A Appendix

A.1 Proof of Theorem 1

We initiate our discussion by introducing the performance disparity between the offline dataset and reality (represented by the universal uncertainty set) in Lemma 1. This serves as a foundation for the subsequent proof presented in Theorem 1.

Lemma (1). *[**Reality Gap:** Performance Gap between Offline Dataset and Reality(the universal uncertainty set)] The value of any policy π learned from P_B on the universal uncertainty set $\mathcal{U}$ and the induced offline dataset transition kernel P_B satisfies:*

$$J_{\rho_0^B}(\pi, P_B) \geq \mathbb{E}_{P_0\sim\mathcal{U}}(J_{\rho_0}(\pi, P_0)) - \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{TV}(\rho_0, \rho_0^B)] - \frac{2\gamma R_{max}}{(1-\gamma)^2}\beta - \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{T^\pi_{\mathcal{V}}}] \tag{11}$$

$$J_{\rho_0^B}(\pi, P_B) \leq \mathbb{E}_{P_0\sim\mathcal{U}}(J_{\rho_0}(\pi, P_0)) + \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{TV}(\rho_0, \rho_0^B)] + \frac{2\gamma R_{max}}{(1-\gamma)^2}\beta + \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{T^\pi_{\mathcal{V}}}], \tag{12}$$

where $\mathcal{V}$ is an unknown state action set defined as: $(s, a) \in \mathcal{V}$ iff (s, a) is not in the offline dataset and $T^\pi_{\mathcal{V}}$ denotes the hitting time of unknown states.

Proof. The inequalities provide a comparison of a policy π's performance under two distinct dynamics models: P_B and $P_0 \sim \mathcal{U}$. To further understand these differences, it's beneficial to categorize the states into two groups: those present in the dataset (known state-actions) and those absent from it (unknown state-actions).

For state-action pairs present in the dataset, the primary objective is to concurrently couple the trajectory of any chosen policy on both the offline dataset MDP, M_B, and the reality MDP, M. Given an initial successful coupling, we consider the following constraint

$$\mathbb{E}_{P_0\sim\mathcal{U}}\left[\|P_B(s, a) - P_0(s, a)\|_1\right] \leq \beta.$$

One can verify that this coupling can be consistently maintained in subsequent steps with a probability of $1 - \beta$. The likelihood of decoupling at time t is at most $1 - (1 - \beta)^t$.

For state-action pairs not present in the datase, the divergence peaks: within M, the return upper bound for cumulative rewards post this encounter is $\frac{R_{\max}}{1-\gamma}$, whereas in M_B, the corresponding return lower bound is $-\frac{R_{\max}}{1-\gamma}$. This divergence can be quantified using the discount factor $\mathbb{E}[\gamma^{\mathrm{T}^\pi_{\mathcal{V}}}]$, resulting in a measure of disparity introduced by these unidentified state-action pairs: $\frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{\mathrm{T}^\pi_{\mathcal{V}}}]$.

So the total difference in the values of the policy π on the two MDPs can be upper bounded as:

$$\left|J_{\rho_0^B}(\pi, P_B) - \mathbb{E}_{P_0\sim\mathcal{U}}[J_{\rho_0}(\pi, P_0)]\right| \tag{13}$$

$$\leq \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{\mathrm{TV}}(\rho_0, \rho_0^B)] + \sum_t \gamma^t(1 - (1-\beta)^t)\cdot 2\cdot R_{\max} + \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{\mathrm{T}^\pi_{\mathcal{V}}}] \tag{14}$$

$$\leq \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{\mathrm{TV}}(\rho_0, \rho_0^B)] + \frac{2R_{\max}\gamma\beta}{(1-\gamma)(1-\gamma\cdot(1-\beta))} + \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{\mathrm{T}^\pi_{\mathcal{V}}}] \tag{15}$$

$$\leq \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{\mathrm{TV}}(\rho_0, \rho_0^B)] + \frac{2R_{\max}\gamma}{(1-\gamma)^2}\beta + \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{\mathrm{T}^\pi_{\mathcal{V}}}] \tag{16}$$

□

For ease of exposition, we restate Theorem 1 as follows.

Theorem (1). *For any ϵ_π sub-optimal policy, we have:*

$$\mathbb{E}_{P_0\sim\mathcal{U}}[J_{\rho_0}(\pi^*,P_0)] - \mathbb{E}_{P_0\sim\mathcal{U}}[J_{\rho_0}(\pi,P_0)] \leq \epsilon_\pi + \frac{4R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{TV}(\rho_0,\rho_0^B)] + \frac{4\gamma R_{max}}{(1-\gamma)^2}\beta + \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{T_{\mathcal{V}}^{\pi}}] + \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{T_{\mathcal{V}}^{\pi^*}}]. \tag{17}$$

Proof. By Lemma 1, we have

$$\begin{aligned}
\mathbb{E}_{P_0\sim\mathcal{U}}[J_{\rho_0}(\pi,P_0)] \geq & J_{\rho_0^B}(\pi,P_B) - \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{\mathrm{TV}}(\rho_0,\rho_0^B)] - \frac{2R_{\max}\gamma}{(1-\gamma)^2}\beta - \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi}}] \\
\geq & J_{\rho_0^B}(\pi^*,P_B) - \epsilon_\pi \qquad (18)\\
& - \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{\mathrm{TV}}(\rho_0,\rho_0^B)] - \frac{2R_{\max}\gamma}{(1-\gamma)^2}\beta - \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi}}] \qquad (19)\\
\geq & \mathbb{E}_{P_0\sim\mathcal{U}}[J_{\rho_0}(\pi^*,P_0)] - \epsilon_\pi \qquad (20)\\
& - \frac{4R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{\mathrm{TV}}(\rho_0,\rho_0^B)] - \frac{4R_{\max}\gamma}{(1-\gamma)^2}\beta - \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi}}] - \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi^*}}] \qquad (21)
\end{aligned}$$

□

A.2 Proof of Theorem 2

We begin by highlighting the performance divergence of a risk-aware policy between the offline dataset and the true environment, represented by the universal uncertainty set, in Lemma 3. Within this lemma, the term $\beta - p_r\beta_r$ underscores the narrowed performance gap achieved by incorporating robustness into the modeling. This foundational understanding sets the stage for the detailed proof in Theorem 2.

Lemma 3. *[**Risk-Aware Policy Reality Gap:** Performance Gap between Risk-Aware Uncertainty Set and Reality (the universal uncertainty set)] Given a robust policy π_r such that $\pi_r = \arg\max_\pi \mathbb{E}_{P\sim\mathcal{U}_r} J_\rho(\pi,P)$ and $\mathbb{E}_{P\sim\mathcal{U}_r}(\mathbb{E}_{s,a} D_{TV}(P,P_B)) \leq \beta_r$. Considering the fact that there might be randomness that we cannot capture during training, we assume $\mathcal{U}_r \subseteq \mathcal{U}$, $\beta \geq \beta_r$, and the probabitlity of $P \in \mathcal{U}_r$ for every $P \in \mathcal{U}$ is p_r where $0 \leq p_r \leq 1$. The performance on the uncertain nominal transition kernel set $\mathcal{U}$ and the training transition kernel set $\mathcal{U}_r$ satisfies:*

$$\mathbb{E}_{P_r\sim\mathcal{U}_r}[J_{\rho_0^B}(\pi_r,P_r)] \geq \mathbb{E}_{P_0\sim\mathcal{U}}[J_{\rho_0}(\pi_r,P_0)] - \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{TV}(\rho_0,\rho_0^B)] - \frac{2\gamma R_{max}}{(1-\gamma)^2}(\beta - p_r\beta_r) - \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{T_{\mathcal{V}}^{\pi_r}}] \tag{22}$$

$$\mathbb{E}_{P_r\sim\mathcal{U}_r}[J_{\rho_0^B}(\pi_r,P_r)] \leq \mathbb{E}_{P_0\sim\mathcal{U}}[J_{\rho_0}(\pi_r,P_0)] + \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{TV}(\rho_0,\rho_0^B)] + \frac{2\gamma R_{max}}{(1-\gamma)^2}\beta + \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{T_{\mathcal{V}}^{\pi_r}}], \tag{23}$$

where $\mathcal{V}$ is an unknown state action set defined as: $(s,a) \in \mathcal{V}$ iff (s,a) is not in the offline dataset and $T_{\mathcal{V}}^{\pi_r}$ denotes the hitting time of unknown states.

Proof. Building upon the insights and methodology established in the proof of Lemma 1, we now turn our attention to a new set of dynamics. In this context, we compare the performance of a policy π across two distinct transition dynamics: $P_r \sim \mathcal{U}_r$ and $P_0 \sim \mathcal{U}$. By leveraging the foundational ideas from the aforementioned theorem, we aim to unravel the performance disparities between these two MDPs, especially focusing on the divergence arising from known state-action pairs in the dataset and those that remain unidentified (unknown).

For states that are in the dataset, we can establish a relationship based on the definition of the uncertainty set. Specifically:

$$\mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}_{P_r \sim \mathcal{U}_r} (\|P_0(s,a) - P_r(s,a)\|_1) \le \max(\beta, \beta_r) = \beta.$$

From this relation, it becomes straightforward to deduce the upper-bound performance of a robust policy. On the other hand, when considering the lower-bound performance of a robust policy, defined as

$$\pi_r = \arg\max_{\pi} \mathbb{E}_{P \sim \mathcal{U}_r} J_\rho(\pi, P),$$

subject to the constraint

$$\mathbb{E}_{P \sim \mathcal{U}_r} [\mathbb{E}_{s,a} D_{\mathrm{TV}}(P, P_B)] \le \beta_r.$$

The "untrained" region is quantified by the difference $(1 - p_r)\beta + p_r(\beta - \beta_r) = \beta - p_r\beta_r$.

Under these conditions, it's clear that the probability of disadvantageous scenarios for the robust policy at each step is $1 - (\beta - p_r\beta_r)$. Consequently, the chance of decoupling at a specific time t is at most $1 - (1 - (\beta - p_r\beta_r))^t$. Then we have:

$$\mathbb{E}_{P_r \sim \mathcal{U}_r} J_{\rho_0^B}(\pi_r, P_r) - \mathbb{E}_{P_0 \sim \mathcal{U}} J_{\rho_0}(\pi_r, P_0) \tag{24}$$

$$\ge \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} [D_{\mathrm{TV}}(\rho_0, \rho_0^B)] + \sum_t \gamma^t (1 - (1 - (\beta - p_r\beta_r))^t) \cdot 2 \cdot R_{\max} + \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi_r}}] \tag{25}$$

$$\ge \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} [D_{\mathrm{TV}}(\rho_0, \rho_0^B)] + \frac{2R_{\max}\gamma(\beta - p_r\beta_r)}{(1-\gamma)(1 - \gamma \cdot (1 - (\beta - p_r\beta_r)))} + \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi_r}}] \tag{26}$$

$$\ge \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} [D_{\mathrm{TV}}(\rho_0, \rho_0^B)] + \frac{2R_{\max}\gamma}{(1-\gamma)^2}(\beta - p_r\beta_r) + \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi_r}}] \tag{27}$$

□

For ease of exposition, we restate Theorem 2 as follows.

Theorem (2). *For an ϵ_{π_r} sub-optimal risk-aware policy, we have:*

$$\begin{aligned}\mathbb{E}_{P_0 \sim \mathcal{U}}[J_{\rho_0}(\pi^*, P_0)] - \mathbb{E}_{P_0 \sim \mathcal{U}}[J_{\rho_0}(\pi_r, P_0)] \le & \epsilon_{\pi_r} + \frac{4R_{max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} [D_{TV}(\rho_0, \rho_0^B)] + \frac{4\gamma R_{max}}{(1-\gamma)^2}(\beta - \frac{1}{2}p_r\beta_r) \\ & + \frac{2R_{max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}[\gamma^{T_{\mathcal{V}}^{\pi_r}}] + \frac{2R_{max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}[\gamma^{T_{\mathcal{V}}^{\pi^*}}].\end{aligned} \tag{28}$$

Proof. By Lemma 3, we have:

$$\begin{aligned}\mathbb{E}_{P_0 \sim \mathcal{U}}[J_{\rho_0}(\pi_r, P_0)] \ge & J_{\rho_0^B}(\pi_r, P_B) - \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} [D_{\mathrm{TV}}(\rho_0, \rho_0^B)] - \frac{2R_{\max}\gamma}{(1-\gamma)^2}\beta - \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi_r}}] \\ \ge & J_{\rho_0^B}(\pi^*, P_B) - \epsilon_{\pi_r} \qquad (29) \\ & - \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} [D_{\mathrm{TV}}(\rho_0, \rho_0^B)] - \frac{2R_{\max}\gamma}{(1-\gamma)^2}\beta - \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi_r}}] \qquad (30) \\ \ge & \mathbb{E}_{P_0 \sim \mathcal{U}}[J_{\rho_0}(\pi^*, P_0)] - \epsilon_{\pi_r} \qquad (31) \\ & - \frac{4R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} [D_{\mathrm{TV}}(\rho_0, \rho_0^B)] - \frac{4R_{\max}\gamma}{(1-\gamma)^2}(\beta - \frac{1}{2}p_r\beta_r) - \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi_r}}] \qquad (32) \\ & - \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{P_0 \sim \mathcal{U}} \mathbb{E}[\gamma^{\mathrm{T}_{\mathcal{V}}^{\pi^*}}] \qquad (33)\end{aligned}$$

□

A.3 Proof of Theorem 3

Theorem (3). *[Relaxed State-Adversarial Policy Performance Lower Bound] For an $\epsilon_{\pi_{RA}}$ sub-optimal relaxed state-adversarial Policy policy, we have*

$$\begin{aligned}\mathbb{E}_{P_0\sim\mathcal{U}}[J_{\rho_0}(\pi^*, P_0)] - \mathbb{E}_{P_0\sim\mathcal{U}}[J_{\rho_0}(\pi_{RA}, P_0)] \leq & \epsilon_{\pi_{RA}} + \frac{4R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}[D_{TV}(\rho_0, \rho_0^B)] \\ & + \frac{4\gamma R_{max}}{(1-\gamma)^2}(\beta - \frac{1}{2}p_{RA}(1-\alpha)) + \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{T_{\mathcal{V}}^{\pi_{RA}}}] \\ & + \frac{2R_{max}}{1-\gamma}\mathbb{E}_{P_0\sim\mathcal{U}}\mathbb{E}[\gamma^{T_{\mathcal{V}}^{\pi^*}}]. \end{aligned} \tag{34}$$

Proof. Consider the relaxed state-adversarial policy:

$$\pi_{\text{RA}} = \arg\max_{\pi} \mathbb{E}_{P\sim\mathcal{U}_\epsilon^\pi} J_\rho(\pi, P),$$

subject to the constraint:

$$\mathbb{E}_{P\sim\mathcal{U}_\epsilon^\pi}(\mathbb{E}_{s,a} D_{\text{TV}}(P, P_B)) \leq 1-\alpha.$$

By setting $\beta_r = 1-\alpha$ and $p_r = p_{\text{RA}}$, we can directly apply Theorem 3. This leads us to the desired assertion, thereby completing the proof. □