Appendix: Soft-Robust Actor-Critic Policy-Gradient

A Proofs

A.1 Proposition 3.1

Proof. Fix $x, y \in \mathcal{X}$. For any policy π, we denote by $p(x, y)$ the probability of getting from state x to state y, which can be written as $\mathbb{E}_{a\sim\pi(x)}[p(x, a, y)]$. Since ω is non-diffuse, there exists p_0 such that $\omega(p_0) > 0$. Also, by Assumption 3.1, there exists an integer n such that $p_0^n(x, y) > 0$. We thus have

$$\begin{aligned}
\bar{p}^n(x, y) &= \Big(E_{p\sim\omega}[p]\Big)^n(x, y)\\
\bar{p}^n(x, y) &\geq \Big(p_0\omega(p_0)\Big)^n(x, y)\\
\bar{p}^n(x, y) &\geq p_0^n(x, y)\omega(p_0)^n > 0
\end{aligned}$$

which shows that $\bar{p}$ is irreducible. Using the same reasoning, we show $\{n \in \mathbb{N} : p_0^n(x, x) > 0\} \subset \{n \in \mathbb{N} : \bar{p}^n(x, x) > 0\}$ and then use the fact that p_0 is aperiodic to conclude that $\bar{p}$ is aperiodic too. □

A.2 Proposition 3.2

This recursive equation comes from the same reasoning as in Lemma 3.1 of Xu and Mannor [2012]. We apply it to the average reward criterion.

Proof. For every $p \in \mathcal{P}$, we can apply the Poisson equation to the corresponding model:

$$J_p(\pi) + V_p^\pi(x) = \sum_{a\in\mathcal{A}} \pi(x, a)\left(r(x, a) + \sum_{x'\in\mathcal{X}} p(x, a, x')V_p^\pi(x')\right)$$

By integrating with respect to ω we obtain:

$$\bar{J}(\pi) + \bar{V}^\pi(x) = \sum_{a\in\mathcal{A}} \pi(x, a)\left(r(x, a) + \sum_{x'\in\mathcal{X}} E_{p\sim\omega}[p(x, a, x')V_p^\pi(x')]\right)$$

We then use the statewise independence assumption on ω to make the recursion explicit. We thus have

$$\begin{aligned}
\bar{J}(\pi) + \bar{V}^\pi(x) &\overset{(1)}{=} \sum_{a\in A} \pi(x, a)\left(r(x, a) + \sum_{x'\in X} \int p(x, a, x')V_p^\pi(x')d\omega_x(p_x)d\omega_{x'}(p_{x'})\right)\\
&\overset{(2)}{=} \sum_{a\in A} \pi(x, a)\left(r(x, a) + \sum_{x'\in X} E_{p_x\sim\omega_x}[p(x, a, x')]E_{p_{x'}\sim\omega_{x'}}[V_p^\pi(x')]\right)\\
&= \sum_{a\in A} \pi(x, a)\left(r(x, a) + \sum_{x'\in X} \bar{p}(x, a, x')\bar{V}^\pi(x')\right),
\end{aligned}$$

where (1) results from the rectangularity assumption on ω. (2) Since $p(x, a, x')$ is an element of vector p_x that only depends on the uncertainty set at state x and $V_p^\pi(x')$ depends on the uncertainty set at state x', we can split the integrals. We slightly abuse notation here because a state can be visited multiple times. In fact, we implicitly introduce dummy states and treat multiple visits to a state as visiting different states. More explicitely, we write ω as $\omega = \bigotimes_{t=0}^{+\infty} \omega_{x,t}$ where $\omega_{x,t} = \omega_x$, ω_x being the distribution at state x. This representation is termed as the *stationary model* in Xu and Mannor [2012]. □

A.3 Corollary 3.1

Proof. According to Proposition 3.2 and summing up both sides of the equality with respect to the stationary distribution $\bar{d}^\pi$, we have

$$\begin{aligned}\bar{J}(\pi) + \sum_{x\in\mathcal{X}} \bar{d}^\pi(x)\bar{V}^\pi(x) &= \sum_{x\in\mathcal{X}} \bar{d}^\pi(x) \sum_{a\in\mathcal{A}} \pi(x,a) \left(r(x,a) + \sum_{x'\in\mathcal{X}} E_{p_x\sim\omega_x}[p(x,a,x')]\bar{V}^\pi(x') \right) \\ &= \sum_{x\in\mathcal{X}} \bar{d}^\pi(x) \sum_{a\in\mathcal{A}} \pi(x,a) \left(r(x,a) + \sum_{x'\in\mathcal{X}} \bar{p}(x,a,x')\bar{V}^\pi(x') \right)\end{aligned}$$

Since $\bar{d}^\pi$ is stationary with respect to $\bar{p}$, we can then write

$$\bar{J}(\pi) + \sum_{x\in\mathcal{X}} \bar{d}^\pi(x)\bar{V}^\pi(x) = \sum_{x\in\mathcal{X}} \bar{d}^\pi(x) \sum_{a\in\mathcal{A}} \pi(x,a) r(x,a) + \sum_{x'\in\mathcal{X}} \bar{d}^\pi(x')\bar{V}^\pi(x').$$

It remains to simplify both sides of the equality in order to get the result. □

A.4 Theorem 4.1

We use the same technique as in Mankowitz et al. [2018]; Sutton et al. [2000] in order to prove a soft-robust version of policy-gradient theorem.

Proof.

$$\begin{aligned}\nabla_\theta \bar{V}^\pi(x) &= \nabla_\theta \sum_{a\in\mathcal{A}} \pi(x,a)\bar{Q}^\pi(x,a) \\ \nabla_\theta \bar{V}^\pi(x) &= \sum_{a\in\mathcal{A}} \Big[\nabla_\theta \pi(x,a)\bar{Q}^\pi(x,a) + \pi(x,a)\nabla_\theta \bar{Q}^\pi(x,a) \Big] \\ \nabla_\theta \bar{V}^\pi(x) &\overset{(1)}{=} \sum_{a\in\mathcal{A}} \Big[\nabla_\theta \pi(x,a)\bar{Q}^\pi(x,a) + \pi(x,a)\nabla_\theta \Big[r(x,a) - \bar{J}(\pi) + \sum_{x'\in\mathcal{X}} \bar{p}(x,a,x')\bar{V}^\pi(x') \Big] \Big] \\ \nabla_\theta \bar{V}^\pi(x) &= \sum_{a\in\mathcal{A}} \Big[\nabla_\theta \pi(x,a)\bar{Q}^\pi(x,a) + \pi(x,a) \Big[-\nabla_\theta \bar{J}(\pi) + \sum_{x'\in\mathcal{X}} \bar{p}(x,a,x')\nabla_\theta \bar{V}^\pi(x') \Big] \Big] \\ \nabla_\theta \bar{J}(\pi) &= \sum_{a\in\mathcal{A}} \Big[\nabla_\theta \pi(x,a)\bar{Q}^\pi(x,a) + \pi(x,a) \Big[\sum_{x'\in\mathcal{X}} \bar{p}(x,a,x')\nabla_\theta \bar{V}^\pi(x') \Big] \Big] - \nabla_\theta \bar{V}^\pi(x) \\ \sum_{x\in\mathcal{X}} \bar{d}^\pi(x)\nabla_\theta \bar{J}(\pi) &\overset{(2)}{=} \sum_{x\in\mathcal{X}} \bar{d}^\pi(x) \sum_{a\in\mathcal{A}} \Big[\nabla_\theta \pi(x,a)\bar{Q}^\pi(x,a) + \sum_{a\in\mathcal{A}} \pi(x,a) \sum_{x'\in\mathcal{X}} \bar{p}(x,a,x')\nabla_\theta \bar{V}^\pi(x') \Big] - \sum_{x\in\mathcal{X}} \bar{d}^\pi(x)\nabla_\theta \bar{V}^\pi(x) \\ \sum_{x\in\mathcal{X}} \bar{d}^\pi(x)\nabla_\theta \bar{J}(\pi) &= \sum_{x\in\mathcal{X}} \bar{d}^\pi(x) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(x,a)\bar{Q}^\pi(x,a) + \sum_{x\in\mathcal{X}} \bar{d}^\pi(x) \sum_{a\in\mathcal{A}} \pi(x,a) \sum_{x'\in\mathcal{X}} \bar{p}(x,a,x')\nabla_\theta \bar{V}^\pi(x') \\ &\quad - \sum_{x\in\mathcal{X}} \bar{d}^\pi(x)\nabla_\theta \bar{V}^\pi(x) \\ \sum_{x\in\mathcal{X}} \bar{d}^\pi(x)\nabla_\theta \bar{J}(\pi) &\overset{(3)}{=} \sum_{x\in\mathcal{X}} \bar{d}^\pi(x) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(x,a)\bar{Q}^\pi(x,a) + \sum_{x'\in\mathcal{X}} \bar{d}^\pi(x')\nabla_\theta \bar{V}(x') - \sum_{x\in\mathcal{X}} \bar{d}^\pi(x)\nabla_\theta \bar{V}^\pi(x) \\ \nabla_\theta \bar{J}(\pi) &= \sum_{a\in\mathcal{A}} \nabla_\theta \pi(x,a)\bar{Q}^\pi(x,a)\end{aligned}$$

where (1) occurs thanks to the soft-robust Poisson equation. (2) Multiply both sides of the Equation by $\sum_{x\in\mathcal{X}} \bar{d}^\pi(x)$. (3) Since $\bar{d}^\pi(x)$ is stationary with respect to $\bar{p}$, we have that $\sum_{x\in\mathcal{X}} \bar{d}^\pi(x) \sum_{a\in\mathcal{A}} \pi(x,a)\bar{p}(x,a,x') = \sum_{x'\in\mathcal{X}} \bar{d}^\pi(x')$. □

A.5 Theorem 4.2

Proof. Recall the mean squared error:

$$\mathcal{E}^{\pi}(w) := \sum_{x \in \mathcal{X}} \bar{d}^{\pi}(x) \sum_{a \in \mathcal{A}} \pi(x,a) \Big[\bar{Q}^{\pi}(x,a) - f_w(x,a) \Big]^2$$

with respect to the soft-robust state distribution $\bar{d}^{\pi}(x)$. If we derive this distribution with respect to the parameters w and analyze it when the process has converged to a local optimum as in Sutton et al. [2000], then we get:

$$\sum_{x \in \mathcal{X}} \bar{d}^{\pi}(x) \sum_{a \in \mathcal{A}} \pi(x,a) \Big[\bar{Q}^{\pi}(x,a) - f_w(x,a) \Big] \nabla_w f_w(x,a) = 0$$

Additionally, the compatibility condition $\nabla_w f_w(x,a) = \nabla_\theta \log \pi(x,a)$ yields:

$$\begin{aligned} \sum_{x \in \mathcal{X}} \bar{d}^{\pi}(x) \sum_{a \in \mathcal{A}} \pi(x,a) \Big[\bar{Q}^{\pi}(x,a) - f_w(x,a) \Big] \nabla_\theta \pi(x,a) \frac{1}{\pi(x,a)} &= 0 \\ \sum_{x \in \mathcal{X}} \bar{d}^{\pi}(x) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(x,a) \Big[\bar{Q}^{\pi}(x,a) - f_w(x,a) \Big] &= 0 \end{aligned}$$

Subtract this quantity from the soft-robust policy gradient (Theorem 4.1). We then have:

$$\begin{aligned} \nabla_\theta \bar{J}(\pi) &= \sum_{x \in \mathcal{X}} \bar{d}^{\pi}(x) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(x,a) \bar{Q}^{\pi}(x,a) - \sum_{x \in \mathcal{X}} \bar{d}^{\pi}(x) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(x,a) \Big[\bar{Q}^{\pi}(x,a) - f_w(x,a) \Big] \\ &= \sum_{x \in \mathcal{X}} \bar{d}^{\pi}(x) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(x,a) f_w(x,a). \end{aligned}$$

□

A.6 Convergence Proof for SR-AC

We define as *soft-robust TD-error* at time t the following random quantity:

$$\delta_t := r_{t+1} - \hat{J}_{t+1} + \sum_{x' \in \mathcal{X}} \bar{p}(x_t, a_t, x') \hat{V}_{x'} - \hat{V}_{x_t}$$

where $\hat{V}_{x_t}$ and $\hat{J}_t$ are unbiased estimates that satisfy $E[\hat{V}_{x_t} \mid x_t, \pi] = \bar{V}^{\pi}(x_t)$ and $E[\hat{J}_{t+1} \mid x_t, \pi] = \bar{J}(\pi)$ respectively. We can easily show that this defines an unbiased estimate of the soft-robust advantage function $\bar{A}^{\pi}$ [Bhatnagar et al., 2009]. Thus, using equation (1), an unbiased estimate of the gradient $\nabla_\theta \bar{J}(\pi)$ can be obtained by taking

$$\widehat{\nabla_\theta J(\pi)} := \delta_t \psi_{x_t a_t}.$$

Similarly, recall the *soft-robust TD-error* with linear function approximation at time t:

$$\delta_t := r_{t+1} - \hat{J}_{t+1} + \sum_{x' \in \mathcal{X}} \bar{p}(x_t, a_t, x') v_t^T \varphi_{x'} - v_t^T \varphi_{x_t},$$

where v_t corresponds to the current estimate of the soft-robust value function parameter.

As in regular MDPs, when doing linear TD learning, the function approximation of the value function introduces a bias in the gradient estimate Bhatnagar et al. [2009].

Define the quantity

$$\widetilde{V}^{\pi}(x) = \sum_{a \in \mathcal{A}} \pi(x,a) \Big[r(x,a) - \bar{J}(\pi) + \sum_{x' \in \mathcal{X}} \bar{p}(x,a,x') v_\pi^T \varphi_{x'} \Big]$$

where $v_\pi^T \varphi_{x'}$ is an estimate of the value function upon convergence of a TD recursion, that is $v_\pi = \lim_{t\to\infty} v_t$. Also, define as δ_t^π the associated error upon convergence:

$$\delta_t^\pi := r_{t+1} - \hat{J}_{t+1} + \sum_{x' \in \mathcal{X}} \bar{p}(x_t, a_t, x') v_\pi^T \varphi_{x'} - v_\pi^T \varphi_{x_t}.$$

Similarly to Lemma 4 of Bhatnagar et al. [2009], the bias of the soft-robust gradient estimate is given by

$$e^\pi := \sum_{x \in \mathcal{X}} \bar{d}^\pi(x) \left[\nabla_\theta \widetilde{V}^\pi(x) - \nabla_\theta v_\pi^T \varphi_x \right],$$

that is $E[\widehat{\nabla_\theta J}(\pi) \mid \theta] = \nabla_\theta \bar{J}(\pi) + e^\pi$. This error term then needs to be small enough in order to ensure convergence of the algorithm.

Let denote as $\bar{V}(v) := \Phi v$ the linear approximation to the soft-robust differential value function defined earlier, where $\Phi \in \mathbb{R}^{n \times d_2}$ is a matrix and each feature vector $\varphi_x(k)$ corresponds to the k^{th} column in Φ. We make the following assumption:

Assumption A.1. *The basis functions $\varphi_x \in \mathbb{R}^{d_2}$ are linearly independent. In particular, Φ has full rank. We also have $\Phi v \neq e$ for all value function parameters $v \in \mathbb{R}^{d_2}$ where e is a vector of all ones.*

The learning rates α_t and β_t (Lines 7 and 8 in Algorithm 1) are established such that $\alpha_t \to 0$ slower than $\beta_t \to 0$ as $t \to \infty$. In addition, $\sum_t \alpha_t = \sum_t \beta_t = \infty$ and $\sum_t \alpha_t^2, \sum_t \beta_t^2 < \infty$. We also set the soft-robust average reward step-size $\xi_t = c\alpha_t$ for a positive constant c. The soft-robust average reward, TD-error and critic will all operate on the faster timescale α_t and therefore converge faster. Eventually, define a diagonal matrix D where the steady-state distribution $\bar{d}^\pi$ forms the diagonal of this matrix. We write the soft-robust transition probability matrix as:

$$\bar{P}^\pi(x, x') = \sum_{a \in \mathcal{A}} \pi(x, a) \bar{p}(x, a, x'),$$

where $x, x' \in \mathcal{X}$ and $\bar{p}$ designates the average transition model. By denoting $R^\pi \in \mathbb{R}^n$ the column vector of rewards $(\sum_{a\in\mathcal{A}} \pi(x_1, a) r(x_1, a), \cdots \sum_{a\in\mathcal{A}} \pi(x_n, a) r(x_n, a))^T$ where $(x_1, \ldots, x_n)$ is a numbered representation of the state-space and using the following operator $T : \mathbb{R}^n \to \mathbb{R}^n$, we can express the soft-robust Poisson equation as:

$$T(J) = R^\pi - \bar{J}(\pi) e + \bar{P}^\pi J$$

The soft-robust average reward iterates (Line 5) and the critic iterates (Line 7) defined in Algorithm 1 converge almost surely, as stated in the following Lemma which is a straightforward application of Lemma 5 from Bhatnagar et al. [2009].

Lemma A.1. *For any given policy π and $\{\hat{J}_t\}, \{v_t\}$ as in the soft-robust average reward and critic updates, we have $\hat{J}_t \to \bar{J}(\pi)$ and $v_t \to v^\pi$ almost surely, where*

$$\bar{J}(\pi) = \sum_{x \in \mathcal{X}} \bar{d}^\pi(x) \sum_{a \in \mathcal{A}} \pi(x, a) r(x, a)$$

is the average reward under policy π and v^π is the unique solution to

$$\Phi^T D \Phi v^\pi = \Phi^T D T(\Phi v^\pi)$$

Thanks to all the previous results, convergence of Algorithm 1 can be established by applying Theorem 2 from Bhatnagar et al. [2009] which exploits Borkar's work on two-timescale algorithms [1997]. For simplicity, we assume that the iterates resulting from the actor update (Line 10 of Algorithm 1) in SR-AC remain bounded, although one could prove convergence without such an assumption by incorporating an operator that projects any policy parameter to a compact set, as described in Kushner and Clark [1978]. The resulting actor update would then be the projected value of the predefined iterate. The convergence result is presented as Theorem A.1.

Theorem A.1. *Under all the previous assumptions, given $\epsilon > 0$, there exists $\delta > 0$ such that for a parameter vector $\theta_t, t \geq 0$ obtained using the algorithm, if $\sup_{\pi_t} \|e^{\pi_t}\| < \delta$, then the SR-AC algorithm converges almost surely to an ϵ-neighborhood of a local maximum of $\bar{J}$.*

B Experiments

B.1 One-step MDP

Model Parameters	Value
Nominal probability of success	0.8
Uncertainty set for probabilities of success	$[0.1, 0.7, 0.8, 0.3, 0.5]$
Weighting Distribution 1	$[0.47, 0.22, 0.1, 0.09, 0.12]$
Weighting Distribution 2	$[0.63, 0.04, 0.05, 0.02, 0.26]$
Aggressive rewards	$10^5; -10^5$
Soft robust rewards	$5000; -100$
Robust rewards	$2000; 0$

Hyperparameters	Value
Critic Learning rate α	5e-3
Actor Learning rate β	5e-5
Step size ξ	3α
Number of linear features	5
Number of episodes for training M_{train}	3000
Number of episodes for testing M_{test}	600

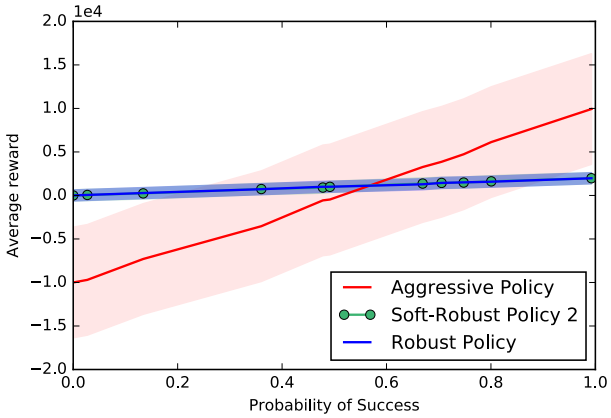

Figure 1: Average reward for different probabilities of success (distribution 2). Soft-robust policy interpolates between aggressive and robust strategies.

B.2 Cart-Pole example

Hyperparameters	Value
Discount factor γ	0.9
Learning rate	1e-4
Mini-batch size	256
Final epsilon	1e-5
Target update interval	10
Max number of episodes for training M_{train}	3000
Number of episodes for testing M_{test}	600

We trained a soft-robust agent on a different weighting over the uncertainty set. Figure 2 shows the performance of the resulting strategy that presents a similar performance as the robust agent. This stronger form of robustness demonstrates

the flexibility we have on the way we fix the weights, which leads to more or less aggressive behaviors.

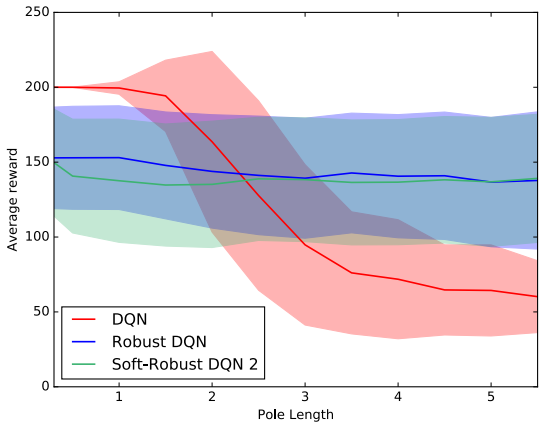

Figure 2: Average reward performance for DQN, robust DQN and soft-robust DQN (distribution 2). Soft-robust policy interpolates between aggressive and robust strategies.

B.3 Pendulum

Hyperparameters	**Value**
Discount factor γ	0.99
Actor learning rate	1e-5
Critic learning rate	1e-3
Mini-batch size	64
Soft target update	$\tau = 0.001$
Max number of episodes for training M_{train}	5000
Number of episodes for testing M_{test}	800

References

Shalabh Bhatnagar, Richard Sutton, Mohammad Ghavamzadeh, and Mark Lee. *Natural Actor-Critic Algorithms*. Automatica, elsevier edition, 2009.

Vivek S. Borkar. Stochastic Approximation with Two Timescales. *Systems and Control Letters*, 29:291–294, 1997.

Harold J. Kushner and Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer Verlag, 1978.

Daniel J Mankowitz, Timothy A Mann, Shie Mannor, Doina Precup, and Pierre-Luc Bacon. Learning Robust Options. In *AAAI*, 2018.

Richard S. Sutton, David McAllester, Satinger Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, volume 12, pages 1057–1063, 2000.

Huan Xu and Shie Mannor. Distributionally Robust Markov Decision Processes. *Mathematics of Operations Research*, 37(2):288–300, 2012.