Exploration via Empowerment Gain: Combining Novelty, Surprise and Learning Progress

Anonymous Authors[1]

Abstract

Exploration in the absence of a concrete task is a key characteristic of autonomous agents and vital for the emergence of intelligent behaviour. Various intrinsic motivation frameworks have been suggested, such as novelty seeking, surprise maximisation or empowerment. Here we focus on the latter, empowerment, an agent-centric and information-theoretic measure of an agent's perceived influence on the world. By considering improving one's empowerment estimator – we call it *empowerment gain (EG)* – we derive a novel exploration criterion that focuses directly on the desired goal: exploration in order to help the agent recognise its capability to interact with the world. We propose a new theoretical framework based on improving a parametrised estimation of empowerment and show how it integrates novelty, surprise and learning progress into a single formulation. Empirically, we validate our theoretical findings on some simple but instructive grid world environments. We show that while such an agent is still novelty seeking, i. e. interested in exploring the whole state space, it focuses on exploration where its perceived influence is greater, avoiding areas of greater stochasticity or traps that limit its control.

1. Introduction

In reinforcement learning (RL) we tend to give an agent a specific task to solve, and use exploration heuristics to speed up training. While these heuristics may be useful, biological systems have a more efficient approach. Even when faced with no concrete task, natural autonomous agents (like children) explore the world playfully to acquire new skills that may be used later. This exploratory behaviour is an autonomous and active endeavour guided by intrinsic motivation which forms the core of a system for task-independent learning (Oudeyer et al., 2007; Oudeyer & Kaplan, 2009).

Intrinsic motivations have been formalised in RL literature in various ways (Aubret et al., 2019). In particular, the concepts of *novelty* and *surprise* are believed to be vital to exploration (Berlyne, 1960; Barto et al., 2013). We follow a common distinction (Berlyne, 1960; Barto et al., 2013; Xu et al., 2021) in defining the key difference of the terms, although the formalisation of novelty is particularly difficult and controversial. Following common intuition, novelty may be defined as being a statistical outlier; something is completely novel if we have not seen it before. One way this has typically been formalised in machine learning (Bellemare et al., 2016; Ostrovski et al., 2017; Tang et al., 2017) is

$$\text{Novelty}(s) \propto -\log p(s), \tag{1}$$

where $p(s)$ models the visitation frequency of state s based on previous experience.

However, not everything that is novel is necessarily surprising. And complex states may reveal something surprising even after having encountered them often before. Surprise (also called *contextual novelty* or *curiosity*) requires an internal world model that formulates an expectation about the future. The deviation between these predictions and the observed reality quantifies the amount of surprise. One way to fomalize this is to use the the forward model's prediction $p_\xi(s_{t+1} \mid s_t, a_t)$ for computing the inverse likelihood of the observation s_{t+1} (Schmidhuber, 1991a; Lopes et al., 2012; Stadie et al., 2015; Achiam & Sastry, 2017; Pathak et al., 2017):

$$\text{Surprise}(s_{t+1} \mid s_t, a_t) \propto p_\xi(s_{t+1} \mid s_t, a_t)^{-1}. \tag{2}$$

But just seeking novelty or surprise faces a problem. Consider static TV noise: it is highly entropic and unpredictable, hence remains highly surprising and novel. To address this, it has been suggested to consider an agent's *learning progress* (Schmidhuber, 1991a; Storck et al., 1995; Lopes et al., 2012; Achiam & Sastry, 2017; Pathak et al., 2019). Since collecting more data in this environment will not lead to better prediction of the noisy observation, an explorer

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

should avoid these areas. We should rather track and optimize the learning progress of the agent, which has been defined as

$$\log p_{\xi'}(s_{t+1} \mid s_t, a_t) - \log p_{\xi}(s_{t+1} \mid s_t, a_t) \tag{3}$$

where ξ' and ξ are the updated and old parameters, after cq. before observing some new data. A related formulation of learning progress can be derived from a Bayesian definition of surprise (information gain) that similarly has been used for exploration (Houthooft et al., 2016).

While learning progress is an important insight, we would like to offer an alternative reason as to why the noisy TV screen is uninteresting for an exploring agent: it is really the failure of increasing its capability to interact with the world. While improving one's world model is an absolutely vital part of that process, it alone is not sufficient – the goal is not just to become a perfect simulator. Even a predictable pattern (e. g. a movie) remains uninteresting without meaningful interaction (e. g. TV remote) or without informing the agent about how to act in the world. As another example, consider an agent exploring a chair. It may not be necessary to predict every nuance of its physical appearance, but it is crucial to find out about its important practical uses such as exploring which surface is suitable for sitting.

One formalism to measure an agent's capability to interact with the world is called *empowerment* (Klyubin et al., 2005a;b). It is an information-theoretic concept that measures the maximal flow of information over an agent's perception–action loop. It is formally defined as the channel capacity between reachable terminal states S' and the possible action sequences $A_{1:T}$ over a horizon T:

$$\mathcal{E}^T(s) = \max_{A_{1:T}} \mathcal{I}(S', A_{1:T} \mid s) \tag{4}$$

where $\mathcal{I}$ denotes the mutual information. Intuitively, an agent is empowered if it can predictably reach many states, and it has low empowerment if its actions have little to no perceived influence on the world.

In this work, we propose a novel criterion for exploratory behaviour in autonomous agents which we call *Empowerment Gain (EG)*. Its goal can be stated as follows: in the absence of a concrete task, an agent should take those actions in the environment which provide the most information for improving its empowerment estimator, i. e. it should take those actions that maximise the increase in perceived influence over its environment after observing new data. *New experiences should help me recognise my capability to interact with the world.* We show how EG combines and puts into relation common notions of novelty seeking, surprise maximisation and learning progress in a single formulation. This formulation takes into account an agent's limitations in actuation and sensation as well as inherent stochasticity of the environment. We provide a number of illustrative experiments in simple and discrete environments which support our theoretical findings and provide some more intuition about the EG criterion. In particular, we show how EG guides an agent towards those areas of the state and action space where it has more potential for improving its influence in the world. Since EG considers an extended time horizon T instead of a single time step, we show that it tends to avoid inescapable traps in the environment that limit its future control. Both of these behaviours cannot be achieved by novelty seeking, surprise maximisation or learning progress alone.

2. Background

2.1. Empowerment

Empowerment (Klyubin et al., 2005a;b; Salge et al., 2014) is a measurement of an agent's perceived control over its environment. It is defined in terms of an agent's embodiment; the coupling of sensors and actuators via the environment (perception–action loop).

Definition 1 *Empowerment for state s is defined as the channel capacity between terminal state distribution S' and the possible action sequences $A_{1:T}$ over a time horizon T:*

$$\mathcal{E}^T(s) = \max_{A_{1:T}} \mathcal{I}(S', A_{1:T} \mid s) \tag{5}$$

$$= \max_{A_{1:T}} [\mathbb{H}(S' \mid s) - \mathbb{H}(S' \mid s, A_{1:T})] \tag{6}$$

$$= \max_{A_{1:T}} [\mathbb{H}(A_{1:T} \mid s) - \mathbb{H}(A_{1:T} \mid s, S')] \tag{7}$$

where $\mathcal{I}$ denotes mutual information and $\mathbb{H}$ denotes entropy.

Interpreting the environment as a communication channel, the agent finds a *source distribution* over action sequences $A_{1:T}$ such that its sensors at time $T+1$ can recover the most information about them. Writing the mutual information as differences of entropies gives an intuitive understanding. Interpreting eq. (6), empowerment finds balance between diversity and predictability. It wants to maximise the diversity of reachable states (entropy of final states S') while still being able to predict the outcome when conditioned on the taken action sequence. In highly stochastic environments, our empowerment will always be relatively small, because even though we (accidentally) reach a lot of states, we can not predict the outcome. By symmetry of mutual information, in eq. (7) we want to maximise the diversity of action sequences, but every action sequence should ideally lead to a unique final state s' such that the action sequence can be recovered from first and final state. For discrete and deterministic environments, empowerment simplifies significantly and becomes proportional to the number of reachable states within a horizon T and the source distribution learns how to reach those states uniformly. In this sense, we can view an action sequence $a_{1:T}$ as a skill that transforms the

world state from s to s' and our source distribution as a distribution over those skills.

3. Empowerment Gain (EG)

In this section, we develop an exploration criterion based on expected improvement of an agent's empowerment estimator. We show how it encapsulates and relates other intrinsic motivations such as novelty seeking, surprise maximisation, learning progress and information gain. We start out by looking at the components that make up empowerment estimation:

$$\hat{\mathcal{E}}_\theta^T(s) = \max_{\omega_\phi(a_{1:T}|s)} \mathbb{E}_{\omega_\phi(a_{1:T}|s)p_\xi(s_{T+1}|s,a_{1:T})}\big[\log p_\xi(s_{T+1} \mid s, a_{1:T}) - \log p_{\phi,\xi}(s_{T+1} \mid s)\big]. \tag{8}$$

This estimation consists of 3 distributions, the *source distribution* $\omega_\phi(a_{1:T} \mid s)$, the transition or *forward model* $p_\xi(s_{T+1} \mid s, a_{1:T})$ and the *final state marginal* distribution $p_{\phi,\xi}(s_{T+1} \mid s)$ where we have split our parameters $\theta = \{\phi, \xi\}$.

The idea of empowerment gain is that an agent should act such that it maximises its expected improvement of its empowerment estimator. Concretely, in any given state s, it should perform action a^* such that it maximises its expected improvement of its empowerment estimator $\hat{\mathcal{E}}_\theta^T(s)$ after updating the parameters θ using the newly collected data d. Formally, the objective is defined as

$$a^* = \arg\max_a \mathbb{E}_{d=(s,a,s'\sim p(s'|s,a))}\Big[\hat{\mathcal{E}}_{\theta'}^T(s) - \hat{\mathcal{E}}_\theta^T(s)\Big] \tag{9}$$

where θ' are the updated parameters after observing new data d. By rewriting this objective in various ways, we develop a deeper understanding and discover that novelty seeking and learning progress of a forward or inverse model are already contained as part of the EG objective.

3.1. Empowerment Gain in a Maximum Likelihood Setting

Let us rewrite EG in a setting that is broadly applicable to any parametrised estimate of one's empowerment:

$$\hat{\mathcal{E}}_{\theta'}^T(s) - \hat{\mathcal{E}}_\theta^T(s) \tag{10}$$

$$\approx \mathbb{H}\Big(S_{T+1}^{\phi',\xi'} \;\Big|\; s\Big) - \mathbb{H}\Big(S_{T+1}^{\phi,\xi} \;\Big|\; s\Big) \tag{11}$$

$$+ \mathbb{E}_{\omega^*_{\phi'}(a_{1:T}|s)}\big[\mathrm{KL}(p_{\xi'}(s_{T+1} \mid s, a_{1:T}) \,||\, p_\xi(s_{T+1} \mid s, a_{1:T}))\big] \tag{12}$$

where $S_{T+1}^{\phi',\xi'}$ is defined by pdf

$$\int \omega^*_{\phi'}(a_{1:T} \mid s) p_{\xi'}(s_{T+1} \mid s, a_{1:T})\, \mathrm{d}a_{1:T}$$

where the approximation is relatively tight for small updates of ϕ and ξ (for derivation see Appendix C). We see that the empowerment estimate can be improved in two distinct ways. First by (local) novelty seeking, i. e. by increasing reachable state entropy (eq. (10)), which can be achieved by either detecting new states or by realising how to reach them more uniformly. Second, by improving the forward model (learning progress) in eq. (12). Different to other methods focussing on learning progress on a forward model, the divergence between current and past forward model is put into context by the empowerment-realising source distribution $\omega_{\phi'}(a_{1:T} \mid s)$. That is, the improvement of the forward model is weighted in so far as it helps the agent realise its influence in the world. So even if learning progress of a forward model is large, if it does not help the agent in attaining states more reliably, empowerment gain may still be small. Inversely, a small update in the forward model can still lead to a large empowerment gain if the improved transition is central to many skills (samples of $\omega_{\phi'}(a_{1:T} \mid s)$). Jointly optimising these two terms results in a tradeoff between novelty seeking and learning progress.

4. Experiments

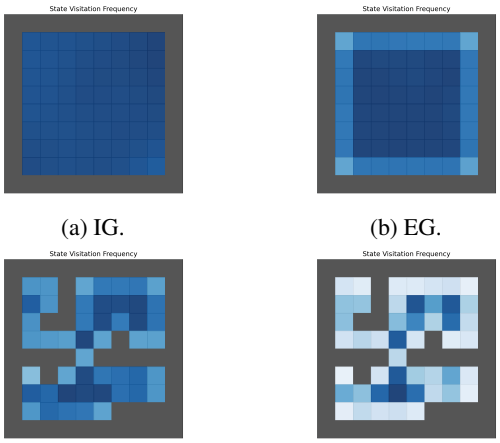
In our experiments we investigated the exploratory behaviour of an EG-maximising agent in various simple but illustrative grid world scenarios. We looked at the effect of various types of action noise, the shape of the state space, traps (sinks in the state space the agent cannot get out of) and empowerment's computation horizon. We compared this to an agent maximising its information gain on a forward model (IG explorer) and to a surprise-based explorer maximising the prediction error (PE explorer). Our main goal was to illustrate and compare what these exploration criteria would do in a perfect information setting, but it should be noted that efficient and scalable algorithms may be developed on top of existing work on empowerment approximation (Mohamed & Rezende, 2015; Karl et al., 2019). This, however, is not the focus of our work. Thus, for our purposes we remain in a simple grid world setting that allows for exact computation of empowerment gain, information gain and closed-form solutions for model updates. This setting also leads to the exploration algorithm (see Algorithm 1 in Appendix A) where we made some simplifying assumptions. In particular, when we explore, we first try out each action an agent could take in the real world, update our parametrised model(s), and compute either expected empowerment gain or information gain. Then we only execute and count the best action according to the respective criterion. While this is impractical, we think this study is very illustrative of the behaviour induced by the various exploration criteria and gives and intuitive understanding of the theoretical results of the previous section.

Our grid worlds, built on top of MiniGrid (Chevalier-Boisvert et al., 2018), have a discrete state space $\mathcal{S}$ and

action space $\mathcal{A}$. The agent may perform five different actions that move to a neighbouring cell (left, right, up, down) or keep the agent where it is (stay). An action that leads into a wall has no effect on the agent's position. As observations we use the global x and y coordinates unless stated otherwise.

For estimating EG and IG, we maintain a Bayesian forward model. For each cell, the forward model is realised by a Categorical distribution with a Dirichlet prior. The parametrisation of the forward model is, of course, essential to the resulting behaviour of these exploration criteria and should be taken into account when interpreting the results. With our choice, the agent cannot generalise information among states as would be the case with e. g. a neural network. We use the iterative Blahut-Arimoto algorithm to directly compute an empowerment estimate, including the distribution parameters of the source $\omega_\phi(a_{1:T} \mid s)$ and inverse model $p_\psi(a_{1:T} \mid s, s_{T+1})$. Information gain can be computed in closed form since the Dirichlet distribution is a conjugate prior to the Categorical distribution. For more details, we refer to Appendix A.

4.1. Deterministic Environments



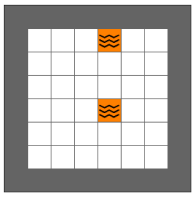
(a) IG. (b) EG.

(c) EG, horizon $T = 1$. (d) EG, horizon $T = 2$.

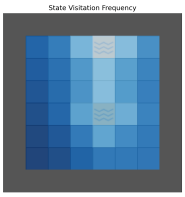
Figure 1: Comparison of state visitation frequencies of the EG and IG explorer in deterministic environments.

We start out by discussing the fully observable and deterministic setting. For IG, in every state an agent performs each action uniformly which also leads to a uniform visitation frequency of every state (see Figure 1a). In contrast, rather than performing each action uniformly, the EG explorer, from any state, visits each reachable state uniformly. This leads to a visitation frequency that qualitatively resembles the empowerment landscape for the environment in that more empowered states (centre of the room) are visited more frequently during exploration than less empowered states (bordering walls). Importantly, the EG explorer still makes sure to visit every state and does not remain strictly on high empowered states, it just shifts the exploration focus towards them. This remains true for more complicated wall structures and is also affected by the empowerment horizon as shown in Figures 1c and 1d. Since a state being more empowerment means having more ways to interact with the world, exploring these states more thoroughly should in general be more helpful for downstream tasks.

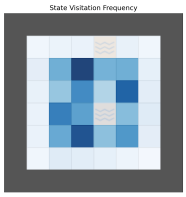
4.2. Sinks in the State Space



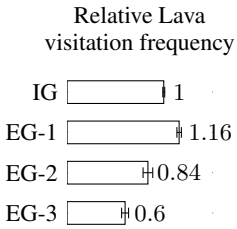
(a) Environment.



(b) IG.



(c) EG, horizon $T = 3$.



(d) Comparison plot.

Figure 2: (b) IG does not avoid lava, its visitation frequency skews away from the lava in so far as the random walk resets randomly whenever the agent enters the lava. Similarly, EG over horizon $T = 1$ cannot foresee and avoid the lava trap. (c) EG over horizon $T = 3$ actively avoids the lava trap to some extent. (d) shows how often lava states are visited compared to all other states (normalized by the ratio of the IG explorer).

What happens if we have detrimental states in the environment that severely limit control? Avoiding such states should generally be helpful for exploration. We investigated the extreme scenario of perfect sinks where an agent, once entered, cannot get out again. In our grid world, we model these states as lava cells (Figure 2a). We can see in Figure 2b that, while the IG explorer skews away from the lava cells to some degree, this is just due to the random walk being reset randomly whenever an agent enters the lava. For the EG explorer, avoidance of lava is actively pursued for an empowerment horizon greater than 1. One may ask: why is EG not fully avoiding the lava since the agent's empowerment in this state is 0 as its actions have no influence at all? This is because the lava cell is still a state in and of itself, and it is still more empowering to know that you can reach that state more reliably. However, improving empowerment

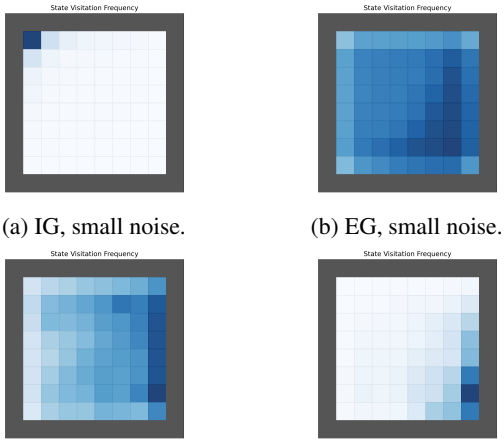
after having entered a lava cell is impossible and hence the increase of the estimated empowerment for the rest of the trajectory is 0. That's why a horizon of 1 is not enough to realise the detrimental effect and longer horizon skew more and more away from the lava as shown in Figure 2d. Interestingly, a purely novelty seeking agent would still aim for a uniform visitation frequency, visiting lava cells even more frequently than the IG explorer.

4.3. Actions of Varying Reliability

Next, we investigated the influence of varying action noises, i. e. in any state different actions are augmented with different level of noise. Concretely, when taking action a, there is a probability p_a that a random action $a' \in \mathcal{A}$ is performed instead. We share the same action noise model for every state in the environment. For IG exploration, we observe that the more noisy action is taken more frequently (Figure 3). This is because occasionally we observe an unexpected event which leads to a bigger change to the model and thus to high information gain. Similarly, surprise-based exploration (PE explorer) prefers the noisy action as well because the prediction error for noisy actions remains larger than for less noisy ones. Interestingly, expected EG skews in the opposite direction, generally preferring the more reliable actions as they have greater potential of increasing one's empowerment estimate. This effect augments the results we found in the deterministic setting and also depends on the empowerment's horizon, which push the agent away from the walls (see Figures 3b to 3d). Again, we argue that this exploration behaviour is more desirable as exploring the actions that have a more reliable effect should in general be more helpful for interacting with the world.

4.4. State-Dependent Action Noise

Let us now look at state-dependent action noise, in which all actions in one state share the same noise model, but differ amongst different states. For the IG and PE explorer we observe the same behaviour as in the deterministic setting, as these criteria cannot see ahead enough; after all, all actions in any state have an identical effect (except next to a wall). Conversely, we found that an EG explorer skews towards the part of the state space with less noise (Figure 4). While it still explores the whole state space, it focuses more on areas with greater potential for influencing the world. However, as with the previously discussed sinks (lava), EG needs to be computed over at least a horizon of 2 in order to realise that going towards the less noisy region helps us with increasing our empowerment estimate to a greater degree. For this experiment, we used a slightly different objective that considered the impact of action sequences on the empowerment estimate instead of just an individual actions as we detail in Appendix B.



(a) IG, small noise. (b) EG, small noise.

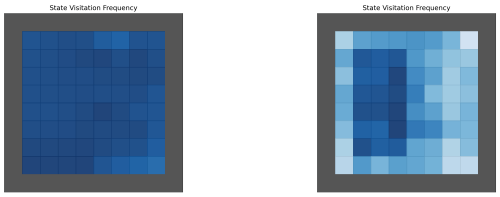
(c) EG, medium noise. (d) EG, large noise.

Figure 3: Comparison of different action noises, there is no noise for actions going down and right, but there is noise going up (with probability p_a) and left (with probability $2p_a$). (a) IG is much more attracted in performing the more noisy actions. (b) When introducing a bit of noise, EG prefers going to the bottom right, but still stays away from the walls as in the deterministic setting. (c, d) When noise on the actions is increased even further, at some point this effect dominates.

5. Related work

Numerous ways of motivating exploration have been suggested over the years. Among the first was Schmidhuber (1991b) which based reward signals on prediction errors of predictive models. This prediction error has been used in various ways using function representation learning (Stadie et al., 2015; Pathak et al., 2017). Focussing purely on prediction error has the downside of being attracted unpredictable traps (static TV noise problem), where the error remains large even if it has been observed countless times. Hence, learning progress of a predictive model has been suggested and formalised in various works (Schmidhuber, 1991a; Lopes et al., 2012; Stadie et al., 2015; Achiam & Sastry, 2017). Houthooft et al. (2016) formulates learning progress in a Bayesian setting as information gain. More recently, ensembles of predictors have been used to form some kind of internal disagreement metric as opposed to directly comparing to the real world difference (Pathak et al., 2019; Shyam et al., 2019; Sekar et al., 2020).

Different from these surprise based approaches, various ways of novelty seeking have similarly been formalised over the years. Bellemare et al. (2016) (extended by Ostrovski et al. (2017)) proposed using a pseudo-count using density estimation for novelty estimation in high-dimensional state spaces. Tang et al. (2017) suggested a computationally simple but effective generalization of count-based approaches



(a) IG. (b) EG, horizon $T = 2$.

Figure 4: (a) - (b) The left half of the room is without noise, while the right half is augmented by linearly increasing state-dependent action noise, i. e. the column right of the center line has noise probability $p_a = 0.1$ and the rightmost column of the room $p_a = 0.4$. State visitation frequency for EG is skewed towards the less noisy part of the state space while IG exploration cannot account for this effect.

using hashing. Lee et al. (2019) recast exploration as a problem of state marginal matching, where they aim to learn a policy for which the state marginal distribution matches a given target state distribution (often a uniform distribution). Savinov et al. (2018) defines novelty through reachability within a certain time horizon from observations in a memory.

However, there are also other exploration approaches based on skill acquisition that are related to our method. Sharma et al. (2019) encourages novel skill acquisition via mutual information maximisation between skills and states. DIAYN (Eysenbach et al., 2018) encourages exploration by incentivising different skills that lead to distinguishable outcomes. Empowerment (Klyubin et al., 2005a;b) as an intrinsic motivation has been gaining a more attention over the recent years and various ways of approximation have been suggested (Mohamed & Rezende, 2015; Gregor et al., 2016; Karl et al., 2019; Zhao et al., 2021).

6. Conclusion

In this work, we propose a novel framework for embodiment driven exploration. Building on the universally applicable and information-theoretic measure of an agent's perceived influence in the world – empowerment – we suggest an exploration criterion based on the expected improvement of one's estimation thereof. Theoretically, we show that it captures and puts into relation many previously suggested exploration criteria such as novelty seeking, surprise maximisation and learning progress. Different to the use of these individual criteria, EG accounts for the agent's current capabilities and boundedness, and focuses directly on what we are interested in – recognising one's capability to interact with the world. In various discrete grid world environments featuring different noise models, we showcase our theoretical findings where an agent's exploration is focused on areas with greater potential for increasing its influence. For future work, the focus will be on smart approximations that can capture the spirit of EG while being scalable to scenarios of interest for current state of the art research.

References

Achiam, J. and Sastry, S. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*, 2017.

Aubret, A., Matignon, L., and Hassas, S. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019.

Barto, A., Mirolli, M., and Baldassarre, G. Novelty or surprise? *Frontiers in psychology*, 4:907, 2013.

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying Count-Based Exploration and Intrinsic Motivation. In *Advances in Neural Information Processing Systems 29*, pp. 1471–1479, 2016.

Berlyne, D. E. Conflict, arousal, and curiosity. 1960.

Chevalier-Boisvert, M., Willems, L., and Pal, S. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. February 2018.

Gregor, K., Rezende, D. J., and Wierstra, D. Variational Intrinsic Control. *arXiv preprint arXiv:1611.07507*, 2016.

Houthooft, R., Chen, X., Duan, Y., Schulman, J., Turck, F. D., and Abbeel, P. Vime: Variational information maximizing exploration. May 2016.

Karl, M., Becker-Ehmck, P., Soelch, M., Benbouzid, D., van der Smagt, P., and Bayer, J. Unsupervised Real-Time Control through Variational Empowerment. In *International Symposium on Robotics Research*, 2019.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. Empowerment: A universal agent-centric measure of control. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, volume 1, pp. 128–135. IEEE, 2005a.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. All else being equal be empowered. In *European Conference on Artificial Life*, pp. 744–753. Springer, 2005b.

Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S., and Salakhutdinov, R. Efficient exploration via state marginal matching. 2019.

Lopes, M., Lang, T., Toussaint, M., and Oudeyer, P.-Y. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Neural Information Processing Systems (NIPS)*, 2012.

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

Ostrovski, G., Bellemare, M. G., Oord, A., and Munos, R. Count-based exploration with neural density models. In *International conference on machine learning*, pp. 2721–2730. PMLR, 2017.

Oudeyer, P.-Y. and Kaplan, F. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2): 265–286, 2007.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pp. 2778–2787. PMLR, 2017.

Pathak, D., Gandhi, D., and Gupta, A. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, pp. 5062–5071. PMLR, 2019.

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

Savinov, N., Raichuk, A., Marinier, R., Vincent, D., Pollefeys, M., Lillicrap, T., and Gelly, S. Episodic curiosity through reachability. *arXiv preprint arXiv:1810.02274*, 2018.

Schmidhuber, J. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pp. 1458–1463, 1991a.

Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pp. 222–227, 1991b.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pp. 8583–8592. PMLR, 2020.

Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

Shyam, P., Jaśkowski, W., and Gomez, F. Model-based active exploration. In *International Conference on Machine Learning*, pp. 5779–5788. PMLR, 2019.

Stadie, B. C., Levine, S., and Abbeel, P. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

Storck, J., Hochreiter, S., and Schmidhuber, J. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the international conference on artificial neural networks, Paris*, volume 2, pp. 159–164. Citeseer, 1995.

Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. # exploration: A study of count-based exploration for deep reinforcement learning. In *31st Conference on Neural Information Processing Systems (NIPS)*, volume 30, pp. 1–18, 2017.

Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., and Herzog, M. H. Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *bioRxiv*, pp. 2020–09, 2021.

Zhao, R., Lu, K., Abbeel, P., and Tiomkin, S. Efficient empowerment estimation for unsupervised stabilization. In *International Conference on Learning Representations*, 2021.

A. Experiments

Algorithm 1 describes the exploration procedure in pseudocode form. We usually perform just one long rollout without resetting the environment, however in the scenario with lava, we reset the agent randomly once the agent is stuck (the agent can not observe this reset). Overall, the agent takes $50,000$ steps in the environment in a whole experiment. For IG exploration the algorithm works analogously, but naturally we compute information gain instead of empowerment gain in line 11. To limit the effect of randomness in stochastic environments, we collect 10 instead of just 1 sample in line 8 to collect data for the model update. For all experiments, we executed 10 separate runs to ensure reproducible results. As the results are quite stable across runs and our results are mostly qualitative in nature, we picked representative plots from individual runs mostly at random. Where quantitative measures are mentioned, they are averaged across runs. For the environments in Section 4.3, the noise levels are: small noise $p_a = 0.05$, medium noise $p_a = 0.1$ and large noise $p_a = 0.2$. There is

Algorithm 1 Exploration Algorithm

Initialise forward model $m = p_\xi(s_{T+1} \mid s, a_{1:T})$ with $Dirichlet(\alpha)$ prior, $\alpha = 0.01$.
for each episode **do**
  $s =$ `env.reset()`
  for t=1..T **do**
    `empow = blahut_arimoto(m)`
    for a in Actions **do**
      # Try out all actions
      `s' = env.step(a)`
      `m' = m.update(s, a, s')`
      `empow' = blahut_arimoto(m')`
      `gain[a] = empow' - empow`
    end for
    # Perform the best action for exploration
    `a = argmax`$_a$ `gain`
    `s' = env.step(best_action)`
    `m = m.update(s, a, s')`
  end for
end for

noise when performing action "up" (with probability p_a) and "left" (with probability $2p_a$). The forward model of each cell is modelled by a Categorical distribution with a Dirichlet(α) prior with $\alpha = 0.01$. For empowerment estimation, Blahut-Arimoto is run until the change of empowerment from one iteration to the next is less than $1\mathrm{e}{-6}$ or after at most 1000 iterations, whichever comes first.

B. An alternative formulation

As an alternative objective to eq. (9), we want to discuss a different formulation that focuses on action sequences instead of individual actions:

$$a^*_{1:T} = \arg\max_{a_{1:T}} \mathbb{E}_{d=(s_1,a_1,\ldots,a_T,s_{T+1})}\Big[\hat{\mathcal{E}}^T_{\theta'}(s_1) - \hat{\mathcal{E}}^T_{\theta}(s_1)\Big].$$

Different to the original formulation, here we are comparing and choosing the best action sequence over an horizon T that matches the empowerment's horizon. Intuitively, as we compute the empowerment over a specific horizon, it might be helpful to consider how execution of a whole skill over the same horizon instead of a single action impacts our empowerment estimate. In particular when our forward model does not have the capacity to generalise over the state-action space, comparing EG after and before experiencing a single state action transition modifies empowerment only in so far as it changes the transition probabilities from the starting state. While choosing among skills may be desirable, it problematically comes with greater computational cost. Naively computing the argmax becomes exponentially more expensive with the horizon T as we now need to compare all possible action sequences instead of individual actions.

C. Derivations

The approximation of eq. (10) is relatively tight for small updates of $\omega^*_\phi(a_{1:T} \mid s)$ and $p_\xi(s_{T+1} \mid s, a_{1:T})$:

$$\begin{aligned}
&\hat{\mathcal{E}}^T_{\theta'}(s) - \hat{\mathcal{E}}^T_{\theta}(s) \\
&= \max_{\omega_{\phi'}(a_{1:T}|s)} \int_{a_{1:T}} \int_{s_{T+1}} \omega_{\phi'}(a_{1:T} \mid s) p_{\xi'}(s_{T+1} \mid s, a_{1:T}) \\
&\qquad \log \frac{p_{\xi'}(s_{T+1} \mid s, a_{1:T})}{p_{\phi',\xi'}(s_{T+1} \mid s)} \\
&\quad - \max_{\omega_{\phi}(a_{1:T}|s)} \int_{a_{1:T}} \int_{s_{T+1}} \omega_{\phi}(a_{1:T} \mid s) p_{\xi}(s_{T+1} \mid s, a_{1:T}) \\
&\qquad \log \frac{p_{\xi}(s_{T+1} \mid s, a_{1:T})}{p_{\phi,\xi}(s_{T+1} \mid s)} \\
&\quad \omega^*_{\phi'}(a_{1:T} \mid s) \text{ and } \omega^*_{\phi}(a_{1:T} \mid s) \text{ be the resp. max. dist.} \\
&= \int_{a_{1:T}} \int_{s_{t+1}} \omega^*_{\phi'}(a_{1:T} \mid s) p_{\xi'}(s_{T+1} \mid s, a_{1:T}) \\
&\quad \log p_{\xi'}(s_{T+1} \mid s, a_{1:T}) + \mathbb{H}\Big(S^{\phi',\xi'}_{T+1} \,\Big|\, s\Big) \\
&\quad - \Bigg(\int_{a_{1:T}} \int_{s_{T+1}} \omega^*_{\phi}(a_{1:T} \mid s) p_{\xi}(s_{T+1} \mid s, a_{1:T}) \\
&\quad \log p_{\xi}(s_{T+1} \mid s, a_{1:T}) + \mathbb{H}\Big(S^{\phi,\xi}_{T+1} \,\Big|\, s\Big) \Bigg) \\
&= \mathbb{H}\Big(S^{\phi',\xi'}_{T+1} \,\Big|\, s\Big) - \mathbb{H}\Big(S^{\phi,\xi}_{T+1} \,\Big|\, s\Big) \\
&\quad + \int_{a_{1:T}} \int_{s_{T+1}} \omega^*_{\phi'}(a_{1:T} \mid s) p_{\xi'}(s_{T+1} \mid s, a_{1:T}) \\
&\qquad \log p_{\xi'}(s_{T+1} \mid s, a_{1:T}) \\
&\quad - \int_{a_{1:T}} \int_{s_{T+1}} \omega^*_{\phi'}(a_{1:T} \mid s) p_{\xi'}(s_{T+1} \mid s, a_{1:T}) \Bigg[\\
&\quad \frac{\omega^*_{\phi}(a_{1:T} \mid s) p_{\xi}(s_{T+1} \mid s, a_{1:T})}{\omega^*_{\phi'}(a_{1:T} \mid s) p_{\xi'}(s_{T+1} \mid s, a_{1:T})} \log p_{\xi}(s_{T+1} \mid s, a_{1:T}) \Bigg] \\
&\quad \text{with } \alpha = \frac{\omega^*_{\phi}(a_{1:T} \mid s) p_{\xi}(s_{T+1} \mid s, a_{1:T})}{\omega^*_{\phi'}(a_{1:T} \mid s) p_{\xi'}(s_{T+1} \mid s, a_{1:T})} \\
&= \mathbb{H}\Big(S^{\phi',\xi'}_{T+1} \,\Big|\, s\Big) - \mathbb{H}\Big(S^{\phi,\xi}_{T+1} \,\Big|\, s\Big) \\
&\quad + \int_{a_{1:T}} \omega^*_{\phi'}(a_{1:T} \mid s) \int_{s_{T+1}} p_{\xi'}(s_{T+1} \mid s, a_{1:T}) \big[\\
&\qquad \log p_{\xi'}(s_{T+1} \mid s, a_{1:T}) - \alpha \log p_{\xi}(s_{T+1} \mid s, a_{1:T}) \big] \\
&= \mathbb{H}\Big(S^{\phi',\xi'}_{T+1} \,\Big|\, s\Big) - \mathbb{H}\Big(S^{\phi,\xi}_{T+1} \,\Big|\, s\Big) \\
&\quad + \mathbb{E}_{\omega^*_{\phi'}(a_{1:T}|s)} [\mathrm{KL}(p_{\xi'}(s_{T+1} \mid s, a_{1:T}) \,||\, p_{\xi}(s_{T+1} \mid s, a_{1:T}))] \\
&\quad + \mathbb{E}_{\omega^*_{\phi'}(a_{1:T}|s) p_{\xi'}(s_{T+1}|s,a_{1:T})} [(\alpha - 1) \log p_{\xi}(s_{T+1} \mid s, a_{1:T})] \\
&\approx \mathbb{H}\Big(S^{\phi',\xi'}_{T+1} \,\Big|\, s\Big) - \mathbb{H}\Big(S^{\phi,\xi}_{T+1} \,\Big|\, s\Big) \\
&\quad + \mathbb{E}_{\omega^*_{\phi'}(a_{1:T}|s)} [\mathrm{KL}(p_{\xi'}(s_{T+1} \mid s, a_{1:T}) \,||\, p_{\xi}(s_{T+1} \mid s, a_{1:T}))]
\end{aligned}$$