OFVF

Anonymous Authors[1]

Abstract

This document provides a basic paper template and submission guidelines. Abstracts must be a single paragraph, ideally between 4–6 sentences long. Gross violations will trigger corrections at the camera-ready phase.

1. Introduction

Markov decision processes (MDPs) provide a versatile methodology for modeling dynamic decision problems under uncertainty. MDPs assume that transition probabilities are known precisely, but this is rarely the case in reinforcement learning. Errors in transition probabilities often results in probabilities often results in policies that are brittle and fail in real-world deployments. The agent has to learn the true dynamics of the MDP as it optimize the performance while interacts with its environment. The key to evaluate RL algorithms is to check how they balance between exploration that gains information about unknown states (actions) and exploitation to achieve near-term performance.

OFU-RL

Posterior sampling

Our work

2. Problem formulation

We consider the problem of learning and solving an uncertain MDP :(S, A, P^M, R^M) where $S = \{1, ..., S\}$ is the state space, $A = \{1, ..., A\}$ is the action space, $R^M(a, s)$ is the believe distribution over true reward when take action a at state s, $P^M(s'|s, a)$ is the believe distribution over the true transition probability of transitioning to state s' when take action a at state s.

value function

regret definition

3. OFVF and Bayes UCRL

Algorithm 1: Bayesian Confidence Interval (BCI)

```
Input: Distribution $\theta$ over $p^\star_{s,a}$, confidence level $\delta$, sample
       count $m$
Output: Nominal point $\bar{p}_{s,a}$ and $L_1$ norm size $\psi_{s,a}$
Sample $X_1, \ldots, X_m \in \Delta^S$ from $\theta$: $X_i \sim \theta$;
Nominal point: $\bar{p}_{s,a} \leftarrow (1/m) \sum_{i=1}^m X_i$;
Compute distances $d_i \leftarrow \bar{p}_{s,a} - X_{i1}$ and sort increasingly;
Norm size: $\psi_{s,a} \leftarrow d_{(1-\delta)\,m}$;
return $\bar{p}_{s,a}$ and $\psi_{s,a}$;
```

Algorithm 2: Bayes UCRL

```
Input: Desired confidence level $\delta$ and prior distribution
Output: Policy with an optimistic return estimate
repeat
    Initialize MDP: $M$;
    Compute posterior: $\tilde{p} \leftarrow$ compute_posterior(prior,
      samples) ;
    foreach $s \in \mathcal{S}, a \in \mathcal{A}$ do
        $\bar{p}_{s,a}, \psi_{s,a} \leftarrow$ Invoke Algorithm 1 with $\tilde{p}$, $\delta$;
        $M \leftarrow addtransitionwith \bar{p}_{s,a}, \psi_{s,a}$;
    Compute policy by solving MDP: $\hat{\pi} \leftarrow$ Solve $M$;
    Collect samples by executing the policy: samples $\leftarrow$
      execute $\hat{\pi}$;
    prior $\leftarrow$ posterior;
until num episodes;
return $(\pi_k, p_0^\mathsf{T} v_k)$ ;
```

Description about Bayes UCRL

pseudocode of OFVF and description

4. Some shortcomings of existing UCRL2 and PSRL

PSRL only have bound on Bayes Regret

UCRL2 have bound on regular Regret but loose

OFVF have better bound on regular Regret.

OFVF have better performance on worst case scenario.

On average case, OFVF would require less samples to produce same performance.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

5. Analysis

Theoretical proof:

Definition 1

Theorem 1

Lemma 1

Conjecture 1

6. Simulation results

something about RiverSwim

something about Inventory

something about MountainCar

7. Conclusion

Summarize the paper

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

A. Do *not* have an appendix here

Do not put content after the references. Put anything that you might normally include after the references in a separate supplementary file.

We recommend that you build supplementary material in a separate document. If you must create one PDF and cut it up, please be careful to use a tool that doesn't alter the margins, and that doesn't aggressively rewrite the PDF file. pdftk usually works fine.

Please do not use Apple's preview to cut off supplementary material. In previous years it has altered margins, and created headaches at the camera-ready stage.

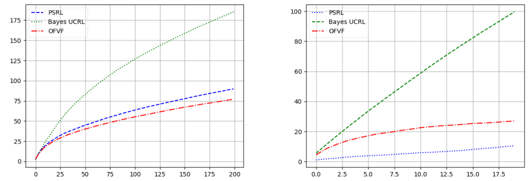

Figure 1. Cumulative regrets of PSRL and Bayes UCRL: left) above described simple problem, right) RiverSwim Problem described in (**?**)