A APPENDIX

A.1 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.

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

$$\begin{aligned}\text{At } t=1, \quad \text{Q(s,a)} &= \text{Q(grid(1,2), UP)} = r\text{(grid(1,2), UP)} + \gamma \times \min_{s' \in N(grid(1,3))} V(s') \\ &= -1 + 0.99 \times \min(\text{V(grid(1,3)), V(grid(0,3)), V(grid(2,3)), V(grid(1,2))} \\ &= -1 + 0.99 \times \min(-1, 0, -1, -1) = -1.9.\end{aligned}$$

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

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 $t = 1, \dots, T$ do
    for state s, action a do
        $Q_t(s,a) = r(s,a) + \gamma \sum_{s'} P_0(s'|s,a) \underset{s'' \in \mathcal{N}(s')}{min} (\max_a Q_{t-1}(s'', a))$
    end
end
```

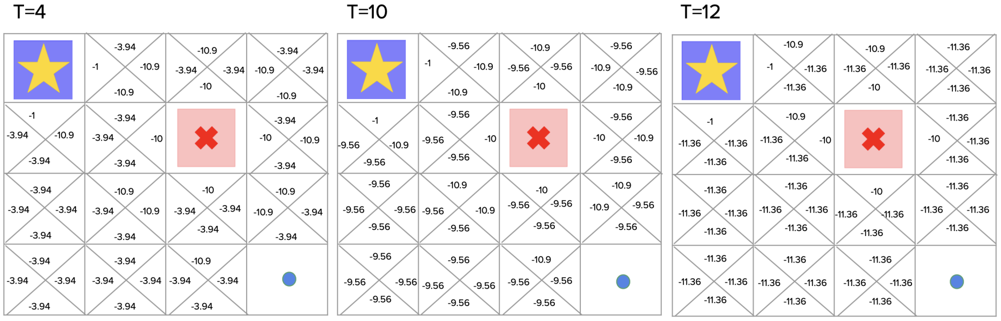

Figure 4: The 4×4 shortest-path grid world. The dot, star, and cross icons indicate the initial, goal, and trap states, respectively. The agent can move either up, down, left, or right at each step and earn $+0$ and -10 rewards when reaching the goal and trap states, respectively. In addition, the agent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

A.2 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

$$V_\epsilon^{\pi,\alpha}(s) := \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P_\epsilon^{\pi,\alpha}(\cdot|s_0,a_0)}\Big[\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}r(s_1,a_1) \tag{10}$$

$$+ \gamma\mathbb{E}_{s_2\sim P_\epsilon^{\pi,\alpha}(\cdot|s1,a1)}\Big[\mathbb{E}_{a_2\sim\pi(\cdot|s_2)}r(s_2,a_2) + ...\Big]\Big]\Big] \tag{11}$$

The corresponding Bellman operator is derived as

$$\mathcal{T}_\epsilon^{\pi,\alpha}V(s) = \mathbb{E}_{a\sim\pi}\Big[r(s,a) + \gamma\mathbb{E}_{s'\sim P_0(\cdot|s,a)}\Big(\alpha V(s') + (1-\alpha)\min_{s''\in\mathcal{N}(s')} V(s'')\Big)\Big] \tag{12}$$

Proof.

$$V_\epsilon^{\pi,\alpha}(s_0) = \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P_\epsilon^{\pi,\alpha}(\cdot|s_0,a_0)}\Big[\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}r(s_1,a_1) \tag{13}$$

$$+ \gamma\mathbb{E}_{s_2\sim P_\epsilon^{\pi,\alpha}(\cdot|s1,a1)}\Big[\mathbb{E}_{a_2\sim\pi(\cdot|s_2)}r(s_2,a_2) + ...\Big]\Big]\Big] \tag{14}$$

$$= \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0)}\Big[\alpha\Big(\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}r(s_1,a_1) \tag{15}$$

$$+ \gamma\mathbb{E}_{s_2\sim P_\epsilon^{\pi,\alpha}(\cdot|s1,a1)}\Big[\mathbb{E}_{a_2\sim\pi(\cdot|s_2)}r(s_2,a_2) + ...\Big) \tag{16}$$

$$+ (1-\alpha)\Big(\min_{s_1'\in\mathcal{N}_\epsilon(s_1)}\mathbb{E}_{a_1'\sim\pi(\cdot|s_1')}r(s_1',a_1') \tag{17}$$

$$+ \gamma\mathbb{E}_{s_2'\sim P_\epsilon^{\pi,\alpha}(\cdot|s_1',a_1')}\Big[\mathbb{E}_{a_2'\sim\pi(\cdot|s_2')}r(s_2',a_2') + ...\Big)\Big]\Big]\Big] \tag{18}$$

$$= \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0)}\Big(\alpha V_\epsilon^{\pi,\alpha}(s_1) + (1-\alpha)\min_{s_1'\in\mathcal{N}_\epsilon(s_1)} V_\epsilon^{\pi,\alpha}(s_1')\Big)\Big], \tag{19}$$

□

A.3 PROOF OF LEMMA 1

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. The reward function J satisfies*

$$J(\pi|P_\sigma^\pi) \geq J(\pi|P_{\sigma+x}^\pi), \quad \forall\pi. \tag{20}$$

Proof.

$$V^\pi(s_0|P^\pi_\sigma) = \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1^\sigma\sim P^\pi_\sigma(\cdot|s_0,a_0)}\Big[V^\pi(s_1^\sigma|P^\pi_\sigma)\Big]\Big] \tag{21}$$

$$= \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0), s_1^\sigma=\text{argmin}V^\pi(s), s\in\mathcal{N}_\sigma(s_1)}\Big[V^\pi(s_1^\sigma|P^\pi_\sigma)\Big]\Big] \tag{22}$$

$$\geq \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0), s_1^{\sigma+x}=\text{argmin}V^\pi(s), s\in\mathcal{N}_{\sigma+x}(s_1)}\Big[V^\pi(s_1^{\sigma+x}|P^\pi_\sigma)\Big]\Big] \tag{23}$$

$$= \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0), s_1^{\sigma+x}=\text{argmin}V^\pi(s), s\in\mathcal{N}_{\sigma+x}(s_1), a_1\sim\pi(\cdot|s_1^{\sigma+x})}\Big[r(s_1^{\sigma+x},a_1) \tag{24}$$

$$+ \gamma\mathbb{E}_{s_2^\sigma\sim P^\pi_\sigma(\cdot|s_1^{\sigma+x},a_1)}[V^\pi(s_2^\sigma|P^\pi_\sigma)]\Big]\Big] \tag{25}$$

$$= \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0), s_1^{\sigma+x}=\text{argmin}V^\pi(s), s\in\mathcal{N}_{\sigma+x}(s_1), a_1\sim\pi(\cdot|s_1^{\sigma+x})}\Big[r(s_1^{\sigma+x},a_1) \tag{26}$$

$$+ \gamma\mathbb{E}_{s_2\sim P_0(\cdot|s_1^{\sigma+x},a_1), s_2^\sigma=\text{argmin}V^\pi(s), s\in\mathcal{N}_\sigma(s_2)}[V^\pi(s_2^\sigma|P^\pi_\sigma)]\Big]\Big] \tag{27}$$

$$\geq \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P_0(\cdot|s_0,a_0), s_1^{\sigma+x}=\text{argmin}V^\pi(s), s\in\mathcal{N}_{\sigma+x}(s_1), a_1\sim\pi(\cdot|s_1^{\sigma+x})}\Big[r(s_1^{\sigma+x},a_1) \tag{28}$$

$$+ \gamma\mathbb{E}_{s_2\sim P_0(\cdot|s_1^{\sigma+x},a_1), s_2^{\sigma+x}=\text{argmin}V^\pi(s), s\in\mathcal{N}_{\sigma+x}(s_2)}[V^\pi(s_2^{\sigma+x}|P^\pi_\sigma)]\Big]\Big] \tag{29}$$

$$\geq \mathbb{E}_{a_i\sim\pi, s_i\sim P^\pi_{\sigma+x})}\Big[r(s_0,a_0) + \gamma r(s_1^{\sigma+x},a_1) + \gamma^2 r(s_2^{\sigma+x},a_2) + ...\Big] \tag{30}$$

$$= \mathbb{E}_{a_0\sim\pi}\Big[r(s_0,a_0) + \gamma\mathbb{E}_{s_1^{\sigma+x}\sim P^\pi_{\sigma+x}(\cdot|s_0,a_0)}\Big[V^\pi(s_1^{\sigma+x}|P^\pi_{\sigma+x})\Big]\Big] \tag{31}$$

$$= V^\pi(s_0|P^\pi_{\sigma+x}) \tag{32}$$

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

$$J(\pi|P^\pi_\sigma) = \mathbb{E}_{s_0\sim\mu}[V^\pi(s_0|P^\pi_\sigma)] \tag{33}$$

$$\geq \mathbb{E}_{s_0\sim\mu}[V^\pi(s_0|P^\pi_{\sigma+x})] \tag{34}$$

$$= J(\pi|P^\pi_{\sigma+x}). \tag{35}$$

□

A.4 PROOF OF LEMMA 2

We prove Lemma 2 based on the continuity assumption 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 implies the similar returns. Hence, for any $\epsilon_c > 0$, there exist a $\delta_c > 0$, such that $|\alpha_1 - \alpha_2| < \delta_c$ and $|J(\pi|P^{\pi,\alpha_1}_\epsilon) - J(\pi|P^{\pi,\alpha_2}_\epsilon)| < \epsilon_c$.

Lemma (Relaxation parameter α as a prior distribution D in domain randomization)**.** *For any distribution $\mathcal{D}$ over the state-adversarial uncertainty set $\mathcal{P}^\pi_\epsilon$, there must be 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

$$J(\pi|P^{\pi,0}_\epsilon) = J(\pi|P^\pi_\epsilon) \leq \mathbb{E}_{p\sim\mathbb{P}^\pi_\epsilon}[J(\pi|p)] \leq J(\pi|P^{\pi,1}_\epsilon) = J(\pi|P_0) \tag{36}$$

Under the condition that $J(\pi|P^{\pi,\alpha}_\epsilon)$ is a continuous function, by Intermediate Value Theorem, we obtain

$$\exists\alpha \in [0, 1], \ni \mathbb{E}_{p\sim\mathbb{P}^\pi_\epsilon}[J(\pi|p)] = J(\pi|P^{\pi,\alpha}_\epsilon) \tag{37}$$

□

A.5 PROOF OF THEOREM 1

Theorem (A naive connection between the average-case and the worst-case returns). *Given a nominal MDP with state adversaries, when updating the current policy π to a new policy $\tilde{\pi}$, the following bound holds Jiang et al. (2021):*

$$J(\tilde{\pi}|P_\epsilon^\pi) \geq \mathbb{E}_{P\sim\mathcal{D}}[J(\tilde{\pi}|P)] - 2R_{max}\frac{\gamma\mathbb{E}_{P\sim\mathcal{D}}[d_{TV}(P_\epsilon^\pi\|P)]}{(1-\gamma)^2} - 4R_{max}\frac{d_{TV}(\pi,\tilde{\pi})}{(1-\gamma)^2}, \tag{38}$$

where R_{max} is the maximum return, $d_{TV}(\pi,\tilde{\pi})$ indicates the total variation divergence between π and $\tilde{\pi}$, P_ϵ^π is the worst state-adversarial transition kernels.

Proof.

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) = J(\tilde{\pi}|P_\epsilon^\pi) - J(\pi|P_\epsilon^\pi) + J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) \tag{39}$$

For the last two term of Equation 39,

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

$$= |\sum_t \gamma^t \sum_{s,a}(p_\pi^t(s,a|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s,a|P_\epsilon^{\pi,\alpha}))R(s,a)| \tag{41}$$

$$\leq \sum_t \gamma^t \sum_{s,a}|(p_\pi^t(s,a|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s,a|P_\epsilon^{\pi,\alpha})|R(s,a) \tag{42}$$

$$\leq 2R_{\text{max}}\sum_t \gamma^t[d_{\text{TV}}(p_\pi^t(s,a|P_\epsilon^\pi)\|p_{\tilde{\pi}}^t(s,a|P_\epsilon^{\pi,\alpha}))] \tag{43}$$

$$\text{because } p_\pi^t(s,a|P_\epsilon^\pi) = \pi(a|s)p_\pi^t(s|P_\epsilon^\pi) \text{ and } p_{\tilde{\pi}}^t(s,a|P_\epsilon^{\pi,\alpha}) = \tilde{\pi}(a|s)p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha}) \tag{44}$$

$$\leq 2R_{\text{max}}[\mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}d_{\text{TV}}(\pi(a|s')\|\tilde{\pi}(a|s')) \tag{45}$$

$$+ d_{\text{TV}}(p_\pi^t(s|P_\epsilon^\pi)\|p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha}))] \tag{46}$$

For the second term of Equation 46,

$$d_{\text{TV}}(p_\pi^t(s|P_\epsilon^\pi)\|p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})) \tag{47}$$

$$\leq t\max_t\mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}d_{\text{TV}}(p_\pi(s|s',a,P_\epsilon^\pi)\|p_{\tilde{\pi}}(s|s',a,P_\epsilon^{\pi,\alpha})) \tag{48}$$

$$\text{because } p_\pi(s|s',a,P_\epsilon^\pi) = \sum_a T(s|s',a,P_\epsilon^\pi)\pi(a|s') \tag{49}$$

$$\leq t\max_t\mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}\mathbb{E}_{a\sim\pi(\cdot|s')}d_{\text{TV}}(T(s|s',a,P_\epsilon^\pi)\|T(s|s',a,P_\epsilon^{\pi,\alpha}) \tag{50}$$

$$+ t\max_t\mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}d_{\text{TV}}(\pi(s|s')\|\tilde{\pi}(a|s')) \tag{51}$$

Then we can rewrite Equation 46as:

$$J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) \tag{52}$$

$$\geq -2R_{\text{max}}\sum_t \gamma^t[(t+1)\max_t\mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}d_{\text{TV}}(\pi(a|s')\|\tilde{\pi}(a|s') \tag{53}$$

$$- t\max_t\mathbb{E}_{s'\sim p_\pi^t(\cdot|P_\epsilon^\pi)}\mathbb{E}_{a\sim\pi(\cdot|s')}d_{\text{TV}}(T(s|s',a,P_\epsilon^\pi)\|T(s|s',a,P_\epsilon^{\pi,\alpha\text{‘}})) \tag{54}$$

Similar to the derivation of Equation 46,

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\pi|P_\epsilon^\pi) \tag{55}$$

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

and rewrite Equation 39 as following,

$$\begin{aligned}
& J(\tilde{\pi}|P^{\pi}_{\epsilon}) - J(\tilde{\pi}|P^{\pi,\alpha}_{\epsilon}) \qquad (57)\\
& \geq -2R_{\text{max}} \sum_{t} \gamma^{t} [2(t+1) \max_{t} \mathbb{E}_{s' \sim p^{t}_{\pi}(\cdot|P^{\pi}_{\epsilon})} d_{\text{TV}}(\pi(a|s') \| \tilde{\pi}(a|s') \qquad (58)\\
& - t \max_{t} \mathbb{E}_{s' \sim p^{t}_{\pi}(\cdot|P^{\pi}_{\epsilon})} \mathbb{E}_{a \sim \pi(\cdot|s')} d_{\text{TV}}(T(s|s', a, P^{\pi}_{\epsilon})) \| T(s|s', a, P^{\pi,\alpha}_{\epsilon})] \qquad (59)\\
& = -2R_{\text{max}} \sum_{t} \gamma^{t} [2(t+1) \max_{t} \mathbb{E}_{s' \sim p^{t}_{\pi}(\cdot|P^{\pi}_{\epsilon})} d_{\text{TV}}(\pi(a|s') \| \tilde{\pi}(a|s') \qquad (60)\\
& - t \mathbb{E}_{P \sim \mathcal{D}}[d_{\text{TV}}(P^{\pi}_{\epsilon} \| P)]] \qquad (61)\\
& = -2R_{\text{max}} \frac{\gamma \mathbb{E}_{P \sim \mathcal{D}}[d_{\text{TV}}(P^{\pi}_{\epsilon} \| P)]}{(1-\gamma)^2} - 4R_{\text{max}} \frac{d_{\text{TV}}(\pi, \tilde{\pi})}{(1-\gamma)^2} \qquad (62)
\end{aligned}$$

□

A.6 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 $R(s) \coloneqq \sum_a \pi(a|s) R(s, a)$.

Theorem (Connecting Worst-Case and Average-Case Returns)**.** *Given a nominal MDP with two properties: (1) Reward function of corresponding MRP with respect to any policy is an L_r-Lipschitz function. (2) Nominal transition kernel P_0 has the smooth transition property δ, where $\|s - s'\|_2 \leq \delta, \forall a$ and $\forall P_0(s'|s, a) > 0$. Then after updating the current policy π to a new policy $\tilde{\pi}$, the following bound holds:*

$$J(\tilde{\pi}|P^{\pi}_{\epsilon}) \geq J(\tilde{\pi}|P^{\pi,\alpha}_{\epsilon}) - \frac{4\gamma(\epsilon+\delta)L_r\alpha}{(1-\gamma)^3} - \frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{max}) d_{TV}(\pi, \tilde{\pi})}{(1-\gamma)^3}, \tag{63}$$

where $d_{TV}(\pi, \tilde{\pi})$ is total variation divergence between π and $\tilde{\pi}$,$P^{\pi,\alpha}_{\epsilon}$ is a relaxed state-adversarial transition kernel, and P^{π}_{ϵ} is a worst-case state-adversarial transition kernel.

We first introduce the supporting lemma before proving the Theorem.

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^{π}_{ϵ}, 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

$$\begin{aligned}
\|s_{\tau+1} - \tilde{s}_{\tau+1}\| &= \|s_{\tau+1} - s_\tau + s_\tau - \tilde{s}_\tau + \tilde{s}_\tau - \tilde{s}_{\tau+1}\| && (64)\\
&\leq \|s_{\tau+1} - s_\tau\| + \|s_\tau - \tilde{s}_\tau\| + \|\tilde{s}_\tau - \tilde{s}_{\tau+1}\| && (65)\\
&\leq \delta + 2\tau(\epsilon+\delta) + (\epsilon+\delta) && (66)\\
&\leq 2(\tau+1)(\epsilon+\delta), && (67)
\end{aligned}$$

where Equation 64 holds by the triangle inequality, Equation 65 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

$$J(\tilde{\pi}|P^{\pi}_{\epsilon}) - J(\tilde{\pi}|P^{\pi,\alpha}_{\epsilon}) = J(\tilde{\pi}|P^{\pi}_{\epsilon}) - J(\pi|P^{\pi}_{\epsilon}) + J(\pi|P^{\pi}_{\epsilon}) - J(\tilde{\pi}|P^{\pi,\alpha}_{\epsilon}) \tag{68}$$

For the last two term of Equation 68,

$$
\begin{aligned}
&|J(\pi|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha})| \tag{69}\\
=&|\sum_t \gamma^t \sum_{s,a} \pi(a|s) p_\pi^t(s|P_\epsilon^\pi) R(s,a) - \tilde{\pi}(a|s) p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha}) R(s,a)| \tag{70}\\
=&|\sum_t \gamma^t \sum_{s,a} \pi(a|s)[p_\pi^t(s|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})]R(s,a) + (\pi(a|s) - \tilde{\pi}(a|s)) p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha}) R(s,a)| \tag{71}\\
\leq&\sum_t \gamma^t \sum_{s,a} |\pi(a|s)[p_\pi^t(s|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})]R(s,a)| + |(\pi(a|s) - \tilde{\pi}(a|s)) p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha}) R(s,a)| \tag{72}
\end{aligned}
$$

For the first term of Equation 72, we have the t-step initial state probability:

$$
\begin{aligned}
&|p_\pi^t(s|P_\epsilon^\pi) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})| \tag{73}\\
\leq&|p_\pi^t(s|P_\epsilon^\pi) - p_\pi^t(s|P_\epsilon^{\pi,\alpha})| + |p_\pi^t(s|P_\epsilon^{\pi,\alpha}) - p_{\tilde{\pi}}^t(s|P_\epsilon^{\pi,\alpha})| \tag{74}\\
& \tag{75}
\end{aligned}
$$

Now we prove the following inequality.

$$
\begin{aligned}
&\sum_s |p_\pi^t(s|P_\epsilon^\pi) - p_\pi^t(s|P_\epsilon^{\pi,\alpha})| \tag{76}\\
=&\sum_s |\sum_{s'} p_\pi^{t-1}(s'|P_\epsilon^\pi)(\sum_{k,Z_{ks}=1} P_0(k|s')) \tag{77}\\
&-(1-\alpha)\sum_{s'} p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})(\sum_{k,Z_{ks}=1} P_0(k|s')) \tag{78}\\
&-\alpha \sum_{s'} p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha}) P_0(s|s'), \tag{79}\\
\leq&\sum_s \sum_{s'} |p_\pi^{t-1}(s'|P_\epsilon^\pi) - p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})|(\sum_{k,Z_{ks}=1} P_0(k|s')) \tag{80}\\
&+\alpha \sum_s \sum_{s'} p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})|(\sum_{k,Z_{ks}=1} P_0(k|s')) - P_0(s|s')| \tag{81}\\
\leq&\sum_{s'} |p_\pi^{t-1}(s'|P_\epsilon^\pi) - p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})| \tag{82}\\
&+\alpha \cdot \underset{s'}{max} \sum_s |(\sum_{k,Z_{ks}=1} P_0(k|s')) - P_0(s|s')| \tag{83}\\
\leq&\sum_{s'} |p_\pi^{t-1}(s'|P_\epsilon^\pi) - p_\pi^{t-1}(s'|P_\epsilon^{\pi,\alpha})| + 2\alpha \tag{84}\\
=&2\alpha t \tag{85}
\end{aligned}
$$

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 77 to 79 follow 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 82 holds because $\sum_s \sum_{k,Z_{ks}=1} P_0(k|s') = 1$. In addition, Equation 85 is obtained by recursively applying Equations 77 to 84 to the first term of Equation 84.

For the first two terms of Equation 74, we have

$$
\begin{aligned}
&|p_\pi^t(s|P_\epsilon^\pi) - p_\pi^t(s|P_\epsilon^{\pi,\alpha})| \tag{86}\\
\leq&\sum_s |p_\pi^t(s|P_\epsilon^\pi) - p_\pi^t(s|P_\epsilon^{\pi,\alpha})| \tag{87}\\
\leq&2\alpha t, \tag{88}
\end{aligned}
$$

For the last two terms of Equation 74, we have

$$
\begin{aligned}
&|p^t_{\pi}(s|P^{\pi,\alpha}_{\epsilon}) - p^t_{\tilde{\pi}}(s|P^{\pi,\alpha}_{\epsilon})| && (89)\\
\leq &\sum_{s} |p^t_{\pi}(s|P^{\pi,\alpha}_{\epsilon}) - p^t_{\tilde{\pi}}(s|P^{\pi,\alpha}_{\epsilon})| && (90)\\
= &\sum_{s}\sum_{s',a} \Big(|p^{t-1}_{\pi}(s'|P^{\pi,\alpha}_{\epsilon})\pi(a|s') - p^{t-1}_{\tilde{\pi}}(s'|P^{\pi,\alpha}_{\epsilon})\tilde{\pi}(a|s')|\Big)\Big(\alpha P_0(s|s',a) + (1-\alpha)\sum_{k,Z_{ks}=1} P_0(k|s',a)\Big) && (91)\\
\leq &\sum_{s',a} |p^{t-1}_{\pi}(s'|P^{\pi,\alpha}_{\epsilon})\pi(a|s') - p^{t-1}_{\tilde{\pi}}(s'|P^{\pi,\alpha}_{\epsilon})\tilde{\pi}(a|s')| && (92)\\
\leq &\sum_{s',a} |p^{t-1}_{\pi}(s'|P^{\pi,\alpha}_{\epsilon})\pi(a|s') - p^{t-1}_{\pi}(s'|P^{\pi,\alpha}_{\epsilon})\tilde{\pi}(a|s')| + \sum_{s',a} |p^{t-1}_{\pi}(s'|P^{\pi,\alpha}_{\epsilon})\tilde{\pi}(a|s') - p^{t-1}_{\tilde{\pi}}(s'|P^{\pi,\alpha}_{\epsilon})\tilde{\pi}(a|s')| && (93)\\
= &\sum_{s',a} |p^{t-1}_{\pi}(s'|P^{\pi,\alpha}_{\epsilon})(\pi(a|s') - \tilde{\pi}(a|s'))| + \sum_{s',a} |(p^{t-1}_{\pi}(s'|P^{\pi,\alpha}_{\epsilon}) - p^{t-1}_{\tilde{\pi}}(s'|P^{\pi,\alpha}_{\epsilon}))\tilde{\pi}(a|s')| && (94)\\
\leq &2t d_{\text{TV}}(\pi,\tilde{\pi}) && (95)
\end{aligned}
$$

Hence, we can rewrite Equation 73 as:

$$
|p^t_{\pi}(s|P^{\pi}_{\epsilon}) - p^t_{\tilde{\pi}}(s|P^{\pi,\alpha}_{\epsilon})| \leq 2\alpha t + 2t d_{\text{TV}}(\pi,\tilde{\pi}) \qquad (96)
$$

where Equation 96 holds by applying Equation 95 and 88.

Under the condition that the reward function of the MRP under policy π is L_r-Lipschitz, $|R(s_1) - R(s_2)| \leq 2t(\epsilon+\delta)L_r$ if $|s_1 - s_2| \leq 2t(\epsilon+\delta)$. By Lemma 3, for every probability density in $p^t_{\pi}(s|P^{\pi}_{\epsilon})$, there exists a corresponding density point transited by $P^{\pi,\alpha}_{\epsilon}$, 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 96, for every state, the total probability density difference is bounded by $2\alpha t + 2t d_{\text{TV}}(\pi,\tilde{\pi})$. The total reward difference at time t will be

$$
\begin{aligned}
&|\sum_{s,a} \pi(a|s)[p^t_{\pi}(s|P^{\pi}_{\epsilon}) - p^t_{\tilde{\pi}}(s|P^{\pi,\alpha}_{\epsilon})]R(s,a)| && (97)\\
=&|\sum_{s} [p^t_{\pi}(s|P^{\pi}_{\epsilon}) - p^t_{\tilde{\pi}}(s|P^{\pi,\alpha}_{\epsilon})]R(s)| && (98)\\
\leq&(2\alpha t + 2t d_{\text{TV}}(\pi,\tilde{\pi})) \cdot 2t(\epsilon+\delta)L_r && (99)
\end{aligned}
$$

Combining Equations 69and 99, we have

$$
\begin{aligned}
&|J(\pi|P^{\pi}_{\epsilon}) - J(\tilde{\pi}|P^{\pi,\alpha}_{\epsilon})| && (100)\\
\leq&\sum_{t} \gamma^t \Big((2\alpha t + 2t d_{\text{TV}}(\pi,\tilde{\pi})) \cdot 2t(\epsilon+\delta)L_r + 2R_{\text{max}} d_{\text{TV}}(\pi,\tilde{\pi})\Big) && (101)\\
=&\sum_{t} \gamma^t 4\alpha t^2(\epsilon+\delta)L_r + \sum_{t} \gamma^t 4t^2(\epsilon+\delta)L_r d_{\text{TV}}(\pi,\tilde{\pi}) + \sum_{t} \gamma^t 2R_{\text{max}} d_{\text{TV}}(\pi,\tilde{\pi}) && (102)\\
=&\frac{\gamma(4\alpha(\epsilon+\delta)L_r)}{(1-\gamma)^3} + \frac{4\gamma(\epsilon+\delta)L_r d_{\text{TV}}(\pi,\tilde{\pi})}{(1-\gamma)^3} + 2\frac{R_{\text{max}} d_{\text{TV}}(\pi,\tilde{\pi})}{(1-\gamma)} && (103)
\end{aligned}
$$

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

$$
\begin{aligned}
&J(\pi|P^{\pi}_{\epsilon}) - J(\tilde{\pi}|P^{\pi,\alpha}_{\epsilon}) && (104)\\
\geq& -\frac{\gamma(4\alpha(\epsilon+\delta)L_r)}{(1-\gamma)^3} - \frac{4\gamma(\epsilon+\delta)L_r d_{\text{TV}}(\pi,\tilde{\pi})}{(1-\gamma)^3} - \frac{2R_{\text{max}} d_{\text{TV}}(\pi,\tilde{\pi})}{(1-\gamma)} && (105)
\end{aligned}
$$

Similar to the derivation of Equation 69

$$
\begin{aligned}
&|J(\tilde{\pi}|P^{\pi}_{\epsilon}) - J(\pi|P^{\pi}_{\epsilon})| && (106)\\
\leq&|\sum_{t} \gamma^t \sum_{s,a} (\tilde{\pi}(a|s) - \pi(a|s)) p^t_{\tilde{\pi}}(s|P^{\pi}_{\epsilon}) R(s,a)| && (107)\\
\leq&\frac{2R_{\text{max}} d_{\text{TV}}(\pi,\tilde{\pi})}{(1-\gamma)} && (108)
\end{aligned}
$$

Hence we have

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\pi|P_\epsilon^\pi) \geq -\frac{2R_{\max} d_{\text{TV}}(\pi, \tilde{\pi})}{(1-\gamma)} \tag{109}$$

By combining Equations 105, 109, we rewrite Equation 68 as

$$J(\tilde{\pi}|P_\epsilon^\pi) - J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) \tag{110}$$

$$\geq -\frac{4\gamma(\epsilon+\delta)L_r\alpha}{(1-\gamma)^3} - \frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{\max})d_{\text{TV}}(\pi, \tilde{\pi})}{(1-\gamma)^3} \tag{111}$$

By combining Equations of PPO,

$$J(\tilde{\pi}|P_\epsilon^\pi) \tag{112}$$

$$\geq J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) - \frac{4\gamma(\epsilon+\delta)L_r\alpha}{(1-\gamma)^3} - \frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{\max})d_{\text{TV}}(\pi, \tilde{\pi})}{(1-\gamma)^3} \tag{113}$$

$$\geq J(\tilde{\pi}|P_\epsilon^{\pi,\alpha}) - \frac{8\gamma(1-\alpha)\cdot(\epsilon+\delta)L_r}{(1-\gamma)^3} - \frac{4\gamma(\epsilon+\delta)L_r\alpha}{(1-\gamma)^3} - \frac{4(\gamma(\epsilon+\delta)L_r + (1-\gamma)^2 R_{\max})d_{\text{TV}}(\pi, \tilde{\pi})}{(1-\gamma)^3} \tag{114}$$

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

$\square$

A.7 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 A.2. 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

$$\frac{\partial J_{\text{meta}}(\alpha_t;\theta_{t+1})}{\partial\alpha_t} = \frac{\partial J_{\text{meta}}(\alpha_t;\theta_{t+1})}{\partial\theta_{t+1}}\frac{\partial\theta_{t+1}}{\partial\alpha_t}, \tag{116}$$

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 $-\frac{4(\gamma(\epsilon+\delta)L_r+(1-\gamma)^2R_{\max})}{(1-\gamma)^3}$ (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 9 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.

A.8 DYNAMICS OF RELAXATION PARAMETER α DURING TRAINING

We leverage Theorem 2 to address both the average-case and the worst-case performance. Specifically, we maximize the lower bound of the worst-case performance (i.e., RHS of Equation 9) when training

Algorithm 3: Practical Implementation of Relaxed State-Adversarial Policy Optimization

```
Input: MDP $(\mathcal{S}, \mathcal{A}, P_0, r, \gamma)$, number of iterations $T$, nominal transition kernel $P_0$,
       hyperparameter for RAPPO $C_1$ and $C_2$, $\epsilon$-Neighborhood
Initialize the policy $\pi_{\theta_0}$, the value function $V_{\phi_0}$
for $t = 0, \dots, T-1$ 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 $J(\pi_{\theta_t}|P_\epsilon^{\pi_{\theta_{t-1}}, \alpha_t}) = \sum_{j=0}^{T_{\text{upd}}}[r_j + \gamma[\alpha_t V_{\phi_t}(s'_j) - (1-\alpha_t)(\underset{s''_j \in \mathcal{N}_\epsilon(s'_j)}{min} V_{\phi_t}(s''_j))]]$
    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 $J(\pi_{\theta_{t+1}}|P^{\pi_{\theta_t}, \alpha_t}) = \sum_{j=0}^{T'_{\text{upd}}}[r_j + \gamma[\alpha_t V_{\phi_{t+1}}(s'_j) - (1-\alpha_t)(\underset{s''_j \in \mathcal{N}_\epsilon(s'_j)}{min} V_{\phi_{t+1}}(s''_j))]]$
    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 $\alpha_t$ 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
```

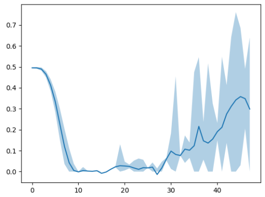

Figure 5: We show the dynamics of α during training the Halfcheetah environment, in which α was initially set to 0.5 and then updated by the meta objective function. The solid line and shading area are the mean and standard deviation of α determined from 5 seeds. As can be seen, α first decreased to maximize rewards in the worst-case environment and then increased to maximize average-case rewards.

policies. Since the unknowns α and π in Theorem 2 are correlated, we solve the upper-level and the lower-level tasks iteratively and alternatively to update their values. Figure 5 shows the dynamics of α during training. As indicated, the meta-objective moved α to 0 to maximize the worst-case performance initially. It then increased the value of α to improve the agent's performance close to nominal environments. The strategy was reasonable because maximizing rewards in the worst case could also increase rewards in the average case (by Lemma 1). Once the worst-case rewards are larger than a specific bound (i.e., Theorem 2 holds), the meta-objective increased the value of α for further increasing the rewards close to nominal environments. Note that the variance of α is reasonable. Since π was randomly initialized, the correlations between the variants of π and α were not necessarily the same.

We point out that π and α in Theorem 2 are correlated, and they should be considered simultaneously to maximize the lower bound of the worst-case performance. In other words, simply designing a schedule for α is insufficient to optimize a policy. To verify this claim, we conducted experiments, in which α linearly increases (Scheduler+) and decrease (Scheduler+) during training. Table 3 shows the experiment results. As indicated, RAPPO outperformed Scheduler+ and Scheduler- in many environments, except when the attacks to ants were weak. The results were not surprising because the schedules did not consider the correlation between α and π. As long as the loss landscape of the objective function was not convex, the scheduler was difficult to find a good solution.

A.9 HIGH VARIANCE OF THE TOTAL EXPECTED RETURNS

In the experiments of robustness against states adversaries (Table 1), we perturbed policies every step, and each attack aimed to push the policy to the worst neighboring state in the experiment. The high variance of the total returns was inevitable. This phenomenon did not appear only in our results but also in all the baselines, such as PPO and SCPPO. In addition, a stronger attack also led to a higher variance of the return. Take the performance of PPO in the HalfCheetah environment as an example, the mean-variance ratio was 5.26 (5286/1004) when the environment was nominal, and the ratio decreased to 0.81 (819/1003) when the magnitude σ of attack became 0.025. The performance of the SCPPO in the HalfCheetah environment also had a similar phenomenon. The mean-variance ratio decreased from 8.68 (6157/709) to 0.08 (60/791).

Environment		Nominal	$\sigma = 0.005$	$\sigma = 0.01$	$\sigma = 0.015$	$\sigma = 0.02$	$\sigma = 0.025$
HalfCheetah	Scheduler +	5744.16±883.70	4591.63±1395.03	3542.69±1730.48	2001.48±1910.79	1096.29±1859.76	408.94±1644.85
	Scheduler -	5865.87±770.69	5176.21±829.77	3456.30±1903.02	1970.26±2033.79	941.72±1483.99	-158.06±922.55
	RAPPO	**6146.74±742.52**	**5519.39±774.32**	**4353.17±1510.70**	**3087.78±1568.22**	**1878.98±1287.87**	**846.85±951.06**
		Nominal	$\sigma = 0.0008$	$\sigma = 0.0016$	$\sigma = 0.002$	$\sigma = 0.0024$	$\sigma = 0.003$
Hopper	Scheduler +	3032.28±763.94	1156.97±614.91	829.46±353.28	691.56±250.66	591.01±210.47	479.65±238.97
	Scheduler -	3032.31±822.57	1276.19±592.80	789.55±428.12	758.48±456.64	617.28±332.39	486.93±225.36
	RAPPO	**3301.48±520.76**	**2198.12±859.08**	**1457.89±537.07**	**1244.16±584.26**	**1067.22±605.42**	**1014.58±779.92**
		Nominal	$\sigma = 0.001$	$\sigma = 0.0015$	$\sigma = 0.002$	$\sigma = 0.0025$	$\sigma = 0.003$
Walker2d	Scheduler +	4051.98±1130.50	1548.57±1068.91	994.26±600.17	736.46±314.88	597.01±237.82	618.84±287.75
	Scheduler -	4043.19±1238.57	2410.64±1581.69	1601.00±1206.41	1601.00±1206.41	771.69±768.81	771.69±768.81
	RAPPO	**4608.12±962.96**	**3998.66±1487.38**	**3298.44±1478.25**	**2160.74±1408.89**	**1470.07±1013.95**	**1173.75±783.89**
		Nominal	$\sigma = 0.01$	$\sigma = 0.02$	$\sigma = 0.03$	$\sigma = 0.04$	$\sigma = 0.05$
Ant	Scheduler +	5847.72±1039.84	4482.75±1306.05	**2630.83±1235.01**	**1395.59±704.53**	705.89±381.12	496.82±401.96
	Scheduler -	**6131.02±606.15**	**4719.54±1191.67**	2153.95±1165.94	1051.75±544.08	705.12±233.81	610.91±236.39
	RAPPO	6022.20±698.78	4381.52±1357.64	2284.53±1225.28	1038.94±553.84	**733.02±255.95**	**672.69±219.99**
		Nominal	$\sigma = 0.003$	$\sigma = 0.004$	$\sigma = 0.005$	$\sigma = 0.006$	$\sigma = 0.007$
Humanoid	Scheduler +	5162.77±1714.34	2903.53±2094.77	2365.29±1840.93	1689.72±1580.01	1347.32±1282.06	915.06±694.24
	Scheduler -	5039.48±1798.54	3316.45±2153.78	2655.50±2043.42	1847.50±1527.36	1322.75±958.38	1022.42±676.20
	RAPPO	**5355.23±1491.76**	**3768.12±1972.79**	**3227.09±1883.64**	**2537.66±1698.87**	**1747.90±1274.77**	**1350.13±1133.66**

Table 3: We compared the performance of agents trained using RAPPO and two simple schedulers of α in Mujoco environments under multiple degrees of state perturbation. Scheduler+ and Scheduler- indicate that α linearly transited from 0 to 1 and from 1 to 0, respectively. Mean and Standard deviations are reported.

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

Table 4: 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.

To compare the performance of RAPPO and the baselines objectively, we averaged the results from 250 trajectories (50 initialization over 5 seeds). Although the standard deviations of the total expected returns were high, the results in Table 1 were statistically significant. To verify this claim from the statistical perspective, we computed the test of significance. The null hypothesis was that the difference of the mean between RAPPO and SCPPO was zero. The alternative hypothesis was that the difference was larger than zero. The following table shows the lower bound of 95% confidence interval (CI). The value implies that one can have 95% confidence that the interval included the true value of the difference between two methods. If the 95% confidence interval did not cover 0, it meant that the difference between SCPPO and RAPPO was statistically significant. We will attach the statistical analysis in the supplemental material in the revised manuscript.