A Bayesian Approach to Robust Reinforcement Learning - Appendix

A Theoretical Proofs

Recall the assumptions made in the paper:

Assumption A.1. *For any episode, the graph resulting from a worst-case transition model is directed and acyclic.*

Assumption A.2. *For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, the rewards are bounded: $-R_{\max} \leq r_{sa} \leq R_{\max}$. This implies that the robust Q-value is bounded as well: $\mid Q^h_{sa} \mid \leq HR_{\max} =: Q_{\max}$.*

Recall also the worst-case transition from a posterior uncertainty set:

$$\widehat{Q}^h_{sa} = r^h_{sa} + \inf_{p \in \widehat{\mathcal{P}}^h_{sa}} \sum_{s',a'} \pi^h_{s'a'} p_{sas'} \widehat{Q}^{h+1}_{s'a'} \quad ,$$

with $\widehat{Q}^{H+1}_{sa} = 0$ and

$$\widehat{p}^h_{sa} \in \arg \min_{p \in \widehat{\mathcal{P}}^h_{sa}} \sum_{s',a'} \pi^h_{s'a'} p_{sas'} \widehat{Q}^{h+1}_{s'a'} \tag{1}$$

is a worst-case transition at step h.

A.1 Proof of Lemma 4.1

Lemma A.1. *Under Assumptions A.1 and A.2, for any worst-case transition $\widehat{p}$ as defined in equation (1), the conditional variance of the robust Q-values under the posterior distribution satisfies the robust Bellman inequality:*

$$\mathbf{var}_t \widehat{Q}^h_{sa} \leq \nu^h_{sa} + \sum_{s',a'} \pi^h_{s'a'} \mathbb{E}_t \left(\widehat{p}^h_{sas'} \right) \mathbf{var}_t \widehat{Q}^{h+1}_{s'a'},$$

with $\mathbf{var}_t \widehat{Q}^{H+1} = 0$ and $\nu^h_{sa} := Q^2_{\max} \sum_{s' \in \mathcal{S}} \frac{\mathbf{var}_t \widehat{p}^h_{sas'}}{\mathbb{E}_t \widehat{p}^h_{sas'}}$.

Proof. The proof for the robust setup follows the same line as in O'Donoghue et al. [2018] and is given here for completeness.

First rewrite the conditional variance:

$$\begin{aligned}
\mathrm{var}_t(\widehat{Q}^h_{sa}) &:= \mathbb{E}_t \left(\widehat{Q}^h_{sa} - \mathbb{E}_t \widehat{Q}^h_{sa} \right)^2 \\
&= \mathbb{E}_t \left(\inf_{p \in \widehat{\mathcal{P}}^h_{sa}} \sum_{s',a'} \pi^h_{s'a'} p^h_{sas'} \widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t \inf_{p \in \widehat{\mathcal{P}}^h_{sa}} \sum_{s',a'} \pi^h_{s'a'} p^h_{sas'} \widehat{Q}^{h+1}_{s'a'} \right)^2 \\
&= \mathbb{E}_t \left(\sum_{s',a'} \pi^h_{s'a'} \widehat{p}^h_{sas'} \widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t \sum_{s',a'} \pi^h_{s'a'} \widehat{p}^h_{sas'} \widehat{Q}^{h+1}_{s'a'} \right)^2 \\
&= \mathbb{E}_t \left(\sum_{s',a'} \pi^h_{s'a'} \left(\widehat{p}^h_{sas'} \widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t \widehat{p}^h_{sas'} \widehat{Q}^{h+1}_{s'a'} \right) \right)^2 ,
\end{aligned}$$

where we used the following definitions:

$$\begin{aligned}
\widehat{Q}^h_{sa} &= r^h_{sa} + \inf_{p \in \widehat{\mathcal{P}}^h_{sa}} \sum_{s',a'} \pi^h_{s'a'} p^h_{sas'} \widehat{Q}^{h+1}_{s'a'} \\
\widehat{p}^h_{sa} &\in \arg \inf_{p \in \widehat{\mathcal{P}}^h_{sa}} \sum_{s',a'} \pi^h_{s'a'} p^h_{sas'} \widehat{Q}^{h+1}_{s'a'} .
\end{aligned}$$

Assume that $\mathbb{E}_t\widehat{p}_{sas'} > 0$ for all h, s, a, s' belonging to the adequate sets. Since any worst-case transition satisfies $\sum_{s'} \widehat{p}^h_{sas'} = 1$, we have $\sum_{s',a'} \pi_{s'a'}\mathbb{E}_t\widehat{p}^h_{sas'} = 1$ and $\pi_{s'a'}\mathbb{E}_t\widehat{p}^h_{sas'}$ defines a probability distribution over states and actions. Thus,

$$\begin{aligned}\mathbb{E}_t\left(\sum_{s',a'}\pi^h_{s'a'}\left(\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'}\right)\right)^2 &= \mathbb{E}_t\left(\sum_{s',a'}\pi^h_{s'a'}\frac{\mathbb{E}_t\widehat{p}^h_{sas'}}{\mathbb{E}_t\widehat{p}^h_{sas'}}(\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t\sum_{s',a'}\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'})\right)^2\\ &\leq \sum_{s',a'}\pi^h_{s'a'}\frac{\mathbb{E}_t\widehat{p}^h_{sas'}}{\left(\mathbb{E}_t\widehat{p}^h_{sas'}\right)^2}\mathbb{E}_t\left(\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t\sum_{s',a'}\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'}\right)^2,\end{aligned}$$

by applying Jensen's inequality to the convex function $x \mapsto x^2$. Therefore,

$$\mathrm{var}_t(\widehat{Q}^h_{sa}) \leq \sum_{s',a'}\pi^h_{s'a'}\frac{\mathbb{E}_t\widehat{p}^h_{sas'}}{\left(\mathbb{E}_t\widehat{p}^h_{sas'}\right)^2}\mathbb{E}_t\left(\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'}\right)^2$$

Rewriting $\widehat{Q}^{h+1}_{s'a'} = r^{h+1}_{s'a'} + \inf_{p\in\widehat{\mathcal{P}}^{h+1}_{s'a'}}\sum_{s'',a''}\pi^{h+1}_{s''a''}p^{h+1}_{s''s'a'}\widehat{Q}^{h+2}_{s''a''}$ and $\widehat{p}^h_{sa} = \arg\inf_{p\in\widehat{\mathcal{P}}^h_{sa}}\sum_{s',a'}\pi^h_{s'a'}p^h_{sas'}\widehat{Q}^{h+1}_{s'a'}$ enables us to claim that under Assumption A.1, $\widehat{p}^h_{sa}$ is independent of $\widehat{Q}^{h+1}_{sa}$ conditionally on $\mathcal{F}_t$, because $\widehat{Q}^{h+1}_{s'a'}$ depends on downstream uncertainty sets. Note that this claim relies on the rectangular structure of the uncertainty set. Thus,

$$\begin{aligned}\mathbb{E}_t\left(\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t\widehat{p}^h_{sas'}\widehat{Q}^{h+1}_{s'a'}\right)^2 &= \mathbb{E}_t\left((\widehat{p}^h_{sas'} - \mathbb{E}_t\widehat{p}^h_{sas'})\widehat{Q}^{h+1}_{s'a'} + \mathbb{E}_t\widehat{p}^h_{sas'}(\widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t\widehat{Q}^{h+1}_{s'a'})\right)^2\\ &= \mathbb{E}_t\left((\widehat{p}^h_{sas'} - \mathbb{E}_t\widehat{p}^h_{sas'})\widehat{Q}^{h+1}_{s'a'}\right)^2 + \mathbb{E}_t\left(\widehat{p}^h_{sas'}(\widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t\widehat{Q}^{h+1}_{s'a'})\right)^2.\end{aligned}$$

We use the conditional independence property again and Assumption A.2 in order to deduce the following:

$$\mathbb{E}_t\left((\widehat{p}^h_{sas'} - \mathbb{E}_t\widehat{p}^h_{sas'})\widehat{Q}^{h+1}_{s'a'}\right)^2 = \mathbb{E}_t(\widehat{p}^h_{sas'} - \mathbb{E}_t\widehat{p}^h_{sas'})^2\mathbb{E}_t(\widehat{Q}^{h+1}_{s'a'})^2 \leq Q^2_{\max}\mathbf{var}_t\widehat{p}^h_{sas'},$$

$$\text{and } \mathbb{E}_t\left(\widehat{p}^h_{sas'}(\widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t\widehat{Q}^{h+1}_{s'a'})\right)^2 = \mathbb{E}_t(\widehat{p}^h_{sas'})^2\mathbb{E}_t(\widehat{Q}^{h+1}_{s'a'} - \mathbb{E}_t\widehat{Q}^{h+1}_{s'a'})^2 = \mathbb{E}_t(\widehat{p}^h_{sas'})^2\mathbf{var}_t\widehat{Q}^{h+1}_{s'a'}.$$

Finally,

$$\begin{aligned}\mathrm{var}_t(\widehat{Q}^h_{sa}) &\leq \sum_{s',a'}\pi^h_{s'a'}\frac{\mathbb{E}_t\widehat{p}^h_{sas'}}{\left(\mathbb{E}_t\widehat{p}^h_{sas'}\right)^2}(Q^2_{\max}\mathbf{var}_t\widehat{p}^h_{sas'} + \mathbb{E}_t(\widehat{p}^h_{sas'})^2\mathbf{var}_t\widehat{Q}^{h+1}_{s'a'})\\ &\leq \sum_{s',a'}\pi^h_{s'a'}\frac{\mathbb{E}_t\widehat{p}^h_{sas'}}{\left(\mathbb{E}_t\widehat{p}^h_{sas'}\right)^2}Q^2_{\max}\mathbf{var}_t\widehat{p}^h_{sas'} + \sum_{s',a'}\pi^h_{s'a'}\mathbb{E}_t\widehat{p}^h_{sas'}\mathbf{var}_t\widehat{Q}^{h+1}_{s'a'}\\ &\leq Q^2_{\max}\sum_{s'}\frac{\mathbf{var}_t\widehat{p}^h_{sas'}}{\mathbb{E}_t\widehat{p}^h_{sas'}} + \sum_{s',a'}\pi^h_{s'a'}\mathbb{E}_t\widehat{p}^h_{sas'}\mathbf{var}_t\widehat{Q}^{h+1}_{s'a'}\\ &\leq \nu^h_{sa} + \sum_{s',a'}\pi^h_{s'a'}\mathbb{E}_t\widehat{p}^h_{sas'}\mathbf{var}_t\widehat{Q}^{h+1}_{s'a'},\end{aligned}$$

where ν^h_{sa} is given by $\nu^h_{sa} := Q^2_{\max}\sum_{s'}\frac{\mathbf{var}_t\widehat{p}^h_{sas'}}{\mathbb{E}_t\widehat{p}^h_{sas'}}$. □

A.2 Proof of Theorem 4.1

Theorem A.1 (Solution of URBE)**.** *For any worst-case transition $\widehat{p}$ as defined in equation (1) and any policy π, under Assumptions A.1 and A.2, there exists a unique mapping w that satisfies the uncertainty robust Bellman equation:*

$$w^h_{sa} = \nu^h_{sa} + \sum_{s'\in\mathcal{S},a'\in\mathcal{A}}\pi^h_{s'a'}\mathbb{E}_t(\widehat{p}^h_{sas'})w^{h+1}_{s'a'}, \tag{2}$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $h = 1, \cdots, H$ where $w^{H+1} = 0$. Furthermore, $w \geq \mathbf{var}_t\widehat{Q}$.

Proof. Denote by $\mathcal{W}^h$ the robust Bellman operator underlying equation (2) and rewrite is as $\mathcal{W}^h w^{h+1} = w^h$. We can easily see that the robust Bellman operator is non-decreasing. Also, it has a unique solution, as stated in the following lemma, which is the policy evaluation version of the Min-Max Problem addressed in Bertsekas [2000] (Exercise 1.5).

Lemma A.2. *For every $(s, a) \in \mathcal{S} \times \mathcal{A}$, for all step $h = 1, \cdots, H$, w_{sa}^h is given by the subsequent steps of the following algorithm which proceeds backwards from $H+1$ to h:*

$$\begin{cases} w_{sa}^{H+1} = 0 \textit{ for all } (s,a) \in \mathcal{S} \times \mathcal{A} \\ w_{sa}^h = \nu_{sa}^h + \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \pi_{s'a'}^h \mathbb{E}_t(\widehat{p}_{sas'}^h) w_{s'a'}^{h+1} \end{cases}$$

Therefore, there exists a unique solution to $\mathcal{W}^h w^{h+1} = w^h, h = 1, \cdots, H$.

The lower-bound then follows from induction reasoning. At step H, we have $\mathbf{var}_t \widehat{Q}^{H+1} = 0 = w^{H+1}$. Assume that for some $h \leq H$ we have $w^{h+1} \geq \mathbf{var}_t \widehat{Q}^{h+1}$. Then, by assumption and using Lemma A.1, we get:

$$\mathbf{var}_t \widehat{Q}^h \leq \mathcal{W}^h \mathbf{var}_t \widehat{Q}^{h+1} \leq \mathcal{W}^h w^{h+1} = w^h.$$

The induction property is hereditary, which concludes the proof of the theorem. □

B DQN-URBE Experiments

Table 1: System's dynamics

	MARSROVER	**CARTPOLE**
Nominal model	$p = 0.005$	Length = 0.75, Mass = 1
Size of uncertainty set	15 samples	15 samples

Table 2: Networks

DQN-URBE NETWORKS	**MARSROVER**	**CARTPOLE**
Q-network	ReLu(2 hidden layers of size 10)	ReLu(3 hidden layers of size 128)
U(R)BE-network	ReLu(1 hidden layer of size 15), linear activation function for the output	ReLu(1 hidden layer of size 100), linear activation function for the output

Table 3: Hyper-parameters

DQN-URBE HYPERPARAMETERS	**MARSROVER**	**CARTPOLE**
Discount factor γ	0.9	0.9
Q-learning rate	1e-4	1e-4
U(R)BE network learning rate	1e-4	1e-4
Initial variance coefficient μ	1e-2	1e-2
Posterior parameter β	0.5	0.5
Mini-batch size	100	256
Final epsilon	1e-3	1e-5
Target update interval	10	10
Max number of episodes for training M_{train}	3000	4000
Number of episodes for testing M_{test}	200	200

References

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, $2^d edition,$ 2000.

Brendan O'Donoghue, Ian Osband, Remi Munos, and Volodymyr Mnih. The Uncertainty Bellman Equation and Exploration. *Proceedings of the 35th International Conference on Machine Learning*, 2018.