# Appendix

## Table of Contents



## A 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 bounded concentrability coefficients (Chen & Jiang, 2019; Munos, 2005) and uniformly lower bounded data visitation distribution (Yin & 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., 2022), 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 & Mehrotra, 2019; Gao, 2020; Bertsimas et al., 2018; Duchi & Namkoong, 2018; Blanchet & Murthy, 2019), distributionally robust dynamic programming has also attracted considerable amount of attention, e.g. Iyengar (2005); Nilim & Ghaoui (2003); Xu & Mannor (2012); Nilim & El Ghaoui (2005), where natural robust extensions to the standard Bellman 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 & Mannor, 2012; Wolff et al., 2012; Kaufman & Schaefer, 2013; Ho et al., 2018; Smirnova et al., 2019; Ho et al., 2021; Goyal & Grand-Clement, 2022; Derman & Mannor, 2020; Tamar et al., 2014; Badrinath & Kalathil, 2021). Nonetheless, the majority of prior theoretical analyses focus on planning with an exact knowledge of the uncertainty set (Iyengar, 2005; Xu & 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 & 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 & Kalathil (2021) proposed a model-free algorithm for the online setting with linear function approximation to cope with large state spaces. Yang et al. (2021); Panaganti & 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.

## B PRELIMINARIES

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

### B.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 2** ((Hu & 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*

$$\sup_{\mathcal{P}\in\mathcal{U}^\sigma(P^0)} \mathbb{E}_{X\sim\mathcal{P}}[f(X)] = \inf_{\lambda\geq 0}\left\{\lambda\log\mathbb{E}_{X\sim P^0}\left[\exp\left(\frac{f(X)}{\lambda}\right)\right] + \lambda\sigma\right\}. \tag{30}$$

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

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0)} \mathcal{P}X = \sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^0\exp\left(-\frac{X}{\lambda}\right)\right) - \lambda\sigma\right\}. \tag{31}$$

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 3** ((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*

$$\lambda^\star \in \arg\max_{\lambda\geq 0} f(\lambda, P), \qquad \textit{where } f(\lambda, P) := \left\{-\lambda\log\mathbb{E}_{X\sim P}\left[\exp\left(\frac{-X}{\lambda}\right)\right] - \lambda\sigma\right\}. \tag{32}$$

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

$$\lambda^\star \in \left[0, \frac{M}{\sigma}\right], \tag{33}$$

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

$$\log\left(\mathbb{P}(X = \operatorname{essinf} X)\right) + \sigma \geq 0. \tag{34}$$

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

$$\lim_{\lambda \to 0} f(\lambda, P) = \lim_{\lambda \to 0} \left\{ -\lambda \log \mathbb{E}_{X \sim P}\left[\exp\left(\frac{-X}{\lambda}\right)\right] - \lambda\sigma \right\} = \operatorname{essinf} X. \tag{35}$$

### B.2 Concentration inequalities

In light of Lemma 3 (cf. 35), 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 4** ((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*

$$\min_{i \in [n]} X_i = \operatorname{essinf} \widehat{X} = \operatorname{essinf} X, \tag{36}$$

*as long as*

$$n \geq -\frac{\log(2/\delta)}{\log(1 - P_{\mathsf{min},X})}. \tag{37}$$

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

**Lemma 5** (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*

$$\mathbb{P}\left(|N/n - p| \geq pt\right) \leq \exp\left(-c_{\mathsf{f}} npt^2\right), \qquad \forall t \in [0, 1]. \tag{38}$$

**Lemma 6** ((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*

$$N \geq \frac{np}{8\log\left(\frac{1}{\delta}\right)} \qquad \text{if } np \geq 8\log\left(\frac{1}{\delta}\right), \tag{39a}$$

$$N \leq \begin{cases} e^2 np & \text{if } np \geq \log\left(\frac{1}{\delta}\right), \\ 2e^2 \log\left(\frac{1}{\delta}\right) & \text{if } np \leq 2\log\left(\frac{1}{\delta}\right) \end{cases} \tag{39b}$$

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

### B.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

$$\mathsf{KL}\big(\mathsf{Ber}(p) \parallel \mathsf{Ber}(q)\big) \coloneqq p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}, \tag{40}$$

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

**Lemma 7.** *For any* $p, q \in \left[\frac{1}{2}, 1\right)$ *and* $p > q$*, it holds that*

$$\mathsf{KL}\big(\mathsf{Ber}(p) \parallel \mathsf{Ber}(q)\big) \leq \mathsf{KL}\big(\mathsf{Ber}(q) \parallel \mathsf{Ber}(p)\big) \leq \frac{(p-q)^2}{p(1-p)}. \tag{41}$$

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

$$\mathsf{KL}\left(\mathsf{Ber}\left(x\right) \parallel \mathsf{Ber}(q)\right) > \mathsf{KL}\left(\mathsf{Ber}\left(y\right) \parallel \mathsf{Ber}(q)\right). \tag{42}$$

*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 $\frac{\partial f(x,q)}{\partial x} = \log\frac{x}{q} + \log\frac{1-q}{1-x} < 0$. □

# C Analysis: episodic finite-horizon RMDPs

## C.1 Proof of the upper bound: Theorem 1

In this section, we outline the proof of Theorem 1. 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 $\pi^{\mathsf{b}}$ in the nominal model $P^0$ as

$$\mathcal{C}^{\mathsf{b}} = \left\{(h,s,a) : d_h^{\mathsf{b},P^0}(s,a) > 0\right\}. \tag{43}$$

Moreover, we recall the definition in (22) and define a similar one based on the exact nominal model $P^0$ as

$$P_{\mathsf{min},h}(s,a) \coloneqq \min_{s'}\left\{P_h^0(s' \mid s,a) :\ P_h^0(s' \mid s,a) > 0\right\}. \tag{44}$$

Clearly, by comparing with the definitions (23) and (24), it holds that

$$P_{\mathsf{min}}^\star = \min_{h,s}\ P_{\mathsf{min},h}(s,\pi_h^\star(s)), \qquad P_{\mathsf{min}}^{\mathsf{b}} = \min_{(h,s,a)\in\mathcal{C}^{\mathsf{b}}}\ P_{\mathsf{min},h}(s,a). \tag{45}$$

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

$$\mathcal{D}_h^\star \coloneqq \left\{\left[d_h^{\star,P}(s)\right]_{s\in\mathcal{S}} : P \in \mathcal{U}^\sigma\left(P^0\right)\right\} = \left\{\left[d_h^{\star,P}\left(s,\pi_h^\star(s)\right)\right]_{s\in\mathcal{S}} : P \in \mathcal{U}^\sigma\left(P^0\right)\right\}, \tag{46}$$

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 1 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 C.2.

**Lemma 8.** *Instate the assumptions in Theorem 1. Then for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$, consider any vector $V \in \mathbb{R}^S$ independent of $\widehat{P}_{h,s,a}^0$ obeying $\|V\|_\infty \le H$. With probability at least $1-\delta$, one has*

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}_{h,s,a}^0)} \mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{h,s,a}^0)} \mathcal{P}V\right| \le b_h(s,a) \tag{47}$$

*with $b_h(s,a)$ given in (21). Moreover, for all $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$, with probability at least $1-\delta$, one has*

$$\frac{P_{\mathsf{min},h}(s,a)}{8\log(KHS/\delta)} \le \widehat{P}_{\mathsf{min},h}(s,a) \le e^2 P_{\mathsf{min},h}(s,a). \tag{48}$$

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

$$\forall(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H+1]: \qquad \widehat{Q}_h(s,a) \le Q_h^{\widehat{\pi},\sigma}(s,a), \qquad \widehat{V}_h(s) \le V_h^{\widehat{\pi},\sigma}(s), \tag{49}$$

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

$$\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a) \le \max_a Q_h^{\widehat{\pi},\sigma}(s,a) = V_h^{\widehat{\pi},\sigma}(s). \tag{50}$$

Therefore, the remainder of this step focuses on verifying the former assertion in (49) by induction.

- To begin, the claim (49) 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 1 (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

$$\begin{aligned}
\widehat{Q}_h(s,a) &= r_h(s,a) + \sup_{\lambda \ge 0}\left\{-\lambda \log\left(\widehat{P}^0_{h,s,a} \cdot \exp\left(\frac{-\widehat{V}_{h+1}}{\lambda}\right)\right) - \lambda\sigma\right\} - b_h(s,a) \\
&\overset{\text{(i)}}{=} r_h(s,a) + \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} - b_h(s,a) \\
&\le r_h(s,a) + \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} - b_h(s,a) \qquad (51) \\
&\quad + \left|\inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} - \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1}\right| \\
&\overset{\text{(ii)}}{\le} r_h(s,a) + \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V_{h+1}^{\widehat{\pi},\sigma} + 0 \overset{\text{(iii)}}{=} Q_h^{\widehat{\pi},\sigma}(s,a), \qquad (52)
\end{aligned}$$

  where (i) rewrites the update rule back to its primal form (cf. (18)), (ii) holds by applying (47) with the condition (27) satisfied and the induction hypothesis $\widehat{V}_{h+1} \le V_{h+1}^{\widehat{\pi},\sigma}$, and lastly, (iii) follows by the robust Bellman consistency equation (8).

Putting them together, we have verified the claim (49) by induction.

**Step 2: bounding $V_h^{\star,\sigma}(s) - V_h^{\widehat{\pi},\sigma}(s)$.** With the pessimism property (49) in place, we observe that the following relation holds

$$0 \le V_h^{\star,\sigma}(s) - V_h^{\widehat{\pi},\sigma}(s) \le V_h^{\star,\sigma}(s) - \widehat{V}_h(s) \le Q_h^{\star,\sigma}\big(s,\pi_h^\star(s)\big) - \widehat{Q}_h\big(s,\pi_h^\star(s)\big), \qquad (53)$$

where the last inequality follows from $\widehat{Q}_h\big(s,\pi_h^\star(s)\big) \le \max_a \widehat{Q}_h(s,a) = \widehat{V}_h(s)$. Then, by the robust Bellman optimality equation in (9) and the primal version of the update rule (cf. (18))

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

we arrive at

$$\begin{aligned}
V_h^{\star,\sigma}(s) - \widehat{V}_h(s) &\le Q_h^{\star,\sigma}\big(s,\pi_h^\star(s)\big) - \widehat{Q}_h\big(s,\pi_h^\star(s)\big) \\
&= \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}V_{h+1}^{\star,\sigma} - \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(\widehat{P}^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}\widehat{V}_{h+1} + b_h\big(s,\pi_h^\star(s)\big) \\
&\le \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}V_{h+1}^{\star,\sigma} - \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}\widehat{V}_{h+1} \\
&\quad + \left|\inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(\widehat{P}^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}\widehat{V}_{h+1} - \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}\widehat{V}_{h+1}\right| + b_h\big(s,\pi_h^\star(s)\big) \\
&\overset{\text{(i)}}{\le} \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}V_{h+1}^{\star,\sigma} - \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{h,s,\pi_h^\star(s)}\right)} \mathcal{P}\widehat{V}_{h+1} + 2b_h\big(s,\pi_h^\star(s)\big) \\
&\overset{\text{(ii)}}{\le} \widehat{P}^{\inf}_{h,s,\pi_h^\star(s)}\big(V_{h+1}^{\star,\sigma} - \widehat{V}_{h+1}\big) + 2b_h\big(s,\pi_h^\star(s)\big), \qquad (54)
\end{aligned}$$

where (i) holds by applying Lemma 2 (cf. (47)) 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

$$\widehat{P}^{\inf}_{h,s,\pi^\star_h(s)} \coloneqq \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} \tag{55}$$

and consequently,

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

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

$$\left[\widehat{P}^{\inf}_h\right]_{s,\cdot} = \widehat{P}^{\inf}_{h,s,\pi^\star_h(s)}, \qquad \text{and} \qquad b^\star_h(s) = b_h\big(s,\pi^\star_h(s)\big). \tag{56}$$

Applying (54) recursively over the time steps $h, h+1, \cdots, H$ using the above notation gives

$$\begin{aligned} 0 \le V^{\star,\sigma}_h - \widehat{V}_h &\le \widehat{P}^{\inf}_h\big(V^{\star,\sigma}_{h+1} - \widehat{V}_{h+1}\big) + 2b^\star_h \\ &\le \widehat{P}^{\inf}_h \widehat{P}^{\inf}_{h+1}\big(V^{\star,\sigma}_{h+2} - \widehat{V}_{h+2}\big) + 2\widehat{P}^{\inf}_h b^\star_{h+1} + 2b^\star_h \le \cdots \le 2\sum_{i=h}^{H}\left(\prod_{j=h}^{i-1}\widehat{P}^{\inf}_j\right) b^\star_i, \end{aligned} \tag{57}$$

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

For any $d^\star_h \in \mathcal{D}^\star_h$ (cf. (46)), taking inner product with (57) leads to

$$\left\langle d^\star_h, V^{\star,\sigma}_h - \widehat{V}_h\right\rangle \le \left\langle d^\star_h, 2\sum_{i=h}^{H}\left(\prod_{j=h}^{i-1}\widehat{P}^{\inf}_j\right) b^\star_i\right\rangle = 2\sum_{i=h}^{H}\left\langle d^\star_i, b^\star_i\right\rangle, \tag{58}$$

where

$$d^\star_i \coloneqq \left[\left(d^\star_h\right)^\top\left(\prod_{j=h}^{i-1}\widehat{P}^{\inf}_j\right)\right]^\top \in \mathcal{D}^\star_i \tag{59}$$

by the definition of $\mathcal{D}^\star_i$ (cf. (46)) for all $i = h+1, \cdots, H$.

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

- For $s\in S$ where $\max_{P\in\mathcal{U}^\sigma(P^0)} d^{\star,P}_i\big(s,\pi^\star_i(s)\big) = \max_{P\in\mathcal{U}^\sigma(P^0)} d^{\star,P}_i(s) = 0$, it follows from the definition (cf. (46)) that for any $d^\star_i \in \mathcal{D}^\star_i$, it satisfies that

$$d^\star_i(s) = 0. \tag{60}$$

- For $s\in S$ where $\max_{P\in\mathcal{U}^\sigma(P^0)} d^{\star,P}_i\big(s,\pi^\star_i(s)\big) = \max_{P\in\mathcal{U}^\sigma(P^0)} d^{\star,P}_i(s) > 0$, by the assumption in (12)

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

it implies that

$$d^{\mathsf{b},P^0}_i\big(s,\pi^\star_i(s)\big) > 0 \quad \text{and} \quad \big(i,s,\pi^\star_i(s)\big) \in \mathcal{C}^{\mathsf{b}}. \tag{61}$$

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

$$N_i\big(s,\pi_i^\star(s)\big) \geq \frac{Kd_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)}{8} - 5\sqrt{Kd_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)\log\frac{KH}{\delta}} \overset{\text{(i)}}{\geq} \frac{Kd_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)}{16}$$
$$\overset{\text{(ii)}}{\geq} \frac{K\max_{P\in\mathcal{U}^\sigma(P^0)}\min\left\{d_i^{\star,P}\big(s,\pi_i^\star(s)\big),\frac{1}{S}\right\}}{16C^\star_{\mathsf{rob}}} \geq \frac{K\min\left\{d_i^\star(s),\frac{1}{S}\right\}}{16C^\star_{\mathsf{rob}}}, \tag{62}$$

where (i) holds due to

$$Kd_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big) \geq c_1\frac{d_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)\log(KHS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}}P^{\mathsf{b}}_{\mathsf{min}}} \geq \frac{c_1\log\frac{KH}{\delta}}{P^{\mathsf{b}}_{\mathsf{min}}} \geq c_1\log\frac{KH}{\delta} \tag{63}$$

for some sufficiently large $c_1$, where the first inequality follows from Condition (27), the second inequality follows from

$$d^{\mathsf{b}}_{\mathsf{min}} = \min_{h,s,a}\left\{d_h^{\mathsf{b},P^0}(s,a) : d_h^{\mathsf{b},P^0}(s,a)>0\right\} \leq d_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big) \tag{64}$$

and the last inequality follows from $P^{\mathsf{b}}_{\mathsf{min}}\leq 1$. In addition, (ii) follows from Assumption 1. With this in place, we observe that the pessimistic penalty (see (21)) obeys

$$b_i^\star(s) \leq c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{\widehat{P}_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)N_i\big(s,\pi_i^\star(s)\big)}} \overset{\text{(i)}}{\leq} 4c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log^2(\frac{KHS}{\delta})}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)N_i\big(s,\pi_i^\star(s)\big)}}$$
$$\leq 16c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{C^\star_{\mathsf{rob}}\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K\min\left\{d_i^\star(s),\frac{1}{S}\right\}}}, \tag{65}$$

where (i) holds by applying (48) in view of the fact that $\big(i,s,\pi_i^\star(s)\big)\in\mathcal{C}^{\mathsf{b}}$ by (61), and the last inequality holds by (62).

Combining the results in the above two cases leads to

$$\sum_{s\in\mathcal{S}} d_i^\star(s)b_i^\star(s) \leq \sum_{s\in\mathcal{S}} 16d_i^\star(s)c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{C^\star_{\mathsf{rob}}\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K\min\left\{d_i^\star(s),\frac{1}{S}\right\}}}$$
$$\overset{\text{(i)}}{\leq} 16c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\sum_{s\in\mathcal{S}} d_i^\star(s)\frac{C^\star_{\mathsf{rob}}\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K\min\left\{d_i^\star(s),\frac{1}{S}\right\}}}\sqrt{\sum_{s\in\mathcal{S}} d_i^\star(s)}$$
$$\leq 32c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{SC^\star_{\mathsf{rob}}\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K}}, \tag{66}$$

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

$$\sum_{s\in\mathcal{S}}\frac{d_i^\star(s)}{\min\left\{d_i^\star(s),\frac{1}{S}\right\}} \leq \sum_{s\in\mathcal{S}} d_i^\star(s)\left(\frac{1}{d_i^\star(s)}+\frac{1}{1/S}\right) = \sum_{s\in\mathcal{S}}1+\frac{1}{S}\sum_{s\in\mathcal{S}}d_i^\star(s) \leq 2S. \tag{67}$$

**Step 4: finishing up the proof.** Then, inserting (66) back into (58) with $h=1$ shows

$$\left\langle d_1^\star, V_1^{\star,\sigma}-\widehat{V}_1\right\rangle \leq 2\sum_{i=1}^H\langle d_i^\star,b_i^\star\rangle \leq \sum_{i=1}^H 64c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{SC^\star_{\mathsf{rob}}\log^2\frac{KH}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K}} \leq c_2\frac{H^2}{\sigma}\sqrt{\frac{SC^\star_{\mathsf{rob}}\log^2\frac{KH}{\delta}}{P^\star_{\mathsf{min}}K}}, \tag{68}$$

where the last inequality holds by plugging in the relation $P^\star_{\mathsf{min}}\leq P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)$ for $i=1,\ldots,H$ by the definition in (23) (see also (45)), and choosing $c_2$ to be large enough. The proof is completed.

### C.2 Proof of Lemma 8

To begin, we shall introduce the following fact that

$$\forall (h,s,a) \in \mathcal{C}^{\mathsf{b}}: \qquad N_h(s,a) \geq \frac{c_1 \log \frac{KHS}{\delta}}{16 P_{\mathsf{min},h}(s,a)} \geq -\frac{\log \frac{2KHS}{\delta}}{\log(1-P_{\mathsf{min},h}(s,a))}, \tag{69}$$

as long as Condition (27) holds. The proof is postponed to Appendix C.2.3. With this in mind, we shall first establish the simpler bound (48) and then move on to show (47).

#### C.2.1 Proof of (48)

To begin, recall that (69) is satisfied for all $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$. By Lemma 6 and the union bound, it holds that with probability at least $1-\delta$ that for all $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$:

$$\forall s' \in \mathcal{S}: \qquad P_h^0(s' \,|\, s,a) \geq \frac{\widehat{P}_h^0(s' \,|\, s,a)}{e^2} \geq \frac{P_h^0(s' \,|\, s,a)}{8e^2 \log(\frac{KHS}{\delta})}. \tag{70}$$

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

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

where (i) and (ii) follow from (70).

#### C.2.2 Proof of (47)

The main goal of (47) is to control the gap between robust Bellman operations based on the nominal transition kernel $P_{h,s,a}^0$ and the estimated kernel $\widehat{P}_{h,s,a}^0$ 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, (47) can be easily verified that

$$\left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}_{h,s,a}^0)} \mathcal{P}V - \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P_{h,s,a}^0)} \mathcal{P}V \right| \overset{\text{(i)}}{=} \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P_{h,s,a}^0)} \mathcal{P}V \leq \|V\|_\infty \overset{\text{(ii)}}{\leq} H \overset{\text{(iii)}}{=} b_h(s,a), \tag{71}$$

where (i) follows from the fact $\widehat{P}_{h,s,a}^0 = 0$ when $N_h(s,a) = 0$ (see (15)), (ii) arises from the assumption $\|V\|_\infty \leq H$, and (iii) holds by the definition of $b_h(s,a)$ in (21). Therefore, the remainder of the proof will focus on verifying (47) for $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$. Rewriting the term of interest via duality (cf. Lemma 2) yields

$$\begin{aligned} &\left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}_{h,s,a}^0)} \mathcal{P}V - \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P_{h,s,a}^0)} \mathcal{P}V \right| \\ &= \left| \sup_{\lambda \geq 0} \left\{ -\lambda \log\left( \widehat{P}_{h,s,a}^0 \exp\left( \frac{-V}{\lambda} \right) \right) - \lambda\sigma \right\} - \sup_{\lambda \geq 0} \left\{ -\lambda \log\left( P_{h,s,a}^0 \exp\left( \frac{-V}{\lambda} \right) \right) - \lambda\sigma \right\} \right|. \end{aligned} \tag{72}$$

Denoting

$$\widehat{\lambda}_{h,s,a}^\star := \arg\max_{\lambda \geq 0} \left\{ -\lambda \log\left( \widehat{P}_{h,s,a}^0 \exp\left( \frac{-V}{\lambda} \right) \right) - \lambda\sigma \right\}, \tag{73a}$$

$$\lambda_{h,s,a}^\star := \arg\max_{\lambda \geq 0} \left\{ -\lambda \log\left( P_{h,s,a}^0 \exp\left( \frac{-V}{\lambda} \right) \right) - \lambda\sigma \right\}, \tag{73b}$$

Lemma 3 (cf. (33)) then gives that

$$\lambda^\star_{h,s,a} \in \left[0, \frac{H}{\sigma}\right], \qquad \widehat{\lambda}^\star_{h,s,a} \in \left[0, \frac{H}{\sigma}\right], \tag{74}$$

due to $\|V\|_\infty \le H$. We shall control (72) 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) $\lambda^\star_{h,s,a} \neq 0$ or $\widehat{\lambda}^\star_{h,s,a} \neq 0$.

**Case (a): $\lambda^\star_{h,s,a} = 0$ and $\widehat{\lambda}^\star_{h,s,a} = 0$.** Applying Lemma 3 and Lemma 4 to (72) gives that, with probability at least $1 - \frac{\delta}{KH}$,

$$\begin{aligned}\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V\right| &\overset{\text{(i)}}{=} \left|\operatorname{essinf}_{s\sim\widehat{P}^0_{h,s,a}} V(s) - \operatorname{essinf}_{s\sim P^0_{h,s,a}} V(s)\right| \\ &\overset{\text{(ii)}}{=} \left|\operatorname{essinf}_{s\sim P^0_{h,s,a}} V(s) - \operatorname{essinf}_{s\sim P^0_{h,s,a}} V(s)\right| \\ &= 0 \le b_h(s,a).\end{aligned} \tag{75}$$

where (i) holds by Lemma 3 (cf. (35)) and (ii) arises from Lemma 4 (cf. (36)) given (69).

**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 (73):

$$\sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\} \ge -\widehat{\lambda}^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right) - \widehat{\lambda}^\star_{h,s,a}\sigma, \tag{76a}$$

$$\sup_{\lambda\ge 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\} \ge -\lambda^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda^\star_{h,s,a}}\right)\right) - \lambda^\star_{h,s,a}\sigma. \tag{76b}$$

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 3 (cf. (35)) and Lemma 4 (cf. (36)) that

$$\begin{aligned}\sup_{\lambda\ge 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\} &\ge \lim_{\lambda\to 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\} \\ &= \operatorname{essinf}_{s\sim\widehat{P}^0_{h,s,a}} V(s) = \operatorname{essinf}_{s\sim P^0_{h,s,a}} V(s) \\ &= \sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\},\end{aligned} \tag{77}$$

leading to

$$\begin{aligned}&\left|\sup_{\lambda\ge 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\} - \sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\}\right| \\ &\overset{\text{(i)}}{\le} \left(-\widehat{\lambda}^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right) - \widehat{\lambda}^\star_{h,s,a}\sigma\right) \\ &\quad - \left(-\widehat{\lambda}^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right) - \widehat{\lambda}^\star_{h,s,a}\sigma\right) \\ &\le \widehat{\lambda}^\star_{h,s,a}\left|\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right) - \log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right)\right|,\end{aligned} \tag{78}$$

where (i) follows from the definition of $\widehat{\lambda}^\star_{h,s,a}$ in (73) and the fact in (76a).

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

$$\forall (h,s,a) \in \mathcal{C}^{\mathsf{b}},\, V \in \mathbb{R}^S: \quad \frac{\left|\left(\widehat{P}^0_{h,s,a} - P^0_{h,s,a}\right)\cdot \exp\left(\frac{-V}{\lambda}\right)\right|}{P^0_{h,s,a}\cdot \exp\left(\frac{-V}{\lambda}\right)} \le \sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}} N_h(s,a) P_{\mathsf{min},h}(s,a)}} \le \frac{1}{2}. \tag{79}$$

The proof is postponed to Appendix C.2.4. With (79) in place, we can further bound (78) (which is plugged into (72)) as

$$\begin{aligned}
\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V\right| &\le \widehat{\lambda}^\star_{h,s,a}\left|\log\left(1 + \frac{\left(\widehat{P}^0_{h,s,a} - P^0_{h,s,a}\right)\cdot \exp\left(\frac{-V}{\lambda}\right)}{P^0_{h,s,a}\cdot \exp\left(\frac{-V}{\lambda}\right)}\right)\right| \\
&\overset{\text{(i)}}{\le} 2\widehat{\lambda}^\star_{h,s,a}\frac{\left|\left(\widehat{P}^0_{h,s,a} - P^0_{h,s,a}\right)\cdot \exp\left(\frac{-V}{\lambda}\right)\right|}{P^0_{h,s,a}\cdot \exp\left(\frac{-V}{\lambda}\right)} \\
&\overset{\text{(ii)}}{\le} \frac{2H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}} N_h(s,a) P_{\mathsf{min},h}(s,a)}} \\
&\le \frac{2eH}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}} N_h(s,a) \widehat{P}_{\mathsf{min},h}(s,a)}} \\
&\le c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{\widehat{P}_{\mathsf{min},h}(s,a) N_h(s,a)}},
\end{aligned} \tag{80}$$

where (i) follows from $\log(1+x) \le 2|x|$ for any $|x| \le \frac{1}{2}$ in view of (79), (ii) follows from (74) as well as (79), and the last line follows from (48) and choosing $c_{\mathrm{b}}$ to be sufficiently large.

Moreover, note that it can be easily verified that

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

due to the assumption $\|V\|_\infty \le H$. Plugging in the definition of $b_h(s,a)$ in (21), combined with the above bounds, we have that with probability at least $1-\delta$,

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V\right| \le \min\left\{c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{N_h(s,a)\widehat{P}_{\mathsf{min},h}(s,a)}},\, H\right\} =: b_h(s,a). \tag{81}$$

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

$$\begin{aligned}
&\left|\sup_{\lambda\ge 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\} - \sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\}\right| \\
&\le \lambda^\star_{h,s,a}\left|\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda^\star_{h,s,a}}\right)\right) - \log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda^\star_{h,s,a}}\right)\right)\right|,
\end{aligned} \tag{82}$$

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

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

$$\left(-\lambda^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot e^{\frac{-V}{\lambda^\star_{h,s,a}}}\right)-\lambda^\star_{h,s,a}\sigma\right)-\left(-\lambda^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot e^{\frac{-V}{\lambda^\star_{h,s,a}}}\right)-\lambda^\star_{h,s,a}\sigma\right)\Bigg\}$$

$$\le \max_{\lambda\in\{\lambda^\star_{h,s,a},\widehat{\lambda}^\star_{h,s,a}\}}\lambda\left|\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)\right)-\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)\right)\right|, \tag{83}$$

where (i) can be verified by applying the facts in (76). Hence, the above term (83) can be controlled again in a similar manner as (78); 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.

### C.2.3 Proof of (69)

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

$$Kd_h^{\mathsf{b},P^0}(s,a)\overset{\text{(i)}}{\ge}\frac{c_1 d_h^{\mathsf{b},P^0}(s,a)\log(KHS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}}P^{\mathsf{b}}_{\mathsf{min}}}\overset{\text{(ii)}}{\ge}\frac{c_1\log(KHS/\delta)}{P^{\mathsf{b}}_{\mathsf{min}}}\overset{\text{(iii)}}{\ge}\frac{c_1\log(KHS/\delta)}{P_{\mathsf{min},h}(s,a)}, \tag{84}$$

where (i) follows from Condition (27), (ii) follows from the definition that $d^{\mathsf{b}}_{\mathsf{min}}\le d_h^{\mathsf{b},P^0}(s,a)$ for $(h,s,a)\in\mathcal{C}^{\mathsf{b}}$, and (iii) comes from (45).

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

$$\begin{aligned}N_h(s,a)&\ge\frac{Kd_h^{\mathsf{b},P^0}(s,a)}{8}-5\sqrt{Kd_h^{\mathsf{b},P^0}(s,a)\log\frac{KH}{\delta}}\\&\ge\frac{Kd_i^{\mathsf{b},P^0}(s,a)}{16}\ge\frac{c_1\log\frac{KH}{\delta}}{16P_{\mathsf{min},h}(s,a)},\end{aligned} \tag{85}$$

where the second line follows from the above relation as long as $c_1$ is sufficiently large. The last inequality of (69) then follows from

$$\frac{c_1\log\frac{KHS}{\delta}}{16P_{\mathsf{min},h}(s,a)}\ge-\frac{\log\frac{2KHS}{\delta}}{\log(1-P_{\mathsf{min},h}(s,a))}, \tag{86}$$

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

### C.2.4 Proof of (79)

Denoting

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

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

$$\begin{aligned}\frac{\left|\left(\widehat{P}^0_{h,s,a}-P^0_{h,s,a}\right)\cdot\exp\left(\frac{-V}{\lambda}\right)\right|}{P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)}&\le\frac{\sum_{s'\in\mathsf{supp}\left(P^0_{h,s,a}\right)}\left|\widehat{P}^0_h(s'\,|\,s,a)-P^0_h(s'\,|\,s,a)\right|\exp\left(\frac{-V(s')}{\lambda}\right)}{\sum_{s'\in\mathsf{supp}\left(P^0_{h,s,a}\right)}P^0_h(s'\,|\,s,a)\exp\left(\frac{-V(s')}{\lambda}\right)}\\&\le\max_{s'\in\mathsf{supp}\left(P^0_{h,s,a}\right)}\frac{\left|\widehat{P}^0_h(s'\,|\,s,a)-P^0_h(s'\,|\,s,a)\right|}{P^0_h(s'\,|\,s,a)},\end{aligned} \tag{87}$$

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 $\mathsf{Binomial}\big(N_h(s,a), P^0_h(s' \,|\, s,a)\big)$. Thus, applying Lemma 5 with $t = \sqrt{\frac{\log\left(\frac{KHS}{\delta}\right)}{c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a)}}$ yields

$$\mathbb{P}\left(\left|\widehat{P}^0_h(s' \,|\, s,a) - P^0_h(s' \,|\, s,a)\right| \geq P^0_h(s' \,|\, s,a)t\right) \leq \exp\left(-c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a) t^2\right) \leq \frac{\delta}{KHS}, \tag{88}$$

as soon as $t \leq \frac{1}{2}$, which can be verified by the fact (69) and $P_{\mathsf{min},h}(s,a) \leq P^0_h(s' \,|\, s,a)$ (cf. (44)), namely,

$$N_h(s,a) \geq \frac{c_1 \log \frac{KHS}{\delta}}{16 P_{\mathsf{min},h}(s,a)} \geq \frac{\log\left(\frac{KHS}{\delta}\right)}{4 c_{\mathsf{f}} P_{\mathsf{min},h}(s,a)} \geq \frac{\log\left(\frac{KHS}{\delta}\right)}{4 c_{\mathsf{f}} P^0_h(s' \,|\, s,a)} \tag{89}$$

as long as $c_1$ is sufficiently large.

Applying (88) and taking the union bound over $s \in \mathsf{supp}\big(P^0_{h,s,a}\big)$ lead to that with probability at least $1 - \frac{\delta}{KH}$,

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

where the last line uses again (89). Plugging this back into (87) and applying the union bound over $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$ then completes the proof.

### C.3 Proof of the lower bound: Theorem 2

The proof of Theorem 2 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. We shall first construct some hard instances and then characterize the sample complexity requirements over these instances.

#### C.3.1 Construction of hard problem instances

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

$$\left\{\mathcal{M}_\phi = \left(\mathcal{S}, \mathcal{A}, P^\phi = \{P^\phi_h\}_{h=1}^H, \{r_h\}_{h=1}^H, H\right) \,|\, \phi = \{0,1\}\right\}, \tag{90}$$

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

$$P^\phi_1(s' \,|\, s,a) = \begin{cases} p\mathbb{1}(s'=0) + (1-p)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,\phi) \\ q\mathbb{1}(s'=0) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,1-\phi) \\ \mathbb{1}(s'=1) & \text{if} \quad s=1 \\ q\mathbb{1}(s'=s) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad s>1 \end{cases} \tag{91a}$$

and

$$P^\phi_h(s' \,|\, s,a) = \mathbb{1}(s'=s), \qquad \forall (h,s,a) \in \{2,\ldots,H\} \times \mathcal{S} \times \mathcal{A}. \tag{91b}$$

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 (91). Here, $p$ and $q$ are set to be

$$p = 1 - \frac{1}{H} + \Delta \qquad \text{and} \qquad q = 1 - \frac{1}{H} \tag{92}$$

for some $H \geq e^8$ and $\Delta$ (whose value will be specified later) obeying

$$\frac{1}{H} \leq \frac{1}{H^{1-3/\beta}} \leq \frac{1}{2} \qquad \text{and} \qquad \Delta \leq \frac{1}{2H}, \tag{93}$$

where $\beta$ is set as

$$\beta := \frac{\log H}{2} \geq 4. \tag{94}$$

The assumption (93) immediately indicates the facts

$$1 > p > q \geq \frac{1}{2}. \tag{95}$$

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

$$r_h(s, a) = \begin{cases} 1 & \text{if } s = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{96}$$

**Construction of the history/batch dataset.** In the nominal environment $\mathcal{M}_\phi$, 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 $\rho^{\mathsf{b}}$ and behavior policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{h=1}^H$:

$$\rho^{\mathsf{b}}(s) = \mu(s) \quad \text{and} \quad \pi_h^{\mathsf{b}}(a \,|\, s) = \frac{1}{2}, \qquad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{97}$$

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

$$\mu(s) = \frac{1}{CS}\mathbb{1}(s = 0) + \Big(1 - \frac{1}{CS}\Big)\mathbb{1}(s = 1), \tag{98}$$

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

$$\frac{1}{CS} \leq \frac{1}{4}. \tag{99}$$

As it turns out, 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:

$$d_1^{\mathsf{b}, P^\phi}(0, a) = \frac{1}{2}\mu(0), \qquad \forall a \in \mathcal{A}, \tag{100a}$$

$$\frac{\mu(s)}{2} \leq d_h^{\mathsf{b}, P^\phi}(s) \leq 2\mu(s), \qquad \frac{\mu(s)}{4} \leq d_h^{\mathsf{b}, P^\phi}(s, a) \leq \mu(s), \qquad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{100b}$$

In addition, we choose the following initial state distribution

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{if } s > 0 \end{cases}. \tag{101}$$

The proof of the claim (100) is postponed to Appendix C.4.1.

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

$$P^\phi_{h,s,a} := P^\phi_h(\cdot \,|\, s, a) \in [0, 1]^{1 \times S}. \tag{102}$$

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

$$\mathcal{U}^\sigma(P^\phi) := \otimes \mathcal{U}^\sigma\left(P^\phi_{h,s,a}\right), \qquad \mathcal{U}^\sigma(P^\phi_{h,s,a}) := \left\{P_{h,s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{h,s,a} \,\|\, P^\phi_{h,s,a}\right) \leq \sigma\right\}, \tag{103}$$

where the radius of the uncertainty set $\sigma$ obeys

$$\left(1-\frac{3}{\beta}\right)\log(H) \le \sigma \le \left(1-\frac{2}{\beta}\right)\log(H). \tag{104}$$

Before continuing, we shall introduce some notation for convenience. For any $P_h^\phi(\cdot\,|\,s,a)$ in (91), we define the limit of the perturbed kernel transiting to the next state $s'$ from the current state-action pair $(s,a)$ by

$$\underline{P}_h^\phi(s'\,|\,s,a) \coloneqq \inf_{P_{h,s,a}\in\mathcal{U}^\sigma(P^\phi_{h,s,a})} P_h(s'\,|\,s,a), \tag{105}$$

and in particular, denote

$$\underline{p} \coloneqq \underline{P}_1^\phi(0\,|\,0,\phi), \qquad \underline{q} \coloneqq \underline{P}_1^\phi(0\,|\,0,1-\phi). \tag{106}$$

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

**Lemma 9.** *When $\beta$ satisfies (94) and the uncertainty level $\sigma$ satisfies (234), the perturbed transition kernels obey*

$$\underline{p} \ge \underline{q} \ge \frac{1}{\beta}. \tag{107}$$

*Proof.* See Appendix C.4.2. □

**Value functions and optimal policies.** We take a moment to derive the corresponding value functions and identify the optimal policies. With some abuse of the notation, 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 $\pi$ (resp. $\pi^{\star,\phi}$) at step $h$ with uncertainty radius $\sigma$. Armed with these notation, we introduce the following lemma which collects the properties concerning the value functions and optimal policies.

**Lemma 10.** *For any $\phi=\{0,1\}$ and any policy $\pi$, defining*

$$z_\phi^\pi \coloneqq \underline{p}\pi_1(\phi\,|\,0) + \underline{q}\pi_1(1-\phi\,|\,0), \tag{108}$$

*it holds that*

$$V_1^{\pi,\sigma,\phi}(0) = 1 + z_\phi^\pi(H-1). \tag{109}$$

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

$$V_1^{\star,\sigma,\phi}(0) = 1+\underline{p}(H-1), \tag{110a}$$

$$\forall h\in[H]\setminus\{1\}: \quad V_h^{\star,\sigma,\phi}(0) = H-h+1, \tag{110b}$$

$$\forall h\in[H]: \quad \pi_h^{\star,\phi}(\phi\,|\,0)=1, \quad \pi_h^{\star,\phi}(\phi\,|\,1)=1, \qquad V_h^{\star,\sigma,\phi}(1)=0. \tag{110c}$$

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

$$2C \le C^\star_{\mathsf{rob}} \le 4C. \tag{111}$$

*Proof.* See Appendix C.4.3. □

In view of Lemma 10, 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

$$P^\star_{\mathsf{min}} \coloneqq \min_{h,s,s'}\left\{P_h^\phi\left(s'|s,\pi_h^{\star,\phi}(s)\right):\ P_h^\phi\left(s'|s,\pi_h^{\star,\phi}(s)\right)>0\right\} = P_1^\phi\left(1|0,\phi\right) = 1-p. \tag{112}$$

### C.3.2 Establishing the minimax lower bound

We are now ready to establish the sample complexity lower bound. With the choice of the initial distribution $\rho$ in (101), 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 $\phi$.** We make the following claim which shall be verified in Appendix C.4.4: given $\varepsilon \leq \frac{H}{256e^6 \log H}$, choosing

$$\Delta = \frac{128e^6\sigma(1-q)\varepsilon}{H} \leq \frac{\sigma}{2H\log H} \leq \frac{1}{2H}, \tag{113}$$

which satisfies (93) with the aid of (234) and (92), it holds that for any policy $\widehat{\pi}$,

$$\big\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\widehat{\pi},\sigma,\phi} \big\rangle \geq 2\varepsilon\big(1 - \widehat{\pi}_1(\phi\,|\,0)\big). \tag{114}$$

Armed with this relation between the policy $\widehat{\pi}$ and its sub-optimality gap, we are positioned to construct an estimate of $\phi$. 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

$$\mathbb{P}_\phi\left\{\big\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\widehat{\pi},\sigma,\phi} \big\rangle \leq \varepsilon\right\} \geq \frac{7}{8}, \tag{115}$$

then in view of (114), 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 $\phi \in \{0,1\}$:

$$\widehat{\phi} = \arg\max_{a\in\{0,1\}} \widehat{\pi}_1(a\,|\,0), \tag{116}$$

which obeys

$$\mathbb{P}_\phi\big\{\widehat{\phi} = \phi\big\} \geq \mathbb{P}_\phi\big\{\widehat{\pi}_1(\phi\,|\,0) > 1/2\big\} \geq \frac{7}{8}. \tag{117}$$

In what follows, we would like to show (117) cannot happen without enough samples, which would in turn contradict (114).

**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:

$$p_{\mathrm{e}} \coloneqq \inf_{\psi} \max\big\{\mathbb{P}_0(\psi \neq 0),\, \mathbb{P}_1(\psi \neq 1)\big\}, \tag{118}$$

where the infimum is taken over all possible tests $\psi$ 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 & Zaiats (2009, Theorem 2.2) and the additivity of the KL divergence (cf. Tsybakov & Zaiats (2009, Page 85)), we obtain

$$\begin{aligned} p_{\mathrm{e}} &\geq \frac{1}{4}\exp\Big(-K\mathsf{KL}\big(\mu^{\mathsf{b},0} \,\|\, \mu^{\mathsf{b},1}\big)\Big) \\ &\geq \frac{1}{4}\exp\bigg\{-\frac{1}{2}K\mu(0)\Big(\mathsf{KL}\big(P_1^0(\cdot\,|\,0,0) \,\|\, P_1^1(\cdot\,|\,0,0)\big) + \mathsf{KL}\big(P_1^0(\cdot\,|\,0,1) \,\|\, P_1^1(\cdot\,|\,0,1)\big)\Big)\bigg\}, \end{aligned} \tag{119}$$

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 (100).

It remains to control the KL divergence terms in (119). Given $p \geq q \geq 1/2$ (cf. (95)), applying Lemma 7 (cf. (41)) yields

$$\begin{aligned}\mathsf{KL}\big(P_1^0(\cdot\,|\,0,0)\;\|\;P_1^1(\cdot\,|\,0,0)\big) = \mathsf{KL}\left(p\;\|\;q\right) \leq \frac{(p-q)^2}{(1-p)p} &\overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\ &\overset{\text{(ii)}}{=} \frac{128^2 e^{12}\sigma^2(1-q)^2\varepsilon^2}{H^2 p(1-p)} \\ &\overset{\text{(iii)}}{\leq} \frac{c_1\sigma^2 P_{\mathsf{min}}^\star \varepsilon^2}{H^2},\end{aligned} \tag{120}$$

where (i) follows from the definition (92), (ii) holds by plugging in the expression of $\Delta$ in (113), (iii) arises from $1-q \leq 2(1-p) = 2P_{\mathsf{min}}^\star$ (see (93) and (112)), $p > \frac{1}{2}$, 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 (120) back into (119) demonstrates that: if the sample size is chosen as

$$KH \leq \frac{H^3 S C_{\mathsf{rob}}^\star \log 2}{4c_1 P_{\mathsf{min}}^\star \sigma^2 \varepsilon^2}, \tag{121}$$

then one necessarily has

$$\begin{aligned}p_{\mathrm{e}} &\geq \frac{1}{4}\exp\left\{-\frac{1}{2}K\mu(0)\cdot 2\frac{c_1\sigma^2 P_{\mathsf{min}}^\star\varepsilon^2}{H^2}\right\} \overset{\text{(i)}}{=} \frac{1}{4}\exp\left\{-K\frac{c_1\sigma^2 P_{\mathsf{min}}^\star\varepsilon^2}{SCH^2}\right\} \\ &\overset{\text{(ii)}}{\geq} \frac{1}{4}\exp\left\{-K\frac{4c_1\sigma^2 P_{\mathsf{min}}^\star\varepsilon^2}{SC_{\mathsf{rob}}^\star H^2}\right\} \geq \frac{1}{8},\end{aligned} \tag{122}$$

where (i) follows from (98) and (ii) holds by (111).

**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 (116) 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 (121) to avoid contradition with (122). The proof is thus finished.

### C.4 Proof of auxiliary results

#### C.4.1 Proof of (100)

With the initial state distribution and behavior policy defined in (97), 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

$$\forall a \in \mathcal{A}: \quad d_1^{\mathsf{b},P^\phi}(0,a) = \frac{1}{2}\mu(0). \tag{123}$$

In view of (91a), 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

$$d_2^{\mathsf{b},P^\phi}(1) = \mathbb{P}\left\{s_2 = 1\,|\,s_1 \sim d_1^{\mathsf{b},P^\phi};\pi^{\mathsf{b}}\right\}$$

$$= \mu(0)\left[\pi_1^{\mathsf{b}}(\phi\,|\,0)(1-p) + \pi_1^{\mathsf{b}}(1-\phi\,|\,0)(1-q)\right] + \mu(1) = \mu(1) + \frac{(2-p-q)\mu(0)}{2}.$$

With the above result in mind and recalling the assumption in (95), we arrive at

$$\frac{\mu(0)}{2} \le d_2^{\mathsf{b},P^\phi}(0) \le \mu(0), \qquad \mu(1) \le d_2^{\mathsf{b},P^\phi}(1) \overset{\text{(i)}}{\le} 2\mu(1), \tag{124}$$

where (i) holds by applying (95) and (99) (which implies $\mu(0) \le \mu(1)$ by the assumption in (99))

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

Finally, from the definitions of $P_h^\phi(\cdot\,|\,s,a)$ in (91b) and the Markov property, we arrive at for any $(h,s) \in [H] \times \mathcal{S}$,

$$\frac{\mu(s)}{2} \le d_h^{\mathsf{b},P^\phi}(s) \le 2\mu(s), \tag{125}$$

which directly leads to

$$\frac{\mu(s)}{4} \le d_h^{\mathsf{b},P^\phi}(s,a) = \pi_1^{\mathsf{b}}(a\,|\,s) d_h^{\mathsf{b},P^\phi}(s) \le \mu(s). \tag{126}$$

### C.4.2 Proof of Lemma 9

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 (107). Towards this, invoking the definition in (40), let $\sigma'$ be the KL divergence from $\mathsf{Ber}\left(\frac{1}{\beta}\right)$ to $\mathsf{Ber}(q)$, defined as follows

$$\begin{aligned}\sigma' &:= \mathsf{KL}\left(\mathsf{Ber}\left(\frac{1}{\beta}\right) \,\|\, \mathsf{Ber}(q)\right) = \frac{1}{\beta}\log\frac{\frac{1}{\beta}}{q} + \left(1-\frac{1}{\beta}\right)\log\frac{\left(1-\frac{1}{\beta}\right)}{1-q} \\ &= \left(\frac{1}{\beta}\right)\log\left(\frac{1}{\beta}\right) - \left(\frac{1}{\beta}\right)\log(q) + \left(1-\frac{1}{\beta}\right)\log(H) + \left(1-\frac{1}{\beta}\right)\log\left(1-\frac{1}{\beta}\right),\end{aligned} \tag{127}$$

where the second line uses the definition of $q$ in (92). We claim that $\sigma'$ satisfies the following relation with $\sigma$, which will be proven at the end of this proof:

$$\left(1-\frac{3}{\beta}\right)\log(H) \le \sigma \le \left(1-\frac{2}{\beta}\right)\log(H) \le \sigma' \le \left(1-\frac{1}{\beta}\right)\log(H). \tag{128}$$

Recalling the definition of the transition kernel in (91a)

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

the uncertainty set of the transition kernel with radius $\sigma$ is thus given as

$$\begin{aligned}&\mathcal{U}^\sigma(P^\phi_{1,0,1-\phi}) \\ &= \{P_{1,0,1-\phi} \in \Delta(\mathcal{S}) : P(0\,|\,0,1-\phi) = q', P(1\,|\,0,1-\phi) = 1-q', \mathsf{KL}\left(\mathsf{Ber}\left(q'\right) \,\|\, \mathsf{Ber}(q)\right) \le \sigma\}.\end{aligned} \tag{129}$$

Recalling the definition of $\underline{q}$ in (106), we can bound

$$\begin{aligned}\underline{q} &= \inf_{P_{1,0,1-\phi} \in \mathcal{U}^\sigma(P^\phi_{1,0,1-\phi})} P(0\,|\,0,1-\phi) = \inf_{q':\mathsf{KL}(\mathsf{Ber}(q')\|\mathsf{Ber}(q)) \le \sigma} q' \\ &\overset{\text{(i)}}{\ge} \inf_{q':\mathsf{KL}(\mathsf{Ber}(q')\|\mathsf{Ber}(q)) \le \sigma'} q' = \frac{1}{\beta},\end{aligned}$$

where (i) holds by $\sigma \le \sigma'$ (cf. (128)) and the last equality follows from applying Lemma 7 (cf. (42)) and (127) to arrive at

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

**Proof of (128).** To control $\sigma'$, we plug in the assumptions in (95) 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

$$\sigma' \leq \left(1-\frac{1}{\beta}\right)\log(H). \tag{130}$$

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

$$\sigma' \geq -1 + \left(1-\frac{1}{\beta}\right)\log(H) \geq \left(1-\frac{2}{\beta}\right)\log(H) \tag{131}$$

as long as $\log H \geq \beta$ (cf. (94)). With (130) and (131) in hand, it is straightforward to see that the choice of the uncertainty radius $\sigma$ in (234) obeys the advertised bound (128).

### C.4.3 Proof of Lemma 10

For notational conciseness, we shall drop the superscript $\phi$ 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 $\pi$. 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 (91a), 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

$$V_h^{\star,\sigma}(1) = V_h^{\pi,\sigma}(1) = 0, \qquad \forall h \in [H]. \tag{132}$$

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 (91b), 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

$$V_h^{\star,\sigma}(0) = V_h^{\pi,\sigma}(0) = H - h + 1, \qquad 2 \leq h \leq H. \tag{133}$$

Taking (132) and (133) together, it follows that

$$\forall\ 2 \leq h \leq H: \qquad V_h^{\pi,\sigma}(0) > V_h^{\pi,\sigma}(1). \tag{134}$$

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

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

$$\begin{aligned}
&\overset{\text{(ii)}}{=} 1+\pi_1(\phi\,|\,0)\Big[\underline{p}V_2^{\pi,\sigma}(0)+\big(1-\underline{p}\big)\,V_2^{\pi,\sigma}(1)\Big]\\
&\qquad+\pi_1(1-\phi\,|\,0)\Big[\underline{q}V_2^{\pi,\sigma}(0)+\big(1-\underline{q}\big)\,V_2^{\pi,\sigma}(1)\Big]\\
&\overset{\text{(iii)}}{=} 1+V_2^{\pi,\sigma}(1)+z_\phi^\pi\left[V_2^{\pi,\sigma}(0)-V_2^{\pi,\sigma}(1)\right]\\
&=1+z_\phi^\pi V_2^{\pi,\sigma}(0)
\end{aligned}\tag{135}$$

where (i) uses the definition of the reward function in (96), (ii) uses (134) so that the infimum is attained by picking the choice specified in (106) with a smallest probability mass imposed on the transition to state 0. Finally, we plug in the definition (108) of $z_\phi^\pi$ in (iii), and the last line follows from (132).

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

$$V_1^{\star,\sigma}(0)=1+z_\phi^{\pi^{\star,\phi}}V_2^{\star,\sigma}(0)=1+z_\phi^{\pi^{\star,\phi}}(H-1),\tag{136}$$

where the second equality follows from (133). Observing that the function $(H-1)z$ is increasing in $z$ and that $z_\phi^\pi$ is increasing in $\pi_1(\phi\,|\,0)$ (due to the fact $\underline{p}\geq\underline{q}$ in (107)). As a result, the optimal policy obeys

$$\pi_1^{\star,\phi}(\phi\,|\,0)=1\tag{137}$$

at state 0, and plugging back to (136) 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

$$\forall h\in[H]:\qquad \pi_h^{\star,\phi}(\phi\,|\,0)=1,\quad \pi_h^{\star,\phi}(\phi\,|\,1)=1.\tag{138}$$

**Proof of claim (111).** 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

$$\begin{aligned}
d_2^{\star,P}(0,\phi)&=d_2^{\star,P}(0)\pi_2^{\star,\phi}(\phi\,|\,0)=d_2^{\star,P}(0)=\mathbb{P}_{s_2\sim P(\cdot\,|\,s_1,\pi_1^{\star,\phi}(s_1))}\big\{s_2=0\,|\,s_1\sim\rho;\pi^{\star,\phi}\big\}\\
&=P_1(0\,|\,0,\phi)\overset{\text{(i)}}{\geq}\underline{P}_1^\phi(0\,|\,0,\phi)\overset{\text{(ii)}}{=}\underline{p}\geq\frac{1}{\beta},
\end{aligned}\tag{139}$$

which (i) holds by plugging in the definition (105), (ii) follows from the definition (106), and the final inequality arises from Lemma 9. Hence, for all $2\leq h\leq H$, by the Markov property and $P_h^\phi(0\,|\,0,\phi)=1$, we have

$$d_h^{\star,P}(0,\phi)=d_2^{\star,P}(0,\phi)\geq\frac{1}{\beta}.\tag{140}$$

Examining the definition of $C_{\mathsf{rob}}^\star$ in (12), we make the following observations.

- For $h=1$, we have

$$\begin{aligned}
\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)}\frac{\min\left\{d_1^{\star,P}(s,a),\frac{1}{S}\right\}}{d_1^{\mathsf{b},P^\phi}(s,a)}&\overset{\text{(i)}}{=}\max_{P\in\mathcal{U}^\sigma(P^\phi)}\frac{\min\left\{d_1^{\star,P}(0,\phi),\frac{1}{S}\right\}}{d_1^{\mathsf{b},P^\phi}(0,\phi)}\\
&\overset{\text{(ii)}}{=}\max_{P\in\mathcal{U}^\sigma(P^\phi)}\frac{1}{Sd_1^{\mathsf{b},P^\phi}(0,\phi)}\overset{\text{(iii)}}{=}\frac{2}{S\mu(0)}=2C,
\end{aligned}\tag{141}$$

  where (i) holds by $d_1^{\star,P}(s)=\rho(s)=0$ for all $s\in\mathcal{S}\setminus\{0\}$ (see (101)) 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 (100), and the last equality arises from the definition in (98).

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

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d_2^{\star,P}(s,a),\frac{1}{S}\right\}}{d_2^{\mathsf{b},P^\phi}(s,a)} \overset{\text{(i)}}{=} \max_{s\in\{0,1\},P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d_2^{\star,P}(s,\phi),\frac{1}{S}\right\}}{d_2^{\mathsf{b},P^\phi}(s,\phi)}$$
$$\leq \max_{s\in\{0,1\},P\in\mathcal{U}^\sigma(P^\phi)} \frac{1}{Sd_2^{\mathsf{b},P^\phi}(s,\phi)} \overset{\text{(ii)}}{\leq} \frac{4}{S\mu(0)} = 4C, \tag{142}$$

  where (i) holds by the optimal policy in (110) and the trivial fact that $d_2^{\star,P}(s) = 0$ for all $s \in \mathcal{S}\setminus\{0,1\}$ (see (101) and (91a)), (ii) arises from (100), and the last equality comes from (98).

- For all other steps $h = 3,\ldots,H$, observing from the deterministic transition kernels in (91b), it can be easily verified that

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d_h^{\star,P}(s,a),\frac{1}{S}\right\}}{d_h^{\mathsf{b},P^\phi}(s,a)} = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d_2^{\star,P}(s,a),\frac{1}{S}\right\}}{d_2^{\mathsf{b},P^\phi}(s,a)} \leq 4C. \tag{143}$$

Combining the above cases, we complete the proof by

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

### C.4.4 Proof of the claim (114)

Recall that by virtue of (108) and (110), we arrive at

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

Applying (109) yields

$$\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\pi,\sigma,\phi}\right\rangle = V_h^{\star,\sigma,\phi}(0) - V_h^{\pi,\sigma,\phi}(0) = \left(\underline{p} - z_\phi^\pi\right)(H-1) = \left(\underline{p}-\underline{q}\right)(H-1)\left(1-\pi_1(\phi\,|\,0)\right), \tag{144}$$

where the last equality uses the definition (108). 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

$$\overline{V}(0) = H-1 \quad \text{and} \quad \overline{V}(s) = 0, \quad \forall s\in\mathcal{S}\setminus\{0\}. \tag{145}$$

With this in hand, applying Lemma 2 gives

$$\begin{aligned}
(H-1)\left(\underline{p}-\underline{q}\right) &\overset{\text{(i)}}{=} \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{1,0,\phi})} \mathcal{P}\overline{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{1,0,1-\phi})} \mathcal{P}\overline{V} \\
&= \sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^\phi_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right\} - \sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^\phi_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right\} \\
&\overset{\text{(ii)}}{\geq} \left\{-\lambda^\star\log\left(P^\phi_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right) - \lambda^\star\sigma\right\} - \left\{-\lambda^\star\log\left(P^\phi_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right) - \lambda^\star\sigma\right\} \\
&= -\lambda^\star\left[\log\left(P^\phi_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right) - \log\left(P^\phi_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right)\right],
\end{aligned} \tag{146}$$

where (i) follows from (see the definition of $\underline{p}$ in (106))

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

Here, (ii) holds by letting

$$\lambda^\star := \arg\max_{\lambda\geq 0} f(\lambda) := \arg\max_{\lambda\geq 0}\left\{-\lambda\log\left(P^\phi_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right\}. \tag{147}$$

The rest of the proof is then to control (146). We start with the observation that $\lambda^\star > 0$; this is because in view of Lemma 3 (cf. (34)), it suffices to verify that

$$\log(1-q)+\sigma \overset{\text{(i)}}{\leq} \log\left(\frac{1}{H}\right)+\left(1-\frac{2}{\beta}\right)\log H = -\frac{2}{\beta}\log H < 0, \tag{148}$$

where (i) holds by (234). We now claim the following bound for $\lambda^\star$ holds, whose proof is postponed to the end:

$$\frac{H}{16\sigma} \leq \frac{H-1}{\log(\beta H)} \leq \lambda^\star \leq \frac{H-1}{\left(1-\frac{3}{\beta}\right)\log(H)}, \tag{149}$$

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

$$\frac{1}{\beta H} \leq e^{-(H-1)/\lambda^\star} \leq \frac{1}{H^{1-3/\beta}}. \tag{150}$$

Moving to the second term of (146), it follows that

$$\begin{aligned}
&\log\left(P^{\phi}_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right)-\log\left(P^{\phi}_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^\star}\right)\right)\\
&\overset{\text{(i)}}{=}\log\frac{pe^{-(H-1)/\lambda^\star}+(1-p)}{qe^{-(H-1)/\lambda^\star}+(1-q)}\\
&=\log\left(1+\frac{(p-q)\left(e^{-(H-1)/\lambda^\star}-1\right)}{qe^{-(H-1)/\lambda^\star}+(1-q)}\right)\\
&\overset{\text{(ii)}}{<}-\frac{\Delta\left(1-e^{-(H-1)/\lambda^\star}\right)}{qe^{-(H-1)/\lambda^\star}+(1-q)}\\
&\overset{\text{(iii)}}{\leq}-\frac{1}{2}\frac{\Delta}{H^{\frac{3}{\beta}}(1-q)+(1-q)}\\
&\leq-\frac{\Delta}{4e^6(1-q)},
\end{aligned} \tag{151}$$

where (i) follows from the definitions in (91) and (145), (ii) holds by $\log(1+x) < x$ for $x \in (-1,\infty)$, (iii) can be verified by (150) and (93):

$$1-e^{-(H-1)/\lambda^\star} \geq 1-\frac{1}{H^{1-3/\beta}} \geq \frac{1}{2},$$

and the last line uses $H^{3/\beta} = H^{6/\log H} = e^6$ by the definition of $\beta$ in (94). Plugging (149) and (151) back into (146) and (144), we arrive at

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

where (i) holds by the definition of $\beta$ in (94) and the last inequality follows directly from the choice of $\Delta$ in (113).

**Proof of inequality (149).** Applying (33) in Lemma 3 to $\lambda^\star$ in (147) leads to the upper bound in (149):

$$\lambda^\star \leq \frac{H-1}{\sigma} \leq \frac{H-1}{\left(1-\frac{3}{\beta}\right)\log(H)}, \tag{152}$$

where the last inequality holds by (234). As a result, we shall focus on showing the lower bounds in (149) in the remainder of the proof.

Recalling the definition of $q$ in (92), we can reparameterize $1-q$ using two positive variables $c_q$ and $\lambda_q$ (whose choices will be made clearer soon) as follows:

$$1-q=\frac{1}{H}=c_q e^{-(H-1)/\lambda_q}. \tag{153}$$

Deriving the first derivative of the function of interest $f(\lambda)$ in (147) as follows:

$$\begin{aligned}\nabla_\lambda f(\lambda) &= \nabla_\lambda \left(-\lambda \log\left(P^{\phi}_{1,0,1-\phi}\cdot \exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right)\\ &\overset{\text{(i)}}{=} \nabla_\lambda \left(-\lambda \log\left(qe^{-(H-1)/\lambda}+1-q\right) - \lambda\sigma\right)\\ &= -\sigma - \log\left(qe^{-(H-1)/\lambda}+1-q\right) - \frac{1}{\lambda}\cdot\frac{q(H-1)e^{-(H-1)/\lambda}}{qe^{-(H-1)/\lambda}+1-q},\end{aligned} \tag{154}$$

where (i) holds by the chosen transition kernels in (91) 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

$$\begin{aligned}\nabla_\lambda f(\lambda)\,|_{\,\lambda=\lambda_q} &= -\sigma - \log\left((1-q)\frac{q}{c_q}+1-q\right) + \frac{(1-q)\frac{q}{c_q}\log\frac{1-q}{c_q}}{(1-q)\frac{q}{c_q}+1-q}\\ &= -\sigma - \log(1-q) - \log\left(1+\frac{q}{c_q}\right) + \frac{\frac{q}{c_q}\log\frac{1-q}{c_q}}{\frac{q}{c_q}+1}\\ &= -\sigma - \log(1-q)\left(1-\frac{q/c_q}{q/c_q+1}\right) - \log\left(1+\frac{q}{c_q}\right) - \frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q}\\ &\overset{\text{(i)}}{=} -\sigma + \log H\left(1-\frac{q/c_q}{q/c_q+1}\right) - \log\left(1+\frac{q}{c_q}\right) - \frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q}\\ &\overset{\text{(ii)}}{\geq} \log H\left(\frac{2}{\beta}-\frac{q/c_q}{q/c_q+1}\right) - \log\left(1+\frac{q}{c_q}\right) - \frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q}\\ &\overset{\text{(iii)}}{\geq} \frac{1}{\beta}\log H - \log(1+\frac{1}{\beta}) - 1\\ &\geq \frac{1}{\beta}\log H - 2 = 0,\end{aligned} \tag{155}$$

$$\tag{156}$$

where (i) holds by (153), (ii) follows from the bound of $\sigma$ in (234), (iii) arises from letting $c_q = \beta \geq 4$ and noting the fact $1/2 \leq q < 1$ (see (95)), leading to

$$\frac{1}{2\beta} \leq \frac{q}{c_q} < \frac{1}{\beta}, \qquad \frac{q/c_q}{q/c_q+1} \leq \frac{1}{\beta}, \qquad \frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q} < 1. \tag{157}$$

Finally, the last line holds by $1/\beta \leq \frac{1}{4}$ and $\log H = 2\beta$ (see (94)).

To proceed, note that the function $f(\lambda)$ is concave with respect to $\lambda$. Therefore, observing $\nabla_\lambda f(\lambda)\,|_{\,\lambda=\lambda_q} \geq 0$ with $c_q = \beta$, we have $\lambda_q \leq \lambda^\star$, which implies (see (153))

$$1-q = \frac{1}{H} = \beta e^{-(H-1)/\lambda_q} \leq \beta e^{-(H-1)/\lambda^\star}. \tag{158}$$

The above assertion directly gives

$$\lambda^\star \geq \frac{H-1}{\log(\beta H)}.$$

The proof is completed by noticing

$$\frac{H-1}{\log(\beta H)} = \frac{H-1}{\log(H)+\log\beta} \overset{\text{(i)}}{\geq} \frac{H-1}{2\log H} \geq \frac{\left(1-\frac{3}{\beta}\right)(H-1)}{2\sigma} \geq \frac{H}{16\sigma},$$

where (i) follows from (94), the penultimate inequality follows from (234), and the last inequality follows from $\beta \in [4,\infty)$.

## D PROBLEM FORMULATION: DISCOUNTED INFINITE-HORIZON RMDPS

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

### D.1 BASICS ABOUT DISCOUNTED INFINITE-HORIZON MDPS

A discounted infinite-horizon MDP can be denoted by $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \gamma, P, r\}$. 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, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ represent the transition kernel of the MDP, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the intermediate reward function.

**Policy, value/Q function and occupancy distribution.** 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 $\pi$ when $\pi$ is a deterministic policy. With this in mind, let $\rho$ 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 $\pi$, namely

$$\forall s \in \mathcal{S}: \qquad d^{\pi,P}(s) \coloneqq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \,|\, s_0 \sim \rho, \pi, P), \tag{159a}$$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad d^{\pi,P}(s, a) \coloneqq \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \,|\, s_0 \sim \rho, \pi, P)\, \pi(a \,|\, s). \tag{159b}$$

Here, the occupancy distributions are conditioned on $s_0 \sim \rho$ and the sequence of actions and states are generated based on policy $\pi$ and transition kernel $P$.

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

$$\forall s \in \mathcal{S}: \qquad V^{\pi,P}(s) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\Big|\, s_0 = s\right], \tag{160}$$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \qquad Q^{\pi,P}(s, a) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\Big|\, s_0 = s, a_0 = a\right], \tag{161}$$

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

### D.2 DISTRIBUTIONALLY ROBUST DISCOUNTED INFINITE-HORIZON MDPS

Before continuing, we introduce the vector of a transition kernel $P$ at $(s, a)$ as

$$P_{s,a} \coloneqq P(\cdot \,|\, s, a) \in \mathbb{R}^{1 \times S}, \tag{162}$$

which is used throughout Appendix D to Appendix F.

In this work, instead of the standard MDP introduced above, 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, $\mathcal{U}^\sigma(P^0)$ denote the set of possible transition kernels within an uncertainty set centered around a nominal kernel $P^0$ 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

$$\mathcal{U}^\sigma(P^0) \coloneqq \otimes\, \mathcal{U}^\sigma(P^0_{s,a}), \qquad \mathcal{U}^\sigma(P^0_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{s,a} \,\|\, P^0_{s,a}\right) \le \sigma\right\}. \tag{163}$$

Armed with this, we define the *robust value function* $V^{\pi,\sigma}$ and the *robust Q-function* $Q^{\pi,\sigma}$ in the discounted infinite-horizon setting respectively as

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

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** 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 & 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

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) \coloneqq V^{\pi^\star,\sigma}(s) = \max_{\pi} V^{\pi,\sigma}(s), \tag{164a}$$

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad Q^{\star,\sigma}(s,a) \coloneqq Q^{\pi^\star,\sigma}(s,a) = \max_{\pi} Q^{\pi,\sigma}(s,a). \tag{164b}$$

Next, applying (159) with $\pi = \pi^\star$, we adopt the the following short-hand notation for the occupancy distributions associated with the optimal policy:

$$\forall s \in \mathcal{S}: \qquad d^{\star,P}(s) \coloneqq d^{\pi^\star,P}(s), \tag{165a}$$

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad d^{\star,P}(s,a) \coloneqq d^{\pi^\star,P}(s,a) = d^{\star,P}(s)\,\mathbb{1}\{a = \pi^\star(s)\}. \tag{165b}$$

In addition, the Bellman's optimality principle can also be extended to the infinite-horizon robust MDPs, which is essential. Specifically, for any policy $\pi$ (resp. optimal policy $\pi^\star$), the robust value function and Q-function obey the following *robust Bellman consistency equation* (resp. *robust Bellman optimality equation*):

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad Q^{\pi,\sigma}(s,a) = r(s,a) + \gamma \inf_{P\in\mathcal{U}^\sigma(P^0)} V^{\pi,\sigma}, \tag{166}$$

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{P\in\mathcal{U}^\sigma(P^0)} V^{\star,\sigma}. \tag{167}$$

### D.3 Distributionally robust offline RL

We observe a batch dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{1\le i\le N}$ consisting of $N$ sample transitions. These transitions are independently generated based on some behavior distribution $d^{\mathsf{b}} \in \Delta(\mathcal{S}\times\mathcal{A})$ over the nominal transition kernel $P^0$, i.e.,

$$(s_i, a_i) \overset{\text{i.i.d.}}{\sim} d^{\mathsf{b}} \quad \text{and} \quad s_i' \overset{\text{i.i.d.}}{\sim} P^0(\cdot \,|\, s_i, a_i), \qquad 1 \le i \le N. \tag{168}$$

Similar to Assumption 1, we design the following *robust single-policy clipped concentrability* 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*

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b},P^0}(s,a)} \le C^\star_{\mathsf{rob}} \tag{169}$$

*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 (169), and refer to it as the robust single-policy clipped concentrability coefficient.*

Armed with these, we are ready to introduce the goal in the infinite-horizon case. 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

$$V^{\widehat{\pi},\sigma}(\rho) \ge V^{\star,\sigma}(\rho) - \varepsilon \tag{170}$$

in a sample-efficient manner.

## E Algorithm and theory: discounted infinite-horizon RMDPs

In this section, we present both the model-based algorithm —DRVI-LCB—for robust offline RL and its performance guarantees for the discounted infinite-horizon setting.

### E.1 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 $(s,a)$ as

$$N(s,a) \coloneqq \sum_{i=1}^{N} \mathbb{1}\big\{(s_i,a_i) = (s,a)\big\}. \tag{171}$$

**Algorithm 2:** Robust value iteration with LCB (DRVI-LCB) for infinite-horizon RMDPs.

**input:** a dataset $\mathcal{D}$; reward function $r$; uncertainty level $\sigma$.
**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 (172);
Compute the penalty term $b(s,a)$ according to (176);
**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 (179);
  **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 $s\in\mathcal{S}$.

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:

$$\widehat{P}^0(s'\,|\,s,a) := \begin{cases} \frac{1}{N(s,a)}\sum_{i=1}^{N}\mathbb{1}\big\{(s_i,a_i,s_i')=(s,a,s')\big\}, & \text{if } N(s,a)>0 \\ 0, & \text{else} \end{cases} \tag{172}$$

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

### E.2 Algorithm: DRVI-LCB for 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 taking the uncertainty into consideration by incorporating some penalty term inside the value estimation, summarized in Algorithm 2.

**The pessimistic robust Bellman operator.** To begin with, recall the classical distributionally robust Bellman operator (Zhou et al., 2021; Iyengar, 2005; Nilim & El Ghaoui, 2005) in the infinite-horizon case $\mathcal{T}^\sigma(\cdot):\mathbb{R}^{SA}\to\mathbb{R}^{SA}$,

$$\begin{aligned} \forall(s,a)\in\mathcal{S}\times\mathcal{A}: \quad & \mathcal{T}^\sigma(Q)(s,a) := r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}V, \\ \text{where} \quad & V=[V(s)]_{s\in\mathcal{S}}, \quad V(s):=\max_a Q(s,a). \end{aligned} \tag{173}$$

The contraction property of the above robust Bellman operator plays a fundamental role in the convergence and performance of prior works, e.g. Yang et al. (2021); Zhou et al. (2021); Panaganti & Kalathil (2022). In view of this, to apply the pessimistic principle in RMDPs, we define 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:

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \quad \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q)(s,a)=\max\left\{r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}V-b\big(s,a\big),\,0\right\}, \tag{174}$$

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

To specify the tailored penalty term $b(s,a)$ in (174), we first introduce an additional term

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \quad \widehat{P}_{\min}(s,a):=\min_{s'}\Big\{\widehat{P}^0(s'\,|\,s,a):\ \widehat{P}^0(s'\,|\,s,a)>0\Big\}, \tag{175}$$

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 $c_{\mathrm{b}}>0$ and any $(s,a)\in\mathcal{S}\times\mathcal{A}$, $b(s,a)$ is defined as

$$b(s,a)=\begin{cases} \min\left\{\frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\min}(s,a)N(s,a)}}+\frac{4}{\sigma N(1-\gamma)}\ ,\ \frac{1}{1-\gamma}\right\}+\frac{2}{\sigma N} & \text{if } N(s,a)>0, \\ \frac{1}{1-\gamma}+\frac{2}{\sigma N} & \text{otherwise.} \end{cases} \tag{176}$$

Our proposed pessimistic robust Bellman operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ (cf. (174)) will play an important role in DRVI-LCB. Encouragingly, although the pessimistic robust Bellman operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ involves an additional penalty term $b(s,a)$ compared to the classical robust Bellman operator $\mathcal{T}^{\sigma}(\cdot)$ used in DRVI (Zhou et al., 2021), it still enjoys the celebrated $\gamma$-contractive property. Before continuing, we summarize the contraction property below, whose proof is postponed to Appendix F.3.1.

**Lemma 11** ($\gamma$-Contraction). *For any $\gamma \in \left[\frac{1}{2}, 1\right)$, the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ (cf. (174)) is a $\gamma$-contraction w.r.t. $\|\cdot\|_{\infty}$. Namely, for any $Q_1, Q_2 \in \mathbb{R}^{SA}$ s.t. $Q_1(s,a), Q_2(s,a) \in \left[0, \frac{1}{1-\gamma}\right]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, one has*

$$\left\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q_2)\right\|_{\infty} \leq \gamma \left\|Q_1 - Q_2\right\|_{\infty}. \tag{177}$$

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

**Our algorithm DRVI-LCB for infinite-horizon robust offline RL.** Armed with the $\gamma$-contraction of the pessimistic robust Bellman operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$, we are positioned to introduce DRVI-LCB for infinite-horizon RMDPs, summarized in Algorithm 2. Specifically, DRVI-LCB can be seen as a value iteration algorithm w.r.t. $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ (cf. (174)), whose update rule at the $m$-th iteration can be formulated as

$$\widehat{Q}_m(s,a) = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_{m-1})(s,a) = \max\left\{r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}_{m-1} - b\big(s,a\big),\, 0\right\} \tag{178}$$

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

$$\widehat{Q}_m(s,a) = \max\left\{r(s,a) + \sup_{\lambda \geq 0}\left\{-\lambda \log\left(\widehat{P}^0_{s,a} \cdot \exp\left(\frac{-\widehat{V}_{m-1}}{\lambda}\right)\right) - \lambda\sigma\right\} - b\big(s,a\big),\, 0\right\}, \tag{179}$$

which can be solved efficiently (Iyengar, 2005; Yang et al., 2021; Panaganti & Kalathil, 2022). The computational efficiency is due to the fact that the variable to be optimized is a scalar and the dimension of it won't explode as the size of the state space $S$ increases (independent w.r.t. the size of the state-action space).

To continue, 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,

$$\widehat{\pi}(s) = \arg\max_{a} \widehat{Q}_M(s,a) \quad \text{for all } s \in \mathcal{S}. \tag{180}$$

Finally, we introduce a useful fact that the iterates $\left\{\widehat{Q}_m\right\}_{m \geq 0}$ of our algorithm DRVI-LCB converge linearly to the fixed point $\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}$, summarized in the following lemma; its proof is postponed to Appendix F.3.2.

**Lemma 12.** *Let $\widehat{Q}_0 = 0$. The iterates of Algorithm 2 obey*

$$\forall m \geq 0: \quad \widehat{Q}_m \leq \widehat{Q}_{\mathsf{pe}}^{\star,\sigma} \quad \textit{and} \quad \|\widehat{Q}_m - \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_{\infty} \leq \frac{\gamma^m}{1-\gamma}. \tag{181}$$

*In addition, choosing $M \geq \frac{\log \frac{\sigma N}{1-\gamma}}{\log \frac{1}{\gamma}}$, one has*

$$\|\widehat{Q}_M - \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_{\infty} \leq \frac{1}{\sigma N}. \tag{182}$$

### E.3 Performance guarantees: infinite-horizon RMDPs

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

- $d^{\mathsf{b}}_{\mathsf{min}}$: the smallest positive state-action occupancy distribution of the dataset $\mathcal{D}$ under the nominal model $P^0$, i.e.,

$$d^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{s,a} \left\{ d^{\mathsf{b},P^0}(s,a) : \ d^{\mathsf{b},P^0}(s,a) > 0 \right\}. \tag{183}$$

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

$$P^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{s,a,s'} \left\{ P^0(s' \,|\, s,a) : \ d^{\mathsf{b},P^0}(s,a) > 0, \ P^0(s' \,|\, s,a) > 0 \right\}. \tag{184}$$

  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

$$P^{\star}_{\mathsf{min}} \coloneqq \min_{s,s'} \left\{ P^0\big(s' \,|\, s, \pi^\star(s)\big) : \ P^0\big(s' \,|\, s, \pi^\star(s)\big) > 0 \right\}. \tag{185}$$

  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 of 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 F.1.

**Theorem 3.** *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 2 are chosen as (176) for sufficiently large $c_{\mathrm{b}}$. With probability at least $1-\delta$, the output $\widehat{\pi}$ of Algorithm 2 obeys*

$$V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho) \le \frac{c_0}{\sigma(1-\gamma)^2} \sqrt{\frac{SC^{\star}_{\mathsf{rob}} \log^2\left(\frac{(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{\star}_{\mathsf{min}} N}}, \tag{186}$$

*as long as the number of samples $N$ satisfies*

$$N \ge \frac{c_1 \log(NS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}} P^{\mathsf{b}}_{\mathsf{min}}}. \tag{187}$$

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

$$\underbrace{\frac{SC^{\star}_{\mathsf{rob}}}{P^{\star}_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}}_{\varepsilon\text{-dependent}} + \underbrace{\frac{\log(NS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}} P^{\mathsf{b}}_{\mathsf{min}}}}_{\text{burn-in cost}}. \tag{188}$$

Note that the burn-in cost is independent with the accuracy level $\varepsilon$, which tells us that the sample complexity is on the order of

$$\widetilde{O}\left(\frac{SC^{\star}_{\mathsf{rob}}}{P^{\star}_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}\right) \tag{189}$$

as long as $\varepsilon$ is small enough.

The sample complexity of DRVI-LCB dramatically outperforms prior works, which has been compared in detail in Section 1.2 (cf. Table 1). We highlight that our sample complexity depends linearly on the state space $S$, in sharp contrast to prior works that all suffer from a quadratic dependency $S^2$. To achieve this, we resort to a delicate technique called leave-one-out analysis (Agarwal et al., 2020; Li et al., 2020; 2022) to decouple the statistical dependency introduced across the iterates of robust value iteration, which has potential to be used in deriving tighter sample complexity of other RMDP problems.

In addition, we develop an information-theoretic lower bound for robust offline RL as provided in the following theorem whose proof can be found in Appendix F.2.

**Theorem 4.** *Suppose* $(S, C^\star_{\mathsf{rob}}, \gamma, \sigma, \varepsilon)$ *obeying* $\frac{1}{1-\gamma} \geq e^8$, $S \geq \log(\frac{1}{1-\gamma})/2$, $C^\star_{\mathsf{rob}} \geq 8/S$, $\varepsilon \leq \frac{1}{256e^6(1-\gamma)\log\frac{1}{1-\gamma}}$, *and* $\log\frac{1}{1-\gamma} - 6 \leq \sigma \leq \log\frac{1}{1-\gamma} - 4$, *we consider two infinite-horizon robust MDPs* $\mathcal{M}_0, \mathcal{M}_1$, *an initial state distribution* $\rho$, *and a batch dataset with* $N$ *independent samples. Consequently, denoting* $\mathbb{P}_0$ *(resp.* $\mathbb{P}_1$*) as the probability when the MDP is* $\mathcal{M}_0$ *(resp.* $\mathcal{M}_1$*), one has*

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

*as long as*

$$N \leq \frac{c_1 S C^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\varepsilon^2}.$$

*Here, the infimum is taken over all estimator* $\widehat{\pi}$ *and* $c_1 > 0$ *is some universal constant.*

The sample complexity minimax lower bound shown above indicates that no any algorithm can find an $\varepsilon$-optimal policy if the sample complexity is below the order of $\Omega\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\varepsilon^2}\right)$. It directly confirms that DRVI-LCB is near-optimal up to a polynomial factor of the effective horizon length $\frac{1}{(1-\gamma)}$ (cf. (188)), which is the first provable algorithm with near-optimal sample complexity for infinite-horizon robust offline RL. Moreover, the requirement of the offline history dataset is also much weaker than prior literature on robust offline RL (Yang et al., 2021; Zhou et al., 2021), without the need of full coverage of the state-action space.

# F Analysis: discounted infinite-horizon RMDPs

## F.1 Proof of the upper bound: Theorem 3

In this section, we outline the proof of Theorem 3. Before moving to the main proof, we introduce some notations and important properties that are useful throughout the analysis.

### F.1.1 Notation and preliminary facts

**Notation.** To begin with, we denote the state-action space covered by the batch dataset $\mathcal{D}$ generated in the nominal model $P^0$ as

$$\mathcal{C}^{\mathsf{b}} = \left\{(s,a) : d^{\mathsf{b},P^0}(s,a) > 0\right\}. \tag{190}$$

Armed with it, in view of the definition in (175), we define a similar one based on the exact nominal model $P^0$ as

$$P_{\mathsf{min}}(s,a) \coloneqq \min_{s'}\left\{P^0(s' \,|\, s,a) :\; P^0(s' \,|\, s,a) > 0\right\}, \tag{191}$$

which combined with (184) and (185) directly indicates that

$$P^{\mathsf{b}}_{\mathsf{min}} = \min_{(s,a)\in\mathcal{C}^{\mathsf{b}}} P_{\mathsf{min}}(s,a), \qquad P^\star_{\mathsf{min}} = \min_s P_{\mathsf{min}}(s,\pi^\star(s)). \tag{192}$$

Finally, we denote the robust Q-value and robust value function outputed from Algorithm 2 as

$$\widehat{Q} = \widehat{Q}_M \quad \text{and} \quad \widehat{V} = \widehat{V}_M. \tag{193}$$

**Properties of** $N(s,a)$**.** We recall a key property of the visiting times $N(s,a)$ over the state-action pair $(s,a)$ which has been established as follows:

**Lemma 13** ((Li et al., 2022), Lemma 7)**.** *For any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, the quantities* $\{N(s,a)\}$ *in* (171) *obey*

$$\max\left\{N(s,a), \frac{2}{3}\log\frac{NS}{\delta}\right\} \geq \frac{N d^{\mathsf{b},P^0}(s,a)}{12} \tag{194}$$

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

Armed with Lemma 13, we introduce another fact that is useful throughout the proof. With probability at least $1-\delta$, we have

$$\forall (s,a) \in \mathcal{C}^{\mathsf{b}}: \qquad N(s,a) \geq \frac{N d^{\mathsf{b},P^0}(s,a)}{12} \geq \frac{c_1 \log(NS/\delta)}{12 P_{\min}(s,a)} \geq -\frac{\log \frac{2NS}{\delta}}{\log(1-P_{\min}(s,a))}, \tag{195}$$

as long as $c_1$ is some sufficient large universal constant and (187) holds. The proof is postponed to Appendix F.3.4.

**Properties of $\widehat{Q}$ and $\widehat{V}$.** Invoking Lemma 12 directly leads to

$$\left\| \widehat{Q} - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}} \right\|_\infty \leq \frac{1}{\sigma N} \tag{196}$$

and therefore

$$\left\| \widehat{V} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}} \right\|_\infty = \max_s \left| \max_a \widehat{Q}(s,a) - \max_a \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a) \right| \leq \left\| \widehat{Q} - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}} \right\|_\infty \leq \frac{1}{\sigma N}, \tag{197}$$

where the penultimate inequality arises from that the maximum operator is a 1-contraction.

### F.1.2 Proof of Theorem 3

We are now positioned to outline the proof of Theorem 3, which is separated into several key steps.

**Step 1: controlling the estimation uncertainty.** In view of the access to only finite and partial coverage samples for estimating the nominal transition kernel $P^0$, we need to efficiently control the uncertainty term $\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}$. However, the statistical dependency between the estimated value $\widehat{V}$ and the kernel estimation $\widehat{P}^0_{s,a}$ (since $\widehat{P}^0_{s,a}$ will be reused in the update rule (cf. (179)) for all the iterations) adds daunting challenges of controlling it tightly. This is also why the expensive quadratical dependency w.r.t. state space $S$ appears in the sample complexity of the prior works (Zhou et al., 2021; Panaganti & Kalathil, 2022; Yang et al., 2021), which was addressed by the covering theory. To overcome this challenge, we count on a leave-one-out argument motivated by (Agarwal et al., 2020; Li et al., 2020; 2022) to decouple the dependency. The results are summarized as the following lemma with the proof deferred to Appendix F.3.3.

**Lemma 14.** *Suppose the assumptions in Theorem 3 are satisfied and $\gamma \in \left[\frac{1}{2}, 1\right)$. Then for all vector $\widetilde{V}$ obeying $\left\|\widetilde{V} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty \leq \frac{1}{N}$ and $\|\widetilde{V}\|_\infty \leq \frac{1}{1-\gamma}$, with probability at least $1-\delta$, one has*

$$\begin{aligned} &\left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widetilde{V} \right| \\ &\leq \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta})}{\widehat{P}_{\min}(s,a)N(s,a)}} + \frac{4}{N\sigma(1-\gamma)}\ ,\ \frac{1}{1-\gamma} \right\} \end{aligned} \tag{198}$$

*simultaneously for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. In addition, with probability at least $1-\delta$, we have*

$$\forall (s,a) \in \mathcal{C}^{\mathsf{b}}: \quad \frac{P_{\min}(s,a)}{8\log(NS/\delta)} \leq \widehat{P}_{\min}(s,a) \leq e^2 P_{\min}(s,a). \tag{199}$$

**Step 2: establishing the pessimism property.** Armed with above lemma, we shall show that the estimated $\widehat{Q}(s,a)$ is a lower bound of $Q^{\widehat{\pi},\sigma}(s,a)$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. 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. (174)) obeys

$$\widehat{Q}^{\star,\sigma}_{\mathsf{pe}} = \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}) = \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b(s,a),\, 0 \right\}. \tag{200}$$

With this in mind, we shall show $\widehat{Q} \leq Q^{\widehat{\pi},\sigma}$ in two different conditions of $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$. In the state-action pairs obeying $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a) = 0$, recalling that the initial $\widehat{Q}_0 = 0$ and the definition in (193) gives

$$\widehat{Q}(s,a) = \widehat{Q}_M \leq \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a) = 0, \tag{201}$$

where the inequality holds by applying Lemma 12. As a result, $Q^{\widehat{\pi},\sigma}(s,a) \geq 0$ directly indicates $\widehat{Q}(s,a) \leq Q^{\widehat{\pi},\sigma}(s,a)$.

Consequently, we shall focus on the second case when $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a) > 0$. To continue, we observe that

$$\begin{aligned}
\widehat{Q}(s,a) &\overset{\text{(i)}}{\leq} \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a) + \frac{1}{\sigma N} = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b\big(s,a\big) + \frac{1}{\sigma N} \\
&\leq r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} - b(s,a) + \frac{1}{\sigma N} + \gamma \left| \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} \right| \\
&\overset{\text{(ii)}}{\leq} r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} - b(s,a) + \frac{2}{\sigma N} \\
&\leq r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V} - b(s,a) + \frac{2}{\sigma N} + \gamma \left| \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V} \right| \\
&\leq r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V}, \qquad (202)
\end{aligned}$$

where (i) follows from (196), (ii) arises from (197) and the basic fact that infimum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$, and the final inequality holds by the definition of $b(s,a)$ (see (176)) and Lemma 14.

With above result in mind and invoking the robust Bellman equation $Q^{\widehat{\pi},\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}V^{\widehat{\pi},\sigma}$ (see (166)), we arrive at

$$\begin{aligned}
Q^{\widehat{\pi},\sigma}(s,a) - \widehat{Q}(s,a) &\geq \gamma \left[ \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}V^{\widehat{\pi},\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V} \right] \\
&\overset{\text{(i)}}{=} \gamma \left[ \widetilde{P}_{s,a} V^{\widehat{\pi},\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V} \right] \geq \gamma \widetilde{P}_{s,a} \left( V^{\widehat{\pi},\sigma} - \widehat{V} \right), \qquad (203)
\end{aligned}$$

where (i) holds by letting $\widetilde{P}_{s,a} \coloneqq \arg\min_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})} \mathcal{P}V^{\widehat{\pi},\sigma}$. Consequently, one has

$$\begin{aligned}
\min_{s,a} \left[ Q^{\widehat{\pi},\sigma}(s,a) - \widehat{Q}(s,a) \right] &\geq \min_{s,a} \left[ \gamma \widetilde{P}_{s,a} \left( V^{\widehat{\pi},\sigma} - \widehat{V} \right) \right] \overset{\text{(i)}}{\geq} \gamma \min_{s} \left[ V^{\widehat{\pi},\sigma}(s) - \widehat{V}(s) \right] \\
&= \gamma \min_{s} \left[ Q^{\widehat{\pi},\sigma}\big(s,\widehat{\pi}(s)\big) - \widehat{Q}\big(s,\widehat{\pi}(s)\big) \right] \\
&\geq \gamma \min_{s,a} \left[ Q^{\widehat{\pi},\sigma}\big(s,a\big) - \widehat{Q}\big(s,a\big) \right] \qquad (204)
\end{aligned}$$

where (i) follows from $\widetilde{P}_{s,a} \in \Delta(\mathcal{S})$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Noting that $0 < \gamma < 1$, we conclude $Q^{\widehat{\pi},\sigma}(s,a) - \widehat{Q}(s,a) \geq 0$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ in this case, otherwise (204) won't happen.

Summing up these two cases, we arrive at

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad Q^{\widehat{\pi},\sigma}(s,a) \geq \widehat{Q}(s,a) \qquad (205)$$

and consequently

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) \geq V^{\widehat{\pi},\sigma}(s) = \max_a Q^{\widehat{\pi},\sigma}(s,a) \geq \max_a \widehat{Q}(s,a) = \widehat{V}(s). \qquad (206)$$

**Step 3: bounding $V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s)$.** First, armed with above pessimistic property (cf. (206)), we convert the goal to control another term in view of

$$V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \leq V^{\star,\sigma}(s) - \widehat{V}(s). \qquad (207)$$

Towards this, we observe that

$$\widehat{V}(s) = \max_a \widehat{Q}(s,a) \geq \widehat{Q}\big(s,\pi^\star(s)\big) \overset{\text{(i)}}{\geq} \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\big(s,\pi^\star(s)\big) - \frac{1}{\sigma N}$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{\geq} r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b\big(s,\pi^\star(s)\big) - \frac{1}{\sigma N} \\
&\geq r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - b\big(s,\pi^\star(s)\big) - \frac{1}{\sigma N} \\
&\quad - \gamma \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} \right| \\
&\overset{\text{(iii)}}{\geq} r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - b\big(s,\pi^\star(s)\big) - \frac{2}{\sigma N} \\
&\geq r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - b\big(s,\pi^\star(s)\big) - \frac{2}{\sigma N} \\
&\quad - \gamma \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} \right| \\
&\geq r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - 2b\big(s,\pi^\star(s)\big),
\end{aligned} \tag{208}
$$

where (i) follows from (196), (ii) holds by applying (200), (iii) arises from (197) 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 (176)) and Lemma 14.

To continue, invoking the robust Bellman optimality equation in (167) gives

$$
V^{\star,\sigma}(s) = Q^{\star,\sigma}\big(s,\pi^\star(s)\big) = r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}V^{\star,\sigma}. \tag{209}
$$

Armed with above results and (208), we arrive at

$$
\begin{aligned}
V^{\star,\sigma}(s) - \widehat{V}(s) &\leq Q^{\star,\sigma}\big(s,\pi^\star(s)\big) - \left( r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - 2b\big(s,\pi^\star(s)\big) \right) \\
&\leq \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}V^{\star,\sigma} - \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} + 2b\big(s,\pi^\star(s)\big) \\
&\leq \gamma \widehat{P}^{\inf}_{s,\pi^\star(s)} \left( V^{\star,\sigma} - \widehat{V} \right) + 2b\big(s,\pi^\star(s)\big)
\end{aligned} \tag{210}
$$

where the final inequality holds by introducing the additional notation

$$
\widehat{P}^{\inf}_{s,\pi^\star(s)} \coloneqq \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} \tag{211}
$$

and evidently,

$$
\inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}V^{\star,\sigma} \leq \widehat{P}^{\inf}_{s,\pi^\star(s)} V^{\star,\sigma}, \qquad \text{and} \qquad \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} = \widehat{P}^{\inf}_{s,\pi^\star(s)} \widehat{V}.
$$

Before continuing, for convenience, we 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

$$
\left[\widehat{P}^{\inf}\right]_{s,\cdot} = \widehat{P}^{\inf}_{s,\pi^\star(s)}, \qquad \text{and} \qquad b^\star(s) = b\big(s,\pi^\star(s)\big). \tag{212}
$$

With these notation in mind, applying (210) leads to

$$
\begin{aligned}
\left\langle \rho, V^{\star,\sigma} - \widehat{V} \right\rangle &= \sum_{s\in\mathcal{S}} \rho(s) \left( V^{\star,\sigma}(s) - \widehat{V}(s) \right) \\
&\leq \gamma \sum_{s\in\mathcal{S}} \rho(s) \widehat{P}^{\inf}_{s,\pi^\star(s)} \left( V^{\star,\sigma} - \widehat{V} \right) + 2\sum_{s\in\mathcal{S}} \rho(s) b\big(s,\pi^\star(s)\big)
\end{aligned}
$$

$$= \gamma \rho^\top \widehat{P}^{\inf} \left( V^{\star,\sigma} - \widehat{V} \right) + 2\rho^\top b^\star. \tag{213}$$

Applying the above result recursively gives

$$\begin{aligned}
\left\langle \rho, V^{\star,\sigma} - \widehat{V} \right\rangle &\le \gamma \rho^\top \widehat{P}^{\inf} \left( V^{\star,\sigma} - \widehat{V} \right) + 2\rho^\top b^\star \\
&\le \gamma \left( \gamma \rho^\top \widehat{P}^{\inf} \right) \widehat{P}^{\inf} \left( V^{\star,\sigma} - \widehat{V} \right) + 2 \left( \gamma \rho^\top \widehat{P}^{\inf} \right) b^\star + 2\rho^\top b^\star \\
&\le \cdots \le \left\{ \lim_{i \to \infty} \gamma^i \rho^\top \left( \widehat{P}^{\inf} \right)^i \left( V^{\star,\sigma} - \widehat{V} \right) \right\} + 2\rho^\top \sum_{i=0}^{\infty} \gamma^i \left( \widehat{P}^{\inf} \right)^i b^\star \\
&\overset{\text{(i)}}{\le} 2\rho^\top \sum_{i=0}^{\infty} \gamma^i \left( \widehat{P}^{\inf} \right)^i b^\star = 2\rho^\top \left( I - \gamma \widehat{P}^{\inf} \right)^{-1} b^\star,
\end{aligned} \tag{214}$$

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 the basic fact 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. (159) and (165a)), we arrive at

$$\left( d^{\star,\widehat{P}^{\inf}} \right)^\top = (1-\gamma)\rho^\top \sum_{t=0}^{\infty} \gamma^t \left( \widehat{P}^{\inf} \right)^t = (1-\gamma)\rho^\top \left( I - \gamma \widehat{P}^{\inf} \right)^{-1}. \tag{215}$$

Combined with (207), plugging in above result back into (214) yields

$$\left\langle \rho, V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \right\rangle \le \left\langle \rho, V^{\star,\sigma} - \widehat{V} \right\rangle \le \frac{2}{1-\gamma} \left\langle d^{\star,\widehat{P}^{\inf}}, b^\star \right\rangle. \tag{216}$$

**Step 4: controlling $\left\langle d^{\star,\widehat{P}^{\inf}}, b^\star \right\rangle$ using concentrability.** Before continuing, note that $\widehat{P}^{\inf} \in \mathcal{U}^\sigma(P^0)$ (see (211) and (212)), which in words means $\widehat{P}^{\inf}$ is some transition kernel inside $\mathcal{U}^\sigma(P^0)$ — the uncertainty set around the nominal kernel $P^0$.

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 set of state into two types and control them separately.

- **For $s \in \mathcal{S}$ where $\max_{P \in \mathcal{U}^\sigma(P^0)} d^{\star,P}\big(s, \pi^\star(s)\big) = 0$.** In this case, one has for any $\widetilde{P} \in \mathcal{U}^\sigma\big(P^0\big)$,

$$0 \le d^{\star,\widetilde{P}}(s) = d^{\star,\widetilde{P}}\big(s, \pi^\star(s)\big) \le \max_{P \in \mathcal{U}^\sigma\left(P^0\right)} d^{\star,P}\big(s, \pi^\star(s)\big) = 0, \tag{217}$$

  which consequently indicates

$$d^{\star,\widehat{P}^{\inf}}(s) = 0. \tag{218}$$

- **For $s \in \mathcal{S}$ where $\max_{P \in \mathcal{U}^\sigma\left(P^0\right)} d^{\star,P}\big(s, \pi^\star(s)\big) > 0$.** For such state $s$, we claim that

$$d^{\mathsf{b},P^0}\big(s, \pi^\star(s)\big) > 0 \quad \text{and} \quad \big(s, \pi^\star(s)\big) \in \mathcal{C}^{\mathsf{b}}, \tag{219}$$

  which can be verified by recalling Assumption 2 which requires that the history dataset $\mathcal{D}$ obeys

$$\max_{P \in \mathcal{U}^\sigma(P^0)} \frac{\min\left\{ d^{\star,P}\big(s, \pi^\star(s)\big), \frac{1}{S} \right\}}{d^{\mathsf{b},P^0}\big(s, \pi^\star(s)\big)} = \max_{P \in \mathcal{U}^\sigma(P^0)} \frac{\min\left\{ d^{\star,P}(s), \frac{1}{S} \right\}}{d^{\mathsf{b},P^0}(s)} \le C^\star_{\mathsf{rob}} < \infty. \tag{220}$$

  To continue, invoking the fact in (195) with $\big(s, \pi^\star(s)\big) \in \mathcal{C}^{\mathsf{b}}$ gives

$$N\big(s, \pi^\star(s)\big) \ge \frac{N d^{\mathsf{b},P^0}\big(s, \pi^\star(s)\big)}{12}$$

$$\overset{\text{(i)}}{\geq} \frac{N \max_{P\in\mathcal{U}^{\sigma}(P^0)} \min\left\{d^{\star,P}\big(s,\pi^{\star}(s)\big),\frac{1}{S}\right\}}{12C^{\star}_{\mathsf{rob}}} \geq \frac{N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}{12C^{\star}_{\mathsf{rob}}}, \tag{221}$$

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. (176)) by

$$\begin{aligned}
b^{\star}(s) &\leq \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\big(s,\pi^{\star}(s)\big)}} + \frac{4}{\sigma N(1-\gamma)} + \frac{2}{\sigma N} \\
&\overset{\text{(i)}}{\leq} \frac{4c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\big(s,\pi^{\star}(s)\big)}} + \frac{4}{\sigma N(1-\gamma)} + \frac{2}{\sigma N} \\
&\leq \frac{16c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{C^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}} + \frac{6}{\sigma N(1-\gamma)} \\
&\leq \frac{20c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{C^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}},
\end{aligned}$$

where (i) arises from (199), the penultimate inequality follows from (221), and the last inequality holds as long as $c_{\mathrm{b}}$ is large enough.

Summing up the results in above two cases, we arrive at

$$\begin{aligned}
\left\langle d^{\star,\widehat{P}^{\mathrm{inf}}}, b^{\star}\right\rangle &= \sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s) b^{\star}(s) \\
&\leq \sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)\frac{20c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{C^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}} \\
&\overset{\text{(i)}}{\leq} \frac{20c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)\frac{C^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}}\sqrt{\sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)} \\
&\leq \frac{40c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{SC^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{\star}_{\mathsf{min}}N}},
\end{aligned} \tag{222}$$

where (i) arises from applying the Cauchy-Schwarz inequality, and the last inequality holds by $P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big) \geq P^{\star}_{\mathsf{min}}$ for all $s\in\mathcal{S}$ (see (192)) and the following fact that has been established in (67):

$$\sum_{s\in\mathcal{S}} \frac{d^{\star,\widehat{P}^{\mathrm{inf}}}(s)}{\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}} \leq 2S. \tag{223}$$

**Step 5: finishing up the proof.** Finally, inserting (222) back into (216), we complete the proof: with probability at least $1-2\delta$, one has

$$\left\langle \rho, V^{\star,\sigma} - V^{\widehat{\pi},\sigma}\right\rangle \leq \frac{2}{1-\gamma}\left\langle d^{\star,\widehat{P}^{\mathrm{inf}}}, b^{\star}\right\rangle \leq \frac{80c_{\mathrm{b}}}{\sigma(1-\gamma)^2}\sqrt{\frac{SC^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{\star}_{\mathsf{min}}N}}. \tag{224}$$

### F.2 Proof of the lower bound: Theorem 4

We shall first construct some hard discounted infinite-horizon MDP instances and then characterize the sample complexity requirements over these instances.

### F.2.1 CONSTRUCTION OF HARD PROBLEM INSTANCES

**Construction of a collection of hard MDPs.** Suppose there are two MDPs

$$\left\{\mathcal{M}_\phi = \left(\mathcal{S}, \mathcal{A}, P^\phi, r, \gamma\right) \mid \phi = \{0,1\}\right\}. \tag{225}$$

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

$$P^\phi(s' \mid s, a) = \begin{cases} p\mathbb{1}(s'=2) + (1-p)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,\phi) \\ q\mathbb{1}(s'=2) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,1-\phi) \\ \mathbb{1}(s'=s) & \text{if} \quad s = 1 \text{ or } s = 2 \\ q\mathbb{1}(s'=s) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad s > 2 \end{cases} \tag{226}$$

where $p$ and $q$ are set as

$$p = \gamma + \Delta \quad \text{and} \quad q = \gamma \tag{227}$$

for some $\gamma$ and $\Delta$ obeying

$$1 - \gamma \le 1/e^8 \le \frac{1}{2} \quad \text{and} \quad \Delta \le \frac{1}{2}(1-\gamma). \tag{228}$$

Here, $\Delta$ is some value that will be introduced later. Consequently, applying (227) directly leads to

$$1 \ge p \ge q \ge \gamma \ge \frac{1}{2}. \tag{229}$$

Regarding the introduced transition kernel, in words, state 1 and 2 are absorbing states, and states $s > 2$ will stay without moving or go to state 1. The action will only influence the transitions in state 0. 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

$$r(s,a) = \begin{cases} 1 & \text{if } s = 0 \text{ or } s = 2 \\ 0 & \text{otherwise} \end{cases}. \tag{230}$$

**Uncertainty set of the transition kernels.** Then we introduce the considered robust MDPs with some tailored radius $\sigma$ of the uncertainty set, along with some useful properties.

To begin with, we introduce an important constant $\beta$ defined as

$$\beta := \frac{\log \frac{1}{1-\gamma}}{2} \ge 4. \tag{231}$$

Then, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we denote the transition kernel vector of $\mathcal{M}_\phi$ as

$$P^\phi_{s,a} := P^\phi(\cdot \mid s, a) \in [0,1]^{1 \times S}. \tag{232}$$

Armed with these, the perturbed transition kernels in $\mathcal{M}_\phi$ is limited to the following uncertainty set

$$\mathcal{U}^\sigma(P^\phi) := \otimes \mathcal{U}^\sigma\left(P^\phi_{s,a}\right), \qquad \mathcal{U}^\sigma(P^\phi_{s,a}) := \left\{P_{s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{s,a} \parallel P^\phi_{s,a}\right) \le \sigma\right\}, \tag{233}$$

where the radius of the uncertainty set $\sigma$ obeys

$$\left(1 - \frac{3}{\beta}\right) \log \frac{1}{1-\gamma} \le \sigma \le \left(1 - \frac{2}{\beta}\right) \log \frac{1}{1-\gamma}. \tag{234}$$

Next, to introduce the properties, for any $P^\phi(\cdot \mid s, a)$ in (226), we denote the minimum limit of the perturbed distribution transiting from the current state-action pair $(s,a)$ to the next state $s'$ as

$$\underline{P}^\phi(s' \mid s, a) := \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})} P(s' \mid s, a). \tag{235}$$

As the transition from state 0 to state 2 plays an important role in the analysis, in particular, we denote

$$\underline{p} := \underline{P}^\phi(2 \mid 0, \phi), \qquad \underline{q} := \underline{P}^\phi(2 \mid 0, 1-\phi). \tag{236}$$

With these definitions in mind, we summarize some useful properties of the uncertainty set in the following lemma.

**Lemma 15.** *Suppose $\beta$ satisfies (231) and the uncertainty level $\sigma$ satisfies (234). The perturbed transition kernels obey*

$$\underline{p} \geq \underline{q} \geq \frac{1}{\beta}. \tag{237}$$

*Proof.* The proof follows from exactly the pipeline in Appendix C.4.2 except replacing $H$ with $\frac{1}{1-\gamma}$. We omit the details for brevity. □

**Construction of the history/batch dataset.** Before continuing, we define a useful state distribution (only supported on the state subset $\{0, 1, 2\}$):

$$\mu(s) = \frac{1}{CS}\mathbb{1}(s = 0) + \frac{1}{CS}\mathbb{1}(s = 2) + \Big(1 - \frac{2}{CS}\Big)\mathbb{1}(s = 1), \tag{238}$$

where $C > 0$ is some constant that determines the robust concentrability coefficient $C^{\star}_{\mathsf{rob}}$ (will be introduced momentarily) and obeys

$$\frac{1}{CS} \leq \frac{1}{4}. \tag{239}$$

Generated over the nominal environment $\mathcal{M}_{\phi}$, a batch dataset consists of $N$ i.i.d samples $\{(s_i, a_i, s_i')\}_{1\leq i\leq N}$ according to (168), with the occupancy state (resp. state action) distribution chosen to be:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad d^{\mathsf{b}, P^{\phi}}(s) = \mu(s) \quad \text{and} \quad d^{\mathsf{b}, P^{\phi}}(s, a) = \frac{\mu(s)}{2}. \tag{240}$$

Additionally, we choose the following initial state distribution:

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{otherwise} \end{cases}. \tag{241}$$

**Value functions and optimal policies.** Now we are positioned to derive the corresponding robust value functions and identify the optimal policies. To clarify the notations, for any MDP $\mathcal{M}_{\phi}$, we denote $\pi^{\star}_{\phi}$ as the optimal policy. In addition, we denote the robust value function of any policy $\pi$ (resp. optimal policy $\pi^{\star}_{\phi}$) as $V^{\pi,\sigma}_{\phi}$ (resp. $V^{\star,\sigma}_{\phi}$) with uncertainty radius $\sigma$. Then, we can introduce the following lemma which describes some important properties of the robust value functions and optimal policies.

**Lemma 16.** *For any $\phi = \{0, 1\}$ and any policy $\pi$, one has*

$$V^{\pi,\sigma}_{\phi}(0) = 1 + \gamma z^{\pi}_{\phi}\frac{1}{1-\gamma}, \tag{242}$$

*where $z^{\pi}_{\phi}$ is defined as*

$$z^{\pi}_{\phi} \coloneqq \underline{p}\pi(\phi \,|\, 0) + \underline{q}\pi(1 - \phi \,|\, 0). \tag{243}$$

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

$$V^{\star,\sigma}_{\phi}(0) = 1 + \gamma\underline{p}\frac{1}{1-\gamma}, \quad V^{\star,\sigma}_{\phi}(2) = \frac{1}{1-\gamma}, \quad V^{\star,\sigma}_{\phi}(s) = 0 \quad \textit{for } s = 1 \textit{ or } s > 2 \tag{244a}$$

$$\pi^{\star}_{\phi}(\phi \,|\, 0) = 1, \qquad \textit{for } s \in \mathcal{S}. \tag{244b}$$

*Moreover, choosing $S \geq \beta$, the robust single-policy clipped concentrability coefficient $C^{\star}_{\mathsf{rob}}$ obeys*

$$C^{\star}_{\mathsf{rob}} = 2C. \tag{245}$$

*Proof.* See Appendix F.2.3. □

### F.2.2 ESTABLISHING THE MINIMAX LOWER BOUND

Now we are positioned to provide the sample complexity lower bound. Before starting, we introduce a useful notation representing the smallest positive state transition probability of the optimal policy $\pi_\phi^\star$ under any robust MDP $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$:

$$P_{\mathsf{min}}^\star := \min_{s,s'} \left\{ P^\phi\left(s' \,|\, s, \pi_\phi^\star(s)\right) :\ P^\phi\left(s' \,|\, s, \pi_\phi^\star(s)\right) > 0 \right\} = P^\phi\left(1|0,\phi\right) = 1-p. \tag{246}$$

To continue, the goal is to control the quantity w.r.t. any policy estimator $\widehat{\pi}$ based on the batch dataset and chosen initial distribution $\rho$ in (241)

$$\left\langle \rho, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \right\rangle = V_\phi^{\star,\sigma}(0) - V_\phi^{\widehat{\pi},\sigma}(0). \tag{247}$$

Towards this, we first introduce the following lemma:

**Lemma 17.** *Given* $\epsilon \le \frac{1}{256e^6(1-\gamma)\log\left(\frac{1}{1-\gamma}\right)}$, *choosing*

$$\Delta = 128e^6\sigma(1-q)\epsilon(1-\gamma), \tag{248}$$

*one has* $\Delta$ *obeys (satisfying (228))*

$$\Delta \le \frac{\sigma(1-\gamma)}{2\log\left(\frac{1}{1-\gamma}\right)} \le \frac{1}{2}(1-\gamma), \tag{249}$$

*and for any policy* $\widehat{\pi}$*,*

$$V_\phi^{\star,\sigma}(0) - V_\phi^{\widehat{\pi},\sigma}(0) \ge 2\epsilon\big(1 - \widehat{\pi}(\phi\,|\,0)\big). \tag{250}$$

*Proof.* This lemma can be verified by following the same pipeline in Appendix C.4.4 except replacing $H$ with $\frac{1}{1-\gamma}$ and with the additional condition $\gamma \ge \frac{1}{2}$. □

Armed with this lemma, following the same pipeline in Appendix C.3.2, we can complete the proof by observing that: let $c_1$ be some sufficient large constant, as long as the sample size is chosen as

$$N \le \frac{SC_{\mathsf{rob}}^\star \log 2}{4c_1 P_{\mathsf{min}}^\star \sigma^2 (1-\gamma)^2 \varepsilon^2}, \tag{251}$$

then we necessarily has

$$\inf_{\widehat{\pi}} \max_{\phi\in\{0,1\}} \mathbb{P}_\phi \left\{ \left\langle \rho, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \right\rangle \right\} \ge \frac{1}{8}, \tag{252}$$

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.

### F.2.3 PROOF OF LEMMA 16

To start, for any $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$, it is easily observed that for any policy $\pi$, the robust value function of a state $s > 0$ obey

$$V_\phi^{\pi,\sigma}(2) = \sum_{t=0}^{\infty} \gamma^t \cdot 1 = \frac{1}{1-\gamma}, \tag{253a}$$

$$V_\phi^{\pi,\sigma}(s) = \sum_{t=0}^{\infty} \gamma^t \cdot 0 = 0, \qquad \text{for } s = 1 \text{ or } s > 2. \tag{253b}$$

Similarly, the robust value function of state 0 satisfies

$$V_\phi^{\pi,\sigma}(0) = \mathbb{E}_{a\sim\pi(\cdot\,|\,0)} \left[ r(0,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,a}^\phi)} \mathcal{P} V_\phi^{\pi,\sigma} \right]$$

$$
\begin{aligned}
&\overset{\text{(i)}}{=} 1+\gamma\pi(\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{\phi}_{0,\phi})}\mathcal{P}V^{\pi,\sigma}_{\phi}+\gamma\pi(1-\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{\phi}_{0,1-\phi})}\mathcal{P}V^{\pi,\sigma}_{\phi}\\
&\overset{\text{(ii)}}{=} 1+\gamma\pi(\phi\,|\,0)\Big[\underline{p}V^{\pi,\sigma}_{\phi}(2)+\big(1-\underline{p}\big)\,V^{\pi,\sigma}_{\phi}(1)\Big]\\
&\qquad+\gamma\pi(1-\phi\,|\,0)\Big[\underline{q}V^{\pi,\sigma}_{\phi}(2)+\big(1-\underline{q}\big)\,V^{\pi,\sigma}_{\phi}(1)\Big]\\
&\overset{\text{(iii)}}{=} 1+\gamma V^{\pi,\sigma}_{\phi}(1)+\gamma z^{\pi}_{\phi}\Big[V^{\pi,\sigma}_{\phi}(2)-V^{\pi,\sigma}_{\phi}(1)\Big]\\
&=1+\gamma z^{\pi}_{\phi}V^{\pi,\sigma}_{\phi}(2)
\end{aligned}
\tag{254}
$$

where (i) holds by the reward function defined in (230), (ii) arises from (253) which indicates $V^{\pi,\sigma}_{\phi}(2)\geq V^{\pi,\sigma}_{\phi}(1)$ so that the infimum is obtained by picking the smallest possible mass on the transition to state 2 which reaches (236), (iii) follows from plugging in the definition of $z^{\pi}_{\phi}$ in (243), and the last identity is due to (253). Consequently, taking $\pi=\pi^{\star}_{\phi}$, we directly arrive at

$$
V^{\star,\sigma}_{\phi}(0)=1+\gamma z^{\pi^{\star}}_{\phi}V^{\pi,\sigma}_{\phi}(2)=1+\gamma z^{\pi^{\star}}_{\phi}\frac{1}{1-\gamma},
\tag{255}
$$

which holds by (253). Observing that the function $z\frac{\gamma}{1-\gamma}$ is increasing in $z$ and $z^{\pi}_{\phi}$ is also increasing in $\pi(\phi\,|\,0)$ (see the fact $\underline{p}\geq\underline{q}$ in (237)), the optimal policy in state 0 obeys

$$
\pi^{\star}_{\phi}(\phi\,|\,0)=1.
\tag{256}
$$

Finally, plugging the above fact back into (243) leads to

$$
z^{\star}_{\phi}:=z^{\pi^{\star}}_{\phi}=\underline{p}\pi^{\star}_{\phi}(\phi\,|\,0)+\underline{q}\pi^{\star}_{\phi}(1-\phi\,|\,0)=\underline{p},
\tag{257}
$$

which combined with (255) yields

$$
V^{\star,\sigma}_{\phi}(0)=1+\gamma\underline{p}\frac{1}{1-\gamma}.
\tag{258}
$$

For the rest of states $s>0$, since the action does not influence the transition, without loss of generality, we choose the optimal policy to obey

$$
\forall s>0:\quad \pi^{\star}_{\phi}(\phi\,|\,s)=1.
\tag{259}
$$

**Proof of (245).** To begin with, for any MDP $\mathcal{M}_{\phi}$ with $\phi\in\{0,1\}$, recall the definition of $C^{\star}_{\mathsf{rob}}$

$$
C^{\star}_{\mathsf{rob}}=\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^{\sigma}(P^{\phi})}\frac{\min\big\{d^{\star,P}(s,a),\frac{1}{S}\big\}}{d^{\mathsf{b},P^{\phi}}(s,a)}.
\tag{260}
$$

Regarding $\pi^{\star}_{\phi}(\phi\,|\,s)=1$ 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

$$
d^{\star,P}(0,\phi)=(1-\gamma)\rho(0)\pi^{\star}_{\phi}(\phi\,|\,0)=(1-\gamma),
\tag{261}
$$

which holds by that the agent transits from state 0 to some other states at the first step and then will never go back to state 0. In addition, one has for any $P\in\mathcal{U}^{\sigma}(P^{\phi})$,

$$
\begin{aligned}
d^{\star,P}(2,\phi)&=(1-\gamma)\rho(0)P(2\,|\,0,\phi)\sum_{t=0}^{\infty}\gamma^{t}\big(P(2\,|\,2,\phi)\big)^{t}\\
&=(1-\gamma)P(2\,|\,0,\phi)\sum_{t=0}^{\infty}\gamma^{t}\overset{\text{(i)}}{\geq}\underline{p}\geq\frac{1}{\beta}
\end{aligned}
\tag{262}
$$

where (i) holds by (236) and the final inequality follows from (237). Armed with above facts, we observe that

$$
\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^{\sigma}(P^{\phi})}\frac{\min\big\{d^{\star,P}(s,a),\frac{1}{S}\big\}}{d^{\mathsf{b},P^{\phi}}(s,a)}=\max_{s\in\{0,1,2\},P\in\mathcal{U}^{\sigma}(P^{\phi})}\frac{\min\big\{d^{\star,P}(s,\phi),\frac{1}{S}\big\}}{d^{\mathsf{b},P^{\phi}}(s,\phi)}
\tag{263}
$$

which follows from the properties of optimal policy in (259) 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:

$$\max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(2,\phi), \frac{1}{S}\right\}}{d^{\mathsf{b},P^\phi}(2,\phi)} \overset{\text{(i)}}{=} \frac{1}{Sd^{\mathsf{b},P^\phi}(2,\phi)} \overset{\text{(ii)}}{=} \frac{2}{S\mu(2)} = 2C, \tag{264}$$

$$\max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(0,\phi), \frac{1}{S}\right\}}{d^{\mathsf{b},P^\phi}(0,\phi)} \leq \frac{1}{Sd^{\mathsf{b},P^\phi}(0,\phi)} \overset{\text{(iii)}}{=} \frac{2}{S\mu(0)} = 2C, \tag{265}$$

$$\max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(1,\phi), \frac{1}{S}\right\}}{d^{\mathsf{b},P^\phi}(1,\phi)} \leq \frac{1}{Sd^{\mathsf{b},P^\phi}(1,\phi)} \overset{\text{(iv)}}{=} \frac{2}{S\left(1-\frac{2}{CS}\right)} \overset{\text{(v)}}{\leq} \frac{4}{S} \leq C, \tag{266}$$

where (i) holds by (262) by choosing $S \geq \beta$, (ii), (iii) and (iv) follow from the definition in (240) and (238), and (v) and the final inequality of (266) arise from the assumption in (239). Plugging in above result back into (263) directly completes the proof by

$$C^\star_{\mathsf{rob}} = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b},P^\phi}(s,a)} = 2C. \tag{267}$$

### F.3 Proof of auxiliary lemmas

#### F.3.1 Proof of Lemma 11

We shall provide the proof to show that the operator $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$ (cf. (174)) is a $\gamma$-contraction and the existence of the unique fixed point of $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$ subsequently.

Before starting, suppose $Q, Q', Q_1, Q_2 \in \mathbb{R}^{SA}$ obey $Q(s,a), Q'(s,a), Q_1(s,a), Q_2(s,a) \in \left[0, \frac{1}{1-\gamma}\right]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Then we introduce the following notations:

$$\begin{aligned} \forall s \in \mathcal{S}: \quad & V(s) \coloneqq \max_a Q(s,a), \quad V'(s) \coloneqq \max_a Q'(s,a), \\ & V_1(s) \coloneqq \max_a Q_1(s,a), \quad V_2(s) \coloneqq \max_a Q_2(s,a). \end{aligned} \tag{268}$$

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

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad \widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q)(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V - b\big(s,a\big), \tag{269}$$

which consequently leads to

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q)(s,a) = \max\left\{\widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q)(s,a), 0\right\}. \tag{270}$$

With this in mind, we observe that

$$\begin{aligned} \left\|\widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_2)\right\|_\infty &= \gamma \left\| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V_1 - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V_2 \right\|_\infty \overset{\text{(i)}}{\leq} \gamma \left\|V_1 - V_2\right\|_\infty \\ &\overset{\text{(ii)}}{=} \gamma \max_s \left| \max_a Q_1(s,a) - \max_a Q_2(s,a) \right| \\ &\leq \gamma \max_{(s,a)} |Q_1(s,a) - Q_2(s,a)| = \gamma \left\|Q_1 - Q_2\right\|_\infty \end{aligned} \tag{271}$$

where the first equality holds by applying the definition of $b(s,a)$ (cf. (176)) and (269), (i) follows from that the infimum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$ and $\|PV_1 - PV_2\|_\infty \leq \|V_1 - V_2\|_\infty$ for all $P \in \Delta(\mathcal{S})$, (ii) arises from the definitions in (268), and the last inequality is due to the maximum operator is also a 1-contraction w.r.t. $\|\cdot\|_\infty$.

Taking the above result with (270), we verify the desired assertion by

$$\left\|\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_1) - \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_2)\right\|_\infty \leq \left\|\widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}^\sigma_{\mathsf{pe}}(Q_2)\right\|_\infty \leq \gamma \left\|Q_1 - Q_2\right\|_\infty, \tag{272}$$

where the first inequality follows from the basic fact that the maximum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$.

**Existence of the unique fixed-point.** To continue, we shall firstly show that there exist at least one fixed-point of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$. Recalling the definition of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ (cf. (174))

$$\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q)(s,a) = \max\left\{r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V - b\big(s,a\big),\, 0\right\}, \tag{273}$$

one has as long as $0 \le Q \le \frac{1}{1-\gamma}\cdot 1$, it is easily verified $0 \le \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q) \le \frac{1}{1-\gamma}\cdot 1$. Then, we construct the following sequence of Q-function recursively

$$Q^{(0)} = 0, \quad \text{and} \quad Q^{(t+1)} = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q^{(t)}) \quad \text{for all } t \ge 0, \tag{274}$$

which mimic the iterations of our algorithm DRVI-LCB. As a result, the proof for the Banach fixed-point theorem (Agarwal et al., 2001, Theorem 1) gives that as $t \to \infty$, $Q^{(t)}$ converges to some point $Q^{(\infty)}$. It can also be verified that $0 \le Q^{(\infty)} \le \frac{1}{1-\gamma}\cdot 1$, which indicates the existence of the fixed points. Then, to prove the uniqueness of the fixed points of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$, we suppose that there exists another point $Q'$ obeying $Q' = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q')$. The definition of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ directly gives $Q' \ge 0$ and if $\|Q'\|_\infty > \frac{1}{1-\gamma}$, then

$$\begin{aligned}\|Q'\|_\infty = \left\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q')\right\|_\infty &\le \|r\|_\infty + \gamma \max_{(s,a)}\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V'\right| \\ &\le 1 + \gamma\|V'\|_\infty \le 1 + \gamma\|Q'\|_\infty < (1-\gamma)\|Q'\|_\infty + \gamma\|Q'\|_\infty = \|Q'\|_\infty\end{aligned} \tag{275}$$

gives contraction. Therefore, we have $0 \le Q' \le \frac{1}{1-\gamma}\cdot 1$, which yields

$$\left\|Q' - Q^{(\infty)}\right\|_\infty = \left\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q') - \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}\left(Q^{(\infty)}\right)\right\|_\infty \le \gamma\left\|Q' - Q^{(\infty)}\right\|_\infty. \tag{276}$$

However, (276) can't happen given $\gamma \in \left[\frac{1}{2}, 1\right)$, indicating the uniqueness of the fixed points of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$.

### F.3.2 Proof of Lemma 12

To begin with, considering any $Q, Q'$ obeying $Q \le Q', 0 \le Q \le \frac{1}{1-\gamma}\cdot 1$, and $0 \le Q' \le \frac{1}{1-\gamma}\cdot 1$, we observe that the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ (cf. (174)) has the monotone non-decreasing property, namely,

$$\begin{aligned}\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q) &= \max\left\{r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V - b\big(s,a\big),\, 0\right\} \\ &= \max\left\{r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\max_{a'} Q(\cdot,a') - b\big(s,a\big),\, 0\right\} \\ &\le \max\left\{r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\max_{a'} Q'(\cdot,a') - b\big(s,a\big),\, 0\right\} = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q').\end{aligned} \tag{277}$$

In addition, armed with (277) and the initial rule $\widehat{Q}_0 = 0$, we also observe that the fixed-point $\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}$ of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ obeys $0 \le \widehat{Q}_{\mathsf{pe}}^{\star,\sigma} \le \frac{1}{1-\gamma}\cdot 1$. Consequently, we arrive at

$$\widehat{Q}_1 = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_0) \le \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}) = \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}. \tag{278}$$

Implementing the above result recursively gives

$$\text{for all } m \ge 0: \quad \widehat{Q}_m \le \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}. \tag{279}$$

To continue, applying Lemma 11 yields that for any $m \ge 0$,

$$\|\widehat{Q}_m - \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_\infty = \left\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_{m-1}) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_{\mathsf{pe}}^{\star,\sigma})\right\|_\infty \le \gamma\|\widehat{Q}_{m-1} - \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_\infty \tag{280}$$

$$\leq \cdots \leq \gamma^m \|\widehat{Q}_0 - \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_\infty = \gamma^m \|\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_\infty \leq \frac{\gamma^m}{1-\gamma}, \tag{281}$$

where the last inequality holds by the fact $\|\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_\infty \leq \frac{1}{1-\gamma}$ (see Lemma 11). The final assertion can be directly achieved with the above result by observing

$$\left\|\widehat{Q}_M - \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\right\|_\infty \leq \frac{\gamma^M}{1-\gamma} \leq \frac{1}{\sigma N} \tag{282}$$

when $M \geq \frac{\log \frac{\sigma N}{1-\gamma}}{\log \frac{1}{\gamma}}$.

### F.3.3 Proof of Lemma 14

We first note that the second assertion in (199) is a counterpart of (48) which can be verified following the same argument in Appendix C.2.1 except the set of notations are adapted to the infinite-horizon case. Therefore, the rest of the proof will focus on verifying (198).

To begin with, we consider the situation when $N(s,a) = 0$. In this case, (198) can be easily verified that

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}V\right| \overset{\text{(i)}}{=} \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}V \leq \|V\|_\infty \overset{\text{(ii)}}{\leq} \frac{1}{1-\gamma}, \tag{283}$$

where (i) follows from the fact $\widehat{P}^0_{s,a} = 0$ when $N(s,a) = 0$ (see (172)), and (ii) arises from the assumption $\|V\|_\infty \leq \frac{1}{1-\gamma}$. Consequently, in the remainder of the proof, we focus on verifying (198) when $N(s,a) > 0$.

Before continuing, we introduce a counterpart of the fact (47) in Lemma 8 as follows:

**Lemma 18.** *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*

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}V\right| \leq \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log(\frac{NS}{\delta})}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}}. \tag{284}$$

*Proof.* The proof follows from the same pipeline of the proof in Appendix C.2.2. The only difference is the upper bound on $\|V\|_\infty$ is $\frac{1}{1-\gamma}$ (as opposed to $H$), the union bound is taken over $N$ (as opposed to $KH$), and some notations are exchanged to that of the infinite-horizon case. We omit the proof details for conciseness. □

Armed with above point-wise results, we are now ready to derive the union bound over all $\widetilde{V}$ desired in Lemma 14, counting on a leave-one-out argument separated into the following steps.

**Step 1: construction of auxiliary robust MDPs with state-absorbing nominal transitions.** To begin with, we 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 $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ so that it is the same as $\widehat{\mathcal{M}}_{\mathsf{rob}}$ except the properties in state $s$. Specifically, the reward function of the auxiliary robust MDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is denoted as $r^{s,u}$ which obeys

$$\begin{cases} r^{s,u}(s,a) = u & \text{for all } a \in \mathcal{A}, \\ r^{s,u}(\widetilde{s},a) = r(\widetilde{s},a) & \text{for all } (\widetilde{s},a) \in \mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s} \neq s. \end{cases} \tag{285}$$

In addition, the nominal transition kernel of $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is denoted as $P^{s,u}$ such that

$$\begin{cases} P^{s,u}(s' \,|\, s,a) = \mathbb{1}(s'=s) & \text{for all } (s',a) \in \mathcal{S}\times\mathcal{A}, \\ P^{s,u}(\cdot \,|\, \widetilde{s},a) = \widehat{P}^0(\cdot \,|\, \widetilde{s},a) & \text{for all } (\widetilde{s},a) \in \mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s} \neq s. \end{cases} \tag{286}$$

It can be observed that the nominal transition kernel $P^{s,u}$ of the auxiliary $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ drops all the randomness of $\widehat{P}_{s,a}^0$ for all $a \in \mathcal{A}$ in state $s$ and makes $s$ an absorbing state, while keeps other parts the same as $\widehat{P}^0$.

With the robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ in hand, we can define the corresponding penalty term for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ as follows

$$b^{s,u}(s,a) \coloneqq \begin{cases} \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}^{s,u}(s,a)N(s,a)}} + \frac{4}{N\sigma(1-\gamma)} \ , \ \frac{1}{1-\gamma} \right\} + \frac{2}{\sigma N} & \text{if } N(s,a) > 0, \\ \frac{1}{1-\gamma} + \frac{2}{\sigma N} & \text{otherwise,} \end{cases} \tag{287}$$

where $P_{\mathsf{min}}^{s,u}$ is defined as the smallest positive state transition probability over the nominal kernel $P^{s,u}$ as follows:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad P_{\mathsf{min}}^{s,u}(s,a) \coloneqq \min_{s'} \Big\{ P^{s,u}(s' \,|\, s,a) : \ P^{s,u}(s' \,|\, s,a) > 0 \Big\}. \tag{288}$$

Armed with the penalty term, the pessimistic robust Bellman operator $\widehat{\mathcal{T}}_{s,u}^{\sigma}(Q)(\cdot)$ w.r.t. $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ is defined as

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{\mathcal{T}}_{s,u}^{\sigma}(Q)(s,a) = \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(P_{s,a}^{s,u})} \mathcal{P}V - b^{s,u}(s,a),\, 0 \right\}. \tag{289}$$

**Step 2: verifying the relation between $\widehat{\mathcal{M}}_{\mathsf{rob}}$ and the auxiliary robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$.** Recall that $\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}$ is the unique fixed-point of operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ with the corresponding value $\widehat{V}_{\mathsf{pe}}^{\star,\sigma}$. In particular, given a state $s$, we introduce a special reward

$$u^\star \coloneqq (1-\gamma)\widehat{V}_{\mathsf{pe}}^{\star,\sigma}(s) + \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}^{s,u}(s,a)N(s,a)}} + \frac{4}{N\sigma(1-\gamma)} \ , \ \frac{1}{1-\gamma} \right\} + \frac{2}{\sigma N}. \tag{290}$$

With it in mind, we shall justify that there exists a fixed-point $\widehat{Q}_{s,u^\star}^{\star,\sigma}$ of the operator $\widehat{\mathcal{T}}_{s,u^\star}^{\sigma}(\cdot)$ whose corresponding value $\widehat{V}_{s,u^\star}^{\star,\sigma}$ is identical to $\widehat{V}_{\mathsf{pe}}^{\star,\sigma}$. Towards this, we shall show the facts in two different cases:

- **For state $s' \neq s$.** In this case, for any $s' \neq s$ and $a \in \mathcal{A}$, it can be verified that

$$\begin{aligned} &\max\left\{ r^{s,u^\star}(s',a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(P_{s',a}^{s,u^\star})} \mathcal{P}\widehat{V}_{\mathsf{pe}}^{\star,\sigma} - b^{s,u^\star}(s',a),\, 0 \right\} \\ &= \max\left\{ r(s',a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}_{s',a}^0)} \mathcal{P}\widehat{V}_{\mathsf{pe}}^{\star,\sigma} - b(s',a),\, 0 \right\} \\ &= \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_{\mathsf{pe}}^{\star,\sigma})(s',a) = \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}(s',a), \end{aligned} \tag{291}$$

  where the first identity follows from the definitions in (285) and (286), the penultimate equality arises from (174), and the final equality holds by that $\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}$ 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. (286)), applying (288) yields

$$P_{\mathsf{min}}^{s,u}(s,a) = 1. \tag{292}$$

  Plugging above fact into (287) leads to

$$b^{s,u}(s,a) = \begin{cases} \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{N(s,a)}} + \frac{4}{N\sigma(1-\gamma)} \ , \ \frac{1}{1-\gamma} \right\} + \frac{2}{\sigma N} & \text{if } N(s,a) > 0, \\ \frac{1}{1-\gamma} & \text{otherwise} \end{cases} \tag{293}$$

for all $a \in \mathcal{A}$. As a result, we have for any $a \in \mathcal{A}$:

$$\begin{aligned}
&\max\left\{r^{s,u^\star}(s,a)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^{s,u^\star}_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b^{s,u^\star}(s,a),\, 0\right\} \\
&= \max\left\{u^\star + \gamma \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) - b^{s,u^\star}(s,a),\, 0\right\} \\
&\overset{\text{(i)}}{=} \max\left\{(1-\gamma)\widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) + \gamma \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s),\, 0\right\} = \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s),
\end{aligned} \tag{294}$$

where (i) follows from plugging in the definition of $u^\star$ in (290) and $b^{s,u^\star}(s,a)$ in (293).

Summing up the above results, we observe 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

$$\begin{cases}
\widehat{Q}^{\star,\sigma}_{s,u^\star}(s,a) = \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) & \text{for all } a \in \mathcal{A} \\
\widehat{Q}^{\star,\sigma}_{s,u^\star}(s',a) = \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s',a) & \text{for all } s' \neq s \text{ and } a \in \mathcal{A}.
\end{cases} \tag{295}$$

Consequently, we already 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 $\varepsilon$-net for all rewards $u$.** Before continuing, it is easily verified that the reward obeys

$$u^\star \le 1 + \min\left\{\frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{s,u}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{4}{\sigma N(1-\gamma)},\ \frac{1}{1-\gamma}\right\} + \frac{2}{\sigma N} \le \frac{2}{\sigma} + \frac{2}{(1-\gamma)}. \tag{296}$$

As a result, we construct an $\varepsilon$-net (Vershynin, 2018) of the range $\left[0, \frac{2}{\sigma} + \frac{2}{(1-\gamma)}\right]$ with $\varepsilon = \frac{1}{\sigma N}$ as follows:

$$\mathcal{U}_\varepsilon \coloneqq \left\{\frac{i}{\sigma N} \,|\, 1 \le i \le \sigma N\left(\frac{2}{\sigma} + \frac{2}{(1-\gamma)}\right)\right\}. \tag{297}$$

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 pipeline in the proof of Lemma 11 (cf. Appendix F.3.1), 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 $0 \le \widehat{Q}^{\star,\sigma}_{s,u} \le \frac{1}{1-\gamma}\cdot 1$. Consequently, the corresponding value function also satisfies $\|\widehat{V}^{\star,\sigma}_{s,u}\|_\infty \le \frac{1}{1-\gamma}$.

To continue, in view of the definitions in (285) and (286), we notice that 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 $\widehat{P}^0_{s,a}$. So invoking Lemma 18 and taking the union bound over all samples $N$ and $u \in \mathcal{U}_\varepsilon$ give that, with probability at least $1-\delta$,

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u}\right| \le \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} \tag{298}$$

hold simultaneously for all $(s,a,u) \in \mathcal{S}\times\mathcal{A}\times\mathcal{U}_\varepsilon$ with $N(s,a) > 0$.

**Step 4: implementing a covering argument.** To continue, we shall control the gap between the value functions of the fixed-points of $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ and the auxiliary operator $\widehat{\mathcal{T}}^{\sigma}_{s,u}(\cdot)$, i.e., $\left\|\widehat{V}^{\star,\sigma}_{s,u} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty$. First, recalling that $u^\star \in \left[0, \frac{2}{\sigma} + \frac{2}{(1-\gamma)}\right]$ (see (290)), 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 (289) yields

$$\forall Q \in \mathbb{R}^{SA}: \quad \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(Q) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(Q)\right\|_\infty \overset{\text{(i)}}{\le} |\widetilde{u} - u^\star| \le \frac{1}{\sigma N}, \tag{299}$$

where (i) holds by $b^{s,\widetilde{u}}(s,a) = b^{s,u^\star}(s,a)$ for $s$ (see (293)) and $b^{s,\widetilde{u}}(s',a) = b^{s,u^\star}(s',a) = b(s',a)$ for all $s' \neq s$.

With this in mind, we observe that the fixed-points obey

$$\begin{aligned}\left\|\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty &= \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}}) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty \\ &\le \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}}) - \widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty + \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star}) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty \\ &\le \gamma\left\|\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty + \frac{1}{\sigma N}, \end{aligned} \tag{300}$$

which directly indicates that

$$\left\|\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \frac{1}{(1-\gamma)\sigma N} \tag{301}$$

and then

$$\left\|\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{V}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \left\|\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \frac{1}{(1-\gamma)\sigma N}. \tag{302}$$

Armed with above facts, invoking the identity between $\widehat{V}^{\star,\sigma}_{\mathsf{pe}}$ and $\widehat{V}^{\star,\sigma}_{s,u^\star}$ established in Step 2 gives

$$\begin{aligned}&\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right| = \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star}\right| \\ &\overset{\text{(i)}}{\le} \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}}\right| \\ &\quad + \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star}\right| + \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star}\right| \\ &\overset{\text{(ii)}}{\le} \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}}\right| + \frac{2}{N\sigma(1-\gamma)} \\ &\le \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{2}{N\sigma(1-\gamma)}, \end{aligned} \tag{303}$$

where (i) holds by applying the triangle inequality, (ii) arises from (302) and the basic fact that infimum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$, and the final inequality follows from (298).

**Step 5: union bound for all $\widetilde{V}$.** Now we are positioned to show the union bound for all vector $\widetilde{V}$ obeying $\left\|\widetilde{V} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty \le \frac{1}{\sigma N}$ and $\|\widetilde{V}\|_\infty \le \frac{1}{1-\gamma}$. For any $\widetilde{V}$ mentioned above, we invoke (303) and apply the triangle inequality to reach

$$\begin{aligned}&\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widetilde{V}\right| \le \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right| \\ &\quad + \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right| + \left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right| \end{aligned} \tag{304}$$

$$\le \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{4}{N\sigma(1-\gamma)}. \tag{305}$$

Finally, we complete the proof by verifying that

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})}\mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})}\mathcal{P}\widetilde{V}\right| \le \left\|\widetilde{V}\right\|_\infty \le \frac{1}{1-\gamma} \tag{306}$$

which holds by that the infimum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$ and the assumption $\|\widetilde{V}\|_\infty \le \frac{1}{1-\gamma}$.

### F.3.4 PROOF OF (195)

For all $(s,a) \in \mathcal{C}^{\mathsf{b}}$, one has

$$Nd^{\mathsf{b},P^0}(s,a) \overset{\text{(i)}}{\geq} \frac{c_1 d^{\mathsf{b},P^0}(s,a)\log(NS/\delta)}{d^{\mathsf{b}}_{\min} P^{\mathsf{b}}_{\min}} \overset{\text{(ii)}}{\geq} \frac{c_1 \log(NS/\delta)}{P^{\mathsf{b}}_{\min}} \overset{\text{(iii)}}{\geq} \frac{c_1 \log(NS/\delta)}{P_{\min}(s,a)}, \tag{307}$$

where (i) follows from the condition (187), (ii) arises from the definition that $d^{\mathsf{b}}_{\min} \leq d^{\mathsf{b},P^0}(s,a)$ for all $(s,a) \in \mathcal{C}^{\mathsf{b}}$, and (iii) follows from the definition in (192).

Armed with above result, when $c_1$ is large enough, one has $\frac{2}{3}\log\frac{NS}{\delta} < \frac{Nd^{\mathsf{b},P^0}(s,a)}{12}$. Consequently, Lemma 13 tells us that with probability at least $1-\delta$,

$$N(s,a) \geq \frac{Nd^{\mathsf{b},P^0}(s,a)}{12} \geq \frac{c_1 \log(NS/\delta)}{12 P_{\min}(s,a)}. \tag{308}$$

Regarding the basic fact $x \leq -\log(1-x)$ for all $x \in [0,1]$, the last inequality of (195) can be verified by

$$\frac{c_1 \log(NS/\delta)}{12 P_{\min}(s,a)} \geq -\frac{\log\frac{2NS}{\delta}}{\log(1-P_{\min}(s,a))}. \tag{309}$$