A. Discounted Markov Games

A.1. Preliminaries

We define the framework of discounted, two-player zero-sum Markov Games (MG) with finite state space and continuous action space. A MG is determined by the 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, P, R, \gamma)$ (Patek, 1997). Here $\mathcal{S}$ is a finite state space, $\mathcal{A}$ and $\mathcal{B}$ are compact subsets of $\mathbb{R}^A$, which represent the agent and adversary, respectively. For any $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ let the dynamics $P = P(\cdot \mid s, a, b)$ be a probability measure on $\mathcal{S}$, and let the reward function $r(s, a, b)$ be a bounded measureable function on $\mathcal{A} \times \mathcal{B}$ for any $s \in \mathcal{S}$. Consider a strategy of the players μ, ν, where both are probability measures over Borel sets of $\mathcal{A}, \mathcal{B}$, respectively. Let $r^{\mu,\nu} \in \mathbb{R}^{|\mathcal{S}|}$ where $r^{\mu,\nu}(s) \overset{\text{def}}{=} \mathbb{E}_{a\sim\mu, b\sim\nu}[r(s, \mu, \nu)]$, and the dynamics $P^{\mu,\nu} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, where $P^{\mu,\nu}_{i,j} \overset{\text{def}}{=} \mathbb{E}_{a\sim\mu, b\sim\nu}[P(s_j \mid s_i, \mu, \pi_B)]$ and is a stochastic matrix. Following notation from Maitra & Parthasarathy (1970), we denote P_A and P_B as the set of probability measures on the Borel sets of $\mathcal{A}$ and $\mathcal{B}$, respectively.

Definition 1. *The value of fixed strategy μ, ν is given by $v^{\mu,\nu} = \sum_{t=0}^{\infty} \gamma^t (P^{\mu,\nu})^t r^{\mu,\nu}$. Given a fixed $\nu \in P_B$ the value of the optimal counter strategy of player A is $v^{\nu} = \sup_{\mu\in P_A} v^{\mu,\nu}$. Accordingly, for a fixed $\mu \in P_A$ the value of the optimal counter strategy of player B is $v^{\mu} = \inf_{\nu\in P_B} v^{\mu,\nu}$. Furthermore, if the $\sup$ and $\inf$ are attainable, we refer to $\arg\min_{\nu\in P_B} v^{\mu,\nu}$ and $\arg\max_{\mu\in P_A} v^{\mu,\nu}$ as optimal counter strategies to μ and ν, respectively.*

We make the following assumptions on the dynamics and reward functions.

Assumption 1.

- *Both $\mathcal{A}, \mathcal{B}$ are compact metric spaces.*
- *For any $s \in \mathcal{S}$ the reward r is continuous and bounded function on $\mathcal{A} \times \mathcal{B}$.*
- *For any $s \in \mathcal{S}$, whenever $(a_n, b_n) \to (a, b)$, where $(a_n, b_n), (a, b) \in \mathcal{A} \times \mathcal{B}$, then $P(\cdot \mid s, a_n, b_n)$ converges weakly to $P(\cdot \mid s, a, b)$.*

In the rest of the section we follow (Patek, 1997)[Section 2-3] that analyzed zero-sum MG for stochastic shortest paths, while performing minor modifications for the discounted and continuous action-space setup.

Define the following Bellman operators.

Definition 2. *Let P_A and P_B be the set of all probability measures on the Borel Sets of $\mathcal{A}$ and $\mathcal{B}$, respectively, $\mu \in P_A, \nu \in P_B$, and let $v \in \mathbb{R}^{|\mathcal{S}|}$. The Bellman operator, and Fixed-Policy Bellman operators are according to the following.*

$$T^{\mu,\nu} v = r^{\mu,\nu} + \gamma P^{\mu,\nu} v,$$
$$T^{\mu} v = \min_{\nu\in P_B} \left(r^{\mu,\nu} + \gamma P^{\mu,\nu} v\right), \ \bar{T}^{\nu} v = \max_{\mu\in P_A} \left(r^{\mu,\nu} + \gamma P^{\mu,\nu} v\right)$$
$$T v = \max_{\mu\in P_A} \min_{\nu\in P_B} \left(r^{\mu,\nu} + \gamma P^{\mu,\nu} v\right), \ \bar{T} v = \min_{\nu\in P_B} \max_{\mu\in P_A} \left(r^{\mu,\nu} + \gamma P^{\mu,\nu} v\right),$$

where equality holds component-wise.

Notice that the $\max$ and $\min$ are attainable since P_A, P_B are compact sets. Furthermore, by Maitra & Parthasarathy (1970)[Lemma 2.2] and under Assumption 1, both the $\max$ and $\min$ are continuous and bounded. Thus, we can replace $\sup\inf$ and $\inf\sup$ by corresponding $\max$ and $\min$.

We have the following important lemma.

Lemma 1. *For any bounded $v \in \mathbb{R}^{|\mathcal{S}|}$, $Tv = \bar{T}v$.*

Proof. Following similar arguments as in Maitra & Parthasarathy (1970), Equation 2, and using Sion's minimax theorem (Sion et al., 1958)[Theorem 3.4], for any $s \in \mathcal{S}$ we have that,

$$\sup_{\mu\in P_A} \inf_{\nu\in P_B} r^{\mu,\nu}(s) + P^{\mu,\nu} v(s) = \inf_{\nu\in P_B} \sup_{\mu\in P_A} r^{\mu,\nu}(s) + P^{\mu,\nu} v(s).$$

Since P_A, P_B are compact and $r^{\mu,\nu} + P^{\mu,\nu} v$ is bounded and continuous on $\mathcal{A} \times \mathcal{B}$ for any $s \in \mathcal{S}$, the $\sup, \inf$ can be replaced by $\min, \max$ (e.g., by Maitra & Parthasarathy (1970)[Lemma 2.2]). □

The analysis in Patek (1997) is based on assumption R, which results in $Tv = \bar{T}v$. Since we allow the agents to use mixed-strategies, according to Lemma 1, we obtain $Tv = \bar{T}v$ in our setup as well. Furthermore, since we use discounted MG, assumption SSP in Patek (1997) is also satisfied. Every strategy (μ, ν) is proper; it terminates with probability one, as the discount factor (γ) is smaller than 1.

Lemma 2. *$T^{\mu,\nu}$, T^{μ}, $\bar{T}^{\nu}$, T are γ contractions in the sup-norm.*

Proof. We follow similar technique as in Patek (1997), adjusted to our setup. Let $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$. Then,

$$T^{\mu,\nu}v_1 - T^{\mu,\nu}v_2 = \gamma P^{\mu,\nu}(v_1 - v_2) \leq \gamma P^{\mu,\nu}\mathbf{1}||v_1 - v_2||_\infty = \gamma \mathbf{1}||v_1 - v_2||_\infty,$$

where $\mathbf{1}$ is the one vector. The last relation holds since $P^{\mu,\nu}$ is a stochastic matrix and thus $P^{\mu,\nu}\mathbf{1} = \mathbf{1}$. By repeating the same argument for $T^{\mu,\nu}v_2 - T^{\mu,\nu}v_1$ and taking the sup-norm we conclude that $||T^{\mu,\nu}v_1 - T^{\mu,\nu}v_2||_\infty \leq \gamma||v_1 - v_2||_\infty$.

We now prove similar result on T^{μ}. Let $\nu, \nu' \in P_B$ such that $T^{\mu}v_1 = T^{\mu,\nu}v_1,\ T^{\mu}v_2 = T^{\mu,\nu'}v_2$. Then,

$$\begin{aligned} T^{\mu}v_1 - T^{\mu}v_2 &\leq T^{\mu,\nu}v_1 - T^{\mu,\nu}v_2, \\ T^{\mu}v_2 - T^{\mu}v_1 &\leq T^{\mu,\nu'}v_1 - T^{\mu,\nu'}v_2. \end{aligned}$$

By taking the sup-norm and using the fact $T^{\mu,\nu}$ is a γ-contraction, we conclude that T^{μ} is also a γ-contraction. Similar argument establishes that $\bar{T}^{\nu}$ is a γ-contraction.

Lastly, let $\mu \in P_A$ such that $Tv_2 = T^{\mu}v_2$, and $\nu \in P_B$ such that $T^{\mu}v_1 = T^{\mu,\nu}v_1$. Then,

$$\begin{aligned} Tv_1 - Tv_2 &= Tv_1 - T^{\mu}v_2 \\ &\leq T^{\mu}v_1 - T^{\mu}v_2 \\ &= T^{\mu,\nu}v_1 - T^{\mu}v_2 \\ &\leq T^{\mu,\nu}v_1 - T^{\mu,\nu}v_2. \end{aligned}$$

Similar argument leads to $Tv_2 - Tv_1 \leq T^{\mu,\nu}v_2 - T^{\mu,\nu}v_1$ for properly defined μ, ν. Again, by taking the sup norm and using the fact that $T^{\mu,\nu}$ is a γ-contraction we conclude the proof. □

The following propositions relate the fixed-point of $T_{\mu}, \bar{T}^{\nu}$ to the values and policies defined in 1. Furthermore, the last one establishes the fact the zero-sum MG has value.

Proposition 3. *The following claims hold.*

- *Let $\mu \in P_A, \nu \in P_B$ be stationary policies. The value $v^{\mu,\nu}$ is the fixed point of the operator $T^{\mu,\nu}$, $v^{\mu,\nu} = T^{\mu,\nu}v^{\mu,\nu}$.*
- *Given a policy $\nu \in P_B$, $v^{\nu} = \sup_{\mu \in P_A}$ is the unique fixed point of $\bar{T}^{\nu}$. Furthermore, the* sup *is attainable in the set A.*
- *Given a policy $\mu \in P_A$, $v^{\mu} = \inf_{\nu \in P_B}$ is the unique fixed point of T_{μ}. Furthermore, the* inf *is attainable in the set B.*

Proof. The proof of the first claim is standard, e.g., Puterman (1994)[Section 6.1]. By fixing a policy for any of the players the problem amounts for solving a single agent MDP (e.g., Puterman (1994)). Due to Assumption 1, the reward and dynamics of the MDP are also continuous and bounded. Since the action set in compact for both player A and B, we can use Puterman (1994)[Theorem 6.2.10] and conclude the proof. □

Proposition 4. *The unique fixed point $v^* = Tv^*$ is also the equilibrium value of the zero-sum MG, $v^* = \sup_{\mu \in P_A} \inf_{\nu \in P_B} v^{\mu,\nu} = \inf_{\nu \in P_B} \sup_{\mu \in P_A} v^{\mu,\nu}$, thus, the MG has a well defined value.*

Furthermore, the stationary policies $\mu \in P_A, \nu \in P_B$ for which $v^* = \bar{T}v^* = Tv^* = T^{\mu,\nu}v^*$ are in Nash-Equilibrium, and satisfy $v^{\mu',\nu^*} \leq v^* \leq v^{\mu^*,\nu}$ for any $\nu' \in P_B,\ \mu' \in P_A$.

Proof. See proof Patek (1997)[Proposition 3.2]. □

Algorithm 1 Zero-Sum Markov-Game PI

```
Initialize: $\nu_0, k = 0$
while stopping criterion is not satisfied do
    $\mu_k \in \arg\max_\mu v^{\mu,\nu_k}$
    $\nu_{k+1} \in \arg\min_\nu \bar{T}^\nu v^{\mu_k,\nu_k}$
    $k \leftarrow k+1$
end while
Return $\pi_{k-1}$
```

Algorithm 2 Soft Zero-Sum Markov-Game PI

```
Initialize: $\nu_0, k = 0, \eta \in (0, 1]$
while stopping criterion is not satisfied do
    $\mu_k \in \arg\max_\mu v^{\mu,\nu_k}$
    $\nu' \in \arg\min_\nu \bar{T}^\nu v^{\mu_k,\nu_k}$
    $\nu_{k+1} = (1-\eta)\nu_k + \eta\nu'$
    $k \leftarrow k+1$
end while
Return $\pi_{k-1}$
```

A.2. Policy Iteration and Soft Policy Iteration for Zero-Sum Markov Games

In this section, we formulate two PI schemes that solve a zero-sum MG. The Zero-Sum MG PI scheme (see Alg. 1) is a well known one (Hoffman & Karp, 1966; Rao et al., 1973; Hansen et al., 2013).

The Soft Zero-Sum MG PI (see Alg. 2) generalizes the usual PI. Instead of updating with a 1-step greedy policy it updates softly w.r.t. the 1-step greedy policy. Although this generalization has been analyzed extensively for a single-agent PI (e.g., (Kakade & Langford, 2002; Scherrer, 2014)), to the best of our knowledge, it was not analyzed in the context of Markov-Games.

By generalizing arguments from (Scherrer, 2014) to framework of Zero-Sum MG (defined in Section A.1) we prove the following result.

Theorem 5. *The sequence* $v_k \overset{def}{=} v^{\mu_k,\nu_k}$ *contracts toward* v^* *with rate of* $1-\eta+\gamma\eta$*, i.e.,*

$$||v_k - v^*_\alpha|| \le (1-\eta+\gamma\eta)||v_{k-1} - v^*_\alpha|| \ .$$

As a corollary, and by plugging $\eta = 1$, we get the convergence rate of Zero-Sum MG PI. Notice that although the action space is continuous the proof follows using standard machinery, since the state space is still finite. We now give the proof of the theorem.

The proof has two steps. We first show $v^* \le v_{k+1} \le v_k$, where $v_k \overset{\text{def}}{=} v^{\mu_k,\nu_k}$. Building on this fact, we prove the contraction property by generalizing technique from (Scherrer, 2014)[Theorem 1], to two player game.

Lemma 6. $v^* \le v_{k+1} \le v_k$.

Proof. We have that $v_k = v^{\mu_k,\nu_k}$.

$$\begin{aligned}
v^{\mu_k,\nu_k} &= \bar{T}^{\nu_k} v^{\mu_k,\nu_k} \\
&= (1-\eta)\bar{T}^{\nu_k} v^{\mu_k,\nu_k} + \eta\bar{T}^{\nu_k} v^{\mu_k,\nu_k} \\
&\ge (1-\eta)\bar{T}^{\nu_k} v^{\mu_k,\nu_k} + \min_{\nu\in P_B} \eta\bar{T}^{\nu} v^{\mu_k,\nu_k} \\
&= (1-\eta)\bar{T}^{\nu_k} v^{\mu_k,\nu_k} + \eta\bar{T}^{\nu'} v^{\mu_k,\nu_k} \\
&= \max_{\mu\in P_A} ((1-\eta)T^{\mu,\nu_k} v^{\mu_k,\nu_k}) + \max_{\mu\in P_A} \left(\eta\bar{T}^{\mu,\nu'} v^{\mu_k,\nu_k}\right) \\
&\ge \max_{\mu\in P_A} \left((1-\eta)T^{\mu,\nu_k} v^{\mu_k,\nu_k} + \eta\bar{T}^{\mu,\nu'} v^{\mu_k,\nu_k}\right) \\
&= \max_{\mu\in P_A} T^{\mu,(1-\eta)\nu_k+\eta\nu'} v^{\mu_k,\nu_k} = \bar{T}^{(1-\eta)\nu_k+\eta\nu'} v^{\mu_k,\nu_k}.
\end{aligned} \tag{1}$$

The first relation holds due to Proposition 3, the forth relation holds by construction of ν', $\min_{\nu\in P_B} \bar{T}^\nu v^{\mu_k,\nu_k} = \bar{T}^{\nu'} v^{\mu_k,\nu_k}$, the fifth relation is by Definition 2, the sixth relation holds since sum of maximum elements is bigger than the maximum of a sum, and the seventh relation holds since the fixed-policy Bellman operator satisfies $T^{\mu,(1-\eta)\nu_1+\eta\nu_2} = (1-\eta)T^{\mu,\nu_1} + \eta T^{\mu,\nu_2}$.

Due to the monotonicity of $\bar{T}^{(1-\eta)\nu_k+\eta\nu'}$ (e.g, Patek (1997)[Appendix A]), we can repeatedly use (1),

$$v_k \geq \bar{T}^{(1-\eta)\nu_k+\eta\nu'} v_k \geq \cdots \geq \lim_{n\to\infty} (\bar{T}^{(1-\eta)\nu_k+\eta\nu'})^n v_k = v_{k+1},$$

where $v_{k+1} = v^{\mu_{k+1},\nu_{k+1}}$. Indeed, $\bar{T}^{(1-\eta)\nu_k+\eta\nu'}$ is the optimal Bellman operator given a fixed adversary strategy, $(1-\eta)\nu_k + \eta\nu'$.

Lastly, we show that in each iteration $v^* \leq v_k$. For any adversarial strategy ν_k,

$$v_k = \max_{\mu\in P_A} v^{\mu,\nu_k} \geq \min_{\nu\in P_B} \max_{\mu\in P_A} v^{\mu,\nu} = v^*.$$

Where the third relation holds by Proposition 4. □

We are now ready to prove Theorem 5.

Proof. As before, define $v_k \overset{\text{def}}{=} v^{\mu_k,\nu_k}$. We have that,

$$\begin{aligned} v^* - v_{k+1} &= v^* - T^{\mu_{k+1},(1-\eta)\nu_k+\eta\nu'} v_{k+1} \\ &\geq v^* - T^{\mu_{k+1},(1-\eta)\nu_k+\eta\nu'} v_k \\ &= (1-\eta)(v^* - T^{\mu_{k+1},\nu_k} v_k) + \eta(v^* - T^{\mu_{k+1},\nu'} v_k), \end{aligned} \tag{2}$$

where the first relation holds since $v_{k+1} = v^{\mu_{k+1},(1-\eta)\nu_k+\eta\nu'}$ and the second relation holds since $T^{\mu,\nu}$ is a monotone operator and $v_{k+1} \leq v_k$ by Lemma 6.

Consider the first term in (2).

$$v^* - T^{\mu_{k+1},\nu_k} v_k \geq v^* - T^{\mu_k,\nu_k} v_k = v_k. \tag{3}$$

The first relation holds since $T^{\mu_k,\nu_k} v_k = \max_{\mu\in P_A} T^{\mu,\nu_k} v_k$ and the second relation holds since by definition $v_k = v^{\mu_k,\nu_k} = T^{\mu_k,\nu_k} v^{\mu_k,\nu_k}$ (due to Proposition 3).

Remember that $\nu' \in \arg\min_{\nu\in P_B} \bar{T}^\nu v_k$ (as in the update of Alg. 2). Thus,

$$\bar{T}^{\nu'} v_k = \min_{\nu\in P_B} \bar{T}^\nu v_k = \min_{\nu\in P_B} \max_{\mu\in P_A} T^{\mu,\nu} v_k = \max_{\mu\in P_A} \min_{\nu\in P_B} T^{\mu,\nu} = T v_k, \tag{4}$$

where the third relation is due to Lemma 1.

Now, for the second term in (2) we have that,

$$\begin{aligned} v^* - T^{\mu_{k+1},\nu'} v_k &= T v^* - T^{\mu_{k+1},\nu'} v_k \\ &\geq T^{\mu^*,\nu^*} v^* - \max_{\mu\in P_A} T^{\mu,\nu'} v_k \\ &= T v^* - T v_k. \end{aligned} \tag{5}$$

The first relation holds since v^* is the fixed point of T, and the third relation holds by (4).

Plugging (3) and (5) to (2) yields,

$$v^* - v_{k+1} \geq (1-\eta)(v^* - v_k) + \eta(T v^* - T v_k).$$

Since $0 \geq v^* - v_{k+1}$ by Lemma 6, we can take the max-norm and conclude the proof,

$$\begin{aligned} ||v^* - v_{k+1}||_\infty &\leq (1-\eta)||v^* - v_k||_\infty + \eta||T v^* - T v_k||_\infty \\ &\leq (1-\eta)||v^* - v_k||_\infty + \eta\gamma||v^* - v_k||_\infty, \end{aligned}$$

where the first relation holds by the triangle inequality and the second holds since T is a γ-contraction by Proposition 2. □

B. Probabilistic Action Robust MDP

In this section, we focus on PR-MDPs (Section 3) and map the problem of solving the optimal probabilistic robust policy to solving a Zero-Sum MG. We then continue and provide the proofs of Section 3, which are mostly corollaries to the results in Section A.

For simplicity, we provide the definition of PR-MDPs as given in Section 3.

Definition 3. *Let $\alpha \in [0,1]$. A Probabilistic Action Robust MDP is defined by the 5-tuple of an MDP (see Section 2.1). Let $\pi, \bar{\pi}$ be policies of an agent an adversary. We define their probabilistic joint policy $\pi^{\text{mix}}_{P,\alpha}(\pi,\bar{\pi})$ as $\forall s \in \mathcal{S},\ \pi^{\text{mix}}_{P,\alpha}(\mathbf{a} \mid \mathbf{s}) \equiv (1-\alpha)\pi(\mathbf{a} \mid \mathbf{s}) + \alpha\bar{\pi}(\mathbf{a} \mid \mathbf{s})$.*

Let π be an agent policy. As opposed to standard MDPs, the value of the policy is defined by $v^{\pi}_{P,\alpha} = \min_{\bar{\pi}\in\Pi} \mathbb{E}^{\pi^{\text{mix}}_{P,\alpha}(\pi,\bar{\pi})}[\sum_t \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)]$, where $\mathbf{a}_t \sim \pi^{\text{mix}}_{P,\alpha}(\pi(\mathbf{s}_t), \bar{\pi}(\mathbf{s}_t))$. The optimal probabilistic robust policy is the optimal policy of the PR-MDP

$$\pi^*_{P,\alpha} \in \arg\max_{\pi\in\mathcal{P}(\Pi)} \min_{\bar{\pi}\in\Pi} \mathbb{E}^{\pi^{\text{mix}}_{P,\alpha}(\pi,\bar{\pi})}[\sum_t \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)]. \tag{6}$$

The optimal probabilistic robust value is $v^*_{P,\alpha} = v^{\pi^*_{P,\alpha}}_{P,\alpha}$.

B.1. Probabilistic Action Robust MDP as a Zero-Sum Markov Game

Consider the single agent MDP on which the PR-MDP is defined, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$.

Assumption 2.

- *$\mathcal{A}$ is compact metric space.*
- *For any $s \in \mathcal{S}$ the reward r is continuous and bounded function on $\mathcal{A}$.*
- *For any $s \in \mathcal{S}$, whenever $(a_n) \to (a)$, where $(a_n), (a) \in \mathcal{A}$, then $P(\cdot \mid s, a_n)$ converges weakly to $P(\cdot \mid s, a)$.*

Solving the optimal probabilistic robust policy can be equivalently viewed as solving a Zero-Sum MG $\mathcal{M}_{P,\alpha}$. Let $\mathcal{M}_{P,\alpha} = (\mathcal{S}, \mathcal{A}, \mathcal{A}, P_{P,\alpha}, R_{P,\alpha}, \gamma)$. Meaning, its state-space is equal to that of the original MDP, the action space of the two players is the action space of the original MDP, and its discount factor is equal to the discount factor of $\mathcal{M}$. Its reward and dynamics are given as follows,

$$r_{P,\alpha}(s,a,b) = (1-\alpha)r(s,a) + \alpha r(s,b),\ P_{P,\alpha}(s' \mid s,a,b) = (1-\alpha)P(s' \mid s,a) + \alpha P(s' \mid s,b). \tag{7}$$

By Assumption 2 on $\mathcal{M}$, Assumption 1 on the MG is satisfied.

It is easy to prove that a value $v^{\pi^{\text{mix}}_{P,\alpha}(\pi_1,\pi_2)}$ defined on $\mathcal{M}$ is equal to the value v^{π_1,π_2} defined on $\mathcal{M}_{P,\alpha}$. Since there is a one-to-one correspondence between the problems, solving the later is equivalent to solving the first.

B.2. Proof of Proposition 1

Consider the Zero-Sum MG $\mathcal{M}_{P,\alpha}$, and let P_A be the set of all probability measures on the Borel Sets of $\mathcal{A}$. We see that the Bellman operators of $\mathcal{M}_{P,\alpha}$ (Definition 2) decouples to two terms due to (7),

$$\begin{aligned} Tv &= \max_{\mu\in P_A} \min_{\nu\in P_A} r^{\mu,\nu} + \gamma P^{\mu,\nu} v \\ &= (1-\alpha)\left(\max_{\mu\in P_A} r^{\mu} + P^{\mu} v\right) + \alpha\left(\min_{\nu\in P_A} r^{\mu} + P^{\mu} v\right), \end{aligned} \tag{8}$$

and similarly for $T^{\mu}, \bar{T}^{\nu}$ and $T^{\mu,\nu}$.

According to Proposition 4 the the optimal policy for the $\max$-agent μ^* satisfies $v^* = Tv^* = T^{\mu^*}v^*$. Thus, μ^* should satisfy

$$\begin{aligned} &(1-\alpha)\left(\max_{\mu\in P_A} r^{\mu} + P^{\mu}v^*\right) + \alpha\left(\min_{\nu\in P_A} r^{\mu} + P^{\mu}v^*\right) = (1-\alpha)\left(r^{\mu^*} + P^{\mu^*}v^*\right) + \alpha\left(\min_{\nu\in P_A} r^{\mu} + P^{\mu}v^*\right) \\ \iff &\max_{\mu\in P_A} r^{\mu} + P^{\mu}v^* = r^{\mu^*} + P^{\mu^*}v^* \end{aligned}$$

meaning, $\mu^* \in \max_{\mu \in P_A} r^\mu + P^\mu v^*$ which can always be solved by a deterministic policy.

B.3. Probabilistic Action Robust and Robust MDPs

Based on the mapping between a PR-MDP to a corresponding Zero-Sum MG B.1 the relation to Robust MDPs becomes apparent. Instead for the adversary to pick an action which induces a change in the dynamics and reward 7, the adversary can directly choose the dynamics and reward. Obviously, the value of such a policy is similar under this equivalent view. We conclude the result since the adversary is defined on the class of stochastic policies $\mathcal{P}(\Pi)$.

B.4. Proof of Proposition 2

Repeating the same arguments as in Policy Gradient Theorem (Sutton et al., 2000)[Theorem 1] for continuous action space we have that for any $s \in \mathcal{S}$ and $\pi \in \mathcal{P}(\Pi)$, i.e., any stochastic stationary policy,

$$\nabla_\pi v^\pi(s) = \sum_s d^\pi(s) \int_{\mathbf{a} \in \mathcal{A}} \nabla_\pi \pi(s, \mathbf{a}) q^\pi(s, \mathbf{a}) d\,\mathbf{a}$$

Notice that we can replace the integration and differentiation order by Leibniz integral rule since $\nabla_\pi v^\pi(s)$ exists and is bounded. Let $h(\cdot \mid s)$ be a deterministic probability measure on A. Similarly to (Scherrer & Geist, 2014) for any $s \in \mathcal{S}$,

$$\begin{aligned}\langle \nabla_\pi v^\pi(s), h \rangle &= \sum_s d^\pi(s) \int_{\mathbf{a} \in \mathcal{A}} \langle \nabla_\pi \pi(s, \mathbf{a}), h \rangle q^\pi(s, \mathbf{a}) d\,\mathbf{a} \\ &= \sum_s d^\pi(s) q^\pi(s, h(s)).\end{aligned}$$

To minimize $\langle \nabla_\pi v^\pi(s), h \rangle$ we choose for any $s \in \mathcal{S}$, $\mathbf{a}_h \in \arg\min_a q^\pi(\cdot, a) = \arg\min_{\pi'} r^{\pi'} + \gamma P^{\pi'} v^\pi$.

B.5. Proof of Theorem 3

The theorem is a corollary of Theorem 5 and Proposition 2, while using the structure of the defined zero-sum MG for PR-MDP in Section B.1, $\mathcal{M}_{P,\alpha}$.

Specifically, the first stage of the general Soft Zero-Sum MG PI 2 is similar to the first stage of Soft Probabilistic Robust PI 2. Furthermore, for $\mathcal{M}_{P,\alpha}$ it holds for any bounded $v \in \mathbb{R}^{|\mathcal{S}|}$,

$$\begin{aligned}\arg\min_{\nu \in P_A} \bar{T}^\nu v &= \arg\min_{\nu \in P_A} \max_{\mu \in P_A} T^{\mu,\nu} v \\ &= \arg\min_{\nu \in P_A} \max_{\mu \in P_A} (1-\alpha)(r^\mu + \gamma P^\mu v) + \alpha(r^\nu + \gamma P^\nu v) \\ &= \arg\min_{\nu \in P_A} (r^\nu + \gamma P^\nu v),\end{aligned}$$

where the first relation holds by definition 2, the second relation holds due to the specific form of the Bellman operators similarly to (8), and the third relation holds since the first term does not depend on ν.

By using Proposition 2 we get that Soft Probabilistic Robust PI 2 is an instance of the more general Soft Zero-Sum MG PI 2, and prove the Theorem as a corollary of Theorem 5.

C. Noisy Action Robust MDP as a Zero-Sum Markov Game

We focus on NR-MDPs (Section 4) and map the problem of solving the optimal noisy robust policy to solving a Zero-Sum MG. As in previous section, the proofs of Section 4, are mostly corollaries to the results in Section A.

For simplicity, we provide the definition of NR-MDPs as given in Section 3.

Definition 2. *Let $\alpha \in [0, 1]$. A Noisy Action Robust MDP is defined by the 5-tuple of an MDP (see Section 2.1). Let $\pi, \bar{\pi}$ be policies of an agent and an adversary. We define their noisy joint policy $\pi^{\text{mix}}_{N,\alpha}(\pi, \bar{\pi})$ as*

$$\forall s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}, \pi^{\text{mix}}_{N,\alpha}(\mathbf{a} \mid \mathbf{s}) \equiv \mathbb{E}_{\substack{\mathbf{b} \sim \pi(\cdot|s) \\ \bar{\mathbf{b}} \sim \bar{\pi}(\cdot|s)}}[\mathbb{1}_{\mathbf{a} = (1-\alpha)\,\mathbf{b} + \alpha \bar{\mathbf{b}}}],$$

the relation is obtained by the fact that $\mathbf{a} \sim \pi, \bar{\mathbf{a}} \sim \bar{\pi}$.

Let π be an agent policy. For NR-MDP, its value is defined by $v^{\pi}_{N,\alpha} = \min_{\bar{\pi} \in \Pi} \mathbb{E}^{\pi^{\text{mix}}_{N,\alpha}(\pi,\bar{\pi})}[\sum_t \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)]$, where $\mathbf{a}_t \sim \pi^{\text{mix}}_{N,\alpha}(\pi(\mathbf{s}_t), \bar{\pi}(\mathbf{s}_t))$. The optimal α-noisy robust policy is the optimal policy of the NR-MDP

$$\pi^*_{N,\alpha} \in \arg\max_{\pi \in \mathcal{P}(\Pi)} \min_{\bar{\pi} \in \Pi} \mathbb{E}^{\pi^{\text{mix}}_{N,\alpha}(\pi,\bar{\pi})}[\sum_t \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)]. \tag{9}$$

The optimal noisy robust value is $v^*_{N,\alpha} = v^{\pi^*_{N,\alpha}}_{N,\alpha}$.

C.1. Noisy Action Robust MDP as a Zero-Sum Markov Game

Consider the single agent MDP on which the NR-MDP is defined, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and assume it satisfies Assumption 2. Solving the optimal probabilistic robust policy can be equivalently viewed as solving a Zero-Sum MG $\mathcal{M}_{N,\alpha}$. Let $\mathcal{M}_{N,\alpha} = (\mathcal{S}, \mathcal{A}, \mathcal{A}, P_{N,\alpha}, R_{N,\alpha}, \gamma)$. Meaning, its state-space is equal to that of the original MDP, the action space of the two players is the action space of the original MDP, and its discount factor is equal to the discount factor of $\mathcal{M}$. Its reward and dynamics are given as follows,

$$r_{N,\alpha}(s,a,b) = r(s, (1-\alpha)a + \alpha b),\ P_{P,\alpha}(s' \mid s,a,b) = P(s' \mid s, (1-\alpha)a + \alpha b). \tag{10}$$

Since the single agent MDP satisfies Assumption Assumption 2, the MG game defined by $\mathcal{M}_{N,\alpha}$ satisfies 1.

It is easy to prove that a value $v^{\pi^{\text{mix}}_{N}(\pi_1,\pi_2)}$ defined on the induced NR-MDP from $\mathcal{M}$ is equal to the value v^{π_1,π_2} defined on the MG $\mathcal{M}_{N,\alpha}$. Since there is a one-to-one correspondence between the problems, solving the later is equivalent to solving the first.

C.2. Proof of Proposition 4

Consider an MDP with a single state a quadratic reward of the form $r(a) = a^2$ where $a \in [-1,1]$. In this case, the solution does not depend on the horizon and an optimal action w.r.t. a single time step will be the solution for the discounted reward. Denote $\mathcal{P}([-1,1])$ as the set of all probability measures on the Borel sets of $[-1,1]$.

If both of the players are only allowed to take deterministic actions, then the min-max and max-min values are not equivalent,

$$\max_{a \in [-1,1]} \min_{b \in [-1,1]} ((1-\alpha)a + \alpha b)^2 = \begin{cases} (1-2\alpha)^2,\ \alpha \leq 0.5 \\ 0,\ \alpha > 0.5 \end{cases}$$
$$\min_{b \in [-1,1]} \max_{a \in [-1,1]} ((1-\alpha)a + \alpha b)^2 = (1-\alpha)^2.$$

Thus, for this example, strong duality on the sets of deterministic policies does not hold,

$$\max_{a \in [-1,1]} \min_{b \in [-1,1]} ((1-\alpha)a + \alpha b)^2 < \min_{b \in [-1,1]} \max_{a \in [-1,1]} ((1-\alpha)a + \alpha b)^2 = (1-\alpha)^2.$$

Furthermore, we now show that considering random policies can increase the value. Let the policy of the $\max$-player be $P(a=-1) = P(a=1) = 0.5$, obviously, $P \in \mathcal{P}([-1,1])$. For this policy, we have that,

$$\min_{b \in [-1,1]} \mathbb{E}_{a \sim P(\cdot)}[(1-\alpha)a + \alpha b)^2] = \min_{b \in [-1,1]} (1-\alpha)^2 + \alpha^2 b = (1-\alpha)^2.$$

We conclude that for this example

$$\max_{a \in [-1,1]} \min_{b \in [-1,1]} ((1-\alpha)a + \alpha b)^2 < \max_{P \in \mathcal{P}([-1,1])} \min_{b \in [-1,1]} \mathbb{E}_{a \sim P}[((1-\alpha)a + \alpha b)^2].$$

C.3. Policy Iteration of NR-MDP

We can use the Soft Zero-Sum MG PI (see Algorithm 2), or, by fixing $\eta = 1$, Zero-Sum MG PI.

The algorithm repeats two stages of (i) solving an MDP by fixing the adversary policy, (ii) solving a 1-step greedy minimax decision problem on the set of stochastic policies. This comes in contrast to the corresponding PI algorithm that solves PR-MDP. There, stage (ii) involved in solving a *single* agent, 1-step greedy, decision problem. This problem can be more easily solved by function maximization.

Furthermore, this fact suggest that a simple Frank-Wolfe update (Frank & Wolfe, 1956), as was performed in Soft Probabilistic Robust PI (Algorithm 2) would not work, at least not using the analysis we suggested here. Meaning, a relation between the maximal projection on the gradient $\nabla_\pi v^\pi$ and the 1-step greedy minimax decision problem, as shown to hold in Proposition 2, would not exists.

D. Actor Gradients Proof

Proof. Our proof follows the proof of the deterministic policy gradients (DPG) (Silver et al., 2014).

In order to retain consistency with (Silver et al., 2014), we denote the deterministic policy π by $\mu : S \mapsto A$. The parametrized policies μ_θ and $\bar{\mu}_{\bar{\theta}}$ are, respectively, the actor and adversary policies. We refer to the α-mixture policy $\pi^{\text{mix}}_{N/P,\alpha}(\mu_\theta, \bar{\mu}_{\bar{\theta}})$ simply as $\pi^{\text{mix}}_{N/P,\alpha}(\theta, \bar{\theta})$, for ease of notation.

Assumption 3. $p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}), \nabla_{\mathbf{a}} p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}), \mu_\theta(\mathbf{s}), \nabla_\theta \mu_\theta(\mathbf{s}), \bar{\mu}_{\bar{\theta}}(\mathbf{s}), \nabla_{\bar{\theta}} \bar{\mu}_{\bar{\theta}}(\mathbf{s}), r(\mathbf{s}, \mathbf{a}), \nabla_a r(\mathbf{s}, \mathbf{a}), p_1(\mathbf{s})$ *are continuous in all parameters and variables* $\mathbf{s}, \mathbf{a}, \mathbf{s}'$ *and* x.

Assumption 4. *There exists a* b *and* L *such that* $\sup_{\mathbf{s}} p_1(\mathbf{s}) < b, \sup_{\mathbf{a},\mathbf{s},\mathbf{s}'} p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) < b, \sup_{\mathbf{a},\mathbf{s}} r(\mathbf{s}, \mathbf{a}) < b, \sup_{\mathbf{a},\mathbf{s},\mathbf{s}'} ||\nabla_{\mathbf{a}} p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})|| < L$, *and* $\sup_{\mathbf{s},\mathbf{a}} ||\nabla_{\mathbf{a}} r(\mathbf{s}, \mathbf{a})|| < L$.

NR-MDP:

$$
\begin{aligned}
\nabla_\theta v^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})} &= \nabla_\theta Q^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}, \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})) \\
&= \nabla_\theta \left(r(\mathbf{s}, \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})) + \int_S \gamma p(\mathbf{s}' \mid \mathbf{s}, \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})) v^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') \right) d\,\mathbf{s}' \\
&= \nabla_\theta \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s}) \nabla_{\mathbf{a}} r(\mathbf{s}, \mathbf{a}) \mid_{\mathbf{a}=\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})} + \nabla_\theta \int_S \gamma p(\mathbf{s}' \mid \mathbf{s}, \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})) v^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \\
&= \nabla_\theta \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s}) \nabla_\theta r(\mathbf{s}, \mathbf{a}) \mid_{\mathbf{a}=\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})} \\
&\quad + \int_S \gamma \left(p(\mathbf{s}' \mid \mathbf{s}, \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})) \nabla_\theta v^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') + \nabla_\theta \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s}) \nabla_{\mathbf{a}} p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) \mid_{\mathbf{a}=\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})} v^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') \right) d\,\mathbf{s}' \\
&= \nabla_\theta \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s}) \nabla_{\mathbf{a}} \left(r(\mathbf{s}, \mathbf{a}) + \int_S \gamma p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) v^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \right) \mid_{\mathbf{a}=\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})} \\
&\quad + \int_S \gamma p(\mathbf{s}' \mid \mathbf{s}, \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})) \nabla_\theta v^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \\
&= \nabla_\theta \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}, \mathbf{a}) \mid_{\mathbf{a}=\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})(\mathbf{s})} + \int_S \gamma p(\mathbf{s} \to \mathbf{s}', 1, \pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})) \nabla_\theta v^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \quad .
\end{aligned}
$$

Where $p(\mathbf{s} \to \mathbf{s}', t, \pi)$ denotes the density at state $\mathbf{s}'$ after transitioning for t steps from state $\mathbf{s}$. Iterating this formula leads to the following result:

$$
\begin{aligned}
\nabla_\theta v^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)} &= \nabla_\theta \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s})} \\
&+ \int_S \gamma p(\mathbf{s} \to \mathbf{s}', 1, \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)) \nabla_\theta \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}') \nabla_{\mathbf{a}} Q^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s}',\mathbf{a}) \mid_{\mathbf{a}=\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}')} d\,\mathbf{s}' \\
&+ \int_S \gamma p(\mathbf{s} \to \mathbf{s}', 1, \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)) \int_S \gamma p(\mathbf{s}' \to \mathbf{s}'', 1, \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)) \nabla_\theta v^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s}'') d\,\mathbf{s}''\, d\,\mathbf{s}' \\
&= \nabla_\theta \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s})} \\
&+ \int_S \gamma p(\mathbf{s} \to \mathbf{s}', 1, \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)) \nabla_\theta \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}') \nabla_{\mathbf{a}} Q^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s}',\mathbf{a}) \mid_{\mathbf{a}=\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}')} d\,\mathbf{s}' \\
&+ \int_S \gamma^2 p(\mathbf{s} \to \mathbf{s}', 2, \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)) \nabla_\theta v^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s}') d\,\mathbf{s}' \\
&= \int_S \sum_{t=0}^{\infty} \gamma^t p(\mathbf{s} \to s', t, \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)) \nabla_\theta \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}') \nabla_{\mathbf{a}} Q^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s}',\mathbf{a}) \mid_{\mathbf{a}=\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}')} d\,\mathbf{s}' \quad .
\end{aligned}
$$

Taking the expectation over S_1:

$$
\begin{aligned}
\nabla_\theta J(\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)) &= \nabla_\theta \int_S p_1(\mathbf{s}) v^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s}) d\,\mathbf{s} \\
&= \int_S p_1(\mathbf{s}) \nabla_\theta v^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s}) d\,\mathbf{s} \\
&= \int_S \int_S \sum_{t=0}^{\infty} \gamma^t p_1(\mathbf{s}) p(\mathbf{s} \to \mathbf{s}', t, \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)) \nabla_\theta \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}') \nabla_{\mathbf{a}} Q^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s}',\mathbf{a}) \mid_{\mathbf{a}=\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}')} d\,\mathbf{s}'\, d\,\mathbf{s} \\
&= \int_S \rho^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)} \nabla_\theta \pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s})} d\,\mathbf{s} \\
&= \int_S \rho^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)} \nabla_\theta ((1-\alpha)\mu_\theta(\mathbf{s}) + \alpha \bar\mu_{\bar\theta}(\mathbf{s})) \nabla_{\mathbf{a}} Q^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s})} d\,\mathbf{s} \\
&= (1-\alpha) \int_S \rho^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)} \nabla_\theta \mu_\theta(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\pi^{\mathrm{mix}}_{N,\alpha}(\theta,\bar\theta)(\mathbf{s})} d\,\mathbf{s}
\end{aligned}
$$

notice that compared to the standard DPGs (Silver et al., 2014), the gradient is w.r.t. the actor's (adversary's) policy and is weighted by $1-\alpha$ (α). Similar to the DPG, the gradient of the action-value function is taken w.r.t. the action taken (the mixture policy).

PR-MDP: The PR-MDP, constructed by two deterministic policies μ_θ and $\bar\mu_{\bar\theta}$ can be defined as follows:

$$\pi^{\mathrm{mix}}_{P,\alpha}(u \mid s; \theta, \bar\theta) = (1-\alpha)\delta(u - \mu_\theta(s)) + \alpha\delta(u - \bar\mu_{\bar\theta}(s)).$$

$$v^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)} = \int_A \pi^{\mathrm{mix}}_{P,\alpha}(u \mid s; \theta, \bar\theta) Q^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)}(\mathbf{s}, \pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)(\mathbf{s})) du$$

$$
\begin{aligned}
\nabla_\theta v^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)} &= \nabla_\theta \int_A \pi^{\mathrm{mix}}_{P,\alpha}(u \mid s; \theta, \bar\theta) Q^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)}(\mathbf{s}, u) du \\
&= \nabla_\theta [(1-\alpha) Q^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)}(\mathbf{s}, \mu_\theta(\mathbf{s})) + \alpha Q^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)}(\mathbf{s}, \bar\mu_{\bar\theta}(\mathbf{s}))] \\
&= (1-\alpha) \nabla_\theta Q^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)}(\mathbf{s}, \mu_\theta(\mathbf{s})) + \alpha \nabla_\theta Q^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)}(\mathbf{s}, \bar\mu_{\bar\theta}(\mathbf{s}))
\end{aligned}
$$

we address each element, (1) $\nabla_\theta Q^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)}(\mathbf{s}, \mu_\theta(\mathbf{s}))$ and (2) $Q^{\pi^{\mathrm{mix}}_{P,\alpha}(\theta,\bar\theta)}(\mathbf{s}, \bar\mu_{\bar\theta}(\mathbf{s}))$, individually:

(1):

$$\begin{aligned}
\nabla_\theta Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\mu_\theta(\mathbf{s})) &= \nabla \left(r(\mathbf{s},\mu_\theta(\mathbf{s})) + \int_S \gamma p(\mathbf{s}' \mid \mathbf{s},\mu_\theta(\mathbf{s})) v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') \right) d\,\mathbf{s}' \\
&= \nabla_\theta \mu_\theta(\mathbf{s}) \nabla_{\mathbf{a}} r(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} + \nabla_\theta \int_S \gamma p(\mathbf{s}' \mid \mathbf{s},\mu_\theta(\mathbf{s})) v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \\
&= \nabla_\theta \mu_\theta(\mathbf{s}) \nabla_\theta r(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} \\
&\quad + \int_S \gamma \left(p(\mathbf{s}' \mid \mathbf{s},\mu_\theta(\mathbf{s})) \nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') + \nabla_\theta \mu_\theta(\mathbf{s}) \nabla_{\mathbf{a}} p(\mathbf{s}' \mid \mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') \right) d\,\mathbf{s}' \\
&= \nabla_\theta \mu_\theta(\mathbf{s}) \nabla_{\mathbf{a}} \left(r(\mathbf{s},\mathbf{a}) + \int_S \gamma p(\mathbf{s}' \mid \mathbf{s},\mathbf{a}) v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \right) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} \\
&\quad + \int_S \gamma p(\mathbf{s}' \mid \mathbf{s},\mu_\theta(\mathbf{s})) \nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \\
&= \nabla_\theta \mu_\theta(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} + \int_S \gamma p(\mathbf{s} \to \mathbf{s}', 1, \mu_\theta) \nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \quad .
\end{aligned}$$

Where $p(\mathbf{s} \to \mathbf{s}', t, \pi)$ denotes the density at state $\mathbf{s}'$ after transitioning for t steps from state s.

(2):

$$\begin{aligned}
\nabla_\theta Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\bar{\mu}_{\bar{\theta}}(\mathbf{s})) &= \nabla_\theta \left(r(\mathbf{s},\bar{\mu}_{\bar{\theta}}(\mathbf{s})) + \int_S \gamma p(\mathbf{s}' \mid \mathbf{s},\bar{\mu}_{\bar{\theta}}(\mathbf{s})) v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') \right) d\,\mathbf{s}' \\
&= \nabla_\theta \bar{\mu}_{\bar{\theta}}(\mathbf{s}) \nabla_{\mathbf{a}} r(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\bar{\mu}_{\bar{\theta}}(\mathbf{s})} + \nabla_\theta \int_S \gamma p(\mathbf{s}' \mid \mathbf{s},\bar{\mu}_{\bar{\theta}}(\mathbf{s})) v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \\
&= \int_S \gamma \left(p(\mathbf{s}' \mid \mathbf{s},\bar{\mu}_{\bar{\theta}}(\mathbf{s})) \nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') + \nabla_\theta \bar{\mu}_{\bar{\theta}}(\mathbf{s}) \nabla_{\mathbf{a}} p(\mathbf{s}' \mid \mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\bar{\mu}_{\bar{\theta}}(\mathbf{s})} v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') \right) d\,\mathbf{s}' \\
&= \int_S \gamma p(\mathbf{s}' \mid \mathbf{s},\bar{\mu}_{\bar{\theta}}(\mathbf{s})) \nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \quad .
\end{aligned}$$

Hence:

$$\begin{aligned}
\nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})} &= (1-\alpha) \nabla_\theta Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\mu_\theta(\mathbf{s})) + \alpha \nabla_\theta Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\bar{\mu}_{\bar{\theta}}(\mathbf{s})) \\
&= (1-\alpha) \nabla_\theta \mu_\theta(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} \\
&\quad + (1-\alpha) \int_S \gamma p(\mathbf{s} \to \mathbf{s}', 1, \mu_\theta) \nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' + \alpha \int_S \gamma p(\mathbf{s}' \mid \mathbf{s},\bar{\mu}_{\bar{\theta}}(\mathbf{s})) \nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}' \\
&= (1-\alpha) \nabla_\theta \mu_\theta(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} + \int_S \gamma p(\mathbf{s}' \mid \mathbf{s}, \pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})(\mathbf{s})) \nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}') d\,\mathbf{s}'
\end{aligned}$$

Applying this iteratively:

$$
\begin{aligned}
\nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})} &= (1-\alpha)\nabla_\theta\mu_\theta(\mathbf{s})\nabla_{\mathbf{a}}Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} \\
&+ \int_S \gamma p(\mathbf{s}' \mid \mathbf{s}, \pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})(\mathbf{s}))\nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}')d\,\mathbf{s}' \\
&= (1-\alpha)\nabla_\theta\mu_\theta(\mathbf{s})\nabla_{\mathbf{a}}Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} \\
&\quad + \int_S \gamma p(\mathbf{s}\to\mathbf{s}', 1, \pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta}))\nabla_\theta\mu_\theta(\mathbf{s}')\nabla_{\mathbf{a}}Q^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}',\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s}')} \, d\,\mathbf{s}' \\
&\quad + \int_S \gamma p(\mathbf{s}\to\mathbf{s}', 1, \pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})) \int_S \gamma p(\mathbf{s}'\to\mathbf{s}'', 1, \pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta}))\nabla_\theta v^{\pi^{\text{mix}}_{N,\alpha}(\theta,\bar{\theta})}(\mathbf{s}'')d\,\mathbf{s}''\,d\,\mathbf{s}' \\
&= (1-\alpha)\nabla_\theta\mu_\theta(\mathbf{s})\nabla_{\mathbf{a}}Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s})} \\
&\quad + (1-\alpha)\int_S \gamma p(\mathbf{s}\to\mathbf{s}', 1, \pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta}))\nabla_\theta\mu_\theta(\mathbf{s}')\nabla_{\mathbf{a}}Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}',\mathbf{a}) \mid_{\mathbf{a}=\mu_\theta(\mathbf{s}')} \, d\,\mathbf{s}' \\
&\quad + \int_S \gamma^2 p(\mathbf{s}\to\mathbf{s}', 2, \pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta}))\nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}')d\,\mathbf{s}' \\
&= (1-\alpha)\int_S \sum_{t=0}^{\infty} \gamma^t p(\mathbf{s}\to s', t, \pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta}))\nabla_\theta\mu_\theta(\mathbf{s}')\nabla_{\mathbf{a}}Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}',\mathbf{a}) \mid_{\mathbf{a}=\mu\theta(\mathbf{s}')} \, d\,\mathbf{s}' \quad .
\end{aligned}
$$

Taking the expectation over S_1:

$$
\begin{aligned}
\nabla_\theta J(\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})) &= \nabla_\theta \int_S p_1(\mathbf{s}) v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s})d\,\mathbf{s} \\
&= \int_S p_1(\mathbf{s})\nabla_\theta v^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s})d\,\mathbf{s} \\
&= (1-\alpha)\int_S\int_S \sum_{t=0}^{\infty}\gamma^t p_1(\mathbf{s}) p(\mathbf{s}\to\mathbf{s}', t, \pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta}))\nabla_\theta\pi_\theta(\mathbf{s}')\nabla_{\mathbf{a}}Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s}',\mathbf{a}) \mid_{\mathbf{a}=\pi_\theta(\mathbf{s}')} \, d\,\mathbf{s}'\,d\,\mathbf{s} \\
&= (1-\alpha)\int_S \rho^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}\nabla_\theta\pi_\theta(\mathbf{s})\nabla_{\mathbf{a}}Q^{\pi^{\text{mix}}_{P,\alpha}(\theta,\bar{\theta})}(\mathbf{s},\mathbf{a}) \mid_{\mathbf{a}=\pi_\theta(\mathbf{s})} \, d\,\mathbf{s}
\end{aligned}
$$

the resulting gradient update for the actor does not directly take into consideration the policy of the adversary, thus resulting in a gradient rule similar (weighted by $(1-\alpha)$ for the actor and α for the adversary) to that seen in Silver et al. (2014).

Intuitively, as the action is sampled w.p. $(1-\alpha)$ from the actor and w.p, α from the adversary, each player acts greedily at the immediate step ignoring potential perturbations. The mutual effect of the actor and adversary is attained through the Q value which captures the long term return of the mixture policy. □

Algorithm 3 Action-Robust DDPG

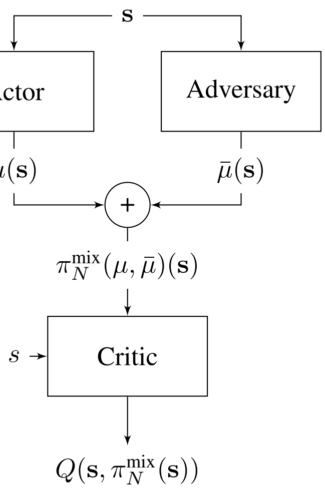
```
**Input:** Actor update steps ($N$), uncertainty value $\alpha$ and discount factor $\gamma$
Randomly initialize critic network $Q(\mathbf{s}, \mathbf{a}; \phi)$, actor $f(\mathbf{s}; \theta)$ and adversary $\bar{f}(\mathbf{s}; \bar{\theta})$
Initialize target networks with weights $\phi^-, \theta^-, \bar{\theta}^-$
Initialize replay buffer $R$
**for** episode in $0 \ldots M$ **do**
  Receive initial state $\mathbf{s}_0$
  **for** t in $0 \ldots T$ **do**
    Sample action $\mathbf{a}_t = \begin{cases} f(\mathbf{s}; \theta_\pi) \text{ w.p. } (1-\alpha) \text{ and } \bar{f}(s; \theta_{\bar{\pi}}) \text{ otherwise} & \text{, PR-MDP} \\ (1-\alpha) f(\mathbf{s}; \theta_\pi) + \alpha \bar{f}(\mathbf{s}; \bar{\theta}_{\bar{\pi}}) & \text{, NR-MDP} \end{cases}$
    $\tilde{\mathbf{a}}_t = \mathbf{a}_t$ + exploration noise
    Execute action $\tilde{\mathbf{a}}_t$ and observe reward $r_t$ and new state $s_{t+1}$
    Store transition $(\mathbf{s}_t, \tilde{\mathbf{a}}_t, r_t, \mathbf{s}_{t+1})$ in $R$
    **for** i in $0 \ldots N$ **do**
      Sample batch from replay buffer
      Update actor:
      $\theta \leftarrow \begin{cases} \nabla_\theta (1-\alpha) Q(\mathbf{s}, f(\mathbf{s}; \theta)) & \text{, PR-MDP} \\ \nabla_\theta Q(\mathbf{s}, (1-\alpha) f(\mathbf{s}; \theta) + \alpha \bar{f}(\mathbf{s}; \bar{\theta})) & \text{, NR-MDP} \end{cases}$
      Update critic:
      $\phi \leftarrow \begin{cases} \nabla_\phi ||r + \gamma[(1-\alpha) Q(\mathbf{s}', f(\mathbf{s}'; \theta^-)) + \alpha Q(\mathbf{s}', f(\mathbf{s}'; \bar{\theta}^-))]||_2^2 & \text{, PR-MDP} \\ \nabla_\phi ||r + \gamma[Q(\mathbf{s}', (1-\alpha) f(\mathbf{s}'; \theta^-) + \alpha f(\mathbf{s}'; \bar{\theta}^-))]||_2^2 & \text{, NR-MDP} \end{cases}$
    **end for**
    Sample batch from replay buffer
    Update adversary:
    $\bar{\theta} \leftarrow \begin{cases} \nabla_{\bar{\theta}} \alpha Q(\mathbf{s}, \bar{f}(\mathbf{s}; \bar{\theta})) & \text{, PR-MDP} \\ \nabla_{\bar{\theta}} Q(\mathbf{s}, (1-\alpha) f(\mathbf{s}; \theta) + \alpha \bar{f}(\mathbf{s}; \bar{\theta})) & \text{, NR-MDP} \end{cases}$
    Update critic
    Update the target networks:
        $\theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$
        $\bar{\theta}^- \leftarrow \tau\bar{\theta} + (1-\tau)\bar{\theta}^-$
        $\phi^- \leftarrow \tau\phi + (1-\tau)\phi^-$
  **end for**
**end for**
```

Algorithm 3 presents our Action Robust approach adapted to the DDPG algorithm (Lillicrap et al., 2015). The action we play during exploration is based on the exploration scheme selected, OU noise adds noise at the action level whereas in parameter space noise we pertube the parameters θ and $\bar{\theta}$.

Notice that the critic update is different, in both scenarios, from the default DDPG update rule. The reason is that the critic is updated based on the expectation over the policy, which in the NR-MDP results in the α mixture policy and in the PR-MDP a convex sum of Q values.

Figure 1 presents a block diagram of our approach for the NR-MDP scenario:

Figure 1. Action Robust DDPG, NR-MDP



s
Actor
Adversary
$\mu(\mathbf{s})$
$\bar{\mu}(\mathbf{s})$
+
$\pi_N^{\text{mix}}(\mu, \bar{\mu})(\mathbf{s})$
s
Critic
$Q(\mathbf{s}, \pi_N^{\text{mix}}(\mathbf{s}))$

We improve the actor (adversary) by taking the gradient of Q w.r.t. $\theta(\bar{\theta})$ and performing backpropagation through the critic. Autograd engines (Baydin et al., 2018) automatically ensure that the gradients propagate directly to the actor (adversary) without affecting the adversary (actor) or the critic. During exploration we simply play π_N^{mix} a deterministic policy (as it is a convex sum of two deterministic values).

For the PR-MDP the schema is similar to the standard DDPG approach.

Figure 2. Action Robust DDPG, PR-MDP

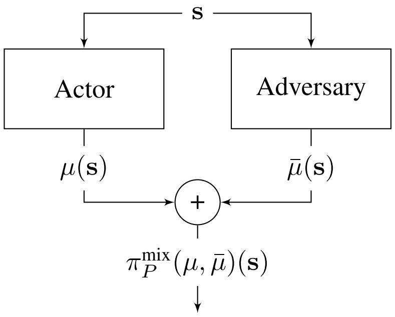


Figure 2 depicts the block diagram during exploration. π_P^{mix} defines a stochastic policy over μ and $\bar{\mu}$. Thus, with probability $1 - \alpha$ we sample action $\mu(\mathbf{s})$ and otherwise $\bar{\mu}(\mathbf{s})$.

Figure 3. Action Robust DDPG, PR-MDP

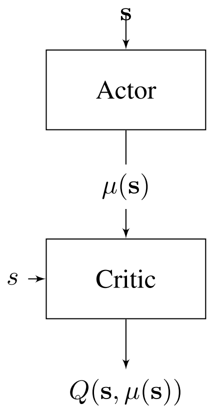


Figure 3 presents the approach during training. This approach is identical to the standard DDPG approach, except that once taking the gradient $\nabla_\theta Q(s, \mu_\theta(\mathbf{s}))$, we multiply the loss (similar to a change of learning rate) by $1 - \alpha$.

The critic is trained on the expectation over the mixture policies, which in the case of DDPG results in $Q(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + \gamma[(1 - \alpha)Q(\mathbf{s}', \mu(\mathbf{s}')) + \alpha Q(\mathbf{s}', \bar{\mu}(\mathbf{s}'))]$.

E. Empirical Results

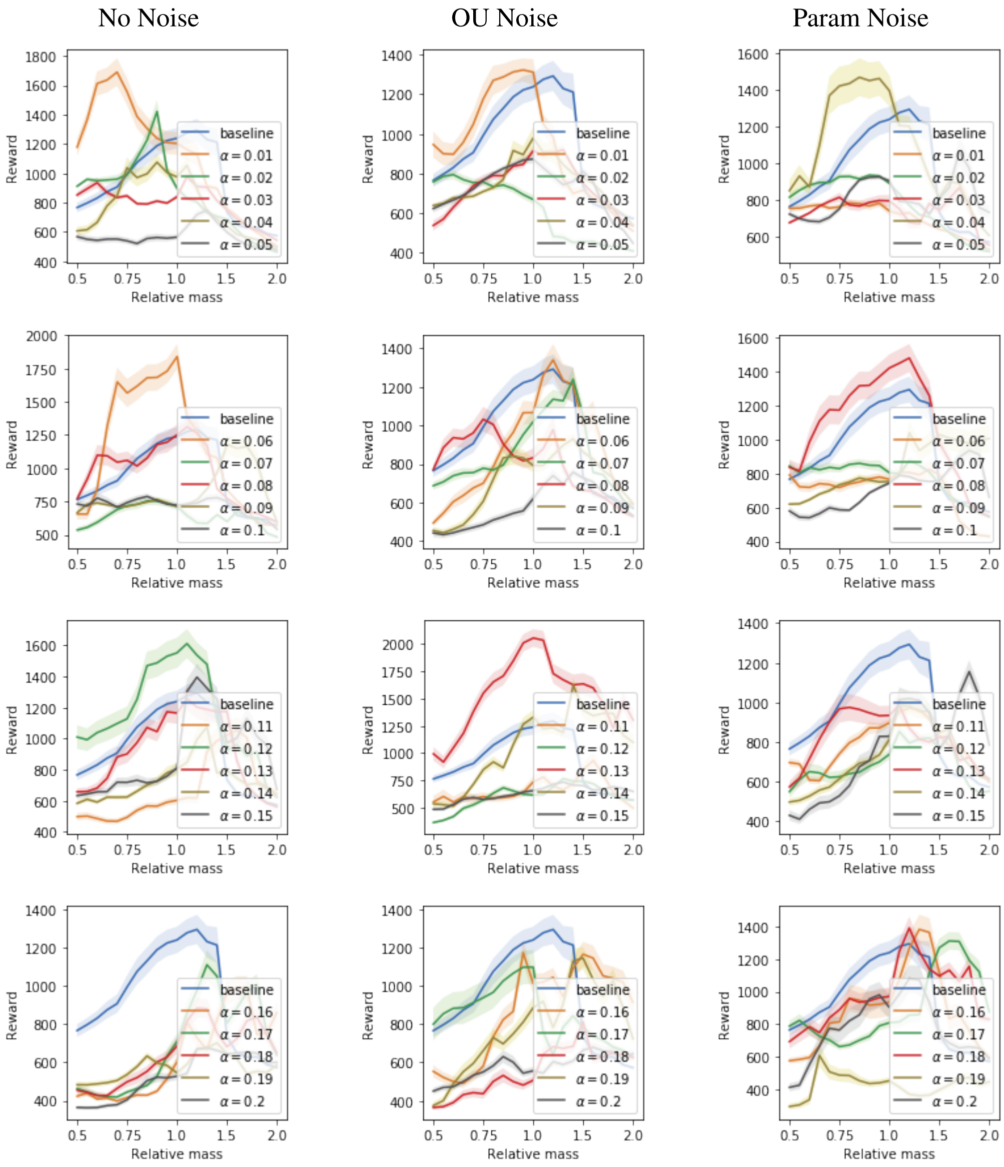


Figure 4. NR-MDP: exploration and α ablation.

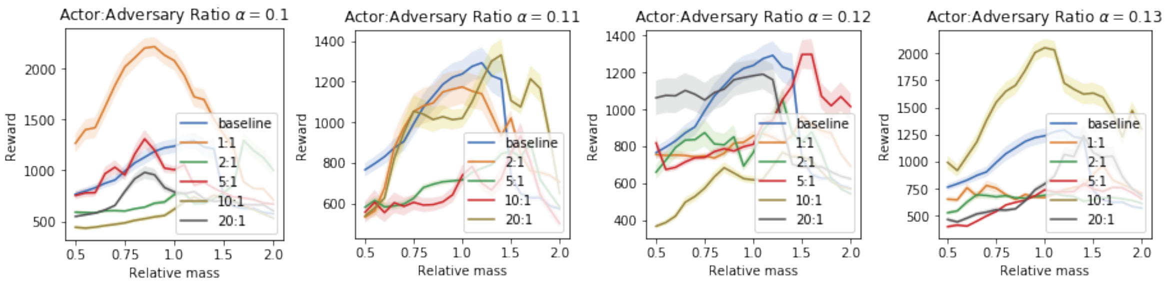


Figure 5. NR-MDP: α and training ratio ablation.

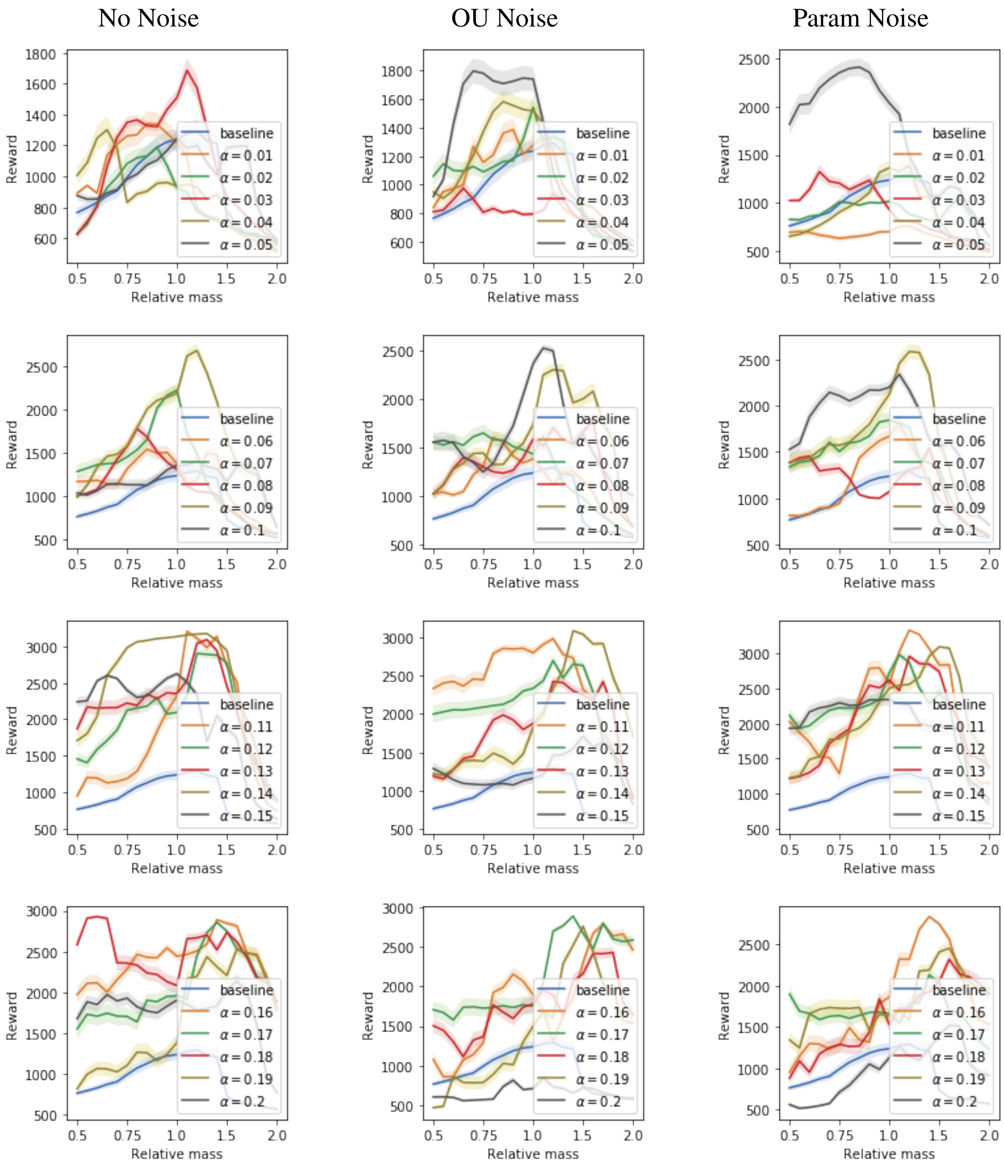


Figure 6. PR-MDP: exploration and α ablation.

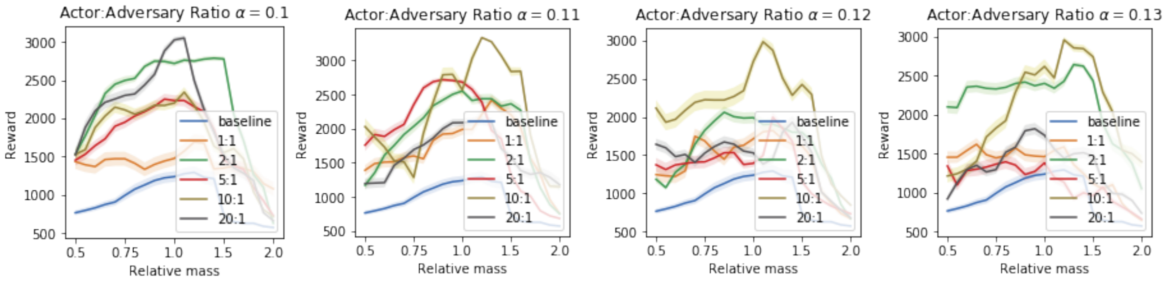


Figure 7. PR-MDP: α and training ratio ablation.

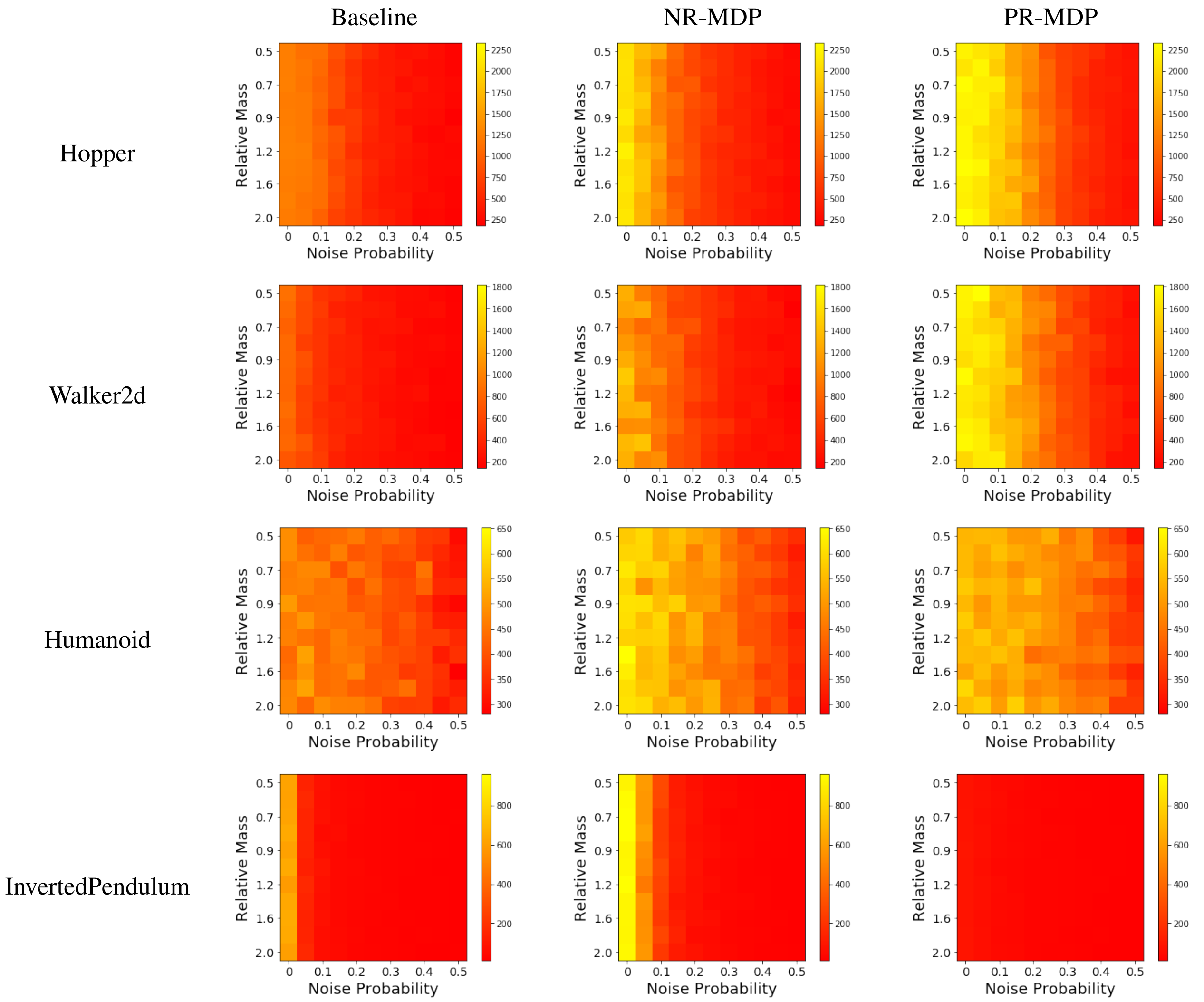


Figure 8. Robustness to model uncertainty. Noise probability denotes the probability of a randomly sampled noise being played instead of the selected action.

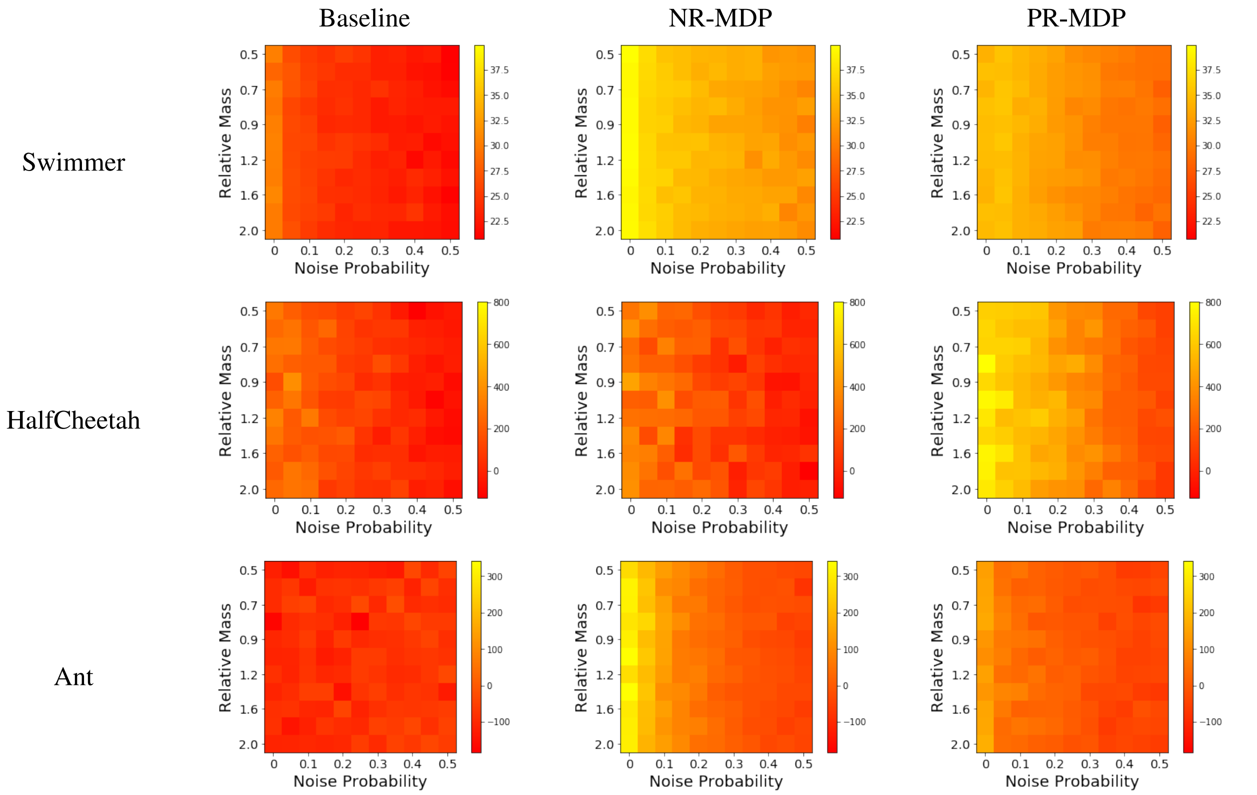


Figure 9. Robustness to model uncertainty continued. Noise probability denotes the probability of a randomly sampled noise being played instead of the selected action.

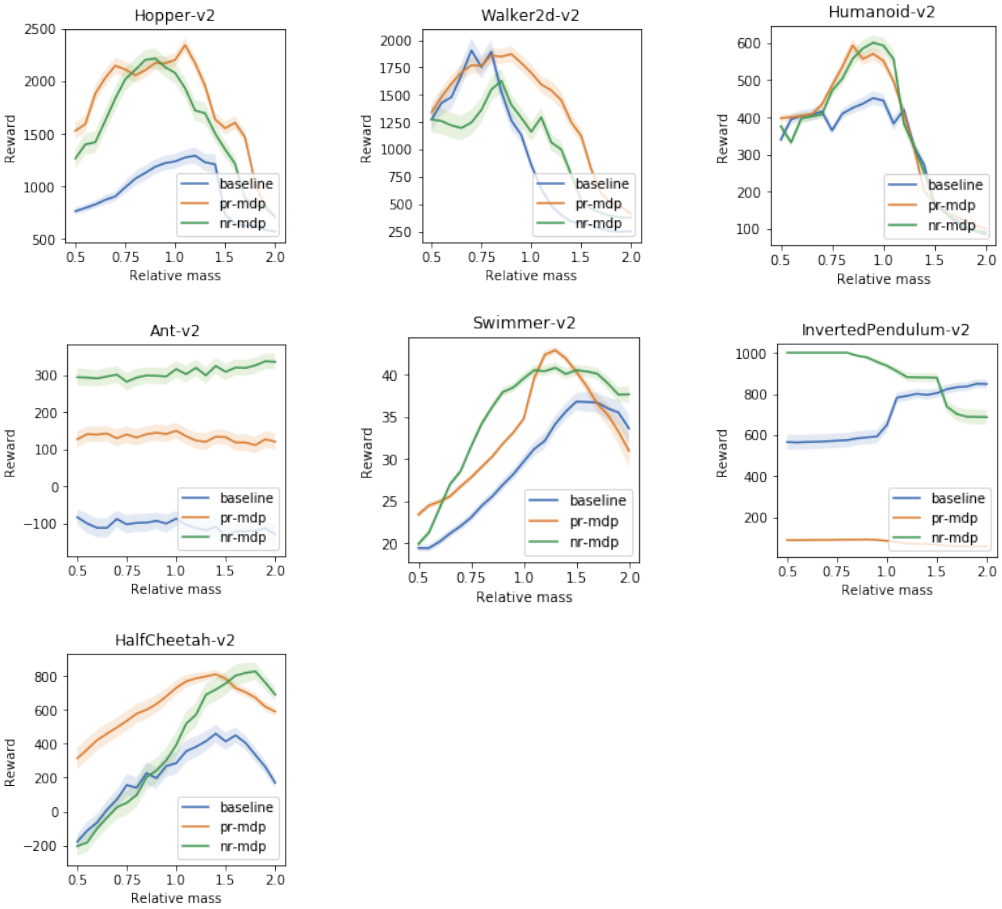


Figure 10. Robustness to mass uncertainty.

References

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. Automatic differentiation in machine learning: a survey. *Journal of Marchine Learning Research*, 18:1–43, 2018.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Hansen, T. D., Miltersen, P. B., and Zwick, U. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1, 2013.

Hoffman, A. J. and Karp, R. M. On nonterminating stochastic games. *Management Science*, 12(5):359–370, 1966.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pp. 267–274, 2002.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Maitra, A. and Parthasarathy, T. On stochastic games. *Journal of Optimization Theory and Applications*, 5(4):289–300, 1970.

Patek, S. D. *Stochastic and shortest path games: theory and algorithms*. PhD thesis, Massachusetts Institute of Technology, 1997.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

Rao, S., Chandrasekaran, R., and Nair, K. Algorithms for discounted stochastic games. *Journal of Optimization Theory and Applications*, 11(6):627–637, 1973.

Scherrer, B. Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, pp. 1314–1322, 2014.

Scherrer, B. and Geist, M. Local policy search in a convex space and conservative policy iteration as boosted policy search. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2014.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *ICML*, 2014.

Sion, M. et al. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.