POLICY-CONDITIONED UNCERTAINTY SETS FOR ROBUST MARKOV DECISION PROCESSES

A. TIRINZONI, X. CHEN, M. PETRIK, AND B. ZIEBART

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

POLITECNICO MILANO 1863 · amazon · University of New Hampshire · UIC

PROBLEM

- Compute a **robust policy** π for an MDP $\langle \mathcal{S}, \mathcal{A}, \tau, R \rangle$ whose transition probabilities $\tau(s_{t+1}|s_t, a_t)$ are *unknown*
- Only a *limited* number of trajectories generated from a **reference policy** $\widetilde{\pi}$ is available
- **Robust optimization** approach:
  - Define uncertainty sets Ξ based on samples such that, with high probability, $\tau \in \Xi$
  - Find the optimal policy against the worst-case dynamics in Ξ:

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

MOTIVATION

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

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

- **Rectangular** RMDPs:
  - Polynomial-time optimization 🙂
  - Robust Bellman optimality equation 🙂
  - Very conservative solutions ☹
- **Non-rectangular** RMDPs:
  - **NP-hard** optimization problem in general [e.g., Mannor et al., 2012]

CONTRIBUTIONS

1. We propose **policy-conditioned uncertainty sets**:
   - **Non-rectangular** uncertainty sets via *marginal statistics* of the given trajectories
   - **Off-policy robustness**: the impact of the reference policy on the desired control policy is considered in the learning process
   - **Tractable** and **convex** optimization by shifting to parameterized control problems
2. We provide **empirical results** showing the benefits of our approach over rectangular RMDPs

MARGINALLY-CONSTRAINED ROBUST CONTROL PROCESSES

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**:

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

BENEFITS

- Non-rectangular
- Constrain whole trajectories
- Dependence on the reference policy
- **Generalization** across the state-space

MARGINALLY-CONSTRAINED ROBUST MDP

$$\max_{\pi} \min_{\tau\in\Xi^{\phi}_{\widetilde{\pi}}} \left\{\rho(\pi, \tau) - \lambda^{-1} H(\tau)\right\} \quad \longrightarrow \quad \max_{\omega} \left\{\max_{\pi} \operatorname{softmin}_{\tau} \left(\rho(\pi, \tau) + \omega \cdot \kappa_\phi(\widetilde{\pi}, \tau)\right) - \omega \cdot \widehat{\kappa}_\phi\right\}$$

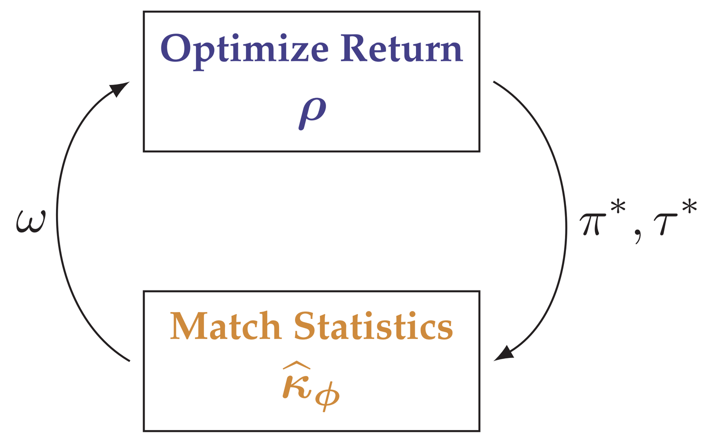
ALTERNATED OPTIMIZATION

1. **Optimize return ρ**. 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 $\widehat{\kappa}_\phi$**. Update parameters ω so that τ^* matches the sample statistics under the reference policy $\widetilde{\pi}$:

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



MIXED-OBJECTIVE MINIMAX OPTIMAL CONTROL

- **Issue**. Solving the zero-sum game at step 1 requires finding dynamics τ that minimize the sum of two expected returns under different policies
  - **NP-hard** problem [Petrik et al., 2016] $\rightarrow$ *Non-Markovian* solution
- **Main result**. Markovian solution when augmenting the state space with a continuous **belief state** to keep track of the relative importance of the two policies:

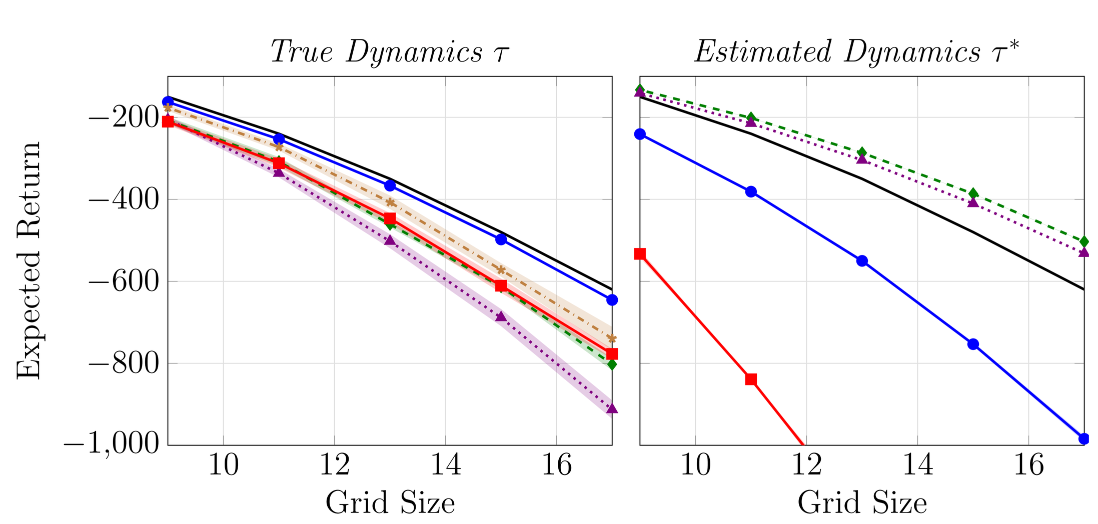
$$b_t = \frac{\prod_{i=1}^{t} \pi(a_i|h_i)}{\prod_{i=1}^{t} \pi(a_i|h_i) + \prod_{i=1}^{t} \widetilde{\pi}(a_i|s_i)} \quad \rightarrow \quad b_{t+1} = \frac{b_t \pi(a_{t+1}|h_{t+1})}{b_t \pi(a_{t+1}|h_{t+1}) + (1-b_t)\widetilde{\pi}(a_{t+1}|s_{t+1})}$$

- The equilibrium can be found by solving a **min-max dynamic program** using discretized belief states:

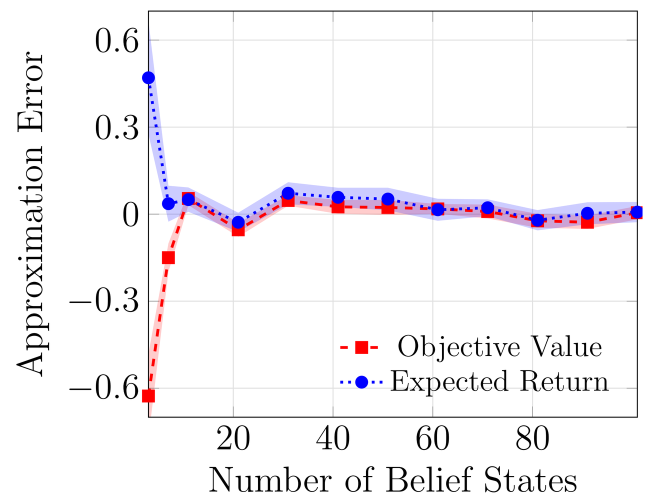
$$\tau^*(s_{t+1}|s_t, a_t, b_t) = \frac{e^{-\lambda Q(s_t, a_t, b_t, s_{t+1})}}{\sum_{s'_{t+1}} e^{-\lambda Q(s_t, a_t, b_t, s'_{t+1})}} \qquad \pi^*(s_t, b_{t-1}) = \arg\max_{a_t} V'(s_t, a_t, b_t)$$

RESULTS

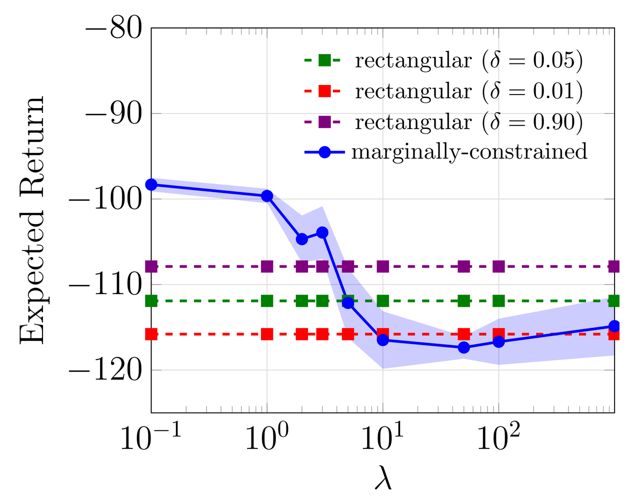
GRID WORLD



BELIEF DISCRETIZATION



ENTROPY REGULARIZATION



Optimal — Marginally-constrained — Rectangular — Logistic regression — Maximum likelihood — Fitted Q-Iteration

REFERENCES

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

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 2005.

Shie Mannor, Ofir Mebel, and Huan Xu. Lightning does not strike twice: Robust mdps with coupled uncertainty. *arXiv preprint arXiv:1206.4643*, 2012.

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

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.