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:
  $$\mathcal{P}_{s,a} = \{ p \in \Delta^S : \| p - \bar{p}_{s,a} \|_1 \leq \psi_{s,a} \}.$$  
  Note: $\bar{p}_{s,a}$ is the nominal transition probability.

Plausibility 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
  $$\psi_{s,a} = \left\{ \| \bar{p}_{s,a} - \bar{p}_{s,a} \|_1 \leq \sqrt{\frac{2}{n_{s,a}} \log SA^2} \right\}$$

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

OFVF

- Optimistic algorithms solve an optimistic version of Bellman update:
  $$V^\pi_h(s,a) := \max_{p \in \mathcal{P}_{s,a}} \sum_{s'} P_{s,a}(s') (r_{s,a} + V^\pi_{s'}(s'))$$

- OFVF uses samples from a posterior distribution and computes an optimal plausibility set for a singleton $\mathcal{V}$ as:
  $$g = \max_{k \in \mathcal{V}} \left\{ k : \mathbb{P}_{\mathcal{P}_{s,a}}[k \leq v^{s,a}] \geq 1 - \delta/(SA) \right\}$$
  For $\mathcal{V} = \{v_1,v_2,\ldots,v_l\}$, OFVF solves the following linear program:
  $$\psi_{s,a} = \min_{p \in \Delta^S} \left\{ \max_{1 \leq i \leq k} \{ |q_i - p|_1 : v_i q_i = g_i^0, \right\}$$

- 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 up to time $T$ for a policy $\pi^*$:
  $$\sup_{s \in S} \left\{ \sum_{h=0}^T p_h(s)(V^*(s) - V^{\pi^*}(s)) \right\}$$

- We compare OFVF with BayesUCRL and OFVF.

Problem Statement

- Finite horizon Markov Decision Process $\mathcal{M}$ with states $S = \{1, \ldots, S\}$ and actions $A = \{1, \ldots, A\}$.
  $$\bar{p}_{s,a} : S \times A \rightarrow \Delta^S$$ 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:
  $$V^\pi_h(s) := \sum_{s'} P_{s,a}(s') [r_{s,a} + V^\pi_h(s')]$$

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

Contact Information

{russel, gu, mpetrik}@cs.unh.edu

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
