Information Directed Reinforcement Learning

Andrea Zanette, Rahul Sarkar

Institute for Computational and Mathematical Engineering, Stanford University

Abstract

We consider the problem of minimizing the expected cumulative regret in an undiscounted episodic MDP using a model-based Bayesian framework. We propose a practical algorithm that aims at quantifying the cost of exploration by relating the expected regret to the variance of the policies over the posterior distribution. This approach is shown to outperform state-of-the art exploration strategies like Posterior Sampling Reinforcement Learning on numerical experiments.

Introduction

Exploration is widely acknowledged as a key difficulty in Reinforcement Learning. In this work we consider the exploration problem in an episodic undiscounted Markov Decision Process (MDP). The goal is to minimize the cumulative regret defined as the difference between the maximum possible expected reward and that accumulated by the agent. The dominant paradigms for efficient exploration in Reinforcement Learning are:

- Probability Matching
- Optimism in the face of Uncertainty

Recently, a new approach called Information-Directed Sampling (IDS) was proposed in [1]. The idea of IDS for the Bandit problem is to minimize the expected "cost" of acquiring information about the optimal action, i.e., it measures the cost of exploration. However such an approach may be computationally intractable even for the Bandit problem. Our goal is to capture the main idea of IDS and extend it to Reinforcement Learning with a practical algorithm.

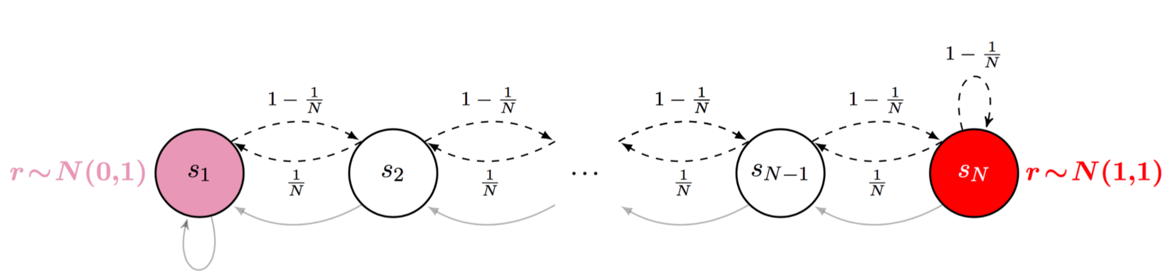

Figure: An MDP that requires efficient exploration

Motivating Idea

Posterior Sampling for Reinforcement Learning (PSRL) is very sample-efficient because it is optimistic "in the right amount". Our idea is to improve PSRL by introducing some bias towards policies μ with low *Information Ratio*. The Information Ratio is defined as the ratio between the expected regret in the next episode and the information gained about the optimal policy:

$$\text{InfoRatio}(\mu) = \frac{\text{Regret}(\mu)^2}{\text{InfoGain}(\mu)} \leq \frac{\text{Regret}(\mu)^2}{2\text{Variance}(\mu)}$$

where the inequality follows from Pinsker's inequality. The Information Ratio measures the cost of exploration: the agent is willing to incur a higher regret if the policy is informative, i.e., it has high variance. By minimizing the (empirical) Information Ratio at the beginning of each episode we hope to reduce the overall cost of exploration.

Algorithm

Information Directed Reinforcement Learning (IDRL) proceeds as follows:

Information-Directed Reinforcement Learning

1: **Input:** Prior distribution
2: **for** episode $t = 1$ **to** T **do**
3: Sample k MDPs from the Posterior
4: Compute Optimal Policies $\mu_{1,\dots,k}$
5: **for** Policy $\mu_{1,\dots,k}$ **do**
6: Compute Expected Regret $\Delta_{1,\dots,k}$
7: Compute Policy Variance $\sigma_{1,\dots,k}$
8: **end for**
9: Select the Policy that Minimizes $\frac{\Delta}{\sigma}$
10: Execute Policy in the Environment
11: Update Posterior Distribution
12: **end for**

Numerical Results

Multiarmed Bandit Problem (1000 Episodes)

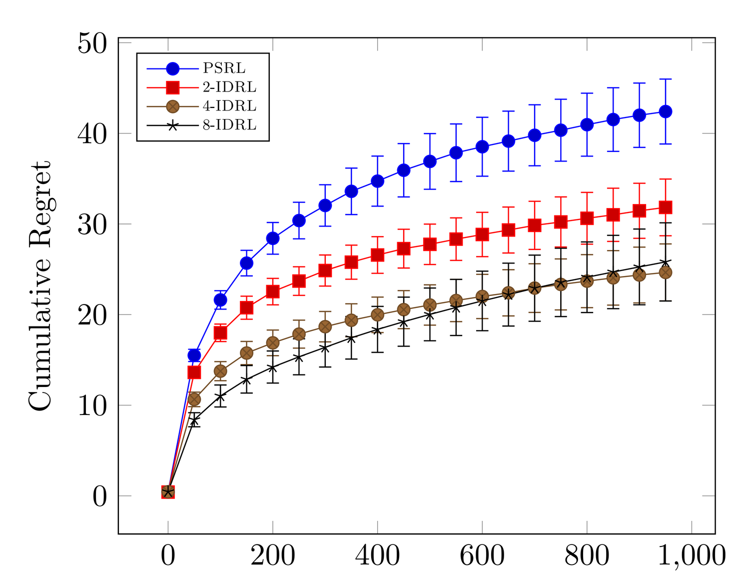

Figure: 20 Actions

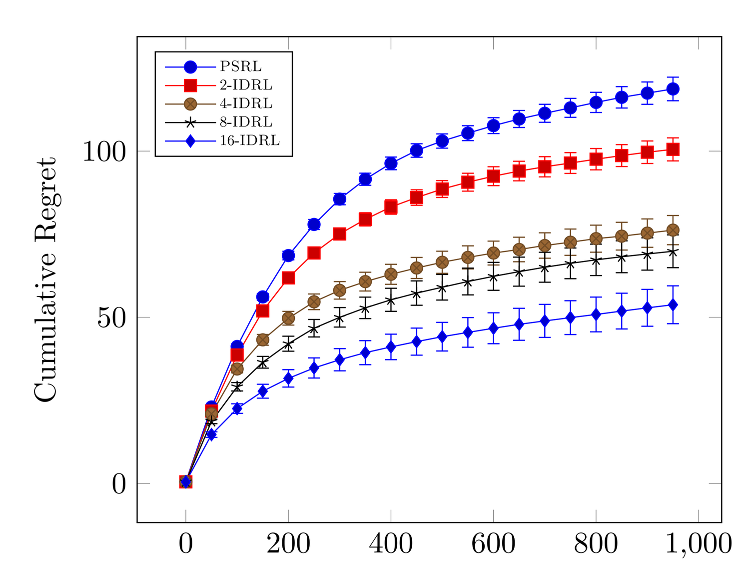

Figure: 100 Actions

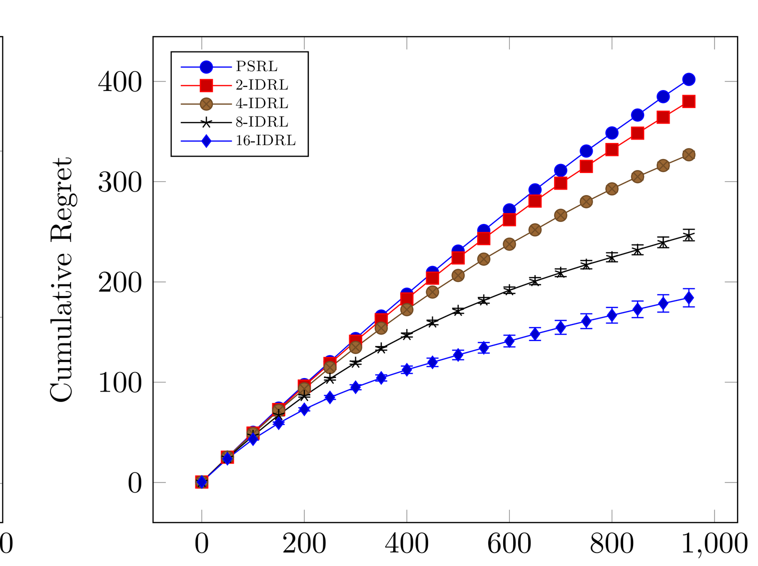

Figure: 1000 Actions

Chain Environment (Time Horizon = Number of States, 2000 Episodes)

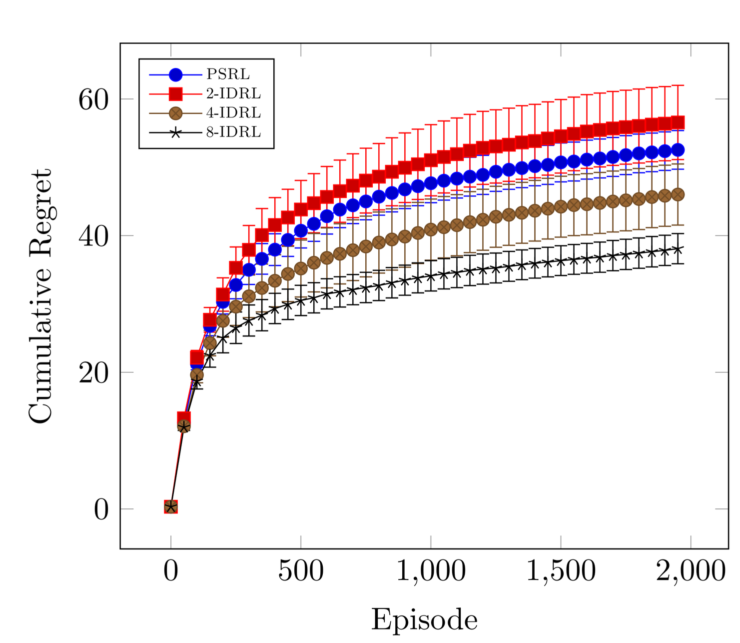

Figure: 4 States

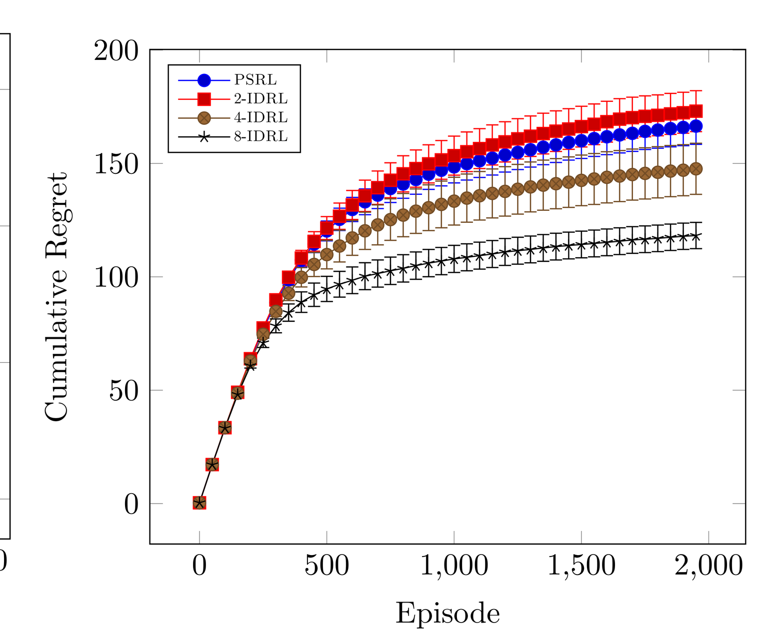

Figure: 8 States

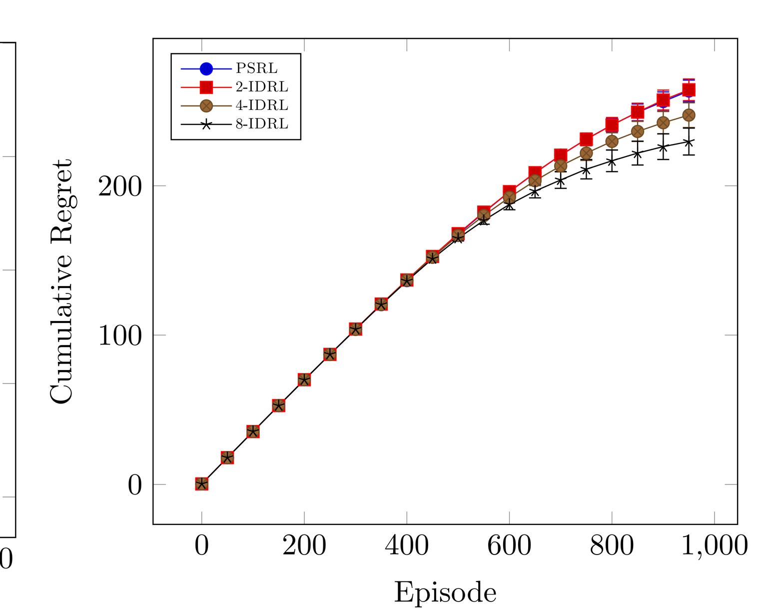

Figure: 12 States

Why does it Work?

Our approach exploits more aggressively than PSRL. This reduces the cumulative regret especially when a short time horizon does not allow extensive exploration of the state-action space.

Theorem (Informal)

Assume that the transition probabilities for an episodic undiscounted MDP are known. Let S be the number of states, A the number of actions, H the time horizon and T the number of episodes. If the information ratio is minimized exactly then the expected cumulative regret for IDRL is upper bounded by

$$\mathcal{O}\left(S\sqrt{HA\log(A)T}\right)$$

for any prior distribution of the rewards.

Future Work

As future work, we would like to obtain bounds for the Bayesian regret without the assumption that the transition probabilities are known.

Acknowledgements

We would like to thank Prof. Emma Brunskill for being our project mentor and for her valuable suggestions on our project.

References

[1] Daniel Russo, Benjamin Van Roy. *Learning to Optimize Via Information-Directed Sampling.* NIPS, 2014.

[2] Ian Osband, Benjamin Van Roy *Why is Posterior Sampling Better than Optimism for Reinforcement Learning.* EWRL, 2016