Journal of Machine Learning Research 25 (2024) 1-91 Submitted 12/22; Revised 12/23; Published 7/24

Distributionally Robust Model-Based Offline Reinforcement Learning with Near-Optimal Sample Complexity

Laixi Shi LAIXIS@CALTECH.EDU
Computing Mathematical Sciences
California Institute of Technology
Pasadena, CA, 91125, USA

Yuejie Chi YUEJIECHI@CMU.EDU
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Editor: Shipra Agrawal

Abstract

This paper concerns the central issues of model robustness and sample efficiency in offline reinforcement learning (RL), which aims to learn to perform decision making from history data without active exploration. Due to uncertainties and variabilities of the environment, it is critical to learn a robust policy—with as few samples as possible—that performs well even when the deployed environment deviates from the nominal one used to collect the history dataset. We consider a distributionally robust formulation of offline RL, focusing on tabular robust Markov decision processes with an uncertainty set specified by the Kullback-Leibler divergence in both finite-horizon and infinite-horizon settings. To combat with sample scarcity, a model-based algorithm that combines distributionally robust value iteration with the principle of pessimism in the face of uncertainty is proposed, by penalizing the robust value estimates with a carefully designed data-driven penalty term. Under a mild and tailored assumption of the history dataset that measures distribution shift without requiring full coverage of the state-action space, we establish the finite-sample complexity of the proposed algorithms. We further develop an information-theoretic lower bound, which suggests that learning RMDPs is at least as hard as the standard MDPs when the uncertainty level is sufficient small, and corroborates the tightness of our upper bound up to polynomial factors of the (effective) horizon length for a range of uncertainty levels. To the best our knowledge, this provides the first provably near-optimal robust offline RL algorithm that learns under model uncertainty and partial coverage.

Keywords: offline/batch reinforcement learning, distributional robustness, pessimism, model-based reinforcement learning, finite-sample analysis

1. Introduction

Reinforcement learning (RL) concerns about finding an optimal policy that maximizes an agent's expected total reward in an unknown environment. A fundamental challenge of deploying RL to real-world applications is the limited ability to explore or interact with the environment, due to resources, time, or safety constraints. Offline RL, or batch RL,

seeks to circumvent this challenge by resorting to history data—which are often collected by executing some possibly unknown behavior policy in the past—with the hope that the history data might already provide significant insights about the targeted optimal policy without further exploration (Levine et al., 2020).

Besides maximizing the expected total reward, perhaps an equally important goal—to say the least—for an RL agent is safety and robustness (Garcıa and Fernández, 2015), especially in high-stake applications such as robotics, autonomous driving, clinical trials, financial investments, and so on (Choi et al., 2009; Schulman et al., 2013). It has been observed that a standard RL agent trained in an ideal environment might be extremely sensitive and fail catastrophically when the deployed environment is subject to small adversarial perturbations (Zhang et al., 2020). Consequently, robust RL has attracted a surge of attentions with the goal to learn an optimal policy that is robust to environment perturbations. In fact, providing robustness guarantees becomes even more relevant in the offline setting, which can be formulated as *robust offline RL*, since the history data is often inevitably collected from a timeframe where it is no longer reasonable to assume model stillness, due to the highly non-stationary and time-varying dynamics of many real-world applications. Altogether, this naturally leads to a question:

> *Can we learn a near-optimal policy which is robust with respect to uncertainties and variabilities of the environments using as few history samples as possible?*

1.1 Challenges and premises in robust offline RL

Despite significant amount of recent activities in robust RL and offline RL, addressing model uncertainty and sample efficiency simultaneously remains challenging due to several key issues that we single out below.

- *Distribution shift.* The history data is generated by following some behavior policy in an outdated environment, which can result in a data distribution that is heavily deviated from the desired one, i.e., induced by the target policy in the deployed environment.
- *Partial and limited coverage.* The history data might only provide partial and limited coverage over the entire state-action space, where the limited sample size leads to a poor estimate of the associated model parameters, and consequently, unreliable policy learning outcomes.

Understanding the implications of—and designing algorithms that work around—these challenges play a major role in advancing the state-of-the-art of robust offline RL. In particular, two prevalent algorithmic ideas, distributional robustness and pessimism, are called out as our guiding principles.

- *Distributional robustness.* Instead of finding an optimal policy in a fixed environment, motivated by the literature in distributionally robust optimization (Delage and Ye, 2010), one might seek to find a policy that achieves the best worst-case performance for all the environments in some uncertainty set around the offline environment, as formalized in the framework of robust RL (Iyengar, 2005; Nilim and El Ghaoui, 2005).

- *Pessimism.* When the samples are scarce, it is wise to act with caution based on the principle of pessimism, where one subtracts a penalty term—representing the confidence of the corresponding estimate—from the value functions to avoid excessive risk. Encouragingly, pessimism has been recently shown as an indispensable ingredient to achieve sample efficiency in offline RL without requiring full coverage (Jin et al., 2021; Rashidinejad et al., 2021; Li et al., 2022), as long as the trajectory of the behavior policy provides sufficient overlap with that of the target policy.

While these two ideas have been proven useful for robust RL and offline RL *separately*, tackling robust offline RL needs novel ingredients that go significantly beyond a naïve combination of existing techniques. This is because, in robust offline RL, one needs to handle the distribution shift induced not only by the behavior policy, but also by model perturbations, thus the penalty term derived from the pessimism principle in standard offline RL is no longer applicable. Indeed, while the value function of standard RL depends linearly with respect to the transition kernel, the dependency between the nominal transition kernel and the robust value function unfortunately becomes highly nonlinear—even without a closed-form expression—making the control of statistical uncertainty extremely challenging in robust offline RL.

1.2 Main contributions

In this work, we provide an affirmative answer to the question raised earlier, by developing a provably efficient model-based algorithm that learns a near-optimal *distributionally-robust* policy from a minimal number of offline samples. Specifically, we consider a Robust Markov Decision Process (RMDP) with S states, A actions in both the nonstationary finite-horizon setting (with horizon length H) and the discounted infinite-horizon setting (with discount factor γ). Different from standard MDPs, RMDPs specify a family transition kernels, which lie within an uncertainty set taken as a small ball of size σ around a nominal transition kernel with respect to the Kullback-Leibler (KL) divergence. Given K episodes (resp. N transitions) of history data drawn by following some behavior policy π^{b} under the nominal transition kernel in the finite-horizon (resp. infinite-horizon) setting, our goal is to learn the optimal robust policy $\pi^\star$ in the maximin sense, which has the best worst-case value for all the transition kernels within the uncertainty set (Iyengar, 2005; Nilim and El Ghaoui, 2005). Our main results are summarized below.

- We introduce a notion called *robust single-policy clipped concentrability coefficient* $C^\star_{\mathsf{rob}} \in [1/S, \infty]$ to quantify the quality of history data, which measures the distribution shift between the behavior policy π^{b} and the optimal robust policy $\pi^\star$ in the presence of model perturbations, without requiring full coverage of the entire state-action space by the behavior policy. In contrast, prior algorithms (Yang et al., 2022; Zhou et al., 2021; Panaganti and Kalathil, 2022)—using simulator or offline data—all require full coverage of the entire state-action space.

- We propose a novel pessimistic variant of distributionally robust value iteration with a plug-in estimate of the nominal transition kernel (Iyengar, 2005; Nilim and El Ghaoui, 2005), called DRVI-LCB, by penalizing the robust value estimates with a carefully designed data-driven penalty term. We demonstrate that DRVI-LCB finds an ε-optimal

robust policy as soon as the sample size is above $\widetilde{O}\left(\frac{SC^\star_{\mathsf{rob}}H^5}{P^\star_{\mathsf{min}}\sigma^2\varepsilon^2}\right)$ for the finite-horizon setting and $\widetilde{O}\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}\sigma^2(1-\gamma)^4\varepsilon^2}\right)$ for the infinite-horizon setting, up to some logarithmic factor after a burn-in cost independent of ε. Here, $P^\star_{\mathsf{min}}$ is the smallest positive state transition probability of the optimal robust policy $\pi^\star$ under the nominal kernel.

- To complement the upper bound, we further develop information-theoretic lower bounds for a range of uncertainty levels, showing there exists some transition kernel such that at least $\Omega\left(\frac{SC^\star_{\mathsf{rob}}H^4}{\varepsilon^2}\right)$ samples (resp. $\Omega\left(\frac{SC^\star_{\mathsf{rob}}}{(1-\gamma)^3\varepsilon^2}\right)$ samples) are needed to find an ε-optimal robust policy when the uncertainty level $\sigma \lesssim 1/H$ (resp. $\sigma \lesssim (1-\gamma)$), and at least $\Omega\left(\frac{SC^\star_{\mathsf{rob}}H^3}{P^\star_{\mathsf{min}}\sigma^2\varepsilon^2}\right)$ samples (resp. $\Omega\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}\sigma^2(1-\gamma)^2\varepsilon^2}\right)$ samples) are needed to find an ε-optimal robust policy when the uncertainty level $\sigma \asymp \log(1/P^\star_{\mathsf{min}})$, regardless of the choice of algorithms in the finite-horizon (resp. infinite-horizon) setting. Hence, this suggests that learning RMDPs is at least as hard as the standard MDP (Li et al., 2022) when the uncertainty level is sufficiently small, and corroborates the near-optimality of DRVI-LCB with respect to all key parameters up to a polynomial factor of the horizon length H (resp. the effective horizon length $\frac{1}{1-\gamma}$) for a range of uncertainty levels ($\sigma \asymp \log(1/P^\star_{\mathsf{min}})$).

To the best of our knowledge, our paper is the first work to execute the principle of pessimism in a data-driven manner for robust offline RL, leading to the first provably efficient algorithm that learns under simultaneous model uncertainty and partial coverage of the history dataset. See Table 1 for a summary.

Comparison with prior art under full coverage. Prior works (Yang et al., 2022; Zhou et al., 2021; Panaganti and Kalathil, 2022) have only addressed the infinite-horizon setting under full coverage of the history data. Fortunately, our results also seamlessly cover this easier scenario, by replacing $C^\star_{\mathsf{rob}}$ with A. Specializing our result to this setting to facilitate comparison, DRVI-LCB finds an ε-optimal robust policy with at most $\widetilde{O}\left(\frac{SA}{P^\star_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}\right)$ samples, which depends linearly with respect to the size of the state space S (ignoring other parameters). In contrast, all prior works (Yang et al., 2022; Zhou et al., 2021; Panaganti and Kalathil, 2022) incur sample complexities that scale at least quadratically with respect to the size of the state space S. In addition, our bound improves the *exponential* dependency on $\frac{1}{1-\gamma}$ of Zhou et al. (2021); Panaganti and Kalathil (2022) to a *polynomial* dependency, as well as the *quadratic* dependency on $1/P_{\mathsf{min}}$ (which satisfies $P_{\mathsf{min}} \le P^\star_{\mathsf{min}}$) of Yang et al. (2022) to a *linear* one on $1/P^\star_{\mathsf{min}}$. These improvements further corroborate the benefit of the proposed DRVI-LCB even under full coverage. See Table 2 for detailed comparisons.

1.3 Related works

We shall focus on the closely related works on offline RL and distributionally robust RL.

Offline RL. Focusing on the task of learning an optimal policy from offline data, a significant amount of prior arts sets to understand the sample complexity and efficacy of offline RL under different assumptions of the history dataset. A bulk of prior results requires the history data to cover all the state-action pairs, under assumptions such as uniformly

Horizon	Algorithm	Coverage	Sample complexity	Uncertainty level
infinite-horizon	DRVI-LCB **(this work)**	partial	$\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}$	full range
	Lower bound **(this work)**	partial	$\frac{SC^\star_{\mathsf{rob}}}{(1-\gamma)^3\varepsilon^2}$	$\sigma \lesssim (1-\gamma)$
	Lower bound **(this work)**	partial	$\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\varepsilon^2}$	$\sigma \asymp \log(1/P^\star_{\mathsf{min}})$
finite-horizon	DRVI-LCB **(this work)**	partial	$\frac{SC^\star_{\mathsf{rob}}H^5}{P^\star_{\mathsf{min}}\sigma^2\varepsilon^2}$	full range
	Lower bound **(this work)**	partial	$\frac{SC^\star_{\mathsf{rob}}H^4}{\varepsilon^2}$	$\sigma \lesssim 1/H$
	Lower bound **(this work)**	partial	$\frac{SC^\star_{\mathsf{rob}}H^3}{P^\star_{\mathsf{min}}\sigma^2\varepsilon^2}$	$\sigma \asymp \log(1/P^\star_{\mathsf{min}})$

Table 1: Our results for finding an ε-optimal robust policy in the infinite/finite-horizon robust MDPs with an uncertainty set measured with respect to the KL divergence using history data under partial coverage. The sample complexities included in the table are valid for sufficiently small ε, with all logarithmic factors omitted. Here, σ is the uncertainty level, S is the size of the state space, H is the horizon length for the finite-horizon setting, γ is the discount factor for the infinite-horizon setting, $C^\star_{\mathsf{rob}}$ is the robust single-policy clipped concentrability coefficient, and $P^\star_{\mathsf{min}}$ is the smallest positive state transition probability of the nominal kernel *visited by the optimal robust policy* $\pi^\star$.

bounded concentrability coefficients (Chen and Jiang, 2019; Munos, 2005) and uniformly lower bounded data visitation distribution (Yin and Wang, 2021; Yin et al., 2021), where the latter assumption is also related to studies of asynchronous Q-learning (Li et al., 2021). More recently, the principle of pessimism has been investigated for offline RL in both model-based (Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021; Li et al., 2022) and model-free algorithms (Kumar et al., 2020; Shi et al., 2022; Yan et al., 2023; Woo et al., 2024), without the stringent requirement of full coverage. In particular, Li et al. (2022) established the near-minimax optimality of a pessimistic variant of value iteration under the single-policy clipped concentrability of history data, which inspired our algorithm design in the distributionally robust setting.

Distributionally robust RL. While distributionally robust optimization has been mainly investigated in the context of supervised learning (Rahimian and Mehrotra, 2019; Gao, 2022; Bertsimas et al., 2018; Duchi and Namkoong, 2021; Blanchet and Murthy, 2019; Sinha et al., 2018), distributionally robust dynamic programming has also attracted considerable amount of attention, e.g. Iyengar (2005); Nilim and Ghaoui (2003); Xu and Mannor (2012); Nilim and El Ghaoui (2005), where natural robust extensions to the standard Bell-

Problem type	Algorithm	Coverage	Sample complexity
infinite-horizon	DRVI (Zhou et al., 2021)	full	$\frac{S^2A\exp\left(O(\frac{1}{1-\gamma})\right)}{(1-\gamma)^4\sigma^2\varepsilon^2}$
	REVI/DRVI (Panaganti and Kalathil, 2022)	full	$\frac{S^2A\exp\left(O(\frac{1}{1-\gamma})\right)}{(1-\gamma)^4\sigma^2\varepsilon^2}$
	DRVI (Yang et al., 2022)	full	$\frac{S^2A}{P_{\mathsf{min}}^2(1-\gamma)^4\sigma^2\varepsilon^2}$
	DRVI-LCB **(this work)**	full	$\frac{SA}{P_{\mathsf{min}}^\star(1-\gamma)^4\sigma^2\varepsilon^2}$
finite-horizon	DRVI-LCB **(this work)**	full	$\frac{SAH^5}{P_{\mathsf{min}}^\star\sigma^2\varepsilon^2}$

Table 2: Comparisons between our results and prior arts for finding an ε-optimal robust policy in the infinite/finite-horizon robust MDPs with an uncertainty set measured with respect to the KL divergence under full coverage of the history data. The sample complexities included in the table are valid for sufficiently small ε, with all logarithmic factors omitted. Here, σ is the uncertainty level, S is the size of the state space, A is the size of the action space, H is the horizon length for the finite-horizon setting, γ is the discount factor for the infinite-horizon setting, $P_{\mathsf{min}}^\star$ is the smallest positive state transition probability of the nominal kernel *visited by the optimal robust policy* $\pi^\star$, and P_{min} is the smallest positive state transition probability of the nominal kernel; it holds $P_{\mathsf{min}} \le P_{\mathsf{min}}^\star$.

man machineries are developed under mild assumptions. Targeting robust MDPs, empirical and theoretical works have been widely explored under different forms of uncertainty sets (Iyengar, 2005; Xu and Mannor, 2012; Wolff et al., 2012; Kaufman and Schaefer, 2013; Ho et al., 2018; Smirnova et al., 2019; Ho et al., 2021; Goyal and Grand-Clement, 2022; Derman and Mannor, 2020; Tamar et al., 2014; Badrinath and Kalathil, 2021; Abdullah et al., 2019; Hou et al., 2020; Song and Zhao, 2020; Yang, 2017; Wang et al., 2022; Ding et al., 2023). Nonetheless, the majority of prior theoretical analyses focus on planning with an exact knowledge of the uncertainty set (Iyengar, 2005; Xu and Mannor, 2012; Tamar et al., 2014), or are asymptotic in nature (Roy et al., 2017).

A number of robust RL algorithms were proposed recently with an emphasis on finite-sample performance guarantees under different data generating mechanisms. Wang and Zou (2021) proposed a robust Q-learning algorithm with an R-contamination uncertain set for the online setting, which achieves a similar bound as its non-robust counterpart. Badrinath and Kalathil (2021) proposed a model-free algorithm for the online setting with linear function approximation to cope with large state spaces. Yang et al. (2022); Panaganti and Kalathil (2022) developed sample complexities for a model-based robust RL algorithm with a variety of uncertainty sets where the data are collected using a generative model. In addition, Zhou et al. (2021) examined the uncertainty set defined by the KL divergence

for offline data with uniformly lower bounded data visitation distribution. These works all require full coverage of the state-action space, whereas ours is the first one to leverage the principle of pessimism in robust offline RL.

Since the first appearance of our paper on arXiv in August 2022, a few more papers have emerged that also tackle the sample complexity of robust RL algorithms. For example, Wang et al. (2023a,b) developed finite-sample complexity bounds for robust variants of Q-learning with the generative model when the uncertainty set is measured by KL divergence; in particular, the improved bound of variance-reduced robust Q-learning (Wang et al., 2023b) becomes independent of the size of the uncertainty set when it is sufficiently small with respect to the minimal support probability of the nominal kernel at a price of worse dependency with $1/P^{\star}_{\mathsf{min}}$. Shi et al. (2023) provided near-optimal sample complexity bounds for model-based robust RL algorithms with the generative model when the uncertainty set is measured by the total variation or chi-square distances, which highlighted that different uncertainty sets can lead to drastically different sample complexities, and hence, statistical consequences. Wang et al. (2024) studied the sample complexity of offline robust RL with the total variation uncertainty set under linear function approximation, and Shi et al. (2024) extended the study of robust MDPs to the setting of robust Markov games.

1.4 Notation and paper organization

Throughout this paper, we denote by $\Delta(\mathcal{S})$ the probability simplex over a set $\mathcal{S}$, and introduce the notation $[H] \coloneqq \{1, \cdots, H\}$ for any positive integer $H > 0$. In addition, for any vector $x = \big[x(s,a)\big]_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{SA}$ (resp. $x = \big[x(s)\big]_{s\in\mathcal{S}} \in \mathbb{R}^{S}$) that constitutes certain values for each state-action pair (resp. state), we overload the notation by letting $x^2 = \big[x(s,a)^2\big]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ (resp. $x^2 = \big[x(s)^2\big]_{s\in\mathcal{S}}$). 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 $1 \le i \le n$. Finally, the Kullback-Leibler (KL) divergence for any two distributions P and Q is denoted as $\mathsf{KL}(P \,\|\, Q)$.

The rest of this paper is organized as follows. Section 2 provides the backgrounds and introduces the distributionally robust formulation of finite-horizon MDPs in the offline setting under partial coverage. Section 3 presents the proposed algorithm and provides sample complexity guarantees. Section 4 develops the corresponding results for the infinite-horizon setting. Section 5 demonstrate the performance of the proposed algorithm through numerical experiments. Finally, we conclude in Section 6. The detailed proofs are postponed to the appendix.

2. Problem formulation: episodic finite-horizon RMDPs

2.1 Basics of finite-horizon episodic tabular MDPs

Consider an episodic finite-horizon MDP, represented by $\mathcal{M} = \big(\mathcal{S}, \mathcal{A}, H, P \coloneqq \{P_h\}_{h=1}^{H}, \{r_h\}_{h=1}^{H}\big)$, where $\mathcal{S} = \{1, \cdots, S\}$ and $\mathcal{A} = \{1, \cdots, A\}$ are the finite state and action spaces, respectively, H is the horizon length, $P_h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ (resp. $r_h : \mathcal{S}\times\mathcal{A} \to [0,1]$)

denotes the probability transition kernel (resp. reward function) at step h ($1 \le h \le H$).[1] For any transition kernel P, we introduce the S-dimensional distribution vectors

$$P_{h,s,a} \coloneqq P_h(\cdot \,|\, s,a) \in [0,1]^{1\times S}, \qquad \forall (h,s,a) \in [H]\times \mathcal{S}\times \mathcal{A} \tag{1}$$

to represent the probability transition vector in state s when taking action a at step h.

Denote by $\pi = \{\pi_h\}_{h=1}^H$ as the policy or action selection rule of an agent, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the action selection probability over the action space; when the policy is deterministic, we slightly abuse the notation and refer to $\pi_h(s)$ as the action selected by policy π in state s at step h. The value function $V^{\pi,P} = \{V_h^{\pi,P}\}_{h=1}^H$ of policy π with a transition kernel P is defined by

$$\forall (h,s) \in [H]\times \mathcal{S}: \qquad V_h^{\pi,P}(s) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=h}^{H} r_t\big(s_t,a_t\big) \,\Big|\, s_h = s\right], \tag{2}$$

where the expectation is taken over the randomness of the trajectory $\{s_h,a_h,r_h\}_{h=1}^H$ generated by executing policy π, namely, $a_t \sim \pi_t(s_t)$, and $s_{t+1} \sim P_t(\cdot \,|\, s_t,a_t)$. Similarly, the Q-function $Q^{\pi,P} = \{Q_h^{\pi,P}\}_{h=1}^H$ of policy π is defined as

$$\forall (h,s,a) \in [H]\times \mathcal{S}\times \mathcal{A}: \qquad Q_h^{\pi,P}(s,a) \coloneqq r_h(s,a) + \mathbb{E}_{\pi,P}\left[\sum_{t=h+1}^{H} r_t(s_t,a_t) \,\Big|\, s_h = s, a_h = a\right], \tag{3}$$

where the expectation is again taken over the randomness of the trajectory.

Moreover, when the initial state s_1 is drawn from a given distribution ρ, let $d_h^{\pi,P}(s \,|\, \rho)$ and $d_h^{\pi,P}(s,a \,|\, \rho)$ denote respectively the state occupancy distribution and the state-action occupancy distribution induced by π at time step $h \in [H]$. In particular, we often dropped the dependency with respect to ρ whenever it is clear from the context, by simply writing $d_h^{\pi,P}(s) := d_h^{\pi,P}(s \,|\, \rho)$ and $d_h^{\pi,P}(s,a) := d_h^{\pi,P}(s,a \,|\, \rho)$, i.e.,

$$\forall (h,s) \in [H]\times \mathcal{S}: \qquad d_h^{\pi,P}(s) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \rho, \pi, P), \tag{4a}$$

$$\forall (h,s,a) \in [H]\times \mathcal{S}\times \mathcal{A}: \quad d_h^{\pi,P}(s,a) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \rho, \pi, P)\, \pi_h(a \,|\, s), \tag{4b}$$

which are conditioned on $s_1 \sim \rho$ and the event that all actions and states are drawn according to policy π and transition kernel P.

2.2 Distributionally robust MDPs

In this section, we focus on finite-horizon episodic distributionally robust MDPs (RMDPs), denoted by $\mathcal{M}_{\mathsf{rob}} = \big(\mathcal{S},\mathcal{A},H,\mathcal{U}^{\sigma}(P^0),\{r_h\}_{h=1}^H\big)$. Different from standard MDPs, we now consider an ensemble of probability transition kernels or models within an uncertainty set centered around a nominal one $P^0 = \{P_h^0\}_{h=1}^H$, where the distance between the transition kernels is measured in terms of the Kullback-Leibler (KL) divergence. Specifically, given

1. Without loss of generality, we assume the reward function is deterministic, fixed, and normalized to be within $[0,1]$; it is straightforward to generalize our framework to incorporate random rewards with uncertainties.

an uncertainty level $\sigma > 0$, the uncertainty set around P^0, which satisfies the so-called (s,a)-rectangularity condition (Wiesemann et al., 2013), is specified as

$$\mathcal{U}^\sigma(P^0) \coloneqq \otimes\, \mathcal{U}^\sigma(P^0_{h,s,a}), \qquad \mathcal{U}^\sigma(P^0_{h,s,a}) \coloneqq \left\{P_{h,s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{h,s,a} \,\|\, P^0_{h,s,a}\right) \le \sigma\right\}, \quad (5)$$

where $\otimes$ denote the Cartesian product. In words, the KL divergence between the true transition probability vector and the nominal one at each state-action pair is at most σ; moreover, the RMDP reduces to the standard MDP when $\sigma = 0$.

Instead of evaluating a policy in a fixed MDP, the performance of a policy in the RMDP is evaluated based on its worst-case—i.e., smallest—value function over all the instances in the uncertainty set. That is, we define the *robust value function* $V^{\pi,\sigma} = \{V_h^{\pi,\sigma}\}_{h=1}^H$ and the *robust Q-function* $Q^{\pi,\sigma} = \{Q_h^{\pi,\sigma}\}_{h=1}^H$ respectively as

$$\begin{aligned}\forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}: \quad V_h^{\pi,\sigma}(s) &\coloneqq \inf_{P\in\mathcal{U}^\sigma(P^0)} V_h^{\pi,P}(s),\\ Q_h^{\pi,\sigma}(s,a) &\coloneqq \inf_{P\in\mathcal{U}^\sigma(P^0)} Q_h^{\pi,P}(s,a),\end{aligned}$$

where the infimum is taken over the uncertainty set of transition kernels.

Optimal robust policy. For finite-horizon RMDPs, it has been established that there exists at least one deterministic policy that maximizes the robust value function and the robust Q-function simultaneously (Iyengar, 2005; Nilim and El Ghaoui, 2005). In view of this, we shall denote a deterministic policy $\pi^\star = \{\pi_h^\star\}_{h=1}^H$ as an optimal robust policy throughout this paper. The resulting *optimal robust value function* $V^{\star,\sigma} = \{V_h^{\star,\sigma}\}_{h=1}^H$ and *optimal robust Q-function* $Q^{\star,\sigma} = \{Q_h^{\star,\sigma}\}_{h=1}^H$ are denoted by

$$\forall (h,s) \in [H]\times\mathcal{S}: \quad V_h^{\star,\sigma}(s) \coloneqq V_h^{\pi^\star,\sigma}(s) = \max_\pi V_h^{\pi,\sigma}(s), \quad (6a)$$

$$\forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}: \quad Q_h^{\star,\sigma}(s,a) \coloneqq Q_h^{\pi^\star,\sigma}(s,a) = \max_\pi Q_h^{\pi,\sigma}(s,a). \quad (6b)$$

Similar to (4), we adopt the following short-hand notation for the occupancy distributions associated with the optimal policy:

$$\forall (h,s) \in [H]\times\mathcal{S}: \qquad d_h^{\star,P}(s) \coloneqq d_h^{\pi^\star,P}(s), \quad (7a)$$

$$\forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}: \qquad d_h^{\star,P}(s,a) \coloneqq d_h^{\pi^\star,P}(s,a) = d_h^{\star,P}(s)\,\mathbb{1}\{a = \pi_h^\star(s)\}. \quad (7b)$$

Robust Bellman equations. It turns out the Bellman's principle of optimality can be extended naturally to its robust counterpart (Iyengar, 2005; Nilim and El Ghaoui, 2005), which plays a fundamental role in solving the RMDP. To begin with, for any policy π, the robust value function and robust Q-function satisfy the following *robust Bellman consistency equation*:

$$\forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}: \qquad Q_h^{\pi,\sigma}(s,a) = r_h(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V_{h+1}^{\pi,\sigma}. \quad (8)$$

Additionally, the optimal robust Q-function obeys the *robust Bellman optimality equation*:

$$\forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}: \qquad Q_h^{\star,\sigma}(s,a) = r_h(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V_{h+1}^{\star,\sigma}, \quad (9)$$

which can be solved efficiently via a robust variant of value iteration when the RMDP is known (Iyengar, 2005; Nilim and El Ghaoui, 2005).

2.3 Distributionally robust offline RL

Let $\mathcal{D}$ be a history/batch dataset, which consists of a collection of K *independent* episodes generated based on executing a behavior policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{h=1}^H$ in some nominal MDP $\mathcal{M}^0 = \big(\mathcal{S}, \mathcal{A}, H, P^0 := \{P_h^0\}_{h=1}^H, \{r_h\}_{h=1}^H\big)$. More specifically, for $1 \le k \le K$, the k-th episode $\big(s_1^k, a_1^k, \ldots, s_H^k, a_H^k, s_{H+1}^k\big)$ is generated according to

$$s_1^k \sim \rho^{\mathsf{b}}, \qquad a_h^k \sim \pi_h^{\mathsf{b}}(\cdot \,|\, s_h^k) \qquad \text{and} \qquad s_{h+1}^k \sim P_h^0(\cdot \,|\, s_h^k, a_h^k), \qquad 1 \le h \le H. \tag{10}$$

Throughout the paper, ρ^{b} represents for some initial distribution associated with the history dataset. Then, we introduce the following short-hand notation for the occupancy distribution w.r.t. π^{b}:

$$\forall (h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}: \qquad d_h^{\mathsf{b}}(s) \coloneqq d_h^{\pi^{\mathsf{b}}, P^0}(s), \qquad d_h^{\mathsf{b}}(s, a) \coloneqq d_h^{\pi^{\mathsf{b}}, P^0}(s, a). \tag{11}$$

Goal. With the history dataset $\mathcal{D}$ in hand, our goal is to find a near-optimal robust policy $\widehat{\pi}$, which satisfies

$$V_1^{\widehat{\pi}, \sigma}(\rho) \ge V_1^{\star, \sigma}(\rho) - \varepsilon \tag{12}$$

using as few samples as possible, where ε is the target accuracy level, and

$$V_1^{\pi, \sigma}(\rho) \coloneqq \mathbb{E}_{s_1 \sim \rho}\big[V_1^{\pi, \sigma}(s_1)\big] \qquad \text{and} \qquad V_1^{\star, \sigma}(\rho) \coloneqq \mathbb{E}_{s_1 \sim \rho}\big[V_1^{\star, \sigma}(s_1)\big] \tag{13}$$

are evaluated when the initial state s_1 is drawn from a given distribution ρ.

Robust single-policy clipped concentrability. To quantify the quality of the history dataset to achieve the set goal, it is desirable to capture the distribution mismatch between the history dataset and the desired ones, inspired by the *single-policy clipped concentrability* assumption recently proposed by Li et al. (2022), we introduce a tailored assumption for robust MDPs as follows.

Assumption 1 (Robust single-policy clipped concentrability) *The behavior policy of the history dataset $\mathcal{D}$ satisfies*

$$\max_{(s,a,h,P) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{U}^\sigma(P^0)} \frac{\min\big\{d_h^{\star, P}(s, a), \frac{1}{S}\big\}}{d_h^{\mathsf{b}}(s, a)} \le C_{\mathsf{rob}}^\star \tag{14}$$

for some quantity $C_{\mathsf{rob}}^\star \in \big[\frac{1}{S}, \infty\big]$. Here, we take $C_{\mathsf{rob}}^\star$ to be the smallest quantity satisfying (14)*, and refer to it as the robust single-policy clipped concentrability coefficient. In addition, we follow the convention $0/0 = 0$.*

In words, $C_{\mathsf{rob}}^\star$ measures the worst-case discrepancy—between the optimal robust policy $\pi^\star$ (from initial state distribution ρ) in any model $P \in \mathcal{U}^\sigma(P^0)$ within the uncertainty set and the behavior policy π^{b} (from initial state distribution ρ^{b}) in the nominal model P^0—in terms of the clipped maximum density ratio of the state-action occupancy distributions.

- *Distribution shift.* When the uncertainty level $\sigma = 0$, Assumption 1 reduces back to the single-policy clipped concentrability in Li et al. (2022) for standard offline RL, a weaker notion that can be S times smaller than the single-policy concentrability adopted in, e.g., Rashidinejad et al. (2021); Xie et al. (2021); Shi et al. (2022). On the other end, whenever $\sigma > 0$, the proposed robust single-policy clipped concentrability accounts for the distribution shift not only due to the policies in use ($\pi^\star$ versus π^{b}) with respect to the respective initial state distributions, but also the underlying environments ($P \in \mathcal{U}^\sigma(P^0)$ versus P^0), and therefore, is generally larger than that in the non-robust counterpart.

- *Partial coverage.* As long as $C^\star_{\mathsf{rob}}$ is finite, i.e., $C^\star_{\mathsf{rob}} < \infty$, it admits the scenarios when the history dataset only provides *partial coverage* over the entire state-action space, as long as the behavior policy π^{b} visits the state-action pairs that are visited by the optimal robust policy $\pi^\star$ under at least one model in the uncertainty set.

Remark 1 *To facilitate comparison with prior works assuming full coverage, we can bound $C^\star_{\mathsf{rob}}$ when the batch dataset is generated using a simulator (Yang et al., 2022; Panaganti and Kalathil, 2022); namely, we can generate sample state transitions based on the transition kernel of the nominal MDP for all state-action pairs at all time steps. In this case, it amounts to that $d_h^{\mathsf{b}}(s,a) = \frac{1}{SA}$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, which directly leads to the bound $C^\star_{\mathsf{rob}} = \max_{(s,a,h,P) \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b}}(s,a)} \le \frac{1/S}{1/(SA)} = A$.*

3. Algorithm and theory: episodic finite-horizon RMDPs

In this section, we present a model-based algorithm—namely DRVI-LCB—for robust offline RL in the finite-horizon setting, along with its performance guarantees.

3.1 Building an empirical nominal MDP

For a moment, imagine we have access to N *independent* sample transitions $\mathcal{D}_0 := \{(h_i, s_i, a_i, s_i')\}_{i=1}^N$ drawn from the transition kernel P^0 of the nominal MDP $\mathcal{M}^0$, where each sample (h_i, s_i, a_i, s_i') indicates the transition from state s_i to state s_i' when action a_i is taken at step h_i, drawn according to $s_i' \sim P_{h_i}^0(\cdot \mid s_i, a_i)$. It is then natural to build an empirical estimate $\widehat{P}^0 = \{\widehat{P}_h^0\}_{h=1}^H$ of P^0 based on the empirical frequencies of state transitions, where

$$\widehat{P}_h^0(s' \mid s,a) := \begin{cases} \frac{1}{N_h(s,a)} \sum_{i=1}^N \mathbb{1}\big\{(h_i,s_i,a_i,s_i') = (h,s,a,s')\big\}, & \text{if } N_h(s,a) > 0 \\ 0, & \text{else} \end{cases} \tag{15}$$

for any $(h,s,a,s') \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Here, $N_h(s,a)$ denotes the total number of sample transitions from (s,a) at step h as

$$N_h(s,a) := \sum_{i=1}^N \mathbb{1}\big\{(h_i,s_i,a_i) = (h,s,a)\big\}. \tag{16}$$

Algorithm 1: Two-fold subsampling trick for the finite-horizon setting.

input: a dataset $\mathcal{D}$, probability δ.

data splitting: split $\mathcal{D}$ into two $\mathcal{D}^{\mathsf{main}}$ and $\mathcal{D}^{\mathsf{aux}}$, where each contain $K/2$ trajectories.

lower bounding the number of transitions in $\mathcal{D}^{\mathsf{main}}$: denote the number of transitions from state s at step h in $\mathcal{D}^{\mathsf{main}}$ (resp. $\mathcal{D}^{\mathsf{aux}}$) as $N_h^{\mathsf{main}}(s)$ (resp. $N_h^{\mathsf{aux}}(s)$), construct

$$N_h^{\mathsf{trim}}(s) \coloneqq \max\left\{N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s)\log\frac{HS}{\delta}},\, 0\right\}; \tag{17}$$

generate the subsampled dataset $\mathcal{D}^{\mathsf{trim}}$: randomly sample the transitions (i.e., the quadruples taking the form (s,a,h,s')) from $\mathcal{D}^{\mathsf{main}}$ uniformly at random, such that for each $(s,h)\in\mathcal{S}\times[H]$, $\mathcal{D}^{\mathsf{trim}}$ contains $\min\{N_h^{\mathsf{trim}}(s), N_h^{\mathsf{main}}(s)\}$ sample transitions.

output: set $\mathcal{D}_0 = \mathcal{D}^{\mathsf{trim}}$.

While it is possible to directly break down the history dataset $\mathcal{D}$ into sample transitions, unfortunately, the sample transitions from the same episode are not independent, significantly hurdling the analysis. To alleviate this, Li et al. (2022, Algorithm 2) proposed a simple two-fold subsampling scheme to preprocess the history dataset $\mathcal{D}$ and decouple the statistical dependency, resulting into a distributionally equivalent dataset $\mathcal{D}_0$ with independent samples; for completeness, we provide the procedure in Algorithm 1. We have the following lemma paraphrased from Li et al. (2022) for the obtained dataset $\mathcal{D}_0$.

Lemma 2 ((Li et al., 2022)) *With probability at least $1-8\delta$, the output dataset from the two-fold subsampling scheme in Li et al. (2022) is distributionally equivalent to $\mathcal{D}_0$, where $\{N_h(s,a)\}$ are independent of the sample transitions in $\mathcal{D}^0$ and obey*

$$N_h(s,a) \geq \frac{Kd_h^{\mathsf{b}}(s,a)}{8} - 5\sqrt{Kd_h^{\mathsf{b}}(s,a)\log\frac{KH}{\delta}}. \tag{18}$$

for all $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$.

Therefore, by invoking the two-fold sampling trick from Li et al. (2022), it is sufficient to treat the dataset $\mathcal{D}_0$ with independent samples onwards with Lemma 2 in place, which greatly simplifies the analysis.

3.2 DRVI-LCB: a pessimistic variant of robust value iteration

Armed with the estimate $\widehat{P}^0$ of the nominal transition kernel P^0, we are positioned to introduce our algorithm DRVI-LCB, summarized in Algorithm 2.

Distributionally robust value iteration. Before proceeding, let us recall the update rule of the classical distributionally robust value iteration (DRVI), which serves as the basis

Algorithm 2: Robust value iteration with LCB (DRVI-LCB) for robust offline RL.

input: a dataset $\mathcal{D}_0$; reward function r; uncertainty level σ.
initialization: $\widehat{Q}_{H+1} = 0$, $\widehat{V}_{H+1} = 0$.
for $h = H, \cdots, 1$ **do**
  Compute the empirical nominal transition kernel $\widehat{P}_h^0$ according to (15);
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    Compute the penalty term $b_h(s,a)$ according to (22);
    Set $\widehat{Q}_h(s,a)$ according to (21);
  for $s \in \mathcal{S}$ **do**
    Set $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a)$ and $\widehat{\pi}_h(s) = \arg\max_a \widehat{Q}_h(s,a)$;
output: $\widehat{\pi} = \{\widehat{\pi}_h\}_{1 \le h \le H}$.

of our algorithmic development. Given an estimate of the nominal MDP $\widehat{P}^0$ and the radius σ of the uncertainty set, DRVI updates the robust value functions according to

$$\widehat{Q}_h(s,a) = r_h(s,a) + \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1}, \qquad \text{and} \qquad \widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a), \tag{19}$$

which works backwards from $h = H$ to $h = 1$, with the terminal condition $\widehat{Q}_{H+1} = 0$. Due to strong duality (Hu and Hong, 2013), the update rule of the robust Q-functions in (19) can be equivalently reformulated in its dual form as

$$\widehat{Q}_h(s,a) = r_h(s,a) + \sup_{\lambda \ge 0} \left\{ -\lambda \log\left(\widehat{P}^0_{h,s,a} \exp\left(\frac{-\widehat{V}_{h+1}}{\lambda} \right) \right) - \lambda\sigma \right\}, \tag{20}$$

which can be solved efficiently (Iyengar, 2005; Yang et al., 2022; Panaganti and Kalathil, 2022).

Our algorithm DRVI-LCB**.** Motivated by the principle of pessimism in standard offline RL (Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021; Li et al., 2022), we propose to perform a pessimistic variant of DRVI, where the update rule of DRVI-LCB at step h is modified as

$$\widehat{Q}_h(s,a) = \max\left\{ r_h(s,a) + \sup_{\lambda \ge 0} \left\{ -\lambda \log\left(\widehat{P}^0_{h,s,a} \cdot \exp\left(\frac{-\widehat{V}_{h+1}}{\lambda} \right) \right) - \lambda\sigma \right\} - b_h(s,a),\, 0 \right\}. \tag{21}$$

Here, the robust Q-function estimate is adjusted by subtracting a carefully designed data-driven penalty term $b_h(s,a)$ that measures the uncertainty of the value estimates. Specifically, for some $\delta \in (0,1)$ and any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the penalty term $b_h(s,a)$ is defined as

$$b_h(s,a) = \begin{cases} \min\left\{ c_{\mathrm{b}} \frac{H}{\sigma} \sqrt{\frac{\log(\frac{KHS}{\delta})}{\widehat{P}_{\mathsf{min},h}(s,a) N_h(s,a)}},\ H \right\} & \text{if } N_h(s,a) > 0, \\ H & \text{otherwise,} \end{cases} \tag{22}$$

where c_{b} is some universal constant, and

$$\widehat{P}_{\mathsf{min},h}(s,a) \coloneqq \min_{s'} \Big\{ \widehat{P}_h^0(s' \,|\, s,a) : \ \widehat{P}_h^0(s' \,|\, s,a) > 0 \Big\}. \tag{23}$$

The penalty term is novel and different from the one used in standard (no-robust) offline RL (Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021; Li et al., 2022; Shi et al., 2022), by taking into consideration the unique problem structure pertaining to robust MDPs. In particular, it tightly upper bounds the statistical uncertainty which carries a non-linear and implicit dependency w.r.t. the estimated nominal transition kernel induced by the uncertainty set $\mathcal{U}(P^0)$, addressing unique challenges not present for the standard MDP case.

3.3 Performance guarantees

Before stating the main theorems, let us first introduce several important metrics.

- $P^\star_{\mathsf{min}}$, which only depends on the state-action pairs covered by the optimal robust policy $\pi^\star$ under the nominal model P^0:

$$P^\star_{\mathsf{min}} \coloneqq \min_{h,s,s'} \Big\{ P_h^0\left(s' | s, \pi_h^\star(s)\right) : \ P_h^0\left(s' | s, \pi_h^\star(s)\right) > 0 \Big\}. \tag{24}$$

  In words, $P^\star_{\mathsf{min}}$ is the smallest positive state transition probability of the optimal robust policy $\pi^\star$ under the nominal kernel P^0.

- Similarly, we introduce $P^{\mathsf{b}}_{\mathsf{min}}$ which only depends on the state-action pairs covered by the behavior policy π^{b} under the nominal model P^0:

$$P^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{h,s,a,s'} \Big\{ P_h^0\left(s' | s,a\right) : \ d_h^{\mathsf{b}}(s,a) > 0,\ P_h^0\left(s' \,|\, s,a\right) > 0 \Big\}. \tag{25}$$

  In words, $P^{\mathsf{b}}_{\mathsf{min}}$ is the smallest positive state transition probability of the behavior policy π^{b} under the nominal kernel P^0.

- Finally, let $d^{\mathsf{b}}_{\mathsf{min}}$ denote the smallest positive state-action occupancy distribution of the behavior policy π^{b} under the nominal model P^0:

$$d^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{h,s,a} \Big\{ d_h^{\mathsf{b}}(s,a) : \ d_h^{\mathsf{b}}(s,a) > 0 \Big\}. \tag{26}$$

We are now positioned to present the performance guarantees of DRVI-LCB for robust offline RL.

Theorem 3 *Given an uncertainty level $\sigma > 0$, suppose that the penalty terms in Algorithm 2 are chosen as (22) for sufficiently large c_{b}. With probability at least $1-\delta$, the output $\widehat{\pi}$ of Algorithm 2 obeys*

$$V_1^{\star,\sigma}(\rho) - V_1^{\widehat{\pi},\sigma}(\rho) \le c_0 \frac{H^2}{\sigma} \sqrt{\frac{SC^\star_{\mathsf{rob}} \log^2(KHS/\delta)}{P^\star_{\mathsf{min}} K}}, \tag{27}$$

as long as the number of episodes K satisfies

$$K \geq \frac{c_1 \log(KHS/\delta)}{d_{\mathsf{min}}^{\mathsf{b}} P_{\mathsf{min}}^{\mathsf{b}}}, \tag{28}$$

where c_0 and c_1 are some sufficiently large universal constants.

Our theorem is the first to characterize the sample complexities of robust offline RL under *partial coverage*, to the best of our knowledge (cf. Table 2). Theorem 3 shows that DRVI-LCB finds an ε-optimal robust policy as soon as the sample size $T = KH$ is above the order of

$$\underbrace{\frac{SC_{\mathsf{rob}}^{\star} H^5}{P_{\mathsf{min}}^{\star} \sigma^2 \varepsilon^2}}_{\epsilon\text{-dependent}} + \underbrace{\frac{H}{d_{\mathsf{min}}^{\mathsf{b}} P_{\mathsf{min}}^{\mathsf{b}}}}_{\text{burn-in cost}}, \tag{29}$$

up to some logarithmic factor, where the burn-in cost is independent of the accuracy level ε. For sufficiently small accuracy level ε, this results in a sample complexity of

$$\widetilde{O}\left(\frac{SC_{\mathsf{rob}}^{\star} H^5}{P_{\mathsf{min}}^{\star} \sigma^2 \varepsilon^2}\right). \tag{30}$$

Our theorem suggests that the sample efficiency of robust offline RL critically depends on the problem structure of the given RMDP (i.e. coverage of the optimal robust policy $\pi^\star$ as measured by $P_{\mathsf{min}}^{\star}$) as well as the quality of the history dataset (as measured by $C_{\mathsf{rob}}^{\star}$). Given that $C_{\mathsf{rob}}^{\star}$ can be as small as on the order of $1/S$, the sample complexity requirement can exhibit a much weaker dependency with the size of the state space S.

On the flip side, to assess the optimality of Theorem 3, we develop an information-theoretic lower bound for robust offline RL as provided in the following theorem.

Theorem 4 *For any $(H, S, C, P_{\mathsf{min}}^{\star}, \sigma, \varepsilon)$ obeying $H \geq 2e^8$, $C \geq 4/S$, $P_{\mathsf{min}}^{\star} \in (0, \frac{1}{H}]$, and $\varepsilon \leq \frac{H}{384e^6 \log(1/P_{\mathsf{min}}^{\star})}$, we can construct a collection of finite-horizon RMDPs $\{\mathcal{M}_\theta \,|\, \theta \in \Theta\}$, an initial state distribution ρ, and a batch dataset with K independent sample trajectories each with length H satisfying $2C \leq C_{\mathsf{rob}}^{\star} \leq 4C$, such that*

$$\inf_{\widehat{\pi}} \max_{\theta \in \Theta} \mathbb{P}_\theta \left\{ V_1^{\star,\sigma}(\rho) - V_1^{\widehat{\pi},\sigma}(\rho) > \varepsilon \right\} \geq \frac{1}{8},$$

provided that

$$T = KH \leq \begin{cases} \frac{c_1 SC_{\mathsf{rob}}^{\star} H^4}{\varepsilon^2} & \text{if} \quad 0 < \sigma \leq \frac{1}{20H}, \\ \frac{c_1 SC_{\mathsf{rob}}^{\star} H^3}{P_{\mathsf{min}}^{\star} \sigma^2 \varepsilon^2} & \text{if} \quad \log(1/P_{\mathsf{min}}^{\star}) - 6 \leq \sigma \leq \log(1/P_{\mathsf{min}}^{\star}) - 5 \end{cases}. \tag{31}$$

Here, $c_1 > 0$ is some universal constant, the infimum is taken over all estimators $\widehat{\pi}$, and $\mathbb{P}_\theta$ denotes the probability when the RMDP is $\mathcal{M}_\theta$.

The messages of Theorem 4 are two-fold.

- When the uncertainty level $\sigma \lesssim 1/H$ is relatively small, Theorem 4 shows that no algorithm can succeed in finding an ε-optimal robust policy when the sample complexity falls below the order of
$$\Omega\left(\frac{SC^{\star}_{\mathsf{rob}}H^4}{\varepsilon^2}\right),$$
which is at least as large as the sample complexity requirement of non-robust offline RL (Li et al., 2022). Consequently, this leads to new insights regarding the statistical hardness of learning robust RMDPs with the KL uncertainty set: it can be at least as hard as the standard MDPs (which corresponds to $\sigma = 0$), for sufficiently small uncertainty levels.

- When the uncertainty level $\sigma \asymp \log(1/P^{\star}_{\mathsf{min}})$, Theorem 4 shows that no algorithm can succeed in finding an ε-optimal robust policy when the sample complexity falls below the order of
$$\Omega\left(\frac{SC^{\star}_{\mathsf{rob}}H^3}{P^{\star}_{\mathsf{min}}\sigma^2\varepsilon^2}\right),$$
which confirms the near-optimality of DRVI-LCB up to a factor of H^2 ignoring logarithmic factors. Therefore, DRVI-LCB is the first provable algorithm for robust offline RL with a near-optimal sample complexity without requiring the stringent full coverage assumption.

4. Robust offline RL for discounted infinite-horizon RMDPs

In this section, we turn to the studies of robust offline RL for discounted infinite-horizon MDPs.

4.1 Backgrounds on discounted infinite-horizon RMDPs

Similar to the finite-horizon setting, we consider the discounted infinite-horizon robust MDPs (RMDPs) represented by $\mathcal{M}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^{\sigma}(P^0), r\}$. Here, $\mathcal{S} = \{1, 2, \cdots, S\}$ is the state space, $\mathcal{A} = \{1, 2, \cdots, A\}$ is the action space, $\gamma \in [0, 1)$ is the discounted factor, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the intermediate reward function. Different from the standard MDPs, $\mathcal{U}^{\sigma}(P^0)$ denote the set of possible transition kernels within an uncertainty set centered around a nominal kernel $P^0 : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ using the distance measured in terms of the KL divergence. In particular, given an uncertainty level $\sigma > 0$, the uncertainty set around P^0 is specified as

$$\mathcal{U}^{\sigma}(P^0) \coloneqq \otimes\, \mathcal{U}^{\sigma}(P^0_{s,a}), \qquad \mathcal{U}^{\sigma}(P^0_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{s,a} \parallel P^0_{s,a}\right) \le \sigma\right\}, \tag{32}$$

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

$$P_{s,a} \coloneqq P(\cdot \mid s, a) \in \mathbb{R}^{1\times S}, \qquad P^0_{s,a} \coloneqq P^0(\cdot \mid s, a) \in \mathbb{R}^{1\times S}. \tag{33}$$

Note that at any time step, the adversary of the nature chooses a history-independent component within the fixed uncertainty set $\mathcal{U}^{\sigma}(P^0_{s,a})$ defined in (30), conditioned only on the current state-action pair (s, a). This is to ensure the computation tractability of finding such adversary.

Policy and robust value/Q functions. A (possibly random) stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ represents the selection rule of the agent, namely, $\pi(a \,|\, s)$ denote the probability of choosing a in state s. With some abuse of notation, let $\pi(s)$ represent the action chosen by π when π is a deterministic policy. We define the *robust value function* $V^{\pi,\sigma}$ and *robust Q-function* $Q^{\pi,\sigma}$ respectively as

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

where the value function $V^{\pi,P}$ and Q-function $Q^{\pi,P}$ w.r.t. policy π and transition kernel P are defined respectively by

$$\forall s \in \mathcal{S}: \qquad V^{\pi,P}(s) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\Big|\, s_0 = s\right], \tag{34}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q^{\pi,P}(s,a) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\Big|\, s_0 = s, a_0 = a\right], \tag{35}$$

where the expectation is taken over the randomness of the trajectory. In words, the robust value/Q functions characterize the worst case over all the instances in the uncertainty set.

Optimal policy and robust Bellman equation. Similar to the finite-horizon RMDPs, it is well-known that there exists at least one deterministic policy that maximizes the robust value function and Q-function simultaneously in the infinite-horizon setting as well (Iyengar, 2005; Nilim and El Ghaoui, 2005). With this in mind, we denote the optimal policy as $\pi^\star$ and the corresponding *optimal robust value function* (resp. *optimal robust Q-function*) as $V^{\star,\sigma}$ (resp. $Q^{\star,\sigma}$), namely

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) \coloneqq V^{\pi^\star,\sigma}(s) = \max_{\pi} V^{\pi,\sigma}(s), \tag{36a}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^{\star,\sigma}(s,a) \coloneqq Q^{\pi^\star,\sigma}(s,a) = \max_{\pi} Q^{\pi,\sigma}(s,a). \tag{36b}$$

In addition, we continue to admit the Bellman's optimality principle, resulting in the following *robust Bellman consistency equation* (resp. *robust Bellman optimality equation*):

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^{\pi,\sigma}(s,a) = r(s,a) + \gamma \inf_{P \in \mathcal{U}^\sigma(P^0_{s,a})} PV^{\pi,\sigma}, \tag{37a}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{P \in \mathcal{U}^\sigma(P^0_{s,a})} PV^{\star,\sigma}. \tag{37b}$$

Occupancy distributions. To begin, let ρ be some initial state distribution. We denote $d^{\pi,P}(s \,|\, \rho)$ and $d^{\pi,P}(s,a \,|\, \rho)$ respectively as the state occupancy distribution and the state-action occupancy distribution induced by policy π, namely

$$\forall s \in \mathcal{S}: \qquad d^{\pi,P}(s) \coloneqq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \,|\, s_0 \sim \rho, \pi, P), \tag{38a}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad d^{\pi,P}(s,a) \coloneqq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \,|\, s_0 \sim \rho, \pi, P)\, \pi(a \,|\, s). \tag{38b}$$

Here, the occupancy distributions are conditioned on $s_0 \sim \rho$ and the sequence of actions and states are generated based on policy π and transition kernel P. Next, applying (38) with $\pi = \pi^\star$, we adopt the the following short-hand notation for the occupancy distributions associated with the optimal policy:

$$\forall s \in \mathcal{S}: \qquad d^{\star,P}(s) \coloneqq d^{\pi^\star,P}(s), \tag{39a}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad d^{\star,P}(s,a) \coloneqq d^{\pi^\star,P}(s,a) = d^{\star,P}(s)\, \mathbb{1}\{a = \pi^\star(s)\}. \tag{39b}$$

4.2 Data collection and constructing the empirical MDP

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

$$(s_i, a_i) \overset{\text{i.i.d.}}{\sim} d^{\mathsf{b}} \quad \text{and} \quad s_i' \overset{\text{i.i.d.}}{\sim} P^0(\cdot \,|\, s_i, a_i), \qquad 1 \le i \le N. \tag{40}$$

Armed with these, we are ready to introduce the goal in the infinite-horizon setting. Given the history dataset $\mathcal{D}$ obeying Assumption 2, for some target accuracy $\varepsilon > 0$, we aim to find a near-optimal robust policy $\widehat{\pi}$, which satisfies

$$V^{\widehat{\pi},\sigma}(\rho) \ge V^{\star,\sigma}(\rho) - \varepsilon \tag{41}$$

in a sample-efficient manner for some initial state distribution ρ.

Remark 5 *For simplicity, we limit ourselves to the case when the history dataset consists of independent sample transitions. It is not difficult to generalize to the Markovian data case, when we only have access to a single trajectory of data generated by following some behavior policy, by combining the two-fold subsampling trick in Li et al. (2022, Appendix D) with our analysis. We leave this extension to interested readers.*

Similar to Assumption 1, we design the following *robust single-policy clipped concentrability* assumption tailored for infinite-horizon RMDPs to characterize the quality of the history dataset.

Assumption 2 (Robust single-policy clipped concentrability for infinite-horizon MDPs) *The behavior policy of the history dataset $\mathcal{D}$ satisfies*

$$\max_{(s,a,P) \in \mathcal{S} \times \mathcal{A} \times \mathcal{U}^\sigma(P^0)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)} \le C^\star_{\mathsf{rob}} \tag{42}$$

for some finite quantity $C^\star_{\mathsf{rob}} \in \left[\frac{1}{S}, \infty\right)$. Following the convention $0/0 = 0$, we denote $C^\star_{\mathsf{rob}}$ to be the smallest quantity satisfying (42), and refer to it as the robust single-policy clipped concentrability coefficient.

Remark 6 *Similar to Remark 1, we can bound $C^\star_{\mathsf{rob}} \le A$ when the batch dataset is generated using a simulator (Yang et al., 2022; Panaganti and Kalathil, 2022). By combining this bound of $C^\star_{\mathsf{rob}}$ with the theoretical guarantees developed momentarily in Theorem 9, we obtain the comparison in Table 2.*

Algorithm 3: Robust value iteration with LCB (DRVI-LCB) for infinite-horizon RMDPs.

input: a dataset $\mathcal{D}$; reward function r; uncertainty level σ; number of iterations M.
initialization: $\widehat{Q}_0(s,a)=0$, $\widehat{V}_0(s)=0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.
Compute the empirical nominal transition kernel $\widehat{P}^0$ according to (44);
Compute the penalty term $b(s,a)$ according to (48);
for $m=1,2,\cdots,M$ **do**
  for $s\in\mathcal{S}, a\in\mathcal{A}$ **do**
    Set $\widehat{Q}_m(s,a)$ according to (51);
  for $s\in\mathcal{S}$ **do**
    Set $\widehat{V}_m(s)=\max_a \widehat{Q}_m(s,a)$;
output: $\widehat{\pi}$ s.t. $\widehat{\pi}(s)=\arg\max_a \widehat{Q}_M(s,a)$ for all $s\in\mathcal{S}$.

Building an empirical nominal MDP Recalling that we have N independent samples in the dataset $\mathcal{D}=\{(s_i,a_i,s_i')\}_{1\le i\le N}$. First, we denote $N(s,a)$ as the total number of sample transitions from any state-action pair (s,a) as

$$N(s,a)\coloneqq\sum_{i=1}^{N}\mathbb{1}\big\{(s_i,a_i)=(s,a)\big\}. \tag{43}$$

Armed with $N(s,a)$, we construct the empirical estimate $\widehat{P}^0$ of the nominal kernel P^0 by the visiting frequencies of state-action pairs as follows:

$$\widehat{P}^0(s'\,|\,s,a)\coloneqq\begin{cases}\frac{1}{N(s,a)}\sum_{i=1}^{N}\mathbb{1}\big\{(s_i,a_i,s_i')=(s,a,s')\big\}, & \text{if } N(s,a)>0\\ 0, & \text{else}\end{cases} \tag{44}$$

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

4.3 DRVI-LCB for discounted infinite-horizon RMDPs

With the estimate $\widehat{P}^0$ of the nominal transition kernel P^0 in hand, we are positioned to introduce our algorithm DRVI-LCB for infinite-horizon RMDPs, which bears some similarity with the finite-horizon version (cf. Algorithm 2), by taking the uncertainties of the value estimates into consideration throughout the value iterations. The procedure is summarized in Algorithm 3.

The pessimistic robust Bellman operator. At the core of DRVI-LCB is a pessimistic variant of the classical robust Bellman operator in the infinite-horizon setting (Zhou et al., 2021; Iyengar, 2005; Nilim and El Ghaoui, 2005), denoted as $\mathcal{T}^\sigma(\cdot):\mathbb{R}^{SA}\to\mathbb{R}^{SA}$, which we recall as follows:

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}:\quad \mathcal{T}^\sigma(Q)(s,a)\coloneqq r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}V,\quad\text{with}\quad V(s)\coloneqq\max_a Q(s,a). \tag{45}$$

Encouragingly, the robust Bellman operator shares the nice γ-contraction property of the standard Bellman operator, ensuring fast convergence of robust value iteration by applying the robust Bellman operator (45) recursively. In the robust offline setting, instead of recursing using the population robust Bellman operator, we need to construct a pessimistic variant of the robust Bellman operator $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ w.r.t. the empirical nominal kernel $\widehat{P}^0$ as follows:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q)(s,a) = \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V - b\big(s,a\big),\, 0 \right\}, \tag{46}$$

where $b(s,a)$ denotes the penalty term that measures the data-dependent uncertainty of the value estimates.

To specify the tailored penalty term $b(s,a)$ in (46), we first introduce an additional term

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{P}_{\mathsf{min}}(s,a) \coloneqq \min_{s'} \Big\{ \widehat{P}^0(s' \,|\, s,a) :\ \widehat{P}^0(s' \,|\, s,a) > 0 \Big\}, \tag{47}$$

which in words represents the smallest positive transition probability of the estimated nominal kernel $\widehat{P}^0(s' \,|\, s,a)$. Then for some $\delta \in (0,1)$, some universal constant $c_{\mathrm{b}} > 0$, $b(s,a)$ is defined as

$$b(s,a) = \begin{cases} \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3 S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a) N(s,a)}} + \frac{4}{\sigma N(1-\gamma)}\ ,\ \frac{1}{1-\gamma} \right\} + \frac{2}{\sigma N} & \text{if } N(s,a) > 0, \\ \frac{1}{1-\gamma} + \frac{2}{\sigma N} & \text{otherwise.} \end{cases} \tag{48}$$

As shall be illuminated, our proposed pessimistic robust Bellman operator $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ (cf. (46)) plays an important role in DRVI-LCB. Encouragingly, despite the additional data-driven penalty term $b(s,a)$, it still enjoys the celebrated γ-contractive property, which greatly facilitates the analysis. Before continuing, we summarize the γ-contraction property below, whose proof is postponed to Appendix C.1.

Lemma 7 (γ-Contraction) *For any $\gamma \in [0,1)$, the operator $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ (cf. (46)) 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*

$$\left\| \widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q_1) - \widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q_2) \right\|_\infty \le \gamma \left\| Q_1 - Q_2 \right\|_\infty. \tag{49}$$

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

Our algorithm DRVI-LCB for infinite-horizon robust offline RL. Armed with the γ-contraction property of the pessimistic robust Bellman operator $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$, we are positioned to introduce DRVI-LCB for infinite-horizon RMDPs, summarized in Algorithm 3. Specifically, DRVI-LCB can be seen as a value iteration algorithm w.r.t. $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ (cf. (46)), whose update rule at the m-th iteration can be formulated as

$$\widehat{Q}_m(s,a) = \widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\widehat{Q}_{m-1})(s,a) = \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}_{m-1} - b\big(s,a\big),\, 0 \right\}, \tag{50}$$

and $\widehat{V}_m(s) = \max_a \widehat{Q}_m(s,a)$ for all $m = 1, 2, \cdots, M$. In view of strong duality (Hu and Hong, 2013), the above convex problem can be translated into a dual formulation, leading to the following equivalent update rule:

$$\widehat{Q}_m(s,a) = \max\left\{ r(s,a) + \sup_{\lambda \geq 0}\left\{ -\lambda \log\left(\widehat{P}^0_{s,a} \cdot \exp\left(\frac{-\widehat{V}_{m-1}}{\lambda} \right) \right) - \lambda\sigma \right\} - b\big(s,a\big),\, 0 \right\}, \tag{51}$$

which can be solved efficiently (Iyengar, 2005; Yang et al., 2022; Panaganti and Kalathil, 2022) as a one-dimensional optimization problem.

To finish the description, we initialize the estimates of Q-function ($\widehat{Q}_0$) and value function ($\widehat{V}_0$) to be zero and output the greedy policy of the final Q-estimates ($\widehat{Q}_M$) as the final policy $\widehat{\pi}$, namely,

$$\widehat{\pi}(s) = \arg\max_a \widehat{Q}_M(s,a) \quad \text{for all } s \in \mathcal{S}. \tag{52}$$

It turns out that the iterates $\left\{\widehat{Q}_m\right\}_{m\geq 0}$ of DRVI-LCB converge linearly to the fixed point $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$ owing to the nice γ-contraction property outlined in Lemma 7. This fact is summarized in the following lemma, whose proof is postponed to Appendix C.2.

Lemma 8 *Let $\widehat{Q}_0 = 0$. The iterates of Algorithm 3 obey*

$$\forall m \geq 0: \quad \widehat{Q}_m \leq \widehat{Q}^{\star,\sigma}_{\mathsf{pe}} \quad \textit{and} \quad \left\|\widehat{Q}_m - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty \leq \frac{\gamma^m}{1-\gamma}. \tag{53}$$

4.4 Performance guarantees

Before introducing the main theorems, we first define several essential metrics.

- $d^{\mathsf{b}}_{\mathsf{min}}$: the smallest positive entry of the distribution d^{b}, i.e.,

$$d^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{s,a}\left\{ d^{\mathsf{b}}(s,a) :\ d^{\mathsf{b}}(s,a) > 0 \right\}. \tag{54}$$

- $P^{\mathsf{b}}_{\mathsf{min}}$: the smallest positive state transition probability under the nominal kernel P^0 in the region covered by dataset $\mathcal{D}$, i.e.,

$$P^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{s,a,s'}\left\{ P^0\left(s' \,|\, s,a\right) :\ d^{\mathsf{b}}(s,a) > 0,\, P^0\left(s' \,|\, s,a\right) > 0 \right\}. \tag{55}$$

  Note that $P^{\mathsf{b}}_{\mathsf{min}}$ is determined only by the state-action pairs covered by the batch dataset $\mathcal{D}$.

- $P^{\star}_{\mathsf{min}}$: the smallest positive state transition probability of the optimal robust policy $\pi^\star$ under the nominal kernel P^0, namely

$$P^{\star}_{\mathsf{min}} \coloneqq \min_{s,s'}\left\{ P^0\big(s' \,|\, s,\pi^\star(s)\big) :\ P^0\big(s' \,|\, s,\pi^\star(s)\big) > 0 \right\}. \tag{56}$$

  We also note that $P^{\star}_{\mathsf{min}}$ is determined only by the state-action pairs covered by the optimal robust policy $\pi^\star$ under the nominal model P^0.

We are now positioned to introduce the sample complexity upper bound of DRVI-LCB, together with the minimax lower bound, for solving infinite-horizon RMDPs. First, we present the performance guarantees of DRVI-LCB for robust offline RL in the infinite-horizon case, with the proof deferred to Appendix C.3.

Theorem 9 *Let c_0 and c_1 be some sufficiently large universal constants. Given an uncertainty level $\sigma > 0$, suppose that the penalty terms in Algorithm 3 are chosen as* (48) *for sufficiently large c_{b}. With probability at least $1-\delta$, the output $\widehat{\pi}$ of Algorithm 3 obeys*

$$V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho) \le \frac{c_0}{\sigma(1-\gamma)^2}\sqrt{\frac{SC^\star_{\mathsf{rob}}\log^2\left(\frac{(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^\star_{\mathsf{min}}N}}, \tag{57}$$

as long as the number of samples N satisfies

$$N \ge \frac{c_1\log(NS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}}P^{\mathsf{b}}_{\mathsf{min}}}. \tag{58}$$

The result directly indicates that DRVI-LCB can finds an ε-optimal policy as long as the sample size in dataset $\mathcal{D}$ exceeds the order of (ignoring logarithmic factors)

$$\underbrace{\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}}_{\varepsilon\text{-dependent}} + \underbrace{\frac{1}{d^{\mathsf{b}}_{\mathsf{min}}P^{\mathsf{b}}_{\mathsf{min}}}}_{\text{burn-in cost}}. \tag{59}$$

Note that the burn-in cost is independent with the accuracy level ε, which tells us that the sample complexity is no more than

$$\widetilde{O}\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}\right) \tag{60}$$

as long as ε is small enough. The sample complexity of DRVI-LCB still dramatically outperforms prior works under full coverage, which has been compared in detail in Section 1.2 (cf. Table 2). In particular, our sample complexity produces an exponential improvement over Zhou et al. (2021); Panaganti and Kalathil (2022) in terms of the dependency with the effective horizon $\frac{1}{1-\gamma}$, which is especially significant for long-horizon problems. Compared with Yang et al. (2022), our sample complexity is better by at least a factor of $S/P^\star_{\mathsf{min}}$. To achieve the claimed bound, we resort to a delicate technique called the leave-one-out analysis (Agarwal et al., 2020; Li et al., 2020, 2022), by carefully designing an auxiliary set of RMDPs to decouple the statistical dependency introduced across the iterates of pessimistic robust value iteration. This is the first time that the leave-one-out analysis is applied to understanding the sample efficiency of model-based robust RL algorithms, which is of potential independent interest to tighten the sample complexity of other robust RL problems.

To complement the upper bound, we develop an information-theoretic lower bound for robust offline RL as provided in the following theorem whose proof can be found in Appendix C.4.

Theorem 10 *For any* $(S, P^\star_{\mathsf{min}}, C^\star_{\mathsf{rob}}, \gamma, \sigma, \varepsilon)$ *obeying* $\frac{1}{1-\gamma} \geq 2e^8$, $P^\star_{\mathsf{min}} \in \big(0, 1-\gamma\big]$, $S \geq \log\big(1/P^\star_{\mathsf{min}}\big)$, $C^\star_{\mathsf{rob}} \geq 8/S$, *and* $\varepsilon \leq \frac{1}{384e^6(1-\gamma)\log\left(1/P^\star_{\mathsf{min}}\right)}$, *we can construct a collection of infinite-horizon RMDPs* $\{\mathcal{M}_\theta \,|\, \theta \in \Theta\}$, *an initial state distribution* ρ, *and a batch dataset with* N *independent samples, such that*

$$\inf_{\widehat{\pi}} \max_{\theta \in \Theta} P_\theta \left\{ V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho) > \varepsilon \right\} \geq \frac{1}{8},$$

provided that

$$N \leq \begin{cases} \frac{c_1 SC^\star_{\mathsf{rob}}}{(1-\gamma)^3\varepsilon^2} & \text{if} \quad 0 < \sigma \leq \frac{1-\gamma}{20}, \\ \frac{c_1 SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\varepsilon^2} & \text{if} \quad \log\big(1/P^\star_{\mathsf{min}}\big) - 6 \leq \sigma \leq \log\big(1/P^\star_{\mathsf{min}}\big) - 5 \end{cases} \tag{61}$$

Here, $c_1 > 0$ *is some universal constant, the infimum is taken over all estimators* $\widehat{\pi}$, *and* $\mathbb{P}_\theta$ *denotes the probability when the RMDP is* $\mathcal{M}_\theta$.

Similar to the finite-horizon setting, the messages of Theorem 10 are two-fold.

- When the uncertainty level $\sigma \lesssim 1-\gamma$ is relatively small, Theorem 10 shows that no algorithm can succeed in finding an ε-optimal robust policy when the sample complexity falls below the order of

  $$\Omega\left(\frac{SC^\star_{\mathsf{rob}}}{(1-\gamma)^3\varepsilon^2}\right),$$

  which is at least as large as the sample complexity requirement of non-robust offline RL (Li et al., 2022). Consequently, this again suggests that learning robust RMDPs with the KL uncertainty set can be at least as hard as the standard MDPs (which corresponds to $\sigma = 0$), for sufficiently small uncertainty levels.

- When the uncertainty level $\sigma \asymp \log(1/P^\star_{\mathsf{min}})$, Theorem 10 shows that no algorithm can succeed in finding an ε-optimal robust policy when the sample complexity falls below the order of

  $$\Omega\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\varepsilon^2}\right),$$

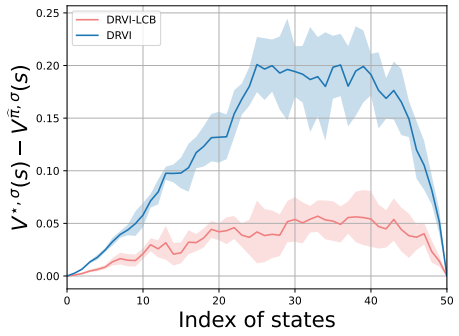
  which directly confirms that DRVI-LCB is near-optimal up to a polynomial factor of the effective horizon length $\frac{1}{1-\gamma}$ (cf. (59)). To the best of our knowledge, DRVI-LCB is the first provable algorithm with near-optimal sample complexity for infinite-horizon robust offline RL. Moreover, the requirement imposed on the history dataset is also much weaker than prior literature on robust offline RL (Yang et al., 2022; Zhou et al., 2021), without the need of full coverage of the state-action space.

5. Numerical experiments

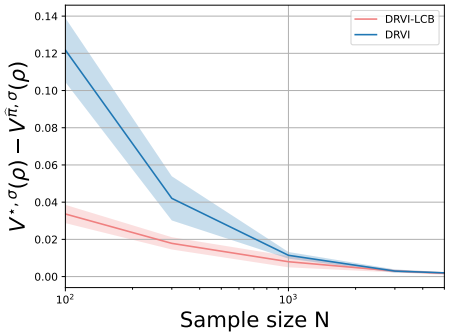
We conduct experiments on the gambler's problem (Sutton and Barto, 2018; Zhou et al., 2021) to evaluate the performance of the proposed algorithm DRVI-LCB, with comparisons to the robust value iteration algorithm DRVI without pessimism (Panaganti and Kalathil, 2022). Our code can be accessed at:

`https://github.com/Laixishi/Robust-RL-with-KL-divergence`.

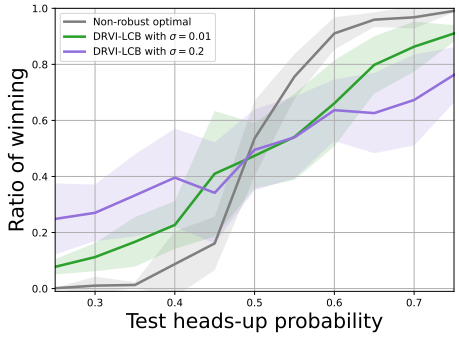
Gambler's problem. In the gambler's game (Sutton and Barto, 2018; Zhou et al., 2021), a gambler bets on a sequence of coin flips, winning the stake with heads and losing with tails. Starting from some initial balance, the game ends when the gambler's balance either reaches 50 or 0, or the total number of bets H is hit. This problem can be formulated as an episodic finite-horizon MDP, with a state space $\mathcal{S} = \{0, 1, \cdots, 50\}$ and the associated possible actions $a \in \{0, 1, \cdots, \min\{s, 50 - s\}\}$ at state s. Here, we set the horizon length $H = 100$. Moreover, the parameter of the transition kernel, which is the probability of heads for the coin flip, is fixed as p_{head} and remains the same in all time steps $h \in [H]$. The reward is set as 1 when the state reaches $s = 50$ and 0 for all other cases. In addition, suppose the initial state (i.e., the gambler's initial balance) distribution ρ is taken uniformly at random within $\mathcal{S}$.



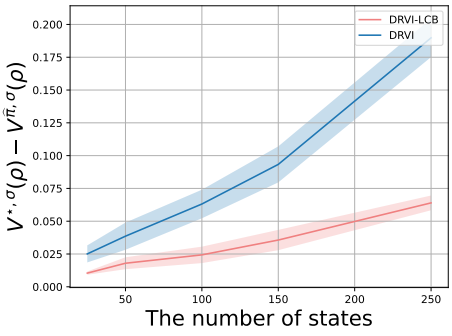
(a) value gap versus states



(b) value gap versus sample size



(c) winning rate versus p_{head}



(d) value gap versus the number of states

Figure 1: Performance of the proposed algorithm DRVI-LCB using independent samples per state-action pair and time step, where it shows better sample efficiency than the baseline algorithm DRVI without pessimism, as well as better robustness in the learned policy compare to its non-robust counterpart.

The benefit of pessimism. We first utilize a history dataset with N independent samples per state-action pair and time step, generated from the nominal MDP with $p^0_{\mathsf{head}} = 0.6$. We evaluate the performance of the learned policy $\widehat{\pi}$ using our proposed method DRVI-LCB with comparison to DRVI without pessimism, where we fix the uncertainty level $\sigma = 0.1$ for learning the robust optimal policy. The experiments are repeated 10 times with the average

and standard deviations reported. To begin with, Figure 1(a) plots the sub-optimality value gap $V_1^{\star,\sigma}(s) - V_1^{\widehat{\pi},\sigma}(s)$ for every $s \in \mathcal{S}$, when a sample size $N = 100$ is used to learn the robust policies. It is shown that DRVI-LCB outperform the baseline DRVI uniformly over the state space when the sample size is small, corroborating the benefit of pessimism in the sample-starved regime. Furthermore, Figure 1(b) shows the sub-optimality gap $V_1^{\star,\sigma}(\rho) - V_1^{\widehat{\pi},\sigma}(\rho)$ with varying sample sizes $N = 100, 300, 1000, 3000, 5000$, where the initial test distribution ρ is generated randomly.[2] While the performance of DRVI-LCB and DRVI both improves with the increase of the sample size, the proposed algorithm DRVI-LCB achieves much better performance with fewer samples.

The benefit of distributional robustness. To corroborate the benefit of distributional robustness, we evaluate the performance of the policy learned from $N = 1000$ samples using DRVI-LCB on perturbed environments with varying model parameters $p_{\mathsf{head}} \in [0.25, 0.75]$. We measure the practical performance based on the ratio of winning (i.e., reaching the state $s = 50$) calculated from 3000 episodes. Figure 1(c) illustrates the ratio of winning against the test probability of heads for the policies learned from DRVI-LCB with $\sigma = 0.01$ and $\sigma = 0.2$, which are benchmarked against the non-robust optimal policy of the nominal MDP using the exact model. It can be seen that the policies learned from DRVI-LCB deviate from the non-robust optimal policy as σ increases, which achieves better worst-case rates of winning across a wide range of perturbed environments. On the other end, while the non-robust policy maximizes the performance when the test environment is close to the history one used for training, its performance degenerates to be much worse than the robust policies when the probability of heads is mismatched significantly, especially when p_{head} drops below, say around, 0.5.

Impact of the number of states. We evaluate the performance of DRVI-LCB and DRVI when the number of states varies within $[25, 50, 100, 150, 200, 250]$, using a fixed sample size $N = 300$. Figure 1(d) show that DRVI-LCB performs consistently better than DRVI as the number of states S increases, where the value gap exhibits a linear scaling with respect to the number of states.

Performance using trajectory data. Instead of using independent samples, we now evaluate the proposed algorithm using a dataset consisting of K sample trajectories generated from a uniform random policy, for the same setting of Figure 1(b). Figure 2 shows the sub-optimality value gap with respect to the number of trajectories K, where the performance of both DRVI-LCB and DRVI improves as K increases, and DRVI-LCB achieves better performance especially when K is small, consistent with the observation under independent data.

6. Conclusion

To accommodate both model robustness and sample efficiency, this paper proposes a distributionally robust model-based algorithm for offline RL with the principle of pessimism. We study the finite-sample complexity of the proposed algorithm DRVI-LCB, and estab-

2. The probability distribution vector $\rho \in \Delta(\mathcal{S})$ is generated as $\rho(s) = u_s / \sum_{s \in \mathcal{S}} u_s$, where u_s is drawn independently from a uniform distribution.

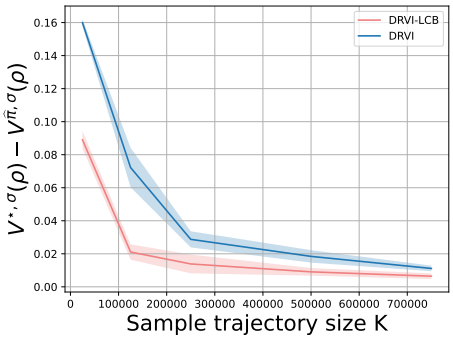


Figure 2: Performance of the proposed algorithm DRVI-LCB compared against DRVI using trajectory data.

lish an information-theoretic lower bound to benchmark its near-optimality for a range of uncertainty levels. Numerical experiments are provided to demonstrate the efficacy of the proposed algorithm. To the best our knowledge, this provides the first provably near-optimal robust offline RL algorithm that learns under model perturbation and partial coverage. This work opens up several interesting directions.

- *Tightening the gap between upper and lower bounds.* Our upper and lower bounds still leave room for future improvements. For example, it is yet to establish the information-theoretic lower bound over the full range of the uncertainty set, and close the gap between the upper and lower bounds with respect to the horizon length.

- *Model-free algorithms for robust offline RL.* Can we design provably efficient model-free algorithms for robust offline RL with partial coverage? Recent works (Wang et al., 2023a,b) in understanding robust variants of Q-learning in the generative model might shed light on how to approach this question.

- *Choice of uncertainty sets.* Moreover, it is possible to extend our framework to handle uncertainty sets defined using other distances such as the chi-square distance and the total variation distance in a similar fashion. Shi et al. (2023) recently established near minimax-optimal sample complexities for the total variation distance and the chi-square distance in the generative model setting, paving ways to study these uncertainty sets in the offline setting.

- *Adaptive tuning of the uncertainty set.* In this work, we treat the radius of the uncertainty set as a fixed, a priori specified parameter, and study the sample complexity of learning a robust optimal policy with respect to the given uncertainty set modeling the sim-to-real gap. It is of great interest to incorporate the tuning of the uncertainty set (both its size and metric) to complete the pipeline of the algorithm design, which will require a different framework than the one adopted in the current paper.

We leave these questions to future investigations.

Acknowledgments

This work 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 authors thank Gen Li for helpful discussions.

Appendix A. Preliminaries

Before starting, let us introduce some additional notation useful throughout the theoretical analysis. Let $\operatorname{ess\,inf} X$ denote the essential infimum of a function/variable X.

A.1 Properties of the robust Bellman operator

To begin with, we introduce the following strong duality lemma which is widely used in distributionally robust optimization when the uncertainty set is defined with respect to the KL divergence.

Lemma 11 ((Hu and Hong, 2013), Theorem 1) *Suppose $f(x)$ has a finite moment generating function in some neighborhood around $x = 0$, then for any $\sigma > 0$ and a nominal distribution P^0, we have*

$$\sup_{\mathcal{P}\in\mathcal{U}^\sigma(P^0)} \mathbb{E}_{X\sim\mathcal{P}}[f(X)] = \inf_{\lambda\geq 0}\left\{\lambda\log\mathbb{E}_{X\sim P^0}\left[\exp\left(\frac{f(X)}{\lambda}\right)\right] + \lambda\sigma\right\}. \tag{62}$$

Armed with the above lemma, it is easily verified that for any positive constant M and a nominal distribution vector $P^0 \in \mathbb{R}^{1\times S}$ supported over the state space $\mathcal{S}$, if $X(s) \in [0, M]$ for all $s \in \mathcal{S}$, then

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0)} \mathcal{P}X = \sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^0\exp\left(-\frac{X}{\lambda}\right)\right) - \lambda\sigma\right\}. \tag{63}$$

For convenience, we introduce the following lemma, paraphrased from Zhou et al. (2021, Lemma 4) and its proof, to further characterize several essential properties of the optimal dual value.

Lemma 12 ((Zhou et al., 2021)) *Let $X \sim P$ be a bounded random variable with $X \in [0, M]$. Let $\sigma > 0$ be any uncertainty level and the corresponding optimal dual variable be*

$$\lambda^\star \in \arg\max_{\lambda\geq 0}\ f(\lambda, P), \qquad \text{where } f(\lambda, P) \coloneqq \left\{-\lambda\log\mathbb{E}_{X\sim P}\left[\exp\left(\frac{-X}{\lambda}\right)\right] - \lambda\sigma\right\}. \tag{64}$$

Then the optimal value $\lambda^\star$ obeys

$$\lambda^\star \in \left[0, \frac{M}{\sigma}\right], \tag{65}$$

where $\lambda^\star = 0$ if and only if

$$\log\big(\mathbb{P}(X = \operatorname{essinf} X)\big) + \sigma \geq 0. \tag{66}$$

Moreover, when $\lambda^\star = 0$, we have

$$\lim_{\lambda\to 0} f(\lambda, P) = \lim_{\lambda\to 0}\left\{-\lambda\log\mathbb{E}_{X\sim P}\left[\exp\left(\frac{-X}{\lambda}\right)\right] - \lambda\sigma\right\} = \operatorname{essinf} X. \tag{67}$$

A.2 Concentration inequalities

In light of Lemma 12 (cf. 67), we are interested in comparing the values of $\operatorname{essinf} X$ when X is drawn from the population nominal distribution or its empirical estimate. This is supplied by the following lemma from Zhou et al. (2021).

Lemma 13 ((Zhou et al., 2021)) *Let $X \sim P$ be a discrete bounded random variable with $X \in [0, M]$. Let P_n denote the empirical distribution constructed from n independent samples $X_1, X_2, \cdots, X_n$, and let $\widehat{X} \sim P_n$. Denote $P_{\mathsf{min},X}$ as the smallest positive probability $P_{\mathsf{min},X} \coloneqq \min\{\mathbb{P}(X = x) : x \in \mathsf{supp}(X)\}$, where $\mathsf{supp}(X)$ is the support of X. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\min_{i\in[n]} X_i = \operatorname{essinf} \widehat{X} = \operatorname{essinf} X, \tag{68}$$

as long as

$$n \geq -\frac{\log(2/\delta)}{\log(1 - P_{\mathsf{min},X})}. \tag{69}$$

We next gather a few elementary facts about the Binomial distribution, which will be useful throughout the proof.

Lemma 14 (Chernoff's inequality) *Suppose $N \sim \mathsf{Binomial}(n, p)$, where $n \geq 1$ and $p \in [0, 1)$. For some universal constant $c_{\mathsf{f}} > 0$, we have*

$$\mathbb{P}\left(|N/n - p| \geq pt\right) \leq \exp\left(-c_{\mathsf{f}} npt^2\right), \qquad \forall t \in [0, 1]. \tag{70}$$

Lemma 15 ((Shi et al., 2022, Lemma 8)) *Suppose $N \sim \mathsf{Binomial}(n, p)$, where $n \geq 1$ and $p \in [0, 1]$. For any $\delta \in (0, 1)$, we have*

$$N \geq \frac{np}{8\log\left(\frac{1}{\delta}\right)} \qquad \text{if } np \geq 8\log\left(\frac{1}{\delta}\right), \tag{71a}$$

$$N \leq \begin{cases} e^2 np & \text{if } np \geq \log\left(\frac{1}{\delta}\right), \\ 2e^2 \log\left(\frac{1}{\delta}\right) & \text{if } np \leq 2\log\left(\frac{1}{\delta}\right) \end{cases} \tag{71b}$$

hold with probability at least $1 - 4\delta$.

A.3 Kullback-Leibler (KL) divergence

We next introduce some useful facts about the Kullback-Leibler (KL) divergence for two distributions P and Q, denoted as $\mathsf{KL}(P \parallel Q)$. Denoting $\mathsf{Ber}(p)$(resp. $\mathsf{Ber}(q)$) as the Bernoulli distribution with mean p (resp. q), we introduce

$$\mathsf{KL}\big(\mathsf{Ber}(p) \parallel \mathsf{Ber}(q)\big) \coloneqq p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}, \tag{72}$$

which represents the KL divergence from $\mathsf{Ber}(p)$ to $\mathsf{Ber}(q)$. We now introduce the following lemma.

Lemma 16 *For any $p, q \in \left[\frac{1}{2}, 1\right)$ and $p > q$, it holds that*

$$\mathsf{KL}\big(\mathsf{Ber}(p) \parallel \mathsf{Ber}(q)\big) \leq \mathsf{KL}\big(\mathsf{Ber}(q) \parallel \mathsf{Ber}(p)\big) \leq \frac{(p-q)^2}{p(1-p)}. \tag{73}$$

Moreover, for any $0 \leq x < y < q$, it holds

$$\mathsf{KL}\left(\mathsf{Ber}\left(x\right) \parallel \mathsf{Ber}(q)\right) > \mathsf{KL}\left(\mathsf{Ber}\left(y\right) \parallel \mathsf{Ber}(q)\right). \tag{74}$$

Proof The first half of this lemma is proven in Li et al. (2022, Lemma 10). For the latter half, it follows from that the function

$$f(x, q) \coloneqq \mathsf{KL}\left(\mathsf{Ber}\left(x\right) \parallel \mathsf{Ber}(q)\right)$$

is monotonically decreasing for all $x \in (0, q]$, since its derivative with respect to x satisfies $\frac{\partial f(x,q)}{\partial x} = \log \frac{x}{q} + \log \frac{1-q}{1-x} < 0$. ■

Appendix B. Analysis: episodic finite-horizon RMDPs

B.1 Proof of Theorem 3

Before starting, we introduce several additional notation that will be useful in the analysis. First, we denote the state-action space covered by the behavior policy π^{b} in the nominal model P^0 as

$$\mathcal{C}^{\mathsf{b}} = \left\{(h, s, a) : d_h^{\mathsf{b}}(s, a) > 0\right\}. \tag{75}$$

Moreover, we recall the definition in (23) and define a similar one based on the exact nominal model P^0 as

$$P_{\mathsf{min},h}(s, a) \coloneqq \min_{s'} \left\{P_h^0(s' \,|\, s, a) : \; P_h^0(s' \,|\, s, a) > 0\right\}. \tag{76}$$

Clearly, by comparing with the definitions (24) and (25), it holds that

$$P_{\mathsf{min}}^{\star} = \min_{h,s} \; P_{\mathsf{min},h}(s, \pi_h^{\star}(s)), \qquad P_{\mathsf{min}}^{\mathsf{b}} = \min_{(h,s,a) \in \mathcal{C}^{\mathsf{b}}} \; P_{\mathsf{min},h}(s, a). \tag{77}$$

For any time step $h \in [H]$, we denote the set of possible state occupancy distributions associated with the optimal policy $\pi^{\star}$ in a model within the uncertainty set $P \in \mathcal{U}^{\sigma}\left(P^0\right)$ as

$$\mathcal{D}_h^{\star} \coloneqq \left\{\left[d_h^{\star,P}(s)\right]_{s \in \mathcal{S}} : P \in \mathcal{U}^{\sigma}\left(P^0\right)\right\} = \left\{\left[d_h^{\star,P}\big(s, \pi_h^{\star}(s)\big)\right]_{s \in \mathcal{S}} : P \in \mathcal{U}^{\sigma}\left(P^0\right)\right\}, \tag{78}$$

where the second equality is due to the fact that $\pi^{\star}$ is chosen to be deterministic.

With these in place, the proof of Theorem 3 is separated into several key steps, as outlined below.

Step 1: establishing the pessimism property. To achieve this claim, we heavily count on the following lemma whose proof can be found in Appendix B.2.

Lemma 17 *Instate the assumptions in Theorem 3. Then for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, consider any vector $V \in \mathbb{R}^S$ independent of $\widehat{P}^0_{h,s,a}$ obeying $\|V\|_\infty \le H$. With probability at least $1-\delta$, one has*

$$\left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}V - \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V \right| \le b_h(s,a) \tag{79}$$

with $b_h(s,a)$ given in (22). Moreover, for all $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$, with probability at least $1-\delta$, one has

$$\frac{P_{\mathsf{min},h}(s,a)}{8\log(KHS/\delta)} \le \widehat{P}_{\mathsf{min},h}(s,a) \le e^2 P_{\mathsf{min},h}(s,a). \tag{80}$$

Armed with the above lemma, with probability at least $1-\delta$, we shall show the following relation holds

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H+1]: \qquad \widehat{Q}_h(s,a) \le Q_h^{\widehat{\pi},\sigma}(s,a), \qquad \widehat{V}_h(s) \le V_h^{\widehat{\pi},\sigma}(s), \tag{81}$$

which means that $\widehat{Q}_h$ (resp. $\widehat{V}_h$) is a pessimistic estimate of $Q_h^{\widehat{\pi},\sigma}$ (resp. $V_h^{\widehat{\pi},\sigma}$). Towards this, it is easily verified that the latter assertion concerning $V_h^{\widehat{\pi},\sigma}$ is implied by the former, since

$$\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a) \le \max_a Q_h^{\widehat{\pi},\sigma}(s,a) = V_h^{\widehat{\pi},\sigma}(s). \tag{82}$$

Therefore, the remainder of this step focuses on verifying the former assertion in (81) by induction.

- To begin, the claim (81) holds at the base case when $h = H+1$, by invoking the trivial fact $\widehat{Q}_{H+1}(s,a) = Q_{H+1}^{\widehat{\pi},\sigma}(s,a) = 0$.

- Then, suppose that $\widehat{Q}_{h+1}(s,a) \le Q_{h+1}^{\widehat{\pi},\sigma}(s,a)$ holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ at some time step $h \in [H]$, it boils down to show $\widehat{Q}_h(s,a) \le Q_h^{\widehat{\pi},\sigma}(s,a)$.

  By the update rule of $\widehat{Q}_h(s,a)$ in Algorithm 2 (cf. line 7), the above relation holds immediately if $\widehat{Q}_h(s,a) = 0$ since $\widehat{Q}_h(s,a) = 0 \le Q_h^{\widehat{\pi},\sigma}(s,a)$. Otherwise, $\widehat{Q}_h(s,a)$ is updated via

$$\begin{aligned}
&\widehat{Q}_h(s,a) \\
&= r_h(s,a) + \sup_{\lambda \ge 0} \left\{ -\lambda \log\left(\widehat{P}^0_{h,s,a} \cdot \exp\left(\frac{-\widehat{V}_{h+1}}{\lambda} \right) \right) - \lambda\sigma \right\} - b_h(s,a) \\
&\overset{\text{(i)}}{=} r_h(s,a) + \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} - b_h(s,a) \\
&\le r_h(s,a) + \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} + \left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} - \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} \right| - b_h(s,a)
\end{aligned}$$

$$\overset{\text{(ii)}}{\leq} r_h(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V^{\widehat{\pi},\sigma}_{h+1} + 0 \overset{\text{(iii)}}{=} Q^{\widehat{\pi},\sigma}_h(s,a), \tag{83}$$

where (i) rewrites the update rule back to its primal form (cf. (19)), (ii) holds by applying (79) with the condition (28) satisfied and the induction hypothesis $\widehat{V}_{h+1} \leq V^{\widehat{\pi},\sigma}_{h+1}$, and lastly, (iii) follows by the robust Bellman consistency equation (8).

Putting them together, we have verified the claim (81) by induction.

Step 2: bounding $V^{\star,\sigma}_h(s) - V^{\widehat{\pi},\sigma}_h(s)$. With the pessimism property (81) in place, we observe that the following relation holds

$$0 \leq V^{\star,\sigma}_h(s) - V^{\widehat{\pi},\sigma}_h(s) \leq V^{\star,\sigma}_h(s) - \widehat{V}_h(s) \leq Q^{\star,\sigma}_h\big(s,\pi^\star_h(s)\big) - \widehat{Q}_h\big(s,\pi^\star_h(s)\big), \tag{84}$$

where the last inequality follows from $\widehat{Q}_h\big(s,\pi^\star_h(s)\big) \leq \max_a \widehat{Q}_h(s,a) = \widehat{V}_h(s)$. Then, by the robust Bellman optimality equation in (9) and the primal version of the update rule (cf. (19))

$$\begin{aligned}
Q^{\star,\sigma}_h\big(s,\pi^\star_h(s)\big) &= r_h\big(s,\pi^\star_h(s)\big) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}V^{\star,\sigma}_{h+1}, \\
\widehat{Q}_h\big(s,\pi^\star_h(s)\big) &= r_h\big(s,\pi^\star_h(s)\big) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} - b_h\left(s,\pi^\star_h(s)\right),
\end{aligned}$$

we arrive at

$$\begin{aligned}
V^{\star,\sigma}_h(s) - \widehat{V}_h(s) &\leq Q^{\star,\sigma}_h\big(s,\pi^\star_h(s)\big) - \widehat{Q}_h\big(s,\pi^\star_h(s)\big) \\
&= \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}V^{\star,\sigma}_{h+1} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} + b_h\big(s,\pi^\star_h(s)\big) \\
&\leq \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}V^{\star,\sigma}_{h+1} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} \\
&\quad + \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} \right| + b_h\big(s,\pi^\star_h(s)\big) \\
&\overset{\text{(i)}}{\leq} \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}V^{\star,\sigma}_{h+1} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} + 2b_h\big(s,\pi^\star_h(s)\big) \\
&\overset{\text{(ii)}}{\leq} \widehat{P}^{\inf}_{h,s,\pi^\star_h(s)}\big(V^{\star,\sigma}_{h+1} - \widehat{V}_{h+1}\big) + 2b_h\big(s,\pi^\star_h(s)\big),
\end{aligned} \tag{85}$$

where (i) holds by applying Lemma 2 (cf. (79)) since $\widehat{V}_{h+1}$ is independent of $P^0_{h,s,\pi^\star_h(s)}$ by construction, and (ii) arises from introducing the notation

$$\widehat{P}^{\inf}_{h,s,\pi^\star_h(s)} \coloneqq \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} \tag{86}$$

and consequently,

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}V_{h+1}^{\star,\sigma} \le \widehat{P}^{\inf}_{h,s,\pi_h^\star(s)} V_{h+1}^{\star,\sigma}, \qquad \text{and} \qquad \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}\widehat{V}_{h+1} = \widehat{P}^{\inf}_{h,s,\pi_h^\star(s)} \widehat{V}_{h+1}.$$

To continue, let us introduce some additional notation for convenience. Define a sequence of matrices $\widehat{P}_h^{\inf} \in \mathbb{R}^{S\times S}$ and vectors $b_h^\star \in \mathbb{R}^S$ for $h \in [H]$, where their s-th rows (resp. entries) are given by

$$\left[\widehat{P}_h^{\inf}\right]_{s,\cdot} = \widehat{P}^{\inf}_{h,s,\pi_h^\star(s)}, \qquad \text{and} \qquad b_h^\star(s) = b_h\big(s,\pi_h^\star(s)\big). \tag{87}$$

Applying (85) recursively over the time steps $h, h+1, \cdots, H$ using the above notation gives

$$\begin{aligned} 0 \le V_h^{\star,\sigma} - \widehat{V}_h &\le \widehat{P}_h^{\inf}\big(V_{h+1}^{\star,\sigma} - \widehat{V}_{h+1}\big) + 2b_h^\star \\ &\le \widehat{P}_h^{\inf}\widehat{P}_{h+1}^{\inf}\big(V_{h+2}^{\star,\sigma} - \widehat{V}_{h+2}\big) + 2\widehat{P}_h^{\inf} b_{h+1}^\star + 2b_h^\star \le \cdots \le 2\sum_{i=h}^{H}\left(\prod_{j=h}^{i-1}\widehat{P}_j^{\inf}\right) b_i^\star, \end{aligned} \tag{88}$$

where we let $\left(\prod_{j=i}^{i-1}\widehat{P}_j^{\inf}\right) = I$ for convenience.

For any $d_h^\star \in \mathcal{D}_h^\star$ (cf. (78)), taking inner product with (88) leads to

$$\left\langle d_h^\star, V_h^{\star,\sigma} - \widehat{V}_h\right\rangle \le \left\langle d_h^\star, 2\sum_{i=h}^{H}\left(\prod_{j=h}^{i-1}\widehat{P}_j^{\inf}\right) b_i^\star\right\rangle = 2\sum_{i=h}^{H}\left\langle d_i^\star, b_i^\star\right\rangle, \tag{89}$$

where

$$d_i^\star \coloneqq \left[\left(d_h^\star\right)^\top\left(\prod_{j=h}^{i-1}\widehat{P}_j^{\inf}\right)\right]^\top \in \mathcal{D}_i^\star \tag{90}$$

by the definition of $\mathcal{D}_i^\star$ (cf. (78)) for all $i = h+1, \cdots, H$.

Step 3: controlling $\langle d_i^\star, b_i^\star\rangle$ using concentrability. Since $\langle d_i^\star, b_i^\star\rangle = \sum_{s\in\mathcal{S}} d_i^\star(s) b_i^\star(s)$, we shall divide the discussion in two different cases.

- For $s \in S$ where $\max_{P\in\mathcal{U}^\sigma(P^0)} d_i^{\star,P}\big(s,\pi_i^\star(s)\big) = \max_{P\in\mathcal{U}^\sigma(P^0)} d_i^{\star,P}(s) = 0$, it follows from the definition (cf. (78)) that for any $d_i^\star \in \mathcal{D}_i^\star$, it satisfies that

$$d_i^\star(s) = 0. \tag{91}$$

- For $s \in S$ where $\max_{P\in\mathcal{U}^\sigma(P^0)} d_i^{\star,P}\big(s,\pi_i^\star(s)\big) = \max_{P\in\mathcal{U}^\sigma(P^0)} d_i^{\star,P}(s) > 0$, by the assumption in (14)

$$\max_{P\in\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d_i^{\star,P}\big(s,\pi_i^\star(s)\big), \frac{1}{S}\right\}}{d_i^{\mathsf{b}}\big(s,\pi_i^\star(s)\big)} = \max_{P\in\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d_i^{\star,P}(s), \frac{1}{S}\right\}}{d_i^{\mathsf{b}}\big(s,\pi_i^\star(s)\big)} \le C^\star_{\mathsf{rob}} < \infty,$$

it implies that

$$d_i^{\mathsf{b}}\big(s,\pi_i^\star(s)\big)>0 \quad \text{and} \quad \big(i,s,\pi_i^\star(s)\big)\in\mathcal{C}^{\mathsf{b}}. \tag{92}$$

Lemma 2 tells that with probability at least $1-8\delta$,

$$\begin{aligned} N_i\big(s,\pi_i^\star(s)\big) &\geq \frac{Kd_i^{\mathsf{b}}\big(s,\pi_i^\star(s)\big)}{8}-5\sqrt{Kd_i^{\mathsf{b}}\big(s,\pi_i^\star(s)\big)\log\frac{KH}{\delta}} \overset{\text{(i)}}{\geq} \frac{Kd_i^{\mathsf{b}}\big(s,\pi_i^\star(s)\big)}{16} \\ &\overset{\text{(ii)}}{\geq} \frac{K\max_{P\in\mathcal{U}^\sigma(P^0)}\min\Big\{d_i^{\star,P}\big(s,\pi_i^\star(s)\big),\frac{1}{S}\Big\}}{16C_{\mathsf{rob}}^\star} \geq \frac{K\min\big\{d_i^\star(s),\frac{1}{S}\big\}}{16C_{\mathsf{rob}}^\star}, \end{aligned} \tag{93}$$

where (i) holds due to

$$Kd_i^{\mathsf{b}}\big(s,\pi_i^\star(s)\big)\geq c_1\frac{d_i^{\mathsf{b}}\big(s,\pi_i^\star(s)\big)\log(KHS/\delta)}{d_{\mathsf{min}}^{\mathsf{b}}P_{\mathsf{min}}^{\mathsf{b}}}\geq\frac{c_1\log\frac{KH}{\delta}}{P_{\mathsf{min}}^{\mathsf{b}}}\geq c_1\log\frac{KH}{\delta} \tag{94}$$

for some sufficiently large c_1, where the first inequality follows from Condition (28), the second inequality follows from

$$d_{\mathsf{min}}^{\mathsf{b}}=\min_{h,s,a}\Big\{d_h^{\mathsf{b}}(s,a):d_h^{\mathsf{b}}(s,a)>0\Big\}\leq d_i^{\mathsf{b}}\big(s,\pi_i^\star(s)\big) \tag{95}$$

and the last inequality follows from $P_{\mathsf{min}}^{\mathsf{b}}\leq 1$. In addition, (ii) follows from Assumption 1.

With this in place, we observe that the pessimistic penalty (see (22)) obeys

$$\begin{aligned} b_i^\star(s) &\leq c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{\widehat{P}_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)N_i\big(s,\pi_i^\star(s)\big)}} \overset{\text{(i)}}{\leq} 4c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log^2(\frac{KHS}{\delta})}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)N_i\big(s,\pi_i^\star(s)\big)}} \\ &\leq 16c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{C_{\mathsf{rob}}^\star\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K\min\big\{d_i^\star(s),\frac{1}{S}\big\}}}, \end{aligned} \tag{96}$$

where (i) holds by applying (80) in view of the fact that $\big(i,s,\pi_i^\star(s)\big)\in\mathcal{C}^{\mathsf{b}}$ by (92), and the last inequality holds by (93).

Combining the results in the above two cases leads to

$$\begin{aligned} \sum_{s\in\mathcal{S}}d_i^\star(s)b_i^\star(s) &\leq \sum_{s\in\mathcal{S}}16d_i^\star(s)c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{C_{\mathsf{rob}}^\star\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K\min\big\{d_i^\star(s),\frac{1}{S}\big\}}} \\ &\overset{\text{(i)}}{\leq} 16c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\sum_{s\in\mathcal{S}}d_i^\star(s)\frac{C_{\mathsf{rob}}^\star\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K\min\big\{d_i^\star(s),\frac{1}{S}\big\}}}\sqrt{\sum_{s\in\mathcal{S}}d_i^\star(s)} \\ &\leq 32c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{SC_{\mathsf{rob}}^\star\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K}}, \end{aligned} \tag{97}$$

where (i) follows from the Cauchy-Schwarz inequality and the last inequality hold by the trivial fact

$$\sum_{s\in\mathcal{S}}\frac{d_i^\star(s)}{\min\big\{d_i^\star(s),\frac{1}{S}\big\}}\leq\sum_{s\in\mathcal{S}}d_i^\star(s)\left(\frac{1}{d_i^\star(s)}+\frac{1}{1/S}\right)=\sum_{s\in\mathcal{S}}1+\frac{1}{S}\sum_{s\in\mathcal{S}}d_i^\star(s)\leq 2S. \tag{98}$$

Step 4: finishing up the proof. Then, inserting (97) back into (89) with $h = 1$ shows

$$\left\langle d_1^\star, V_1^{\star,\sigma} - \widehat{V}_1 \right\rangle \le 2\sum_{i=1}^{H} \langle d_i^\star, b_i^\star \rangle \le \sum_{i=1}^{H} 64 c_{\mathrm{b}} \frac{H}{\sigma} \sqrt{\frac{SC^\star_{\mathsf{rob}} \log^2 \frac{KH}{\delta}}{P_{\mathsf{min},i}\big(s, \pi_i^\star(s)\big) K}} \le c_2 \frac{H^2}{\sigma} \sqrt{\frac{SC^\star_{\mathsf{rob}} \log^2 \frac{KH}{\delta}}{P^\star_{\mathsf{min}} K}}, \tag{99}$$

where the last inequality holds by plugging in the relation $P^\star_{\mathsf{min}} \le P_{\mathsf{min},i}\big(s, \pi_i^\star(s)\big)$ for $i = 1, \ldots, H$ by the definition in (24) (see also (77)), and choosing c_2 to be large enough. The proof is completed.

B.2 Proof of Lemma 17

To begin, we shall introduce the following fact that

$$\forall (h, s, a) \in \mathcal{C}^{\mathsf{b}} : \qquad N_h(s, a) \ge \frac{c_1 \log \frac{KHS}{\delta}}{16 P_{\mathsf{min},h}(s, a)} \ge -\frac{\log \frac{2KHS}{\delta}}{\log(1 - P_{\mathsf{min},h}(s, a))}, \tag{100}$$

as long as Condition (28) holds. The proof is postponed to Appendix B.2.3. With this in mind, we shall first establish the simpler bound (80) and then move on to show (79).

B.2.1 Proof of (80)

To begin, recall that (100) is satisfied for all $(h, s, a) \in \mathcal{C}^{\mathsf{b}}$. By Lemma 15 and the union bound, it holds that with probability at least $1 - \delta$ that for all $(h, s, a) \in \mathcal{C}^{\mathsf{b}}$:

$$\forall s' \in \mathcal{S} : \qquad P_h^0(s' \,|\, s, a) \ge \frac{\widehat{P}_h^0(s' \,|\, s, a)}{e^2} \ge \frac{P_h^0(s' \,|\, s, a)}{8e^2 \log(\frac{KHS}{\delta})}. \tag{101}$$

To characterize the relation between $P_{\mathsf{min},h}(s, a)$ and $\widehat{P}_{\mathsf{min},h}(s, a)$ for any $(h, s, a) \in \mathcal{C}^{\mathsf{b}}$, we suppose—without loss of generality—that $P_{\mathsf{min},h}(s, a) = P_h^0(s_1 \,|\, s, a)$ and $\widehat{P}_{\mathsf{min},h}(s, a) = \widehat{P}_h^0(s_2 \,|\, s, a)$ for some $s_1, s_2 \in \mathcal{S}$. Then, it follows that

$$\begin{aligned} P_{\mathsf{min},h}(s, a) = P_h^0(s_1 \,|\, s, a) &\overset{\text{(i)}}{\ge} \frac{\widehat{P}_h^0(s_1 \,|\, s, a)}{e^2} \ge \frac{\widehat{P}_{\mathsf{min},h}(s, a)}{e^2} = \frac{\widehat{P}_h^0(s_2 \,|\, s, a)}{e^2} \\ &\overset{\text{(ii)}}{\ge} \frac{P_h^0(s_2 \,|\, s, a)}{8e^2 \log(\frac{KHS}{\delta})} \ge \frac{P_{\mathsf{min},h}(s, a)}{8e^2 \log(\frac{KHS}{\delta})}, \end{aligned}$$

where (i) and (ii) follow from (101).

B.2.2 Proof of (79)

The main goal of (79) is to control the gap between robust Bellman operations based on the nominal transition kernel $P^0_{h,s,a}$ and the estimated kernel $\widehat{P}^0_{h,s,a}$ by the constructed penalty term. Towards this, first consider $(h, s, a) \notin \mathcal{C}^{\mathsf{b}}$, which corresponds to the state-action pairs (s, a) that haven't been visited at step h by the behavior policy. In other words, $N_h(s, a) = 0$. In this case, (79) can be easily verified that

$$\left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}V - \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V \right| \overset{\text{(i)}}{=} \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V \le \|V\|_\infty \overset{\text{(ii)}}{\le} H \overset{\text{(iii)}}{=} b_h(s, a), \tag{102}$$

where (i) follows from the fact $\widehat{P}^0_{h,s,a} = 0$ when $N_h(s,a) = 0$ (see (15)), (ii) arises from the assumption $\|V\|_\infty \le H$, and (iii) holds by the definition of $b_h(s,a)$ in (22). Therefore, the remainder of the proof will focus on verifying (79) for $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$. Rewriting the term of interest via duality (cf. Lemma 11) yields

$$\begin{aligned}&\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V\right| \\ &= \left|\sup_{\lambda\ge 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\} - \sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\}\right|.\end{aligned} \tag{103}$$

Denoting

$$\widehat{\lambda}^\star_{h,s,a} \coloneqq \arg\max_{\lambda\ge 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\}, \tag{104a}$$

$$\lambda^\star_{h,s,a} \coloneqq \arg\max_{\lambda\ge 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\}, \tag{104b}$$

Lemma 12 (cf. (65)) then gives that

$$\lambda^\star_{h,s,a} \in \left[0, \frac{H}{\sigma}\right], \qquad \widehat{\lambda}^\star_{h,s,a} \in \left[0, \frac{H}{\sigma}\right], \tag{105}$$

due to $\|V\|_\infty \le H$. We shall control (103) in three different cases separately: (a) $\lambda^\star_{h,s,a} = 0$ and $\widehat{\lambda}^\star_{h,s,a} = 0$; (b) $\lambda^\star_{h,s,a} > 0$ and $\widehat{\lambda}^\star_{h,s,a} = 0$ or $\lambda^\star_{h,s,a} = 0$ and $\widehat{\lambda}^\star_{h,s,a} > 0$; and (c) $\lambda^\star_{h,s,a} \ne 0$ or $\widehat{\lambda}^\star_{h,s,a} \ne 0$.

Case (a): $\lambda^\star_{h,s,a} = 0$ and $\widehat{\lambda}^\star_{h,s,a} = 0$. Applying Lemma 12 and Lemma 13 to (103) gives that, with probability at least $1 - \frac{\delta}{KH}$,

$$\begin{aligned}\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V\right| &\overset{\text{(i)}}{=} \left|\operatorname{essinf}_{s\sim\widehat{P}^0_{h,s,a}} V(s) - \operatorname{essinf}_{s\sim P^0_{h,s,a}} V(s)\right| \\ &\overset{\text{(ii)}}{=} \left|\operatorname{essinf}_{s\sim P^0_{h,s,a}} V(s) - \operatorname{essinf}_{s\sim P^0_{h,s,a}} V(s)\right| \\ &= 0 \le b_h(s,a).\end{aligned} \tag{106}$$

where (i) holds by Lemma 12 (cf. (67)) and (ii) arises from Lemma 13 (cf. (68)) given (100).

Case (b): $\lambda^\star_{h,s,a} > 0$ and $\widehat{\lambda}^\star_{h,s,a} = 0$ or $\lambda^\star_{h,s,a} = 0$ and $\widehat{\lambda}^\star_{h,s,a} > 0$. Towards this, note that two trivial facts are implied by the definition (104):

$$\sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\} \ge -\widehat{\lambda}^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right) - \widehat{\lambda}^\star_{h,s,a}\sigma, \tag{107a}$$

$$\sup_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\geq -\lambda^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda^\star_{h,s,a}}\right)\right)-\lambda^\star_{h,s,a}\sigma. \tag{107b}$$

To continue, first, we consider a subcase when $\lambda^\star_{h,s,a}=0$ and $\widehat{\lambda}^\star_{h,s,a}>0$. With probability at least $1-\frac{\delta}{KH}$, it follows from Lemma 12 (cf. (67)) and Lemma 13 (cf. (68)) that

$$\begin{aligned}\sup_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}&\geq \lim_{\lambda\to 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\\&=\operatorname{essinf}_{s\sim\widehat{P}^0_{h,s,a}}V(s)=\operatorname{essinf}_{s\sim P^0_{h,s,a}}V(s)\\&=\sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\},\end{aligned} \tag{108}$$

leading to

$$\begin{aligned}&\left|\sup_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}-\sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\right|\\&\overset{\text{(i)}}{\leq}\left(-\widehat{\lambda}^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right)-\widehat{\lambda}^\star_{h,s,a}\sigma\right)\\&\quad-\left(-\widehat{\lambda}^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right)-\widehat{\lambda}^\star_{h,s,a}\sigma\right)\\&\leq\widehat{\lambda}^\star_{h,s,a}\left|\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right)-\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right)\right|,\end{aligned} \tag{109}$$

where (i) follows from the definition of $\widehat{\lambda}^\star_{h,s,a}$ in (104) and the fact in (107a).

We pause to claim that with probability at least $1-\delta$, the following bound holds

$$\forall(h,s,a)\in\mathcal{C}^{\mathsf{b}},\,V\in\mathbb{R}^S:\quad\frac{\left|\left(\widehat{P}^0_{h,s,a}-P^0_{h,s,a}\right)\cdot\exp\left(\frac{-V}{\lambda}\right)\right|}{P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)}\leq\sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}}N_h(s,a)P_{\mathsf{min},h}(s,a)}}\leq\frac{1}{2}. \tag{110}$$

The proof is postponed to Appendix B.2.4. With (110) in place, we can further bound (109) (which is plugged into (103)) as

$$\begin{aligned}&\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})}\mathcal{P}V-\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})}\mathcal{P}V\right|\\&\leq\widehat{\lambda}^\star_{h,s,a}\left|\log\left(1+\frac{\left(\widehat{P}^0_{h,s,a}-P^0_{h,s,a}\right)\cdot\exp\left(\frac{-V}{\lambda}\right)}{P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)}\right)\right|\\&\overset{\text{(i)}}{\leq}2\widehat{\lambda}^\star_{h,s,a}\frac{\left|\left(\widehat{P}^0_{h,s,a}-P^0_{h,s,a}\right)\cdot\exp\left(\frac{-V}{\lambda}\right)\right|}{P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)}\end{aligned}$$

$$\overset{\text{(ii)}}{\leq} \frac{2H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}}N_h(s,a)P_{\mathsf{min},h}(s,a)}}$$
$$\leq \frac{2eH}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}}N_h(s,a)\widehat{P}_{\mathsf{min},h}(s,a)}} \leq c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{\widehat{P}_{\mathsf{min},h}(s,a)N_h(s,a)}}, \tag{111}$$

where (i) follows from $\log(1+x) \leq 2|x|$ for any $|x| \leq \frac{1}{2}$ in view of (110), (ii) follows from (105) as well as (110), and the last line follows from (80) and choosing c_{b} to be sufficiently large.

Moreover, note that it can be easily verified that

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{h,s,a})}\mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{h,s,a})}\mathcal{P}V\right| \leq H$$

due to the assumption $\|V\|_\infty \leq H$. Plugging in the definition of $b_h(s,a)$ in (22), combined with the above bounds, we have that with probability at least $1-\delta$,

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{h,s,a})}\mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{h,s,a})}\mathcal{P}V\right| \leq \min\left\{c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{N_h(s,a)\widehat{P}_{\mathsf{min},h}(s,a)}}\ ,\ H\right\} =: b_h(s,a). \tag{112}$$

The other subcase when $\lambda^\star_{h,s,a} > 0$ and $\widehat{\lambda}^\star_{h,s,a} = 0$ follows similarly from the bound

$$\left|\sup_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\} - \sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\right|$$
$$\leq \lambda^\star_{h,s,a}\left|\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda^\star_{h,s,a}}\right)\right) - \log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda^\star_{h,s,a}}\right)\right)\right|, \tag{113}$$

and therefore, will be omitted for simplicity.

Case (c): $\lambda^\star_{h,s,a} > 0$ and $\widehat{\lambda}^\star_{h,s,a} > 0$. It follows that

$$\left|\sup_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\} - \sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\right|$$
$$\overset{\text{(i)}}{\leq} \max\Bigg\{\bigg(-\widehat{\lambda}^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot e^{\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}}\right)$$
$$-\widehat{\lambda}^\star_{h,s,a}\sigma\bigg) - \bigg(-\widehat{\lambda}^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot e^{\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}}\right) - \widehat{\lambda}^\star_{h,s,a}\sigma\bigg),$$
$$\left(-\lambda^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot e^{\frac{-V}{\lambda^\star_{h,s,a}}}\right) - \lambda^\star_{h,s,a}\sigma\right) - \left(-\lambda^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot e^{\frac{-V}{\lambda^\star_{h,s,a}}}\right) - \lambda^\star_{h,s,a}\sigma\right)\Bigg\}$$
$$\leq \max_{\lambda\in\{\lambda^\star_{h,s,a},\widehat{\lambda}^\star_{h,s,a}\}}\lambda\left|\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)\right) - \log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)\right)\right|, \tag{114}$$

where (i) can be verified by applying the facts in (107). Hence, the above term (114) can be controlled again in a similar manner as (109); we omit the details for simplicity.

Summing up. Combining the previous results in different cases by the union bound, with probability at least $1-10\delta$, it is satisfied that for all $(h,s,a)\in\mathcal{C}^{\mathsf{b}}$:

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{h,s,a})}\mathcal{P}V-\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{h,s,a})}\mathcal{P}V\right|\le b_h(s,a),$$

which concludes the proof.

B.2.3 Proof of (100)

Observe that for all $(h,s,a)\in\mathcal{C}^{\mathsf{b}}$:

$$Kd_h^{\mathsf{b}}(s,a)\overset{\text{(i)}}{\ge}\frac{c_1 d_h^{\mathsf{b}}(s,a)\log(KHS/\delta)}{d_{\mathsf{min}}^{\mathsf{b}}P_{\mathsf{min}}^{\mathsf{b}}}\overset{\text{(ii)}}{\ge}\frac{c_1\log(KHS/\delta)}{P_{\mathsf{min}}^{\mathsf{b}}}\overset{\text{(iii)}}{\ge}\frac{c_1\log(KHS/\delta)}{P_{\mathsf{min},h}(s,a)},\tag{115}$$

where (i) follows from Condition (28), (ii) follows from the definition that $d_{\mathsf{min}}^{\mathsf{b}}\le d_h^{\mathsf{b}}(s,a)$ for $(h,s,a)\in\mathcal{C}^{\mathsf{b}}$, and (iii) comes from (77).

Lemma 2 then tells that with probability at least $1-8\delta$,

$$\begin{aligned}N_h(s,a)&\ge\frac{Kd_h^{\mathsf{b}}(s,a)}{8}-5\sqrt{Kd_h^{\mathsf{b}}(s,a)\log\frac{KH}{\delta}}\\&\ge\frac{Kd_i^{\mathsf{b}}(s,a)}{16}\ge\frac{c_1\log\frac{KH}{\delta}}{16P_{\mathsf{min},h}(s,a)},\end{aligned}\tag{116}$$

where the second line follows from the above relation as long as c_1 is sufficiently large. The last inequality of (100) then follows from

$$\frac{c_1\log\frac{KHS}{\delta}}{16P_{\mathsf{min},h}(s,a)}\ge-\frac{\log\frac{2KHS}{\delta}}{\log(1-P_{\mathsf{min},h}(s,a))},\tag{117}$$

since $x\le-\log(1-x)$ for all $x\in[0,1]$.

B.2.4 Proof of (110)

Denoting

$$\mathsf{supp}\big(P^0_{h,s,a}\big):=\big\{s'\in\mathcal{S}:P_h^0(s'\,|\,s,a)>0\big\}$$

as the support of $P^0_{h,s,a}$, we observe that

$$\begin{aligned}\frac{\left|\left(\widehat{P}^0_{h,s,a}-P^0_{h,s,a}\right)\cdot\exp\left(\frac{-V}{\lambda}\right)\right|}{P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)}&\le\frac{\sum_{s'\in\mathsf{supp}\left(P^0_{h,s,a}\right)}\left|\widehat{P}_h^0(s'\,|\,s,a)-P_h^0(s'\,|\,s,a)\right|\exp\left(\frac{-V(s')}{\lambda}\right)}{\sum_{s'\in\mathsf{supp}\left(P^0_{h,s,a}\right)}P_h^0(s'\,|\,s,a)\exp\left(\frac{-V(s')}{\lambda}\right)}\\&\le\max_{s'\in\mathsf{supp}\left(P^0_{h,s,a}\right)}\frac{\left|\widehat{P}_h^0(s'\,|\,s,a)-P_h^0(s'\,|\,s,a)\right|}{P_h^0(s'\,|\,s,a)},\end{aligned}\tag{118}$$

where the second line follows from $\sum_i a_i=\sum_i b_i\frac{a_i}{b_i}\le(\max_i\frac{a_i}{b_i})\sum_i b_i$ for any positive sequences $\{a_i,b_i\}_i$ obeying $a_i,b_i>0$.

To continue, note that for any $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$ and $s' \in \mathsf{supp}\big(P^0_{h,s,a}\big)$, $N_h(s,a)\widehat{P}^0_h(s' \,|\, s,a)$ follows the binomial distribution $\mathsf{Binomial}\left(N_h(s,a), P^0_h(s' \,|\, s,a)\right)$. Thus, applying Lemma 14 with $t = \sqrt{\frac{\log\left(\frac{KHS}{\delta}\right)}{c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a)}}$ yields

$$\mathbb{P}\left(\left|\widehat{P}^0_h(s' \,|\, s,a) - P^0_h(s' \,|\, s,a)\right| \geq P^0_h(s' \,|\, s,a)t\right) \leq \exp\left(-c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a) t^2\right) \leq \frac{\delta}{KHS}, \tag{119}$$

as soon as $t \leq \frac{1}{2}$, which can be verified by the fact (100) and $P_{\mathsf{min},h}(s,a) \leq P^0_h(s' \,|\, s,a)$ (cf. (76)), namely,

$$N_h(s,a) \geq \frac{c_1 \log\frac{KHS}{\delta}}{16 P_{\mathsf{min},h}(s,a)} \geq \frac{\log\left(\frac{KHS}{\delta}\right)}{4 c_{\mathsf{f}} P_{\mathsf{min},h}(s,a)} \geq \frac{\log\left(\frac{KHS}{\delta}\right)}{4 c_{\mathsf{f}} P^0_h(s' \,|\, s,a)} \tag{120}$$

as long as c_1 is sufficiently large.

Applying (119) and taking the union bound over $s \in \mathsf{supp}\big(P^0_{h,s,a}\big)$ lead to that with probability at least $1 - \frac{\delta}{KH}$,

$$\begin{aligned}
\max_{s' \in \mathsf{supp}\left(P^0_{h,s,a}\right)} \frac{\left|\widehat{P}^0_h(s' \,|\, s,a) - P^0_h(s' \,|\, s,a)\right|}{P^0_h(s' \,|\, s,a)} &\leq \max_{s' \in \mathsf{supp}\left(P^0_{h,s,a}\right)} \frac{P^0_h(s' \,|\, s,a)\sqrt{\frac{\log\left(\frac{KHS}{\delta}\right)}{c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a)}}}{P^0_h(s' \,|\, s,a)} \\
&= \max_{s' \in \mathsf{supp}\left(P^0_{h,s,a}\right)} \sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a)}} \\
&\leq \sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}} N_h(s,a) P_{\mathsf{min},h}(s,a)}} \leq \frac{1}{2},
\end{aligned}$$

where the last line uses again (120). Plugging this back into (118) and applying the union bound over $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$ then completes the proof.

B.3 Proof of Theorem 4

The proof of Theorem 4 is inspired by the construction in Li et al. (2022) for standard MDPs, but is considerably more involved to handle the uncertainty set unique in robust MDPs. In particular, we construct two different classes of hard instances for different range of the uncertainty level σ to achieve a tighter σ-dependent lower bound. In what follows, we start with the lower bound for the case when the uncertainty level is relatively small, by first constructing some hard instances and then characterizing the sample complexity requirements over these instances. We then move onto the case when the uncertainty level is relatively large, and carry out a similar argument.

B.3.1 Construction of hard problem instances: small uncertainty level

Construction of a collection of hard MDPs To begin, let's consider a collection $\Theta \subseteq \{0,1\}^H$, consisting of vectors with H dimensions. The Gilbert-Varshamov lemma

(Gilbert, 1952) tells that there exists a set $\Theta \subseteq \{0,1\}^H$ such that:

$$|\Theta| \geq e^{H/8} \qquad \text{and} \qquad \|\theta - \widetilde{\theta}\|_1 \geq \frac{H}{8} \quad \text{for any } \theta, \widetilde{\theta} \in \Theta \text{ obeying } \theta \neq \widetilde{\theta}. \tag{121}$$

Armed with Θ, we then generate a collection of MDPs

$$\mathsf{MDP}(\Theta) = \left\{ \mathcal{M}^\theta = \left(\mathcal{S}, \mathcal{A}, P^\theta = \{P_h^{\theta_h}\}_{h=1}^H, \{r_h\}_{h=1}^H, H\right) \mid \theta = [\theta_h]_{1 \leq h \leq H} \in \Theta \right\}, \tag{122}$$

where

$$\mathcal{S} = \{0, 1, \ldots, S-1\}, \qquad \text{and} \qquad \mathcal{A} = \{0, 1\}.$$

The transition kernel $P^\theta = \{P_h^{\theta_h}\}_{h=1}^H$ of the MDP $\mathcal{M}^\theta$ is specified as follows:

$$P_h^{\theta_h}(s' \,|\, s, a) = \begin{cases} p\mathbb{1}(s'=0) + (1-p)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0, \theta_h) \\ q\mathbb{1}(s'=0) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0, 1-\theta_h) \\ \mathbb{1}(s'=1) & \text{if} \quad s = 1 \\ q\mathbb{1}(s'=s) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad s > 1 \end{cases} \tag{123}$$

for any $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Here, p and q are set according to

$$p = 1 - \frac{c_1}{H} \qquad \text{and} \qquad q = p - \frac{c_2 \varepsilon}{H^2} \tag{124}$$

for $c_1 = 1/8$ and some c_2 that satisfies

$$\frac{c_2 \varepsilon}{H^2} \leq \frac{c_1}{2H} \leq \frac{1}{8}. \tag{125}$$

Clearly, it follows that

$$p > q \geq \frac{1}{2} \tag{126}$$

by construction. Furthermore, the MDP will stay in the state subset $\{0,1\}$ if its initial state falls in $\{0,1\}$. The reward function of these MDPs is set as

$$r_h(s,a) = \begin{cases} 1 & \text{if} \quad s = 0 \\ 0 & \text{if} \quad s > 0 \end{cases} \tag{127}$$

for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

Uncertainty set of the transition kernels. Denote the transition kernel vector as

$$P_{h,s,a}^\theta \coloneqq P_h^\theta(\cdot \,|\, s, a) \in [0,1]^{1 \times S}. \tag{128}$$

For any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the perturbation of the transition kernels in $\mathcal{M}^\theta$ is restricted to the following uncertainty set

$$\mathcal{U}^\sigma(P^\theta) \coloneqq \otimes \mathcal{U}^\sigma\left(P_{h,s,a}^\theta\right), \qquad \mathcal{U}^\sigma(P_{h,s,a}^\theta) \coloneqq \left\{ P_{h,s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{h,s,a} \parallel P_{h,s,a}^\theta\right) \leq \sigma \right\} \tag{129}$$

with the uncertainty level σ satisfying

$$0 < \sigma \le \frac{1}{20H}. \tag{130}$$

Before continuing, we shall introduce some notation for convenience. For any $P_h^{\theta_h}(\cdot \,|\, s, a)$ in (123), we define the limit of the perturbed kernel transiting to the next state s' from the current state-action pair (s, a) by

$$\underline{P}_h^{\theta_h}(s' \,|\, s, a) \coloneqq \inf_{P_{h,s,a} \in \mathcal{U}^\sigma(P_{h,s,a}^{\theta_h})} P_h(s' \,|\, s, a), \tag{131}$$

and in particular, denote

$$\underline{p}_h \coloneqq \underline{P}_h^{\theta_h}(0 \,|\, 0, \theta_h), \qquad \underline{q}_h \coloneqq \underline{P}_h^{\theta_h}(0 \,|\, 0, 1-\theta_h). \tag{132}$$

Armed with the above definitions, we introduce the following lemma which implies some useful properties of the uncertainty set.

Lemma 18 *The perturbed transition kernels obey*

$$\underline{p}_1 = \underline{p}_2 = \cdots \underline{p}_H, \quad \underline{q}_1 = \underline{q}_2 = \cdots \underline{q}_H. \tag{133}$$

Denoting $\underline{p}_1 \coloneqq p^\star$ and $\underline{q}_1 \coloneqq q^\star$, we have that when the uncertainty level σ satisfies (130)*,*

$$\underline{p}_\star \ge \underline{q}_\star \ge 1 - \frac{c_3}{H} \quad \text{and} \quad \underline{p}_\star - \underline{q}_\star \ge p - q \ge 0 \tag{134}$$

for constant $c_3 = 2c_1 = \frac{1}{4}$.

Proof The proof is postponed to Appendix B.3.5. ■

Value functions and optimal policies. We take a moment to derive the corresponding value functions and identify the optimal policies. With some abuse of notation, for any MDP $\mathcal{M}_\theta$, we denote $\pi^{\star,\theta} = \{\pi_h^{\star,\theta}\}_{h=1}^H$ as the optimal policy, and let $V_h^{\pi,\sigma,\theta}$ (resp. $V_h^{\star,\sigma,\theta}$) represent the robust value function of policy π (resp. $\pi^{\star,\theta}$) at step h with uncertainty radius σ. Armed with these notation, we introduce the following lemma which collects the properties concerning the value functions and optimal policies.

Lemma 19 *Consider any $\theta \in \Theta$ and any policy π. Then it holds that*

$$V_h^{\pi,\sigma,\theta}(0) = 1 + x_h^{\pi,\theta} V_{h+1}^{\pi,\sigma,\theta}(0) \tag{135}$$

for any $h \in [H]$, where

$$x_h^{\pi,\theta} = \underline{p}_\star \pi_h(\theta_h \,|\, 0) + \underline{q}_\star \pi_h(1-\theta_h \,|\, 0). \tag{136}$$

In addition, for any $h \in [H]$ and $s \in \mathcal{S} \setminus \{0\}$, the optimal policies and the optimal value functions obey

$$\pi_h^{\star,\theta}(\theta_h \,|\, 0) = 1, \qquad V_h^{\star,\sigma,\theta}(0) \ge \tfrac{2}{3}(H+1-h), \tag{137a}$$

$$\pi_h^{\star,\theta}(\theta_h \,|\, s) = 1, \qquad V_h^{\star,\sigma,\theta}(s) = 0, \tag{137b}$$

provided that $0 < c_1 \le 1/2$.

Proof See Appendix B.3.6. ■

Construction of the history/batch dataset. In the nominal environment $\mathcal{M}_\theta$, a batch dataset is generated consisting of K *independent* sample trajectories each of length H, where each trajectory is generated according to (10), based on the following initial state distribution ρ^{b} and behavior policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{h=1}^H$:

$$\rho^{\mathsf{b}}(s) = \mu(s) \quad \text{and} \quad \pi_h^{\mathsf{b}}(a \,|\, s) = \frac{1}{2}, \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{138}$$

Here, $\mu(s)$ is defined as the following state distribution supported on the state subset $\{0,1\}$:

$$\mu(s) = \frac{1}{CS}\mathbb{1}(s=0) + \Big(1 - \frac{1}{CS}\Big)\mathbb{1}(s=1), \tag{139}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and $C > 0$ is some constant that will determine the concentrability coefficient $C^\star_{\mathsf{rob}}$ (as we shall detail momentarily) and obeys

$$\frac{1}{CS} \le \frac{1}{4}. \tag{140}$$

As it turns out, for any MDP $\mathcal{M}_\theta$, the occupancy distributions of the above batch dataset are the same (due to symmetry) and admit the following simple characterization:

$$d_1^{\mathsf{b},P^\theta}(0,a) = \frac{1}{2}\mu(0), \qquad \forall a \in \mathcal{A}, \tag{141a}$$

$$\frac{\mu(s)}{2} \le d_h^{\mathsf{b},P^\theta}(s) \le 2\mu(s), \qquad \frac{\mu(s)}{4} \le d_h^{\mathsf{b},P^\theta}(s,a) \le \mu(s), \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{141b}$$

In addition, we choose the following initial state distribution

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{if } s > 0 \end{cases}. \tag{142}$$

With this choice of ρ, the single-policy clipped concentrability coefficient $C^\star_{\mathsf{rob}}$ and the quantity C are intimately connected as follows:

$$2C \le C^\star_{\mathsf{rob}} \le 4C. \tag{143}$$

The proof of the claim (141) and (143) are postponed to Appendix B.3.7.

B.3.2 Establishing the minimax lower bound: small uncertainty level

Towards this, we first make the following claim: for an arbitrary policy π obeying

$$\sum_{h=1}^H \big\|\pi_h(\cdot \,|\, 0) - \pi_h^{\star,\theta}(\cdot \,|\, 0)\big\|_1 \ge \frac{H}{8}, \tag{144}$$

one has

$$\big\langle \rho, V_1^{\star,\sigma,\theta} - V_1^{\pi,\sigma,\theta} \big\rangle > \varepsilon. \tag{145}$$

We shall postpone the proof of this claim to Appendix B.3.8.

Armed with the above claim and following the same arguments in (Li et al., 2022, Section C.3.2), we complete the proof by observing: for some small enough constant c_4, as long as the sample size is beneath

$$N = KH \le \frac{c_4 C^\star_{\mathsf{rob}} S H^4}{\varepsilon^2}, \tag{146}$$

then we necessarily have

$$\inf_{\widehat{\pi}} \max_{\theta \in \Theta} \mathbb{P}_\theta \left\{ V_\theta^{\star,\sigma}(\rho) - V_\theta^{\widehat{\pi},\sigma}(\rho) \ge \varepsilon \right\} \ge \frac{1}{4}, \tag{147}$$

where $\mathbb{P}_\theta$ denote the probability conditioned on that the MDP is $\mathcal{M}_\theta$. We omit the details for brevity and complete the proof; interested readers can referred to (Li et al., 2022, Section C.3.2).

B.3.3 Construction of hard problem instances: large uncertainty level

We now move onto the case when the uncertainty level is relatively large, we construct another class of hard instances, which is almost the same as the previous one except for the transition kernel.

Construction of a collection of hard MDPs. Let us introduce two MDPs

$$\left\{ \mathcal{M}_\phi = \left(\mathcal{S}, \mathcal{A}, P^\phi = \{P_h^\phi\}_{h=1}^H, \{r_h\}_{h=1}^H, H \right) \mid \phi = \{0,1\} \right\}, \tag{148}$$

where the state space is $\mathcal{S} = \{0, 1, \ldots, S-1\}$, and the action space is $\mathcal{A} = \{0,1\}$. The transition kernel P^ϕ of the constructed MDP $\mathcal{M}_\phi$ is defined as

$$P_1^\phi(s' \mid s, a) = \begin{cases} p\mathbb{1}(s'=0) + (1-p)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,\phi) \\ q\mathbb{1}(s'=0) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,1-\phi) \\ \mathbb{1}(s'=1) & \text{if} \quad s=1 \\ q\mathbb{1}(s'=s) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad s>1 \end{cases} \tag{149a}$$

and

$$P_h^\phi(s' \mid s, a) = \mathbb{1}(s'=s), \qquad \forall (h,s,a) \in \{2, \ldots, H\} \times \mathcal{S} \times \mathcal{A}. \tag{149b}$$

In words, except at step $h=1$, the MDP always stays in the same state. Additionally, the MDP will always stay in the state subset $\{0,1\}$ if the initial distribution is supported only on $\{0,1\}$, in view of (149). Here, p and q are set to be

$$p = 1-\alpha \qquad \text{and} \qquad q = 1-\alpha-\Delta \tag{150}$$

for some $H \ge 2e^8$, α and Δ obeying

$$0 < \alpha \le \frac{1}{H} \le \frac{1}{2e^8} \qquad \text{and} \qquad \Delta \le \frac{\alpha}{2} \le \frac{1}{2H} \le \frac{1}{4e^8}, \tag{151}$$

where β is set as

$$\beta \coloneqq \frac{\log \frac{1}{\alpha+\Delta}}{2} \geq \frac{\log(2H/3)}{2} \geq 4. \tag{152}$$

The assumption (151) immediately indicates the facts

$$1 > p > q \geq \frac{1}{2}. \tag{153}$$

Moreover, for any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, the reward function is defined as

$$r_h(s,a) = \begin{cases} 1 & \text{if } s = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{154}$$

Construction of the history/batch dataset. We utilize the same batch dataset described in Appendix B.3.1 and choose the same initial state distribution ρ in (142) As a result, for any MDP $\mathcal{M}_\phi$, the occupancy distributions of the above batch dataset are the same (due to symmetry) and admit the following simple characterization:

$$d_1^{\mathsf{b},P^\phi}(0,a) = \frac{1}{2}\mu(0), \qquad \forall a \in \mathcal{A}, \tag{155a}$$

$$\frac{\mu(s)}{2} \leq d_h^{\mathsf{b},P^\phi}(s) \leq 2\mu(s), \qquad \frac{\mu(s)}{4} \leq d_h^{\mathsf{b},P^\phi}(s,a) \leq \mu(s), \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{155b}$$

The proof of the claim (155) is postponed to Appendix B.3.9.

Uncertainty set of the transition kernels. Denote the transition kernel vector as

$$P_{h,s,a}^\phi \coloneqq P_h^\phi(\cdot \,|\, s,a) \in [0,1]^{1\times S}. \tag{156}$$

For any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, the perturbation of the transition kernels in $\mathcal{M}_\phi$ is restricted to the following uncertainty set

$$\mathcal{U}^\sigma(P^\phi) \coloneqq \otimes\, \mathcal{U}^\sigma\left(P_{h,s,a}^\phi\right), \qquad \mathcal{U}^\sigma(P_{h,s,a}^\phi) \coloneqq \left\{P_{h,s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{h,s,a} \,\|\, P_{h,s,a}^\phi\right) \leq \sigma\right\}, \tag{157}$$

where the radius of the uncertainty set σ obeys:

$$\left(1 - \frac{3}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) \leq \sigma \leq \left(1 - \frac{2}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right). \tag{158}$$

Before continuing, we shall introduce some notation for convenience. For any $P_h^\phi(\cdot \,|\, s,a)$ in (149), we define the limit of the perturbed kernel transiting to the next state s' from the current state-action pair (s,a) by

$$\underline{P}_h^\phi(s' \,|\, s,a) \coloneqq \inf_{P_{h,s,a} \in \mathcal{U}^\sigma(P_{h,s,a}^\phi)} P_h(s' \,|\, s,a), \tag{159}$$

and in particular, denote

$$\underline{p} \coloneqq \underline{P}_1^\phi(0 \,|\, 0,\phi), \qquad \underline{q} = \underline{P}_1^\phi(0 \,|\, 0, 1-\phi). \tag{160}$$

Armed with the above definitions, we introduce the following lemma which implies some useful properties of the uncertainty set.

Lemma 20 *When β satisfies* (152) *and the uncertainty level σ satisfies* (158)*, the perturbed transition kernels obey*

$$\underline{p} \geq \underline{q} \geq \frac{1}{\beta}. \tag{161}$$

Proof See Appendix B.3.10. ■

Value functions and optimal policies. Similar to Appendix B.3.1, for any MDP $\mathcal{M}_\phi$, we denote $\pi^{\star,\phi} = \{\pi_h^{\star,\phi}\}_{h=1}^H$ as the optimal policy, and let $V_h^{\pi,\sigma,\phi}$ (resp. $V_h^{\star,\sigma,\phi}$) represent the robust value function of policy π (resp. $\pi^{\star,\phi}$) at step h with uncertainty radius σ. Then we introduce the following lemma which collects the properties concerning the value functions and optimal policies.

Lemma 21 *For any $\phi = \{0,1\}$ and any policy π, defining*

$$z_\phi^\pi \coloneqq \underline{p}\pi_1(\phi \,|\, 0) + \underline{q}\pi_1(1-\phi \,|\, 0), \tag{162}$$

it holds that

$$V_1^{\pi,\sigma,\phi}(0) = 1 + z_\phi^\pi(H-1). \tag{163}$$

In addition, the optimal policies and the optimal value functions obey

$$V_1^{\star,\sigma,\phi}(0) = 1 + \underline{p}(H-1), \tag{164a}$$

$$\forall h \in [H] \setminus \{1\}: \quad V_h^{\star,\sigma,\phi}(0) = H - h + 1, \tag{164b}$$

$$\forall h \in [H]: \quad \pi_h^{\star,\phi}(\phi \,|\, 0) = 1, \quad \pi_h^{\star,\phi}(\phi \,|\, 1) = 1, \qquad V_h^{\star,\sigma,\phi}(1) = 0. \tag{164c}$$

The robust single-policy clipped concentrability coefficient $C^\star_{\mathsf{rob}}$ obeys

$$2C \leq C^\star_{\mathsf{rob}} \leq 4C. \tag{165}$$

Proof See Appendix B.3.11. ■

In view of Lemma 21, we note that the smallest positive state transition probability of the optimal policy $\pi^\star$ under any MDP $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$ thus can be given by

$$P^\star_{\mathsf{min}} \coloneqq \min_{h,s,s'} \left\{ P_h^\phi\left(s' | s, \pi_h^{\star,\phi}(s)\right) : \; P_h^\phi\left(s' | s, \pi_h^{\star,\phi}(s)\right) > 0 \right\} = P_1^\phi\left(1 | 0, 1-\phi\right) = 1-p, \tag{166}$$

which obeys

$$\alpha = P^\star_{\mathsf{min}} \in (0, 1/H]$$

according to (150) and (151).

B.3.4 Establishing the minimax lower bound: large uncertainty level

We are now ready to establish the sample complexity lower bound. With the choice of the initial distribution ρ in (142), for any policy estimator $\widehat{\pi}$ computed based on the batch dataset, we plan to control the quantity

$$\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\widehat{\pi},\sigma,\phi} \right\rangle = V_1^{\star,\sigma,\phi}(0) - V_1^{\widehat{\pi},\sigma,\phi}(0).$$

Step 1: converting the goal to estimate ϕ. We make the following claim which shall be verified in Appendix B.3.12: given $\varepsilon \leq \frac{H}{384e^6 \log\left(\frac{1}{\alpha}\right)} \leq \frac{H}{384e^6 \log\left(\frac{1}{\alpha+\Delta}\right)}$, choosing

$$\Delta = \frac{128e^6\sigma(1-q)\varepsilon}{H} = \frac{128e^6\sigma(\alpha+\Delta)\varepsilon}{H} \leq \frac{128e^6(\alpha+\Delta)\varepsilon \log\left(\frac{1}{\alpha+\Delta}\right)}{H} \leq \frac{\alpha}{2}, \tag{167}$$

which satisfies (151) with the aid of (158) and (150), it holds that for any policy $\widehat{\pi}$,

$$\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\widehat{\pi},\sigma,\phi} \right\rangle \geq 2\varepsilon\big(1 - \widehat{\pi}_1(\phi \,|\, 0)\big). \tag{168}$$

Armed with this relation between the policy $\widehat{\pi}$ and its sub-optimality gap, we are positioned to construct an estimate of ϕ. We denote $\mathbb{P}_\phi$ as the probability distribution when the MDP is $\mathcal{M}_\phi$, for any $\phi \in \{0,1\}$.

Suppose for the moment that a policy estimate $\widehat{\pi}$ achieves

$$\mathbb{P}_\phi\left\{\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\widehat{\pi},\sigma,\phi} \right\rangle \leq \varepsilon\right\} \geq \frac{7}{8}, \tag{169}$$

then in view of (168), we necessarily have $\widehat{\pi}_1(\phi \,|\, 0) \geq \frac{1}{2}$ with probability at least $\frac{7}{8}$. With this in mind, we are motivated to construct the following estimate $\widehat{\phi}$ for $\phi \in \{0,1\}$:

$$\widehat{\phi} = \arg\max_{a \in \{0,1\}} \widehat{\pi}_1(a \,|\, 0), \tag{170}$$

which obeys

$$\mathbb{P}_\phi\big\{\widehat{\phi} = \phi\big\} \geq \mathbb{P}_\phi\big\{\widehat{\pi}_1(\phi \,|\, 0) > 1/2\big\} \geq \frac{7}{8}. \tag{171}$$

In what follows, we would like to show (171) cannot happen without enough samples, which would in turn contradict (168).

Step 2: probability of error in testing two hypotheses. Armed with the above preparation, we shall focus on differentiating the two hypotheses $\phi \in \{0,1\}$. Towards this, consider the minimax probability of error defined as follows:

$$p_{\mathrm{e}} \coloneqq \inf_{\psi} \max\big\{\mathbb{P}_0(\psi \neq 0),\, \mathbb{P}_1(\psi \neq 1)\big\}, \tag{172}$$

where the infimum is taken over all possible tests ψ constructed from the batch dataset.

Let $\mu^{\mathsf{b},\phi}$ (resp. $\mu_h^{\mathsf{b},\phi}(s_h)$) be the distribution of a sample trajectory $\{s_h, a_h\}_{h=1}^H$ (resp. a sample (a_h, s_{h+1}) conditional on s_h) for the MDP $\mathcal{M}_\phi$. Following standard results from Tsybakov and Zaiats (2009, Theorem 2.2) and the additivity of the KL divergence (cf. Tsybakov and Zaiats (2009, Page 85)), we obtain

$$\begin{aligned} p_{\mathrm{e}} &\geq \frac{1}{4}\exp\Big(-K\mathsf{KL}\big(\mu^{\mathsf{b},0} \,\|\, \mu^{\mathsf{b},1}\big)\Big) \\ &\geq \frac{1}{4}\exp\bigg\{-\frac{1}{2}K\mu(0)\Big(\mathsf{KL}\big(P_1^0(\cdot \,|\, 0,0) \,\|\, P_1^1(\cdot \,|\, 0,0)\big) + \mathsf{KL}\big(P_1^0(\cdot \,|\, 0,1) \,\|\, P_1^1(\cdot \,|\, 0,1)\big)\Big)\bigg\}, \end{aligned} \tag{173}$$

where we also use the independence of the K trajectories in the batch dataset in the first line. Here, the second line arises from the chain rule of the KL divergence (Duchi, 2018, Lemma 5.2.8) and the Markov property of the sample trajectories (recall that $d_h^{\mathsf{b},P^0} = d_h^{\mathsf{b},P^1}$) according to

$$\begin{aligned}
\mathsf{KL}\big(\mu^{\mathsf{b},0} \,\|\, \mu^{\mathsf{b},1}\big) &= \sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h \sim d_h^{\mathsf{b},P^0}} \Big[\mathsf{KL}\big(\mu_h^{\mathsf{b},0}(s_h) \,\|\, \mu_h^{\mathsf{b},1}(s_h)\big)\Big] \\
&= \sum_{a\in\{0,1\}} d_1^{\mathsf{b},P^0}(0,a)\mathsf{KL}\left(P_1^0(\cdot\,|\,0,a) \,\|\, P_1^1(\cdot\,|\,0,a)\right) \\
&= \frac{1}{2}\mu(0) \sum_{a\in\{0,1\}} \mathsf{KL}\big(P_1^0(\cdot\,|\,0,a) \,\|\, P_1^1(\cdot\,|\,0,a)\big),
\end{aligned}$$

where the penultimate equality holds by the fact that $P_h^0(\cdot\,|\,s,a)$ and $P_h^1(\cdot\,|\,s,a)$ only differ when $h=1$ and $s=0$, and the last equality follows from (155).

It remains to control the KL divergence terms in (173). Given $p \geq q \geq 1/2$ (cf. (153)), applying Lemma 16 (cf. (73)) yields

$$\begin{aligned}
\mathsf{KL}\big(P_1^0(\cdot\,|\,0,0) \,\|\, P_1^1(\cdot\,|\,0,0)\big) = \mathsf{KL}\left(p \,\|\, q\right) \leq \frac{(p-q)^2}{(1-p)p} &\overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\
&\overset{\text{(ii)}}{=} \frac{128^2 e^{12}\sigma^2(1-q)^2\varepsilon^2}{H^2 p(1-p)} \\
&\overset{\text{(iii)}}{\leq} \frac{c_1\sigma^2 P_{\mathsf{min}}^\star \varepsilon^2}{H^2},
\end{aligned} \tag{174}$$

where (i) follows from the definition (150), (ii) holds by plugging in the expression of Δ in (167), (iii) arises from $1-q \leq 2(1-p) = 2P_{\mathsf{min}}^\star$ (see (151) and (166)), $p > \frac{1}{2}$, as long as c_1 is a large enough constant. It can be shown that $\mathsf{KL}\big(P_1^0(\cdot\,|\,0,1) \,\|\, P_1^1(\cdot\,|\,0,1)\big)$ can be upper bounded in the same way. Substituting (174) back into (173) demonstrates that: if the sample size is chosen as

$$KH \leq \frac{H^3 S C_{\mathsf{rob}}^\star \log 2}{4c_1 P_{\mathsf{min}}^\star \sigma^2 \varepsilon^2}, \tag{175}$$

then one necessarily has

$$\begin{aligned}
p_{\mathrm{e}} \geq \frac{1}{4}\exp\left\{-\frac{1}{2}K\mu(0)\cdot 2\frac{c_1\sigma^2 P_{\mathsf{min}}^\star\varepsilon^2}{H^2}\right\} &\overset{\text{(i)}}{=} \frac{1}{4}\exp\left\{-K\frac{c_1\sigma^2 P_{\mathsf{min}}^\star\varepsilon^2}{SCH^2}\right\} \\
&\overset{\text{(ii)}}{\geq} \frac{1}{4}\exp\left\{-K\frac{4c_1\sigma^2 P_{\mathsf{min}}^\star\varepsilon^2}{SC_{\mathsf{rob}}^\star H^2}\right\} \geq \frac{1}{8},
\end{aligned} \tag{176}$$

where (i) follows from (139) and (ii) holds by (165).

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

$$\mathbb{P}_0\big\{\big\langle \rho, V_1^{\star,\sigma,0} - V_1^{\widehat{\pi},\sigma,0}\big\rangle > \varepsilon\big\} < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1\big\{\big\langle \rho, V_1^{\star,\sigma,1} - V_1^{\widehat{\pi},\sigma,1}\big\rangle > \varepsilon\big\} < \frac{1}{8}.$$

Then Step 1 tells us that the estimator $\widehat{\phi}$ defined in (170) must satisfy

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

which cannot happen under the sample size condition in (175) to avoid contradition with (176). The proof is thus finished.

B.3.5 Proof of Lemma 18

First, (133) can be easily verified by the definition of $\underline{p}_h$ and $\underline{q}_h$ in (132) and the transition $P_h^{\theta_h}$ in (123).

Proof of the first inequality in (134)**.** It is observed that

$$\begin{aligned}
&\mathsf{KL}\left(\mathsf{Ber}\left(1-\frac{2c_1}{H}\right)\,\|\,\mathsf{Ber}(q)\right)\\
&=\left(1-\frac{2c_1}{H}\right)\log\left(\frac{1-\frac{2c_1}{H}}{q}\right)+\frac{2c_1}{H}\log\left(\frac{\frac{2c_1}{H}}{1-q}\right)\\
&=\left(1-\frac{2c_1}{H}\right)\log\left(1+\frac{1-q-\frac{2c_1}{H}}{q}\right)+\frac{2c_1}{H}\log\left(1+\frac{q-1+\frac{2c_1}{H}}{1-q}\right)\\
&\overset{\text{(i)}}{\geq}\left(1-\frac{2c_1}{H}\right)\frac{\frac{1-q-\frac{2c_1}{H}}{q}}{2q}\frac{1+q-\frac{2c_1}{H}}{1-\frac{2c_1}{H}}+\frac{2c_1}{H}\cdot 2\frac{q-1+\frac{2c_1}{H}}{1-q+\frac{2c_1}{H}}\\
&=(q-1+\frac{2c_1}{H})\left[-\frac{1+q-\frac{2c_1}{H}}{2q^2}+\frac{4c_1}{H}\frac{1}{1-q+\frac{2c_1}{H}}\right]\\
&\overset{\text{(ii)}}{\geq}\frac{c_1}{2H}\left[\frac{-1}{(1-\frac{3c_1}{2H})^2}+\frac{8}{7}\right]\geq\frac{1}{20H}\geq\sigma,
\end{aligned}\tag{177}$$

where (i) holds by $\log(1+x)\geq\frac{x}{2(1+x)}$ when $0\leq x<\infty$ and $\log(1+x)\geq\frac{x}{2}\frac{2+x}{1+x}$ when $-1<x\leq 0$ (Topsøe, 2007), and the penultimate inequality holds by $1-\frac{3c_1}{2H}\geq 1-\frac{3}{256}\geq\frac{63}{64}$. Here, (ii) can be verified by

$$\begin{aligned}
-\frac{1+q-\frac{2c_1}{H}}{2q^2}\geq-\frac{1+q}{2q^2}\geq\frac{-2}{2q^2}&\overset{\text{(iii)}}{\geq}\frac{-1}{(1-\frac{3c_1}{2H})^2},\\
\frac{4c_1}{H}\frac{1}{1-q+\frac{2c_1}{H}}&\overset{\text{(iv)}}{\geq}\frac{4c_1}{H}\frac{1}{\frac{3c_1}{2H}+\frac{2c_1}{H}}\geq\frac{8}{7},
\end{aligned}$$

where (iii) holds by $q\geq 1-\frac{3c_1}{2H}$, and (iv) arises from $0\leq 1-q\leq\frac{3c_1}{2H}$.

With above fact and $\mathsf{KL}\left(\mathsf{Ber}\left(\underline{q}_\star\right)\,\|\,\mathsf{Ber}(q)\right)=\sigma$ in mind, applying Lemma 16 leads to $\underline{q}_\star\geq 1-\frac{2c_1}{H}$.

Proof of the second inequality in (134)**.** First, observing the first claim in (134), combined with Lemma 16, we know that for any uncertainty level $\sigma \leq \frac{1}{20H}$, there exists a unique $\underline{q}_\star$ obeying $\underline{q}_\star \geq 1 - \frac{2c_1}{H} > \frac{1}{2}$ such that

$$\sigma = \mathsf{KL}\left(\mathsf{Ber}\left(\underline{q}_\star\right) \parallel \mathsf{Ber}(q)\right). \tag{178}$$

Then let us define the following function for $0 < x \leq \frac{1-q}{3}$ (i.e., $p = q + x$)

$$g(x,q) = \mathsf{KL}\left(\mathsf{Ber}\left(\underline{q}_\star + x\right) \parallel \mathsf{Ber}(q+x)\right) = (\underline{q}_\star + x)\log\frac{\underline{q}_\star + x}{q+x} + (1 - \underline{q}_\star - x)\log\frac{1-\underline{q}_\star - x}{1-q-x} \tag{179}$$

The first derivative $\nabla_x g(x,q)$ is

$$\begin{aligned}
&\nabla_x g(x,q) \\
&= \log\left(\frac{\underline{q}_\star + x}{q+x}\right) + (q+x)\frac{q+x-(\underline{q}_\star+x)}{(q+x)^2} - \log\left(\frac{1-\underline{q}_\star - x}{1-q-x}\right) \\
&\quad + (1-q-x)\frac{-(1-q-x)+(1-\underline{q}_\star - x)}{(1-q-x)^2} \\
&= \log\left(\frac{\underline{q}_\star + x}{q+x}\right) - \log\left(\frac{1-\underline{q}_\star - x}{1-q-x}\right) + \frac{q-\underline{q}_\star}{q+x} + \frac{q-\underline{q}_\star}{1-q-x} \\
&= \log\left(1 + \frac{\underline{q}_\star - q}{q+x}\right) - \log\left(1 + \frac{q-\underline{q}_\star}{1-q-x}\right) + \frac{q-\underline{q}_\star}{q+x} + \frac{q-\underline{q}_\star}{1-q-x} \\
&\overset{\text{(i)}}{\geq} \frac{\underline{q}_\star - q}{2(q+x)}\frac{2+\frac{\underline{q}_\star - q}{q+x}}{1+\frac{\underline{q}_\star - q}{q+x}} - \frac{q-\underline{q}_\star}{2(1-q-x)}\frac{2+\frac{q-\underline{q}_\star}{1-q-x}}{1+\frac{q-\underline{q}_\star}{1-q-x}} + \frac{q-\underline{q}_\star}{q+x} + \frac{q-\underline{q}_\star}{1-q-x} \\
&= \frac{q-\underline{q}_\star}{2(q+x)}\left(2 - \frac{2q+2x+\underline{q}_\star - q}{q+x+\underline{q}_\star - q}\right) + \frac{q-\underline{q}_\star}{2(1-q-x)}\left(2 - \frac{2-2q-2x+q-\underline{q}_\star}{1-q-x+q-\underline{q}_\star}\right) \\
&= \frac{q-\underline{q}_\star}{2(q+x)}\frac{\underline{q}_\star - q}{x+\underline{q}_\star} + \frac{q-\underline{q}_\star}{2(1-q-x)}\frac{q-\underline{q}_\star}{1-\underline{q}_\star - x} \\
&= \frac{(q-\underline{q}_\star)^2\left[(q+x)(\underline{q}_\star+x) - (1-q-x)(1-\underline{q}_\star - x)\right]}{2(q+x)(\underline{q}_\star+x)(1-q-x)(1-\underline{q}_\star - x)} \geq 0
\end{aligned} \tag{180}$$

where (i) holds by $\log(1+x) \leq \frac{x}{2}\frac{2+x}{1+x}$ when $0 \leq x < \infty$ and $\log(1+x) \geq \frac{x}{2}\frac{2+x}{1+x}$ when $-1 < x \leq 0$ (Topsøe, 2007), and the last inequality always holds for any $0 < x \leq \frac{1-q}{3}$ and $\underline{q}_\star \geq \frac{1}{2}$.

The above fact shows that $g(p-q,q) \geq g(0,q) = \sigma$, and thus

$$\mathsf{KL}\left(\mathsf{Ber}\left(p - q + \underline{q}_\star\right) \parallel \mathsf{Ber}(p)\right) \geq \sigma, \tag{181}$$

which complete the proof by observing that $\underline{p}_\star \geq p - q + \underline{q}_\star$ via applying Lemma 16 with (178).

B.3.6 Proof of Lemma 19

Ordering the value function for different states. First, note that for any policy π at the final step $H+1$, we have

$$\forall s \in \mathcal{S}: \quad V_{H+1}^{\pi,\sigma,\theta}(s) = 0. \tag{182}$$

Then for any $\theta \in \Theta$ and any policy π, it is easily verified that

$$\begin{aligned} V_H^{\pi,\sigma,\theta}(0) &= \mathbb{E}_{a\sim\pi_h(\cdot\,|\,0)}\left[r_h(0,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{h,0,a}^{\theta_h})} \mathcal{P}V_{H+1}^{\pi,\sigma,\theta}\right] = 1, \\ V_H^{\pi,\sigma,\theta}(s) &= \mathbb{E}_{a\sim\pi_h(\cdot\,|\,s)}\left[r_h(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{h,s,a}^{\theta_h})} \mathcal{P}V_{H+1}^{\pi,\sigma,\theta}\right] = 0, \quad \forall s \in \mathcal{S}\setminus\{0\}, \end{aligned} \tag{183}$$

which directly indicates that

$$\forall s \in \mathcal{S}\setminus\{0,1\}: \quad V_H^{\pi,\sigma,\theta}(0) > V_H^{\pi,\sigma,\theta}(s) = 0. \tag{184}$$

Then we provide the following claim which will be proved momentarily using induction: For any $\theta \in \Theta$ and any policy π, the following equation holds

$$\forall (h,s) \in [H]\times\mathcal{S}\setminus\{0\}: \quad V_h^{\pi,\sigma,\theta}(0) > V_h^{\pi,\sigma,\theta}(s) = 0. \tag{185}$$

The above result leads to the following immediate fact for state $s \in \mathcal{S}\setminus\{0\}$:

$$\forall (h,s) \in [H]\times\mathcal{S}\setminus\{0\}: \quad V_h^{\star,\sigma,\theta}(s) = \max_\pi V_h^{\pi,\sigma,\theta}(s) = 0, \tag{186}$$

since (185) holds for any π and $h \in [H]$. Therefore, for any state $s \in \mathcal{S}\setminus\{0\}$, without loss of generality, we choose the optimal policy obeying

$$\forall (h,s) \in [H]\times\mathcal{S}\setminus\{0\}: \qquad \pi_h^{\star,\theta}(\theta_h\,|\,s) = 1. \tag{187}$$

Then the rest of the proof will focus on deriving the value function and optimal policy over state $s=0$. To begin with, recalling the value function in (192)

$$V_h^{\pi,\sigma,\theta}(0) = 1 + x_h^{\pi,\theta} V_{h+1}^{\pi,\sigma,\theta}(0) \tag{188}$$

and observing that the function $V_h^{\pi,\sigma,\theta}(0)$ is increasing in $x_h^{\pi,\theta}$ and that $x_h^{\pi,\theta}$ is increasing in $\pi_h(\theta_h\,|\,0)$ (due to the fact $\underline{p}_\star \geq \underline{q}_\star$ in (134)). As a result, the optimal policy obeys

$$\pi_h^{\star,\theta}(\theta_h\,|\,0) = 1 \tag{189}$$

at state 0. Plugging it back to (188) gives

$$\begin{aligned} V_h^{\star,\sigma,\theta}(0) &= 1 + x_h^{\pi^\star,\theta} V_{h+1}^{\star,\sigma,\theta}(0) \\ &\overset{\text{(i)}}{=} 1 + \underline{p}_\star V_{h+1}^{\star,\sigma,\theta}(0) \geq \sum_{j=0}^{H-h} \underline{p}_\star^j \geq \sum_{j=0}^{H-h} \left(1 - \frac{c_3}{H}\right)^j = \frac{1-\left(1-\frac{c_3}{H}\right)^{H-h+1}}{c_3/H} \end{aligned}$$

$$\geq \frac{2}{3}(H+1-h). \tag{190}$$

where (i) holds by $x_h^{\pi^\star,\theta} = \underline{p}_\star \pi_h^{\star,\theta}(\theta_h \,|\, 0) + \underline{q}_\star \pi_h^{\star,\theta}(1-\theta_h \,|\, 0) = \underline{p}_\star$. Here, the last inequality holds since we observe that

$$\Big(1-\frac{c_3}{H}\Big)^{H-h+1} \leq \exp\Big(-\frac{c_3}{H}(H-h+1)\Big) \leq 1 - \frac{2c_3(H-h+1)}{3H},$$

as long as $c_3 \leq 0.5$, which follows due to the elementary inequalities $1-x \leq \exp(-x)$ for any $x \geq 0$ and $\exp(-x) \leq 1-2x/3$ for any $0 \leq x \leq 1/2$.

Proof of claim in (185)**.** We shall (185) through induction. Towards this, assuming that at time step $h+1$, the following holds

$$\forall s \in \mathcal{S} \setminus \{0,1\}: \quad V_{h+1}^{\pi,\sigma,\theta}(0) > V_{h+1}^{\pi,\sigma,\theta}(s) = 0. \tag{191}$$

Observing that the base case when $h = H$ has already been confirmed in (184), now we move on to prove the same property for time step h.

To start with, the robust value function of state 0 at step h satisfies

$$\begin{aligned}
V_h^{\pi,\sigma,\theta}(0) &= \mathbb{E}_{a\sim\pi_h(\cdot\,|\,0)}\left[r_h(0,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{h,0,a}^{\theta_h})} \mathcal{P}V_{h+1}^{\pi,\sigma,\theta}\right] \\
&\overset{\text{(i)}}{=} 1 + \pi_h(\theta_h \,|\, 0)\Big(\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{h,0,\theta}^{\theta_h})} \mathcal{P}V_{h+1}^{\pi,\sigma,\theta}\Big) + \pi_h(1-\theta_h \,|\, 0)\Big(\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{h,0,1-\theta_h}^{\theta_h})} \mathcal{P}V_{h+1}^{\pi,\sigma,\theta}\Big) \\
&\overset{\text{(ii)}}{=} 1 + \pi_h(\theta_h \,|\, 0)\Big[\underline{p}V_{h+1}^{\pi,\sigma,\theta}(0) + \big(1-\underline{p}\big)\,V_{h+1}^{\pi,\sigma,\theta}(1)\Big] \\
&\quad + \pi_h(1-\theta_h \,|\, 0)\Big[\underline{q}V_{h+1}^{\pi,\sigma,\theta}(0) + \big(1-\underline{q}\big)\,V_{h+1}^{\pi,\sigma,\theta}(1)\Big] \\
&\overset{\text{(iii)}}{=} 1 + V_{h+1}^{\pi,\sigma,\theta}(1) + x_h^{\pi,\theta}\left[V_{h+1}^{\pi,\sigma,\theta}(0) - V_{h+1}^{\pi,\sigma,\theta}(1)\right] \\
&= 1 + x_h^{\pi,\theta} V_{h+1}^{\pi,\sigma,\theta}(0)
\end{aligned} \tag{192}$$

where (i) uses the definition of the reward function in (127), (ii) uses the induction assumption in (191) so that the infimum is attained by picking the choice specified in (132) with a smallest probability mass imposed on the transition to state 0. Finally, we plug in the definition (136) of $x_h^{\pi,\theta}$ in (iii), and the last line follows from (191).

$$V_h^{\pi,\sigma,\theta}(s) = \mathbb{E}_{a\sim\pi_h(\cdot\,|\,s)}\left[r_h(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{h,s,a}^{\theta_h})} \mathcal{P}V_{h+1}^{\pi,\sigma,\theta}\right] = 0,$$

where the last inequality holds by the reward and transition function in (127) and (123) with the induction assumption (191).

Combining (192) and (193), we complete the proof:

$$\forall s \in \mathcal{S} \setminus \{0\}: \quad V_h^{\pi,\sigma,\theta_h}(0) \geq 1 > V_h^{\pi,\sigma,\theta}(s) = 0. \tag{193}$$

B.3.7 Proof of claim (141) and (143)

Proof of the claim (141)**.** With the initial state distribution and behavior policy defined in (138), we have for any MDP $\mathcal{M}_\theta$,

$$d_1^{\mathsf{b},P^\theta}(s) = \rho^{\mathsf{b}}(s) = \mu(s),$$

which leads to

$$\forall a \in \mathcal{A}: \quad d_1^{\mathsf{b},P^\theta}(0,a) = \frac{1}{2}\mu(0). \tag{194}$$

In view of (149a), the state occupancy distribution at any step $h = 2, 3, \cdots, H$ obeys

$$\begin{aligned} d_h^{\mathsf{b},P^\theta}(0) &\geq \mathbb{P}\left\{s_h = 0 \,|\, s_{h-1} = 0; \pi^{\mathsf{b}}\right\} \geq d_{h-1}^{\mathsf{b},P^\theta}(0)\left[\pi_{h-1}^{\mathsf{b}}(\theta_{h-1}\,|\,0)\underline{p}_\star + \pi_{h-1}^{\mathsf{b}}(1-\theta_{h-1}\,|\,0)\underline{q}_\star\right] \\ &\geq d_{h-1}^{\mathsf{b},P^\theta}(0)\underline{q}_\star \geq \cdots \geq d_1^{\mathsf{b},P^\theta}(0)\prod_{j=0}^{h-1}\underline{q}_\star \geq d_1^{\mathsf{b},P^\theta}(0)\Big(1-\frac{c_3}{H}\Big)^H > \frac{\mu(0)}{2}, \end{aligned} \tag{195}$$

where the last line makes use of the properties $\underline{q}_\star \geq 1 - c_3/H$ in Lemma 18 and

$$\Big(1-\frac{c_3}{H}\Big)^H \geq \Big(1-\frac{1}{2H}\Big)^H > \frac{1}{2}$$

provided that $0 < c_3 = 1/4 < 1/2$. In addition, as state 1 is an absorbing state and state 0 will only transfer to itself or state 1 at each time step, we directly achieve that

$$d_h^{\mathsf{b},P^\theta}(0) \leq d_{h-1}^{\mathsf{b},P^\theta}(0) \leq \cdots \leq d_1^{\mathsf{b},P^\theta}(0) \leq \mu(0). \tag{196}$$

For state 1, as it is absorbing, we directly have

$$d_h^{\mathsf{b},P^\theta}(1) = \mathbb{P}\left\{s_h = 1 \,|\, s_{h-1} = 1; \pi^{\mathsf{b}}\right\} \geq d_{h-1}^{\mathsf{b},P^\theta}(1) \geq \cdots \geq d_1^{\mathsf{b},P^\theta}(1) = \mu(1). \tag{197}$$

According to the assumption in (140), it is easily verified that

$$d_h^{\mathsf{b},P^\theta}(1) \leq 1 \leq 2\mu(1). \tag{198}$$

Finally, combining (195), (196), (197), (198), the definitions of $P_h^{\theta_h}(\cdot\,|\,s,a)$ in (123) and the Markov property, we arrive at for any $(h,s) \in [H] \times \mathcal{S}$,

$$\frac{\mu(s)}{2} \leq d_h^{\mathsf{b},P^\theta}(s) \leq 2\mu(s), \tag{199}$$

which directly leads to

$$\frac{\mu(s)}{4} \leq d_h^{\mathsf{b},P^\theta}(s,a) = \pi_1^{\mathsf{b}}(a\,|\,s) d_h^{\mathsf{b},P^\theta}(s) \leq \mu(s). \tag{200}$$

Proof of the claim (143)**.** Examining the definition of $C^\star_{\mathsf{rob}}$ in (14), we make the following observations.

- For $h = 1$, we have

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d_1^{\star,P}(s,a), \frac{1}{S}\right\}}{d_1^{\mathsf{b},P^\theta}(s,a)} \overset{\text{(i)}}{=} \max_{P\in\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d_1^{\star,P}(0,\theta_h), \frac{1}{S}\right\}}{d_1^{\mathsf{b},P^\theta}(0,\theta_h)} \overset{\text{(ii)}}{=} \max_{P\in\mathcal{U}^\sigma(P^\theta)} \frac{1}{Sd_1^{\mathsf{b},P^\theta}(0,\theta_h)} \overset{\text{(iii)}}{=} \frac{2}{S\mu(0)} = 2C, \tag{201}$$

  where (i) holds by $d_1^{\star,P}(s) = \rho(s) = 0$ for all $s \in \mathcal{S}\setminus\{0\}$ (see (142)) and $\pi_h^{\star,\theta}(\theta_h \,|\, 0) = 1$ for all $h \in [H]$, (ii) follows from the fact $d_1^{\star,P}(0,\theta) = 1$, (iii) is verified in (141), and the last equality arises from the definition in (139).

- Similarly, for $h = 2, 3, \cdots, H$, we arrive at

$$\begin{aligned}\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b},P^\theta}(s,a)} &\overset{\text{(i)}}{=} \max_{s\in\{0,1\},P\in\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d_h^{\star,P}(s,\theta_h), \frac{1}{S}\right\}}{d_h^{\mathsf{b},P^\theta}(s,\theta_h)} \\ &\le \max_{s\in\{0,1\},P\in\mathcal{U}^\sigma(P^\theta)} \frac{1}{Sd_h^{\mathsf{b},P^\theta}(s,\theta_h)} \overset{\text{(ii)}}{\le} \frac{4}{S\mu(0)} = 4C,\end{aligned} \tag{202}$$

  where (i) holds by the optimal policy in (137) and the trivial fact that $d_h^{\star,P}(s) = 0$ for all $s \in \mathcal{S}\setminus\{0,1\}$ (see (142) and (123)), (ii) arises from (141), and the last equality comes from (139).

Combining the above cases, we complete the proof by

$$2C \le C^\star_{\mathsf{rob}} = \max_{(h,s,a,P)\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b},P^\theta}(s,a)} \le 4C.$$

B.3.8 Proof of the claim (145)

By virtue of (136) and (137), we see that $x_h^{\pi^{\star,\theta},\theta} = \underline{p}_\star$ for all $h \in [H]$, which combined with (135) gives

$$\begin{aligned}\left\langle \rho, V_h^{\star,\sigma,\theta} - V_h^{\pi,\sigma,\theta}\right\rangle &= V_h^{\star,\theta}(0) - V_h^{\pi,\theta}(0) \\ &= \underline{p}_\star V_{h+1}^{\star,\sigma,\theta}(0) - x_h^{\pi,\theta} V_{h+1}^{\pi,\theta}(0) \\ &= x_h^{\pi,\theta}\left(V_{h+1}^{\star,\sigma,\theta}(0) - V_{h+1}^{\pi,\sigma,\theta}(0)\right) + (\underline{p}_\star - x_h^{\pi,\theta}) V_{h+1}^{\star,\sigma,\theta}(0) \\ &\overset{\text{(i)}}{\ge} \underline{q}_\star\left(V_{h+1}^{\star,\sigma,\theta}(0) - V_{h+1}^{\pi,\sigma,\theta}(0)\right) + (\underline{p}_\star - x_h^{\pi,\theta}) V_{h+1}^{\star,\sigma,\theta}(0) \\ &\overset{\text{(ii)}}{\ge} \underline{q}_\star\left(V_{h+1}^{\star,\sigma,\theta}(0) - V_{h+1}^{\pi,\sigma,\theta}(0)\right) + \frac{1}{2}(p-q)\left\|\pi_h^{\star,\theta}(\cdot \,|\, 0) - \pi_h(\cdot \,|\, 0)\right\|_1 V_{h+1}^{\star,\sigma,\theta}(0)\end{aligned}$$

$$\overset{\text{(iii)}}{\geq} \underline{q}_\star \left(V_{h+1}^{\star,\sigma,\theta}(0) - V_{h+1}^{\pi,\sigma,\theta}(0)\right) + \frac{c_2\varepsilon}{3H^2}(H+1-h)\left\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\right\|_1 \tag{203}$$

where (i) follows from the fact that $x_h^\pi \geq \underline{q}_\star$ for any π and $h \in [H]$, and (iii) holds by the facts (137) and the choice (124) of (p, q). Here, (ii) arises from

$$\begin{aligned}
\underline{p}_\star - x_h^{\pi,\theta} &= (\underline{p}_\star - \underline{q}_\star)\big(1 - \pi_h(\theta_h\,|\,0)\big) \\
&\geq (p-q)\big(1 - \pi_h(\theta_h\,|\,0)\big) \\
&= \frac{1}{2}(p-q)\big(1 - \pi_h(\theta_h\,|\,0) + \pi_h(1-\theta_h\,|\,0)\big) = \frac{1}{2}(p-q)\left\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\right\|_1,
\end{aligned} \tag{204}$$

where the first inequality holds by applying Lemma 18. With the fact of (203) in mind, combined with the fact $\underline{q}_\star \geq 1 - \frac{c_3}{H}$, following the same proof pipeline of (Li et al., 2022, (276) to (278)) leads to

$$\left\langle \rho, V_h^{\star,\sigma,\theta} - V_h^{\pi,\sigma,\theta} \right\rangle > \varepsilon. \tag{205}$$

We omit the proof here for conciseness.

B.3.9 Proof of (155)

With the initial state distribution and behavior policy defined in (138), we have for any MDP $\mathcal{M}_\phi$ with $\phi \in \{0, 1\}$,

$$d_1^{\mathsf{b},P^\phi}(s) = \rho^{\mathsf{b}}(s) = \mu(s),$$

which leads to

$$\forall a \in \mathcal{A}: \quad d_1^{\mathsf{b},P^\phi}(0,a) = \frac{1}{2}\mu(0). \tag{206}$$

In view of (149a), the state occupancy distribution at step $h = 2$ obeys

$$d_2^{\mathsf{b},P^\phi}(0) = \mathbb{P}\left\{s_2 = 0\,|\,s_1 \sim d_1^{\mathsf{b},P^\phi}; \pi^{\mathsf{b}}\right\} = \mu(0)\left[\pi_1^{\mathsf{b}}(\phi\,|\,0)p + \pi_1^{\mathsf{b}}(1-\phi\,|\,0)q\right] = \frac{(p+q)\mu(0)}{2},$$

and

$$\begin{aligned}
d_2^{\mathsf{b},P^\phi}(1) &= \mathbb{P}\left\{s_2 = 1\,|\,s_1 \sim d_1^{\mathsf{b},P^\phi}; \pi^{\mathsf{b}}\right\} \\
&= \mu(0)\left[\pi_1^{\mathsf{b}}(\phi\,|\,0)(1-p) + \pi_1^{\mathsf{b}}(1-\phi\,|\,0)(1-q)\right] + \mu(1) = \mu(1) + \frac{(2-p-q)\mu(0)}{2}.
\end{aligned}$$

With the above result in mind and recalling the assumption in (153), we arrive at

$$\frac{\mu(0)}{2} \leq d_2^{\mathsf{b},P^\phi}(0) \leq \mu(0), \qquad \mu(1) \leq d_2^{\mathsf{b},P^\phi}(1) \overset{\text{(i)}}{\leq} 2\mu(1), \tag{207}$$

where (i) holds by applying (153) and (140) (which implies $\mu(0) \le \mu(1)$ by the assumption in (140))

$$d_2^{\mathsf{b},P^\phi}(1) = \mu(1) + \frac{(2-p-q)\mu(0)}{2} \le \mu(1) + \mu(0) \le 2\mu(1).$$

Finally, from the definitions of $P_h^\phi(\cdot \,|\, s,a)$ in (149b) and the Markov property, we arrive at for any $(h,s) \in [H] \times \mathcal{S}$,

$$\frac{\mu(s)}{2} \le d_h^{\mathsf{b},P^\phi}(s) \le 2\mu(s), \tag{208}$$

which directly leads to

$$\frac{\mu(s)}{4} \le d_h^{\mathsf{b},P^\phi}(s,a) = \pi_1^{\mathsf{b}}(a \,|\, s) d_h^{\mathsf{b},P^\phi}(s) \le \mu(s). \tag{209}$$

B.3.10 Proof of Lemma 20

Note that $\underline{p} \ge \underline{q}$ can be easily verified since $p > q$, which indicates that the first assertion is true. So we will focus on the second assertion in (161). Towards this, invoking the definition in (72), let σ' be the KL divergence from $\mathsf{Ber}\left(\frac{1}{\beta}\right)$ to $\mathsf{Ber}(q)$, defined as follows

$$\begin{aligned}\sigma' &\coloneqq \mathsf{KL}\left(\mathsf{Ber}\left(\frac{1}{\beta}\right) \,\|\, \mathsf{Ber}(q)\right) = \frac{1}{\beta}\log\frac{\frac{1}{\beta}}{q} + \left(1-\frac{1}{\beta}\right)\log\frac{\left(1-\frac{1}{\beta}\right)}{1-q} \\ &= \left(\frac{1}{\beta}\right)\log\left(\frac{1}{\beta}\right) - \left(\frac{1}{\beta}\right)\log(q) + \left(1-\frac{1}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) + \left(1-\frac{1}{\beta}\right)\log\left(1-\frac{1}{\beta}\right),\end{aligned} \tag{210}$$

where the second line uses the definition of q in (150). We claim that σ' satisfies the following relation with σ, which will be proven at the end of this proof:

$$0 < \sigma \le \left(1-\frac{2}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) \le \sigma' \le \left(1-\frac{1}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right). \tag{211}$$

Recalling the definition of the transition kernel in (149a)

$$P_1^\phi(0 \,|\, 0, 1-\phi) = q, \quad P_1^\phi(1 \,|\, 0, 1-\phi) = 1-q, \quad P_1^\phi(s \,|\, 0, 1-\phi) = 0, \quad \forall s \in \mathcal{S} \setminus \{0,1\},$$

the uncertainty set of the transition kernel with radius σ is thus given as

$$\begin{aligned}\mathcal{U}^\sigma(P^\phi_{1,0,1-\phi}) = \Big\{ P_{1,0,1-\phi} \in \Delta(\mathcal{S}) : &P(0 \,|\, 0, 1-\phi) = q', P(1 \,|\, 0, 1-\phi) = 1-q', \\ &\mathsf{KL}\left(\mathsf{Ber}\left(q'\right) \,\|\, \mathsf{Ber}(q)\right) \le \sigma \Big\}.\end{aligned} \tag{212}$$

Recalling the definition of $\underline{q}$ in (160), we can bound

$$\underline{q} = \inf_{P_{1,0,1-\phi} \in \mathcal{U}^\sigma(P^\phi_{1,0,1-\phi})} P(0 \,|\, 0, 1-\phi) = \inf_{q' : \mathsf{KL}(\mathsf{Ber}(q') \| \mathsf{Ber}(q)) \le \sigma} q'$$

$$\overset{\text{(i)}}{\geq} \inf_{q':\mathsf{KL}(\mathsf{Ber}(q')\|\mathsf{Ber}(q))\leq\sigma'} q' = \frac{1}{\beta},$$

where (i) holds by $\sigma \leq \sigma'$ (cf. (211)) and the last equality follows from applying Lemma 16 (cf. (74)) and (210) to arrive at

$$\forall 0 \leq q' < \frac{1}{\beta}: \qquad \mathsf{KL}\left(\mathsf{Ber}\left(q'\right) \parallel \mathsf{Ber}(q)\right) > \mathsf{KL}\left(\mathsf{Ber}\left(\frac{1}{\beta}\right) \parallel \mathsf{Ber}(q)\right) = \sigma'.$$

Proof of (211)**.** To control σ', we plug in the assumptions in (153) and $\beta \geq 4$ and arrive at the trivial facts

$$\left(\frac{1}{\beta}\right)\log\left(\frac{1}{\beta}\right) - \left(\frac{1}{\beta}\right)\log(q) < 0, \quad \left(1-\frac{1}{\beta}\right)\log\left(1-\frac{1}{\beta}\right) < 0.$$

The above facts directly lead to

$$\sigma' \leq \left(1-\frac{1}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right). \tag{213}$$

Similarly, observing

$$-1 \leq \left(\frac{1}{\beta}\right)\log\left(\frac{1}{\beta}\right) + \left(1-\frac{1}{\beta}\right)\log\left(1-\frac{1}{\beta}\right) \leq 0, \quad -\left(\frac{1}{\beta}\right)\log(q) \geq 0,$$

we arrive at

$$\sigma' \geq -1 + \left(1-\frac{1}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) \geq \left(1-\frac{2}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) \tag{214}$$

as long as $\log\left(\frac{1}{\alpha+\Delta}\right) \geq \beta$ (cf. (152)). With (213) and (214) in hand, it is straightforward to see that the choice of the uncertainty radius σ in (158) obeys the advertised bound (211).

B.3.11 Proof of Lemma 21

For notational conciseness, we shall drop the superscript ϕ and use the shorthand $V_h^{\pi,\sigma} = V_h^{\pi,\sigma,\phi}$ and $V_h^{\star,\sigma} = V_h^{\star,\sigma,\phi}$ whenever it is clear from the context. We begin by deriving the robust value function for any policy π. Starting with state 1, at any step $h \in [H]$, it obeys

$$V_h^{\pi,\sigma}(1) = \mathbb{E}_{a\sim\pi_h(\cdot\,|\,1)}\left[r_h(1,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{h,1,a})} \mathcal{P}V_{h+1}^{\pi,\sigma}\right] = 0 + V_{h+1}^{\pi,\sigma}(1),$$

where the first equality follows from the robust Bellman consistency equation (cf. (8)), and the second equality follows from the observation that the distribution $P^\phi_{h,1,a}$ is supported solely on state 1 in view of (149a), therefore $\mathcal{U}^\sigma(P^\phi_{h,1,a}) = P^\phi_{h,1,a}$. Leveraging the terminal condition $V_{H+1}^{\pi,\sigma}(1) = 0$, and recursively applying the previous relation, we have

$$V_h^{\star,\sigma}(1) = V_h^{\pi,\sigma}(1) = 0, \qquad \forall h \in [H]. \tag{215}$$

Similarly, turning to state 0, at any step $h > 1$, the robust value function satisfies

$$V_h^{\pi,\sigma}(0) = \mathbb{E}_{a\sim\pi_h(\cdot\,|\,0)}\left[r_h(0,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{h,0,a})} \mathcal{P}V_{h+1}^{\pi,\sigma}\right] = 1 + V_{h+1}^{\pi,\sigma}(0),$$

which again uses the fact that the distribution $P^\phi_{h,0,a}$ is supported solely on state 0 in view of (149b), therefore $\mathcal{U}^\sigma(P^\phi_{h,0,a}) = P^\phi_{h,0,a}$. Leveraging the terminal condition $V_{H+1}^{\pi,\sigma}(0) = 0$, and recursively applying the previous relation, we have

$$V_h^{\star,\sigma}(0) = V_h^{\pi,\sigma}(0) = H - h + 1, \qquad 2 \le h \le H. \tag{216}$$

Taking (215) and (216) together, it follows that

$$\forall\ 2 \le h \le H: \qquad V_h^{\pi,\sigma}(0) > V_h^{\pi,\sigma}(1). \tag{217}$$

Consequently, the robust value function of state 0 at step $h = 1$ satisfies

$$\begin{aligned}
V_1^{\pi,\sigma}(0) &= \mathbb{E}_{a\sim\pi_1(\cdot\,|\,0)}\left[r_1(0,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{1,0,a})} \mathcal{P}V_2^{\pi,\sigma}\right] \\
&\overset{\text{(i)}}{=} 1 + \pi_1(\phi\,|\,0)\Big(\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{1,0,\phi})} \mathcal{P}V_2^{\pi,\sigma}\Big) + \pi_1(1-\phi\,|\,0)\Big(\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{1,0,1-\phi})} \mathcal{P}V_2^{\pi,\sigma}\Big) \\
&\overset{\text{(ii)}}{=} 1 + \pi_1(\phi\,|\,0)\Big[\underline{p}V_2^{\pi,\sigma}(0) + \big(1-\underline{p}\big)\,V_2^{\pi,\sigma}(1)\Big] + \pi_1(1-\phi\,|\,0)\Big[\underline{q}V_2^{\pi,\sigma}(0) \\
&\quad + \big(1-\underline{q}\big)\,V_2^{\pi,\sigma}(1)\Big] \\
&\overset{\text{(iii)}}{=} 1 + V_2^{\pi,\sigma}(1) + z_\phi^\pi\left[V_2^{\pi,\sigma}(0) - V_2^{\pi,\sigma}(1)\right] \\
&= 1 + z_\phi^\pi V_2^{\pi,\sigma}(0)
\end{aligned} \tag{218}$$

where (i) uses the definition of the reward function in (154), (ii) uses (217) so that the infimum is attained by picking the choice specified in (160) with a smallest probability mass imposed on the transition to state 0. Finally, we plug in the definition (162) of z_ϕ^π in (iii), and the last line follows from (215).

Therefore, taking $\pi = \pi^{\star,\phi}$ in the previous relation directly leads to

$$V_1^{\star,\sigma}(0) = 1 + z_\phi^{\pi^{\star,\phi}} V_2^{\star,\sigma}(0) = 1 + z_\phi^{\pi^{\star,\phi}}(H-1), \tag{219}$$

where the second equality follows from (216). Observing that the function $(H-1)z$ is increasing in z and that z_ϕ^π is increasing in $\pi_1(\phi\,|\,0)$ (due to the fact $\underline{p} \ge \underline{q}$ in (161)). As a result, the optimal policy obeys

$$\pi_1^{\star,\phi}(\phi\,|\,0) = 1 \tag{220}$$

at state 0, and plugging back to (219) gives

$$V_1^{\star,\sigma}(0) = 1 + z_\phi^{\pi^{\star,\phi}}(H-1) = 1 + \underline{p}(H-1),$$

where $z_\phi^{\pi^{\star,\phi}} = \underline{p}\pi_1^{\star,\phi}(\phi\,|\,0) + \underline{q}\pi_1^{\star,\phi}(1-\phi\,|\,0) = \underline{p}$. For the rest of the states, without loss of generality, we choose the optimal policy obeying

$$\forall h \in [H]: \qquad \pi_h^{\star,\phi}(\phi\,|\,0) = 1, \quad \pi_h^{\star,\phi}(\phi\,|\,1) = 1. \tag{221}$$

Proof of claim (165). Given that $\pi_h^{\star,\phi}(\phi\,|\,0)=1$ for all $h\in[H]$ and $\rho(0)=1$, for any $P\in\mathcal{U}^\sigma(P^\phi)$, we have

$$\begin{aligned} d_2^{\star,P}(0,\phi) &= d_2^{\star,P}(0)\pi_2^{\star,\phi}(\phi\,|\,0) = d_2^{\star,P}(0) = \mathbb{P}_{s_2\sim P(\cdot\,|\,s_1,\pi_1^{\star,\phi}(s_1))}\big\{s_2=0\,|\,s_1\sim\rho;\pi^{\star,\phi}\big\} \\ &= P_1(0\,|\,0,\phi) \overset{\text{(i)}}{\geq} \underline{P}_1^{\phi}(0\,|\,0,\phi) \overset{\text{(ii)}}{=} \underline{p} \geq \frac{1}{\beta}, \end{aligned} \tag{222}$$

which (i) holds by plugging in the definition (159), (ii) follows from the definition (160), and the final inequality arises from Lemma 20. Hence, for all $2\leq h\leq H$, by the Markov property and $P_h^\phi(0\,|\,0,\phi)=1$, we have

$$d_h^{\star,P}(0,\phi) = d_2^{\star,P}(0,\phi) \geq \frac{1}{\beta}. \tag{223}$$

Examining the definition of $C^\star_{\mathsf{rob}}$ in (14), we make the following observations.

- For $h=1$, we have

$$\begin{aligned} \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_1^{\star,P}(s,a),\frac{1}{S}\big\}}{d_1^{\mathsf{b},P^\phi}(s,a)} &\overset{\text{(i)}}{=} \max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_1^{\star,P}(0,\phi),\frac{1}{S}\big\}}{d_1^{\mathsf{b},P^\phi}(0,\phi)} \\ &\overset{\text{(ii)}}{=} \max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{1}{Sd_1^{\mathsf{b},P^\phi}(0,\phi)} \overset{\text{(iii)}}{=} \frac{2}{S\mu(0)} = 2C, \end{aligned} \tag{224}$$

  where (i) holds by $d_1^{\star,P}(s)=\rho(s)=0$ for all $s\in\mathcal{S}\setminus\{0\}$ (see (142)) and $\pi_h^{\star,\phi}(\phi\,|\,0)=1$ for all $h\in[H]$, (ii) follows from the fact $d_1^{\star,P}(0,\phi)=1$, (iii) is verified in (155), and the last equality arises from the definition in (139).

- Similarly, for $h=2$, we arrive at

$$\begin{aligned} \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_2^{\star,P}(s,a),\frac{1}{S}\big\}}{d_2^{\mathsf{b},P^\phi}(s,a)} &\overset{\text{(i)}}{=} \max_{s\in\{0,1\},P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_2^{\star,P}(s,\phi),\frac{1}{S}\big\}}{d_2^{\mathsf{b},P^\phi}(s,\phi)} \\ &\leq \max_{s\in\{0,1\},P\in\mathcal{U}^\sigma(P^\phi)} \frac{1}{Sd_2^{\mathsf{b},P^\phi}(s,\phi)} \overset{\text{(ii)}}{\leq} \frac{4}{S\mu(0)} = 4C, \end{aligned} \tag{225}$$

  where (i) holds by the optimal policy in (164) and the trivial fact that $d_2^{\star,P}(s)=0$ for all $s\in\mathcal{S}\setminus\{0,1\}$ (see (142) and (149a)), (ii) arises from (155), and the last equality comes from (139).

- For all other steps $h=3,\ldots,H$, observing from the deterministic transition kernels in (149b), it can be easily verified that

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_h^{\star,P}(s,a),\frac{1}{S}\big\}}{d_h^{\mathsf{b},P^\phi}(s,a)} = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_2^{\star,P}(s,a),\frac{1}{S}\big\}}{d_2^{\mathsf{b},P^\phi}(s,a)} \leq 4C. \tag{226}$$

Combining the above cases, we complete the proof by

$$2C \le C^\star_{\mathsf{rob}} = \max_{(h,s,a,P)\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b},P^\phi}(s,a)} \le 4C.$$

B.3.12 Proof of the claim (168)

Recall that by virtue of (162) and (164), we arrive at

$$z_\phi^\star \coloneqq z_\phi^{\pi^\star,\phi} = \underline{p}\pi_1^{\star,\phi}(\phi\,|\,0) + \underline{q}\pi_1^{\star,\phi}(1-\phi\,|\,0) = \underline{p}.$$

Applying (163) yields

$$\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\pi,\sigma,\phi}\right\rangle = V_h^{\star,\sigma,\phi}(0) - V_h^{\pi,\sigma,\phi}(0) = \left(\underline{p} - z_\phi^\pi\right)(H-1) = \left(\underline{p}-\underline{q}\right)(H-1)\left(1-\pi_1(\phi\,|\,0)\right), \tag{227}$$

where the last equality uses the definition (162). Therefore, it boils down to control $\underline{p}-\underline{q}$.

To continue, we define an auxiliary value function vector $\overline{V} \in \mathbb{R}^{S\times 1}$ obeying

$$\overline{V}(0) = H-1 \quad \text{and} \quad \overline{V}(s) = 0, \quad \forall s \in \mathcal{S}\setminus\{0\}. \tag{228}$$

With this in hand, applying Lemma 11 gives

$$\begin{aligned}
&(H-1)\left(\underline{p}-\underline{q}\right)\\
&\overset{\text{(i)}}{=} \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{1,0,\phi})} \mathcal{P}\overline{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{1,0,1-\phi})} \mathcal{P}\overline{V}\\
&= \sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^\phi_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right\} - \sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^\phi_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right\}\\
&\overset{\text{(ii)}}{\ge} \left\{-\lambda^\star\log\left(P^\phi_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right) - \lambda^\star\sigma\right\} - \left\{-\lambda^\star\log\left(P^\phi_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right) - \lambda^\star\sigma\right\}\\
&= -\lambda^\star\left[\log\left(P^\phi_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right) - \log\left(P^\phi_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right)\right],
\end{aligned} \tag{229}$$

where (i) follows from (see the definition of $\underline{p}$ in (160))

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{1,0,\phi})} \mathcal{P}\overline{V} = \underline{P}_1^\phi(0\,|\,0,\phi)\overline{V}(0) = (H-1)\underline{p},$$
$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{1,0,1-\phi})} \mathcal{P}\overline{V} = \underline{P}_1^\phi(0\,|\,0,1-\phi)\overline{V}(0) = (H-1)\underline{q}.$$

Here, (ii) holds by letting

$$\lambda^\star \coloneqq \arg\max_{\lambda\ge 0} f(\lambda) \coloneqq \arg\max_{\lambda\ge 0}\left\{-\lambda\log\left(P^\phi_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right\}. \tag{230}$$

The rest of the proof is then to control (229). We start with the observation that $\lambda^\star > 0$; this is because in view of Lemma 12 (cf. (66)), it suffices to verify that

$$\log(1-q) + \sigma \overset{\text{(i)}}{\le} \log(\alpha+\Delta) + \left(1-\frac{2}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) = -\frac{2}{\beta}\log\left(\frac{1}{\alpha+\Delta}\right) < 0, \tag{231}$$

where (i) holds by (158). We now claim the following bound for $\lambda^\star$ holds, whose proof is postponed to the end:

$$\frac{H}{16\sigma} \le \frac{H-1}{\log\left(\frac{\beta}{\alpha+\Delta}\right)} \le \lambda^\star \le \frac{H-1}{\left(1-\frac{3}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right)}, \tag{232}$$

which immediately implies the following by taking exponential maps given $\lambda^\star > 0$:

$$\frac{\alpha+\Delta}{\beta} \le e^{-(H-1)/\lambda^\star} \le (\alpha+\Delta)^{1-3/\beta}. \tag{233}$$

Moving to the second term of (229), it follows that

$$\begin{aligned}
&\log\left(P^{\phi}_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right) - \log\left(P^{\phi}_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right)\\
&\overset{\text{(i)}}{=} \log\frac{pe^{-(H-1)/\lambda^\star}+(1-p)}{qe^{-(H-1)/\lambda^\star}+(1-q)}\\
&= \log\left(1+\frac{(p-q)\left(e^{-(H-1)/\lambda^\star}-1\right)}{qe^{-(H-1)/\lambda^\star}+(1-q)}\right)\\
&\overset{\text{(ii)}}{<} -\frac{\Delta\left(1-e^{-(H-1)/\lambda^\star}\right)}{qe^{-(H-1)/\lambda^\star}+(1-q)}\\
&\overset{\text{(iii)}}{\le} -\frac{1}{2}\frac{\Delta}{\left(\frac{1}{\alpha+\Delta}\right)^{\frac{3}{\beta}}(1-q)+(1-q)}\\
&\le -\frac{\Delta}{4e^6(1-q)},
\end{aligned} \tag{234}$$

where (i) follows from the definitions in (149) and (228), (ii) holds by $\log(1+x) < x$ for $x \in (-1,\infty)$, (iii) can be verified by (233), $\beta \ge 4$, and (151):

$$1-e^{-(H-1)/\lambda^\star} \ge 1-(\alpha+\Delta)^{1-3/\beta} \ge 1-(\alpha+\Delta)^{1/4} \ge 1-\left(\frac{3}{2H}\right)^{1/4} \ge \frac{1}{2},$$

and the last line uses $\left(\frac{1}{\alpha+\Delta}\right)^{\frac{3}{\beta}} = \left(\frac{1}{\alpha+\Delta}\right)^{6/\log\left(\frac{1}{\alpha+\Delta}\right)} = e^6$ by the definition of β in (152). Plugging (232) and (234) back into (229) and (227), we arrive at

$$\begin{aligned}
\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\pi,\sigma,\phi}\right\rangle &= (H-1)\left(\underline{p}-\underline{q}\right)\left(1-\pi_1(\phi\,|\,0)\right)\\
&\overset{\text{(i)}}{\ge} \frac{H\Delta}{64e^6\sigma(1-q)}\left(1-\pi_1(\phi\,|\,0)\right) = 2\varepsilon\left(1-\pi_1(\phi\,|\,0)\right),
\end{aligned}$$

where (i) holds by (232) and the last equality follows directly from the choice of Δ in (167).

Proof of inequality (232)**.** Applying (65) in Lemma 12 to $\lambda^\star$ in (230) leads to the upper bound in (232):

$$\lambda^\star \leq \frac{H-1}{\sigma} \leq \frac{H-1}{\left(1-\frac{3}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right)}, \tag{235}$$

where the last inequality holds by (158). As a result, we shall focus on showing the lower bounds in (232) in the remainder of the proof.

Recalling the definition of q in (150), we can reparameterize $1-q$ using two positive variables c_q and λ_q (whose choices will be made clearer soon) as follows:

$$1-q = \alpha = c_q e^{-(H-1)/\lambda_q}. \tag{236}$$

Deriving the first derivative of the function of interest $f(\lambda)$ in (230) as follows:

$$\begin{aligned}
\nabla_\lambda f(\lambda) &= \nabla_\lambda\left(-\lambda\log\left(P^\phi_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right)-\lambda\sigma\right)\\
&\overset{\text{(i)}}{=} \nabla_\lambda\left(-\lambda\log\left(qe^{-(H-1)/\lambda}+1-q\right)-\lambda\sigma\right)\\
&= -\sigma-\log\left(qe^{-(H-1)/\lambda}+1-q\right)-\frac{1}{\lambda}\cdot\frac{q(H-1)e^{-(H-1)/\lambda}}{qe^{-(H-1)/\lambda}+1-q},
\end{aligned} \tag{237}$$

where (i) holds by the chosen transition kernels in (149) and the last line arises from basic calculus. To continue, when $\lambda=\lambda_q$, the derivative of the function $f(\lambda)$ can be expressed as

$$\begin{aligned}
\nabla_\lambda f(\lambda)\,|_{\,\lambda=\lambda_q} &= -\sigma-\log\left((1-q)\frac{q}{c_q}+1-q\right)+\frac{(1-q)\frac{q}{c_q}\log\frac{1-q}{c_q}}{(1-q)\frac{q}{c_q}+1-q}\\
&= -\sigma-\log(1-q)-\log\left(1+\frac{q}{c_q}\right)+\frac{\frac{q}{c_q}\log\frac{1-q}{c_q}}{\frac{q}{c_q}+1}\\
&= -\sigma-\log(1-q)\left(1-\frac{q/c_q}{q/c_q+1}\right)-\log\left(1+\frac{q}{c_q}\right)-\frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q}\\
&\overset{\text{(i)}}{=} -\sigma+\log\left(\frac{1}{\alpha+\Delta}\right)\left(1-\frac{q/c_q}{q/c_q+1}\right)-\log\left(1+\frac{q}{c_q}\right)-\frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q}\\
&\overset{\text{(ii)}}{\geq} \log\left(\frac{1}{\alpha+\Delta}\right)\left(\frac{2}{\beta}-\frac{q/c_q}{q/c_q+1}\right)-\log\left(1+\frac{q}{c_q}\right)-\frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q}\\
&\overset{\text{(iii)}}{\geq} \frac{1}{\beta}\log\left(\frac{1}{\alpha+\Delta}\right)-\log(1+\frac{1}{\beta})-1\\
&\geq \frac{1}{\beta}\log\left(\frac{1}{\alpha+\Delta}\right)-2=0,
\end{aligned} \tag{238}$$

$$\tag{239}$$

where (i) holds by (236), (ii) follows from the bound of σ in (158), (iii) arises from letting $c_q=\beta\geq 4$ and noting the fact $1/2\leq q<1$ (see (153)), leading to

$$\frac{1}{2\beta}\leq\frac{q}{c_q}<\frac{1}{\beta},\qquad \frac{q/c_q}{q/c_q+1}\leq\frac{1}{\beta},\qquad \frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q}<1. \tag{240}$$

Finally, the last line holds by $1/\beta \le \frac{1}{4}$ and $\log\left(\frac{1}{\alpha+\Delta}\right) = 2\beta$ (see (152)).

To proceed, note that the function $f(\lambda)$ is concave with respect to λ. Therefore, observing $\nabla_\lambda f(\lambda)\,|_{\,\lambda=\lambda_q} \ge 0$ with $c_q = \beta$, we have $\lambda_q \le \lambda^\star$, which implies (see (236))

$$1 - q = \alpha + \Delta = \beta e^{-(H-1)/\lambda_q} \le \beta e^{-(H-1)/\lambda^\star}. \tag{241}$$

The above assertion directly gives

$$\lambda^\star \ge \frac{H-1}{\log\left(\frac{\beta}{\alpha+\Delta}\right)}.$$

The proof is completed by noticing

$$\frac{H-1}{\log\left(\frac{\beta}{\alpha+\Delta}\right)} = \frac{H-1}{\log\left(\frac{1}{\alpha+\Delta}\right) + \log\beta} \overset{\text{(i)}}{\ge} \frac{H-1}{2\log\left(\frac{1}{\alpha+\Delta}\right)} \ge \frac{H}{16\sigma},$$

where (i) follows from (152), and the last inequality follows from (158) and the fact $\beta \in [4, \infty)$.

Appendix C. Analysis: discounted infinite-horizon RMDPs

C.1 Proof of Lemma 7

We shall first show that the operator $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$ (cf. (46)) is a γ-contraction, which will in turn imply the existence of the unique fixed point of $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$. Before starting, suppose that the entries of $Q_1, Q_2 \in \mathbb{R}^{SA}$ are all bounded in $\left[0, \frac{1}{1-\gamma}\right]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Denote that

$$\forall s \in \mathcal{S}: \quad V_1(s) \coloneqq \max_a Q_1(s,a), \quad V_2(s) \coloneqq \max_a Q_2(s,a). \tag{242}$$

Proof of γ-contraction. We first show that $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$ is a γ-contraction. Towards this, instead of $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$, we begin with a simpler operator $\widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$, defined as follows:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad \widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q)(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V - b\big(s,a\big), \tag{243}$$

which consequently leads to

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q)(s,a) = \max\left\{\widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q)(s,a), 0\right\}. \tag{244}$$

It follows straightforwardly that

$$\left\|\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_1) - \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_2)\right\|_\infty \le \left\|\widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_2)\right\|_\infty, \tag{245}$$

and hence it suffices to establish the γ-contraction of $\widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$. With this in mind, we observe that

$$\left\|\widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_2)\right\|_\infty = \gamma \left\|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V_1 - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V_2\right\|_\infty \overset{\text{(i)}}{\le} \gamma \left\|V_1 - V_2\right\|_\infty$$

$$\overset{\text{(ii)}}{=} \gamma \max_s \left| \max_a Q_1(s,a) - \max_a Q_2(s,a) \right| \\ \leq \gamma \max_{(s,a)} |Q_1(s,a) - Q_2(s,a)| = \gamma \left\| Q_1 - Q_2 \right\|_\infty, \tag{246}$$

where the first equality holds by the definition of $\widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$ (cf. (243)), (i) follows from that the infimum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$ and $\|\mathcal{P}V_1 - \mathcal{P}V_2\|_\infty \leq \|V_1 - V_2\|_\infty$ for all $\mathcal{P} \in \Delta(\mathcal{S})$, (ii) arises from the definitions in (242), and the last inequality is due to the maximum operator is also a 1-contraction w.r.t. $\|\cdot\|_\infty$. Combining the above two inequalities establish the desired statement.

Existence of the unique fixed point. To continue, we shall first claim that there exists at least one fixed point of $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$. This is a standard argument, which we omit for brevity; interested readers are encouraged to refer to, e.g. Li et al. (2022), for details.

To prove the uniqueness of the fixed points of $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$, suppose that there exist two fixed points Q' and Q'' obeying obeying $Q' = \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q')$ and $Q'' = \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q'')$. Moreover, the definition of $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$ directly implies $0 \leq Q', Q'' \leq \frac{1}{1-\gamma}$, since for any $0 \leq Q \leq \frac{1}{1-\gamma}$, it follows that $0 \leq \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q) \leq \frac{1}{1-\gamma}$. By the γ-contraction property, it follows that

$$\left\| Q' - Q'' \right\|_\infty = \left\| \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q') - \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q'') \right\|_\infty \leq \gamma \left\| Q' - Q'' \right\|_\infty. \tag{247}$$

However, (247) can't happen given $\gamma \in \left[\frac{1}{2}, 1\right)$, indicating the uniqueness of the fixed points of $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$.

C.2 Proof of Lemma 8

To begin with, considering any Q, Q' obeying $Q \leq Q'$, and $0 \leq Q, Q' \leq \frac{1}{1-\gamma}$. We observe that the operator $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$ (cf. (46)) has the monotone non-decreasing property, namely,

$$\begin{aligned} \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q)(s,a) &= \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V - b\big(s,a\big),\, 0 \right\} \\ &= \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P} \max_{a'} Q(\cdot, a') - b\big(s,a\big),\, 0 \right\} \\ &\leq \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P} \max_{a'} Q'(\cdot, a') - b\big(s,a\big),\, 0 \right\} = \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q')(s,a), \end{aligned} \tag{248}$$

where the last line uses $Q \leq Q'$. Recalling the fixed point $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$ of $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$, armed with (248) and the initialization $\widehat{Q}_0 = 0$, we arrive at

$$\widehat{Q}_1 = \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\widehat{Q}_0) \leq \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}) = \widehat{Q}^{\star,\sigma}_{\mathsf{pe}},$$

where the inequality follows from $\widehat{Q}_0 = 0 \leq \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$. Implementing the above result recursively gives

$$\forall\, m \geq 0: \quad \widehat{Q}_m \leq \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}.$$

Applying the γ-contraction property in Lemma 7 thus yields that for any $m \geq 0$,

$$\begin{aligned}\|\widehat{Q}_m - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\|_\infty &= \left\|\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\widehat{Q}_{m-1}) - \widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\widehat{Q}^{\star,\sigma}_{\mathsf{pe}})\right\|_\infty \\ &\leq \gamma\|\widehat{Q}_{m-1} - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\|_\infty \leq \cdots \leq \gamma^m\|\widehat{Q}_0 - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\|_\infty = \gamma^m\|\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\|_\infty \leq \frac{\gamma^m}{1-\gamma},\end{aligned}$$

where the last inequality holds by the fact $\|\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\|_\infty \leq \frac{1}{1-\gamma}$ (see Lemma 7).

C.3 Proof of Theorem 9

To begin, we introduce some additional notation that will be useful throughout the analysis. We denote the state-action space covered by the batch dataset $\mathcal{D}$ as

$$\mathcal{C}^{\mathsf{b}} = \left\{(s,a) : d^{\mathsf{b}}(s,a) > 0\right\}. \tag{249}$$

In addition, recalling the definition in (47), we define a similar one based on the true nominal model P^0 as

$$P_{\mathsf{min}}(s,a) \coloneqq \min_{s'}\left\{P^0(s' \,|\, s,a) :\ P^0(s' \,|\, s,a) > 0\right\}, \tag{250}$$

which directly indicates that

$$P^{\star}_{\mathsf{min}} = \min_{s}\ P_{\mathsf{min}}(s,\pi^\star(s)), \qquad P^{\mathsf{b}}_{\mathsf{min}} = \min_{(s,a)\in\mathcal{C}^{\mathsf{b}}}\ P_{\mathsf{min}}(s,a). \tag{251}$$

Next, we denote the set of possible state occupancy distributions associated with the optimal policy $\pi^\star$ in a model within the uncertainty set $P \in \mathcal{U}^\sigma\left(P^0\right)$ as

$$\mathcal{D}^\star \coloneqq \left\{\left[d^{\star,P}(s)\right]_{s\in\mathcal{S}} : P \in \mathcal{U}^\sigma\left(P^0\right)\right\} = \left\{\left[d^{\star,P}\big(s,\pi^\star(s)\big)\right]_{s\in\mathcal{S}} : P \in \mathcal{U}^\sigma\left(P^0\right)\right\}, \tag{252}$$

where the second equality is due to the fact that $\pi^\star$ is chosen to be deterministic.

We are now ready to embark on the proof of Theorem 9. We first introduce a fact that is used throughout the proof; the proof is postponed to Appendix C.3.2:

$$\forall(s,a)\in\mathcal{C}^{\mathsf{b}}: \qquad N(s,a) \geq \frac{Nd^{\mathsf{b}}(s,a)}{12} \geq \frac{c_1\log(NS/\delta)}{12P_{\mathsf{min}}(s,a)} \geq -\frac{\log\frac{2NS}{\delta}}{\log(1-P_{\mathsf{min}}(s,a))} \tag{253}$$

as long as (58) holds.

For notation simplicity, denote the output Q-function and value function from Algorithm 3 as $\widehat{Q} = \widehat{Q}_M$ and $\widehat{V} = \widehat{V}_M$. Invoking Lemma 8 with $M \geq \frac{\log\frac{\sigma N}{1-\gamma}}{\log\frac{1}{\gamma}}$ directly leads to

$$\left\|\widehat{Q} - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty \leq \frac{1}{\sigma N} \tag{254}$$

and therefore

$$\left\|\widehat{V} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty \leq \max_s\left|\max_a\widehat{Q}(s,a) - \max_a\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a)\right| \leq \left\|\widehat{Q} - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty \leq \frac{1}{\sigma N}. \tag{255}$$

The proof of Theorem 9 is separated into several key steps as follows.

Step 1: controlling the uncertainty via leave-one-out analysis. Given access to only a finite number of samples for estimating the nominal transition kernel P^0, we need to efficiently control

$$\left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V} \right|$$

across the robust value iterations, where $\widehat{V}$ is statistically dependent on $\widehat{P}^0_{s,a}$ (since $\widehat{P}^0_{s,a}$ will be reused in the update rule (cf. (51)) for all the iterations). A naive treatment via the standard covering arguments will unfortunately lead to rather loose bounds (Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022). To overcome this challenge, we resort to the leave-one-out analysis—pioneered by Agarwal et al. (2020); Li et al. (2020, 2022) in the context of model-based RL—to decouple the statistical dependency. The results are summarized in the following lemma, with the proof provided in Appendix C.3.1.

Lemma 22 *Instate the assumptions in Theorem 9. Then for all vector $\widetilde{V}$ obeying $\big\|\widetilde{V} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}}\big\|_\infty \le \frac{1}{\sigma N}$ and $\|\widetilde{V}\|_\infty \le \frac{1}{1-\gamma}$, with probability at least $1-\delta$, one has*

$$\begin{aligned}
&\left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widetilde{V} \right| \\
&\le \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta})}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{4}{N\sigma(1-\gamma)}\ ,\ \frac{1}{1-\gamma} \right\}
\end{aligned} \tag{256}$$

for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. In addition, for all $(s,a)\in\mathcal{C}^{\mathsf{b}}$, with probability at least $1-\delta$, one has

$$\frac{P_{\mathsf{min}}(s,a)}{8\log(NS/\delta)} \le \widehat{P}_{\mathsf{min}}(s,a) \le e^2 P_{\mathsf{min}}(s,a). \tag{257}$$

Step 2: establishing the pessimism property. Armed with Lemma 22, we aim to show the key property that

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \qquad \widehat{Q}(s,a) \le Q^{\widehat{\pi},\sigma}(s,a), \qquad \widehat{V}(s) \le V^{\widehat{\pi},\sigma}(s). \tag{258}$$

Similar to the finite-horizon setting, it suffices to focus on verifying the former assertion in (258). Towards this, we first recall that the fixed point $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$ of the pessimistic robust Bellman operator $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$ (cf. (46)) obeys

$$\widehat{Q}^{\star,\sigma}_{\mathsf{pe}} = \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}) = \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b\big(s,a\big),\, 0 \right\}. \tag{259}$$

If $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a) = 0$. Given the initialization $\widehat{Q}_0 = 0$, invoking Lemma 8 gives

$$\widehat{Q}(s,a) = \widehat{Q}_M(s,a) \le \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a) = 0.$$

As a result, $Q^{\widehat{\pi},\sigma}(s,a) \ge 0 = \widehat{Q}(s,a)$ as desired. Therefore, it boils down to examine the case when $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a) > 0$. One has

$$\widehat{Q}(s,a) \overset{\text{(i)}}{\le} \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a) + \frac{1}{\sigma N} = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b\big(s,a\big) + \frac{1}{\sigma N}$$

$$
\begin{aligned}
&\leq r(s,a)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} - b(s,a) + \frac{1}{\sigma N} + \gamma \left| \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} \right| \\
&\overset{\text{(ii)}}{\leq} r(s,a)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} - b(s,a) + \frac{2}{\sigma N} \\
&\leq r(s,a)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V} - b(s,a) + \frac{2}{\sigma N} + \gamma \left| \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V} \right| \\
&\leq r(s,a)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V}, \qquad (260)
\end{aligned}
$$

where (i) follows from (254), (ii) arises from (255) and the basic fact that infimum operator is 1-contraction w.r.t $\|\cdot\|_\infty$, and the last inequality holds by the definition of $b(s,a)$ (cf. (48)) and Lemma 22. Putting the above inequality together with the robust Bellman equation (cf. (37a)) pertaining to $Q^{\widehat{\pi},\sigma}(s,a)$, we arrive at

$$
\begin{aligned}
Q^{\widehat{\pi},\sigma}(s,a) - \widehat{Q}(s,a) &\geq r(s,a)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}V^{\widehat{\pi},\sigma} - \left(r(s,a)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V} \right) \\
&= \gamma \left(\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}V^{\widehat{\pi},\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V} \right) \\
&\overset{\text{(i)}}{=} \gamma \left(\widetilde{P}_{s,a}V^{\widehat{\pi},\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V} \right) \geq \gamma \widetilde{P}_{s,a}\left(V^{\widehat{\pi},\sigma} - \widehat{V} \right),
\end{aligned}
$$

where (i) holds by setting $\widetilde{P}_{s,a} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}V^{\widehat{\pi},\sigma}$. Consequently, one has

$$
\begin{aligned}
\min_{s,a}\left[Q^{\widehat{\pi},\sigma}(s,a) - \widehat{Q}(s,a) \right] &\geq \min_{s,a}\left[\gamma \widetilde{P}_{s,a}\left(V^{\widehat{\pi},\sigma} - \widehat{V} \right) \right] \overset{\text{(i)}}{\geq} \gamma \min_{s}\left[V^{\widehat{\pi},\sigma}(s) - \widehat{V}(s) \right] \\
&= \gamma \min_{s}\left[Q^{\widehat{\pi},\sigma}\big(s,\widehat{\pi}(s)\big) - \widehat{Q}\big(s,\widehat{\pi}(s)\big) \right] \\
&\geq \gamma \min_{s,a}\left[Q^{\widehat{\pi},\sigma}\big(s,a\big) - \widehat{Q}\big(s,a\big) \right], \qquad (261)
\end{aligned}
$$

where (i) follows from $\widetilde{P}_{s,a} \in \Delta(\mathcal{S})$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Noting that $0 \leq \gamma < 1$, we conclude $Q^{\widehat{\pi},\sigma}(s,a) - \widehat{Q}(s,a) \geq 0$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. This establishes the claim (258).

Step 3: bounding $V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho)$. In view of the pessimistic property (cf. (258)), it follows that

$$
V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \leq V^{\star,\sigma}(s) - \widehat{V}(s). \qquad (262)
$$

Towards this, note that

$$
\begin{aligned}
\widehat{V}(s) = \max_a \widehat{Q}(s,a) \geq \widehat{Q}\big(s,\pi^\star(s)\big) &\overset{\text{(i)}}{\geq} \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\big(s,\pi^\star(s)\big) - \frac{1}{\sigma N} \\
&\overset{\text{(ii)}}{\geq} r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b\big(s,\pi^\star(s)\big) - \frac{1}{\sigma N}
\end{aligned}
$$

$$
\begin{aligned}
&\geq r\big(s,\pi^\star(s)\big)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}-b\big(s,\pi^\star(s)\big)-\frac{1}{\sigma N}\\
&\quad-\gamma\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}-\inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}\right|\\
&\overset{\text{(iii)}}{\geq} r\big(s,\pi^\star(s)\big)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}-b\big(s,\pi^\star(s)\big)-\frac{2}{\sigma N}\\
&\geq r\big(s,\pi^\star(s)\big)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}-b\big(s,\pi^\star(s)\big)-\frac{2}{\sigma N}\\
&\quad-\gamma\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}-\inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}\right|\\
&\geq r\big(s,\pi^\star(s)\big)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}-2b\big(s,\pi^\star(s)\big),
\end{aligned}
\tag{263}
$$

where (i) follows from (254), (ii) holds by applying (259), (iii) arises from (255), and the basic fact that the infimum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$, and the final inequality holds by the definition of $b(s,a)$ (see (48)) and Lemma 22.

To continue, invoking the robust Bellman optimality equation in (37b) gives

$$
V^{\star,\sigma}(s)=Q^{\star,\sigma}\big(s,\pi^\star(s)\big)=r\big(s,\pi^\star(s)\big)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}V^{\star,\sigma}.
$$

Combining the above relation with (263), we arrive at

$$
\begin{aligned}
V^{\star,\sigma}(s)-\widehat{V}(s)&\leq \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}V^{\star,\sigma}-\gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}+2b\big(s,\pi^\star(s)\big)\\
&\leq \gamma \widehat{P}^{\inf}_{s,\pi^\star(s)}\left(V^{\star,\sigma}-\widehat{V}\right)+2b\big(s,\pi^\star(s)\big),
\end{aligned}
\tag{264}
$$

where the final inequality holds evidently, by introducing

$$
\widehat{P}^{\inf}_{s,\pi^\star(s)}:=\operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}.
\tag{265}
$$

Before continuing, for convenience, let us introduce a matrix $\widehat{P}^{\inf}\in\mathbb{R}^{S\times\mathcal{S}}$ and a vector $b^\star\in\mathbb{R}^{\mathcal{S}}$, where their s-th rows (resp. entries) are defined as

$$
\left[\widehat{P}^{\inf}\right]_{s,\cdot}=\widehat{P}^{\inf}_{s,\pi^\star(s)},\qquad\text{and}\qquad b^\star(s)=b\big(s,\pi^\star(s)\big).
\tag{266}
$$

With these notation in hand, averaging (264) over the initial state distribution ρ leads to

$$
\begin{aligned}
V^{\star,\sigma}(\rho)-\widehat{V}(\rho)&=\sum_{s\in\mathcal{S}}\rho(s)\left(V^{\star,\sigma}(s)-\widehat{V}(s)\right)\\
&\leq \gamma\sum_{s\in\mathcal{S}}\rho(s)\widehat{P}^{\inf}_{s,\pi^\star(s)}\left(V^{\star,\sigma}-\widehat{V}\right)+2\sum_{s\in\mathcal{S}}\rho(s)b\big(s,\pi^\star(s)\big)
\end{aligned}
$$

$$= \gamma \rho^\top \widehat{P}^{\inf}\left(V^{\star,\sigma} - \widehat{V}\right) + 2\rho^\top b^\star. \tag{267}$$

Applying the above result recursively gives

$$\begin{aligned} V^{\star,\sigma}(\rho) - \widehat{V}(\rho) &\le \gamma \rho^\top \widehat{P}^{\inf}\left(V^{\star,\sigma} - \widehat{V}\right) + 2\rho^\top b^\star \\ &\le \gamma\left(\gamma \rho^\top \widehat{P}^{\inf}\right)\widehat{P}^{\inf}\left(V^{\star,\sigma} - \widehat{V}\right) + 2\left(\gamma \rho^\top \widehat{P}^{\inf}\right) b^\star + 2\rho^\top b^\star \\ &\le \cdots \le \left\{\lim_{i\to\infty} \gamma^i \rho^\top \left(\widehat{P}^{\inf}\right)^i \left(V^{\star,\sigma} - \widehat{V}\right)\right\} + 2\rho^\top \sum_{i=0}^{\infty} \gamma^i \left(\widehat{P}^{\inf}\right)^i b^\star \\ &\overset{\text{(i)}}{\le} 2\rho^\top \sum_{i=0}^{\infty} \gamma^i \left(\widehat{P}^{\inf}\right)^i b^\star = 2\rho^\top \left(I - \gamma \widehat{P}^{\inf}\right)^{-1} b^\star, \end{aligned} \tag{268}$$

where (i) holds by $\left|\rho^\top \left(\widehat{P}^{\inf}\right)^i \left(V^{\star,\sigma} - \widehat{V}\right)\right| \le \frac{1}{1-\gamma}$ for all $i \ge 0$, and that $\lim_{i\to\infty} \gamma^i \rho^\top \left(\widehat{P}^{\inf}\right)^i \left(V^{\star,\sigma} - \widehat{V}\right) = 0$ since $\lim_{i\to\infty} \gamma^i = 0$ for all $0 \le \gamma < 1$.

To further characterize the above performance gap, invoking the definition of $d^{\star,P}$ (cf. (38) and (39a)), we arrive at

$$\left(d^{\star,\widehat{P}^{\inf}}\right)^\top = (1-\gamma)\rho^\top \sum_{t=0}^{\infty} \gamma^t \left(\widehat{P}^{\inf}\right)^t = (1-\gamma)\rho^\top \left(I - \gamma \widehat{P}^{\inf}\right)^{-1}. \tag{269}$$

Plugging the above expression back into (268), and combining with(262), yields

$$V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho) \le V^{\star,\sigma}(\rho) - \widehat{V}(\rho) \le \frac{2}{1-\gamma}\left\langle d^{\star,\widehat{P}^{\inf}}, b^\star \right\rangle. \tag{270}$$

Step 4: controlling $\left\langle d^{\star,\widehat{P}^{\inf}}, b^\star \right\rangle$ using concentrability. Note that $\widehat{P}^{\inf} \in \mathcal{U}^\sigma(P^0)$ (see (265) and (266)), which in words means $\widehat{P}^{\inf}$ is some transition kernel inside $\mathcal{U}^\sigma(P^0)$ — the uncertainty set around the nominal kernel P^0. Similar to the finite-horizon case, observing that we can express $\left\langle d^{\star,\widehat{P}^{\inf}}, b^\star \right\rangle = \sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\inf}}(s) b^\star(s)$, we divide the states into two cases and control them separately.

- **Case 1: $s \in \mathcal{S}$ where $\max_{P\in\mathcal{U}^\sigma\left(P^0\right)} d^{\star,P}\left(s,\pi^\star(s)\right) = 0$.** Since $\widehat{P}^{\inf} \in \mathcal{U}^\sigma(P^0)$, one has

$$0 \le d^{\star,\widehat{P}^{\inf}}(s) = d^{\star,\widehat{P}^{\inf}}\left(s,\pi^\star(s)\right) \le \max_{P\in\mathcal{U}^\sigma\left(P^0\right)} d^{\star,P}\left(s,\pi^\star(s)\right) = 0,$$

  which consequently indicates

$$d^{\star,\widehat{P}^{\inf}}(s) = 0. \tag{271}$$

- **Case 2: $s \in \mathcal{S}$ where $\max_{P\in\mathcal{U}^\sigma\left(P^0\right)} d^{\star,P}\left(s,\pi^\star(s)\right) > 0$.** For any such state s, we claim that

$$d^{\mathsf{b}}\left(s,\pi^\star(s)\right) > 0 \quad \text{and} \quad \left(s,\pi^\star(s)\right) \in \mathcal{C}^{\mathsf{b}}. \tag{272}$$

This is due to Assumption 2, which requires $C^\star_{\mathsf{rob}}$ to be finite given the numerator is positive:

$$\max_{P\in\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d^{\star,P}\big(s,\pi^\star(s)\big),\frac{1}{S}\right\}}{d^{\mathsf{b}}\big(s,\pi^\star(s)\big)} = \max_{P\in\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d^{\star,P}(s),\frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)} \le C^\star_{\mathsf{rob}} < \infty. \quad (273)$$

To continue, invoking the fact in (253) with $\big(s,\pi^\star(s)\big)\in\mathcal{C}^{\mathsf{b}}$ gives

$$\begin{aligned} N\big(s,\pi^\star(s)\big) &\ge \frac{Nd^{\mathsf{b}}\big(s,\pi^\star(s)\big)}{12} \\ &\overset{\text{(i)}}{\ge} \frac{N\max_{P\in\mathcal{U}^\sigma(P^0)}\min\left\{d^{\star,P}\big(s,\pi^\star(s)\big),\frac{1}{S}\right\}}{12C^\star_{\mathsf{rob}}} \ge \frac{N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}{12C^\star_{\mathsf{rob}}}, \end{aligned} \quad (274)$$

where (i) holds by Assumption 2, and the last inequality holds by $\widehat{P}^{\mathrm{inf}}\in\mathcal{U}^\sigma(P^0)$. With this in mind, we can control the pessimistic penalty $b^\star(s)$ (cf. (48)) by

$$\begin{aligned} b^\star(s) &\le \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}\big(s,\pi^\star(s)\big)N\big(s,\pi^\star(s)\big)}} + \frac{4}{\sigma N(1-\gamma)} + \frac{2}{\sigma N} \\ &\overset{\text{(i)}}{\le} \frac{4c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^\star(s)\big)N\big(s,\pi^\star(s)\big)}} + \frac{4}{\sigma N(1-\gamma)} + \frac{2}{\sigma N} \\ &\le \frac{16c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{C^\star_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^\star(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}} + \frac{6}{\sigma N(1-\gamma)} \\ &\le \frac{20c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{C^\star_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^\star(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}}, \end{aligned}$$

where (i) arises from (257), the penultimate inequality follows from (274), and the last inequality holds as long as c_{b} is large enough.

Summing up the above two cases, we arrive at

$$\begin{aligned} \left\langle d^{\star,\widehat{P}^{\mathrm{inf}}}, b^\star\right\rangle &= \sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)b^\star(s) \\ &\le \sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)\frac{20c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{C^\star_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^\star(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}} \\ &\overset{\text{(i)}}{\le} \frac{20c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)\frac{C^\star_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^\star(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}}\sqrt{\sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)} \end{aligned}$$

$$\leq \frac{40c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{SC^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{\star}_{\mathsf{min}}N}}, \tag{275}$$

where (i) arises from Cauchy-Schwarz inequality, and the last inequality holds since $P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big) \geq P^{\star}_{\mathsf{min}}$ for all $s \in \mathcal{S}$ (see (251)) and the following fact (which has been established in (98)):

$$\sum_{s\in\mathcal{S}} \frac{d^{\star,\widehat{P}^{\mathrm{inf}}}(s)}{\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s), \frac{1}{S}\right\}} \leq 2S.$$

Finally, inserting (275) back into (270), with probability at least $1-2\delta$, one has

$$V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho) \leq \frac{2}{1-\gamma}\left\langle d^{\star,\widehat{P}^{\mathrm{inf}}}, b^{\star}\right\rangle \leq \frac{80c_{\mathrm{b}}}{\sigma(1-\gamma)^2}\sqrt{\frac{SC^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{\star}_{\mathsf{min}}N}},$$

which concludes the proof.

C.3.1 Proof of Lemma 22

We first note that the second assertion in (257) is the counterpart of (80), which can be verified following the same argument in Appendix B.2.1. For brevity, we omit its proof, and shall focus on verifying (256).

To begin with, we consider the situation when $N(s,a)=0$. In this case, (256) can be easily verified since

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}V\right| \overset{\text{(i)}}{=} \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}V \leq \|V\|_{\infty} \overset{\text{(ii)}}{\leq} \frac{1}{1-\gamma}, \tag{276}$$

where (i) follows from the fact $\widehat{P}^0_{s,a}=0$ when $N(s,a)=0$ (see (44)), and (ii) arises from the assumption $\|V\|_{\infty}\leq\frac{1}{1-\gamma}$. Consequently, in the remainder of the proof, we focus on verifying (256) when $N(s,a)>0$. Let us first introduce the counterpart of the claim (79) in Lemma 17 as follows.

Lemma 23 *For all $(s,a)\in\mathcal{S}\times\mathcal{A}$ with $N(s,a)>0$, consider any vector $V\in\mathbb{R}^S$ independent of $\widehat{P}^0_{s,a}$ obeying $\|V\|_{\infty}\leq\frac{1}{1-\gamma}$. With probability at least $1-\delta$, one has*

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}V\right| \leq \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log(\frac{NS}{\delta})}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}}. \tag{277}$$

Proof The proof follows from the same arguments in Appendix B.2.2, with small modifications to adapt to the infinite-horizon setting; we omit the details for conciseness.■

Armed with the above point-wise concentration bound, we are now ready to derive the uniform concentration bound desired as in Lemma 22, counting on a leave-one-out argument divided into the following steps. The crux of the analysis is to construct a set of auxiliary RMDPs, each different from the empirical RMDP only at a single state but possessing crucial statistical independence that facilitates the concentration arguments, which can then be transferred back to the empirical RMDP via a simple triangle inequality.

Step 1: construction of auxiliary RMDPs with state-absorbing empirical nominal transitions. Denote the empirical infinite-horizon robust MDP with the nominal transition kernel $\widehat{P}^0$ as $\widehat{\mathcal{M}}_{\mathsf{rob}}$. Then, for each state s and each scalar $u \geq 0$, we can construct an auxiliary robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ so that it is the same as $\widehat{\mathcal{M}}_{\mathsf{rob}}$ except the properties in state s. To be precise, let the nominal transition kernel and reward function of $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ be $P^{s,u}$ and $r^{s,u}$, which are given respectively as

$$\begin{cases} P^{s,u}(s' \,|\, s,a) = \mathbb{1}(s'=s) & \text{for all } (s',a) \in \mathcal{S}\times\mathcal{A}, \\ P^{s,u}(\cdot \,|\, \widetilde{s},a) = \widehat{P}^0(\cdot \,|\, \widetilde{s},a) & \text{for all } (\widetilde{s},a) \in \mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s}\neq s, \end{cases} \tag{278}$$

and

$$\begin{cases} r^{s,u}(s,a) = u & \text{for all } a \in \mathcal{A}, \\ r^{s,u}(\widetilde{s},a) = r(\widetilde{s},a) & \text{for all } (\widetilde{s},a) \in \mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s}\neq s. \end{cases} \tag{279}$$

Clearly, state s of the auxiliary $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ is absorbing, meaning that the state stays at s once entering it. This removes the randomness of $\widehat{P}^0_{s,a}$ for all $a \in \mathcal{A}$ in state s, a key property we will leverage later.

With the robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ in hand, we still need to complete the design by defining the corresponding penalty term for all $(\widetilde{s},a) \in \mathcal{S}\times\mathcal{A}$, which is given as follows

$$b^{s,u}(\widetilde{s},a) \coloneqq \begin{cases} \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{s,u}_{\mathsf{min}}(s,a)N(\widetilde{s},a)}} + \frac{4}{N\sigma(1-\gamma)}\ ,\ \frac{1}{1-\gamma}\right\} + \frac{2}{\sigma N} & \text{if } N(\widetilde{s},a)>0, \\ \frac{1}{1-\gamma}+\frac{2}{\sigma N} & \text{otherwise,} \end{cases} \tag{280}$$

where $P^{s,u}_{\mathsf{min}}(\widetilde{s},a)$ is defined as the smallest positive state transition probability over the nominal kernel $P^{s,u}(\cdot \,|\, \widetilde{s},a)$:

$$\forall(\widetilde{s},a)\in\mathcal{S}\times\mathcal{A}: \quad P^{s,u}_{\mathsf{min}}(\widetilde{s},a) \coloneqq \min_{s'}\Big\{P^{s,u}(s' \,|\, \widetilde{s},a):\ P^{s,u}(s' \,|\, \widetilde{s},a)>0\Big\}. \tag{281}$$

In view of (278) and (47), it holds that $P^{s,u}_{\mathsf{min}}(\widetilde{s},a) = \widehat{P}_{\mathsf{min}}(\widetilde{s},a)$, and therefore $b^{s,u}(\widetilde{s},a) = b(\widetilde{s},a)$, when $\widetilde{s}\neq s$ for any $u\geq 0$. Armed with the above definitions, the pessimistic robust Bellman operator $\widehat{\mathcal{T}}^{\sigma}_{s,u}(Q)(\cdot)$ of the RMDP $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ is defined as

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \quad \widehat{\mathcal{T}}^{\sigma}_{s,u}(Q)(s,a) = \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{s,u}_{s,a})} \mathcal{P}V - b^{s,u}(s,a),\, 0\right\}. \tag{282}$$

Step 2: fixed-point equivalence between $\widehat{\mathcal{M}}_{\mathsf{rob}}$ and the auxiliary RMDP $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$. Recall that $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$ is the unique fixed point of $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ with the corresponding value $\widehat{V}^{\star,\sigma}_{\mathsf{pe}}$. We claim that there exists some choice of u such that the fixed point of $\widehat{\mathcal{T}}^{\sigma}_{s,u}(Q)(\cdot)$ coincides with that of $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$. In particular, given a state s, we show the following choice of u suffices:

$$u^{\star} \coloneqq (1-\gamma)\widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) + \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{s,u}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{4}{N\sigma(1-\gamma)}\ ,\ \frac{1}{1-\gamma}\right\} + \frac{2}{\sigma N}. \tag{283}$$

Towards this, we shall break our arguments in two different cases.

- **For state $s' \neq s$.** In this case, for any $a \in \mathcal{A}$, it can be verified that

$$\begin{aligned}
&\max\left\{r^{s,u^\star}(s',a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^{s,u^\star}_{s',a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b^{s,u^\star}(s',a),\, 0\right\} \\
&= \max\left\{r(s',a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s',a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b(s',a),\, 0\right\} \\
&= \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\widehat{Q}^{\star,\sigma}_{\mathsf{pe}})(s',a) = \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s',a),
\end{aligned} \tag{284}$$

  where the second line follows from the definitions in (279) and (278) as well as $b^{s,u^\star}(s',a) = b(s',a)$ when $s' \neq s$, the last line arises from the definition of the pessimistic Bellman operator (46), and that $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$ is the fixed point.

- **For state s.** In this case, for any u and $a \in \mathcal{A}$, observing that $P^{s,u}(s' \,|\, s,a)$ has only one positive entry equal to 1 (cf. (278)), applying (281) yields

$$P^{s,u}_{\mathsf{min}}(s,a) = 1. \tag{285}$$

  Plugging the above fact into (280) leads to

$$b^{s,u}(s,a) = \begin{cases} \min\left\{\frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{N(s,a)}} + \frac{4}{N\sigma(1-\gamma)}\ ,\ \frac{1}{1-\gamma}\right\} + \frac{2}{\sigma N} & \text{if } N(s,a) > 0, \\ \frac{1}{1-\gamma} & \text{otherwise} \end{cases} \tag{286}$$

  for all $a \in \mathcal{A}$. As a result, we have for any $a \in \mathcal{A}$:

$$\begin{aligned}
&\max\left\{r^{s,u^\star}(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^{s,u^\star}_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b^{s,u^\star}(s,a),\, 0\right\} \\
&= \max\left\{u^\star + \gamma\widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) - b^{s,u^\star}(s,a),\, 0\right\} \\
&= \max\left\{(1-\gamma)\widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) + \gamma\widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s),\, 0\right\} = \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s),
\end{aligned} \tag{287}$$

  where the second line follows from the fact that $P^{s,u^\star}_{s,a}$ is a singleton distribution at state s, and hence $\mathcal{U}^\sigma(P^{s,u^\star}_{s,a}) = P^{s,u^\star}_{s,a}$ by the definition of the KL uncertainty set, and the second line follows from plugging in the definition of $u^\star$ in (283) and $b^{s,u^\star}(s,a)$ in (286).

Summing up the above two cases, we establish that there exists a fixed point $\widehat{Q}^{\star,\sigma}_{s,u^\star}$ of the operator $\widehat{\mathcal{T}}^\sigma_{s,u^\star}(\cdot)$ if we let

$$\begin{cases} \widehat{Q}^{\star,\sigma}_{s,u^\star}(s,a) = \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) & \text{for all } a \in \mathcal{A}, \\ \widehat{Q}^{\star,\sigma}_{s,u^\star}(s',a) = \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s',a) & \text{for all } s' \neq s \text{ and } a \in \mathcal{A}. \end{cases} \tag{288}$$

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 $\widehat{V}^{\star,\sigma}_{\mathsf{pe}}$.

Step 3: building an ε-net for all reward values u. It is easily verified that the reward $u^\star$ obeys

$$u^\star \le 1 + \min\left\{\frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{s,u}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{4}{\sigma N(1-\gamma)}\ ,\ \frac{1}{1-\gamma}\right\} + \frac{2}{\sigma N} \le \frac{2}{\sigma} + \frac{2}{1-\gamma}. \tag{289}$$

As a result, we construct an ε-net (Vershynin, 2018) of the line segment within the range $\left[0, \frac{2}{\sigma} + \frac{2}{1-\gamma}\right]$ with $\varepsilon = \frac{1}{\sigma N}$ as follows:

$$\mathcal{U}_\varepsilon \coloneqq \left\{\frac{i}{\sigma N} \,|\, 1 \le i \le \left\lfloor \sigma N\left(\frac{2}{\sigma} + \frac{2}{1-\gamma}\right)\right\rfloor\right\}. \tag{290}$$

Armed with this covering net $\mathcal{U}_\varepsilon$, we can construct an auxiliary robust MDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ and its corresponding pessimistic robust Bellman operator for each $u \in \mathcal{U}_\varepsilon$ (see Step 1). Following the same arguments in the proof of Lemma 7 (cf. Appendix C.1), for each $u \in \mathcal{U}_\varepsilon$, 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 $0 \le \widehat{Q}^{\star,\sigma}_{s,u} \le \frac{1}{1-\gamma}\cdot 1$. In turn, the corresponding value function also satisfies $\|\widehat{V}^{\star,\sigma}_{s,u}\|_\infty \le \frac{1}{1-\gamma}$.

In view of the definitions in (278) and (279), for all $u \in \mathcal{U}_\varepsilon$, $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is statistically independent from $\widehat{P}^0_{s,a}$, which indicates the independence between $\widehat{V}^{\star,\sigma}_{s,u}$ and $\widehat{P}^0_{s,a}$. This makes it possible to invoke Lemma 23, and taking the union bound over all samples N and $u \in \mathcal{U}_\varepsilon$ give that, with probability at least $1-\delta$,

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u}\right| \le \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} \tag{291}$$

hold simultaneously for all $(s,a,u) \in \mathcal{S}\times\mathcal{A}\times\mathcal{U}_\varepsilon$ with $N(s,a) > 0$.

Step 4: a covering argument. Recalling that $u^\star \in \left[0, \frac{2}{\sigma} + \frac{2}{1-\gamma}\right]$ (see (289)), we can always find some $\widetilde{u} \in \mathcal{U}_\varepsilon$ such that $|\widetilde{u} - u^\star| \le \frac{1}{\sigma N}$. Consequently, plugging in the operator in (282) yields

$$\forall Q \in \mathbb{R}^{SA}: \quad \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(Q) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(Q)\right\|_\infty \overset{\text{(i)}}{\le} |\widetilde{u} - u^\star| \le \frac{1}{\sigma N}, \tag{292}$$

where (i) holds by $b^{s,\widetilde{u}}(s,a) = b^{s,u^\star}(s,a)$ for s (see (286)) and $b^{s,\widetilde{u}}(s',a) = b^{s,u^\star}(s',a) = b(s',a)$ for all $s' \ne s$.

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

$$\begin{aligned}
\left\|\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty &= \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}}) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty \\
&\le \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}}) - \widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty + \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star}) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty \\
&\le \gamma\left\|\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty + \frac{1}{\sigma N},
\end{aligned} \tag{293}$$

which directly indicates that

$$\left\|\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \frac{1}{(1-\gamma)\sigma N} \tag{294}$$

and

$$\left\|\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{V}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \left\|\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \frac{1}{(1-\gamma)\sigma N}. \tag{295}$$

Armed with the above facts, invoking the identity $\widehat{V}^{\star,\sigma}_{\mathsf{pe}} = \widehat{V}^{\star,\sigma}_{s,u^\star}$ established in Step 2 gives

$$\begin{aligned}
&\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right| = \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star}\right| \\
&\overset{\text{(i)}}{\le} \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}}\right| \\
&\qquad + \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star}\right| + \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star}\right| \\
&\overset{\text{(ii)}}{\le} \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}}\right| + \frac{2}{N\sigma(1-\gamma)} \\
&\le \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{2}{N\sigma(1-\gamma)},
\end{aligned} \tag{296}$$

where (i) holds by applying the triangle inequality, (ii) arises from (295) and the basic fact that infimum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$, and the final inequality follows from (291).

Step 5: finishing up. Now we are positioned to finish up the proof. For all vector $\widetilde{V}$ obeying $\left\|\widetilde{V} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty \le \frac{1}{\sigma N}$ and $\|\widetilde{V}\|_\infty \le \frac{1}{1-\gamma}$, we apply the triangle inequality and invoke (296) to reach

$$\begin{aligned}
&\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widetilde{V}\right| \le \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right| \\
&\qquad + \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right| + \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right| \\
&\le \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{4}{N\sigma(1-\gamma)}.
\end{aligned} \tag{297}$$

Finally, we complete the proof by verifying that

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widetilde{V}\right| \le \left\|\widetilde{V}\right\|_\infty \le \frac{1}{1-\gamma}. \tag{298}$$

C.3.2 Proof of (253)

For all $(s,a) \in \mathcal{C}^{\mathsf{b}}$, one has

$$N d^{\mathsf{b}}(s,a) \overset{\text{(i)}}{\geq} \frac{c_1 d^{\mathsf{b}}(s,a)\log(NS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}} P^{\mathsf{b}}_{\mathsf{min}}} \overset{\text{(ii)}}{\geq} \frac{c_1 \log(NS/\delta)}{P^{\mathsf{b}}_{\mathsf{min}}} \overset{\text{(iii)}}{\geq} \frac{c_1 \log(NS/\delta)}{P_{\mathsf{min}}(s,a)}, \tag{299}$$

where (i) follows from the condition (58), (ii) arises from the definition that $d^{\mathsf{b}}_{\mathsf{min}} \leq d^{\mathsf{b}}(s,a)$ for all $(s,a) \in \mathcal{C}^{\mathsf{b}}$, and (iii) follows from the definition in (251). In particular, when c_1 is large enough, one has $\frac{2}{3}\log\frac{NS}{\delta} < \frac{N d^{\mathsf{b}}(s,a)}{12}$. To continue, we recall a key property of $N(s,a)$ (cf. (43)) in the following lemma.

Lemma 24 ((Li et al., 2022, Lemma 7)) *Fix $\delta \in (0,1)$. With probability at least $1-\delta$, the quantities $\{N(s,a)\}$ in (43) obey*

$$\max\left\{N(s,a),\ \frac{2}{3}\log\frac{NS}{\delta}\right\} \geq \frac{N d^{\mathsf{b}}(s,a)}{12} \tag{300}$$

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

Consequently, Lemma 24 tells us that with probability at least $1-\delta$,

$$N(s,a) \geq \frac{N d^{\mathsf{b}}(s,a)}{12} \geq \frac{c_1 \log(NS/\delta)}{12 P_{\mathsf{min}}(s,a)} \tag{301}$$

as long as c_1 is large enough. Last but not least, taking the basic fact $x \leq -\log(1-x)$ for all $x \in [0,1]$, the last inequality of (253) can be verified by

$$\frac{c_1 \log(NS/\delta)}{12 P_{\mathsf{min}}(s,a)} \geq -\frac{\log\frac{2NS}{\delta}}{\log(1-P_{\mathsf{min}}(s,a))}. \tag{302}$$

C.4 Proof of Theorem 10

Similar to the finite-horizon case, we shall develop the lower bounds for the two cases when the uncertainty levels σ vary separately.

C.4.1 Construction of hard problem instances: small uncertainty level

We first construct some hard discounted infinite-horizon RMDP instances and then characterize the sample complexity requirements over these instances.

Construction of a collection of hard MDPs. Suppose there are two MDPs

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

Here, $\mathcal{S} = \{0,1,\ldots,S-1\}$, and $\mathcal{A} = \{0,1\}$. The transition kernel P^θ of the MDP $\mathcal{M}^\theta$ is specified as follows:

$$P^\theta(s' \mid s,a) = \begin{cases} p\mathbb{1}(s'=0) + (1-p)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,\theta) \\ q\mathbb{1}(s'=0) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,1-\theta) \\ q\mathbb{1}(s'=s) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad s>0 \end{cases}, \tag{303}$$

for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}\times$, where p and q are set to be

$$p = 1 - c_1(1-\gamma), \qquad q = 1 - c_1(1-\gamma) - c_2(1-\gamma)^2\varepsilon, \tag{304}$$

which satisfies

$$\frac{2}{3} \le \gamma < 1 \qquad \text{and} \qquad c_2(1-\gamma)^2\varepsilon \le \frac{c_1(1-\gamma)}{2} \le \frac{1}{8}. \tag{305}$$

for some $c_1 \le \frac{1}{4}$ and small enough c_2. In view of the assumptions (305), one has

$$p > q \ge \frac{1}{2}. \tag{306}$$

Finally, we define the reward function as

$$r(s,a) = \begin{cases} 1 & \text{if } s = 0 \text{ or } s = 2 \\ 0 & \text{otherwise} \end{cases}. \tag{307}$$

Construction of the history/batch dataset. Define a useful state distribution (only supported on the state subset $\{0, 1, 2\}$) as

$$\mu(s) = \frac{1}{CS}\mathbb{1}(s = 0) + \Big(1 - \frac{1}{CS}\Big)\mathbb{1}(s = 1), \tag{308}$$

where $C > 0$ is some constant that determines the robust concentrability coefficient $C^\star_{\mathsf{rob}}$ (which will be made clear soon) and obeys

$$\frac{1}{CS} \le \frac{1}{4}. \tag{309}$$

A batch dataset—consists of N i.i.d samples $\{(s_i, a_i, s_i')\}_{1\le i\le N}$—is generated over the nominal environment $\mathcal{M}_\theta$ according to (40), with the behavior distribution chosen to be:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad d^{\mathsf{b}}(s,a) = \frac{\mu(s)}{2}. \tag{310}$$

Additionally, we choose the following initial state distribution:

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{otherwise} \end{cases}. \tag{311}$$

Uncertainty set of the transition kernels. We next describe the radius σ of the uncertainty set in our construction of the robust MDPs, along with some useful properties, which are similar to the finite-horizon case. The perturbed transition kernels in $\mathcal{M}_\theta$ is limited to the following uncertainty set

$$\mathcal{U}^\sigma(P^\theta) \coloneqq \otimes\, \mathcal{U}^\sigma\left(P^\theta_{s,a}\right), \qquad \mathcal{U}^\sigma(P^\theta_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{s,a} \,\|\, P^\theta_{s,a}\right) \le \sigma\right\}, \tag{312}$$

where $P^\theta_{s,a} \coloneqq P^\theta(\cdot \,|\, s, a) \in [0,1]^{1\times S}$. Moreover, the radius of the uncertainty set σ obeys

$$0 \le \sigma \le \frac{1-\gamma}{20}. \tag{313}$$

For any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we denote the infimum entry of the perturbed transition kernel $P_{s,a} \in \mathcal{U}^\sigma(P^\theta_{s,a})$ moving to the next state s' as

$$\underline{P}^\theta(s' \,|\, s, a) \coloneqq \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^\theta_{s,a})} P(s' \,|\, s, a). \tag{314}$$

As shall be seen, the transition from state 0 to state 2 plays an important role in the analysis, for convenience, we denote

$$\underline{p}_\star \coloneqq \underline{P}^\theta(0 \,|\, 0, \theta), \qquad \underline{q}_\star \coloneqq \underline{P}^\theta(0 \,|\, 0, 1-\theta). \tag{315}$$

With these definitions in place, we summarize some useful properties of the uncertainty set in the following lemma, which parallels Lemma 18 in the finite-horizon case.

Lemma 25 *Suppose the uncertainty level σ satisfies* (313)*. The perturbed transition kernels obey*

$$\underline{p}_\star \geq \underline{q}_\star \geq 1 - c_3(1-\gamma) \quad \text{and} \quad \underline{p}_\star - \underline{q}_\star \geq p - q \geq 0 \tag{316}$$

for constant $c_3 = 2c_1 \leq \frac{1}{4}$.

Proof The proof follows from the same arguments as the proof for Lemma 18 in Appendix B.3.5 by replacing H with $\frac{1}{1-\gamma}$; we omit the details for brevity. ■

Value functions and optimal policies. Now we are positioned to derive the corresponding robust value functions and identify the optimal policies. For any MDP $\mathcal{M}_\theta$ with the above uncertainty set, denote $\pi^\star_\theta$ as the optimal policy. In addition, we denote the robust value function of any policy π (resp. the optimal policy $\pi^\star_\theta$) as $V^{\pi,\sigma}_\theta$ (resp. $V^{\star,\sigma}_\theta$). Then, we introduce the following lemma which describes some important properties of the robust value functions and optimal policies.

Lemma 26 *For any $\theta = \{0, 1\}$ and any policy π, one has*

$$V^{\pi,\sigma}_\theta(0) = \frac{1}{1 - \gamma x^\pi_\theta}, \tag{317}$$

where x^π_θ is defined as

$$x^\pi_\theta \coloneqq \underline{p}_\star \pi(\theta \,|\, 0) + \underline{q}_\star \pi(1-\theta \,|\, 0). \tag{318}$$

In addition, the optimal value functions and the optimal policies obey

$$V^{\star,\sigma}_\theta(0) = \frac{1}{1 - \gamma \underline{p}_\star}, \quad V^{\star,\sigma}_\theta(s) = 0 \quad \text{for } s = 1 \text{ or } s > 2, \tag{319a}$$

$$\pi^\star_\theta(\theta \,|\, s) = 1, \qquad \text{for } s \in \mathcal{S}. \tag{319b}$$

Moreover, the robust single-policy clipped concentrability coefficient $C^\star_{\mathsf{rob}}$ obeys

$$C^\star_{\mathsf{rob}} = 2C. \tag{320}$$

Proof See Appendix C.4.5. ■

C.4.2 Establishing the minimax lower bound: small uncertainty level

Towards this, we first introduce the following lemma, which parallels the claim in (167)-(168) in the finite-horizon case.

Lemma 27 *For any policy $\widehat{\pi}$,*

$$V_\theta^{\star,\sigma}(0) - V_\theta^{\widehat{\pi},\sigma}(0) \geq 2\varepsilon\big(1 - \widehat{\pi}(\theta\,|\,0)\big).$$

Proof This lemma can be directly verified by controlling $V_\theta^{\star,\sigma}(0) - V_\theta^{\widehat{\pi},\sigma}(0)$ with the help of Lemma 25 and Lemma 26; we omit the details for brevity. ■

Armed with this lemma, following the same arguments in Appendix B.3.4, we can complete the proof by observing that: let c_1 be some sufficient large constant, as long as the sample size is beneath

$$N \leq \frac{c_4 S C_{\mathsf{rob}}^\star}{(1-\gamma)^3\varepsilon^2}, \tag{321}$$

then we necessarily have

$$\inf_{\widehat{\pi}} \max_{\theta\in\{0,1\}} \mathbb{P}_\theta\left\{V_\theta^{\star,\sigma}(\rho) - V_\theta^{\widehat{\pi},\sigma}(\rho) \geq \varepsilon\right\} \geq \frac{1}{8}, \tag{322}$$

where $\mathbb{P}_\theta$ denote the probability conditioned on that the MDP is $\mathcal{M}_\theta$. We omit the details for brevity and complete the proof.

C.4.3 Construction of hard problem instances: large uncertainty level

Construction of a collection of hard MDPs. Suppose there are two MDPs

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

Here, γ is the discount parameter, $\mathcal{S} = \{0, 1, \ldots, S-1\}$ is the state space, and $\mathcal{A} = \{0,1\}$ is the action space. The transition kernel P^ϕ of either constructed MDP $\mathcal{M}_\phi$ is defined as

$$P^\phi(s'\,|\,s,a) = \begin{cases} p\mathbb{1}(s'=2) + (1-p)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,\phi) \\ q\mathbb{1}(s'=2) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,1-\phi) \\ \mathbb{1}(s'=s) & \text{if} \quad s=1 \text{ or } s=2 \\ q\mathbb{1}(s'=s) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad s>2 \end{cases}, \tag{323}$$

where p and q are set as

$$p = 1-\alpha \quad \text{and} \quad q = 1-\alpha-\Delta \tag{324}$$

for some γ, α and Δ obeying

$$0 < \alpha \leq 1-\gamma \leq 1/(2e^8) \leq \frac{1}{2} \quad \text{and} \quad \Delta \leq \frac{\alpha}{2}. \tag{325}$$

Here, α and Δ are some values that will be introduced later. Consequently, applying (324) directly leads to

$$1 \geq p \geq q \geq \gamma \geq \frac{1}{2}. \tag{326}$$

Note that state 1 and 2 are absorbing states. In addition, if the initial distribution is supported on states $\{0, 1, 2\}$, the MDP will always stay in the state $\{1, 2\}$ after the first transition.

Finally, we define the reward function as

$$r(s, a) = \begin{cases} 1 & \text{if } s = 0 \text{ or } s = 2 \\ 0 & \text{otherwise} \end{cases}. \tag{327}$$

Construction of the history/batch dataset. Define a useful state distribution (only supported on the state subset $\{0, 1, 2\}$) as

$$\mu(s) = \frac{1}{CS}\mathbb{1}(s = 0) + \frac{1}{CS}\mathbb{1}(s = 2) + \Big(1 - \frac{2}{CS}\Big)\mathbb{1}(s = 1), \tag{328}$$

where $C > 0$ is some constant that determines the robust concentrability coefficient $C^\star_{\mathsf{rob}}$ (which will be made clear soon) and obeys

$$\frac{1}{CS} \leq \frac{1}{4}. \tag{329}$$

A batch dataset—consists of N i.i.d samples $\{(s_i, a_i, s_i')\}_{1 \leq i \leq N}$—is generated over the nominal environment $\mathcal{M}_\phi$ according to (40), with the behavior distribution chosen to be:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad d^{\mathsf{b}}(s, a) = \frac{\mu(s)}{2}. \tag{330}$$

Additionally, we choose the following initial state distribution:

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{otherwise} \end{cases}. \tag{331}$$

Uncertainty set of the transition kernels. We next describe the radius σ of the uncertainty set in our construction of the robust MDPs, along with some useful properties, which are similar to the finite-horizon case. To begin with, with slight abuse of notation, we introduce an important constant β defined as

$$\beta \coloneqq \frac{1}{2}\log\frac{1}{\alpha + \Delta} \geq 4. \tag{332}$$

The perturbed transition kernels in $\mathcal{M}_\phi$ is limited to the following uncertainty set

$$\mathcal{U}^\sigma(P^\phi) \coloneqq \otimes\, \mathcal{U}^\sigma\left(P^\phi_{s,a}\right), \qquad \mathcal{U}^\sigma(P^\phi_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{s,a} \parallel P^\phi_{s,a}\right) \leq \sigma\right\}, \tag{333}$$

where $P^\phi_{s,a} \coloneqq P^\phi(\cdot \,|\, s, a) \in [0, 1]^{1 \times S}$. Moreover, the radius of the uncertainty set σ obeys

$$\left(1 - \frac{3}{\beta}\right)\log\frac{1}{\alpha + \Delta} \leq \sigma \leq \left(1 - \frac{2}{\beta}\right)\log\frac{1}{\alpha + \Delta}. \tag{334}$$

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

$$\underline{P}^\phi(s' \,|\, s, a) \coloneqq \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})} P(s' \,|\, s, a). \tag{335}$$

As shall be seen, the transition from state 0 to state 2 plays an important role in the analysis, for convenience, we denote

$$\underline{p} \coloneqq \underline{P}^\phi(2 \,|\, 0, \phi), \qquad \underline{q} \coloneqq \underline{P}^\phi(2 \,|\, 0, 1-\phi). \tag{336}$$

With these definitions in place, we summarize some useful properties of the uncertainty set in the following lemma, which parallels Lemma 20 in the finite-horizon case.

Lemma 28 *Suppose β satisfies* (332) *and the uncertainty level σ satisfies* (334)*. The perturbed transition kernels obey*

$$\underline{p} \geq \underline{q} \geq \frac{1}{\beta}. \tag{337}$$

Proof The proof follows from the same arguments as Appendix B.3.10 by replacing H with $\frac{1}{1-\gamma}$; we omit the details for brevity. ■

Value functions and optimal policies. Now we are positioned to derive the corresponding robust value functions and identify the optimal policies. For any MDP $\mathcal{M}_\phi$ with the above uncertainty set, denote $\pi^\star_\phi$ as the optimal policy. In addition, we denote the robust value function of any policy π (resp. the optimal policy $\pi^\star_\phi$) as $V^{\pi,\sigma}_\phi$ (resp. $V^{\star,\sigma}_\phi$). Then, we introduce the following lemma which describes some important properties of the robust value functions and optimal policies.

Lemma 29 *For any $\phi = \{0, 1\}$ and any policy π, one has*

$$V^{\pi,\sigma}_\phi(0) = 1 + \frac{\gamma}{1-\gamma} z^\pi_\phi, \tag{338}$$

where z^π_ϕ is defined as

$$z^\pi_\phi \coloneqq \underline{p}\pi(\phi \,|\, 0) + \underline{q}\pi(1-\phi \,|\, 0). \tag{339}$$

In addition, the optimal value functions and the optimal policies obey

$$V^{\star,\sigma}_\phi(0) = 1 + \frac{\gamma}{1-\gamma}\underline{p}, \quad V^{\star,\sigma}_\phi(2) = \frac{1}{1-\gamma}, \quad V^{\star,\sigma}_\phi(s) = 0 \quad \text{for } s = 1 \text{ or } s > 2, \tag{340a}$$

$$\pi^\star_\phi(\phi \,|\, s) = 1, \qquad \text{for } s \in \mathcal{S}. \tag{340b}$$

Moreover, choosing $S \geq 2\beta$, the robust single-policy clipped concentrability coefficient $C^\star_{\mathsf{rob}}$ obeys

$$C^\star_{\mathsf{rob}} = 2C. \tag{341}$$

Proof See Appendix C.4.6. ■

C.4.4 Establishing the minimax lower bound: large uncertainty level

Now we are positioned to provide the sample complexity lower bound. In view of Lemma 29, the smallest positive state transition probability of the optimal policy $\pi_\phi^\star$ under any nominal transition kernel P^ϕ with $\phi \in \{0,1\}$ satisfies:

$$P_{\mathsf{min}}^\star := \min_{s,s'} \Big\{ P^\phi\left(s' \,|\, s, \pi_\phi^\star(s)\right) :\ P^\phi\left(s' \,|\, s, \pi_\phi^\star(s)\right) > 0 \Big\} = P^\phi\left(1|0,\phi\right) = 1-p. \tag{342}$$

Our goal is to control the quantity w.r.t. any policy estimator $\widehat{\pi}$ based on the batch dataset and the chosen initial distribution ρ in (331), which gives

$$V_\phi^{\star,\sigma}(\rho) - V_\phi^{\widehat{\pi},\sigma}(\rho) = V_\phi^{\star,\sigma}(0) - V_\phi^{\widehat{\pi},\sigma}(0). \tag{343}$$

Towards this, we first introduce the following lemma, which parallels the claim in (167)-(168) in the finite-horizon case.

Lemma 30 *Given* $\varepsilon \le \frac{1}{384e^6(1-\gamma)\log\left(\frac{1}{\alpha}\right)} \le \frac{1}{384e^6(1-\gamma)\log\left(\frac{1}{\alpha+\Delta}\right)}$, *choosing* $\Delta = 128e^6\sigma(1-q)\varepsilon(1-\gamma) \le 128e^6(\alpha+\Delta)\varepsilon\log\left(\frac{1}{\alpha+\Delta}\right)(1-\gamma) \le \frac{\alpha}{2}$, *one has for any policy* $\widehat{\pi}$,

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

Proof This lemma follows from the same arguments as Appendix B.3.12 except replacing H with $\frac{1}{1-\gamma}$ under the additional condition $\gamma \ge \frac{1}{2}$; we omit the details for brevity. ■

Armed with this lemma, following the same arguments in Appendix B.3.4, we can complete the proof by observing that: let c_1 be some sufficient large constant, as long as the sample size is beneath

$$N \le \frac{SC_{\mathsf{rob}}^\star \log 2}{4c_1 P_{\mathsf{min}}^\star \sigma^2 (1-\gamma)^2 \varepsilon^2}, \tag{344}$$

then we necessarily have

$$\inf_{\widehat{\pi}} \max_{\phi\in\{0,1\}} \mathbb{P}_\phi\left\{ V_\phi^{\star,\sigma}(\rho) - V_\phi^{\widehat{\pi},\sigma}(\rho) \ge \varepsilon \right\} \ge \frac{1}{8}, \tag{345}$$

where $\mathbb{P}_\phi$ denote the probability conditioned on that the MDP is $\mathcal{M}_\phi$. We omit the details for brevity and complete the proof.

C.4.5 Proof of Lemma 26

First, it is easily verified that for any policy π,

$$\forall s \in \mathcal{S}\setminus\{0\}: \quad V_\theta^{\pi,\sigma}(s) = \sum_{t=0}^{\infty} \gamma^t \cdot 0 = 0 \tag{346}$$

since the reward function $r(s,a) = 0$ for $(s,a) \in \mathcal{S}\setminus\{0\} \times \mathcal{A}$ in (307).

To continue, we observe that the robust value function of state 0 satisfies

$$\begin{aligned} V_\theta^{\pi,\sigma}(0) &= \mathbb{E}_{a\sim\pi(\cdot\,|\,0)}\left[r(0,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\theta_{0,a})} \mathcal{P}V_\theta^{\pi,\sigma}\right] \\ &\overset{\text{(i)}}{=} 1 + \gamma\pi(\theta\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\theta_{0,\theta})} \mathcal{P}V_\theta^{\pi,\sigma} + \gamma\pi(1-\theta\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\theta_{0,1-\theta})} \mathcal{P}V_\theta^{\pi,\sigma} \qquad (347) \\ &\overset{\text{(ii)}}{=} 1 + \gamma\pi(\theta\,|\,0)\left[\underline{p}_\star V_\theta^{\pi,\sigma}(0) + \left(1-\underline{p}_\star\right) V_\theta^{\pi,\sigma}(1)\right] \\ &\quad + \gamma\pi(1-\theta\,|\,0)\left[\underline{q}_\star V_\theta^{\pi,\sigma}(0) + \left(1-\underline{q}_\star\right) V_\theta^{\pi,\sigma}(1)\right] \\ &\overset{\text{(iii)}}{=} 1 + \gamma x_\theta^\pi \left[V_\theta^{\pi,\sigma}(0) - V_\theta^{\pi,\sigma}(1)\right] \\ &= \frac{1}{1-\gamma x_\theta^\pi}, \qquad (348) \end{aligned}$$

where (i) holds by the reward function defined in (307). To see (ii), note that (347) indicates $V_\theta^{\pi,\sigma}(0) \geq 1 \geq V_\theta^{\pi,\sigma}(1) = 0$, so that the infimum is obtained by picking the smallest possible mass on the transition to state 2, provided by the definition in (315), and (iii) follows by plugging in the definition of x_θ^π in (318).

Consequently, observing that the function $\frac{1}{1-\gamma x}$ is increasing in x and x_θ^π is also increasing in $\pi(\theta\,|\,0)$ (see the fact $\underline{p}_\star \geq \underline{q}_\star$ in (316)), the optimal policy in state 0 thus obeys

$$\pi_\theta^\star(\theta\,|\,0) = 1. \qquad (349)$$

Therefore,

$$x_\theta^\star \coloneqq x_\theta^{\pi^\star} = \underline{p}_\star \pi_\theta^\star(\theta\,|\,0) + \underline{q}_\star \pi_\theta^\star(1-\theta\,|\,0) = \underline{p}_\star, \qquad (350)$$

which combined with (348) yields

$$V_\theta^{\star,\sigma}(0) = \frac{1}{1-\gamma\underline{p}_\star}. \qquad (351)$$

Regarding the optimal policy for the remaining states $s > 0$, since the action does not influence the state transition, without loss of generality, we choose the optimal policy to obey

$$\forall s > 0: \quad \pi_\theta^\star(\theta\,|\,s) = 1. \qquad (352)$$

Proof of (320)**.** To begin with, for any MDP $\mathcal{M}_\theta$ with $\theta \in \{0,1\}$, recall the definition of $C^\star_{\mathsf{rob}}$ as

$$C^\star_{\mathsf{rob}} = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)}. \qquad (353)$$

Given $\pi_\theta^\star(\theta\,|\,s) = 1$ for all $s \in \mathcal{S}$ and the initial distribution $\rho(0) = 1$, for any $P \in \mathcal{U}^\sigma(P^\theta)$, we arrive at

$$d^{\star,P}(0,\theta) = (1-\gamma)\rho(0)\sum_{t=0}^{\infty}\gamma^t\left(\underline{P}^\theta(0\,|\,0,\theta)\right)^t$$

$$\overset{\text{(i)}}{=} (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \underline{p}_\star^t = \frac{1-\gamma}{1-\gamma \underline{p}_\star} \overset{\text{(ii)}}{\geq} \frac{1-\gamma}{1-\gamma(1-c_3(1-\gamma))} \geq \frac{1}{2}. \tag{354}$$

where (i) holds by (315) and (ii) follows from (316). In addition, we have

$$d^{\star,P}(0,1-\theta) = 0 \quad \text{and} \quad \forall s>1: \quad d^{\star,P}(s,a) = 0, \tag{355}$$

since $\rho(0)=1$ and state 0 and 1 are absorbing states for all policy and and all $P \in \mathcal{U}^\sigma(P^\theta)$.

Armed with the above facts, we observe that

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)} = \max_{s\in\{0,1\},P\in\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d^{\star,P}(s,\theta), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,\theta)} \tag{356}$$

which follows from the properties of the optimal policy in (319).

Consequently, we control $C^\star_{\mathsf{rob}}$ in states separately:

$$\max_{P\in\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d^{\star,P}(0,\theta), \frac{1}{S}\right\}}{d^{\mathsf{b}}(0,\theta)} \overset{\text{(i)}}{=} \frac{1}{Sd^{\mathsf{b}}(0,\theta)} \overset{\text{(ii)}}{=} \frac{2}{S\mu(0)} = 2C, \tag{357a}$$

$$\max_{P\in\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d^{\star,P}(1,\theta), \frac{1}{S}\right\}}{d^{\mathsf{b}}(1,\theta)} \leq \frac{1}{Sd^{\mathsf{b}}(1,\theta)} \overset{\text{(iii)}}{=} \frac{2}{S\left(1-\frac{1}{CS}\right)} \overset{\text{(iv)}}{\leq} \frac{4}{S} \overset{\text{(v)}}{\leq} C, \tag{357b}$$

where (i) holds by (354) and $S \geq 2$, (ii) and (iii) follow from the definitions in (310) and (308), and (iv) and (v) and arise from the assumption in (309). Plugging the above results back into (356) directly completes the proof of

$$C^\star_{\mathsf{rob}} = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\theta)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)} = 2C.$$

C.4.6 Proof of Lemma 29

For any $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$, we first characterize the robust value function for any policy π over different states. due to state absorbing, the uncertainty set becomes a singleton containing the nominal distribution at state $s=1$ and $s=2$. It is easily observed that for any policy π, the robust value functions at state $s=1$ and $s=2$ obey

$$V_\phi^{\pi,\sigma}(1) = \sum_{t=0}^{\infty}\gamma^t \cdot 0 = 0, \tag{358a}$$

$$V_\phi^{\pi,\sigma}(2) = \sum_{t=0}^{\infty}\gamma^t \cdot 1 = \frac{1}{1-\gamma}, \tag{358b}$$

since $r(1,a)=0$ and $r(2,a)=1$. In addition, for state $s>2$, the perturbed transition kernel is supported on itself and state 1, both of which receive a reward of 0 by design (327), leading to

$$V_\phi^{\pi,\sigma}(s) = \sum_{t=0}^{\infty}\gamma^t \cdot 0 = 0, \qquad \text{for } s>2. \tag{358c}$$

Moving onto the remaining states, the robust value function of state 0 satisfies

$$\begin{aligned}
V_\phi^{\pi,\sigma}(0) &= \mathbb{E}_{a\sim\pi(\cdot\,|\,0)}\left[r(0,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,a})} \mathcal{P}V_\phi^{\pi,\sigma}\right] \\
&\overset{\text{(i)}}{=} 1 + \gamma\pi(\phi\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,\phi})} \mathcal{P}V_\phi^{\pi,\sigma} + \gamma\pi(1-\phi\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,1-\phi})} \mathcal{P}V_\phi^{\pi,\sigma} \\
&\overset{\text{(ii)}}{=} 1 + \gamma\pi(\phi\,|\,0)\Big[\underline{p}V_\phi^{\pi,\sigma}(2) + \big(1-\underline{p}\big)\,V_\phi^{\pi,\sigma}(1)\Big] \\
&\quad + \gamma\pi(1-\phi\,|\,0)\Big[\underline{q}V_\phi^{\pi,\sigma}(2) + \big(1-\underline{q}\big)\,V_\phi^{\pi,\sigma}(1)\Big] \\
&\overset{\text{(iii)}}{=} 1 + \gamma V_\phi^{\pi,\sigma}(1) + \gamma z_\phi^\pi\left[V_\phi^{\pi,\sigma}(2) - V_\phi^{\pi,\sigma}(1)\right] \\
&= 1 + \frac{\gamma}{1-\gamma} z_\phi^\pi,
\end{aligned} \tag{359}$$

where (i) holds by the reward function defined in (327). To see (ii), note that (358) indicates $V_\phi^{\pi,\sigma}(2) \geq V_\phi^{\pi,\sigma}(1)$, so that the infimum is obtained by picking the smallest possible mass on the transition to state 2, provided by the definition in (336). Last but not least, (iii) follows by plugging in the definition of z_ϕ^π in (339), and the last identity is due to (358). Consequently, taking $\pi = \pi_\phi^\star$, we directly arrive at

$$V_\phi^{\star,\sigma}(0) = 1 + \frac{\gamma}{1-\gamma} z_\phi^{\pi^\star}. \tag{360}$$

Observing that the function $z\frac{\gamma}{1-\gamma}$ is increasing in z and z_ϕ^π is also increasing in $\pi(\phi\,|\,0)$ (see the fact $\underline{p} \geq \underline{q}$ in (337)), the optimal policy in state 0 thus obeys

$$\pi_\phi^\star(\phi\,|\,0) = 1. \tag{361}$$

Finally, plugging the above fact back into (339) leads to

$$z_\phi^\star \coloneqq z_\phi^{\pi^\star} = \underline{p}\pi_\phi^\star(\phi\,|\,0) + \underline{q}\pi_\phi^\star(1-\phi\,|\,0) = \underline{p}, \tag{362}$$

which combined with (360) yields

$$V_\phi^{\star,\sigma}(0) = 1 + \frac{\gamma}{1-\gamma}\underline{p}. \tag{363}$$

Regarding the optimal policy for the remaining states $s > 0$, since the action does not influence the state transition, without loss of generality, we choose the optimal policy to obey

$$\forall s > 0: \quad \pi_\phi^\star(\phi\,|\,s) = 1. \tag{364}$$

Proof of (341)**.** To begin with, for any MDP $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$, recall the definition of $C^\star_{\mathsf{rob}}$ as

$$C^\star_{\mathsf{rob}} = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)}. \tag{365}$$

Given $\pi^\star_\phi(\phi \,|\, s) = 1$ for all $s \in \mathcal{S}$ and the initial distribution $\rho(0) = 1$, for any $P \in \mathcal{U}^\sigma(P^\phi)$, we arrive at

$$d^{\star,P}(0,\phi) = (1-\gamma)\rho(0)\pi^\star_\phi(\phi \,|\, 0) = (1-\gamma), \tag{366}$$

which holds due to that the agent transits from state 0 to other states at the first step and then will never go back to state 0. In addition, one has for any $P \in \mathcal{U}^\sigma(P^\phi)$,

$$\begin{aligned} d^{\star,P}(2,\phi) &= (1-\gamma)P(2 \,|\, 0,\phi)\sum_{t=1}^{\infty}\gamma^t\big(P(2 \,|\, 2,\phi)\big)^t \\ &= (1-\gamma)P(2 \,|\, 0,\phi)\sum_{t=1}^{\infty}\gamma^t \overset{\text{(i)}}{\geq} \gamma\underline{p} \geq \frac{1}{2\beta}, \end{aligned} \tag{367}$$

where (i) holds by (336) and the final inequality follows from (337) and $\gamma \geq 1/2$. Armed with the above facts, we observe that

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d^{\star,P}(s,a), \frac{1}{S}\big\}}{d^{\mathsf{b}}(s,a)} = \max_{s\in\{0,1,2\},P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d^{\star,P}(s,\phi), \frac{1}{S}\big\}}{d^{\mathsf{b}}(s,\phi)} \tag{368}$$

which follows from the properties of the optimal policy in (364) and consequently $d^{\star,P}(s) = d^{\star,P}(s,\phi) = 0$ for all $s > 2$ and all $P \in \mathcal{U}^\sigma(P^\phi)$.

To continue, we control the term in states $\{0,1,2\}$ separately:

$$\max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d^{\star,P}(2,\phi), \frac{1}{S}\big\}}{d^{\mathsf{b}}(2,\phi)} \overset{\text{(i)}}{=} \frac{1}{Sd^{\mathsf{b}}(2,\phi)} \overset{\text{(ii)}}{=} \frac{2}{S\mu(2)} = 2C, \tag{369a}$$

$$\max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d^{\star,P}(0,\phi), \frac{1}{S}\big\}}{d^{\mathsf{b}}(0,\phi)} \leq \frac{1}{Sd^{\mathsf{b}}(0,\phi)} \overset{\text{(iii)}}{=} \frac{2}{S\mu(0)} = 2C, \tag{369b}$$

$$\max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d^{\star,P}(1,\phi), \frac{1}{S}\big\}}{d^{\mathsf{b}}(1,\phi)} \leq \frac{1}{Sd^{\mathsf{b}}(1,\phi)} \overset{\text{(iv)}}{=} \frac{2}{S\left(1-\frac{2}{CS}\right)} \overset{\text{(v)}}{\leq} \frac{4}{S} \overset{\text{(vi)}}{\leq} C, \tag{369c}$$

where (i) holds by (367) and $S \geq 2\beta$, (ii), (iii) and (iv) follow from the definitions in (330) and (328), (v) and (vi) arise from the assumption in (329). Plugging the above results back into (368) directly completes the proof of

$$C^\star_{\mathsf{rob}} = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d^{\star,P}(s,a), \frac{1}{S}\big\}}{d^{\mathsf{b}}(s,a)} = 2C.$$

References

Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian Zhang, and Jun Wang. Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*, 2019.

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. *Conference on Learning Theory*, pages 67–83, 2020.

Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR, 2021.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

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

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

James J Choi, David Laibson, Brigitte C Madrian, and Andrew Metrick. Reinforcement learning and savings behavior. *The Journal of finance*, 64(6):2515–2534, 2009.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Esther Derman and Shie Mannor. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

Wenhao Ding, Laixi Shi, Yuejie Chi, and Ding Zhao. Seeing is not believing: Robust reinforcement learning against spurious correlation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

John C Duchi. Introductory lectures on stochastic optimization. *The Mathematics of Data*, 25:99–186, 2018.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Rui Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 2022.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522, 1952.

Vineet Goyal and Julien Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Fast Bellman updates for robust MDPs. In *International Conference on Machine Learning*, pages 1979–1988. PMLR, 2018.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

Linfang Hou, Liang Pang, Xin Hong, Yanyan Lan, Zhiming Ma, and Dawei Yin. Robust reinforcement learning with Wasserstein constraint. *arXiv preprint arXiv:2006.00945*, 2020.

Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pages 1695–1724, 2013.

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

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096, 2021.

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

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc., 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473, 2021.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

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

Arnab Nilim and Laurent Ghaoui. Robustness in markov decision problems with uncertain transition matrices. *Advances in neural information processing systems*, 16, 2003.

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

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Neural Information Processing Systems (NeurIPS)*, 2021.

Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

John Schulman, Jonathan Ho, Alex X Lee, Ibrahim Awwal, Henry Bradlow, and Pieter Abbeel. Finding locally optimal, collision-free trajectories with sequential convex optimization. In *Robotics: science and systems*, volume 9, pages 1–10. Citeseer, 2013.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. 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, 2022.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. *arXiv preprint arXiv:2404.18909*, 2024.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

Jun Song and Chaoyue Zhao. Optimistic distributionally robust policy optimization. *arXiv preprint arXiv:2006.07815*, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust MDPs using function approximation. In *International conference on machine learning*, pages 181–189. PMLR, 2014.

FLEMMING Topsøe. Some bounds for the logarithmic function. *Inequality theory and applications*, 4:137, 2007.

A. B. Tsybakov and V. Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

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

He Wang, Laixi Shi, and Yuejie Chi. Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*, 2024.

Jie Wang, Rui Gao, and Hongyuan Zha. Reliable off-policy evaluation for reinforcement learning. *Operations Research*, 2022.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust Q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3370–3398. PMLR, 2023a.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust Q-learning. *arXiv preprint arXiv:2305.18420*, 2023b.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34, 2021.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Eric M Wolff, Ufuk Topcu, and Richard M Murray. 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, 2012.

Jiin Woo, Laixi Shi, Gauri Joshi, and Yuejie Chi. Federated offline reinforcement learning: Collaborative single-policy coverage suffices. *arXiv preprint arXiv:2402.05876*, 2024.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

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

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous Q-learning. *IEEE Transactions on Information Theory*, 2023.

Insoon Yang. A convex optimization approach to distributionally robust markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169, 2017.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

Ming Yin and Yu-Xiang Wang. Optimal uniform OPE and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings. *arXiv preprint arXiv:2105.06029*, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34, 2021.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020.

Zhengqing Zhou, Qinxun Bai, Zhengyuan Zhou, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR, 2021.