Appendix

A. Grid World Example

We illustrate the over-conservatism problem in Figure 1 using a shortest-path grid world environment. In this environment, the four-adjacent neighbors of each state s are denoted by $\mathcal{N}(s)$. During training, we perturbed states to the nearby worst states to help agents against the uncertainty of environments. The state-adversarial value iteration algorithm is shown in Algorithm 2. To achieve a clear explanation, we first denote the coordinate of the bottom left corner by grid(0, 0). Accordingly, the goal state is located at grid(0, 3), and the trap state is at grid(2, 2). Let s_a and s_b be the states at grid(1, 3) and grid(1, 2), respectively. Since s_a is adjacent to the goal state, the value $V(s_a)$ will increase because of the high reward $R(s_a, a_a)$. However, the value $V(s_a)$ will never propagate to state s_b because only the worst value around s_b is used in the TD update. Since the policy would be penalized by a -1 reward at each step (to learn how to reach the goal state as soon as possible), and the positive reward at the goal state can only propagate to grid(0, 2) and grid(1, 3), the value $V(s_b)$ decreases by the negative (s_b, a_b) after each TD update.

Following the algorithm, we show how the naive state-adversarial method updates the value of (s,a) = (grid(1,2), UP). Initially, all state values are 0.

$$\begin{aligned}
\text{At } t=0, \quad Q(s,a) &= Q(\text{grid(1,2), UP}) = R(\text{grid(1,2), UP}) + \gamma \cdot \min_{s' \in \mathcal{N}(\text{grid(1,3)})} V(s') \\
&= -1 + 0.99 \cdot \min(V(\text{grid(1,3)}), V(\text{grid(0,3)}), V(\text{grid(2,3)}), V(\text{grid(1,2)})) \\
&= -1 + 0.99 \cdot \min(0,0,0,0) = -1.
\end{aligned}$$

$$\begin{aligned}
\text{At } t=1, \quad Q(s,a) &= Q(\text{grid(1,2), UP}) = R(\text{grid(1,2), UP}) + \gamma \cdot \min_{s' \in \mathcal{N}(grid(1,3))} V(s') \\
&= -1 + 0.99 \cdot \min(V(\text{grid(1,3)}), V(\text{grid(0,3)}), V(\text{grid(2,3)}), V(\text{grid(1,2)}) \\
&= -1 + 0.99 \cdot \min(-1,0,-1,-1) = -1.99.
\end{aligned}$$

$$\begin{aligned}
\text{At } t=2, \quad Q(s,a) &= Q(\text{grid(1,2), UP}) = R(\text{grid(1,2), UP}) + \gamma \cdot \min_{s' \in \mathcal{N}(\text{grid(1,3)})} V(s') \\
&= -1 + 0.99 \cdot \min(V(\text{grid(1,3)}), V(\text{grid(0,3)}), V(\text{grid(2,3)}), V(\text{grid(1,2)})) \\
&= -1 + 0.99 \cdot \min(-1,0,-1.99,-1.99) = -2.97.
\end{aligned}$$

As can be seen, although the agent took the action "UP" at grid(1,2) to reach grid(1,3), it considers the minimum value among the neighbours of grid(1, 3) for the robust purpose. Hence, the TD update reduces the value $Q(s,a) = Q$(grid(1,2), UP) at each step. In other words, the agent cannot learn how to move to the goal state because the value of the goal state does not propagate outward during value iteration. Even worse, the agent would move toward the trap state if it is nearby due to the compounding effect of TD updates and the worst-case state-adversarial perturbations. The phenomenon appears after updating state values 12 times.

Algorithm 2 State-Adversarial Perturbation with Greedy Policy

Input : MDP $(\mathcal{S}, \mathcal{A}, P_0, R, \gamma)$, number of iterations T, P_0 is the nominal transition kernel

Initialize the $Q_0(s,a)$ value function with 0.
for $t = 1, \ldots, T$ **do**
  for *state* s*, action* a **do**
    $Q_t(s,a) = R(s,a) + \gamma \sum_{s'} P_0(s'|s,a) \min_{s'' \in \mathcal{N}(s')} (\max_a Q_{t-1}(s'', a))$
  end
end

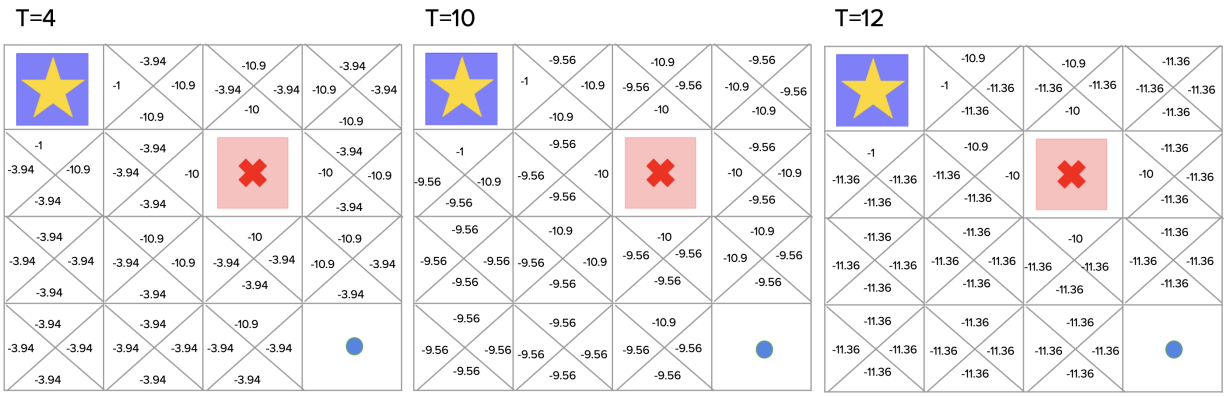

Figure 5. The 4×4 shortest-path grid world. The dot, star, and cross icons indicate the initial, goal, and trap states, respectively. The agent can move either up, down, left, or right at each step and earn $+0$ and -10 rewards when reaching the goal and trap states, respectively. In addition, the agent would be penalized by a -1 reward at each step and learn to reach the goal state as quick as possible.

B. Bellman Equation of Relaxed State-Adversarial Policy Optimization

Given a fixed policy π, we aim to estimate its value using the temporal difference learning. Based on the relaxed state-adversarial transition kernel (Equation 6), we obtain the value function

$$V_\epsilon^{\pi,\alpha}(s) := \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0) + \gamma \mathbb{E}_{s_1\sim P_\epsilon^{\pi,\alpha}(\cdot|s_0,a_0)}\Big[\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}R(s_1,a_1) \tag{11}$$

$$+ \gamma \mathbb{E}_{s_2\sim P_\epsilon^{\pi,\alpha}(\cdot|s1,a1)}\Big[\mathbb{E}_{a_2\sim\pi(\cdot|s_2)}R(s_2,a_2) + ...\Big]\Big]\Big] \tag{12}$$

The corresponding Bellman operator is derived as

$$\mathcal{T}_\epsilon^{\pi,\alpha}V(s) = \mathbb{E}_{a\sim\pi}\Big[R(s,a) + \gamma \mathbb{E}_{s'\sim P_0(\cdot|s,a)}\Big(\alpha V(s') + (1-\alpha)\min_{s''\in\mathcal{N}(s')} V(s'')\Big)\Big] \tag{13}$$

Proof.

$$V_\epsilon^{\pi,\alpha}(s_0) = \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0) + \gamma \mathbb{E}_{s_1\sim P_\epsilon^{\pi,\alpha}(\cdot|s_0,a_0)}\Big[\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}R(s_1,a_1) \tag{14}$$

$$+ \gamma \mathbb{E}_{s_2\sim P_\epsilon^{\pi,\alpha}(\cdot|s1,a1)}\Big[\mathbb{E}_{a_2\sim\pi(\cdot|s_2)}R(s_2,a_2) + ...\Big]\Big]\Big] \tag{15}$$

$$= \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0) + \gamma \mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0)}\Big[\alpha\Big(\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}R(s_1,a_1) \tag{16}$$

$$+ \gamma \mathbb{E}_{s_2\sim P_\epsilon^{\pi,\alpha}(\cdot|s1,a1)}\Big[\mathbb{E}_{a_2\sim\pi(\cdot|s_2)}R(s_2,a_2) + ...\Big)\Big] \tag{17}$$

$$+ (1-\alpha)\Big(\min_{s_1'\in\mathcal{N}_\epsilon(s_1)} \mathbb{E}_{a_1'\sim\pi(\cdot|s_1')}R(s_1',a_1') \tag{18}$$

$$+ \gamma \mathbb{E}_{s_2'\sim P_\epsilon^{\pi,\alpha}(\cdot|s_1',a_1')}\Big[\mathbb{E}_{a_2'\sim\pi(\cdot|s_2')}R(s_2',a_2') + ...\Big)\Big]\Big]\Big] \tag{19}$$

$$= \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0) + \gamma \mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0)}\Big(\alpha V_\epsilon^{\pi,\alpha}(s_1) + (1-\alpha)\min_{s_1'\in\mathcal{N}_\epsilon(s_1)} V_\epsilon^{\pi,\alpha}(s_1')\Big)\Big], \tag{20}$$

□

C. Proof of Lemma 1

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

Lemma (Monotonicity of Average Value in Perturbation Strength)**.** Under the setting of state-adversarial MDP, the value of the local minimum monotonically decreases as the bounded radius σ increases. Let x be a positive real number. Under any

policy π, the total expected return J satisfies

$$J(\pi|P^{\pi}_{\sigma}) \geq J(\pi|P^{\pi}_{\sigma+x}). \tag{21}$$

Proof.

$$\begin{aligned}
V^{\pi}(s_0|P^{\pi}_{\sigma}) =& \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0)+\gamma\mathbb{E}_{s_1^{\sigma}\sim P^{\pi}_{\sigma}(\cdot|s_0,a_0)}\Big[V^{\pi}(s_1^{\sigma}|P^{\pi}_{\sigma})\Big]\Big] && (22)\\
=& \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0)+\gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0),s_1^{\sigma}=\text{argmin}V^{\pi}(s),s\in\mathcal{N}_{\sigma}(s_1)}\Big[V^{\pi}(s_1^{\sigma}|P^{\pi}_{\sigma})\Big]\Big] && (23)\\
\geq& \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0)+\gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0),s_1^{\sigma+x}=\text{argmin}V^{\pi}(s),s\in\mathcal{N}_{\sigma+x}(s_1)}\Big[V^{\pi}(s_1^{\sigma+x}|P^{\pi}_{\sigma})\Big]\Big] && (24)\\
=& \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0)+\gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0),s_1^{\sigma+x}=\text{argmin}V^{\pi}(s),s\in\mathcal{N}_{\sigma+x}(s_1),a_1\sim\pi(\cdot|s_1^{\sigma+x})}\Big[R(s_1^{\sigma+x},a_1) && (25)\\
&+\gamma\mathbb{E}_{s_2^{\sigma}\sim P^{\pi}_{\sigma}(\cdot|s_1^{\sigma+x},a_1)}[V^{\pi}(s_2^{\sigma}|P^{\pi}_{\sigma})]\Big]\Big] && (26)\\
=& \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0)+\gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0),s_1^{\sigma+x}=\text{argmin}V^{\pi}(s),s\in\mathcal{N}_{\sigma+x}(s_1),a_1\sim\pi(\cdot|s_1^{\sigma+x})}\Big[R(s_1^{\sigma+x},a_1) && (27)\\
&+\gamma\mathbb{E}_{s_2\sim P_0(\cdot|s_1^{\sigma+x},a_1),s_2^{\sigma}=\text{argmin}V^{\pi}(s),s\in\mathcal{N}_{\sigma}(s_2)}[V^{\pi}(s_2^{\sigma}|P^{\pi}_{\sigma})]\Big]\Big] && (28)\\
\geq& \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0)+\gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0),s_1^{\sigma+x}=\text{argmin}V^{\pi}(s),s\in\mathcal{N}_{\sigma+x}(s_1),a_1\sim\pi(\cdot|s_1^{\sigma+x})}\Big[R(s_1^{\sigma+x},a_1) && (29)\\
&+\gamma\mathbb{E}_{s_2\sim P_0(\cdot|s_1^{\sigma+x},a_1),s_2^{\sigma+x}=\text{argmin}V^{\pi}(s),s\in\mathcal{N}_{\sigma+x}(s_2)}[V^{\pi}(s_2^{\sigma+x}|P^{\pi}_{\sigma})]\Big]\Big] && (30)\\
\geq& \mathbb{E}_{a_i\sim\pi,s_i\sim P^{\pi}_{\sigma+x})}\Big[R(s_0,a_0)+\gamma R(s_1^{\sigma+x},a_1)+\gamma^2R(s_2^{\sigma+x},a_2)+...\Big] && (31)\\
=& \mathbb{E}_{a_0\sim\pi}\Big[R(s_0,a_0)+\gamma\mathbb{E}_{s_1^{\sigma+x}\sim P^{\pi}_{\sigma+x}(\cdot|s_0,a_0)}\Big[V^{\pi}(s_1^{\sigma+x}|P^{\pi}_{\sigma+x})\Big]\Big] && (32)\\
=& V^{\pi}(s_0|P^{\pi}_{\sigma+x}) && (33)
\end{aligned}$$

where the inequality holds because $\sigma + x$ is a larger perturbation radius than σ. Recall that μ denotes the initial state distribution. Then, we have

$$\begin{aligned}
J(\pi|P^{\pi}_{\sigma}) =& \mathbb{E}_{s_0\sim\mu}[V^{\pi}(s_0|P^{\pi}_{\sigma})] && (34)\\
\geq& \mathbb{E}_{s_0\sim\mu}[V^{\pi}(s_0|P^{\pi}_{\sigma+x})] && (35)\\
=& J(\pi|P^{\pi}_{\sigma+x}). && (36)
\end{aligned}$$

□

D. Proof of Lemma 2

We prove Lemma 2 based on the continuity of the expected discounted return $J(\pi|P^{\pi,\alpha}_{\epsilon})$ with the relaxation parameter $\alpha \in [0,1]$. Based on the continuity of α, the assumption is reasonable because similar values of α imply similar transition kernels (Equation 6). We show this property by the continuity of the epsilon-delta definition as follow. Let $\alpha_1, \alpha_2 \in [0,1]$ be two relaxation parameters. As long as $|\alpha_1 - \alpha_2|$ is small, the state perturbations are similar, which also implies that the total returns would be similar due to bounded rewards. Therefore, by expressing $J(\pi|P^{\pi,\alpha}_{\epsilon}) = \mathbb{E}_{s_0\sim\mu,a_0\sim\pi}[R(s_0,a_0)+\gamma\sum_{s_1}P^{\pi,\alpha}_{\epsilon}(s_1|s_0,a_0)V^{\pi}(s_1|P^{\pi,\alpha}_{\epsilon})]$ and using the boundedness of total return, one can verify that for any $\epsilon_c > 0$, there exist a $\delta_c > 0$, such that $|\alpha_1 - \alpha_2| < \delta_c$ and $|J(\pi|P^{\pi,\alpha_1}_{\epsilon}) - J(\pi|P^{\pi,\alpha_2}_{\epsilon})| < \epsilon_c$. Hence, $J(\pi|P^{\pi,\alpha}_{\epsilon})$ is continuous in α. Now we are ready to prove Lemma 2.

Lemma (Relaxation parameter α as a prior distribution $\mathcal{D}$ in domain randomization)**.** For any distribution $\mathcal{D}$ over the state-adversarial uncertainty set $\mathcal{U}^{\pi}_{\epsilon}$, there must exist an $\alpha \in [0,1]$ such that

$$\mathbb{E}_{P\sim\mathcal{D}}[J(\pi|P)] = J(\pi|P^{\pi,\alpha}_{\epsilon}).$$

Proof. Based on Lemma 1, we have

$$J(\pi|P^{\pi,0}_{\epsilon}) = J(\pi|P^{\pi}_{\epsilon}) \leq \mathbb{E}_{P\sim\mathcal{D}}[J(\pi|P)] \leq J(\pi|P^{\pi,1}_{\epsilon}) = J(\pi|P_0) \tag{37}$$

Under the condition that $J(\pi|P_\epsilon^{\pi,\alpha})$ is a continuous function, by Intermediate Value Theorem, we know that there exists $\alpha \in [0,1]$ such that

$$\mathbb{E}_{P\sim\mathcal{D}}[J(\pi|P)] = J(\pi|P_\epsilon^{\pi,\alpha}). \tag{38}$$

□

E. Proof of Theorem 1

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

Theorem (A Direct Connection Between the Average-Case and the Worst-Case Returns)**.** Given a nominal MDP with transition kernel P_0 along with a state-adversarial uncertainty set $\mathcal{U}_\epsilon^\pi$, for any distribution $\mathcal{D}$ over $\mathcal{U}_\epsilon^\pi$, upon an update from the current policy π to a new policy $\tilde{\pi}$, the following bound holds (Jiang et al., 2021):

$$J(\tilde{\pi}|P_\epsilon^\pi) \geq \mathbb{E}_{P\sim\mathcal{D}}[J(\tilde{\pi}|P)] - 2R_{\max}\frac{\gamma\mathbb{E}_{P\sim\mathcal{D}}[d_{\mathrm{TV}}(P_\epsilon^\pi\|P)]}{(1-\gamma)^2} - 4R_{\max}\frac{d_{\mathrm{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)^2}, \tag{39}$$

where $d_{\mathrm{TV}}(\pi\|\tilde{\pi})$ indicates the total variation divergence between π and $\tilde{\pi}$, and P_ϵ^π is the worst-case state-adversarial transition kernel.

Proof. To begin with, note that

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) = J(\tilde{\pi}|P_\epsilon^\pi) - J(\pi|P_\epsilon^\pi) + J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}). \tag{40}$$

Throughout the proof, we use $p_\pi^t(s|P)$ to denote the state distribution at time t under a transition kernel P and a policy π. For ease of notation, we also define $p_\pi^t(s,a|P) := \pi(a|s)p_\pi^t(s|P)$. For the last two terms of Equation 40,

$$\begin{aligned}
&|J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha})| && (41)\\
&= |\sum_t \gamma^t \sum_{s,a}(p_\pi^t(s,a|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s,a|P_\epsilon^{\pi,\alpha}))R(s,a)| && (42)\\
&\leq \sum_t \gamma^t \sum_{s,a}|(p_\pi^t(s,a|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s,a|P_\epsilon^{\pi,\alpha})|R(s,a) && (43)\\
&\leq 2R_{\max}\sum_t \gamma^t[d_{\mathrm{TV}}(p_\pi^t(s,a|P_\epsilon^\pi)\|p_{\tilde{\pi}}^t(s,a|P_\epsilon^{\pi,\alpha}))] && (44)\\
&\text{because } p_\pi^t(s,a|P_\epsilon^\pi) = \pi(a|s)p_\pi^t(s|P_\epsilon^\pi) \text{ and } p_{\tilde{\pi}}^t(s,a|P_\epsilon^{\pi,\alpha}) = \tilde{\pi}(a|s)p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha}) && (45)\\
&\leq 2R_{\max}\big[\mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}[d_{\mathrm{TV}}(\pi(a|s')\|\tilde{\pi}(a|s'))] && (46)\\
&+ d_{\mathrm{TV}}(p_\pi^t(s|P_\epsilon^\pi)\|p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha}))\big] && (47)
\end{aligned}$$

For the second term of Equation 47,

$$\begin{aligned}
&d_{\mathrm{TV}}(p_\pi^t(s|P_\epsilon^\pi)\|p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})) && (48)\\
&\leq t\cdot\max_t \mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}[d_{\mathrm{TV}}(p_\pi(s|s',a,P_\epsilon^\pi)\|p_{\tilde{\pi}}(s|s',a,P_\epsilon^{\pi,\alpha}))] && (49)\\
&\text{because } p_\pi(s|s',a,P_\epsilon^\pi) = \sum_a P_\epsilon^\pi(s|s',a)\pi(a|s') && (50)\\
&\leq t\cdot\max_t \mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}\mathbb{E}_{a\sim\pi(\cdot|s')}[d_{\mathrm{TV}}(P_\epsilon^\pi(s|s',a)\|P_\epsilon^{\pi,\alpha}(s|s',a,)] && (51)\\
&+ t\cdot\max_t \mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}d_{\mathrm{TV}}(\pi(s|s')\|\tilde{\pi}(a|s')) && (52)
\end{aligned}$$

Then we can rewrite Equation 47as:

$$\begin{aligned}
&J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) && (53)\\
&\geq -2R_{\max}\sum_t \gamma^t[(t+1)\max_t \mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}d_{\mathrm{TV}}(\pi(a|s')\|\tilde{\pi}(a|s') && (54)\\
&- t\max_t \mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}\mathbb{E}_{a\sim\pi(\cdot|s')}d_{\mathrm{TV}}(P_\epsilon^\pi(s|s',a)\|P_\epsilon^{\pi,\alpha}(s|s',a)) && (55)
\end{aligned}$$

Similar to the derivation of Equation 47,

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\pi|P_\epsilon^\pi) \tag{56}$$
$$\geq -2R_{\max} \sum_t \gamma^t [(t+1) \max_t \mathbb{E}_{s' \sim p_\pi^t(\cdot|p_w)} d_{\mathrm{TV}}(\pi(a|s') \| \tilde{\pi}(a|s'))] \tag{57}$$

and rewrite Equation 40 as following,

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) \tag{58}$$
$$\geq -2R_{\max} \sum_t \gamma^t [2(t+1) \max_t \mathbb{E}_{s' \sim p_\pi^t(\cdot|P_\epsilon^\pi)} d_{\mathrm{TV}}(\pi(a|s') \| \tilde{\pi}(a|s') \tag{59}$$
$$- t \max_t \mathbb{E}_{s' \sim p_\pi^t(\cdot|P_\epsilon^\pi)} \mathbb{E}_{a \sim \pi(\cdot|s')} d_{\mathrm{TV}}(P_\epsilon^\pi(s|s',a)) \| P_\epsilon^{\pi,\alpha}(s|s',a)] \tag{60}$$
$$= -2R_{\max} \sum_t \gamma^t [2(t+1) \max_t \mathbb{E}_{s' \sim p_\pi^t(\cdot|P_\epsilon^\pi)} d_{\mathrm{TV}}(\pi(a|s') \| \tilde{\pi}(a|s') \tag{61}$$
$$- t \mathbb{E}_{P \sim \mathcal{D}} [d_{\mathrm{TV}}(P_\epsilon^\pi \| P)]] \tag{62}$$
$$= -2R_{\max} \frac{\gamma \mathbb{E}_{P \sim \mathcal{D}} [d_{\mathrm{TV}}(P_\epsilon^\pi \| P)]}{(1-\gamma)^2} - 4R_{\max} \frac{d_{\mathrm{TV}}(\pi \| \tilde{\pi})}{(1-\gamma)^2} \tag{63}$$

□

F. Proof of Theorem 2

We consider the difference of the expected discounted return under two different state-adversarial transition kernels. To prove this theorem, we utilize the definition of the reward function of the corresponding Markov Reward Process (MRP) with respect to policy π by $R(s) \coloneqq \sum_a \pi(a|s) R(s,a)$. For convenience, we restate Theorem 2 as follows.

Theorem (A Sharper Characterization of the Connection Between Worst-Case and Average-Case Returns)**.** Consider a nominal MDP with a δ-smooth transition kernel and an L_r-Lipschitz reward function (cf. Definitions 4-5). Let $\mathcal{U}_\epsilon^\pi$ be the state-adversarial uncertainty set. For any $\alpha \in [0,1]$, upon an update from the current policy π to a new policy $\tilde{\pi}$, the following bound holds:

$$J(\tilde{\pi}|P_\epsilon^\pi) \geq J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) - \frac{4\gamma(\epsilon+\delta)L_r\alpha}{(1-\gamma)^3} - \frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{\max}) d_{\mathrm{TV}}(\pi \| \tilde{\pi})}{(1-\gamma)^3}, \tag{64}$$

where $d_{\mathrm{TV}}(\pi \| \tilde{\pi})$ is the total variation divergence between π and $\tilde{\pi}$, $P_\epsilon^{\pi,\alpha}$ and P_ϵ^π are the relaxed and worst-case state-adversarial transition kernels within the uncertainty set $\mathcal{U}_\epsilon^\pi$, respectively.

We first introduce the following supporting lemma before proving Theorem 2.

Lemma 3. Given any $\epsilon > 0$, any initial state $s_0 \in \mathcal{S}$, and a policy π, let s_t and $\tilde{s}_t$ denote the state at time step t under the nominal transition kernel P_0 and the state-adversarial transition kernel P_ϵ^π, respectively. Then, we have $\|s_t - \tilde{s}_t\| \leq 2t(\epsilon+\delta)$, with probability one.

Proof of Lemma 3. We prove this by induction. If $t = 1$, we know the difference between s_1 and $\tilde{s}_1$ results from the perturbation at time step 1. Therefore, we have $\|s_1 - \tilde{s}_1\| \leq \epsilon$.

Next, suppose that at time step $t = \tau$, we have $\|s_\tau - \tilde{s}_\tau\| \leq 2\tau(\epsilon+\delta)$. Then, we have

$$\|s_{\tau+1} - \tilde{s}_{\tau+1}\| = \|s_{\tau+1} - s_\tau + s_\tau - \tilde{s}_\tau + \tilde{s}_\tau - \tilde{s}_{\tau+1}\| \tag{65}$$
$$\leq \|s_{\tau+1} - s_\tau\| + \|s_\tau - \tilde{s}_\tau\| + \|\tilde{s}_\tau - \tilde{s}_{\tau+1}\| \tag{66}$$
$$\leq \delta + 2\tau(\epsilon+\delta) + (\epsilon+\delta) \tag{67}$$
$$\leq 2(\tau+1)(\epsilon+\delta), \tag{68}$$

where Equation 65 holds by the triangle inequality, Equation 66 follows the definition of δ, the assumption in the induction step, and the fact that $\tilde{s}_{\tau+1}$ is obtained from $\tilde{s}_\tau$ via the transitions determined by P_0 and the perturbation of strength ϵ. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. To begin with, we have

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) = J(\tilde{\pi}|P_\epsilon^\pi) - J(\pi|P_\epsilon^\pi) + J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) \tag{69}$$

As in the proof of Theorem 1, we use $p_\pi^t(s|P)$ to denote the state distribution at time t under a transition kernel P and a policy π. For ease of notation, we also define $p_\pi^t(s,a|P) := \pi(a|s)p_\pi^t(s|P)$. For the last two term of Equation 69,

$$\begin{aligned}
&|J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha})| && (70)\\
=&|\sum_t \gamma^t \sum_{s,a} \pi(a|s)p_\pi^t(s|P_\epsilon^\pi)R(s,a) - \tilde{\pi}(a|s)p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})R(s,a)| && (71)\\
=&|\sum_t \gamma^t \sum_{s,a} \pi(a|s)[p_\pi^t(s|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})]R(s,a) + (\pi(a|s) - \tilde{\pi}(a|s))p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})R(s,a)| && (72)\\
\leq&\sum_t \gamma^t \sum_{s,a} |\pi(a|s)[p_\pi^t(s|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})]R(s,a)| + |(\pi(a|s) - \tilde{\pi}(a|s))p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})R(s,a)| && (73)
\end{aligned}$$

For the first term of Equation 73, we have the t-step state distribution:

$$\begin{aligned}
&|p_\pi^t(s|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})| && (74)\\
\leq&|p_\pi^t(s|P_\epsilon^\pi) - p_\pi^t(s|P_\epsilon^{\pi,\alpha})| + |p_\pi^t(s|P_\epsilon^{\pi,\alpha}) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})| && (75)\\
& && (76)
\end{aligned}$$

Now we prove the following inequality.

$$\begin{aligned}
&\sum_s |p_\pi^t(s|P_\epsilon^\pi) - p_\pi^t(s|P_\epsilon^{\pi,\alpha})| && (77)\\
=&\sum_s |\sum_{s'} p_\pi^{t-1}(s'|P_\epsilon^\pi)(\sum_{k,Z_{ks}=1} P_0(k|s')) && (78)\\
&- (1-\alpha)\sum_{s'} p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})(\sum_{k,Z_{ks}=1} P_0(k|s')) && (79)\\
&- \alpha \sum_{s'} p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})P_0(s|s'), && (80)\\
\leq&\sum_s \sum_{s'} |p_\pi^{t-1}(s'|P_\epsilon^\pi) - p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})|(\sum_{k,Z_{ks}=1} P_0(k|s')) && (81)\\
&+ \alpha \sum_s \sum_{s'} p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})|(\sum_{k,Z_{ks}=1} P_0(k|s')) - P_0(s|s')| && (82)\\
\leq&\sum_{s'} |p_\pi^{t-1}(s'|P_\epsilon^\pi) - p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})| && (83)\\
&+ \alpha \cdot \max_{s'} \sum_s |(\sum_{k,Z_{ks}=1} P_0(k|s')) - P_0(s|s')| && (84)\\
\leq&\sum_{s'} |p_\pi^{t-1}(s'|P_\epsilon^\pi) - p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})| + 2\alpha && (85)\\
=&2\alpha t && (86)
\end{aligned}$$

where $P_0(s|s') = \sum_a \pi(a|s')P_0(s|s',a)$, $Z_{ks} = Z_\epsilon^{\tilde{\pi}}(k,s)$ is the state perturbation matrix, and Equations 78 to 80 follow from the definition of state perturbation transition kernel. Note that $\sum_{k,Z_{ks}=1} P_0(k|s')$ is the state probability after considering the perturbation, and Equation 83 holds because $\sum_s \sum_{k,Z_{ks}=1} P_0(k|s') = 1$. In addition, Equation 86 is obtained by recursively applying Equations 78 to 85 to the first term of Equation 85.

For the first two terms of Equation 75, we have

$$
\begin{aligned}
&|p_\pi^t(s|P_\epsilon^\pi) - p_\pi^t(s|P_\epsilon^{\pi,\alpha})| && (87)\\
\leq &\sum_s |p_\pi^t(s|P_\epsilon^\pi) - p_\pi^t(s|P_\epsilon^{\pi,\alpha})| && (88)\\
\leq &2\alpha t, && (89)
\end{aligned}
$$

For the last two terms of Equation 75, we have

$$
\begin{aligned}
&|p_\pi^t(s|P_\epsilon^{\pi,\alpha}) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})| && (90)\\
\leq &\sum_s |p_\pi^t(s|P_\epsilon^{\pi,\alpha}) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})| && (91)\\
= &\sum_s \sum_{s',a} \Big(|p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})\pi(a|s') - p_{\tilde{\pi}}^{t-1}(s'|P_\epsilon^{\pi,\alpha})\tilde{\pi}(a|s')|\Big)\Big(\alpha P_0(s|s',a) + (1-\alpha)\sum_{k,Z_{ks}=1} P_0(k|s',a)\Big) && (92)\\
\leq &\sum_{s',a} |p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})\pi(a|s') - p_{\tilde{\pi}}^{t-1}(s'|P_\epsilon^{\pi,\alpha})\tilde{\pi}(a|s')| && (93)\\
\leq &\sum_{s',a} |p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})\pi(a|s') - p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})\tilde{\pi}(a|s')| + \sum_{s',a} |p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})\tilde{\pi}(a|s') - p_{\tilde{\pi}}^{t-1}(s'|P_\epsilon^{\pi,\alpha})\tilde{\pi}(a|s')| && (94)\\
= &\sum_{s',a} |p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})(\pi(a|s') - \tilde{\pi}(a|s'))| + \sum_{s',a} |(p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha}) - p_{\tilde{\pi}}^{t-1}(s'|P_\epsilon^{\pi,\alpha}))\tilde{\pi}(a|s')| && (95)\\
\leq &2td_{\mathrm{TV}}(\pi\|\tilde{\pi}) && (96)
\end{aligned}
$$

Hence, we can rewrite Equation 74 as:

$$
|p_\pi^t(s|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})| \leq 2\alpha t + 2td_{\mathrm{TV}}(\pi\|\tilde{\pi}) \tag{97}
$$

where Equation 97 holds by applying Equation 96 and 89.

Under the condition that the reward function $R(s,a)$ is L_r-Lipschitz continuous in state, we know the reward function of the MRP under policy π is also L_r-Lipschitz continuous, i.e., $|R(s_1) - R(s_2)| \leq 2t(\epsilon+\delta)L_r$ if $\|s_1 - s_2\| \leq 2t(\epsilon+\delta)$. By Lemma 3, for every probability density in $p_\pi^t(s|P_\epsilon^\pi)$, there exists a corresponding density point transited by $P_\epsilon^{\pi,\alpha}$, and the state distance between these two density is less than $2t(\epsilon+\delta)$. Hence, their reward difference is bounded by $2t(\epsilon+\delta)L_r$. By Equation 97, for every state, the total probability density difference is bounded by $2\alpha t + 2td_{\mathrm{TV}}(\pi\|\tilde{\pi})$. The total reward difference at time t will be

$$
\begin{aligned}
&|\sum_{s,a} \pi(a|s)[p_\pi^t(s|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})]R(s,a)| && (98)\\
= &|\sum_s [p_\pi^t(s|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})]R(s)| && (99)\\
\leq &(2\alpha t + 2td_{\mathrm{TV}}(\pi\|\tilde{\pi}))\cdot 2t(\epsilon+\delta)L_r && (100)
\end{aligned}
$$

Combining Equations 70and 100, we have

$$
\begin{aligned}
&|J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha})| && (101)\\
\leq &\sum_t \gamma^t \Big((2\alpha t + 2td_{\mathrm{TV}}(\pi\|\tilde{\pi}))\cdot 2t(\epsilon+\delta)L_r + 2R_{\max}d_{\mathrm{TV}}(\pi\|\tilde{\pi})\Big) && (102)\\
= &\sum_t \gamma^t 4\alpha t^2(\epsilon+\delta)L_r + \sum_t \gamma^t 4t^2(\epsilon+\delta)L_r d_{\mathrm{TV}}(\pi\|\tilde{\pi}) + \sum_t \gamma^t 2R_{\max}d_{\mathrm{TV}}(\pi\|\tilde{\pi}) && (103)\\
= &\frac{\gamma(4\alpha(\epsilon+\delta)L_r)}{(1-\gamma)^3} + \frac{4\gamma(\epsilon+\delta)L_r d_{\mathrm{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)^3} + 2\frac{R_{\max}d_{\mathrm{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)} && (104)
\end{aligned}
$$

When policy π is updated to $\tilde{\pi}$, $J(\pi|P_\epsilon^\pi) \leq J(\tilde{\pi}|P_\epsilon^{\pi,\alpha})$. Then we have

$$J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) \tag{105}$$

$$\geq -\frac{\gamma(4\alpha(\epsilon+\delta)L_r)}{(1-\gamma)^3} - \frac{4\gamma(\epsilon+\delta)L_r d_{\text{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)^3} - \frac{2R_{\max} d_{\text{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)} \tag{106}$$

Similar to the derivation of Equation 70

$$|J(\tilde{\pi}|P_\epsilon^\pi) - J(\pi|P_\epsilon^\pi)| \tag{107}$$

$$\leq |\sum_t \gamma^t \sum_{s,a} (\tilde{\pi}(a|s) - \pi(a|s)) p_{\tilde{\pi}}^t(s|P_\epsilon^\pi) R(s,a)| \tag{108}$$

$$\leq \frac{2R_{\max} d_{\text{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)} \tag{109}$$

Hence we have

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\pi|P_\epsilon^\pi) \geq -\frac{2R_{\max} d_{\text{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)} \tag{110}$$

By combining Equations 106, 110, we rewrite Equation 69 as

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) \tag{111}$$

$$\geq -\frac{4\gamma(\epsilon+\delta)L_r\alpha}{(1-\gamma)^3} - \frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{\max}) d_{\text{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)^3} \tag{112}$$

By combining Equations of PPO,

$$J(\tilde{\pi}|P_\epsilon^\pi) \tag{113}$$

$$\geq J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) - \frac{4\gamma(\epsilon+\delta)L_r\alpha}{(1-\gamma)^3} - \frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{\max}) d_{\text{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)^3} \tag{114}$$

$$\geq J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) - \frac{8\gamma(1-\alpha)\cdot(\epsilon+\delta)L_r}{(1-\gamma)^3} - \frac{4\gamma(\epsilon+\delta)L_r\alpha}{(1-\gamma)^3} - \frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{\max}) d_{\text{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)^3} \tag{115}$$

$$\geq J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) - \frac{8\gamma\cdot(\epsilon+\delta)L_r}{(1-\gamma)^3} + \frac{4\gamma(\epsilon+\delta)L_r\alpha}{(1-\gamma)^3} - \frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{\max}) d_{\text{TV}}(\pi\|\tilde{\pi})}{(1-\gamma)^3} \tag{116}$$

□

G. Additional Experimental Results

Table 4. We extended the SAPPO by adopting the relaxed state adversarial strategy and evaluated whether the extension (i.e., RA-SAPPO) can improve the agents' robustness against state perturbation. Mean and standard deviations are reported.

Environment		Nominal	σ = 0.005	σ = 0.01	σ = 0.015	σ = 0.02	σ = 0.025
HalfCheetah-v2	SAPPO	4928 ± 370	4765 ± 359	4485 ±394	4036 ± 582	3282 ± 1175	2533 ± 1495
	RA-SAPPO	5784 ± 1081	**5371 ± 1323**	**4874 ± 1311**	**4775 ± 933**	**4106 ± 1273**	**3152 ± 1750**
		Nominal	σ = 0.001	σ = 0.0015	σ = 0.002	σ = 0.0025	σ = 0.003
Walker2d-v2	SAPPO	4135 ± 962	2211 ± 1322	940 ± 405	673 ± 318	667 ± 326	614 ± 311
	RA-SAPPO	4539 ± 1014	**3229 ± 1590**	**1564 ± 1410**	**921 ± 789**	**832 ± 806**	**746 ± 772**
		Nominal	σ = 0.003	σ = 0.004	σ = 0.005	σ = 0.006	σ = 0.007
Humanoid-v2	SAPPO	5736 ± 1194	3690 ± 2068	2926 ± 1956	1944 ± 1438	1409 ± 1098	1156 ± 789
	RA-SAPPO	5320 ± 1164	**3960 ± 2082**	**3335 ± 2117**	**2882 ± 2066**	**2129 ± 1776**	**1567 ± 1474**

H. An Illustrative Example for Comparing the Lower Bounds in Theorem 1 and Theorem 2

Recall from Section 4 that Theorem 2 offers a tighter lower bound than Theorem 1. To further substantiate this, let us take the Reacher task in MuJoCo as an example. In Reacher, the goal is to control a robot arm with two joints and move the

robot's fingertip close to the target. Let $s = (s_1, s_2)$ be the position of the fingertip, $s_g = (s_{g_1}, s_{g_2})$ be the position of the target, and $s_1, s_2, s_{g_1}, s_{g_2} \in [0, S]$. Let $a = (a_1, a_2)$ be the action of the joints. In Reacher, the reward function is defined as

$$R(s,a) = -\sqrt{(s_1 - s_{g_1})^2 + (s_2 - s_{g_2})^2} - \kappa(a_1^2 + a_2^2), \tag{117}$$

where κ is some weight factor of the action penalty. Then, one can verify that $R_{\max} = \sqrt{2}S = \mathcal{O}(S)$. Moreover, we have $L_r = 1$ since

$$\left\| \frac{\partial R(s,a)}{\partial s} \right\|_2 = \sqrt{\left(\frac{-(s_1 - s_{g_1})}{\sqrt{(s_1 - s_{g_1})^2 + (s_2 - s_{g_2})^2}}\right)^2 + \left(\frac{-(s_2 - s_{g_2})}{\sqrt{(s_1 - s_{g_1})^2 + (s_2 - s_{g_2})^2}}\right)^2} = 1. \tag{118}$$

Hence, Theorem 2 can reduce the growth rate from $R_{\max} = \mathcal{O}(S)$ to $\frac{2(\epsilon+\delta)L_r\alpha}{(1-\gamma)} = \mathcal{O}(1)$ as ϵ, δ, α, and γ are constants with respect to S.

I. Implementation Details

We apply the online cross-validation (Sutton, 1992) method to update the average-case and worst-case rewards alternatively and iteratively. Specifically, at one step, we update the policy π_{θ_t} using the paths generated by the current relaxation parameter α_t. The Bellman operator used to update the value function is derived in Appendix B. At the other step, we apply the updated model $\pi_{\theta_{t+1}}$ to generate new paths and compute the relaxation parameter α_{t+1} by maximizing the meta objective function. The gradient of relaxation parameter α_t is calculated by

$$\frac{\partial J_{\text{meta}}(\alpha_t; \theta_{t+1})}{\partial \alpha_t} = \frac{\partial J_{\text{meta}}(\alpha_t; \theta_{t+1})}{\partial \theta_{t+1}} \frac{\partial \theta_{t+1}}{\partial \alpha_t}, \tag{119}$$

where $\frac{\partial \theta_{t+1}}{\partial \alpha_t}$ measures how the relaxation parameter affects the updated model parameter. Since $\theta_{t+1} = \theta_t + f(\theta_t, \alpha_t)$, where $f(\theta_t, \alpha_t)$ is the update function for θ_t, we have $\frac{\partial \theta_{t+1}}{\partial \alpha_t} = \frac{\partial f(\theta_t, \alpha_t)}{\partial \alpha_t}$. In our implementation, we use the automatic differentiation package in PyTorch to compute $\frac{\partial J_{\text{meta}}(\alpha_t; \theta_{t+1})}{\partial \theta_{t+1}}$ and $\frac{\partial \theta_{t+1}}{\partial \alpha_t}$. In addition, to avoid the large penalty coefficients $-\frac{4\gamma(\epsilon+\delta)L_r}{(1-\gamma)^3}$ and $-\frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{\max})}{(1-\gamma)^3}$ (Theorem 2), which lead to prohibitively small steps (Jiang et al., 2021), we consider the coefficients to be tunable hyper-parameters C_1 and C_2. We apply the grid search (i.e., $[0.001, 0.01, 0.02]$ for C_1 and $[0.1, 0.5, 1.0, 1.5]$ for C_2) to find the best hyper-parameters.

We use tunable hyperparameters C_1 and C_2 to approximate the coefficients in Equation 10 because this strategy can improve network training. The strategy is commonly used in optimization. Famous examples are TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017). Specifically, TRPO's authors pointed out that the derived penalty coefficient leads to a tiny step at each policy update; and PPO's authors solved the problem by setting the penalty coefficient as (1) a fixed hyperparameter and (2) an adaptive hyperparameter, and (3) by clipping the penalty directly. In our implementation, since α is dynamic, and its value is correlated with π, we set the penalty coefficients of α_t and $d_{\text{TV}}(\pi_{\theta_t}, \pi_{\theta_{t+1}})$ to fixed parameters to achieve a stable network training.

Table 5. We compared the performances of RAPPO and SCPPO from the statistical perspective. The values indicate the lower bound of 95% confidence interval of the test of significance. The null hypothesis was no difference between RAPPO and SCPPO, and the alternative hypothesis was the opposite. Namely, the value larger than 0 indicated that the difference was statistically significant.

	σ = 0.005	σ = 0.01	σ = 0.015	σ = 0.02	σ = 0.025
HalfCheetah-v2	294	717	1046	815	757.1
	σ = 0.0008	σ = 0.0016	σ = 0.002	σ = 0.0024	σ = 0.003
Hopper-v2	718.6	514.3	399.6	216.2	190.1
	σ = 0.001	σ = 0.0015	σ = 0.002	σ = 0.0025	σ = 0.003
Walker2d-v2	1125	1503	1010.8	580.5	474.7
	σ = 0.01	σ = 0.02	σ = 0.03	σ = 0.04	σ = 0.05
Ant-v2	-28	416.3	136	81.4	146.5
	σ = 0.003	σ = 0.004	σ = 0.005	σ = 0.006	σ = 0.007
Humanoid-v2	295	591	534	317	152.3

Algorithm 3 Practical Implementation of Relaxed State-Adversarial Policy Optimization

Input : MDP $(\mathcal{S}, \mathcal{A}, P_0, R, \gamma)$, number of iterations T, number of update samples T_{upd}, nominal transition kernel P_0, hyperparameter for RAPPO C_1 and C_2, uncertainty set radius ϵ

Initialize the policy π_{θ_0}, the value function V_{ϕ_0}

for $t = 0, \ldots, T-1$ **do**

  Sample the tuple $\{s_i, a_i, r_i, s_i'\}_{i=1}^{T_{\text{upd}}}$, where $a_i \sim \pi_{\theta_t}(\cdot|s_i)$, and $s_i' \sim P_0(\cdot|s_i, a_i)$

  Evaluate $J(\pi_{\theta_t}|P_\epsilon^{\pi_{\theta_{t-1}},\alpha_t}) = \sum_{j=0}^{T_{\text{upd}}}[r_j + \gamma[\alpha_t V_{\phi_t}(s_j') - (1-\alpha_t)(\min_{s_j'' \in \mathcal{N}_\epsilon(s_j')} V_{\phi_t}(s_j''))]]$

  Update the policy to $\pi_{\theta_{t+1}}$ and value function to $V_{\phi_{t+1}}$ by PPO

  Sample the tuple $\{s_i, a_i, r_i, s_i'\}_{i=1}^{T_{\text{upd}}}$, where $a_i \sim \pi_{\theta_{t+1}}(\cdot|s_i)$, and $s_i' \sim P_0(\cdot|s_i, a_i)$

  Evaluate $J(\pi_{\theta_{t+1}}|P^{\pi_{\theta_t},\alpha_t}) = \sum_{j=0}^{T_{\text{upd}}}[r_j + \gamma[\alpha_t V_{\phi_{t+1}}(s_j') - (1-\alpha_t)(\min_{s_j'' \in \mathcal{N}_\epsilon(s_j')} V_{\phi_{t+1}}(s_j''))]]$

  Evaluate $J_{\text{meta}}(\alpha_t) = J(\pi_{\theta_{t+1}}|P_\epsilon^{\pi_{\theta_t},\alpha_t}) - C_1\alpha_t - C_2 d_{\text{TV}}(\pi_{\theta_t}\|\pi_{\theta_{t+1}})$

  Update the relaxation parameter α_t via $\frac{\partial J_{\text{meta}}(\alpha_t;\theta_{t+1})}{\partial \alpha_t} = \frac{\partial J_{\text{meta}}(\alpha_t;\theta_{t+1})}{\partial \theta_{t+1}} \frac{\partial \theta_{t+1}}{\partial \alpha_t}$

end

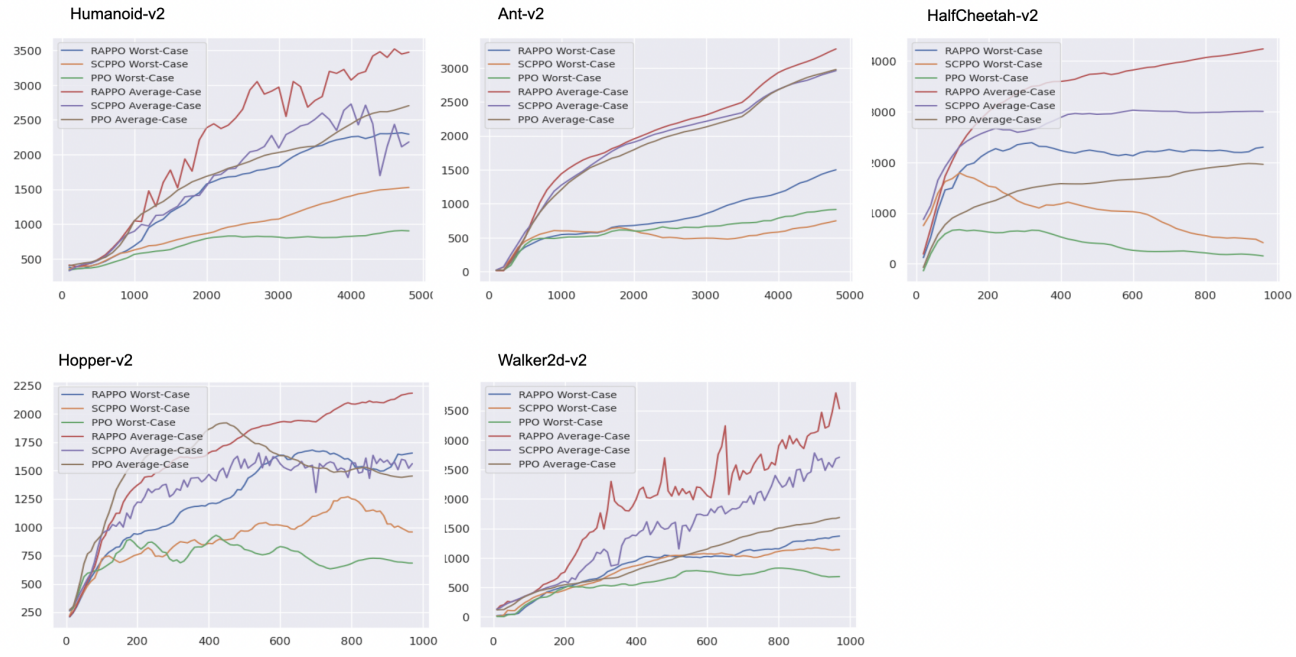

Figure 6. The training curves to facilitate a comparison between the learning performances of PPO, SCPPO, and RAPPO.