Policy-Conditioned Uncertainty Sets for Robust Markov Decision Processes

Andrea Tirinzoni, Xiangli Chen, Marek Petrik, and Brian D. Ziebart

andrea.tirinzoni@polimi.it, cxiangli@amazon.com, mpetrik@cs.unh.edu, bziebart@uic.edu

Planning and Learning Workshop 2018

POLITECNICO
MILANO 1863

amazon

University of
New Hampshire

UIC

Motivation

- MDPs are powerful tools for modeling sequential decision making problems
- In practice, MDP parameters are often **uncertain**
  - Partial observability
  - Incorrect measurements
  - Finite samples
- **Goal**: Find a good control policy in the presence of uncertainty

Problem

- We consider an MDP $\langle \mathcal{S}, \mathcal{A}, \tau, R \rangle$ with **unknown** transition probabilities $\tau(s_{t+1}|s_t, a_t)$
- A limited number of trajectories generated from a given **reference policy** $\widetilde{\pi}$ are available

Problem

- We consider an MDP $\langle \mathcal{S}, \mathcal{A}, \tau, R \rangle$ with **unknown** transition probabilities $\tau(s_{t+1}|s_t, a_t)$
- A limited number of trajectories generated from a given **reference policy** $\widetilde{\pi}$ are available
- **Robust MDPs**:
  - Build uncertainty sets Ξ containing τ with high probability
  - Compute the optimal policy under the **worst-case** parameters in these sets

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

Problem

- We consider an MDP $\langle \mathcal{S}, \mathcal{A}, \tau, R \rangle$ with **unknown** transition probabilities $\tau(s_{t+1}|s_t, a_t)$
- A limited number of trajectories generated from a given **reference policy** $\widetilde{\pi}$ are available
- **Robust MDPs**:
    - Build uncertainty sets Ξ containing τ with high probability
    - Compute the optimal policy under the **worst-case** parameters in these sets

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

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

Rectangular Uncertainty Sets

- The majority of the RMDP literature considers **rectangular** uncertainty sets [Nilim and El Ghaoui, 2005, Wiesemann et al., 2013]:

$$\Xi = \left\{ \tau : \forall s, a \in \mathcal{S} \times \mathcal{A},\ \|\tau(\cdot|s,a) - p_{s,a}\| \leq \epsilon_{s,a} \right\}$$

Rectangular Uncertainty Sets

- The majority of the RMDP literature considers **rectangular** uncertainty sets [Nilim and El Ghaoui, 2005, Wiesemann et al., 2013]:

$$\Xi = \left\{ \tau : \forall s, a \in \mathcal{S} \times \mathcal{A},\ \|\tau(\cdot|s, a) - p_{s,a}\| \leq \epsilon_{s,a} \right\}$$

- **Pros**
    - Polynomial-time optimization
    - Robust Bellman optimality equation
    - Can be easily formed from samples (e.g., via concentration inequalities)

Rectangular Uncertainty Sets

- The majority of the RMDP literature considers **rectangular** uncertainty sets [Nilim and El Ghaoui, 2005, Wiesemann et al., 2013]:

$$\Xi = \left\{ \tau : \forall s, a \in \mathcal{S} \times \mathcal{A},\ \|\tau(\cdot|s,a) - p_{s,a}\| \le \epsilon_{s,a} \right\}$$

- **Pros**
  - Polynomial-time optimization
  - Robust Bellman optimality equation
  - Can be easily formed from samples (e.g., via concentration inequalities)
- **Cons**
  - Very conservative solutions
  - Does not generalize across the state-action space

Non-Rectangular Uncertainty Sets via Marginal Features

- We consider **features** $\phi(s_t, a_t, s_{t+1})$ 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

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

Non-Rectangular Uncertainty Sets via Marginal Features

- We consider **features** $\phi(s_t, a_t, s_{t+1})$ 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

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

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

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

Marginally-Constrained RMDPs

Constrained Problem

$$\max_{\pi} \min_{\tau \in \Xi_{\tilde{\pi}}^{\phi}} \left\{ \rho(\pi, \tau) - \lambda^{-1} H(\tau) \right\}$$

Marginally-Constrained RMDPs

Constrained Problem

$$\max_{\pi} \min_{\tau \in \Xi^{\phi}_{\tilde{\pi}}} \left\{ \rho(\pi, \tau) - \lambda^{-1} H(\tau) \right\}$$

Benefits:

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

Marginally-Constrained RMDPs

Unconstrained Problem

$$\max_{\omega} \left\{ \max_{\pi} \operatorname{softmin}_{\tau} \left(\rho(\pi, \tau) + \omega \cdot \kappa_{\phi}(\widetilde{\pi}, \tau) \right) - \omega \cdot \widehat{\kappa}_{\phi} \right\}$$

Marginally-Constrained RMDPs

Unconstrained Problem

$$\max_{\omega} \left\{ \max_{\pi} \operatorname*{softmin}_{\tau} \left(\rho(\pi, \tau) + \omega \cdot \kappa_\phi(\widetilde{\pi}, \tau) \right) - \omega \cdot \widehat{\kappa}_\phi \right\}$$

1. **Optimize return $\boldsymbol{\rho}$.** Find the equilibrium (π^*, τ^*) of the inner zero-sum game using *min-max dynamic programming*:

$$(\pi^*, \tau^*) \leftarrow \max_{\pi} \operatorname*{softmin}_{\tau} \left\{ \rho(\pi, \tau) + \omega \cdot \kappa_\phi(\widetilde{\pi}, \tau) \right\}$$

2. **Match Statistics $\boldsymbol{\widehat{\kappa}_\phi}$.** Update parameters ω so that τ^* matches the sample statistics with respect to the reference policy $\widetilde{\pi}$:

$$\omega \leftarrow \omega + \eta \left(\kappa_\phi(\widetilde{\pi}, \tau^*) - \widehat{\kappa}_\phi \right)$$

Solving the Zero-Sum Game

- **Issue**. Solving the zero-sum game at step 1 requires finding dynamics τ that minimize the sum of two expected returns under different policies

$$\max_{\pi} \operatorname*{softmin}_{\tau} \left\{ \rho(\pi, \tau) + \omega \cdot \kappa_{\phi}(\widetilde{\pi}, \tau) \right\}$$

  - **NP-hard** problem [Petrik et al., 2016]
  - *Non-Markovian* solution

Solving the Zero-Sum Game

- **Issue**. Solving the zero-sum game at step 1 requires finding dynamics τ that minimize the sum of two expected returns under different policies

$$\max_{\pi} \operatorname{softmin}_{\tau} \left\{ \rho(\pi, \tau) + \omega \cdot \kappa_{\phi}(\widetilde{\pi}, \tau) \right\}$$

  - **NP-hard** problem [Petrik et al., 2016]
  - *Non-Markovian* solution
- **Main result**. A Markovian solution exists when augmenting the state space with a continuous **belief state** which keeps track of the relative importance of the two policies:

$$b_t = \frac{\prod_{i=1}^{t} \pi(a_i|h_i)}{\prod_{i=1}^{t} \pi(a_i|h_i) + \prod_{i=1}^{t} \widetilde{\pi}(a_i|s_i)}$$

Solving the Zero-Sum Game

- **Issue**. Solving the zero-sum game at step 1 requires finding dynamics τ that minimize the sum of two expected returns under different policies

$$\max_{\pi} \operatorname{softmin}_{\tau} \left\{ \rho(\pi, \tau) + \omega \cdot \kappa_{\phi}(\widetilde{\pi}, \tau) \right\}$$

  - **NP-hard** problem [Petrik et al., 2016]
  - *Non-Markovian* solution

- **Main result**. A Markovian solution exists when augmenting the state space with a continuous **belief state** which keeps track of the relative importance of the two policies:

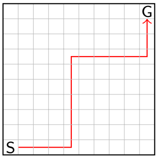
$$b_t = \frac{\prod_{i=1}^{t} \pi(a_i|h_i)}{\prod_{i=1}^{t} \pi(a_i|h_i) + \prod_{i=1}^{t} \widetilde{\pi}(a_i|s_i)}$$

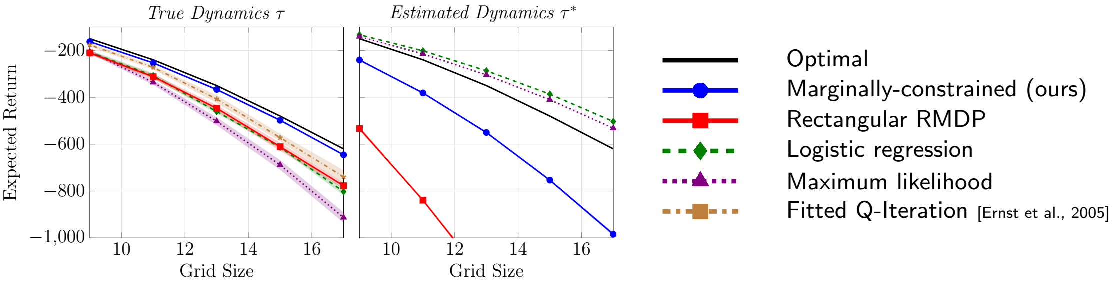
- Approximation by solving a **min-max dynamic program** using discretized belief states

Empirical Evaluation

Grid World

- Different grid sizes
- Fixed number of trajectories under uniform policy





Conclusion

- We proposed a novel class of **uncertainty sets** defined via marginal features of state-action sequences
  - Non-rectangular
  - Policy-conditioned
  - Tractable optimization

- **Future works**:
  - Leverage ideas from POMDPs to improve the optimization
  - Extend the approach to account for multiple reference policies

Contacts

andrea.tirinzoni@polimi.it

https://github.com/AndreaTirinzoni/

References

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

Ernst, D., Geurts, P., and Wehenkel, L. (2005).
Tree-based batch mode reinforcement learning.
Journal of Machine Learning Research.

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.

Petrik, M., Mohammad Ghavamzadeh, and Chow, Y. (2016).
Safe Policy Improvement by Minimizing Robust Baseline Regret.
In *Advances in Neural Information Processing Systems*.

Wiesemann, W., Kuhn, D., and Rustem, B. (2013).
Robust markov decision processes.
Mathematics of Operations Research, 38(1):153–183.