Motivation

• Robust Markov Decision Process (MDP) framework
  – Tackle model mismatch and parameter uncertainty
  – Previously, for state aggregation, performance bound on $||\nu^\pi_R - \nu^*||$ 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 $||v^K - v^*||$ 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 $\Phi$ and representative states $j \in \{1, \ldots, m\}$ to approximate the value function:

$$v = \Phi w$$

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

$\forall j, M(j) := \{i \mid \Phi_{i,j} > 0\}$
Kernel-based approach

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

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

4. Derive $\pi_{w^*}$ in $M$ greedy w.r.t $w^*$, with:
   $$\Phi w^* \leq \nu^{\pi_{w^*}}$$
Theoretical Result

Theorem:

\( w^* \) optimal robust value in \( \tilde{\mathcal{M}} \), \( \pi_{w^*} \) \( \mathcal{M} \) greedy policy w.r.t \( w^* \), \( \nu^* \) optimal value in \( \mathcal{M} \):

\[
||\nu^{\pi_{w^*}} - \nu^*||_\infty \leq \frac{2\epsilon + L_0}{1 - \gamma}
\]

- \( w_0 = \arg\min_w ||\nu^* - \Phi w||_\infty \)
- \( \epsilon = \min_w ||\nu^* - \Phi w||_\infty \) \( \rightarrow \) Function approximator limitations
- \( L_0 = \max_w \{ (j, j') \in \tilde{\mathcal{S}} | \mathcal{M}(j) \cap \mathcal{M}(j') \neq \emptyset \} \) \( \rightarrow \) \( \nu^* \) Smoothness
Practical algorithm

1. Second kernel averager $\Psi$ to approximate the MDP model $(r, P) \rightarrow (\hat{r}, \hat{P})$ from data

2. Solve $\tilde{M}$ with the approximate robust Bellman operator:

$$w^{t+1}(j) \leftarrow \max_a \min_{\beta} \min \frac{\langle p, r^a + \gamma \Phi^a w^t \rangle}{\| p - \tilde{\psi}_a(i_j) \|_1} \leq \beta$$

With Robustness parameter $\beta \in [0, 2]$
Experiments: Acrobot
Acrobot
Experiments: Double Pole Balancing
Double Pole Balancing

![Graphs showing total rewards vs robustness and test environment noise.](image)
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