Kernel-based Reinforcement Learning in Robust Markov Decision Processes

Shiau Hong Lim, Arnaud Autef

Motivation

- Robust Markov Decision Process (MDP) framework
  - Tackle model mismatch and parameter uncertainty
  - Previously, for state aggregation, performance bound on $||v_R^\pi - v^*||$ improved via robust policies:

$$O\left(\frac{1}{(1-\gamma)^2}\right) \rightarrow O\left(\frac{1}{1-\gamma}\right)$$

Contribution

1. Robust performance bound improvement on $||v_R^\pi - v^*||$ extended to the general kernel averager setting
$$O\left(\frac{1}{(1-\gamma)^2}\right) \rightarrow O\left(\frac{1}{1-\gamma}\right)$$
2. Formulation of a practical kernel-based robust algorithm, with empirical results on benchmark tasks

Kernel-based approach

1. MDP to solve $\mathcal{M}$
2. Kernel averager Φ and representative states $j \in \{1, \ldots, m\}$ to approximate the value function:

$$v = \Phi w$$

$$\forall i, j, 0 \leq \Phi_{i,j} \leq 1 \text{ and } \forall i, \sum_j \Phi_{i,j} = 1$$

$$\forall j, M(j) \coloneqq \{i \mid \Phi_{i,j} > 0\}$$

Kernel-based approach

2. Define a non-trivial robust MDP $\tilde{\mathcal{M}}$ with states = representative states

3. Obtain w^* optimal robust value in $\tilde{\mathcal{M}}$

4. Derive π_{w^*} in $\mathcal{M}$ greedy w.r.t w^*, with:

$$\Phi w^* \leq v^{\pi_{w^*}}$$

Theoretical Result

Theorem:
w^* optimal robust value in $\widetilde{\mathcal{M}}$, π_{w^*} $\mathcal{M}$ greedy policy w.r.t w^*, v^* optimal value in $\mathcal{M}$:

$$\left|\left|v^{\pi_{w^*}} - v^*\right|\right|_\infty \leq \frac{2\epsilon + L_0}{1 - \gamma}$$

- $w_0 = \arg\min_w \left|\left|v^* - \Phi w\right|\right|_\infty$
- $\epsilon = \min_w ||v^* - \Phi w||_\infty$ → Function approximator limitations
- $L_0 = \max_{\{(j,j') \in \tilde{S} \mid M(j) \cap M(j') \neq \emptyset\}} |w_0(j) - w_0(j')|$ → v^* Smoothness

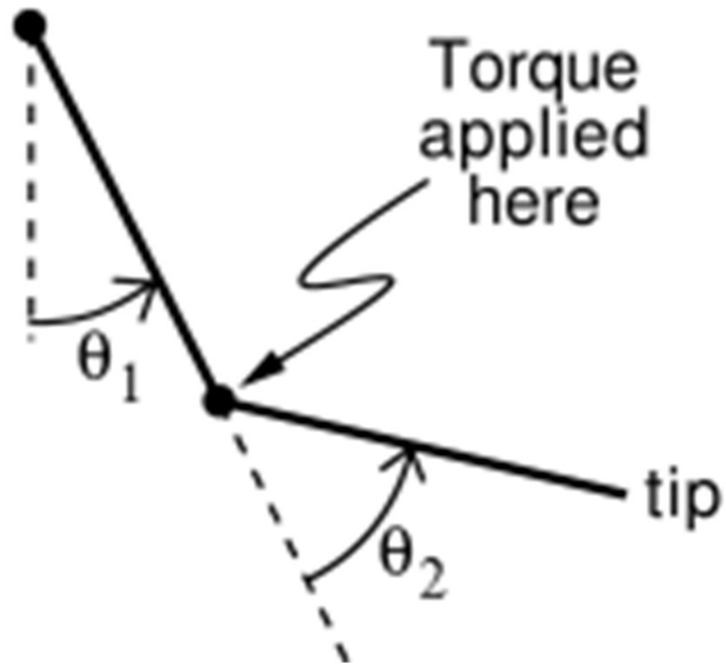
Practical algorithm

1. Second kernel averager Ψ to approximate the MDP model $(r, P) \rightarrow (\hat{r}, \hat{P})$ from data
2. Solve $\widetilde{\mathcal{M}}$ with the approximate robust Bellman operator:

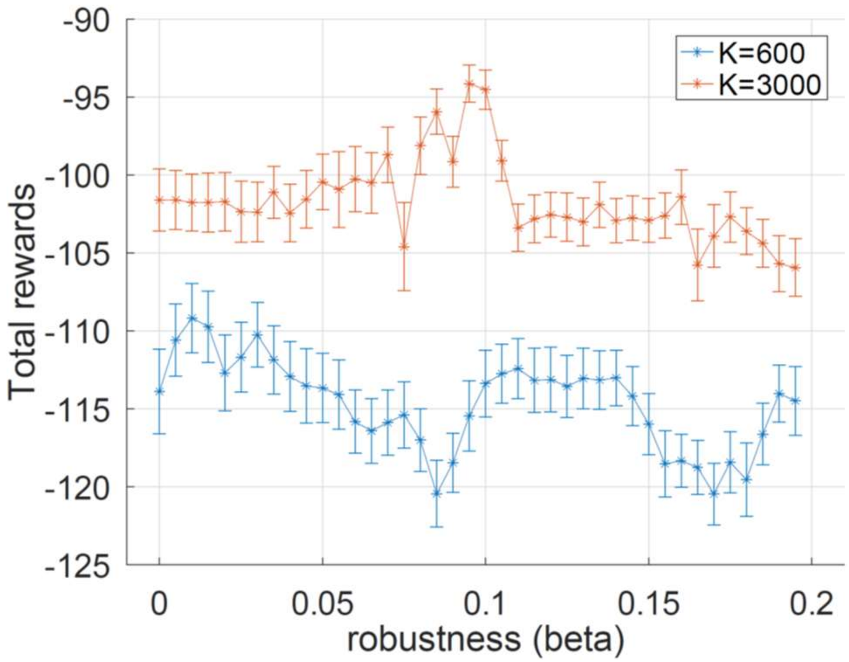
$$w^{t+1}(j) \leftarrow \max_{a} \min_{\left\|p - \hat{\psi}_a(i_j)\right\|_1 \leq \beta} \langle p, r^a + \gamma \Phi^a w^t \rangle$$

With Robustness parameter $\beta \in [0, 2]$

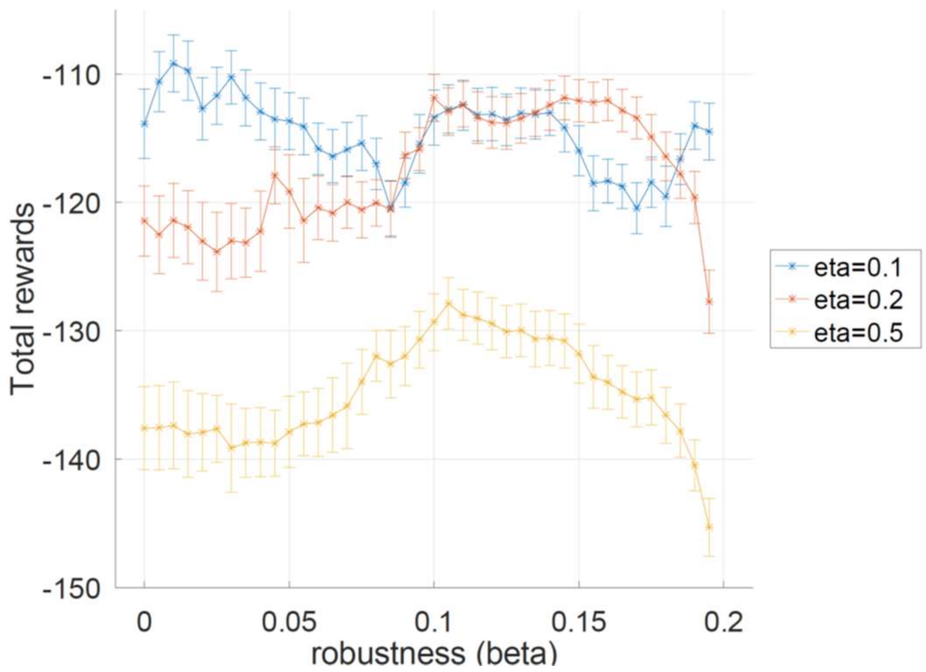
Experiments: Acrobot



Acrobot

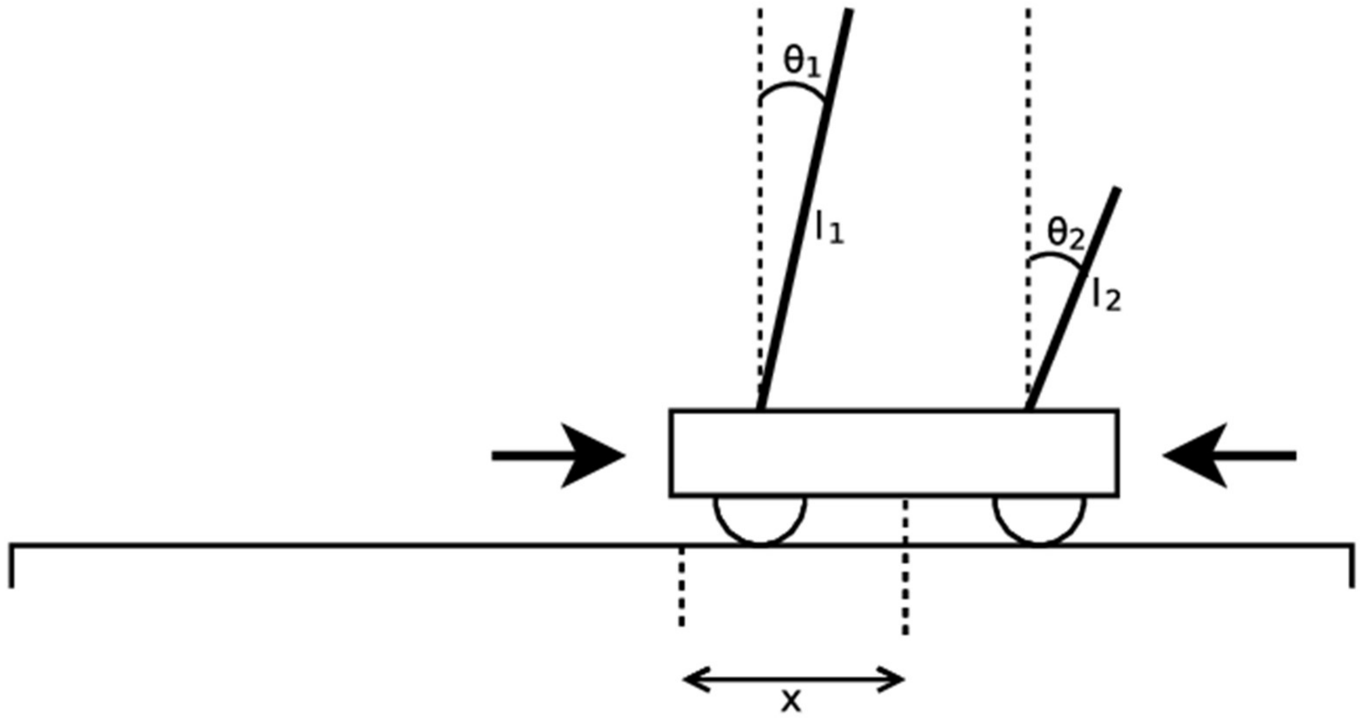


-90
-95
-100
-105
-110
-115
-120
-125
Total rewards
K=600
K=3000
0
0.05
0.1
0.15
0.2
robustness (beta)

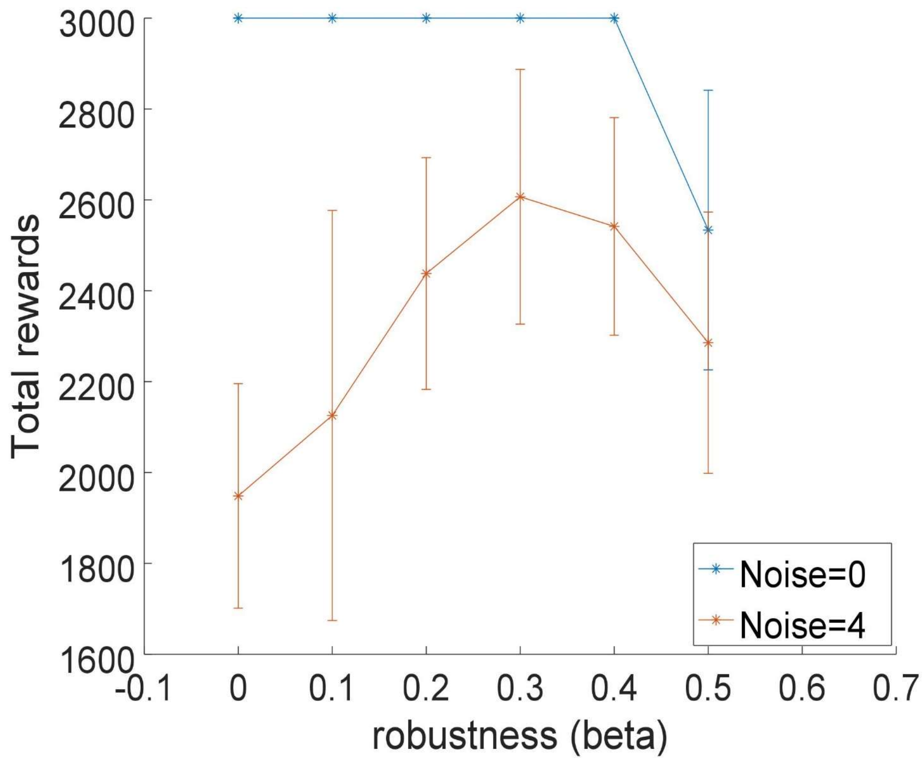


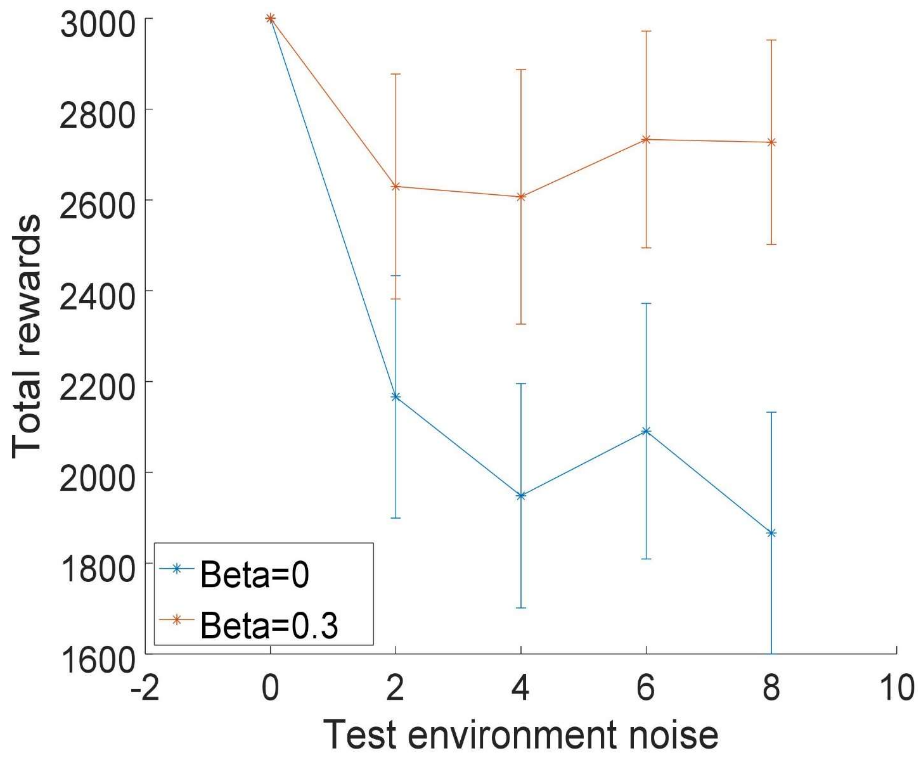
-110
-120
-130
-140
-150
Total rewards
eta=0.1
eta=0.2
eta=0.5
0
0.05
0.1
0.15
0.2
robustness (beta)

Experiments: Double Pole Balancing



Double Pole Balancing





Conclusion

- Theoretical performance guarantees for robust kernel-based reinforcement learning in $O\left(\frac{1}{1-\gamma}\right)$

- Significant empirical benefits from robustness, even stronger with model mismatch (real-world settings)

Thank you!
Please come to see our poster tonight

Shiau Hong Lim, Arnaud Autef