Robust Exploration with Tight Bayesian Plausibility Sets

Reazul Hasan Russel, Tianyi Gu, Marek Petrik

University of New Hampshire

Summary

- **Markov Decision Processes (MDPs)** provide a powerful framework for modeling sequential decision problems under uncertainty.
- **Exploration** of poorly understood states and actions is important for long-term planning and optimization.
- **Optimism in the face of uncertainty (OFU)** is the main driving force of exploration for many RL algorithms.
- We propose **optimism in the face of sensible value functions (OFVF)**- a novel *data-driven* Bayesian algorithm to constructing *Plausibility* sets for exploration in MDPs.

Contribution

- OFVF Computes policies with tighter optimistic estimates for exploration by introducing two new ideas:
  1) It is based on Bayesian posterior distributions.
  2) It uses the structure of the value function to optimize the *location* and *shape* of the plausibility set.
- We showed that, OFU algorithms can be useful and can be competitive to stochastically optimistic algorithms like PSRL.

Problem Statement

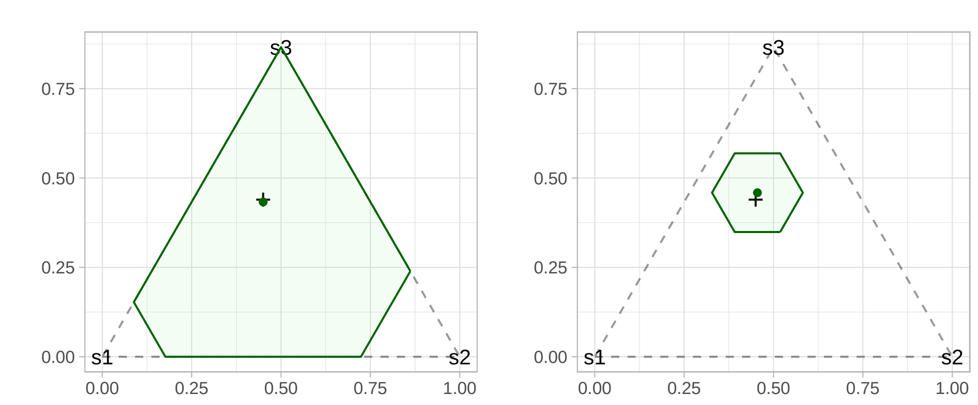
- Finite horizon Markov Decision Process $\mathcal{M}$ with states $\mathcal{S} = \{1, \ldots, S\}$ and actions $\mathcal{A} = \{1, \ldots, A\}$.
- $p_{s,a} : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$ for state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$.
- $R^a_{ss'}$ is reward for taking action $a \in \mathcal{A}$ from state $s \in \mathcal{S}$ and reaching state $s' \in \mathcal{S}$.
- A policy $\pi = (\pi_0, \ldots, \pi_{H-1})$ is a set of functions mapping a state $s \in \mathcal{S}$ to an action $a \in \mathcal{A}$.
- A value function for a policy π as:
$$V^\pi_h(s) := \sum_{s'} P^{\pi(s)}_{ss'}[r_h + V(s')]$$
- **Plausibility set** $\mathcal{P}$: set of possible transition kernels p.

Contact Information

- {rrussel, gu, mpetrik}@cs.unh.edu

Plausibility Sets

- **L_1-constrained (s, a)-rectangular ambiguity set** for state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ is defined as:
$$\mathcal{P}_{s,a} = \{\boldsymbol{p} \in \Delta^S : \|\boldsymbol{p} - \bar{\boldsymbol{p}}_{s,a}\|_1 \le \psi_{s,a}\}.$$
Note: $\bar{\boldsymbol{p}}_{s,a}$ is the **nominal** transition probability.

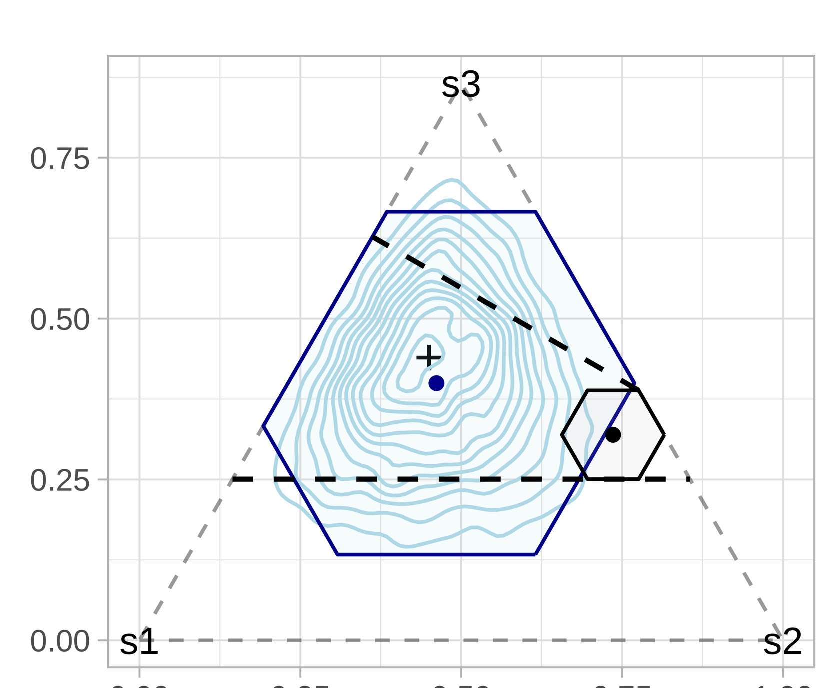


Ambiguity sets with $\psi_{s,a} = 0.5$ (left), and $\psi_{s,a} = 0.15$ (right).

- L_1-norm bounded plausibility set is constructed using Hoeffding's inequality
$$\psi_{sa} = \left\{ \|\tilde{p}_{sa} - \bar{p}_{sa}\|_1 \le \sqrt{\frac{2}{n_{s,a}} \log \frac{SA2^S}{\delta}} \right\}$$
- Bayesian plausibility sets are optimized for the smallest credible region around the mean transition
$$\min_{\psi \in \mathbb{R}_+} \left\{ \psi \ : \ \mathbb{P}\left[\|\tilde{p}_{s,a} - \bar{p}_{s,a}\|_1 > \psi \mid \mathcal{D} \right] < \delta \right\} ,$$

OFVF

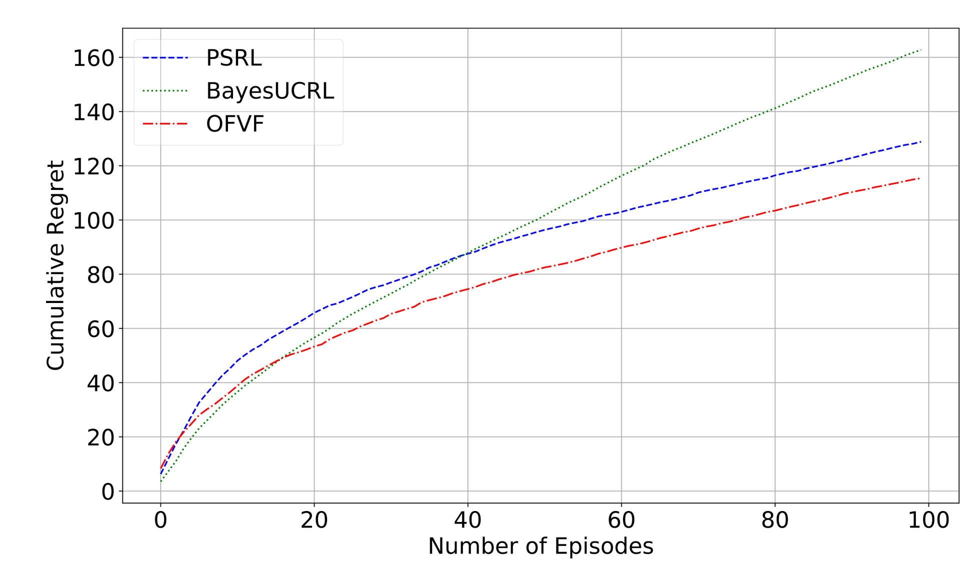
- Optimistic algorithms solve an optimistic version of Bellman update:
$$V^\star_h(s,a) := \max_{p_{sa} \in \mathcal{P}_{sa}} \sum_{s'} p^{\pi(s)}_{ss'}[r_h + V^\star(s')]$$
- OFVF uses samples from a posterior distribution and computes an optimal plausibility set for a singleton $\mathcal{V}$ as:
$$g = \max\left\{ k \ : \ \mathbb{P}_{P^\star}[k \le v^{\mathsf{T}} p^\star_{s,a}] \ge 1 - \delta/(SA) \right\}$$
- For $\mathcal{V} = \{v_1, v_2, \ldots, v_k\}$, OFVF solves the following linear program:
$$\psi_{s,a} = \min_{p \in \Delta^S} \left\{ \max_{i=1,\ldots,k} \|q_i - p\|_1 \ : \ v_i^{\mathsf{T}} q_i = g_i^\star, \ q_i \in \Delta^S, i \in 1, \ldots, k \right\}$$
- OFVF constructs the plausibility set to minimize its radius while still intersecting the hyperplane for each v in $\mathcal{V}$.



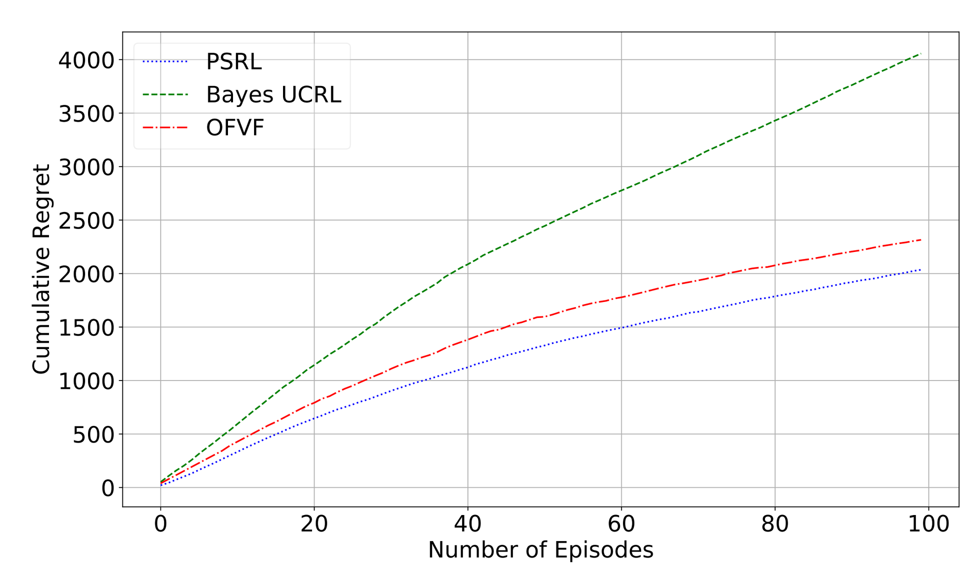
Plausibility sets constructed with Bayesian and OFVF.

Empirical Evaluation

- We evaluate the performance in terms of worst-case *cumulative regret* incurred by the agent upto time T for a policy $\pi^\star_l$:
$$\sup \left[\sum_{s \in \mathcal{S}} p_0(s) \left(V^\star(s) - V^{\pi^\star_l}(s) \right) \right]$$
- We compare OFVF with BayesUCRL and OFVF.



(a) Worst-case cumulative regret for Single state problem



(b) Worst-case cumulative regret for RiverSwim Problem

Conclusion

Empirical results demonstrate that: OFVF outperforms other OFU algorithms like *UCRL* [1]. Rectangularity assumption of OFVF leads to over optimism and PSRL [2] can stand out with the advantage of not having that.

Acknowledgments

This project was supported by NSF under awards No. 1815275, and 1717368.

References

[1]Thomas Jaksch, Ronald Ortner, and Peter Auer. *Near-optimal Regret Bounds for Reinforcement Learning.* Journal of Machine Learning Research, 11(1):1563–1600, 2010.

[2]Ian Osband, Daniel Russo, and Benjamin Van Roy. *(More) Efficient Reinforcement Learning via Posterior Sampling.* Neural Information Processing Systems (NIPS), 2013.