A Bayesian Approach to Intrinsically Motivated Reinforcement Learning

Jeffrey Negrea
Department of Statistical Sciences
University of Toronto
negrea@utstat.toronto.edu

David Duvenaud
Department of Computer Science
University of Toronto
duvenaud@cs.toronto.edu

Abstract

Intrinsically motivated reinforcement learning asks how agents act when motivated by an internal reward system, and how such a reward system should be structured. We suggest a simple unifying approach to deriving intrinsic motivation, as arising from uncertainty about the future goals of an agent. We formalize this approach using Bayesian decision theory, and show that it meets existing heuristic criteria for the characteristics intrinsic motivation should have. We also compare and contrast it to the recently-proposed empowerment metric, and give a simple example where empowerment fails to choose the optimal policy.

1 Introduction and Related Work

A key component of traditional reinforcement learning frameworks is that the mechanism by which the actor receives rewards is fixed and is pre-prescribed as part of the environment. Intrinsically motivated reinforcement learning aims to generalize the notion of reward functions for reinforcement learning problems by shifting the reward system from being part of the environment to being partially or completely part of the actor, and/or by not requiring the reward system to be fixed or pre-prescribed. This is often interpreted as allowing the actor to be motivated not only by achieving standardized goals, but also for having improved its model of the environment, and is often compared to the way in which humans and animals learn through playing and experimentation [1, 2].

However, there is a divergence in the literature as to what it means to be intrinsically motivated; while it is generally established that the core component of intrinsic motivation is that the reward system used to establish the optimal policy is internal to the actor, there exist multiple distinct interpretations of how rewards systems are related to the environment. The first interpretation asserts that there is a fixed underlying external fitness function (e.g. reproductive success) on the space of actor trajectories. This is described in Singh *et al.* [3], where the authors propose that the reward system used to drive the agent's behaviour should be internal to the agent and the focus of their approach is the search for the internal reward function which yields the highest expected fitness. This may be interpreted as a way of smoothing the external reward system (fitness) in a way which generalizes the Q-learning discounted future reward framework in order to avoid overly greedy, short-sighted strategies.

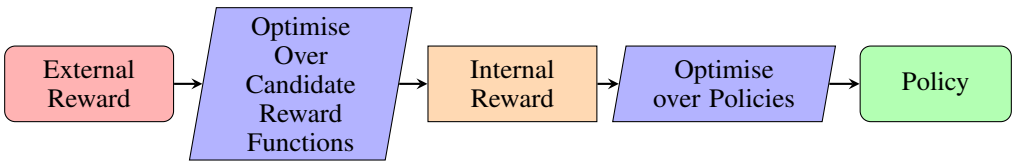

Figure 1: Intrinsic Motivation as per Singh et al.

The second interpretation, as per Schmidhuber [2], posits that reward systems are split (possibly non-linearly) into internal and external components, and that the internal component of motivation is derived from quality of the learned environmental model. The heuristic qualities of the internal reward component in this framework are listed in section 2. Model quality in this framework is, in some instances, identified with the model's efficiency as a compressor of the data history. Incremental internal reward from time t to time $t+1$ is measured based on (a possibly non-linear) notion of the improvement in model quality between the model fit on the first t observations and the model fit on all $t+1$ observations, when performance is assessed with respect to all $t+1$ observations.

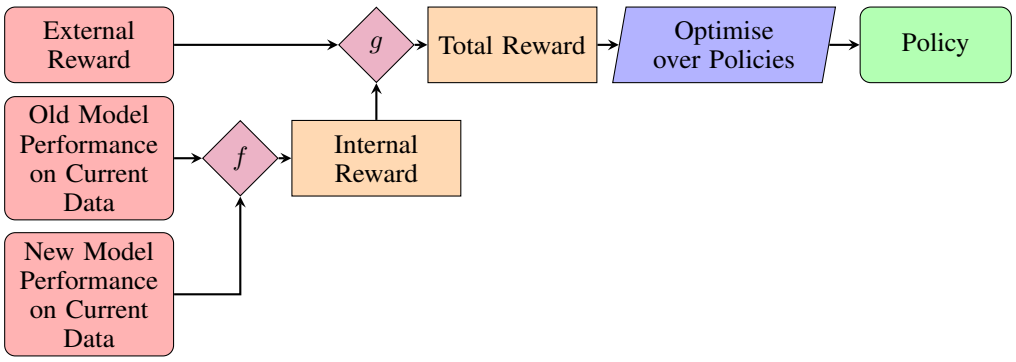

Figure 2: Intrinsic Motivation as per Schmidhuber

The third interpretation, [4, 5], establishes that reward systems exist external to the actor, however they are unknown and stochastic. Internal motivation is then formulated as a utility system that is optimised in order to prepare for an unknown future goal/reward system. It is this notion of intrinsic motivation that we will focus on for this project.

The recently-introduced measure of *empowerment* [4, 5] approaches 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 [4], optimizes the mutual information of the action distribution and the resulting outcome. Heuristically, as described in [5], 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. However, as shown in figure 3, there is no connection between the empowerment-based internal reward system or its resulting policy and the external reward system. This is unsatisfactory as agents will not be driven to achieve real goals through any amount of learning. The expected reward framework we introduce in section 3 addresses the same class of problems as empowerment, but has an architecture which accounts for the external reward system, as shown in figure 4.

Figure 3: Intrinsic Motivation as per Empowerment Maximization

This paper proceeds in the following manner. In section 2 we step back to discuss desiderata for the intrinsic motivation system. In section 3 we formalize the problem we intend to address, formally introduce the notion of empowerment and propose an alternative approach to the unknown/stochastic reward interpretation of intrinsic motivation which puts the analytic formulation of the objective on a more similar basis to the other two interpretations of intrinsic motivation – using expected reward as a starting point. In section 4 we demonstrate that the expected reward framework is no more computationally complex than empowerment. In section 5 we demonstrate that our formulation is non-trivially distinct from empowerment as the optimal strategies may differ for even the smallest non-degenerate problems. In section 6 we briefly describe the posterior update process, by which the model for the reward distribution is improved.

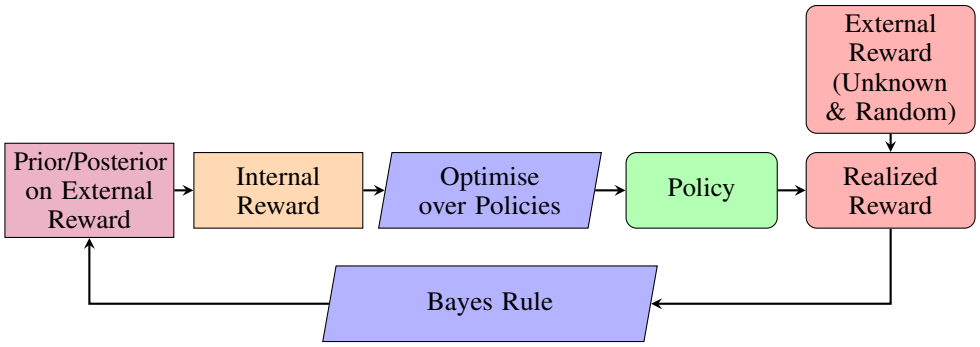

Figure 4: Intrinsic Motivation as per Bayesian Extected Reward

2 Desiderata of Intrinsic Motivation

Previous work has suggested that a wide variety of desirable traits for a hypothetical intrinsically motivated learner.

Schmidhuber [2] posits that intrinsically motivated learners should have the following four traits;

1. An adaptive world model
2. A learning algorithm to update the model
3. An intrinsic reward system which measures the improvement of the learner's model
4. A behavioural policy which aims to optimize the intrinsic reward

Salge *et al.* [5] have characterized empowerment as an intrinsic motivation utility function satisfying three properties; *Local, Universal, and Task-Independent*

1. *Local* means that an agent can determine its level of utility from its local environment as opposed to needing global data
2. *Universal* means the scale on which the level of utility is measured should be independent of the structure of the actor and the environment – utility should be comparable across distinct problem classes.
3. *Task-Independent* means that the level of motivation is independent of any particular goal or reward function.

We demonstrate that a Bayseian approach to unknown/stochastic reward framework for intrinsic motivation, derived via expected reward, can share the characterizing properties of empowerment while maintaining the traits outlined by Schmidhuber.

3 Formalizing Bayesian Intrinsic Motivation

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.

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

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.

3.1.1 Limitations of the empowerment metric

As will be derived analytically in section 5, 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 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. This motivates the Expected Reward framework described below.

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:

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

This definition naturally satisfies the three characterizing properties of empowerment. It is local since calculating expected reward for a state only involves examination of the parts of the state space accessible from that space. It is universal because the scale on which it is measured is probability which is globally comparable. It is task independent because we have 'integrated' over all possible tasks.

Expected Reward may be interpreted in the framework of Bayesian decision theory as the complement of the Bayes risk for the policy 'starting in state s when the prior is π'. This is formalized below:

3.2.1 Bayesian Formulation of Expected Reward

- The possible states of the world is taken as the goal states.
- The action space is taken as the combinations of atomic action in the multiple atomic action framework, or the single stochastic actions in the summarized action framework described above.

- The observations are taken to be the initial and intermediate states in the multiple atomic action formulation, or only the initial state in the summarized action framework.
- The possible decision rule have two components – the second stage selects actions given the initial state, and the first stage selects the initial state.
- The Bayes Loss Function is the complement of the reward function – it is 1 if the goal state is not achieved and 0 if the goal state is achieved.
- The Bayes Risk Function is the expected Loss Function, where expectation is taken over the the possible observations and actions.
- The Bayes Risk is the expectation over possible goal states of the Risk Function.

Determining the Bayes optimal strategy in the above formulation involves optimizing both stages of the decision rule. Given a goal, t, and an initial state, s, the optimal decision rule obviously selects the action policy which gives the highest chance of reaching the goal:

$$\text{argmax}_{\omega \in \Omega_s} p(t|s, a_t)\omega(a_t|s)$$

Hence the Bayes Risk of a decision rule, when considering only decision rules where the second stage is optimal, depends only on the prior distribution on goals and the choice of initial state. and is equal to $(1 - \mathcal{R}_\pi(s))$. Hence the policy of selection the starting state maximizing $\mathcal{R}_\pi(s))$ is equivalent to selecting the Bayes optimal strategy for the above formulation.

4 Computational Complexity

In this section, we show that in the single target state intrinsic motivation problem formulated above, when the transition dynamics are fully known, expected reward maximization is less computationally complex than exact empowerment maximization. The variational approach from [4], simplifies the approximate computation of empowerment based strategies using mini-batches and Monte Carlo approximation when the transition dynamics are not known, hence this is *not* in direct comparison with the cost of the algorithm used in that paper.

We also make explicit that the expected reward strategy is not always guaranteed to simplify as it does 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, 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.

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 probabilistic mixtures of 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 [6] 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 $O(\text{conOpt}(D, E))$. As this must be done for every possible initial state, the total complexity to find the empowerment maximizing initial state is $O(\text{conOpt}(M, MN) \cdot N)$.

4.3 Comparison

Since conOpt 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 for the problem class formulated in section 3. 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.

4.4 Variational Empowerment

In [4] the authors introduce a variational approximation to empowerment. The variational approximation of empowerment is also formulated as an optimization of an expectation over the same distribution as the original empowerment. Hence the analysis of computational complexity of exact evaluation of the variational approximation to empowerment yields the same lower bound on the complexity as the original notion of empowerment. The main advantage of the variational approximation to empowerment, therefore, does not lie in its simplified exact calculation; it lies in the ability to approximate the expectation with Monte-Carlo samples obtained by interaction with the environment, as described in section 4.1 of their paper. This is advantageous when the exact transition dynamics are not known.

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 distinct behaviour from expected reward maximization.

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

The derivation of the empowerment optimal solution is moderately tedious and so is deferred to Appendix A. We state the result below for the reader's convenience. For each state, i, the u which maximizes the mutual information is

$$u_i^\star = \frac{1 - q_i - \text{logistic}\left(\frac{H_{p_i} - H_{q_i}}{p_i - q_i}\right)}{p_i - q_i}$$

where $H_z = -z \log z - (1-z)\log(1-z)$ for all $z \in (0, 1)$.

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 trial and error, 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:

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

Hence the expected reward optimal strategy is to start in state 1.

However, according to the above formulae, the empowerments are

$$\mathcal{E}(0) = 0.223143551$$
$$\mathcal{E}(1) = 0.216016789$$

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 suprememum 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 Learning from Experience

So far, in our discussion of Expected Reward strategies, we have treated the distribution π as being the exact distribution from which rewards are sampled. In reality π would begin as a prior for the probabilities of the multinomial random variable representing the reward distribution. The prior would be chosen based on the agent's beliefs about what types of rewards are likely, and may be chosen to be uniform if the agent has no such information. In an episodic framework the posterior distribution would be updated using Bayes Rule at the end of each episode based on which reward was specified, as outlined in figure 4. Though almost immediately obvious that this is the 'right' thing to do in the expected reward framework, no analogous action exists in the empowerment framework.

This process allows us to claim to achieve the desiderata proposed by Schmidhuber with respect to the knowledge of reward functions. The model for the reward system is multinomial with unknown probabilities. The Bayesian posterior calculations are the learning system which updates the model. The intrinsic reward system is the expectation under the posterior distribution of the reward received. The behavioural policy optimizing intrinsic reward is the policy of starting in the state with highest posterior expected reward.

7 Limitations

The discussion of the expected reward framework in this paper is limited to learning the reward system. We have effectively assumed that the transition dynamics are known. In a realistic setting we would need to learn the transition dynamics simultaneously to learning the reward system. This is, for example, the purpose underlying the variational approximation to empowerment in [4], and it is the motivation underlying (at least a part of) the internal component of the reward system in [2]. In order for the expected reward framework to be useful, we will need to propose a system analogous to the variational approximation of empowerment. We defer this to a later work.

The problem we have considered involves very simple, '0-1' reward systems which only provide positive reward for a single state. In appendix B we provide a slight generalization of this to rewards which can take on positive values other than 1 (and may be random), but are still restricted to being positive in exactly one state. We defer consideration of more complicated reward systems to future work.

We believe that the intrinsic motivation system outlined by Schmidhuber in [2] is the most comprehensive we have encountered in our literature review, and acknowledge that it provides a system by which to learn the transition dynamics as well as the reward system.

8 Conclusions

In summation, we have introduced the expected reward framework for intrinsically motivated reinforcement learning, presented it as a Bayesian decision theory problem, and showed that it satisfies the characterizing properties of empowerment as per Salge. We have compared its computational complexity to that of empowerment and found that it is no more costly. We have also shown that it does indeed provide distinct policies to empowerment. Lastly, we have presented a novel way in which to update our model of reward systems allowing it to satisfy the desiderata posed by Schmidhuber. In the future we intend to extend this framework to a larger class of problems than currently considered and to provide a means by which transition dynamics may be learned alongside the reward system.

References

[1] 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.

[2] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[3] 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.

[4] 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.

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

[6] 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.

[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] 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.

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

[10] R. McAllister and C. E. Rasmussen, "Data-efficient reinforcement learning in continuous-state pomdps," *arXiv preprint arXiv:1602.02523*, 2016.

[11] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," *arXiv preprint arXiv:1610.00696*, 2016.

[12] P. Christiano, Z. Shah, I. Mordatch, J. Schneider, T. Blackwell, J. Tobin, P. Abbeel, and W. Zaremba, "Transfer from simulation to real world through learning deep inverse dynamics model," *arXiv preprint arXiv:1610.03518*, 2016.

[13] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[14] R. G. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models," *arXiv preprint arXiv:1609.09869*, 2016.

Appendix A: Deriving the Empowerment-Optimal Strategy for the Smallest Non-Degenerate Problem

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.

Appendix B: 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.