt{\frac{\sigma}{1+\sigma}\cdot\frac{1}{2(1+\sigma)}}}\right)\Delta \overset{\text{(iv)}}{\le} \left(1+(1+\sigma)\left(1+\frac{1}{1+\sigma}\right)\right)\Delta = (3+\sigma)\Delta, \tag{258}$$

where (i) holds by $\underline{q}=0$ (see (253)), (ii) follows from plugging in $p = q+\Delta = \frac{\sigma}{1+\sigma}+\Delta$, and (iii) and (iv) arises from $\Delta = \min\left\{\frac{1}{4}(1-\gamma), \frac{1}{2(1+\sigma)}\right\} \le 1$ in (222). Combining (257) and (258) yields

$$\frac{\sigma+1}{2}\Delta \le \underline{p} \le (3+\sigma)\Delta. \tag{259}$$

Step 3: combining all the results. Finally, summing up the results for both $\underline{q}$ (in (251) and (253)) and $\underline{p}$ (in (256) and (259)), we arrive at the advertised bound.

E.3.2 Proof of Lemma 20

The robust value function for any policy π. For any $\mathcal{M}_\phi$ with $\phi\in\{0,1\}$, we first characterize the robust value function of any policy π over different states.

Towards this, it is easily observed that for any policy π, the robust value functions at state $s=1$ or any $s\in\{2,3,\cdots,S-1\}$ obey

$$V_\phi^{\pi,\sigma}(1) \overset{\text{(i)}}{=} 1+\gamma V_\phi^{\pi,\sigma}(1) = \frac{1}{1-\gamma} \tag{260a}$$

and

$$\forall s\in\{2,3,\cdots,S\}: \qquad V_\phi^{\pi,\sigma}(s) \overset{\text{(ii)}}{=} 0+\gamma V_\phi^{\pi,\sigma}(1) = \frac{\gamma}{1-\gamma}, \tag{260b}$$

where (i) and (ii) is according to the facts that the transitions defined over states $s\ge 1$ in (61) give only one possible next state 1, leading to a non-random transition in the uncertainty set associated with χ^2 divergence, and $r(1,a)=1$ for all $a\in\mathcal{A}'$ and $r(s,a)=0$ holds all $(s,a)\in\{2,3,\cdots,S-1\}\times\mathcal{A}$.

To continue, the robust value function at state 0 with policy π satisfies

$$\begin{aligned} V_\phi^{\pi,\sigma}(0) &= \mathbb{E}_{a\sim\pi(\cdot\,|\,0)}\left[r(0,a)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,a})} \mathcal{P}V_\phi^{\pi,\sigma}\right] \\ &= 0+\gamma\pi(\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,\phi})}\mathcal{P}V_\phi^{\pi,\sigma}+\gamma\pi(1-\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,1-\phi})}\mathcal{P}V_\phi^{\pi,\sigma} \qquad (261)\\ &\overset{\text{(i)}}{\le} \frac{\gamma}{1-\gamma}, \qquad (262)\end{aligned}$$

where (i) holds by that $\|V_\phi^{\pi,\sigma}\|_\infty \le \frac{1}{1-\gamma}$. Summing up the results in (260b) and (262) leads to

$$\forall s\in\{2,3,\cdots,S\}, \qquad V_\phi^{\pi,\sigma}(1) > V_\phi^{\pi,\sigma}(s) \ge V_\phi^{\pi,\sigma}(0). \tag{263}$$

With the transition kernel in (61) over state 0 and the fact in (263), (261) can be rewritten as

$$\begin{aligned} V_\phi^{\pi,\sigma}(0) &= \gamma\pi(\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,\phi})}\mathcal{P}V_\phi^{\pi,\sigma}+\gamma\pi(1-\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,1-\phi})}\mathcal{P}V_\phi^{\pi,\sigma} \\ &\overset{\text{(i)}}{=} \gamma\pi(\phi\,|\,0)\Big[\underline{p}V_\phi^{\pi,\sigma}(1)+\big(1-\underline{p}\big)V_\phi^{\pi,\sigma}(0)\Big]+\gamma\pi(1-\phi\,|\,0)\Big[\underline{q}V_\phi^{\pi,\sigma}(1)+\big(1-\underline{q}\big)V_\phi^{\pi,\sigma}(0)\Big] \\ &\overset{\text{(ii)}}{=} \gamma z_\phi^\pi V_\phi^{\pi,\sigma}(1)+\gamma\left(1-z_\phi^\pi\right)V_\phi^{\pi,\sigma}(0) \\ &= \frac{\gamma z_\phi^\pi}{(1-\gamma)\Big(1-\gamma\big(1-z_\phi^\pi\big)\Big)}, \qquad (264)\end{aligned}$$

where (i) holds by the definition of $\underline{p}$ and $\underline{q}$ in (224), (ii) follows from the definition of z_ϕ^π in (227), and the last line holds by applying (260a) and solving the resulting linear equation for $V_\phi^{\pi,\sigma}(0)$.

Optimal policy and its optimal value function. To continue, observing that $V_\phi^{\pi,\sigma}(0) =: f(z_\phi^\pi)$ is increasing in z_ϕ^π since the derivative of $f(z_\phi^\pi)$ w.r.t. z_ϕ^π obeys

$$f'(z_\phi^\pi) = \frac{\gamma(1-\gamma)\Big(1-\gamma(1-z_\phi^\pi)\Big) - \gamma^2 z_\phi^\pi(1-\gamma)}{(1-\gamma)^2\Big(1-\gamma(1-z_\phi^\pi)\Big)^2} = \frac{\gamma}{\Big(1-\gamma(1-z_\phi^\pi)\Big)^2} > 0,$$

where the last inequality holds by $0 \leq z_\phi^\pi \leq 1$. Further, z_ϕ^π is also increasing in $\pi(\phi \,|\, 0)$ (see the fact $\underline{p} \geq \underline{q}$ in (224)), the optimal robust policy in state 0 thus obeys

$$\pi_\phi^\star(\phi \,|\, 0) = 1. \tag{265}$$

Considering that the action does not influence the state transition for all states $s > 0$, without loss of generality, we choose the optimal robust policy to obey

$$\forall s > 0: \quad \pi_\phi^\star(\phi \,|\, s) = 1. \tag{266}$$

Taking $\pi = \pi_\phi^\star$ and $z_\phi^{\pi_\phi^\star} = \underline{p}$ in (264), we complete the proof by showing the corresponding optimal robust value function at state 0 as follows:

$$V_\phi^{\star,\sigma}(0) = \frac{\gamma z_\phi^{\pi_\phi^\star}}{(1-\gamma)\left(1-\gamma\left(1-z_\phi^{\pi_\phi^\star}\right)\right)} = \frac{\gamma\underline{p}}{(1-\gamma)\left(1-\gamma\left(1-\underline{p}\right)\right)}.$$

E.3.3 Proof of the claim (232)

Plugging in the definition of φ, we arrive at that for any policy π,

$$\begin{aligned}
\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \big\rangle &= V_\phi^{\star,\sigma}(0) - V_\phi^{\pi,\sigma}(0) \\
&\overset{\text{(i)}}{=} \frac{\gamma\underline{p}}{(1-\gamma)\left(1-\gamma(1-\underline{p})\right)} - \frac{\gamma z_\phi^\pi}{(1-\gamma)\Big(1-\gamma(1-z_\phi^\pi)\Big)} \\
&= \frac{\gamma\Big(\underline{p}-z_\phi^\pi\Big)}{\left(1-\gamma(1-\underline{p})\right)\Big(1-\gamma(1-z_\phi^\pi)\Big)} \overset{\text{(ii)}}{\geq} \frac{\gamma\Big(\underline{p}-z_\phi^\pi\Big)}{\left(1-\gamma\left(1-\underline{p}\right)\right)^2} \overset{\text{(iii)}}{=} \frac{\gamma(\underline{p}-\underline{q})\big(1-\pi(\phi \,|\, 0)\big)}{\left(1-\gamma\left(1-\underline{p}\right)\right)^2},
\end{aligned} \tag{267}$$

where (i) holds by applying Lemma 20, (ii) arises from $z_\phi^\pi \leq \underline{p}$ (see the definition of z_ϕ^π in (227) and the fact $\underline{p} \geq \underline{q} + \frac{3\Delta}{4}$ in (224)), and (iii) follows from the definition of z_ϕ^π in (227).

To further control (267), we consider it in two cases separately:

- Case 1: $\sigma \in \left(0, \frac{1-\gamma}{4}\right)$. In this case, applying Lemma 19 to (267) yields

$$\begin{aligned}
\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \big\rangle &\geq \frac{\gamma(\underline{p}-\underline{q})\big(1-\pi(\phi \,|\, 0)\big)}{\left(1-\gamma\left(1-\underline{p}\right)\right)^2} \geq \frac{\gamma\frac{3\Delta}{4}\big(1-\pi(\phi \,|\, 0)\big)}{\left(1-\gamma\left(1-\frac{5(1-\gamma)}{4}\right)\right)^2} \\
&\geq \frac{\Delta\big(1-\pi(\phi \,|\, 0)\big)}{9(1-\gamma)^2} = 2\varepsilon\big(1-\pi(\phi \,|\, 0)\big),
\end{aligned} \tag{268}$$

where the penultimate inequality follows from $\gamma \geq 3/4$, and the last inequality holds by taking the specification of Δ in (231) as follows:

$$\Delta = 18(1-\gamma)^2\varepsilon. \tag{269}$$

It is easily verified that taking $\varepsilon \leq \frac{1}{72(1-\gamma)}$ as in (230) directly leads to meeting the requirement in (222), i.e., $\Delta \leq \frac{1}{4}(1-\gamma)$.

- Case 2: $\sigma \in \big[\frac{1-\gamma}{4}, \infty\big)$. Similarly, applying Lemma 19 to (267) gives

  $$\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \big\rangle \geq \frac{\gamma\big(\underline{p}-\underline{q}\big)\big(1-\pi(\phi\,|\,0)\big)}{\big(1-\gamma\left(1-\underline{p}\right)\big)^2} \geq \frac{\gamma\frac{\sigma+1}{2}\Delta\big(1-\pi(\phi\,|\,0)\big)}{\min\Big\{1, \left(1-\gamma\left(1-(3+\sigma)\Delta\right)\right)^2\Big\}} \tag{270}$$

  Before continuing, it can be verified that

  $$\begin{aligned} 1-\gamma\left(1-(3+\sigma)\Delta\right) = 1-\gamma+\gamma(3+\sigma)\Delta &\overset{\text{(i)}}{\leq} 1-\gamma+(3+\sigma)\min\left\{\frac{1}{4}(1-\gamma), \frac{1}{2(\sigma+1)}\right\} \\ &\leq \min\left\{2(1+\sigma)(1-\gamma), \frac{3}{2}\right\}, \end{aligned} \tag{271}$$

  where (i) is obtained by $\Delta \leq \min\left\{\frac{1}{4}(1-\gamma), \frac{1}{2(1+\sigma)}\right\}$ (see (221)). Applying the above fact to (270) gives

  $$\begin{aligned} \big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \big\rangle &\geq \frac{\gamma\frac{\sigma+1}{2}\Delta\big(1-\pi(\phi\,|\,0)\big)}{\min\Big\{1, \left(1-\gamma\left(1-(3+\sigma)\Delta\right)\right)^2\Big\}} \overset{\text{(i)}}{\geq} \frac{3(\sigma+1)\Delta\big(1-\pi(\phi\,|\,0)\big)}{8\min\left\{4(1+\sigma)^2(1-\gamma)^2, 1\right\}} \\ &\geq \frac{\Delta\big(1-\pi(\phi\,|\,0)\big)}{\min\Big\{32(1+\sigma)(1-\gamma)^2, \frac{8}{3(1+\sigma)}\Big\}} = 2\varepsilon\big(1-\pi(\phi\,|\,0)\big), \end{aligned} \tag{272}$$

  where (i) holds by $\gamma \geq \frac{3}{4}$ and (270), and the last equality holds by the specification in (231):

  $$\Delta = \begin{cases} 64(1+\sigma)(1-\gamma)^2\varepsilon & \text{if } \sigma \in \Big[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\Big), \\ \frac{16}{3(1+\sigma)}\varepsilon & \text{if } \sigma > \frac{1}{3(1-\gamma)}. \end{cases} \tag{273}$$

  As a result, it is easily verified that the requirement in (222)

  $$\Delta \leq \min\left\{\frac{1}{4}(1-\gamma), \frac{1}{2(1+\sigma)}\right\} \tag{274}$$

  is met if we let

  $$\varepsilon \leq \begin{cases} \frac{1}{256(1+\sigma)(1-\gamma)} & \text{if } \sigma \in \Big[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\Big), \\ \frac{3}{32} & \text{if } \sigma > \frac{1}{3(1-\gamma)}, \end{cases} \tag{275}$$

  as in (230).

The proof is then completed by summing up the results in the above two cases.

F Proof for the offline setting

F.1 Proof of the upper bounds: Corollary 1 and Corollary 3

As the proofs of Corollary 1 and Corollary 3 are similar, without loss of generality, we first focus on Corollary 1 in the case of TV distance.

To begin with, suppose we have access to in total N_{b} independent sample tuples $\{s_i, a_i, a_i', r_i\}_{i=1}^{N_{\mathsf{b}}}$ from either the generative model or a historical dataset. We denote the number of samples generated based on the state-action pair (s,a) as $N(s,a)$, i.e,

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad N(s,a) = \sum_{i=1}^{N_{\mathsf{b}}} \mathbb{1}\big\{s_i = s, a_i = a\big\}. \tag{276}$$

Then according to (13), we can construct an empirical nominal transition for DRVI (Algorithm 1).

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{P}^0(s' \mid s,a) \coloneqq \frac{1}{N(s,a)} \sum_{i=1}^{N(s,a)} \mathbb{1}\big\{s_i = s, a_i = a, s_i' = s'\big\}. \tag{277}$$

Armed with the above estimate of nominal transition kernel, we introduce a slightly general version of Theorem 1, which follows directly from the same proof routine in Appendix 5.2.2.

Theorem 5 (Upper bound under TV distance)**.** *Let the uncertainty set be $\mathcal{U}_\rho^\sigma(\cdot) = \mathcal{U}_{\mathsf{TV}}^\sigma(\cdot)$, as specified by the TV distance (9). Consider any discount factor $\gamma \in \big[\frac{1}{4}, 1\big)$, uncertainty level $\sigma \in (0,1)$, and $\delta \in (0,1)$. Based on the empirical nominal transition kernel in (277), let $\widehat{\pi}$ be the output policy of Algorithm 1 after $T = C_1 \log\big(\frac{N_{\mathsf{b}}}{1-\gamma}\big)$ iterations. Then with probability at least $1-\delta$, one has*

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \le \varepsilon \tag{278}$$

for any $\varepsilon \in \Big(0, \sqrt{1/\max\{1-\gamma,\sigma\}}\Big]$, as long as

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad N(s,a) \ge \frac{C_2}{(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2} \log\left(\frac{SAN_{\mathsf{b}}}{(1-\gamma)\delta}\right). \tag{279}$$

Here, $C_1, C_2 > 0$ are some large enough universal constants.

Furthermore, we invoke a fact derived from basic concentration inequalities (Li et al., 2024) as below.

Lemma 22. *Consider any $\delta \in (0,1)$ and a dataset with N_{b} independent samples satisfying Assumption 1. With probability at least $1-\delta$, the quantities $\{N(s,a)\}$ obey*

$$\max\left\{N(s,a), \frac{2}{3}\log\frac{N_{\mathsf{b}}}{\delta}\right\} \ge \frac{N_{\mathsf{b}}\mu^{\mathsf{b}}(s,a)}{12} \tag{280}$$

simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Now we are ready to verify Corollary 1. Armed with a historical dataset $\mathcal{D}^{\mathsf{b}}$ with N_{b} independent samples that obeys Assumption 1, one has with probability at least $1-\delta$,

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad N(s,a) \ge \frac{N_{\mathsf{b}}\mu^{\mathsf{b}}(s,a)}{12} \ge \frac{N_{\mathsf{b}}\mu_{\min}}{12} \tag{281}$$

as long as $N_{\mathsf{b}} \ge \frac{8\log\frac{N_{\mathsf{b}}}{\delta}}{\mu_{\min}} \ge \frac{8\log\frac{N_{\mathsf{b}}}{\delta}}{\mu^{\mathsf{b}}(s,a)}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Consequently, given $N_{\mathsf{b}} \ge \frac{8\log\frac{N_{\mathsf{b}}}{\delta}}{\mu_{\min}}$, applying Theorem 5 with the fact $N(s,a) \ge \frac{N_{\mathsf{b}}\mu_{\min}}{12}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ (see (281)) directly leads to: DRVI can achieve an ε-optimal policy as long as

$$N(s,a) \ge \frac{N_{\mathsf{b}}\mu_{\min}}{12} \ge \frac{C_2}{(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2} \log\left(\frac{SAN_{\mathsf{b}}}{(1-\gamma)\delta}\right), \tag{282}$$

namely

$$N_{\mathsf{b}} \ge \frac{C_3}{\mu_{\min}(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2} \log\left(\frac{SAN_{\mathsf{b}}}{(1-\gamma)\delta}\right), \tag{283}$$

where C_3 is some large enough universal constant. Note that the above inequality directly implies $N_{\mathsf{b}} \ge \frac{8\log\frac{N_{\mathsf{b}}}{\delta}}{\mu_{\min}}$. This complete the proof of Corollary 1. The same argument holds for Corollary 3.

F.2 Proof of the lower bounds: Corollary 2 and Corollary 4

Analogous to Appendix F.1, without loss of generality, we firstly focus on verifying Corollary 2, where we use the TV distance to measure the uncertainty set.

We stick to the two hard instances $\mathcal{M}_0$ and $\mathcal{M}_1$ (i.e., $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$) constructed in the proof for Theorem 2 (Appendix 5.3.1). Recall that the state space is defined as $\mathcal{S} = \{0,1,2,\cdots,S-1\}$, where the corresponding action space for any state $s \in \{2,3,\cdots,S-1\}$ is $\mathcal{A} = \{0,1,2,\cdots,A-1\}$. For states $s=0$ or $s=1$, the action space is only $\mathcal{A}' = \{0,1\}$. Hence, for a given factor $\mu_{\min} \in (0, \frac{1}{SA}]$, we can construct a historical dataset $\mathcal{D}^{\mathsf{b}}$ with N_{b} samples such that the data coverage becomes the smallest over the state-action pairs $(0,0)$ and $(0,1)$, i.e.,

$$\mu^{\mathsf{b}}(0,0) = \mu^{\mathsf{b}}(0,1) = \mu_{\min} \quad \text{and} \quad \mu^{\mathsf{b}}(s,a) = \frac{1-2\mu_{\min}}{(S-2)A+2}, \quad \forall s \in \{1,2,\cdots,S\}. \tag{284}$$

Armed with the above hard instance and historical dataset, we follow the proof procedure in Appendix 5.3.2 to verify the corollary. Our goal is to distinguish between the two hypotheses $\phi \in \{0,1\}$ by considering the minimax probability of error as follows:

$$p_{\mathrm{e}} \coloneqq \inf_{\psi} \max\big\{\mathbb{P}_0(\psi \neq 0),\, \mathbb{P}_1(\psi \neq 1)\big\}, \tag{285}$$

where the infimum is taken over all possible tests ψ constructed from the samples in $\mathcal{D}^{\mathsf{b}}$.

Recall that we denote μ_ϕ (resp. $\mu_\phi(s)$) as the distribution of a sample tuple (s_i, a_i, s_i') under the nominal transition kernel P^ϕ associated with $\mathcal{M}_\phi$ and the samples are generated independently. Analogous to (80), one has

$$\begin{aligned} p_{\mathrm{e}} &\geq \frac{1}{4}\exp\Big(-N_{\mathsf{b}}\mathsf{KL}\big(\mu_0 \,\|\, \mu_1\big)\Big) \\ &= \frac{1}{4}\exp\Big\{-N_{\mathsf{b}}\mu_{\min}\Big(\mathsf{KL}\big(P^0(\cdot\,|\,0,0) \,\|\, P^1(\cdot\,|\,0,0)\big) + \mathsf{KL}\big(P^0(\cdot\,|\,0,1) \,\|\, P^1(\cdot\,|\,0,1)\big)\Big)\Big\}, \end{aligned} \tag{286}$$

where the last inequality holds by observing that

$$\begin{aligned} \mathsf{KL}\big(\mu_0 \,\|\, \mu_1\big) &= \sum_{s,a,s'} \mu^{\mathsf{b}}(s,a)\mathsf{KL}\big(P^0(s'\,|\,s,a) \,\|\, P^1(s'\,|\,s,a)\big) \\ &= \sum_{a\in\{0,1\}} \mu^{\mathsf{b}}(0,a)\mathsf{KL}\big(P^0(\cdot\,|\,0,a) \,\|\, P^1(\cdot\,|\,0,a)\big) = \mu_{\min}\sum_{a\in\{0,1\}} \mathsf{KL}\big(P^0(\cdot\,|\,0,a) \,\|\, P^1(\cdot\,|\,0,a)\big). \end{aligned} \tag{287}$$

Here, the last line holds by the fact that $P^0(\cdot\,|\,s,a)$ and $P^1(\cdot\,|\,s,a)$ (associated with $\mathcal{M}_0$ and $\mathcal{M}_1$) only differ from each other in state-action pairs $(0,0)$ and $(0,1)$, each has a visitation density of $\mu_{\min}$. Consequently, following the same routine from (81) to the end of Appendix 5.3.2, we applying (82) and (83) with $N = N_{\mathsf{b}}\mu_{\min}$ and complete the proof by showing: if the sample size is selected as

$$N_{\mathsf{b}}\mu_{\min} = N \leq \frac{c_1 \log 2}{8192(1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2}, \tag{288}$$

then one necessarily has

$$p_e = \inf_{\widehat{\pi}} \max\Big\{\mathbb{P}_0\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big),\, \mathbb{P}_1\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big)\Big\} \geq \frac{1}{8}. \tag{289}$$

We can follow the same argument to complete the proof of Corollary 4.