Tight Bayesian Ambiguity Sets for Robust MDPs

Reazul H. Russel and **Marek Petrik**

Why Robustness in Reinforcement Learning

- **Batch RL**: Learn from logged data
- Limited data leads to uncertain transition probabilities
- Brittle policies fail when deployed
- Unacceptable **risk** in high-stakes domains: medicine, industry, ...

Why Robustness in Reinforcement Learning

- **Batch RL**: Learn from logged data
- Limited data leads to uncertain transition probabilities
- Brittle policies fail when deployed
- Unacceptable **risk** in high-stakes domains: medicine, industry, ...

- Compute **robust** policies without being too **conservative**?
  - Optimize **size** and **location** of ambiguity sets in robust MDPs using (hierarchical) Bayesian models

Robust Reinforcement Learning

- Batch of domain samples (log data, no simulator):
  $s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_n, a_n, r_n$

- **Robust policy** π: Guarantee lower bound on **true** return $\rho_{\mathsf{true}}(\pi)$ when deployed

Robust Reinforcement Learning

- Batch of domain samples (log data, no simulator):
  $s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_n, a_n, r_n$

- **Robust policy** π: Guarantee lower bound on **true** return $\rho_{\text{true}}(\pi)$ when deployed

- **Approach**: Estimate return $\rho_{\text{estim}}(\pi)$ of π such that:
  1. Lower bound: $\rho_{\text{estim}}(\pi) \leq \rho_{\text{true}}(\pi)$
  2. Tractable: $\max_\pi \rho_{\text{estim}}(\pi)$
- Solve $\max_\pi \rho_{\text{estim}}(\pi)$

Robust Estimate of Policy Return

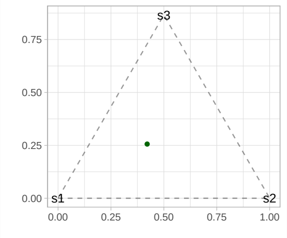
- Use **rectangular robust MDPs** ($\rho_{\text{estim}}(\pi) = p_0^T v_\pi^R$):

$$v^R(s) = \max_a \min_{p_{s,a} \in \mathcal{P}_{s,a}} \left(r_{s,a} + \gamma \cdot p_{s,a}^T v^R \right)$$

- Ambiguity set: $\mathcal{P}_{s,a} = \{ p \in \Delta^s : \|p - \bar{p}_{s,a}\|_1 \leq \psi_{s,a} \}$
- $\approx$ principled regularization

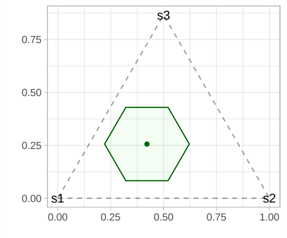
MDP

$p_{s,a} = [0.4, 0.2, 0.2]$



Robust MDP

$\bar{p}_{s,a} = [0.4, 0.2, 0.2], \psi_{s,a} = 0.4$

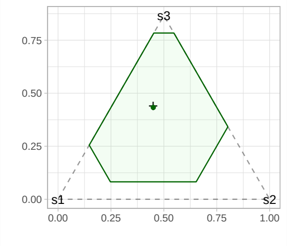


Research Challenge: Data-driven Ambiguity Sets

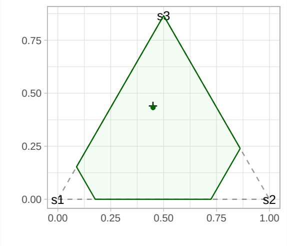
- Too small: not robust, too large: very conservative
- **Standard approach**: Concentration inequality around the max **likelihood estimate** (UCRL, . . .)

Guarantee $\rho_{\text{estim}}(\pi) \leq \rho_{\text{true}}(\pi)$ with

30% confidence



90% confidence

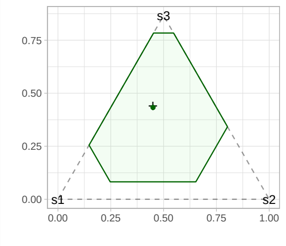


Research Challenge: Data-driven Ambiguity Sets

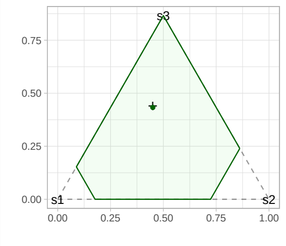
- Too small: not robust, too large: very conservative
- **Standard approach**: Concentration inequality around the max **likelihood estimate** (UCRL, ...)

Guarantee $\rho_{\text{estim}}(\pi) \leq \rho_{\text{true}}(\pi)$ with

30% confidence



90% confidence



Robust but too conservative to be practical!

Getting Robustness Right: Main Insights

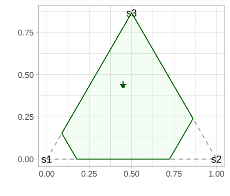
1. Capture prior knowledge using (hierarchical) Bayesian models
2. Optimize size and **location** of ambiguity sets
3. Ambiguity set need **not** be a **confidence interval** (similar to Gupta [2018])

Getting Robustness Right: Main Insights

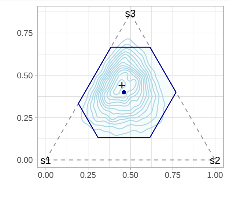
1. Capture prior knowledge using (hierarchical) Bayesian models
2. Optimize size and **location** of ambiguity sets
3. Ambiguity set need **not** be a **confidence interval** (similar to Gupta [2018])

Guarantee $\rho_{\text{estim}}(\pi) \leq \rho_{\text{true}}(\pi)$ with 90% confidence

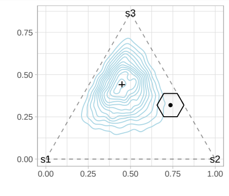
Concentration inequality set



Bayesian credible (confidence) set



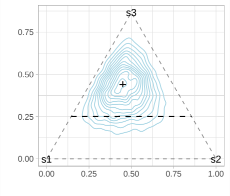
Bayesian **optimized** ambiguity set



RSVF: Optimizing Bayesian Ambiguity Sets

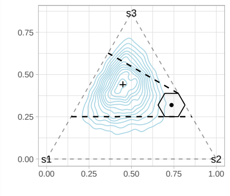
- Fixed value function v^R: Guarantee $\rho_{\text{estim}}(\pi) \leq \rho_{\text{true}}(\pi)$ if ambiguity sets **intersects a hyperplane**
- RSVF: Incrementally grow a set of plausible v^R values

1. Guess v^R



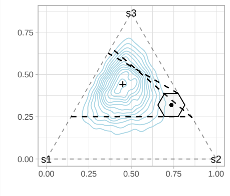
$v^R = [0, 0, 1]$

2...n: Recompute v^R



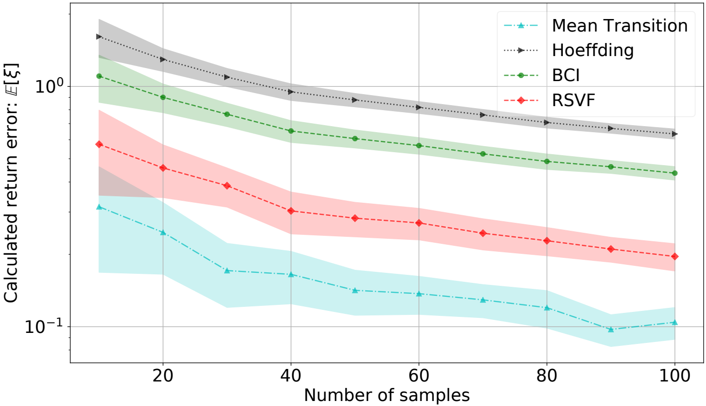
$v^R = [0, 0, 1]$ or $[2, 1, 0]$

n+1: Stop when robust



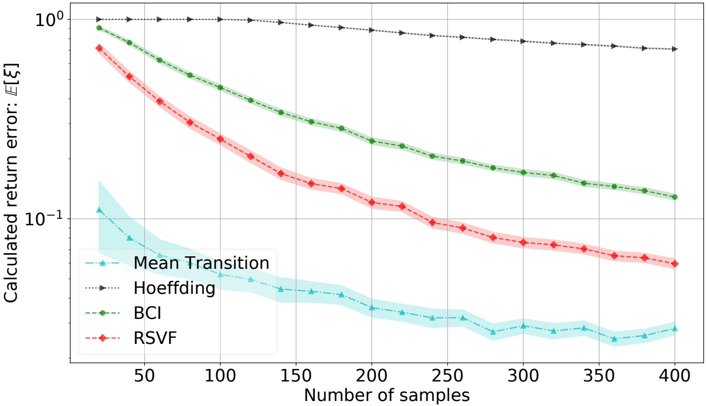
$v^R = [0, 0, 1]$ or $[2, 1, 0]$ or $[3, 1, 0]$

Uninformative Dirichlet Prior (95% confidence)



Smaller error means less conservative solution

Informative Hierarchical Prior (95% confidence)



Smaller error means less conservative solution

Conclusion

- Data-driven construction of robust ambiguity sets
  1. Capture prior knowledge using (hierarchical) Bayesian models
  2. Optimize size and **location** of ambiguity sets
  3. Ambiguity set need **not** be a **confidence interval**
- Pros:
  1. Robust but not too much
  2. Finite-sample guarantees
  3. Easy to define prior knowledge (e.g. Stan, PyMC)
- Cons:
  1. Increased computational complexity

Conclusion

- Data-driven construction of robust ambiguity sets
  1. Capture prior knowledge using (hierarchical) Bayesian models
  2. Optimize size and **location** of ambiguity sets
  3. Ambiguity set need **not** be a **confidence interval**
- Pros:
  1. Robust but not too much
  2. Finite-sample guarantees
  3. Easy to define prior knowledge (e.g. Stan, PyMC)
- Cons:
  1. Increased computational complexity

Thank you