Deriving Intrinsic Motivation From Uncertainty of Future Goals

Jeffrey Negrea[*]
Department of Statistical Sciences
University of Toronto
Toronto, ON M5S 3G3
negrea@utstat.toronto.edu

David Duvenaud
Department of Computer Science
University of Toronto
Toronto, ON M5S 3H5
duvenaud@cs.toronto.edu

Abstract

Intrinsically-motivated reinforcement learning aims to address the question of how actors are motivated in the absence of an external reward system. Among possible behavioural strategies, empowerment-based reasoning is a novel and intuitive strategy for intrinsically motivated reinforcement learning. However, it is unclear from the literature whether empowerment based reasoning is Bayes-optimal. The first contributions of this work are to pose a general reinforcement learning problem in the framework of Bayesian decision theory and to assess the performance of Empowerment-based strategies relative to the Bayes-optimal strategies. Empowerment will be shown to be suboptimal when the Bayesian loss function is the (negative) Expected Reward. A secondary contribution of this work is to demonstrate that Bayes-optimal strategies for problems with non-unit rewards may be determined by examining related problems with unit rewards. The final contribution of this work will present a simple method for an actor to learn the distributions of objectives and rewards through multiple trials and update his Bayes-optimal strategy accordingly.

1 Introduction

1.1 Reinforcement Learning and Intrinsically Motivated Learning Background

PLACE HOLDER

1.2 The Unknown Objective Problem

Suppose an actor is able to traverse a *finite* state space, $\mathcal{S}$. The actor may position itself at any initial state, $s_0 \in \mathcal{S}$. The actor will then be given a target state, $t \in \mathcal{S}$, to which it must travel, chosen randomly from some distribution, $\pi(t)$. The action(s) taken by the actor will be sampled from a (possibly degenerate) distribution, which it may choose from a class of viable action distributions. The resultant state, $s_1 \in \mathcal{S}$, will depend on the action(s) taken, though not deterministically. In [1], the actor is supposed to be allowed a fixed number, K, atomic actions in which it may accomplish its goal. Since the actions are not explicitly conditioned on the intermediate states, then without loss of generality, the paper will consider the possible combinations of K atomic actions in such situations as a single atomic action in a larger space of candidate actions. Show how one can express multiple action joint distribution as a single random action. If the actor succeeds in reaching the target state it receives a unit reward otherwise there is no reward.

[*] MSc. Statistical Sciences Candidate – This paper was submitted as a research project for credit under University of Toronto course code CSC 2541

One way to generalize the above problem would place the actor either in a fixed or random position (not of its choosing) and allow it a fixed number of 'planning stage' actions to position itself before the objective is revealed. The focus of this paper may be viewed as the case in this framework where the actor is almost surely able to position itself in the state of its choosing within the planning stage.

A second generalization of the problem would involve non-uniform, possibly random rewards.

This work assumes that the distribution of the resultant state given the initial state and the actions are known. The problem could be generalized to relax this condition. Some results regarding learning this distribution are presented in [1].

As the actor is uncertain of what his goal will be, any reasonable strategy for selecting the starting state, s_0, will aim to balance the following meta-objectives:

(A) The actor should be able to exert control over the terminal state

(B) The actor should be able to explore a wide range of states

(C) The actor should be able to learn the distribution of objectives (and that of rewards if applicable) over repeated games

Strategies discussed in the literature, such as vanilla empowerment [2, 3, 4] aim to address (A) and (B) above, but provide no means by which to address (C). The generalized empowerment introduced here will be shown to allow Bayesian learning of objective and reward distributions as well accomplishing the other objectives.

2 Empowerment, and Expected Reward

2.1 Empowerment

The empowerment of $s \in \mathcal{S}$ is defined as the maximum over viable action distributions of the mutual information of action random variable and the terminal state random variable, S_1.

$$\mathcal{E}(s) = \max_{(A|s)\sim\omega\in\Omega_s} \mathcal{I}(A, S_1|S) = \max_{(A|s)\sim\omega\in\Omega_s} \mathbb{E}\left[\log \frac{p(S_1, A|s)}{\omega(A|s)p(S_1|s)}\right]$$

Where the expectation is taken with respect to the joint distribution of the action and the resultant state, (A, S_1) conditional on the starting state, s, for a fixed choice of ω — the distribution of the action given the starting state.

2.2 Expected Reward

The Expected Reward when starting in the state of $s \in \mathcal{S}$ is defined assuming that once the goal state is revealed the actor will perform the action which is most likely to allow it to achieve its goal:

$$\mathcal{R}(s) = \sum_{t\in\mathcal{S}} \pi(t) \max_{\omega\in\Omega_s} p(t|s, a_t)\omega(a_t|s)$$

2.3 Computational Complexity

In order to determine the optimal initial state for either Empowerment or Expected Reward, one must compute the empowerment/expected reward of every state. Assume that $|\mathcal{S}| = N$ and that the actor has M atomic actions, $\{a_j\}_{j=1}^{M}$ available and that the class of action distributions from which the actor can sample its action is the simplex defined by the atomic actions.

To determine the complexity of finding the Expected Reward optimal initial state, one notices that $\max_{\omega\in\Omega_s} p(t|s,a)\omega(a|s) = \max_{1\leq j\leq M} p(t|s,a_j)$, since the left problem is a linear program and the basic solutions are the pure strategies. The right problem has run time $O(M)$ for a fixed target state t and initial state s. Since this must be computed for every initial state and every target terminal state, the total complexity to find the Expected Reward maximizing initial state is $O(M \cdot N^2)$.

To determine the complexity of finding the Empowerment optimal initial state, one notes that the mutual information is convex [5] in ω for each candidate initial state. Thus for a particular candidate initial state, computing the Empowerment is equivalent to solving an $M-1$ dimensional convex optimization problem. Suppose that the complexity of this is $\mathfrak{c}(M-1)$. As this must be done for every possible initial state, the total complexity to find the Empowerment maximizing initial state is $O(\mathfrak{c}(M-1)\cdot N)$.

One can definitively say that holding the number of pure actions constant, Empowerment maximization is asymptotically less complex than Expected Reward maximization in the number of states. The complexity of solving a convex optimization problem is dependent on the algorithm used. If the algorithm used for convex optimization is superlinear in the dimensionality then, holding the number of states fixed, Expected Reward is asymptotically less complex than Empowerment maximization with respect to the number of pure actions. If the algorithm used for convex optimization is linear (sublinear) in the dimensionality then, holding the number of states fixed, Expected Reward is asymptotically equally as complex (more complex than) Empowerment maximization with respect to the number of pure actions.

3 Sub-Optimality of Empowerment

Let $\mathcal{S}=\{0,1\}$, $A_p=\begin{bmatrix}p_0 & 1-p_0\\ p_1 & 1-p_1\end{bmatrix}$, and $A_q=\begin{bmatrix}q_0 & 1-q_0\\ q_1 & 1-q_1\end{bmatrix}$. Let ω_u be the action distribution which selects action a_p with probability u and a_q with probability $(1-u)$. When an action, $a_\star$, is taken the actor transitions according to the corresponding Markov kernel, $A_\star$. Let $\Omega=\{\omega_u : u\in[0,1]\}$. Suppose the goal state is chosen uniformly from $\mathcal{S}$.

3.1 Expected Reward Optimal Solution

First one will derive the Expected Reward optimal strategy. To this end, one first notes that for a fixed initial state, once the goal state is known the actor's will move according to the Markov kernel which has higher probability to reach the goal state. If both kernels yield the same probability then the actor is indifferent between them. Let $\mathcal{R}(i)$ denote the Expected Reward when starting in state i. Then

$$\begin{aligned}\mathcal{R}(i) &= \frac{1}{2}\max(p_i,q_i)+\frac{1}{2}\max(1-p_i,1-q_i)\\ &= \begin{cases}\frac{1}{2}p_i+\frac{1}{2}(1-q_i) & p_i\geq q_i\\ \frac{1}{2}q_i+\frac{1}{2}(1-p_i) & p_i<q_i\end{cases}\\ &= \frac{1}{2}(1+|p_i-q_i|)\end{aligned}$$

Hence, in this scenario, the Expected Reward maximizing strategy is to start in the state which has the highest discrepancy between the corresponding possible transition distributions. This is the state that maximizes $|p_i-q_i|$.

Empowerment Optimal Solution

One now derives the Empowerment optimal strategy.

$$\begin{aligned}\mathcal{E}(i) = \max_{u\in[0,1]}\Bigg[& up_i\log\frac{p_i}{up_i+(1-u)q_i}\\ & +(1-u)q_i\log\frac{q_i}{up_i+(1-u)q_i}\\ & +u(1-p_i)\log\frac{(1-p_i)}{u(1-p_i)+(1-u)(1-q_i)}\\ & +(1-u)(1-q_i)\log\frac{(1-q_i)}{u(1-p_i)+(1-u)(1-q_i)}\Bigg]\\ = & -uH_{p_i}-(1-u)H_{q_i}+H_{q_i+u(p_i-q_i)}\end{aligned}$$

Where $H_z=-z\log z-(1-z)\log(1-z)$ for all $z\in(0,1)$. To determine the maximum, we define

$$g(u)=-uH_p-(1-u)H_q+H_{q+u(p-q)}$$

Note that $H'_z := \frac{d}{dz}H_z = -\text{logit}(z) = \text{logit}(1-z)$. Then one can differentiate g to find the u corresponding to the local extrema.

$$\begin{aligned} g'(u) &= -H_p + H_q + (p-q)H'_{q+u(p-q)} \\ &= -H_p + H_q + (p-q)\text{logit}(1-q-u(p-q)) \end{aligned}$$

Thus setting $g'(u^\star) = 0$, assuming $p \neq q$ would imply

$$\begin{aligned} \Big[g'(u^\star) = 0\Big] &\Rightarrow \left[\text{logit}(1-q-u^\star(p-q)) = \frac{H_p - H_q}{p-q}\right] \\ &\Rightarrow \left[1-q-u^\star(p-q) = \text{logistic}\left(\frac{H_p - H_q}{p-q}\right)\right] \\ &\Rightarrow \left[u^\star = \frac{1-q-\text{logistic}\left(\frac{H_p-H_q}{p-q}\right)}{p-q}\right] \end{aligned}$$

Since H is a concave function, g is non-negative on [0,1]. Also, $g(0) = g(1) = 0$. Hence g is maximized at $u^\star$. If $p = q$ then the empowerment is uniformly 0.

3.2 A Worked Counterexample

For Empowerment Maximization to be equivalent to Expected Reward Maximization, $g(u^\star)$ would need to be monotone increasing in $|p-q|$. However one establishes a counter example to this:

Suppose $p_0 = 1$, $q_0 = \frac{1}{2}$, $p_1 = \frac{3}{4}$, and $q_1 = \frac{1}{8}$. Then the expected rewards are:

$$\begin{aligned} \mathcal{R}(0) &= \frac{1}{2}\left(1 + \left(1 - \frac{1}{2}\right)\right) = \frac{3}{4} \\ \mathcal{R}(1) &= \frac{1}{2}\left(1 + \left(\frac{3}{4} - \frac{1}{8}\right)\right) = \frac{13}{16} \end{aligned}$$

Hence the Expected Reward Optimal Strategy is to start in state 1.

However, according to the above formulae, the empowerments are

$$\begin{aligned} \mathcal{E}(0) &= 0.223143551 \\ \mathcal{E}(1) &= 0.216016789 \end{aligned}$$

Hence the Empowerment Optimal Strategy is to start in state 0.

In fact, in this class of two state problems with two pure actions a (locally) supremal gap between the expected reward of the Expected Reward optimal strategy exists near $p_0 = 1$, $q_0 = 0.79$, $p_1 = 0.70$, and $q_1 = 0.30$, where the Expect Reward optimal solution is to start in state 1 and has expected reward 0.70 and the Empowerment optimal Strategy is to start in state 0 and has an expected reward of 0.605 – an absolute gap of 9.5% and a relative gap of 13.6%.

This counter example shows that even in the smallest non-degenerate intrinsically motivated learning problems with a uniform distribution of objectives, there can be material under-performance from Empowerment based decision making. Moreover, for the second example, the empowerment strategy is clearly placing too much emphasis on the control of the resultant state (meta-objective A) while ignoring the need to be able to explore a wide variety of states (meta-objective B).

4 Weighted Objectives

The focus of this paper now shifts to non-unit reward systems. One demonstrates that problem of determining the Expected Reward maximizing strategy for such problems can formulated as an Expected Reward optimization under a re-weighted distribution of the target state.

Theorem 1. *Suppose that the rewards received for various target states are non uniform. As before, let π be the probability distribution over objective states. Let the pay-off for achieving the the target state be $m(t)$. Define the probability measure $\tilde{\pi}$ by*

$$\tilde{\pi}(t) = \frac{m(t)\pi(t)}{\sum_{u\in\mathcal{S}} m(u)\pi(u)}$$

Then the optimal strategy in terms of Expected Reward is maximal $\tilde{\pi}$-empowerment.

Proof. The expected reward for the problem with weights m and objective probabilities π differs from the expected reward for the problem with unit rewards and objective probabilities $\tilde{\pi}$ by a multiplicative constant. Hence any optimal strategy for one of these problems must be optimal for the other. □

Theorem 2. *Suppose that rewards received for various target states are stochastic. Particularly, suppose that the reward received when the objective is $t \in \mathcal{S}$, $R|t$, is distributed as $(R|t) \sim m(r|t)$. As before, let π be the probability distribution over objective states. Define the probability measure $\tilde{\pi}$ by*

$$\tilde{\pi}(t) = \frac{\mathbb{E}(R|t)\pi(t)}{\mathbb{E}(R)}$$

Then the optimal strategy in terms of expected reward is maximal $\tilde{\pi}$-empowerment.

The proof is the same as for the above theorem.

5 Learning from Experience

Suppose now that the distribution of objectives and/or the distribution of rewards are unknown to the actor. One may model the prior beliefs of the actor regarding the post-change-of-measure objective distribution as an $|\mathcal{S}|$-dimensional Dirichlet distribution. The actor may sample his Dirichlet prior at each iteration of the game, play according to the optimal generalized empowerment strategy for the sample, then update its prior between each game based on the observed objectives and the observed rewards.

References

[1] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft, "Learning and policy search in stochastic dynamical systems with bayesian neural networks," *arXiv preprint arXiv:1605.07127*, 2016.

[2] S. Mohamed and D. J. Rezende, "Variational information maximisation for intrinsically motivated reinforcement learning," in *Advances in Neural Information Processing Systems*, pp. 2125–2133, 2015.

[3] C. Salge, C. Glackin, and D. Polani, "Empowerment–an introduction," in *Guided Self-Organization: Inception*, pp. 67–114, Springer, 2014.

[4] N. Chentanez, A. G. Barto, and S. P. Singh, "Intrinsically motivated reinforcement learning," in *Advances in neural information processing systems*, pp. 1281–1288, 2004.

[5] M. Braverman and A. Bhowmick, "Lecture notes in information theory in computer science." `https://www.cs.princeton.edu/courses/archive/fall11/cos597D/L04.pdf`, September 2011.