Beyond Confidence Regions: Tight Bayesian Ambiguity Sets for Robust MDPs

Reazul Russel, Marek Petrik, University of New Hampshire, Durham, NH

Overview and Contributions

- **Motivation**: Compute *safe* policies in batch RL with little data
- **Percentile criterion goal**: Policy π that maximizes [Delage et al., 2010]:

$$\max_{\pi, y} y \text{ s.t. } \mathbb{P}_{P^\star}\left[\mathbb{E}_{S_t\sim P^\star, A_t\sim\pi(S_t)}\left[\sum_{t=0}^{\infty}\gamma^t r(S_t, A_t)\right] \geq y\right] \geq 1-\delta$$

  with confidence $1-\delta$ for *Bayesian uncertainty* over $P^\star$
  - *Important in high-stakes domains*: Confidence when deployed, robust policies, mitigation strategies, training, use a different method
- **Contribution**: Safe RL methods with:
  1. Tighter guarantees on return
  2. Tractable policy optimization

Small Example: 80% Confidence in Batch RL

- **True model**: 4 states: $r(\cdot|s_0) = 0$ and

$$p^\star(s_1, s_2, s_3|s_0) = \begin{bmatrix}0.3 & 0.5 & 0.2\end{bmatrix}, \quad v(s_1, s_2, s_3) = \begin{bmatrix}10 & 5 & -1\end{bmatrix},$$

$$v^\star(s_0) = v^\mathsf{T} p^\star = 6.3$$

- **Batch samples**: $4 \times (s_0 \to s_1)$, $6 \times (s_0 \to s_2)$, $1 \times (s_0 \to s_3)$

  Bayesian posterior: $p \sim \text{Dirichlet}(5, 7, 2)$, samples:

$$p_1^\mathsf{T} = \begin{bmatrix}0.2 & 0.7 & 0.1\end{bmatrix}, \; p_2^\mathsf{T} = \begin{bmatrix}0.6 & 0.3 & 0.1\end{bmatrix}, \ldots$$

- **Estimated model**: Provides no guarantees

$$\tilde{v}(s_0) = {}^4\!/_{11} \cdot 10 + {}^6\!/_{11} \cdot 5 + {}^1\!/_{11} \cdot -1 = 6.27$$

- **Percentile criterion**: lower 80%-quantile of values $v^\mathsf{T} p_i$:

$$\hat{v}(s_0) = \mathrm{V@R}_i^{0.8}[v^\mathsf{T} p_i] = 5.8$$

Prior Work

Challenge: Maximizing percentile criterion is *NP-hard*

1. *Concentration inequalities*: Good return guarantees, but intractable policy optimization e.g. [Thomas et al., 2016]
2. *Second order approximations*: Good solution quality, tractable, but no guarantees and not general e.g. [Delage et al., 2010]
3. *Robust MDPs*: General, tractable, but loose guarantees e.g. [Petrik et al., 2016]

Contribution: Tractable RMDPs with tighter guarantees

Optimizing Percentile Criterion via RMDPs

- Maximize a **tractable lower bound** via Robust MDPs
- Robust Bellman optimality for a **set** $\mathcal{P}_{s,a}$ of transition probabilities

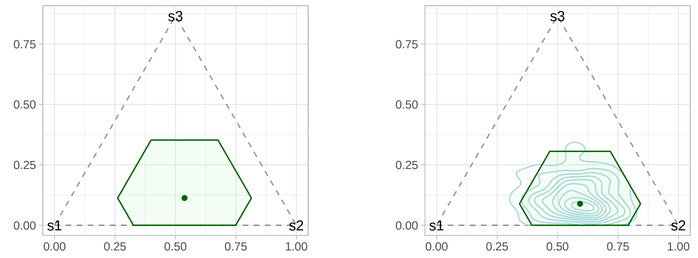
$$\hat{v}(s) = \max_{a\in\mathcal{A}} \min_{p\in\mathcal{P}_{s,a}} \left\{r_{s,a} + \gamma \cdot p^\mathsf{T}\hat{v}\right\}$$

- **Ambiguity set**: e.g. $\mathcal{P}_{s,a} = \{p \in \Delta^3 \; : \; \|\bar{p} - p\|_1 \leq 0.2\}$
- Robust MDPs can be solved fast in polynomial time by VI, PI, ...

Ambiguity Sets as Confidence Regions

- **Prior RMDP methods**: To get 80% confidence in return, construct 80% confidence/credible region for $P^\star$.

Ambiguity sets for $p^\star \in \Delta^3$ around mean $\bar{p}$



Frequentist confidence region | Bayesian credible region

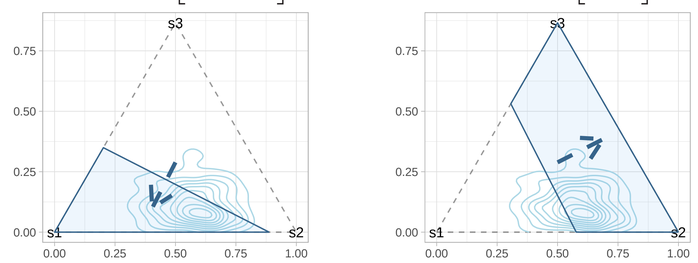
- **Bayesian credible region**: *Ambiguity set* as 80% credible region: ψ such that 80% of p_i satisfy: $\|p_i - \bar{p}\|_1 \leq \psi$

$$\hat{v}(s_0) = \min_{p\in\Delta^3}\{v^\mathsf{T} p : \|\bar{p} - p\|_1 \leq 0.8\} = 2.1$$

Credible regions lead to weak bounds (same posteriors!)

Optimal Bayesian Ambiguity Set: Known v

$v(s_1, s_2, s_3) = \begin{bmatrix}10 & 5 & -1\end{bmatrix}$ $v(s_1, s_2, s_3) = \begin{bmatrix}0 & 1 & 1\end{bmatrix}$



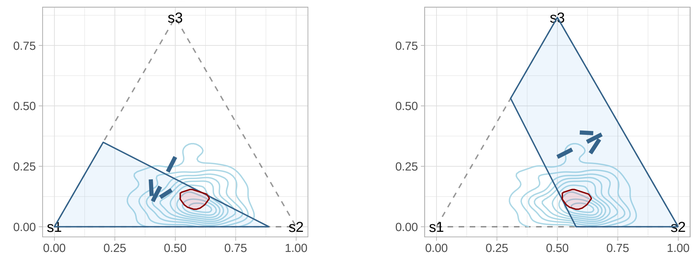
- **Bayes optimal set**: Halfspace $\hat{v}(s_0) = \min_{p\in\mathcal{P}^\star_{s_0,a}} v^\mathsf{T} p = 5.8$

Optimal Bayesian Ambiguity Sets: Unknown $\hat{v}$

- **Observation**: Robust value $\hat{v}$ is **not random** in Bayesian setting

Red set: Intersect half-spaces for all possible value functions

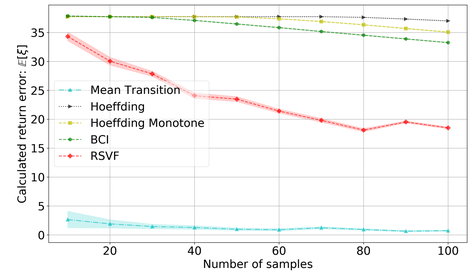
$v(s_1, s_2, s_3) = \begin{bmatrix}10 & 5 & -1\end{bmatrix}$ $v(s_1, s_2, s_3) = \begin{bmatrix}0 & 1 & 1\end{bmatrix}$



- **Theorem**: Intersection (red set) of half-spaces is *optimal* percentile criterion ambiguity set iff $\mathrm{V@R}_i[v^\mathsf{T} p_i]$ is convex in v [Gupta, 2015]
- $\mathrm{V@R}$ is non-convex in general but convex for some distributions
- **RSVF**: New method that computes a *tight outer approximation* of the (red) optimal ambiguity set

Experimental Evaluation: Bound Tightness

Riverswim: Uninformative Dirichlet Prior:



Exponential Population Model: Informative Gaussian Prior