Policy-Conditioned Uncertainty Sets for Robust Markov Decision Processes

Andrea Tirinzoni[1], Xiangli Chen[2], Marek Petrik[3], and Brian D. Ziebart[4]

[1] Politecnico di Milano
[2] Amazon Robotics
[3] University of New Hampshire
[4] University of Illinois at Chicago

Advances in Neural Information Processing Systems 2018

Why Robust MDPs?

- MDPs are powerful tools for modeling sequential decision making problems
- Transition probabilities are often **uncertain**
- Estimation errors can have detrimental effects on the resulting policies
- Unacceptable in applications involving high level of **risk**



Why Robust MDPs?

- MDPs are powerful tools for modeling sequential decision making problems
- Transition probabilities are often **uncertain**
- Estimation errors can have detrimental effects on the resulting policies
- Unacceptable in applications involving high level of **risk**



- Need solutions that are **robust** to this uncertainty

Problem

- **Robust MDPs** given sample trajectories from a reference policy $\widetilde{\pi}$
  - Build **uncertainty sets** Ξ containing the true parameters τ with high probability
  - Compute the optimal policy under the **worst-case** parameters in these sets

$$\max_{\pi} \min_{\tau \in \Xi} \ \mathbb{E}_{\tau,\pi} \left[\sum_{t=1}^{T-1} R(S_t, A_t, S_{t+1}) \right]$$

Problem

- **Robust MDPs** given sample trajectories from a reference policy $\widetilde{\pi}$
  - Build **uncertainty sets** Ξ containing the true parameters τ with high probability
  - Compute the optimal policy under the **worst-case** parameters in these sets

$$\max_{\pi} \min_{\tau \in \Xi} \mathbb{E}_{\tau,\pi} \left[\sum_{t=1}^{T-1} R(S_t, A_t, S_{t+1}) \right]$$

- This problem is **NP-hard** in general [Mannor et al., 2012]

Problem

- **Robust MDPs** given sample trajectories from a reference policy $\widetilde{\pi}$
  - Build **uncertainty sets** Ξ containing the true parameters τ with high probability
  - Compute the optimal policy under the **worst-case** parameters in these sets

$$\max_{\pi} \min_{\tau \in \Xi} \mathbb{E}_{\tau,\pi} \left[\sum_{t=1}^{T-1} R(S_t, A_t, S_{t+1}) \right]$$

- This problem is **NP-hard** in general [Mannor et al., 2012]
- **Rectangular** (independent) constraints [Nilim and El Ghaoui, 2005, Iyengar, 2005] provide tractability, but are too **conservative** and do not generalize

Non-Rectangular Uncertainty Sets via Marginal Features

- We consider **features** $\phi(s, a, s')$ to model the relationships between states and actions
- **Feature expectations** [Abbeel and Ng, 2004] to model the interaction of a policy π with the decision process

$$\boldsymbol{\kappa}_{\phi}(\pi, \tau) = \mathbb{E}_{\tau,\pi}\left[\sum_{t=1}^{T-1} \phi(S_t, A_t, S_{t+1})\right]$$

Non-Rectangular Uncertainty Sets via Marginal Features

- We consider **features** $\phi(s, a, s')$ to model the relationships between states and actions
- **Feature expectations** [Abbeel and Ng, 2004] to model the interaction of a policy π with the decision process

$$\boldsymbol{\kappa}_{\phi}(\pi, \tau) = \mathbb{E}_{\tau,\pi}\left[\sum_{t=1}^{T-1} \phi(S_t, A_t, S_{t+1})\right]$$

- Use feature expectations to define the **uncertainty sets**:

$$\boxed{\Xi_{\widetilde{\pi}}^{\phi}} = \left\{\tau : \boldsymbol{\kappa}_{\phi}(\widetilde{\pi}, \tau) = \widehat{\boldsymbol{\kappa}}\right\} \quad \text{or} \quad \boxed{\widetilde{\Xi}_{\widetilde{\pi}}^{\phi}} = \left\{\tau : \|\boldsymbol{\kappa}_{\phi}(\widetilde{\pi}, \tau) - \widehat{\boldsymbol{\kappa}}\| \leq \epsilon\right\}$$

Some Appealing Properties

- Constrain whole trajectories rather than single states

Some Appealing Properties

- Constrain whole trajectories rather than single states
- Can **generalize** across the state space

Some Appealing Properties

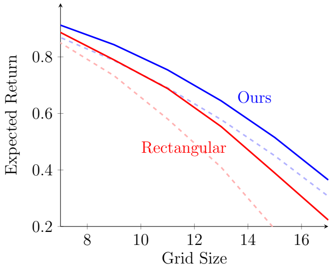
- Constrain whole trajectories rather than single states
- Can **generalize** across the state space
- Uncertainty sets are **policy-conditioned**

Some Appealing Properties

- Constrain whole trajectories rather than single states
- Can **generalize** across the state space
- Uncertainty sets are **policy-conditioned**
- **Tractable** optimization

Some Appealing Properties

- Constrain whole trajectories rather than single states
- Can **generalize** across the state space
- Uncertainty sets are **policy-conditioned**
- **Tractable** optimization
- Less conservative empirical performance than rectangular solutions



Please visit us at poster #168

References

Abbeel, P. and Ng, A. Y. (2004).
Apprenticeship learning via inverse reinforcement learning.
In *Proc. International Conference on Machine Learning*, pages 1–8.

Iyengar, G. N. (2005).
Robust dynamic programming.
Mathematics of Operations Research, 30(2):257–280.

Mannor, S., Mebel, O., and Xu, H. (2012).
Lightning does not strike twice: Robust mdps with coupled uncertainty.
arXiv preprint arXiv:1206.4643.

Nilim, A. and El Ghaoui, L. (2005).
Robust control of markov decision processes with uncertain transition matrices.
Operations Research, 53(5):780–798.