Information Directed reinforcement learning

Andrea Zanette [1] **Rahul Sarkar** [1]

Abstract

Efficient exploration is recognized as a key difficulty in reinforcement learning. We consider an episodic undiscounted MDP where the goal is to minimize the sum of regrets over different episodes. Classical methods are either based on optimism in the face of uncertainty or on probability matching. In this project we explore an approach that aims at quantifying the cost of exploration while remaining computationally tractable.

1. Introduction

In this project, we consider the classical reinforcement learning (RL) problem of an agent interacting with an environment modeled as a Markov decision process (MDP), which is unknown to the agent. As the agent interacts with the world, it learns about the unknown dynamics of the MDP. The goal of the agent is to maximize the expected sum of rewards over time. This naturally leads to a compromise between exploration and exploitation. In order to gain information about poorly understood states an agent must explore these states sacrificing immediate rewards. This tension is widely recognized as a fundamental challenge inherent in reinforcement learning (Sutton & Barto, 1998). Classical temporal-difference algorithms requires that every action be taken at every state infinitely often (Jaakkola, 1994), (Singh, 2000), (Watkins, 1992). To ensure this, a number of heuristic strategies has been developed in practice (Sutton & Barto, 1998).

Several approaches have been developed to address the exploration challenge in a sample-efficient manner. In order to avoid over reliance on point estimates of the system dynamics obtained by few samples, these approaches adopt the paradigm of optimism in the face of uncertainty. This gives a "bonus" to the reward attainable which is as high as statistically plausible. After that, an agent chooses a policy that is optimistic under this environment in order to promote exploration. As uncertainty is reduced, the "optimistic bonus" is reduced and the optimistic view of the world should converge to the underlying MDP. Some of these algorithms have very strong theoretical guarantees (T. Jaksch & Auer., 2010; Kearns & Singh, 2002; Brafman & Tennenholtz, 2003; Bartlett & Tewari, 2009).

Another efficient paradigm for efficient exploration is that of probability matching. Such agents model the unknown MDP with a prior distribution which is updated as the agents experience rewards and transition probabilities. Such idea dates back to the work of Thompson (Thompson, 1933) but recent empirical studies showed its strong numerical performance (Scott, 2010; Chapelle & Li, 2011). This prompted subsequent theoretical studies (Agrawal & Goyal, 2013a;b; E. Kauffmann & Munos, 2012; Daniel Russo, 2013) which provided strong theoretical guarantees for the bandit problem as well as for the reinforcement learning setting (Ian Osband, 2013; 2016) where such approach is commonly referred to as Posterior Sampling for Reinforcement Learning (PSRL).

More recently, a new paradigm for efficient exploration called Information-Directed Sampling (IDS) was proposed in (Daniel Russo, 2013) for the classical multiarmed bandit problem and for linear bandits. The idea is to quantify the cost of exploration via the information gained about the optimal action. Intuitively, the agent is willing to incur a larger instantaneous regret for an action that gives more information about which is the optimal arm to pull. The key quantity of such analysis is the so called information ratio, which is the ratio between the instantaneous square expected regret and the information gained about the optimal action. The hope is that by minimizing the information ratio at each episode the total cost of exploration (to identify the optimal arm) is reduced. The authors in (Daniel Russo, 2013) work with the notion of Bayesian regret which allows them to relate the per-episode expected instantaneous regret computed by the agent to the actual one. Such analysis assumes that the true "world" is sampled from the agent's prior distribution. Unfortunately, despite its elegant and strong theoretical bounds, IDS is difficult to implement mainly because of its computational intractability.

[1]Institute for Computational and Mathematical Engineering, Stanford University, CA. Correspondence to: Andrea Zanette <zanette@stanford.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP.

In this project, we try to to accomplish the following:

- develop a tractable algorithm that captures the main idea of IDS
- extend the core idea of IDS to RL

2. Preliminaries

In this section we recall the idea behind IDS as presented in (Daniel Russo, 2013). We mention that the same analysis can be used to provide an information-theoretic bound for Thompson sampling (Russo, 2016).

2.1. Bayesian Regret

Consider a one state MDP with $|A|$ actions, i.e., a Bandit problem, and define as expected instantaneous reward by pulling arm a:

$$\mathbb{E}_M\left(R_a|\mathcal{F}_t\right)$$

where the random variable M is the unknown single-state MDP and the expectation $\mathbb{E}(\cdot|\mathcal{F}_t)$ is conditioned on the filtration $\mathcal{F}_t$, which is the "history" of action-reward pairs experienced by the agent up to episode $t-1$. Now suppose the agent knew which one is the optimal arm a^* (but not the actual MDP). Conditioned on this information and on the filtration, denote the expected reward by pulling the optimal arm a^*:

$$\mathbb{E}_M\left(R_{a^*}|\mathcal{F}_t, a^*\right).$$

Uncertainty about the optimal action a^* at the current timestep t induces uncertainty about the maximum expected reward attainable by the agent, which can be estimated as:

$$\mathbb{E}_{a^*}\left(\mathbb{E}_M\left(R_{a^*}|a^*\right)|\mathcal{F}_t\right).$$

The authors in (Russo, 2016) define as Bayesian instantaneous regret $\Delta_t(a)$ the difference between the expected reward attainable if the optimal action a^* is known and the expected one attained by the agent that pulls arm a:

$$\Delta_t(a) \stackrel{\text{def}}{=} \mathbb{E}_{a^*}\left(\mathbb{E}_M\left(R_{a^*} - R_a\right)|\mathcal{F}_t, a^*\right). \tag{1}$$

Crucially, this quantity depends on the history observed (that is, the filtration $\mathcal{F}_t$) and it is in principle computable by the agent. By taking the expectation with respect to the filtration and summing over the arms pulled by the agent $\pi = (a_1, ..., a_T)$ up to time T one obtains:

$$\begin{aligned}\text{BayesRegret}(\pi) &\stackrel{\text{def}}{=} \\ &= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{F}_t}\,\Delta_t(a_t) \\ &= \sum_{t=1}^{T} \mathbb{E}_{a^*}\left(\mathbb{E}_M\left(R_{a^*} - R_a\right)|a^*\right)\end{aligned} \tag{2}$$

Equation 2 allows to relate the regret computable by the agent with the actual one, in expectation. This notion of regret assumes that the "true" MDP is sampled from the prior.

2.2. Information Gain

In principle, the IDS agent can compute the probability that an action is optimal at the beginning of each episode. A key quantity is the posterior distribution of the optimal action. Intuitively, as the agent interacts with the environment it updates its beliefs about the optimal action. In the long run, the agent's posterior probability about the optimal action should concentrate around the "true" optimal arm.

The information gain $I_t(a)$ from an action a is the expected reduction in the entropy of the optimal action. Intuitively, an action that can identify the optimal action will have higher $I_t(a)$.

2.3. Information Ratio

A key quantity used in the analysis and implementation of IDS is called information ratio (by pulling arm a) defined as:

$$\Psi(a) = \frac{\Delta_t(a)}{I_t(a)}$$

In (Daniel Russo, 2013) the authors show that, for any algorithm,

$$\text{BayesRegret} = \sum_{t=1}^{T} \mathbb{E}_{\mathcal{F}_t}\,\Delta_t \leq \sqrt{\bar{\Psi} H(\alpha) T} \tag{3}$$

by a simple application of Cauchy-Schwartz inequality. In the above expression, T is the number of episodes, $H(\alpha)$ is the initial entropy of the optimal action and $\bar{\Psi}$ is the average information ratio. This is valid for any learning algorithm; in particular if an algorithm can bound the average information ratio $\bar{\Psi}$ by a constant, then learning occurs upper bounded by $\mathcal{O}(\sqrt{T})$. It seems therefore natural to minimize $\bar{\Psi}$. Since minimizing the average information ratio over all episodes may be intractable, IDS myopically minimizes Ψ at the beginning of each episode.

3. EXACT-IDRL

We now discuss a direct extension of IDS to reinforcement learning. For reference we label it as EXACT-IDRL. We now define the notation, the equivalent notion of regret (for an episode), information gain, and information ratio in the reinforcement learning setting. Another notion to measure the quality of a learning agent in episodic fixed-horizon problems is provided in (Christoph Dann, 2016), but we do not consider this measure here.

3.1. Notation

We assume that the environment can be modeled as an episodic Markov Decision Process (MDP) $M = (S, A, R^M, \mathcal{P}^M, H)$, where S is the state space, A the action space, R^M the rewards as a function of state-action pairs, P^M the transition probabilities and H is the episode length. We denote with V_μ^M the value of the policy μ on MDP M, that is, the value function of the initial state obtained by following policy μ on MDP M. Denote with $\mathcal{F}_t = \{\{(s_{i,j}, a_{i,j}, r_{i,j}, s_{i+1,j})\}_{i=1,\dots,H}\}_{j=1,\dots,t-1}$ the filtration, i.e., the "history" experienced by the agent up to episode $t-1$. The filtration $\mathcal{F}_t$ consists of the sets of all rollouts in terms of states $s_{i,j}$ actions $a_{i,j}$ rewards $r_{i,j}$ and next states $s_{i+1,j}$ experienced by the agent in episode j at timestep i. Let $\varphi^*(\cdot)$ be the operator that acts on an MDP M and returns an optimal policy (with ties randomly broken).

3.2. Instantaneous Regret

At the beginning of episode t the agent chooses a policy μ to follow. The history $\mathcal{F}_t$ observed by the agent changes the prior distribution of the unknown MDP from the prior distribution $f(\cdot)$ to the posterior $f(\cdot|\mathcal{F}_t)$. Accordingly, if the agent knew the optimal policy μ^* then the posterior distribution of the unknown MDP would change to $f(\cdot|\mathcal{F}_t, \mu^*)$. It is natural to extend the definition of instantaneous regret to the RL setting using the value function V_μ^M of a given policy μ in MDP M:

Definition 1. *Define as instantaneous expected regret $\Delta_t(\mu)$ at timestep (t) by following policy (μ):*

$$\Delta_t(\mu) \stackrel{def}{=} \mathbb{E}_{\mu^*}\left(\mathbb{E}_M\left(V_{\mu^*}^M - V_\mu^M\right)|\mathcal{F}_t, \mu^*\right) \tag{4}$$

This notion of regret in equation 4 measures the expected loss by following policy μ instead of the optimal policy μ^* on each MDP and is a natural extension of that given in (Daniel Russo, 2013) to RL. Notice that this quantity is computable by the agent, at least in principle, as it only depends on the MDP prior and on the filtration $\mathcal{F}_t$ through $f(\cdot|\mathcal{F}_t)$.

3.3. Information Gain

Let α_t be the posterior distribution (conditioned on the filtration) of the optimal policy at the beginning of episode t. In other words, let $\alpha_t \in \mathbb{R}^{|A|^{H|S|}}$ be the probability distribution whose i-th component contains the probability that the i-th policy is optimal.

Definition 2. *Define the information gain $I_t(\mu)$ from policy (μ) over an episode:*

$$I_t(\mu) \stackrel{def}{=} \mathbb{E}_{rollout}\left(H(\alpha_t) - H(\alpha_{t+1})|\mathcal{F}_t, \mu\right) \tag{5}$$

In other words, $I_t(\mu)$ is the expected reduction in the entropy of α_t if policy μ is followed in the next episode. The expectation is taken with respect to the rollout experienced by the agent which depends on the posterior distribution of the unknown MDP $\sim f(\cdot|\mathcal{F}_t)$ and is thus computable by the agent.

3.4. Information Ratio

At the beginning of every episode the EXACT-IDRL agent acts according to a policy that seeks to minimize the information ratio described below. Denote with π the probability vector such that the i-th component is the probability that the the agent chooses the i-th policy. With some abuse of notation, define vector Δ_t such that the regret of the i-th policy is in the i-th component and likewise the vector of the information gain I_t. In reinforcement learning we seek to minimize the following quantity, called information ratio:

Definition 3. *Let Δ_t and I_t be the expected instantaneous regret vector and information gain vector, respectively. Define $\Psi(\pi)$ as information ratio obtained by following a policy sampled according to the distribution identified by π:*

$$\Psi(\pi) = \frac{\left(\pi^T \Delta_t\right)^2}{\pi^T I_t} \tag{6}$$

Notice that the argument π of the information ratio is a randomization over policies. For example, PSRL is a randomized algorithm which chooses a vector π such that the i-th component of π is the the probability the that the i-th policy is optimal. Then PSRL sample the policy index from π and follows that policy in the next episode. The idea behind IDS and EXACT-IDRL is to choose a π vector that minimizes the information ratio in order to reduce the cost of exploration. In other words, to select the next policy EXACT-IDRL solves the following optimization program:

$$\begin{aligned} \pi^* = \underset{\pi}{\text{argmin}} \quad & \frac{\left(\pi^T \Delta_t\right)^2}{\pi^T I_t} \\ \text{subject to} \quad & \|\pi\|_1 = 1 \\ & \pi \geq 0 \end{aligned} \tag{7}$$

where the minimization is over all $\pi \in \mathbb{R}^{|A|^{H|S|}}$. Notice that π^* is not a policy, but the probability vector that a policy is optimal. As in (Daniel Russo, 2013), the minimization problem 6 is convex and has a solution with at most two non-zero components. The policy selected by the EXACT-IDRL agent can be finally chosen by sampling from π^*. The same policy is then followed throughout the next episode. The conceptual algorithm is reported in algorithm 1.

Algorithm 1 EXACT-IDRL

Input: Prior distribution f
for episode $t = 1$ **to** T **do**
  Compute π_{IDRL} by solving 7
  Sample μ_{IDRL} according to the distribution π_{IDRL}
  for timestep $i = 1$ **to** H **do**
    Apply action $a = \mu_{IDRL}(s_i, i)$
    Observe reward r_i and nextstate s_{i+1}
  end for
end for

3.5. Bayesian Regret Bound

We immediately have the following observation:

Proposition 1. *Let the transition probabilities $\mathcal{P}(\cdot|s,a)$ be fixed and known to the agent, and assume the agent starts from the same state at the beginning of each episode. Assume the episodic undiscounted unknown MDP M^* is sampled from the agent's prior. Then* EXACT-IDRL *(algorithm 1) achieves a Bayesian expected regret upper bounded by*

$$|S|\sqrt{\frac{1}{2}H|A|\ln(|A|)T}$$

for any choice of the prior distribution of the rewards.

The proof is reported in the appendix. At a very high level, it proceeds as follows. Since the transition probabilities are fixed, it is possible to compute how often a state s is visited, in expectation, by following a fixed policy μ during an episode. While in state s at timestep i, policy μ gives a fixed action a. This allows to compute the expected number of times that a (random) reward $R(s,a)$ is experienced. It is then possible to recast this problem as a linear bandit problem and reuse previous analysis in (Daniel Russo, 2013) which upper bounds the information ratio by that of Thompson sampling to provide a bound on the Bayesian regret.

4. APPROXIMATE-IDRL

The minimization problem 7 is computationally intractable to solve in practice for the following reasons:

1. The space of the optimal policies has dimension $|A|^{H|S|}$, which is too large.
2. The computation of the expected regret Δ_t is also intractable because it requires averaging over infinitely-many MDPs.
3. The computation of the information gain involves computing the reduction in the entropy about the optimal policy, which is intractable.

In order to devise a practical algorithm we make the following approximations:

1. We reduce the number of policies considered at the beginning of each episode
2. We approximate the expected instantaneous regret
3. We approximate the information gain

For brevity, we call APPROXIMATE-IDRL the resulting algorithm. We now discuss all of the above points.

4.1. Restriction on the Space of the Optimization Variables

In an episodic reinforcement learning setting there are $|A|^{H|S|}$ policies, too many to be examined. Our idea is to limit the policies under consideration to a handful in order to retain computational tractability. Instead of sampling these policies uniformly at random we sample them according to the probability that they are optimal. To achieve this, we sample k MDPs from the posterior $f(\cdot|\mathcal{F}_t)$, form the set

$$\widetilde{\mathcal{M}} = \{M_j | M_j \sim f(\cdot|\mathcal{F}_t), j = 1, \dots, k\}$$

and solve for the optimal policy for each of these MDPs:

$$\Phi^* = \{\varphi^*(M_k) | M_k \in \widetilde{\mathcal{M}}\} \tag{8}$$

This gives a set Φ^* of policies to examine that are already very good because they are optimal under statistically plausible MDPs.

4.2. Empirical Instantaneous Regret

Recall the definition of expected instantaneous regret in equation 4. We proceed by doing a change of variables that may simplify the computation:

Lemma 1. *The following holds:*

$$\begin{aligned}\Delta_t(\mu) &\overset{def}{=} \mathbb{E}_{\mu^*}\left(\mathbb{E}_M\left(V_{\mu^*}^M - V_\mu^M\right)|\mathcal{F}_t, \mu^*\right) \\ &= \mathbb{E}_M\left(\max_{\mu^*} V_{\mu^*}^M - V_\mu^M |\mathcal{F}_t\right)\end{aligned} \tag{9}$$

The short proof is reported in the appendix. We use this fact to approximate the expected regret by computing the empirical mean of the value function V_μ^M of policy μ for the MDPs in $\widetilde{\mathcal{M}}$:

$$\widetilde{\Delta}_t(\mu) = \frac{1}{k}\sum_{M\in\widetilde{\mathcal{M}}}\left(\max_{\mu} V_{\mu^*}^M - V_\mu^M\right) \tag{10}$$

4.3. Empirical Information Gain

Now we focus on approximating the information gain of equation 5. As in (Daniel Russo, 2013), we can use Pinsker's inequality (with the assumption that the expected cumulated sum of rewards is between 0 and 1 during an episode) to obtain:

$$\begin{aligned} I_t(\mu) &\geq 2\,\mathbb{E}_{\mu^*}\left(\left(\mathbb{E}_M\left(V_\mu^M|\mu^*\right)-\mathbb{E}_M V_\mu^M\right)^2|\mathcal{F}_t\right) \\ &= 2\,\mathrm{Var}_{\mu^*}\left(\mathbb{E}_M\left(V_\mu^M|\mu^*\right)|\mathcal{F}_t\right) \\ &\overset{\text{def}}{=} 2g(\mu) \end{aligned} \tag{11}$$

where $\mathbb{E}_{\mu^*}$ is the expectation computed with respect to the unknown optimal policy μ^*, $\mathbb{E}_M$ is the expectation with respect to the random MDP $M \sim f(\cdot|\mathcal{F}_t)$ and $\mathbb{E}_M(\cdot|\mu^*)$ is the expectation with respect to the random MDP $M \sim f(\cdot|\mathcal{F}_t, \mu^*)$ conditioned on the fact that μ^* is optimal. Denote via g_t the vector whose i-th component contains the i-th policy. Following equation 11, for any $\pi \geq 0$, $\|\pi\|_1 = 1$ we have that:

$$\Psi(\pi) = \frac{\pi^T\Delta_t}{\pi^T I_t} \leq \frac{\pi^T\Delta_t}{\pi^T g_t}$$

and thus by replacing the information gain by the variance we have an upper bound on Ψ. The idea is to minimize the upper bound which is easier to compute. A bound on the right hand side automatically translates into a bound on the information ratio (left hand side).

The fact that the variance should be computed with respect to the optimal policy implies that we should group together the sampled MDPs which share the same optimal policy. Denote with $\widetilde{M}_{\mu^*} = \{M|\mu^* = \varphi(M)\}$ the set of sampled MDPs that have μ^* as optimal policy, and with $|\cdot|$ the size of a set. The empirical estimate corresponding to $g(\mu)$ can be written as:

$$\widetilde{g}(\mu) \overset{\text{def}}{=} \frac{1}{|\Phi^*|}\sum_{\mu^*\in\Phi^*}\left(\frac{1}{|\widetilde{M}|}\sum_{M\in\widetilde{\mathcal{M}}}V_\mu^M - \frac{1}{|\widetilde{M}_{\mu^*}|}\sum_{M\in\widetilde{M}_{\mu^*}}V_\mu^M\right) \tag{12}$$

Remark on the computation of the variance Notice that by Eve's law (or law of total variance) we have that :

$$\mathrm{Var}_M(V_\mu^M|\mathcal{F}_t) = \mathbb{E}_M\left(\left(V_\mu^M - \mathbb{E}_M V_\mu^M\right)^2|\mathcal{F}_t\right)$$

$$= \mathbb{E}_{\mu^*}\left(\mathrm{Var}_M\left(V_\mu^M|\mu^*\right)|\mathcal{F}_t\right) + \mathrm{Var}_{\mu^*}\left(\mathbb{E}_M\left(V_\mu^M|\mu^*\right)|\mathcal{F}_t\right)$$

$$= \mathbb{E}_{\mu^*}\left(\mathrm{Var}_M\left(V_\mu^M|\mu^*\right)|\mathcal{F}_t\right) + g(\mu).$$

This shows that the approximation

$$g(\mu) \leq \mathrm{Var}_M(V_\mu^M|\mathcal{F}_t) \overset{\text{def}}{=} v(\mu)$$

leads to:

$$\frac{\pi^T\Delta_t}{\pi^T g_t} \geq \frac{\pi^T\Delta_t}{\pi^T v_t}.$$

In this case, minimizing the right hand side does not give an upper bound on the information ratio. In other words, the variance of the value function should be computed by taking as random variable the unknown optimal policy μ^* instead of the unknown MDP M^*, leading to the empirical estimate in equation 12.

4.4. Empirical Information Ratio

Finally, we choose to minimize the information ratio using policies that are optimal over the MDPs considered in $\widetilde{\mathcal{M}}$, using the empirical regret and the empirical variance:

$$\begin{aligned} \widetilde{\pi}^* = \underset{\widetilde{\pi}}{\text{argmin}} \quad & \frac{\left(\widetilde{\pi}^T\widetilde{\Delta}_t\right)^2}{\widetilde{\pi}^T\widetilde{g}_t} \\ \text{subject to} \quad & \|\widetilde{\pi}\|_1 = 1 \\ & \widetilde{\pi} \geq 0. \end{aligned} \tag{13}$$

Solving 13 yields a probability distributions over policies in Φ^*. Sampling from $\widetilde{\pi}^*$ finally gives the policy to follow in the next episode.

The optimization program 13 entails a convex nonlinear objective function subject to linear constraints. Efficient methods exist for solving such problems. Furthermore, it can be shown, similarly to (Daniel Russo, 2013), that the solution $\widetilde{\pi}$ has at most two non-zero components and thus ad hoc methods can be designed to solve 13.

Intuitively, as the number of sampled MDPs increases, both $\widetilde{\Delta}_t$ and $\widetilde{g}_t$ should approximate Δ_t and g_t increasingly well, respectively.

This however does not imply that by solving the minimization problem 13 we get the same solution as the original problem 7. This happens because problem 13 can only produce a policy that is optimal under some MDP. This is in contrast to the original problem 7 where a non-optimal policy can be chosen for the purpose of reducing the uncertainty about the optimal policy.

The practical version of IDRL (algorithm 2) proceeds as follows: at the beginning of each episode it samples $\widetilde{n}$ MDPs and computes the optimal policies under each of the sampled MDPs. Then it approximately computes the expected regret $\widetilde{\Delta}_t$ and variance $\widetilde{g}_t$. Finally, it solves the minimization problem 13 and samples from that solution to obtain a policy to follow in the episode. Similar to PSRL, the policy does not change within the episode.

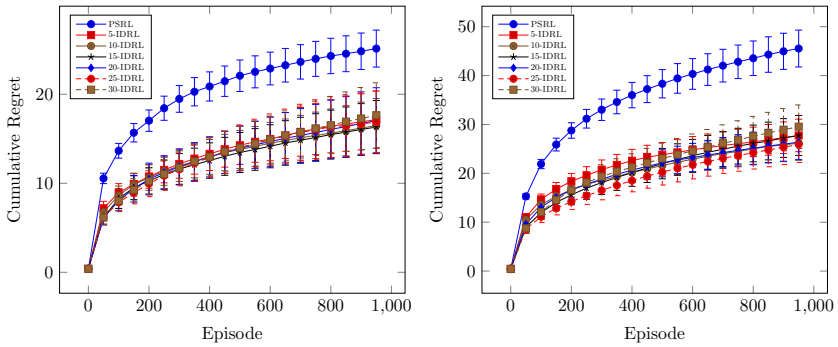

Figure 1. Bandit setting for $|A| \in \{10, 20\}$ respectively (left to right).

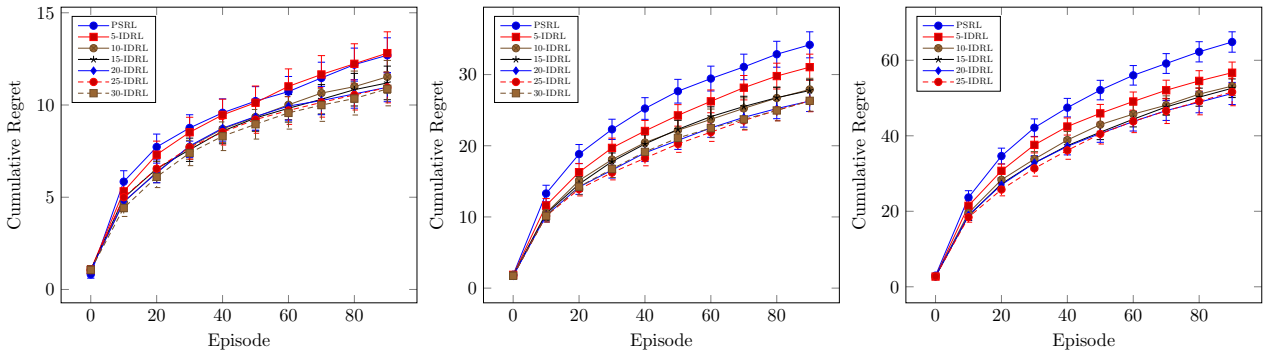

Figure 2. Random MDP setting for $|S| \in \{2, 4, 6\}$ states respectively (left to right).

Algorithm 2 APPROXIMATE-IDRL

Input: Prior distribution f, number of samples $\widetilde{n}$
for episode $t = 1$ **to** T **do**
  Sample $\widetilde{\mathcal{M}} = \{M_k \sim f(\cdot|\mathcal{F}_t), k = 1, ..., \widetilde{n}\}$
  Compute optimal policies $\Phi^* = \{\varphi^*(M_k)|M_k \in \widetilde{\mathcal{M}}\}$
  Compute $\{\widetilde{\Delta}(\mu_k)|\mu_k \in \Phi^*\}$ using equation 10
  Compute $\{\widetilde{g}(\mu_k)|\mu_k \in \Phi^*\}$ using equation 12
  Compute π_{IDRL} by solving 13
  Sample μ_{IDRL} according to the distribution π_{IDRL}
  for timestep $i = 1$ **to** H **do**
    Apply action $a = \mu_{IDRL}(s_i, i)$
    Observe reward r_i and nextstate s_{i+1}
  end for
end for

5. Numerical Experiments

We conduct two numerical experiments to test the APPROXIMATE-IDRL algorithm scheme. For simplicity we test it on a multiarmed bandit problem and on randomly generated MDPs. For each experiment we run 100 simulations. In figures 1,2 we report the mean and the 95% confidence intervals for the mean.

5.1. Multiarmed Bandits

We consider the classical multiarmed bandit problem with $|A| = \{10, 20\}$ arms. Pulling each arm i returns a reward whose distribution follows $\sim$ Bernoulli(p_i). We assume $p_i \sim$ Beta$(1, 1)$, that is a uniform prior distribution for the rewards. For each simulation we sample the bandits from the prior. The same prior is used in Thompson sampling and APPROXIMATE-IDRL. In figure 1, we plot the expected cumulative regret over $T = 1000$ episodes. The results of the numerical experiment are shown in Figure 1. For all cases, we can see from the figures that the APPROXIMATE-IDRL algorithm leads to much lower cumulative regret than PSRL. Also we notice that the cumulative regret decreases in general as $\widetilde{n}$ increases.

5.2. Random MDP

In the second case, we create a random MDP with $|S| = \{2, 4, 6\}$ states, and each state with $|A| = 2$ actions. The length of each episode is taken to be 100 time steps for this example. Taking each action can lead to any of the states with transition probabilities determined from a Dirichlet prior, while the rewards are sampled from a Normal Gamma distribution. However, once determined at the beginning of each episode, the rewards are deterministic

for each action from a state for the entire duration of the episode. The results of the numerical experiment are shown in Figure 2. We see from the figures that APPROXIMATE-IDRL achieves lower cumulative regret than PSRL in all cases.

6. Conclusion

In this project we have proposed a tractable extension of information directed sampling to reinforcement learning.

By extending the notion of Bayesian regret, information gain and information ratio to reinforcement learning, we derive a theoretical algorithm called EXACT-IDRL. For this theoretical algorithm we provide a Bayesian regret bound if the transition probabilities are known to the agent and the rewards are sampled from the agent's prior.

Unfortunately, EXACT-IDRL is computationally intractable and this motivates APPROXIMATE-IDRL which replaces the expected regret and information gain by sample estimates. The policies considered in the minimization of the information ratio are also carefully selected according to the probability that they are optimal.

One can also view APPROXIMATE-IDRL as an approach that starts with policies that PSRL would choose. Then it introduces some bias by relying on an approximately computed information ratio.

Our numerical results indicates that despite the heavy approximations, APPROXIMATE-IDRL performs as well as, and often better than PSRL on simple experiments.

However, several open issues remains. In particular, APPROXIMATE-IDRL needs to evaluate several policies (via policy iteration) on the sampled MDPs in order to compute the variance. This has quadratic cost with the number of samples.

Moreover, it is unclear whether any statement can be made about how the information ratio empirically computed compares with the true one which may be used to obtain Bayesian regret bounds.

It is our hope that we can continue working on this project to address the shortcomings of out approach, improve its numerical performance and better understand it theoretically.

Acknowledgment

The authors would like to thank Prof. Emma Brunskill for teaching a great class and for giving important advice on this project.

References

Agrawal, S. and Goyal, N. Further optimal regret bounds for thompson sampling. In *Proceedings of Machine Learning Research*, 2013a.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of Machine Learning Research*, 2013b.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009.

Brafman, R. I. and Tennenholtz, M. R-max-a general polynomial time algorithm for near- optimal reinforcement learning. In *The Journal of Machine Learning Research*, 2003.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. In *NIPS*, 2011.

Christoph Dann, Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Neural Information Processing Systems*, 2016.

Daniel Russo, Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, 2013.

E. Kauffmann, N. Korda and Munos, R. Thompson sampling: an asymptotically optimal finite time analysis. In *International Conference on Algorithmic Learning Theory*, 2012.

Ian Osband, Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning. In *EWRL*, 2016.

Ian Osband, Daniel Russo, Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Neural Information Processing Systems*, 2013.

Jaakkola, T., Jordan M.I. Singh S.P. On the convergence of stochastic iterative dynamic programming algorithms. In *Neural Computation 6(6)*, 1994.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. In *Machine Learning*, 2002.

Russo, D., Van Roy B. An information-theoretic analysis of thompson sampling. In *Journal of Machine Learning Research*, 2016.

Scott, S.L. A modern bayesian look at the multi-armed bandit. In *Applied Stochastic Models in Business and Industry*, 2010.

Singh, S.P., Jaakkola T. Littman M.L. Szepesvari C. Convergence results for single-step on-policy reinforcement-learning algorithms. In *Machine Learning*, 2000.

Sutton, Richard S. and Barto, Andrew G. The MIT Press, 1998.

T. Jaksch, R. Ortner and Auer., P. Near-optimal regret bounds for reinforcement learning. In *The Journal of Machine Learning Research*, 2010.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. In *Biometrika*, 1933.

Watkins, C.J., Dayan P. *Q*-learning. In *Machine Learning*, 1992.

Appendix

Proposition 1. *Let the transition probabilities $\mathcal{P}(\cdot|s,a)$ be fixed and known to the agent, and assume the agent starts from the same state at the beginning of each episode. Assume the episodic undiscounted unknown MDP M^* is sampled from the agent's prior. Then* EXACT-IDRL *(algorithm 1) achieves a Bayesian expected regret upper bounded by*

$$|S|\sqrt{\frac{1}{2}H|A|\ln(|A|)T}$$

for any choice of the prior distribution of the rewards.

Proof. Without loss of generality, assume that the states are numbered $1,\dots,|S|$ and that the agent starts in state 1 at the beginning of each episodes. Recall that a policy μ identifies a Markov Chain and thus a transition probability matrix P. Denote with (s_{A_i}, a_{A_i}) the (random) position and action taken by the agent, and with $\mathbb{1}\{(s_{A_i}, a_{A_i} = (s,a))\}$ the indicator that the agent takes action a in state s at time i. The value function can be written as the expected sum of rewards at different timesteps:

$$V_\mu^M = \mathbb{E}_M \sum_i^H R(s_{A_i}, a_{A_i})$$

where the rewards are function of the random position experienced by the agent s_{A_i} at time i. Now recast each reward at time i as a sum over all possible states.

$$R(s_{A_i}, a_{A_i}) = \sum_s^{|S|}\sum_a^{|A|} \mathbb{1}\{(s_{A_i}, a_{A_i}) = (s,a)\}R(s,a).$$

In the above expression, $\mathbb{1}\{(s_{A_i}, a_{A_i}) = (s,a)\}$ is the indicator that is 1 if the agent takes action a in position s at timestep i. This allows us to rewrite the value function as:

$$V_\mu^M = \mathbb{E}_M \left(\sum_i^H \sum_s^{|S|}\sum_a^{|A|} \mathbb{1}\{(s_{A_i}, a_{A_i}) = (s,a)\}R(s,a) \right).$$

We recall that the agent is following a fixed policy which provides a deterministic action $\mu(s,i)$ at state s, time step i. This allows to rewrite the indicator as:

$$\mathbb{1}\{(s_{A_i}, a_{A_i}) = (s,a)\} = \mathbb{1}\{(s_{A_i} = s)\}\mathbb{1}\{\mu(s,i) = a\}.$$

We use this fact to rewrite the value function:

$$V_\mu^M = \mathbb{E}_M \left(\sum_i^H \sum_s^{|S|}\sum_a^{|A|} \mathbb{1}\{(s_{A_i} = s)\}\mathbb{1}\{\mu(s,i) = a\} = (s,a)\}R(s,a) \right).$$

Bringing the expectation inside yields:

$$V_\mu^M = \left(\sum_i^H \sum_s^{|S|}\sum_a^{|A|} \mathbb{E}_M\, \mathbb{1}\{(s_{A_i} = s)\}\mathbb{1}\{\mu(s,i) = a\} = (s,a)\}R(s,a) \right).$$

Notice that $\mathbb{1}\{\mu(s,i)=a\}$ is not a random variable but a deterministic quantity. Further notice that the reward distribution $R(s,a)$ for a given state action pair (s,a) is independent of the agent or its trajectory along the MDP. Thus,

$$V_\mu^M = \sum_i^H \sum_s^{|S|} \sum_a^{|A|} \mathbb{E}_M\left(\mathbb{1}\{(s_{A_i}=s)\}\right) \mathbb{1}\{\mu(s,i)=a\} = (s,a)\} \mathbb{E}_M\left(R(s,a)\right),$$

and finally by interchanging the summations:

$$V_\mu^M = \sum_s^{|S|} \sum_a^{|A|} \sum_i^H \mathbb{E}_M\left(\mathbb{1}\{(s_{A_i}=s)\}\right) \mathbb{1}\{\mu(s,i)=a\} = (s,a)\} \mathbb{E}_M\left(R(s,a)\right).$$

Notice that a policy μ induces a Markov Chain and the probability of visiting a state s can be computed as

$$\mathbb{E}_M\, \mathbb{1}\{(s_{A_i}=s)\} = P^j e_1$$

where e_1 is the canonical vector $e_1 = [1,0,\ldots,0]$ indicating that the agent starts from state 1 with probability 1.

Define as $v \in \mathbb{R}^{|S||A|}$ the vector that contains in the position (s,a) the expected number of times the agent takes action a in state s during the episode. We have that

$$v(s,a) = \sum_i^H \mathbb{E}_M\, \mathbb{1}\{(s_{A_i}=s)\} = \sum_i^H P^j e_1.$$

Notice that this is a deterministic quantity computable by the agent. We can now write the expected reward accumulated during the episode as:

$$V_\mu^M = \sum_s^{|S|} \sum_a^{|A|} v(s,a)\, \mathbb{E}_M\, R(s,a)$$

which is an inner product between a fixed vector and a random quantity (the rewards) distributed according to a given probability distribution. This setting is analogous to the linear bandit problem, and proposition 4 of (Daniel Russo, 2013) can be used to give a bound $\mathcal{O}\sqrt{\frac{1}{2}H(\alpha)dT}$ where $H(\alpha)$ is the entropy of the optimal action, d is the dimension of the vector space of the vector of the linear bandit problem. In this case, the agent knows the "action" vector $v(s,a)$ that corresponds to the number of visit to the state action pair for any given policy. Vector v lives in $\mathbb{R}^{|S||A|}$, a space of dimension $d = |S||A|$, and there are at most $|A|^{|S|H}$ such vectors, that is, one per policy. It follows that the entropy of the optimal policy is at most $\log|A|^{|S|H} = |S|H\log|A|$ yielding the bound. □

Lemma 2. *The following holds:*

$$\begin{aligned}\Delta_t(\mu) &\overset{def}{=} \mathbb{E}_{\mu^*}\left(\mathbb{E}_M\left(V_{\mu^*}^M - V_\mu^M|\mathcal{F}_t,\mu^*\right)\right) \\ &= \mathbb{E}_M\left(\max_{\mu^*} V_{\mu^*}^M - V_\mu^M|\mathcal{F}_t\right)\end{aligned} \tag{14}$$

Proof.

$$\begin{aligned}\Delta_t(\mu) &\overset{\text{def}}{=} \mathbb{E}_{\mu^*}\left(\mathbb{E}_M\left(V_{\mu^*}^M - V_\mu^M|\mathcal{F}_t,\mu^*\right)\right) \\ &= \mathbb{E}_{\mu^*}\left(\mathbb{E}_M\left(V_{\mu^*}^M|\mathcal{F}_t,\mu^*\right)\right) - \mathbb{E}_{\mu^*}\left(\mathbb{E}_M\left(V_\mu^M|\mathcal{F}_t,\mu^*\right)\right) \\ &= \mathbb{E}_{\mu^*}\left(\mathbb{E}_M\left(V_{\mu^*}^M|\mathcal{F}_t,\mu^*\right)\right) - \mathbb{E}_M\left(V_\mu^M|\mathcal{F}_t\right) \\ &= \mathbb{E}_{\mu^*}\left(\mathbb{E}_M \max_{\mu^*}\left(V_{\mu^*}^M|\mathcal{F}_t,\mu^*\right)\right) - \mathbb{E}_M\left(V_\mu^M|\mathcal{F}_t\right) \\ &= \mathbb{E}_{\mu^*}\left(\mathbb{E}_M \max_{\widehat{\mu}}\left(V_{\widehat{\mu}}^M|\mathcal{F}_t,\mu^*\right)\right) - \mathbb{E}_M\left(V_\mu^M|\mathcal{F}_t\right) \\ &= \mathbb{E}_M\left(\max_{\widehat{\mu}} V_{\widehat{\mu}}^M - V_\mu^M|\mathcal{F}_t\right) \\ &= \mathbb{E}_M\left(\max_{\mu^*} V_{\mu^*}^M - V_\mu^M|\mathcal{F}_t\right)\end{aligned} \tag{15}$$

□