Deriving Intrinsic Motivation from Uncertainty about 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 asks how agents should act in the absence of an explicit reward system. We suggest a simple unifying approach to deriving intrinsic motivation, as arising from uncertainty about which goal an agent will be asked to achieve in the future. We formalize this approach using Bayesian decision theory, and show that it fulfills previous suggestions about reasonable properties of intrinsic motivation. We also compare and contrast it to the recently-proposed empowerment heuristic, and give a simple example where empowerment fails to choose the optimal action.

1 Introduction

A key component of most reinforcement learning frameworks is that the mechanism by which the actor receives rewards is fixed and pre-prescribed. Intrinsically motivated reinforcement learning [4]. aims to relax the assumption that the reward system is always known. For instance, Singh *et al.* [1] propose that the reward system should be internal to the actor as opposed to prescribed by the environment and describe the open ended behaviour of an actor interacting with an environment. Mohammed and Rezende [2], focus on how to best prepare for an unknown future goal/reward system.

The recently-introduced measure of *empowerment* [2, 3] approach the problem using information-theoretic metrics, aiming to maximize notions of dependence between the actions taken and the resultant state. Specifically, empowerment-based intrinsic motivation, as discussed in [2], optimizes the mutual information of the action distribution and the resulting outcome. Heuristically, empowerment measures the amount by which an actor can influence the outcome of a game. It is 0 when the all of the viable actions have identical influence on the resulting state, and is positive otherwise.

1.0.1 Limitations of the empowerment heuristic

However, empowerment may not sufficiently penalize similarity of available actions. For example, an actor who maximizes empowerment may find himself in a situation where an alternative strategy will yield a higher total probability of achieving a random goal. An empowerment-maximizing actor will also be unable to capture the information gathered by an actor regarding the history of random rewards observed and the history of the effects of actions taken. One may argue in favour of empowerment that the measure is reward-system-independent. However, by virtue of the fact that a goal/reward will be randomly prescribed there must be a distribution from which it is selected.

1.1 Extending reinforcement learning to include uncertainly about future goals

In contrast to these *ad-hoc* heuristics, we show that a simple extension to the standard expected-reward-maximisation framework naturally addresses the same problems as intrinsic motivation. The extension is simple: we make explicit consideration of the distribution from which goals/rewards are sampled. Initially, we consider this distribution to be fixed and known. However, in later sections we discuss how this distribution can be learned in an episodic framework.

2 Desiderata of Intrinsic Motivation

Previous work has suggested that [TODO: Fill in previous suggestions about what exactly people said intrinsic motivation should try to do]

We also propose that an agent should [TODO]

The Bayesian decision theory formulation we define does allow learning of objective and reward distributions and the random impact of actions through repeated trials as well accomplishing the other objectives.

3 Formalizing Bayesian Intrinsic Motivation

[TODO: Formally write down the old expected-reward framework, then the new one]

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. If the actor succeeds in reaching the target state it receives a unit reward, otherwise there is no reward. Since the actions are not explicitly conditioned on the intermediate states, then without loss of generality, we consider the possible combinations of K atomic actions in such situations as a single atomic action in a larger space of candidate actions. Even if the later actions are dependent on the intermediate states, we may view the combination actions and intermediate states as a single action in a larger space of actions which has a distribution based on the intermediate states:

$$\mathbb{P}(\mathbf{a}|s_0) = \sum_{\mathbf{s}\in\mathcal{S}^{K-1}} \mathbb{P}(a_K|\mathbf{s}, \mathbf{a}_{K-1}) \prod_{k=1}^{K-1} \mathbb{P}(a_k|\mathbf{s}_{k-1}, \mathbf{a}_{k-1})\mathbb{P}(s_k|\mathbf{s}_{k-1}, \mathbf{a}_k)$$

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 this example 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. We will consider this case in section 6, where we demonstrate that each such problem is effectively equivalent to a game with unit rewards.

3.1 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.

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

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.

3.2 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:

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

4 Computational Complexity of Expected Reward and Empowerment

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.

4.1 Expected Reward

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-hand-side optimization 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 $O(M \cdot N^2)$.

4.2 Empowerment

To determine the complexity of finding the empowerment optimal initial state, one notes that the mutual information is concave [8] in ω for each candidate initial state. Computing the mutual information for a candidate action distribution and a fixed initial state involves summation of MN logarithmic terms, each with M terms to be summed in the denominator. There are only N distinct denominators which are each repeated M times. Hence the time to compute the mutual information is $O((M + 1)N) = O(MN)$ for a particular candidate initial state and candidate action. Computing the empowerment is then equivalent to solving an $M - 1$ dimensional convex optimization problem where the objective function has $O(MN)$ time to compute. Suppose that the complexity of solving a convex optimization problem with dimension D and time to compute the objective function E is $O(\mathfrak{c}(D, E))$. As this must be done for every possible initial state, the total complexity to find the empowerment maximizing initial state is $O(\mathfrak{c}(M, MN) \cdot N)$.

4.3 Comparison

Since $\mathfrak{c}$ must be at least linear in its second argument and monotone increasing in its first argument, maximizing expected reward is asymptotically less costly than maximizing empowerment. Even in the smallest possible non-degenerate intrinsic motivation problem, as will be seen below, the derivation of the expected reward maximizing strategy is far less difficult than derivation of the empowerment maximizing strategy.

This analysis should be viewed cautiously, however, as the expected reward strategy is not guaranteed to simplify as it did here when the reward type is not restricted to 0-1 rewards for achieving a single randomly selected goal state. In fact, the computation time of empowerment is naturally independent of the random reward system in effect, while expected reward is not. A further goal of our research will be to investigate the computational cost of expected reward maximization in scenarios with more general reward systems.

5 The Smallest Non-Degenerate Problem and the Sub-Optimality of Empowerment Strategies

Here we define the family of smallest non-degenerate intrinsic motivation problems, and demonstrate that in this family, there exists members for which empowerment maximization yields irrational decisions.

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 $\mathcal{S}$. The uniformity of goals is considered a simplifying assumption to place expected reward on a closer basis with empowerment, since it treats all states as equal as on would expect empowerment to do.

This family is considered the smallest non-degenerate group of intrinsic motivation problems because reducing the state space or the action space would yield degenerate problems. In a scenario with a smaller state space there is only one state, any goal is always achieved, and all actions would be identical. In a problem with a smaller action space there is only one action, empowerment is uniformly 0, and expected reward is uniform.

5.1 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 $|p_i - q_i|$. This is intuitive as it aims to best associate one action with each possible goal.

5.2 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

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

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.

5.3 Sub-Optimality Example

For empowerment to be equivalent to expected reward maximization, $g(u^\star)$ would need to be monotone increasing in $|p - q|$. However one establishes two counter examples to this. The first example was found by guess-and-check, and the second example was found by searching for the locally maximal performance gap when initialized from the first example.

5.3.1 Example 1

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.

5.3.2 Example 2

Searching starting from Example 1, 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%.

The actual locally supremal gap is not achieved in the search space as the set of cases where one strategy strictly dominates the other with respect to empowerment while being strictly inferior in terms of expected reward is not closed. The limit point corresponding to the supremum occurs where the empowerment based actor is indifferent between strategies. This is why we have used the term 'supremal' as opposed to 'maximal' or 'optimal' and this is why we have presented a point 'near' the local supremum.

These counter examples show 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). This strategy does not seem rational to the authors and does not seem representative of how 'live' actors would behave.

6 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 equivalent to the optimal expected reward strategy when objectives are drawn from $\tilde{\pi}$ under 0-1 *reward.*

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 equivalent to the optimal expected reward strategy when objectives are drawn from $\tilde{\pi}$ under 0-1 *reward.*

The proof is the same as for the above theorem.

7 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. Singh, R. L. Lewis, A. G. Barto, and J. Sorg, "Intrinsically motivated reinforcement learning: An evolutionary perspective," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 2, pp. 70–82, 2010.

[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] P.-Y. Oudeyer, F. Kaplan, *et al.*, "How can we define intrinsic motivation," in *Proc. 8th Int. Conf. Epigenetic Robot.: Modeling Cogn. Develop. Robot. Syst*, 2008.

[5] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge, 1998.

[6] 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.

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

[8] 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.