Information-Based Exploration for Reinforcement Learning

Editor: –

Abstract

A common approach to the *exploration versus exploitation* trade-off in reinforcement learning is the *exploration bonus* method. Its core idea is to add bonus to actions that encourage exploration. However, previous algorithms using this method defined the bonus as a function of the number of times an action was previously selected, which can lead to inefficient exploration in some cases. Here we propose a new algorithm motivated by information theory, in a different way than previously used in reinforcement learning, to define a more principled bonus function. We simulate our algorithm in a representative scenario and show that it outperforms previous algorithms.

Keywords: Reinforcement Learning, Exploration versus Exploitation Trade-off, Information Theory

1. Introduction

The *exploration versus exploitation* trade-off is a well-known unresolved conflict in reinforcement learning: should an agent select an action that is known to be beneficial or an unfamiliar action that could provide more valuable information. Many algorithms (e.g., Brafman and Tennenholtz (2003); Strehl and Littman (2005, 2008); Kolter and Ng (2009); Lopes et al. (2012)) use the *exploration bonus* method to handle this conflict. In this method the agent assigns a bonus to actions and states that previously were not selected much. Each of the aforementioned algorithms uses a different bonus function, but all of them define the bonus to be some (decreasing) function of the number of times an action was previously selected.

The exploration bonus method was shown to be useful both theoretically and empirically. However, there are some cases where following this method, as was used so far, leads to undesirable results. For example, consider the following extreme case: suppose the agent can select between two actions a_1 and a_2. Its observations so far are that a_1 always led to the same state while a_2 always led to a new state. In this case the agent can reasonably presume that a_1 is much more deterministic than a_2, and thus, for an efficient exploration, the bonus assigned to a_2 should be larger than the bonus assigned to a_1. Therefore, efficient exploration cannot be achieved if the bonus is only a function of the number of times an action was previously selected. Generally, this problem occurs whenever the actions have varied degrees of "randomness".

We propose a bonus function that depends on the states that the agent reached so far, unlike previous works. Inspired by a line of work that combines information theory and control (Tishby and Polani (2011); Rubin et al. (2012); Todorov (2006); Kappen (2005)), we use notion of predictive information to quantify the predictability of our actions (i.e., see Cover and Thomas (2012); Bialek et al. (2001)). Plugging this new measure into the exploration bonus method yields a new algorithm, which we call Information-Based Exploration (IBE). We then show a representative scenario where IBE outperforms the other algorithms (R-max, Bayesian Exploration Bonus, ϵ-greedy, and Model-

based Interval Estimation with Exploration Bonus). Also, we demonstrate that when the degrees of randomness of the transition probabilities are similar to one another, IBE's performance is not much different than the other algorithms.

We mention that while information theory was already used in reinforcement learning (e.g., Tishby and Polani (2011); Rubin et al. (2012); Todorov (2006); Kappen (2005); Van Hoof et al. (2015); Peters et al. (2010); Schulman et al. (2015); Still and Precup (2012)), our results are fundamentally different. This is reflected, for example, in the fact that all these works yield stochastic policies, while our approach enables exploration even using a deterministic policy.

2. Preliminaries

2.1 Markov Decision Process

A (finite) Markov Decision Process (MDP) is a tuple (S, A, R, P) where S is a finite state space; A is a finite action space; $R\colon S \times A \times S \to \mathbb{R}$ is a reward function, and $P(\cdot|s,a)$ is a probability distribution over S for any $a \in A, s \in S$. A (deterministic) *policy* is a mapping from states to actions $\pi\colon S \to A$. Denote the initial state by s_0 and the state at time i by s_{i+1}, i.e., $s_{i+1} \sim P(\cdot|s_i, \pi(s_i))$. In this work, we focus on a setting where the aim is to reach some *terminal state* with maximum expected rewards. This is known in the literature as 'episodic markov decision process'. To this aim, we introduce a new terminal state s_{goal} which is an absorbing state (i.e., $P(s_{goal}|s_{goal}, \cdot) = 1$) We assume here that all rewards are negative (i.e., costs with $R(s,a) < 0$) and the absorbing state is 'cost-free' (i.e., $R(s_{goal}, \cdot) = 0$). In broad terms, the agent's objective is to find a policy that maximizes the expected sum of reward

$$\mathcal{R} = \sum_{i=0}^{\infty} \mathop{\mathbb{E}}_{s_{i+i}} R(s_i, \pi(s_i), s_{i+1}).$$

When the transition probability P is known, an optimal policy can be found using the classical *value-iteration* algorithm (see Sutton and Barto (1998)). This case is easy to solve because all the information is known and thus the exploration versus exploitation trade-off does not arise. Hence, the more interesting case, that we consider in this paper, is when P is unknown (i.e., only the states that the agent has visited, $s_0, s_1, \ldots$, are observed). For simplicity of notation we assume, without loss of generality, that the reward function R, the states S and the actions A are known, as was done in Kolter and Ng (2009).

2.2 Information Theory

Throughout this work we will use a known measure of closeness between distributions, the Kullback-Leibler divergence (see Cover and Thomas (2012)). The divergence from q_2 to q_1 is defined as

$$\mathbb{D}[q_1||q_2] := \sum_x q_1(x) \log \frac{q_1(x)}{q_2(x)}.$$

We will sometimes use this divergence with distributions over paths of actions and states $s_0, a_0, s_1, a_1, \ldots$ which be be denoted by $\mathbb{P}_{p,\pi}$, where the paths are generated using some transition probability $p(s'|s,a)$ and a policy $\pi(a|s)$.

3. Our Results

Our starting point is the framework suggested in Tishby and Polani (2011) for formally describing the perception-action cycle. This framework is based on the relationship between Predictive Information and learning (Bialek et al. (2001)), where the mutual information between the past and the future of a process is shown to be a universal regularizar of learning processes. The predictive information regularization, when applied to MDP processes, adds the flow of information between the agent and its environment to the standard MDP problem. Thus, instead of just maximizing the expected sum of reward using policy π_1 when the transition probability is p_1, the agent is also required to minimize the distance to some prior behavior defined by transition probability $\hat{p}$ and a policy $\hat{\pi}$

$$\begin{aligned}
\mathbb{D}[\mathbb{P}_{p_1,\pi_1}||\mathbb{P}_{p_2,\pi_2}] &= \mathbb{E}\log\frac{\Pr_{p_1\pi_1}(s_0,a_0,s_1,a_1,\ldots)}{\Pr_{\hat{p},\hat{\pi}}(s_0,a_0,s_1,a_1,\ldots)} \\
&= \mathbb{E}\log\frac{\pi_1(a_0|s_0)p_1(s_1|s_0,a_0)\pi_1(a_1|s_1)\ldots}{\hat{\pi}(a_0|s_0)\hat{p}(s_1|s_0,a_0)\hat{\pi}(a_1|s_1)\ldots} \\
&= \sum_t \mathbb{E}\log\frac{\pi_1(a_t|s_t)}{\hat{\pi}(a_t|s_t)} + \mathbb{E}\log\frac{p_1(s_{t+1}|s_t,a_t)}{\hat{p}(s_{t+1}|s_t,a_t)}
\end{aligned}$$

The first term corresponds to the flow of information from the agent to the environment (by means of *action selection*) and the second term corresponds to the flow of information from the environment back to the agent (*sensory perception*). In this work we focus on a deterministic policy which practically eliminates the first term and thus we consider only the information from the environment to the agent.

Since the agent does not know the transition probability and it only knows the observations, then P cannot be used instead of p_1, so we settle with the empirical distribution instead. But how should the empirical distribution P_{emp} be updated. One natural choice is using Bayesian updating of categorical distribution with Dirichlet prior which is

$$P_{emp}(s'|s,a) = \frac{visited(s,a,s') + c}{visited(s,a) + c|S|},$$

where c is some constant, $visited(s,a,s')$ is the number of times that the agent reached state s' after selecting action a from state s, and $visited(s,a) = \sum_{s'} visited(s,a,s')$ is the number of times the agent selected action a from state s. The constant c is some parameter that should be larger than 0; if $c = 0$ then after one observation the agent will mistakenly think that P_{emp} is deterministic. Without loss of generality we set the prior to be the non-informative uniform distribution over the states, i.e., $\hat{p}(s'|s,a) = 1/|S|$. To sum up, the agent will minimize the following free energy

$$\mathbb{F}^{\pi}(s) = \mathbb{E}_{a,s'}\left[-\beta R(s,a) + \log\frac{P_{emp}(s'|s,a)}{\hat{p}(s'|s,a)} + \mathbb{F}^{\pi}(s')\right].$$

Luckily, the agent can easily solve this problem using the value iteration algorithm where now the "reward" is the local free energy

$$f(s,a,s') = \beta R(s,a) - \log\frac{P_{emp}(s'|s,a)}{\hat{p}(s'|s,a)}.$$

It is important to notice that f is different from the usual reward in MDP as it keep changing as a function of the information the agent has on the environment.

3.1 Algorithm

The IBE algorithm is summarized in the following pseudocode. It can be viewed as using the exploration bonus method where now (unlike previous works) the bonus depends on the states that the agent reached so far. As will be seen in the next section, it will preform better than other algorithms when the sensory transition term is quite different for different actions.

Algorithm: Information-Based Exploration

$s \leftarrow$ start-state, $visited \leftarrow 0$ //Initializations
loop
  $f(s,a,s') = \beta R(s,a,s') - \log \frac{P_{emp}(s'|s,a)}{\hat{p}(s'|s,a)}$
  $\pi \leftarrow$ value-iteration(P_{emp}, f)
  choose an action $a \leftarrow \pi(s)$
  get to new state s'
  $visited(s,a,s') \leftarrow visited(s,a,s') + 1$
  for $s'' \in S$ **do**
    $P_{emp}(s''|s,a) \leftarrow \frac{visited(s,a,s'')+c}{visited(s,a)+c|S|}$
  end for
  $s \leftarrow s'$
  if $s =$terminal-state **then**
    $s \leftarrow$ start-state
  end if
end loop
return value-iteration(P_{emp}, R)

4. Experiments

In this section we empirically compare the algorithm we have suggested, IBE, to other algorithms that use the exploration bonus method (Kolter and Ng (2009); Strehl and Littman (2008); Brafman and Tennenholtz (2003)) and to the classical ϵ-greedy algorithm on two representative examples. In the first example there is one transition probability that is much more stochastic than others and, as expected, IBE outperforms the other algorithms. In the second example, which is a kind of a 'sanity-check', all transition probabilities has a similar degree of stochasticity and we observe that IBE has similar (even a slightly better) performance as the other algorithms. We start with a quick review of the previous algorithms.

- ϵ-greedy: In the ϵ-greedy approach (see Sutton and Barto (1998)) the agent chooses the action that it believed to be the best in the long-term with probability $1-\epsilon$, for some parameter of the algorithm ϵ, and it chooses an action uniformly at random, otherwise.

- R-max: Another approach uses the "optimism under uncertainty" principle (see Brafman and Tennenholtz (2003)). This means that if an action was not used enough times the agent will imagine as if the action is extremely valuable. What is considered "enough" time is a parameter of the algorithm. Under this approach exploration is always valuable.

- Bayesian Exploration Bonus (BEB) and Model-based Interval Estimation with Exploration Bonus (MBIE): In (see Strehl and Littman (2005, 2008); Kolter and Ng (2009)) a psewdo rewars is being defined $R(s,a,s') + \zeta(s,a)$ where the exploration bonus, $\zeta(s,a)$, is some decreasing function of number of times the agent selected the action a from state s. In Strehl and Littman (2005, 2008) the exploration bonus $\zeta(s,a) = \frac{a}{\sqrt{visited(s,a)}}$ and in Kolter and Ng (2009) $\zeta(s,a) = \frac{a}{1+visited(s,a)}$, for some parameter a.This approach actually includes the "optimism under uncertainty" principle, if the bonus is large for state-action pairs that have not been observed enough times.

4.1 Simulation: Maze with a Stochastic Square

Our evaluation environment (shown in Figure 1(a)) is a maze which is represented by a discrete MDP with 16 states and four actions: up, down, right, and left. When the agent selects an agent outside the border of the maze, it stays put. The agent starts the maze from the bottom-left corner and ends the maze in the goal state, the bottom-right corner. When the agent reaches the goal state it returns to the start state. All actions, except on the azure square, are deterministic (e.g., when an agent goes *up* it will indeed go to square above it with probability 1). There is a cost of 1 at each action, till reaching the goal state. On the azure square with action "right" the transition probability is completely different: with uniform distribution the agent will reach up, right, left or stays put. The cost if the state reached is the square on the right is 0.5 and if the agent stayed in the same, up, or left squares then the cost is 0.2.

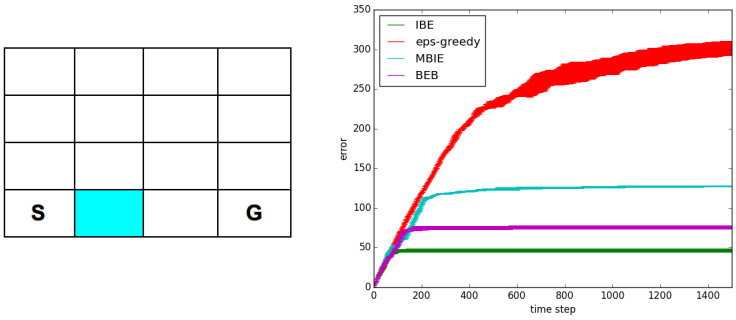
The optimal path is through the azure square, but in order to learn it, the azure square and action "right" must be observed many times while other states-action pairs should be visited much less. We plot the results of the exploration bonus algorithms and ϵ-greedy. We have excluded the R-max algorithm because the time steps required till convergence was significantly higher than the other algorithms. We run each algorithm 5 times and we plot the average cumulative error (i.e., number of time steps that a non-optimal action was taken) and the standard deviation. In this simulation and the next, we have evaluated a wide range of values for the parameters of each algorithm and chose the best for each (the same evaluation strategy was used in previous works Kolter and Ng (2009); Strehl and Littman (2008)). In the IBE algorithm the sensory perception term in the free energy should be less meaningful as the agent gain more observations, we enforce it by choosing, similar to Fox et al. (2016), $\beta := k \cdot \sum_a visited(s,a)$, for some constant k.

4.2 Simulation: Simple Maze

In the second evaluation environment (shown in Figure 2(a)) there is an obstacle (represented by black squares) in the maze. The optimal path is going up twice from the start state, going right, and then down to the goal state. For each action there is a small probability that the agent will follow a different action than the one it selected. We see that IBE has similar performance to R-max and MBIE (maybe even slightly better).

5. Conclusion

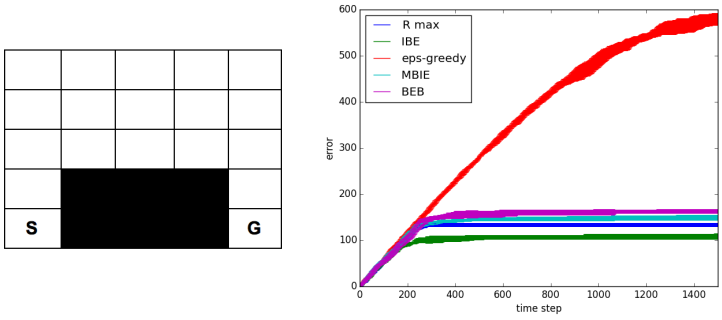
We have proposed the *Information-Based Exploration* algorithm which better utilizes the well-known exploration bonus method. We have observed that previous algorithms that use this method do not consider the states that the agent reached to design the exploration bonus. This observation,



(a) Evaluation environment

(b) Number of non-optimal actions selected versus time for different algorithms: MBIE, BEB, ϵ-greedy, and IBE

Figure 1: Simulation 1 - maze with a stochastic square



(a) Evaluation environment

(b) Number of non-optimal actions selected versus time for different algorithms: R-max, MBIE, BEB, ϵ-greedy, and IBE

Figure 2: Simulation 2 - maze with an obstacle

together with tools from information theory enabled us to design a more sophisticated bonus function. In cases where the randomness of the a transition probabilities (i.e., KL divergence from the uniform distribution) is varied, IBE errs less compared to other algorithms (R-max, MBIE, BEB, and ϵ-greedy). In the algorithm we choose $\beta \to 0$ in order to make sure that as the agent gain more observations the bonus function decreases to 0. Other ways to enforce it can be investigated, for example, by updating the prior using the empirical distribution.

References

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463, 2001.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3:213–231, 2003.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. *Twenty-Fourth National Conference on Artificial Intelligence*, pages 16017–1612, 2016.

Hilbert J Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011, 2005.

J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 513–520. ACM, 2009.

Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in Neural Information Processing Systems*, pages 206–214, 2012.

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In *National Conference on Artificial Intelligence (AAAI)*. Atlanta, 2010.

Jonathan Rubin, Ohad Shamir, and Naftali Tishby. Trading value and information in mdps. In *Decision Making with Imperfect Decision Makers*, pages 57–74. Springer, 2012.

John Schulman, Sergey Levine, Philipp Moritz, Michael I Jordan, and Pieter Abbeel. Trust region policy optimization. *CoRR, abs/1502.05477*, 2015.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Alexander L Strehl and Michael L Littman. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*, pages 856–863. ACM, 2005.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011.

Emanuel Todorov. Linearly-solvable markov decision problems. In *Advances in neural information processing systems*, pages 1369–1376, 2006.

Herke Van Hoof, Jan Peters, and Gerhard Neumann. Learning of non-parametric control policies with high-dimensional state features. In *International Conference on Artificial Intelligence and Statistics*, 2015.