1 Proof of Theorem 3.5

Let Π be the set of all policies. The set Π of policies is convex because a policy can be represented as a probability distribution over actions for each state. Since $\mathcal{S}$ and $\mathcal{A}$ are finite, it is a compact subset of a finite-dimensional Euclidean space. We conclude the theorem if we can show that the mapping $\pi \mapsto \mathbb{E}_{s_0 \sim P_0}[V_{\pi \circ \nu^\star}(s_0)]$ is a continuous function from Π to $\mathbb{R}$, since a continuous function over a compact set attains a maximum in the compact set.

Let π_1 and π_2 be two policies, and let $\pi_\theta = \theta\pi_1 + (1-\theta)\pi_2$ be a convex combination with $\theta \in [0,1]$. For any adversary $\nu \colon \mathcal{S} \to \mathcal{S}$ and for any state $s \in \mathcal{S}$, it holds

$$V_{\pi_\theta \circ \nu}(s) = \theta V_{\pi_1 \circ \nu}(s) + (1-\theta) V_{\pi_2 \circ \nu}(s). \tag{23}$$

This follows from the linearity of the Bellman operator and the definition of the policy π_θ as a convex combination. Next, we consider the adversary that minimizes the value function for a given policy. For any fixed state $s \in \mathcal{S}$, the following holds:

$$\begin{aligned} V_{\pi_\theta \circ \nu^\star} &= \min_\nu V_{\pi_\theta \circ \nu}(s) \\ &= \min_\nu \left[\theta V_{\pi_1 \circ \nu}(s) + (1-\theta) V_{\pi_2 \circ \nu}\right](s) \\ &\geq \theta \min_\nu \left[V_{\pi_1 \circ \nu}(s)\right] + (1-\theta) \min_{\nu'} \left[V_{\pi_2 \circ \nu'}\right] \\ &= \theta V_{\pi_1 \circ \nu^\star}(s) + (1-\theta) V_{\pi_2 \circ \nu^\star}(s). \end{aligned} \tag{24}$$

This inequality uses the fact that the minimum value of a convex combination of functions is greater than or equal to the convex combination of their respective minimum values.

We aggregate the above state-wise results by taking the expectation of both sides over $s_0 \sim P_0$. Since expectation is a linear operator, we arrive at the following.

$$\begin{aligned} &\mathop{\mathbb{E}}_{s_0 \sim P_0}[V_{\pi_\theta \circ \nu^\star}(s_0)] \\ &\geq \theta \mathop{\mathbb{E}}_{s_0 \sim P_0}[V_{\pi_1 \circ \nu^\star}(s_0)] + (1-\theta) \mathop{\mathbb{E}}_{s_0 \sim P_0}[V_{\pi_2 \circ \nu^\star}(s_0)]. \end{aligned} \tag{25}$$

Thus, the mapping $\pi \mapsto \mathop{\mathbb{E}}_{s_0 \sim P_0}[V_{\pi \circ \nu^*}(s_0)]$ is a concave function and, therefore, continuous in the policy space. Since the policy space is compact, it attains a maximum $\pi^*_{\text{rob}} \in \Pi$.

Corollary 1.1 (Optimal Q-value Function in SA-MDPs)**.** *Under the same conditions as Theorem 3.5, if $Q_{\pi \circ \nu^*}$ is the Q-value function corresponding to the policy π under an optimal adversarial perturbation $\nu^*(\pi)$, then, there exists a policy $\pi^*_{rob} = \pi^* \circ \nu^*(\pi^*)$ such that*

$$\mathbb{E}_{s_0 \sim P_0}\left[Q_{\pi^*_{rob}}(s_0, a)\right] \geq \mathbb{E}_{s_0 \sim P_0}\left[Q_{\pi \circ \nu^*(\pi)}(s_0, a)\right], \quad \forall \pi. \tag{26}$$

Proof of Corollary 1.1. This is a natural extension of Theorem 3.5. The mapping $\pi \mapsto \mathop{\mathbb{E}}_{s_0 \sim P_0}[Q_{\pi \circ \nu^*}(s_0, a)]$ inherits the concavity and continuity properties of the value function mapping. Therefore, by the same arguments, the Q-value function attains a maximum at $\pi^*_{\text{rob}} \in \Pi$. ■

2 Proof of the Theorem 4.1

Consider the actual value function under an optimal fixed adversarial perturbation ν^* where the policy π selects actions based on the perturbed state $\nu^*(s_t)$:

$$\mathop{\mathbb{E}}_{s \sim P_0}[V_{\pi \circ \nu^*}(s)] = \mathop{\mathbb{E}}_{s \sim P_0}\left[\mathop{\mathbb{E}}_{\substack{s_{t+1} \sim P(\cdot|s_t, a_t) \\ a_t \sim \pi(\cdot|\nu^*(s_t))}}\left[\sum_{t=0}^{\infty} \gamma^t R_t \Big| s_0 = s\right]\right]. \tag{27}$$

The Bellman operator for a given policy π in this set-up is defined as:

$$\begin{aligned} &(\mathcal{T}^\pi \overline{V})(s_0) \\ &= \mathop{\mathbb{E}}_{s_0 \sim P_0}\left[\mathop{\mathbb{E}}_{\substack{a \sim \pi(\cdot|\nu^*(s)) \\ s' \sim P(\cdot|s,a)}}\left[R(s,a,s') + \gamma V(s')\right]\right]. \end{aligned} \tag{28}$$

For any two policies π_1 and π_2, and their corresponding value functions $\overline{V}_1$ and $\overline{V}_2$, the Bellman updates are:

$$
\begin{aligned}
&(\mathcal{T}^{\pi_1}\overline{V}_1)(s_0) \\
&= \mathop{\mathbb{E}}_{s_0\sim P_0}\left[\mathop{\mathbb{E}}_{\substack{a_1\sim\pi_1(\cdot|\nu^*(s))\\ s'\sim P(\cdot|s,a_1)}}\left[R(s,a_1,s')+\gamma\overline{V}_1(s')\right]\right], \qquad (29)\\
&(\mathcal{T}^{\pi_2}\overline{V}_2)(s_0) \\
&= \mathop{\mathbb{E}}_{s_0\sim P_0}\left[\mathop{\mathbb{E}}_{\substack{a_2\sim\pi_2(\cdot|\nu^*(s))\\ s'\sim P(\cdot|s,a_2)}}\left[R(s,a_2,s')+\gamma\overline{V}_2(s')\right]\right]. \qquad (30)
\end{aligned}
$$

Then, the difference between them is given by

$$
\begin{aligned}
&(\mathcal{T}^{\pi_1}\overline{V}_1)(s_0) - (\mathcal{T}^{\pi_2}\overline{V}_2)(s_0) \\
&= \mathop{\mathbb{E}}_{s_0\sim P_0}\left[\mathop{\mathbb{E}}_{\substack{a_1\sim\pi_1(\cdot|\nu^*(s))\\ s'\sim P(\cdot|s,a_1)}}\left[R(s,a_1,s')+\gamma\overline{V}_1(s')\right]\right] \\
&- \mathop{\mathbb{E}}_{s_0\sim P_0}\left[\mathop{\mathbb{E}}_{\substack{a_2\sim\pi_2(\cdot|\nu^*(s))\\ s'\sim P(\cdot|s,a_2)}}\left[R(s,a_2,s')+\gamma\overline{V}_2(s')\right]\right]. \qquad (31)
\end{aligned}
$$

To understand how policies and value functions behave under adversarial conditions, we need to analyze how much the value function can change as we tweak the policy. A useful fact here is that the difference in value functions, when comparing two different policies, can be bounded by looking at the maximum difference between these policies. This is because for any functions f and g the following holds:

$$
\max_{\pi_1} f(\pi_1) - \max_{\pi_2} g(\pi_2) \le \max_{\pi}\left(f(\pi)-g(\pi)\right). \qquad (32)
$$

$$
\begin{aligned}
&\left|(\mathcal{T}^{\pi_1}\overline{V}_1)(s_0) - (\mathcal{T}^{\pi_2}\overline{V}_2)(s_0)\right| \\
&\le \mathop{\mathbb{E}}_{s_0\sim P_0}\left[\max_{\pi}\left|\mathop{\mathbb{E}}_{\substack{a\sim\pi(\cdot|\nu^*(s))\\ s'\sim P(\cdot|s,a)}}\left[\gamma\left(\overline{V}_1(s')-\overline{V}_2(s')\right)\right]\right|\right] \\
&\le \gamma\mathop{\mathbb{E}}_{s_0\sim P_0}\left[\max_{\pi}\|\overline{V}_1-\overline{V}_2\|_\infty\right]. \qquad (33)
\end{aligned}
$$

In simpler terms, even though the adversary tries to disrupt the learning process, the extent to which it can do so is limited. This bounded difference is a key step in showing that our Bellman operator is a contraction, meaning that it pulls the value function closer to a stable final form with each iteration. As a result, we can confidently say that the learning process will converge to a unique value function that represents the best policy under the worst adversarial conditions. Formally, this is concluded as follows.

Since the maximum difference in value functions is bounded by the infinity norm, by (33) we obtain the following:

$$
\left|(\mathcal{T}^{\pi_1}\overline{V}_1)(s_0) - (\mathcal{T}^{\pi_2}\overline{V}_2)(s_0)\right| \le \gamma\|\overline{V}_1-\overline{V}_2\|_\infty. \qquad (34)
$$

This means that $\mathcal{T}^\pi$ is a contraction mapping with respect to the infinity norm. By the Banach fixed-point theorem, this implies that the sequence of value functions $\{\overline{V}_{\pi^i\circ\nu^*}\}$ with actual state values under fixed optimal adversarial perturbations converges to a unique fixed point. ■

3 Proof of Theorem 4.2

Proof. Let Q_1 and Q_2 be value functions for two optimal policies π_1 and π_2 with corresponding adversaries $\nu^*_{\pi_1}$ and $\nu^*_{\pi_2}$. The difference in their perturbed reward functions is:

$$
\tilde{R}_1(s,a) - \tilde{R}_2(s,a) = \mathbf{w}^*\cdot\left[\phi(\nu^*_{\pi_1}(s)) - \phi(\nu^*_{\pi_2}(s))\right].
$$

By Lipschitz continuity of ϕ and the boundedness of perturbations:

$$
\|\phi(\nu^*_{\pi_1}(s)) - \phi(\nu^*_{\pi_2}(s))\| \le L_\phi\cdot\|\nu^*_{\pi_1}(s) - \nu^*_{\pi_2}(s)\| \le 2L_\phi\delta,
$$

which implies:

$$
\left|\tilde{R}_1(s,a) - \tilde{R}_2(s,a)\right| \le 2L_\phi\delta\cdot\|\mathbf{w}^*\|.
$$

Now consider the difference in Bellman updates:

$$|\mathcal{T}Q_1(s,a) - \mathcal{T}Q_2(s,a)| \leq \left|\tilde{R}_1(s,a) - \tilde{R}_2(s,a)\right| + \gamma \left|\mathbb{E}_{s'}\mathbb{E}_{s_0}\left[Q_1(s',a') - Q_2(s',a')\right]\right|$$
$$\leq 2L_\phi\delta \cdot \|\mathbf{w}^*\| + \gamma \cdot \mathbb{E}_{s'}\mathbb{E}_{s_0}\left[Q_1(s',a') - Q_2(s',a')\right].$$

Since both policies are optimal under their own adversaries:

$$\mathbb{E}_{s_0}\left[V_{\pi_1 \circ \nu^*_{\pi_1}}(s_0)\right] = \mathbb{E}_{s_0}\left[V_{\pi_2 \circ \nu^*_{\pi_2}}(s_0)\right],$$

which implies:

$$\mathbb{E}_{s_0}[Q_1(s_1,a_1)] = \mathbb{E}_{s_0}[Q_2(s_1,a_1)],$$

so the expected Q-value difference vanishes:

$$\gamma\mathbb{E}_{s'}\mathbb{E}_{s_0}[Q_1(s',a') - Q_2(s',a')] = 0.$$

Thus,

$$\|\mathcal{T}Q_1 - \mathcal{T}Q_2\|_\infty \leq 2L_\phi\delta \cdot \|\mathbf{w}^*\|.$$

If $2L_\phi\delta \cdot \|\mathbf{w}^*\| < 1$, then $\mathcal{T}$ is a contraction, and by Banach's Fixed Point Theorem, it has a unique fixed point to which value iteration converges. ■

4 Proof of Lemma 4.3

If $\phi(\cdot)$ is Lipschitz continuous with a constant L, then for any state s_t and its perturbed counterpart $\nu^*(s_t)$,

$$\|\phi(s_t) - \phi(\nu^*(s_t))\|_2 \leq L\|s_t - \nu^*(s_t)\|_2 \leq L\delta. \tag{35}$$

Considering the absolute value of the difference between the actual feature expectation and the believed feature expectations, we have the following.

$$\|\mu_{\pi\circ\nu^*} - \tilde{\mu}_{\pi\circ\nu^*}\|_2 = \left\|\mathbb{E}_{s\sim P_0}\left[\sum_{t=0}^{\infty}\gamma^t\left(\phi(s_t) - \phi(\nu^*(s_t))\right)\right]\right\|_2$$
$$\leq \mathbb{E}_{s\sim P_0}\left[\sum_{t=0}^{\infty}\gamma^t\left\|\phi(s_t) - \phi(\nu^*(s_t))\right\|_2\right]. \tag{36}$$

Using the Lipschitz continuity of ϕ and the bounds on the perturbations ν^*, it follows that

$$\|\mu_{\pi\circ\nu^*} - \tilde{\mu}_{\pi\circ\nu^*}\|_2 \leq \mathbb{E}_{s\sim P_0}\left[\sum_{t=0}^{\infty}\gamma^t L\delta\right]$$
$$= L\delta\sum_{t=0}^{\infty}\gamma^t$$
$$= \frac{L\delta}{1-\gamma}. \tag{37}$$

By choosing $\varepsilon_0 = \frac{L\delta}{1-\gamma}$, we have

$$\|\mu_{\pi\circ\nu^*} - \tilde{\mu}_{\pi\circ\nu^*}\|_2 \leq \varepsilon_0. \tag{38}$$

■

5 Restatement and Proof of Theorem 4.4

Let $\tilde{\mu}_E = \tilde{\mu}_{\pi_E\circ\nu^*}$ be believed perturbed feature expectation of the expert with features $\phi : \mathcal{S} \to [0,1]^k$ in a given SA-MDP under perturbations ν^*. Besides, $\tilde{\mu}_{\pi\circ\nu^*}$ denotes the believed perturbed feature expectation of a policy π. Also, $\pi^{(i+1)} = \pi^*_{rob}$ is the optimal policy for the SA-MDP\R augmented with reward $R(s_{\nu^*}) = (\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}) \cdot \phi(s_{\nu^*})$, i.e.,

$$\pi^{(i+1)} = \arg\max_\pi (\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}) \cdot \tilde{\mu}_{\pi\circ\nu^*}, \tag{39}$$

where $\tilde{\mu}^{(i)}_{\pi\circ\nu^*}$ is the believed perturbed feature expectation at iteration i and $\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*} = \tilde{\mu}(\pi^*_{rob})$. In addition, $\tilde{\mu}^{(i+1)}_{proj}$ is the projection of $\tilde{\mu}_E$ onto the line through $\tilde{\mu}^{(i)}_{\pi\circ\nu^*}$ and $\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*}$. Then

$$\frac{\|\tilde{\mu}_E - \tilde{\mu}^{(i+1)}_{proj}\|_2}{\|\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}\|_2} \leq \frac{k}{\sqrt{k^2 + (1-\gamma)^2\|\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}\|_2^2}}. \tag{40}$$

Let $\tilde{M}_{\text{Co}}$ be the convex hull of the set of believed perturbed feature expectations of all policies π and let $\tilde{\mu}_E \in \tilde{M}_{\text{Co}}$[3] be believed perturbed feature expectation of the expert agent.

For mathematical convenience, we set the origin of the coordinate system at $\tilde{\mu}^{(i)}_{\pi\circ\nu^*}$, the believed perturbed feature expectation at iteration i. This allows us to express all the vectors relative to $\tilde{\mu}^{(i)}_{\pi\circ\nu^*}$. Define the projection $\tilde{\mu}^{(i+1)}_{\text{proj}}$ as the point on the line through $\tilde{\mu}^{(i)}_{\pi\circ\nu^*}$ and $\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*}$ that is closest to $\tilde{\mu}_E$. Formally, this projection is given by:

$$\tilde{\mu}^{(i+1)}_{\text{proj}} = \theta\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*} + (1-\theta)\tilde{\mu}^{(i)}_{\pi\circ\nu^*}, \tag{41}$$

where

$$\theta = \frac{(\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}) \cdot (\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*} - \tilde{\mu}^{(i)}_{\pi\circ\nu^*})}{\|\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*} - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}\|_2^2}. \tag{42}$$

The coefficient θ ensures that $\tilde{\mu}^{(i+1)}_{\text{proj}}$ is the projection which minimizes the squared distance to $\tilde{\mu}_E$.

Next, we calculate the squared distance between $\tilde{\mu}_E$ and $\tilde{\mu}^{(i+1)}_{\text{proj}}$ as

$$\begin{aligned}\|\tilde{\mu}_E - \tilde{\mu}^{(i+1)}_{\text{proj}}\|_2^2 &= \left\|\tilde{\mu}_E - \left(\theta\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*} + (1-\theta)\tilde{\mu}^{(i)}_{\pi\circ\nu^*}\right)\right\|_2^2 \\ &= \|\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}\|_2^2 - \frac{\left((\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}) \cdot (\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*} - \tilde{\mu}^{(i)}_{\pi\circ\nu^*})\right)^2}{\|\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*} - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}\|_2^2}.\end{aligned} \tag{43}$$

By the Cauchy-Schwarz inequality and noting that for any vector $\mathbf{x}$, the norms satisfy $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \sqrt{n}\|\mathbf{x}\|_\infty$, we obtain the following bounds:

$$\|\tilde{\mu}^{(i+1)}_{\pi\circ\nu^*}\|_2 \le \sqrt{k}\left(\frac{1}{1-\gamma}\right), \quad \|\tilde{\mu}_E\|_2 \le \sqrt{k}\left(\frac{1}{1-\gamma}\right). \tag{44}$$

This leads to the inequality:

$$\frac{\|\tilde{\mu}_E - \tilde{\mu}^{(i+1)}_{\text{proj}}\|_2^2}{\|\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}\|_2^2} \le \frac{k^2/(1-\gamma)^2}{k^2/(1-\gamma)^2 + \|\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi\circ\nu^*}\|_2^2}. \tag{45}$$

This concludes the proof. ■

6 Proof of Lemma 4.6

Since the reward function is assumed to be linear in features, for the *actual perturbed feature expectations setting*, the reward at each step is $R(s_t, a_t, s_{t+1}) = \mathbf{w}^* \cdot \phi(s_t)$, leading to:

$$V_{\pi^*_{\text{rob}}}(s) = \mathbb{E}_{\pi^*_{\text{rob}}}\left[\sum_{t=0}^{\infty}\gamma^t\mathbf{w}^*\cdot\phi(s_t)\Big|s_0 = s\right]. \tag{46}$$

Taking expectations over the initial state distribution P_0, we obtain:

$$\mathbb{E}_{s_0\sim P_0}[V_{\pi^*_{\text{rob}}}(s_0)] = \mathbf{w}^*\cdot\mu_\pi. \tag{47}$$

Similarly, under *perturbation*, the reward is given by:

$$\tilde{R}(s_t, a_t, s_{t+1}) = \mathbf{w}^*\cdot\phi(\nu^*(s_t)). \tag{48}$$

Thus, the perturbed value function is:

$$\tilde{V}_{\pi_E}(s) = \mathbb{E}_{\pi^*_{\text{rob}}}\left[\sum_{t=0}^{\infty}\gamma^t\mathbf{w}^*\cdot\phi(\nu^*(s_t))\Big|s_0 = s\right]. \tag{49}$$

$$\mathbb{E}_{s_0\sim P_0}[\tilde{V}_{\pi^*_{\text{rob}}}(s_0)] = \mathbf{w}^*\cdot\tilde{\mu}_\pi. \tag{50}$$

Let π^*_{rob} be an optimal policy with respect to Definition 3.4 in a SA-MDP. Then its expected value function is invariant under perturbations:

$$\mathbb{E}_{s_0\sim P_0}[V_{\pi^*_{\text{rob}}}(s_0)] = \mathbb{E}_{s_0\sim P_0}[\tilde{V}_{\pi^*_{\text{rob}}}(s_0)]. \tag{51}$$

This is due to the fact that in both objectives, the weight vector $\mathbf{w}^*$ stays the same and the only effect of perturbation is on the observed features, not on the underlying reward structure. Substituting the feature expectations:

$$\mathbf{w}^*\cdot\mu_{\pi^*_{\text{rob}}} = \mathbf{w}^*\cdot\tilde{\mu}_{\pi^*_{\text{rob}}}. \tag{52}$$

We conclude:

$$\mu_{\pi^*_{\text{rob}}} = \tilde{\mu}_{\pi^*_{\text{rob}}}. \tag{53}$$

■

[3] We assume feature expectations of the expert can be calculated accurately.

7 Proof of Lemma 4.7

From Lemma 4.6 we know that $\mu_E = \tilde{\mu}_E$. Now, consider the difference between the expert and the observer's feature expectations in perturbed spaces:

$$\|\tilde{\mu}_E - \tilde{\mu}_{\pi \circ \nu^*}\|_2 \leq \varepsilon. \tag{54}$$

Using the expert's feature expectation invariance:

$$\|\tilde{\mu}_E - \tilde{\mu}_{\pi \circ \nu^*}\|_2 = \|\mu_E - \tilde{\mu}_{\pi \circ \nu^*}\|_2. \tag{55}$$

Consider the term inside the norm. Adding and substituting $\mu_{\pi \circ \nu^*}$ gives us:

$$\mu_E - \tilde{\mu}_{\pi \circ \nu^*} = \mu_E - \mu_{\pi \circ \nu^*} + \mu_{\pi \circ \nu^*} - \tilde{\mu}_{\pi \circ \nu^*}. \tag{56}$$

Taking the 2-norm and applying the triangle inequality:

$$\|\mu_E - \tilde{\mu}_{\pi \circ \nu^*}\|_2 \leq \|\mu_E - \mu_{\pi \circ \nu^*}\|_2 + \|\mu_{\pi \circ \nu^*} - \tilde{\mu}_{\pi \circ \nu^*}\|_2. \tag{57}$$

Similarly, by reversing roles:

$$\|\mu_E - \mu_{\pi \circ \nu^*}\|_2 \leq \|\mu_E - \tilde{\mu}_{\pi \circ \nu^*}\|_2 + \|\mu_{\pi \circ \nu^*} - \tilde{\mu}_{\pi \circ \nu^*}\|_2. \tag{58}$$

Since norms are always non-negative, we conclude that:

$$\|\mu_E - \tilde{\mu}_{\pi \circ \nu^*}\|_2 = \|\mu_E - \mu_{\pi \circ \nu^*}\|_2. \tag{59}$$

Thus, we obtain:

$$\|\mu_E - \mu_{\pi \circ \nu^*}\|_2 \leq \varepsilon \quad \iff \quad \|\tilde{\mu}_E - \tilde{\mu}_{\pi \circ \nu^*}\|_2 \leq \varepsilon. \tag{60}$$

This completes the proof. ■

8 Proof of the Theorem 4.8

The goal is for the feature expectations of the learned policy to converge in such a way that the distance between the expert feature expectations μ_E and the actual perturbed feature expectations of the learned policy $\mu_{\pi \circ \nu^*}$ is small, within a specified error margin ε, that is,

$$\|\mu_E - \mu_{\pi \circ \nu^*}\|_2 \leq \varepsilon. \tag{61}$$

However, because we only have access to the believed perturbed feature expectations, we consider convergence in terms of the believed perturbed feature expectations $\|\tilde{\mu}_E - \tilde{\mu}_{\pi \circ \nu^*}\|_2$. By Lemma 4.7 we know that convergence in two domains is equivalent given the robustness of the expert agent with respect to Definition 3.4. Therefore, showing the number of iterations required for convergence in the believed perturbed feature expectation space is equivalent to the actual ones.

Consider the inequality established in Lemma 4.4 and let $d^{(i)} = \|\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi \circ \nu^*}\|_2$ denote the distance at iteration i.

For any step i when $\|\tilde{\mu}_E - \tilde{\mu}_{\pi^{(i)} \circ \nu^*}\|_2 > \varepsilon$,

$$\frac{\|\tilde{\mu}_E - \tilde{\mu}^{(i+1)}_{\text{proj}}\|_2^2}{\|\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi \circ \nu^*}\|_2^2} \leq \frac{k^2/(1-\gamma)^2}{k^2/(1-\gamma)^2 + \varepsilon^2}. \tag{62}$$

By re-arranging, we arrive at

$$\frac{\|\tilde{\mu}_E - \tilde{\mu}^{(i+1)}_{\text{proj}}\|_2}{\|\tilde{\mu}_E - \tilde{\mu}^{(i)}_{\pi \circ \nu^*}\|_2} \leq \frac{k}{\sqrt{k^2 + (1-\gamma)^2 \varepsilon^2}}, \tag{63}$$

Then, the reduction at each step is given by

$$d^{(i+1)} \leq \frac{k}{\sqrt{k^2 + (1-\gamma)^2 \varepsilon^2 d^{(i)}}}. \tag{64}$$

By applying this reduction iteratively over T rounds, we arrive at

$$d^{(T)} \leq \left(\frac{k}{\sqrt{k^2 + (1-\gamma)^2 \varepsilon^2}} \right)^T d^{(0)}. \tag{65}$$

For convergence, we need $d^{(T)} \leq \varepsilon$. Thus, it should hold

$$\left(\frac{k}{\sqrt{k^2+(1-\gamma)^2\varepsilon^2}}\right)^T d^{(0)} \leq \varepsilon. \tag{66}$$

Since $\tilde{M}_{\text{Co}} \in [0,1]^k$, we have $d^{(0)} < \frac{k}{1-\gamma}$. To ensure that dividing ε by $d^{(0)}$ is valid, we recognize that the term $\left(\frac{k}{\sqrt{k^2+(1-\gamma)^2\varepsilon^2}}\right)^T$ represents the cumulative reduction over T iterations. As T increases, this term becomes smaller, eventually making it sufficiently small relative to ε. Then there exists T such that

$$\left(\frac{k}{\sqrt{k^2+(1-\gamma)^2\varepsilon^2}}\right)^T \leq \left(\frac{\varepsilon}{\frac{k}{1-\gamma}}\right). \tag{67}$$

Taking logarithms on both sides results in

$$T \log\left(\frac{k}{\sqrt{k^2+(1-\gamma)^2\varepsilon^2}}\right) \leq \log\left(\frac{\varepsilon}{\frac{k}{1-\gamma}}\right). \tag{68}$$

Since $\log\left(\frac{k}{\sqrt{k^2+(1-\gamma)^2\varepsilon^2}}\right) \approx -\frac{(1-\gamma)^2\varepsilon^2}{2k^2}$ for small ε, we can conclude that

$$T \geq \frac{2k^2}{(1-\gamma)^2\varepsilon^2} \log\left(\frac{k}{(1-\gamma)\varepsilon}\right). \tag{69}$$

Given the reduction factor and the bounds on feature expectations, the number of iterations required for convergence yields

$$T = O\left(\frac{k^2}{(1-\gamma)^2\varepsilon)^2} \log\frac{1}{\varepsilon}\right). \tag{70}$$

■

9 Additional Theoretical Results

9.1 Robust Separability Under Adversarial Perception

Motivated by a large portion of the practical deployment scenarios—such as human demonstrations for robots or legacy datasets captured under clean conditions, we examine whether SAMM-IRL remains applicable when the expert was trained in a non-adversarial environment. Similar concerns have been raised in the literature on robust imitation learning and adversarial robustness [2, 11, 22], where aligning robustness requirements with real-world sensor noise or attacks remains an open challenge. To this end, we derive a sufficient condition under which robust separability still holds despite the observer facing adversarial perturbations unseen by the expert.

SAMM-IRL addresses IRL under adversarial uncertainty, where the observer perceives perturbed expert demonstrations and perturbed versions of its own states. For successful reward recovery, the expert must remain distinguishable i.e., preserve a feature-margin over suboptimal policies under state-adversarial perturbations. This ensures that the max-margin separation remains valid.

In our main formulation, the expert is trained in an SA-MDP against a fixed optimal adversary, and both the expert and the observer operate under the same fixed perturbation during training and inference. During learning, the observer is subjected to a potentially re-optimized adversary that targets its own evolving policy, while the expert perturbations remain fixed. This asymmetric setting models realistic conditions in which sensor interference can adapt to the behavior of a deployed agent but not retroactively alter recorded expert data.

To understand whether SAMM-IRL remains viable when the expert was trained in a clean environment, we derive a sufficient condition under which robust separability still holds.

Lemma 9.1 (Bound on Adversarial Degradation)**.** *Let the adversary perturb any state s within radius ϵ under some norm: $|\nu(s) - s| \leq \epsilon$. Let the feature map ϕ be Lipschitz continuous with constant L. Over a horizon H, the cumulative degradation in features is bounded by $H \cdot L \cdot \epsilon$.*

Proof. Since ϕ is Lipschitz with constant L, we have for each t:

$$\|\phi(\nu(s_t)) - \phi(s_t)\| \leq L \cdot \|\nu(s_t) - s_t\| \leq L \cdot \epsilon.$$

Applying the triangle inequality over all steps $t = 0, 1, \ldots, H$, we get:

$$\left\|\sum_{t=0}^{H} \phi(\nu(s_t)) - \phi(s_t)\right\| \leq \sum_{t=0}^{H} \|\phi(\nu(s_t)) - \phi(s_t)\| \leq (H+1) \cdot L \cdot \epsilon.$$

If we redefine H to be the planning horizon with $H+1$ steps, the result holds as stated. ■

Proposition 9.2 (Robust Separability Condition for SAMM-IRL)**.** *Let $\phi : \mathcal{S} \to \mathbb{R}^d$ be Lipschitz with constant L, and $|\nu(s) - s| \leq \epsilon$. Let H be the effective planning horizon. Suppose the clean expert π_E has margin $\delta > 0$:*

$$\mathbf{w}^\top(\mu_{\pi_E} - \mu_\pi) \geq \delta, \quad \forall \pi \in \Pi, \quad \|\mathbf{w}\| \leq 1. \tag{71}$$

Then SAMM-IRL remains valid under perturbations:

$$\mathbf{w}^\top(\mu_{\pi_E \circ \nu} - \mu_{\pi \circ \nu}) > 0, \quad \forall \pi \in \Pi, \tag{72}$$

if $\delta > 2 \cdot H \cdot L \cdot \epsilon$.

Proof. Let μ_π denote the clean (unperturbed) feature expectation, and $\tilde{\mu}_{\pi \circ \nu}$ the believed perturbed expectation:

$$\tilde{\mu}_{\pi \circ \nu} = \mu_\pi + \Delta_\pi, \quad \text{with } \|\Delta_\pi\| \leq HL\epsilon.$$

Similarly for the expert:

$$\tilde{\mu}_{\pi_E \circ \nu} = \mu_{\pi_E} + \Delta_E, \quad \text{with } \|\Delta_E\| \leq HL\epsilon.$$

Then,

$$\mathbf{w}^\top(\tilde{\mu}_{\pi_E \circ \nu} - \tilde{\mu}_{\pi \circ \nu}) = \mathbf{w}^\top(\mu_{\pi_E} - \mu_\pi) + \mathbf{w}^\top(\Delta_E - \Delta_\pi).$$

Using the triangle inequality and $\|\mathbf{w}\| \leq 1$, we get:

$$\mathbf{w}^\top(\Delta_E - \Delta_\pi) \geq -\|\Delta_E - \Delta_\pi\| \geq -2HL\epsilon.$$

So,

$$\mathbf{w}^\top(\tilde{\mu}_{\pi_E \circ \nu} - \tilde{\mu}_{\pi \circ \nu}) \geq \delta - 2HL\epsilon > 0,$$

which holds whenever $\delta > 2HL\epsilon$. ■

This bound offers a guideline for assessing whether the expert's advantage survives adversarial degradation. It enables practitioners to evaluate whether SAMM-IRL remains applicable under given adversarial conditions and expert quality.

10 Implementation Details

All of our experiments were conducted in a Conda environment using Python, making use of CuPy for GPU acceleration on a 2080 Ti. The source code and evaluations in different environments are provided in the GitHub repository at [3].

10.1 Hyperparameters

Experimental results presented in Table 2(a) and Table 5 use the hyperparameters given in Table 4.

Table 4. Hyperparameters for Reproducing the Results

Hyperparameter	Uniform
Epsilon (ε)	0.1
Gamma (γ)	0.9
Alpha (α)	0.1
Max Episodes	500
Max Iterations per Episode	100
Max Steps per Trajectory	100
Grid Size	5x5

10.2 Full Results for Table 2(a)

Table 5. Feature expectation matching analysis under different adversary types

Feature	Adv. Type	μ_E	μ_{π_ε}	$\mathbf{w}^*$	$\mathbf{w}_\varepsilon$	$\langle \mathbf{w}^*, \mu_{\pi_E \circ \nu^*} \rangle$	$\langle \mathbf{w}^*, \mu_{\pi_\varepsilon \circ \nu^*} \rangle$
goal reached	Uniform	0.12	0.01	0.68	0.24	0.08	0.01
	RS Adv.	0.07	0.00		0.39	0.05	0.00
	Critic Adv.	0.02	0.00		0.31	0.01	0.00
horizontal direction to goal	Uniform	0.44	0.06	0.34	0.51	0.22	0.03
	RS Adv.	0.33	0.03		0.54	0.11	0.02
	Critic Adv.	0.26	0.28		0.04	0.09	0.01
vertical direction to goal	Uniform	0.44	0.44	0.34	0.25	0.11	0.11
	RS Adv.	0.34	0.05		0.55	0.11	0.02
	Critic Adv.	0.29	0.22		0.53	0.10	0.07
danger zones	Uniform	0.00	0.42	-0.51	-0.64	0.00	-0.27
	RS Adv.	0.00	0.00		-0.46	0.00	0.00
	Critic Adv.	0.06	0.09		-0.62	-0.03	-0.06
near boundary	Uniform	1.00	0.56	-0.17	0.17	0.17	0.10
	RS Adv.	0.61	0.97		-0.46	-0.10	-0.45
	Critic Adv.	0.45	0.57		-0.49	-0.22	-0.28