Robust Exploration with Tight Bayesian Plausibility Sets
Reazul Hasan Russel, Tianyi Gu, Marek Petrik

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.

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

- $L_1$-norm bounded plausibility set is constructed using Hoeffding’s inequality:
  \[ \psi_{s,a} = \left\| \bar{p}_{s,a} - \tilde{p}_{s,a} \right\|_1 \leq \sqrt{\frac{2}{n_{s,a}} \log \frac{SA^2}{\delta}}. \]

- Bayesian plausibility sets are optimized for the smallest credible region around the mean transition:
  \[ \min_{\tilde{p} \in \mathbb{R}} \left\{ \psi : P \left( \| p_{s,a} - \tilde{p}_{s,a} \|_1 > \psi \mid \mathcal{D} \right) < \delta \right\}. \]

OFVF
- Optimistic algorithms solve an optimistic version of Bellman update:
  \[ V_h^* (s, a) := \max_{p \in \mathcal{P}_{s,a}} \sum_{s'} P_{s,a}^h (s' \mid p, V^*(s')) \]

- OFVF uses samples from a posterior distribution and computes an optimal plausibility set for a singleton $\mathcal{V}$ as:
  \[ g = \max_k \left\{ k \mid \mathbb{P}_{s,a} (k \leq v^* p_{s,a}) \geq 1 - \delta / (SA) \right\} \]

- For $\mathcal{V} = \{ v_1, v_2, \ldots, v_k \}$, OFVF solves the following linear program:
  \[ \psi_{s,a} = \min_{p \in \Delta S} \left\{ \sum_{i=1}^k \| q_i - p \|_1 : v_i^* q_i = g_i^*, \quad g_i \in \Delta S, i = 1, \ldots, k \right\} \]

- OFVF constructs the plausibility set to minimize its radius while still intersecting the hyperplane for each $v$ in $\mathcal{V}$.

Empirical Evaluation
- We evaluate the performance in terms of worst-case cumulative regret incurred by the agent until time $T$ for a policy $\pi^*$:
  \[ \sup_{s \in S} \left[ \sum_{t=1}^T p_t(s) (V^* (s) - V^T (s)) \right]. \]

- We compare OFVF with BayesUCRL and PSRL.

Problem Statement
- Finite horizon Markov Decision Process $\mathcal{M}$ with states $S = \{ 1, \ldots, S \}$ and actions $A = \{ 1, \ldots, A \}$.
- $p_{s,a} : S \times A \rightarrow \Delta S$ for state $s \in S$ and action $a \in A$.
- $R_{s,a}$ is reward for taking action $a \in A$ from state $s \in S$ and reaching state $s' \in S$.
- A policy $\pi = (\pi_0, \ldots, \pi_{H-1})$ is a set of functions mapping a state $s \in S$ to an action $a \in A$.
- A value function for a policy $\pi$ as:
  \[ V_h^*(s) := \sum_{s'} P_{s,a}^h (s' \mid p, V^*(s')) \]

- Plausibility set $\mathcal{P}$. set of possible transition kernels $p$.

Contact Information
{russel, gu, metrik}@cs.unh.edu

Empirical Evaluation
(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.