EFFICIENT EMPOWERMENT ESTIMATION FOR UNSUPERVISED STABILIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Intrinsically motivated artificial agents learn advantageous behavior without externally-provided rewards. Previously, it was shown that maximizing mutual information between agent actuators and future states, known as the empowerment principle, enables unsupervised stabilization of dynamical systems at upright positions, which is a prototypical intrinsically motivated behavior for upright standing and walking. That follows from the coincidence between the objective of stabilization and the objective of empowerment. Unfortunately, sample-based estimation of this kind of mutual information is challenging. Recently, various variational lower bounds (VLBs) on empowerment have been proposed as solutions; however, they are often biased, have high sample complexity, and are unstable in training. In this work, we propose an alternative solution based on a trainable representation of a dynamical system as a Gaussian channel, which allows us to efficiently calculate an unbiased estimator of empowerment by convex optimization. We demonstrate our solution for sample-based unsupervised stabilization on different dynamical control systems and show the advantages of our method by comparing it to the existing VLB approaches. Specifically, we show that our method has a lower sample complexity, is more stable in training, possesses the essential properties of the empowerment function, and allows, for the first time, to estimate empowerment from images. Consequently, our method opens a path to wider and easier adoption of empowerment for many applications.

1 INTRODUCTION

Intrinsic motivation allows artificial and biological agents to acquire useful behaviours without external knowledge (Barto et al. (2004); Chentanez et al. (2005); Schmidhuber (2010); Barto (2013); Oudeyer et al. (2016)). In the framework of reinforcement learning (RL), this external knowledge is usually provided by an expert through a task-specific reward, which is optimized by an artificial agent towards a desired behavior (Mnih et al. (2013); Schulman et al. (2017)).

In contrast, an intrinsic reward can arise solely from the interaction between the agent and environment, which eliminates, in certain scenarios, the need for domain knowledge and reward engineering (Mohamed & Rezende (2015); Houthooft et al. (2017); Pathak et al. (2017)). Previously, it was shown that maximizing mutual information (Cover & Thomas (2012)) between an agent's actuators and sensors reveals preferable states in the environment that have a higher potential to achieve a larger number of eventual goals (Klyubin et al. (2005); Wissner-Gross & Freer (2013)).

Maximizing this kind of mutual information is known as the empowerment principle (Klyubin et al. (2005); Salge et al. (2014)). Empowerment is a function of the state. Previously, it was found that an agent maximizing its empowerment converges to an unstable equilibrium of the environment which was demonstrated on various dynamical control systems (Jung et al. (2011); Salge et al. (2013); Maximilian et al. (2019)). In this sense, ascending the gradient of the empowerment function towards unstable equilibrium coincides with the objective of optimal control for stabilization at unstable equilibrium (Strogatz (2018)), which is an important task for both engineering (Todorov (2006)) and biological[1] systems. It follows from the above mentioned prior works on empowerment that a plausible estimate of the empowerment function should possess this essential property.

[1]Broadly speaking, an increase in the rate of information flow in the perception-action loop (Tishby & Polani (2011)) could be an impetus for the development of *homosapiens*, as hypothesized in (Yuval (2014)).

Empowerment has been found to be useful for a broad spectrum of applications. Among others are unsupervised skill discovery (Sharma et al. (2020); Eysenbach et al. (2019); Gregor et al. (2017); Maximilian et al. (2019); Campos et al. (2020)), human-agent coordination (Salge & Polani (2017); Guckelsberger et al. (2016)), assistance (Yuqing et al. (2021)), and stabilization (Tiomkin et al. (2017)). In general, in these applications, the agent maximizes mutual information between its skills and one of: (1) its actuators, which generate either single actions (Maximilian et al. (2019); Gregor et al. (2017)) or (2) action sequences (Tiomkin & Tishby (2017); Salge et al. (2014)); (3) options (Gregor et al. (2017)); or (4) skills (Sharma et al. (2020); Eysenbach et al. (2019)).

There exist two possibilities for the estimation and use of empowerment: sequential and concurrent. In sequential mode, one first estimates empowerment before using it and/or its byproducts for downstream tasks (Sharma et al. (2020); Eysenbach et al. (2019)). In concurrent mode, one estimates and uses empowerment simultaneously, which is adopted by our current work. The concurrent mode is suitable for scenarios when the empowerment objective coincides with the control objective, which happens for instance in stabilization.

Estimation of mutual information is hard (McAllester & Statos (2018)). Fortunately, recent advances in deep neural networks (Goodfellow et al. (2016)) have brought new opportunities for the estimation of mutual information from samples by means of variational lower bounds (VLBs) (Poole et al. (2019); Alemi et al. (2017); Tschannen et al. (2020); Mohamed & Rezende (2015)). However, VLB approaches to empowerment in dynamical control systems (Sharma et al. (2020); Achiam et al. (2018)) have high sample complexity, are often unstable in training, and may be biased. Moreover, it was not studied previously whether VLBs on empowerment preserve the essential properties of the empowerment function.

In this work, we introduce a new method for empowerment estimation and utilize the above-mentioned property as an "indicator" for the quality of the estimation. Specifically, we propose a particular representation for dynamical control systems using deep neural networks which is learned from state-action trajectories. This representation admits an efficient estimation of empowerment by convex optimization (Cover & Thomas (2012)), both from raw state and from images. We propose an algorithm for simultaneous estimation and maximization of empowerment by the standard optimizers such as Proximal Policy Optimization (PPO), (Schulman et al. (2017)) and Soft Actor-Critic (SAC) (Haarnoja et al. (2018)). We test our method on the task of unsupervised stabilization of dynamical systems with using only the intrinsic reward, showing our estimator exhibits essential properties of the empowerment function.

We believe our experiments clearly convey the gist of our method without losing the generality our estimator beyond the stabilization task; we discuss more general applications in the future directions. We demonstrate the advantages of our method thorough comparisons to the existing state-of-the-art empowerment estimators in different dynamical systems from the MuJoCo simulator (Todorov et al. (2012)). We found that our method (i) has a lower sample complexity, (ii) is more stable in training, (iii) possesses the essential properties of the empowerment function, and (iv) allows us for the first time to estimate the empowerment function from images. In addition, this work provides for a novel benchmark between different estimators of empowerment, specifically comparing their properties. We hope such a review of the existing methods for empowerment estimation will be helpful to push further this research direction.

The paper is organized as follows. In Section 2 we review the necessary background on reinforcement learning, empowerment, and relevant existing works. In Section 4 we explain our method. In Section 5, we provide numerical simulations of our method and comparison between our method to the existing state-of-the-art. In Section 6 we summarize the paper, discuss the limitations of our method, and propose the continuation which is beyond the scope of the current work.

2 PRELIMINARIES

In this section, we review the necessary background for our method, consisting of the reinforcement learning setting, various empowerment estimators, and the Gaussian channel capacity. We also review the underlying components in relevant prior work which we use for comparison to our method.

2.1 REINFORCEMENT LEARNING

The reinforcement learning (RL) setting is modeled as an infinite-horizon Markov Decision Process (MDP) defined by: the state space $\mathcal{S}$, the action space $\mathcal{A}$, the transition probabilities $p(s'|s,a)$, the initial state distribution $p_0(s)$, the reward function $r(s,a) \in \mathbb{R}$, and the discount factor γ. The goal of RL is to find an optimal control policy $\pi(a|s)$ that maximizes the expected return, i.e. $\max_\pi \ \mathbb{E}_{s_0 \sim p_0, a_t \sim \pi, s_{t+1} \sim p}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$ (Sutton & Barto (2018)).

2.2 EMPOWERMENT

Interaction an agent and its environment is plausibly described by the perception-action cycle (PAC), (Tishby & Polani (2011)), where the agent observes the state of the environment via its sensors and responds with its actuators. Simply put, the maximal information rate from actuators to sensors is in an inherent property of PAC which characterizes *empowerment* of the agent, formally defined below.

Empowerment (Klyubin et al. (2005)) is defined by the maximal mutual information rate (Cover & Thomas (2012)) between agent sensor observations, $o \in \mathcal{O}$, and agent actuators, $a \in \mathcal{A}$, given the current state, $s \in \mathcal{S}$. It is fully specified by a fixed probability distribution of sensor observations conditioned on the actuators and the current state, $p(O \mid A, s)$, denoted by "channel", and by a free probability of the actuator signals conditioned on the current state, $\omega(A \mid s)$, denoted by "source". Empowerment, $\mathcal{E}$, is a capacity of the channel, $p(O \mid A, s)$, which is found by optimization of the source distribution, $\omega(A \mid s)$:

$$\mathcal{E}(s) = \max_{\omega(A|s)} \mathcal{I}[O; A \mid s] = \max_{\omega(A|s)} \sum_{O,A} p(O, A \mid s) \log\left(\frac{p(A \mid O, s)}{\omega(A \mid s)}\right) \tag{1}$$

where $\mathcal{I}[O; A \mid s]$ is the mutual information, reflecting the difference between the entropy of the sensor inputs, $\mathcal{H}(O \mid s)$ and the corresponding conditional entropy, $\mathcal{H}(O \mid A, s)$[2].

In this work, we follow the original formulation of empowerment (Klyubin et al. (2005)), where A is an action sequence of length T started at the state s, and O is the resulting sensor observation sequence. However, our method is applicable to other combinations of length of observation and control sequences.

Remark 1 *The source distribution in equation 1, $\omega^*(A \mid s)$, is used only for the estimation of empowerment. Another policy, $\pi(a \mid s)$, is calculated using empowerment as an intrinsic reward by either reinforcement learning (Maximilian et al. (2019)) or 1-step greedy algorithm (Jung et al. (2011); Salge et al. (2013)) to reach the states with maximal empowerment, as explained. Strictly speaking, the source distribution ω maximizes the mutual information, (i.e. estimates empowerment), while the policy distribution π maximizes expected accumulated empowerment along trajectories.*

In general, the capacity of an arbitrary channel for a continuous random variable is unknown except for a few special cases such as the Gaussian linear channel, as reviewed below.

2.2.1 EMPOWERMENT BY GAUSSIAN CHANNEL CAPACITY

The Gaussian linear channel in our case is given by:

$$O = G(s) \cdot A + \eta, \tag{2}$$

where O and A are the observation and actuator sequences with dimensions $d_O \times 1$ and $d_A \times 1$, respectively; $G(s)$ is a matrix with dimensions $d_O \times d_A$; and η is multivariate Gaussian noise with dimension $d_O \times 1$. Following Salge et al. (2013), we assume the noise is distributed by the normal distribution. For completeness, Appendix A provides the analytical derivation of $G(s)$ in the case of known dynamics for pendulum for the action sequence A with length 3. In contrast, the current work proposes a method how to estimate $G(s)$ for arbitrary dynamics and arbitrary length of the action sequence from samples.

[2] when the agent does not have an access to the true state, s, empowerment is calculated as the maximal mutual information between actuators and sensors given an estimated state.

The capacity of the channel in equation 2 is given by, (Cover & Thomas (2012)):

$$\mathcal{E}(s) = \max_{\substack{p_i \geq 0 \\ \sum_i^k p_i \leq P}} \frac{1}{2} \sum_{i=1}^{k} \log(1 + \sigma_i(s) p_i) \tag{3}$$

where $\sigma_i(s)$, p_i, k, and P, are the i-th singular value of $G(s)$, the corresponding actuator signal power, the number of nonzero singular values, and the total power of the actuator signal, respectively. The optimal solution to equation 3 is given by:

$$p_i^*(\nu) = \max(1/\nu - 1/\sigma_i(s), 0), \quad \sum_i^k p_i^*(\nu) = P, \tag{4}$$

where ν is found by a line search (Cover & Thomas (2012)) to satisfy the total power constraint. The key observation in our method is that it is possible to learn the matrix $G(s)$ from samples, which allows us to estimate empowerment efficiently by the line search in equation 4 even for high-dimensional dynamical systems either from raw state or from images (pixels). In Section 4 we describe how to represent $G(s)$ by deep neural networks and train them from data. The appropriateness of equation 2 for empowerment estimation has been studied in Salge et al. (2013), where the dynamics was known and the state was fully observable. These assumptions allow analytical computation of empowerment for certain dynamical systems, which provides us with the ground truth empowerment landscapes in Section 5.

2.2.2 Empowerment by Variational Lower Bounds

The mutual information in equation 1 can be bounded from below as follows (Mohamed & Rezende (2015); Gregor et al. (2017); Maximilian et al. (2019); Poole et al. (2019)):

$$\begin{aligned}\sum_{O,A} p(O, A \mid s) \log\Big(\frac{p(A \mid O, s)}{\omega(A \mid s)}\Big) &= \sum_{O,A} p(O, A \mid s) \log\Big(\frac{q(A \mid O, s)}{\omega(A \mid s)}\Big) + D_{KL}\left[p(A \mid O, s) \mid\mid q(A \mid O, s)\right] \\ &\leq \sum_{O,A} p(O, A \mid s) \log\Big(\frac{q(A \mid O, s)}{\omega(A \mid s)}\Big)\end{aligned} \tag{5}$$

where $q(A \mid O, s)$ and $\omega(A \mid s)$ are represented by neural networks with parameters ϕ and ψ respectively. The lower bound on empowerment in equation 5 can be estimated from samples by reinforcement learning with the intrinsic reward $r = \sum_t^T \log q_\phi(a_t \mid o_t, s) - \log \omega_\psi(a_t \mid s_t)$, as detailed in (Gregor et al. (2017); Maximilian et al. (2019)). To apply this variational lower bound on empowerment to unsupervised stabilization of dynamical systems one needs to learn three distributions: $q_\phi(a_t \mid o_t, s)$ and $\omega_\psi(a_t \mid s_t)$ for the lower bound in equation 5, and $\pi(a \mid s)$ as the control policy, as mentioned in Remark 1.

Developing the mutual information in equation 1 in the opposite direction, one gets another lower bound on empowerment:

$$\begin{aligned}\sum_{O,A} p(O, A \mid s) \log\Big(\frac{p(O \mid A, s)}{\omega(O \mid s)}\Big) &= \sum_{O,A} p(O, A \mid s) \log\Big(\frac{q(O \mid A, s)}{\omega(O \mid s)}\Big) + D_{KL}\left[p(O \mid A, s) \mid\mid q(O \mid A, s)\right] \\ &\leq \sum_{O,A} p(O, A \mid s) \log\Big(\frac{q(O \mid A, s)}{\omega(O \mid s)}\Big).\end{aligned} \tag{6}$$

An approximation of the lower bound in equation 6 is used in (Sharma et al. (2020)), where A is as a skill z and O is the next state after executing the skill z from the current state s.

For sake of the clearance of the exposition, the above-explained variational empowerment is schematically depicted in Figure 1. The controller policy, $\pi_\theta(a \mid s)$, collects transitions in environment, (s, a, s'). The source policy $\omega_\psi(a \mid s)$ generates the action a at every state s visited by the controller policy, and the environment transits to the next state, s' (yellow block). The inverse channel policy $q_\psi(a \mid s, a')$ is regressed to a given s and and s', generating the intrinsic reward, $C = \log \omega_\phi(a \mid s) - \log q_\psi(a \mid s, a')$, which is used as the reward to improve the controller policy via reinforcement learning. This scheme aims to transfer the intrinsically motivated agent to the state of the maximal empowerment in the environment.

3 Prior Art

In this section we review the relevant prior works, emphasizing the main differences between the existing art and the current work.

The representation of a dynamical system for empowerment estimation given by equation 2 was previously considered in (Salge et al. (2013)), where the channel, $p(O \mid A, s)$ was assumed to be known. In the current work, we remove this assumption and extend the applicability of the efficient channel capacity solution given by equation 3 to arbitrary observations, including images.

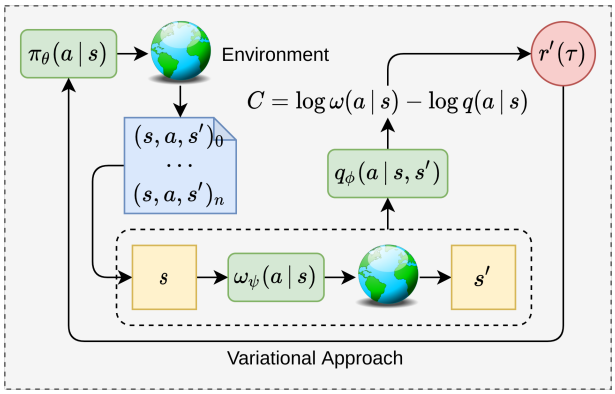


Figure 1: Variational empowerment maximization.

The lower bound in equation 5 is used in (Mohamed & Rezende (2015); Gregor et al. (2017); Maximilian et al. (2019)), where different additional assumptions are made. Specifically, Mohamed & Rezende (2015) estimates empowerment in open loop, while Gregor et al. (2017); Maximilian et al. (2019) provide a solution for closed-loop empowerment, and Maximilian et al. (2019) assumes on differentiability of the dynamics. In the current work, we do not assume differentiable dynamics and we do not resort to variational bounds, aiming to achieve a more stable, sample-efficient, and generalizable estimator of empowerment without losing the essential properties of the empowerment function, as explained in the Introduction. The works by (Mohamed & Rezende (2015); Gregor et al. (2017)) precede that by (Maximilian et al. (2019)). Consequently, Section 5 provides comparisons between our empowerment estimator and the latter representative of this kind of variational lower bound on empowerment.

An approximation of the lower bound in equation 6 is successfully used by Sharma et al. (2020) as described above. Section 5 provides comparisons between our empowerment estimator and Sharma et al. (2020) with regards to stability in training, sample efficiency, and correctness of the empowerment landscape generated by both methods. Although the work by Sharma et al. (2020) aims to discover a diverse set of skills using different machinery – which is not the main focus of our current work – their objective is of the same type us ours: to maximize the mutual information between actuators and states (or sensor). In Section 5 we show that the empowerment landscape derived by Sharma et al. (2020) converges to the true empowerment landscape in simple cases, while it does not converge to a known empowerment landscape in more complicated dynamical systems.

Another relevant and important work is Eysenbach et al. (2019). Citing the authors, their work is about *"maximizing the empowerment of a hierarchical agent whose action space is the set of skills"*. Specifically, they suggest maximizing a variational lower bound in the form of equation 5 on the mutual information between skills and states with an additional term of conditional entropy of actions given skills and states, $\mathcal{H}(A|S, Z)$, (which is equal to zero for a deterministic policy).

These works represent state-of-the-art empowerment estimation. We believe a thorough comparison between our method and this prior art, provided in Section 5, clearly demonstrate the main properties of our method, and clarify our contribution to the improvement of empowerment estimation in dynamical systems from samples. As mentioned in the introduction, empowerment is useful for a broad spectrum of different applications. In this section, we highlighted only the most relevant and updated methods, which we hope are helpful to see the merits and advances of our approach.s

Our improved empowerment estimation possesses the essential properties of the empowerment function; in particular, it converges to the correct empowerment landscape. The training has improved stability. Furthermore, for the first time, it allows for the estimation of the empowerment function from images, which we demonstrate in Section 5, where we show purely intrinsically motivated (empowerment-based) stabilization of inverted pendulum from images. This is significant progress in empowerment estimation which was impossible prior to this work. In the next section, we explain our method, referring to the background reviewed in this section.

4 PROPOSED METHOD

The desiderata for our method as follows:

- to learn the channel matrix ($G(s)$ in equation 2) from samples for an arbitrary dynamics.
- to utilize an existing optimizer for RL problems, such as SAC and/or PPO, to enable reliable large-scale training of deep neural networks.
- to enable empowerment maximization from raw states and/or images.

The rationale of these desired features as follows. The ability to learn $G(s)$ from samples would allow us to estimate empowerment by the efficient line research algorithm in equation 4. Training deep neural networks by RL is a hard problem (Mnih et al. (2013)). Sophisticated optimizers, such PPO and/or SAC, can alleviate the instabilities in training, enhance convergence and exploration. Finally, to push forward the entire field of intrinsically motivated RL, we hope to demonstrate performance that was previously impossible: a purely intrinsically motivated (via empowerment) agent acting from images (pixels).

Schematically, our method is depicted at Figure 2. The controller policy $\pi_\theta(a \mid s)$ collects trajectories in the environment, (s, a, s'), where s is either a raw state or an image. The channel matrix $G_\chi(s_t)$ predicts the final state of a trajectory corresponding to the action sequence a_t^{T-1}, started from s_t. The singular values of $G_\chi(s_t)$ are found by singular value decomposition. The empowerment, $\mathcal{E}(s_t)$, is estimated by equation 3 and equation 4. The estimated values of empowerment provide the intrinsic reward, demoted by $r'(\tau)$, for the update of the controller policy, $\pi_\theta(a \mid s)$, by RL using Soft Actor Critic (SAC). Pseudo-code summarising all the details of our method is provided below in Algorithm 1.

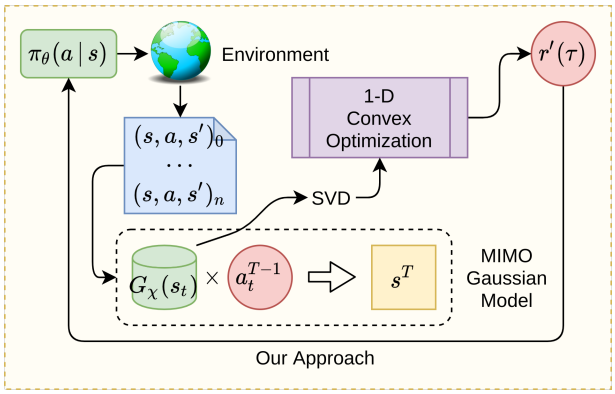


Figure 2: Our approach for empowerment maximization.

Algorithm 1 Estimation and Maximization of Empowerment by Gaussian Channel Capacity

1: **Input:** $\pi_{\theta_0}(a \mid s)$ - control policy, H - prediction horizon.
2: **for** $i = 0$ **to** N **do**
3: $\{\tau_k\}_{k=1}^K \leftarrow$ run policy $\pi_{\theta_i}(a \mid s)$ {data collection in environment}
4: extract tuples $(s_t,\ a_t^{H-1},\ s_{t+H})$ from $\{\tau_k\}_{k=1}^K$ {data arrangement for Gaussian channel}
5: **for** $j = 1$ **to** M **do**
6: $\chi^* \leftarrow \arg\min_\chi \sum_{tuples} \left\| G_\chi(s_t) \cdot a_t^{H-1} - s_{t+H} \right\|_2^2$ {projection to equation 2}
7: **end for**
8: $\{\sigma_i^*(s_t)\}_{i=1}^k \leftarrow$ SingularValueDecomposition$(G_{\chi^*}(s_t))$ {an alternative explained below}
9: $\mathcal{E}(s_t) \leftarrow \frac{1}{2} \sum_{i=1}^k \log(1 + \sigma_i^*(s) p_i^*)$ {equation 3 and equation 4}
10: reinforce $\pi_{\theta_i}(a \mid s)$ with $r'(\tau) = \sum_{t=1}^T \mathcal{E}(s_t)$. {policy update}
11: **end for**

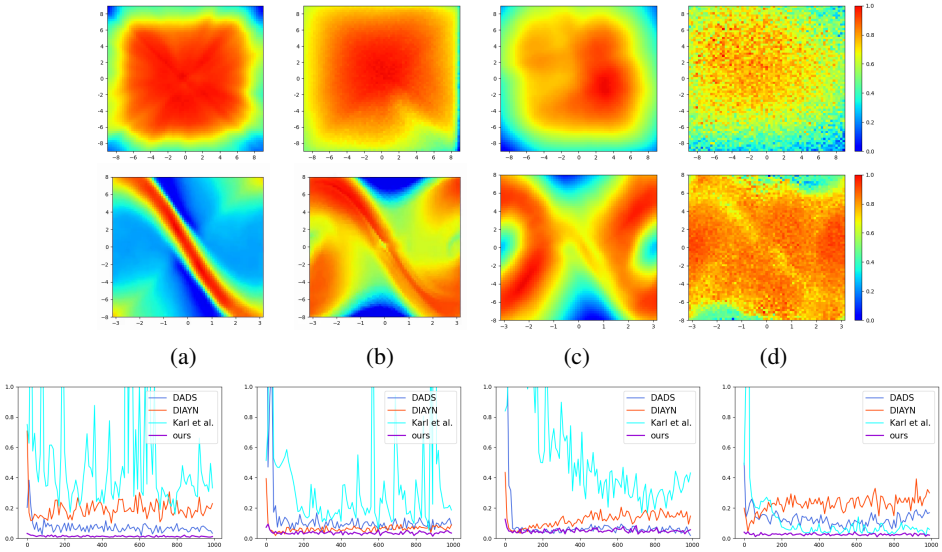
The channel matrix $G_\chi(s_t)$ has dimensions $d_s \times d_a \cdot (H-1)$, where d_s and d_a are the dimensions of the state space (or latent representation) and the action space. In line 6, there is a multiplication between the matrix, $G_\chi(s_t)$, and the action sequence, $a_t^{H-1} \doteq (a_t, a_{t+1}, \ldots, a_{H-1})$. The singular values, $\sigma_i(s_t)$, can be estimated directly without the singular values decomposition by representing the channel matrix as $G_\chi(s_t) = U_\chi(s_t) * Diagm(\Sigma_\chi(s_t)) * V_\chi(s_t)$ with two additional orthonormality constraints: $\left\| U_\chi(s_t) U_\chi^\dagger(s_t) - \mathbf{I}_{\{d_s \times d_s\}} \right\|_2^2$ and $\left\| V_\chi(s_t) V_\chi^\dagger(s_t) - \mathbf{I}_{\{d_a(H-1) \times d_a(H-1)\}} \right\|_2^2$. In our experiments, we did not experience a bottleneck in calculating the singular values decomposition

for both raw states and images. But, the alternative direct calculation can be applied when SVD is undesired. This explains the essence of our algorithm, which can work from images using simple algorithmic modifications (adding an encoder, a decoder, and a regularize for latent space) as described in Appendix B. Note that VLB methods (Figure 1) comprise of three neural networks, (θ, ϕ, ψ), while our method (Figure 2) only uses two (θ, χ), additionally using an efficient 1D line-search in equation 4. Hence our method is simpler, using less "moving parts", generally promising to be more stable. We confirm this with the experiments in the next section.

5 EXPERIMENTS

5.1 CONVERGENCE OF DIFFERENT ESTIMATORS

In this section we compare the convergence of different empowerment estimators to the true empowerment function. We observe that in simple environments such as "ball-in-box", shown in Figure 3, all estimators converge to the correct landscape with high empowerment values in the middle and low values near the boundaries, which correspond to previous results for this environment (Klyubin et al. (2005); Maximilian et al. (2019)). For the pendulum environment, it is clear that while both method (b) and ours succeed in creating results similar to the analytical solution in 6, ours outputs an empowerment landscape that is smoother and more discerning.



(e) Ball-in-box (f) Pendulum (g) Mountain-car (h) Cart-pole

Figure 3: Empowerment convergence by different methods (squashed linearly to [0, 1]). The top row plots the emp value of the 2D ball-in-box environment: x, y-axis correspond to the ball's position. The middle row is from the pendulum environment: x-axis: angle from to top in rad. y-axis: angular velocity in rad/s. (a) - ours, (b) - Maximilian et al. (2019), (c) - Eysenbach et al. (2019), (d) - Sharma et al. (2020). The bottom row plots the standard deviation in average empowerment value across 10 seeds: x-axis marks the number of training episode; each episode contains 1000 environment steps.

Figure 3 shows that the proposed algorithm is superior to the existing sample-based methods with regard to a lower variance in training and its convergence to empowerment.

5.2 TWO PHASES IN EMPOWERMENT-BASED STABILIZATION

Empowerment of the pendulum is highest when it is stationary at the up-right position. In this section, we show the evolution of the empowerment landscape, the corresponding control policy and the state distribution as we run Algorithm 1 without any external reward. In Figure 4, we observe that Algorithm 1 includes two phases: empowerment learning and policy learning. Empowerment converges from epoch 0 to 10 and then, policy converges from epoch 10 to 50.

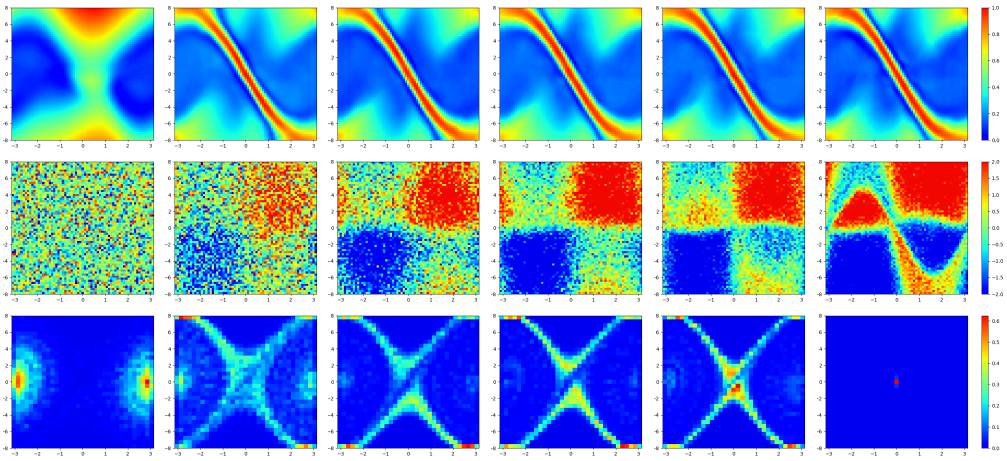


Figure 4: Snapshots of the empowerment landscape (top), the control policy (middle) and the state concentration (bottom). The x-axis is θ (angle) and the y-axis is $\dot{\theta}$ (angular velocity). The angle is measured from the upright position, ($\theta = 0\,\mathrm{rad}$ corresponds to the upright position). Each epoch corresponds to 10^4 steps. The policy corresponds to the optimal policy for the stabilization task with external reward (Doya (2000)).

5.3 Unsupervised stabilization by empowerment from images

In this section we demonstrate for the first time unsupervised stabilization by empowerment from images. Simply put, no prior work succeeded to estimate the empowerment function reliably from images for swinging up the pendulum and stabilizing it at the upright position, which has been demonstrated before only by the analytical empowerment calculation Salge et al. (2013). Figure 5 shows the empowerment landscape computed from images by our approach, and the optimal trajectory for swinging up the pendulum from the bottom position to the upright position. The x-axis is θ (angle) and the y-axis is $\dot{\theta}$ (angular velocity). The black curve is the optimal trajectory for swinging up the pendulum, which starts from the bottom with $\theta = \pi\,\mathrm{rad}$ and $\dot{\theta} = 0\,\mathrm{rad}$. At the beginning the pendulum swings with a small amplitude (the dense black dots around $\theta = \pi\,\mathrm{rad}$, $\dot{\theta} = 0\,\mathrm{rad\,s^{-1}}$), then, once accumulated enough energy, it starts traversing the state space, and arrives at the top position, where it stabilises, (dense black points around $\theta = 0\,\mathrm{rad}$, $\dot{\theta} = 0\,\mathrm{rad\,s^{-1}}$).

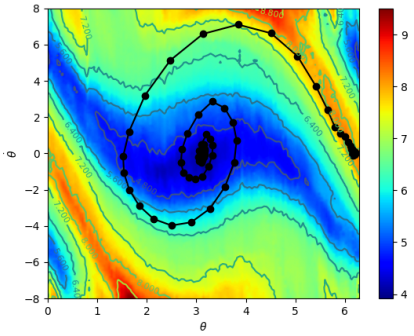


Figure 5: Empowerment landscape and the optimal trajectory, calculated from images of pendulum.

6 Summary and Future Work

In this work we introduce a new sample-based method for empowerment estimation in dynamical systems. The key component of our method is the representation of dynamical system by the Gaussian channel, which allows to estimate empowerment by convex optimization. Our estimator converges to true empowerment landscape both from raw states and images. Empowerment is used in various applications. Our improved estimator has a potential to improve these applications. Specifically, the natural continuation of this work is to learn skills using our representation of empowerment by Gaussian channel, rather than by VLB.

REFERENCES

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms, 2018.

Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

Andrew G Barto. Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems*, pp. 17–47. Springer, 2013.

Andrew G Barto, Satinder Singh, and Nuttapong Chentanez. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*, pp. 112–19, 2004.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giró-i-Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. *International Conference on Machine Learning (ICML)*, 2020.

Nuttapong Chentanez, Andrew G. Barto, and Satinder P. Singh. Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, and L. Bottou (eds.), *Advances in Neural Information Processing Systems 17*, pp. 1281–1288. MIT Press, 2005. URL `http://papers.nips.cc/paper/2552-intrinsically-motivated-reinforcement-learning.pdf`.

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

Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. `https://github.com/openai/baselines`, 2017.

Kenji Doya. Reinforcement learning in continuous time and space. *Neural computation*, 12(1): 219–245, 2000.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=SJx63jRqFm`.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Skc-Fo4Yg`.

Christian Guckelsberger, Christoph Salge, and Simon Colton. Intrinsically motivated general companion npcs via coupled empowerment maximisation. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pp. 1–8. IEEE, 2016.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration, 2017.

Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for continuous agent—environment systems. *Adaptive Behavior*, 19(1):16–39, 2011.

A. S. Klyubin, D. Polani, and C. L. Nehaniv. Empowerment: a universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135 Vol.1, Sep. 2005. doi: 10.1109/CEC.2005.1554676.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. All else being equal be empowered. In *European Conference on Artificial Life*, pp. 744–753. Springer, 2005.

Karl Maximilian, Becker-Ehmck Philip, Soelch Maximilian, Benbouzid Djalel, van-der-Smagt Patrick, and Bayer Justin. Unsupervised real-time control through variational empowerment. *International Symposium on Robotics Research*, 2019.

David McAllester and Karl Statos. Formal limitations on the measurement of mutual information. *arXiv preprint arXiv:1811.04251*, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pp. 2125–2133, 2015.

P-Y Oudeyer, Jacqueline Gottlieb, and Manuel Lopes. Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. In *Progress in brain research*, volume 229, pp. 257–284. Elsevier, 2016.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction, 2017.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information, 2019.

C. Salge and D. Polani. Empowerment as replacement for the three laws of robotics. *Frontiers Robotics AI*, 4:25, 2017.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Approximation of empowerment in the continuous domain. *Advances in Complex Systems*, 16(02n03):1250079, 2013.

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

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HJgLZR4KvH`.

Steven H Strogatz. *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 2018.

Stas Tiomkin and Naftali Tishby. A unified bellman equation for causal information and value in markov decision processes. *arXiv preprint arXiv:1703.01585*, 2017.

Stas Tiomkin, Daniel Polani, and Naftali Tishby. Control capacity of partially observable dynamic systems in continuous time. *arXiv preprint arXiv:1701.04984*, 2017.

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

Emanuel Todorov. Optimal control theory. *Bayesian brain: probabilistic approaches to neural coding*, pp. 269–298, 2006.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rkxoh24FPH`.

Alexander D Wissner-Gross and Cameron E Freer. Causal entropic forces. *Physical review letters*, 110(16):168702, 2013.

Du Yuqing, Tiomkin Stas, Kiciman Emre, Polani Daniel, Abbeel Pieter, and Dragan Anca. Ave: Assistance via empowerment. In *Conference on Neural Information Processing Systems, NeurIPS*. Thirty-fourth annual conference on Neural Information Processing Systems, 2021.

Harari Yuval, Noah. *Sapiens: A Brief History of Humankind*. Harvill Secker, 2014.

A ANALYTICAL DERIVATION OF $G(s)$ FOR INVERTED PENDULUM

The main steps in the derivation of $G(s)$ appear below. The full derivation appears in Section 4.1 at Salge et al. (2013). The current state of the pendulum, s_t, is given by θ_t and $\dot{\theta}_t$:

$$s_t = \begin{bmatrix} \theta_t \\ \dot{\theta}_t \end{bmatrix}, \tag{7}$$

where θ is measured from the upright position. We consider the development of the pendulum for three time steps of duration dt. The behaviour of the pendulum can be approximate linearly at every time step for small enough dt. Then, the next state of the pendulum is given by:

$$s_{t+dt} = s_t + A_t dt + B_t a_t, \tag{8}$$

where a_t is the agent's action, $A_t = \begin{bmatrix} \dot{\theta}_t \\ \frac{g}{l}\sin(\theta_t) \end{bmatrix}$, and $B_t = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. For the next time step we need to calculate a new matrix, A, because the non-linearity of the system causes A_{t+dt} to depend on the new state. Consequently, $A_{t+dt} = \begin{bmatrix} \dot{\theta}_{t+dt} \\ \frac{g}{l}\sin(\theta_{t+dt}) \end{bmatrix}$. A similar matrix can be calculated for A_{t+2dt}. An propagation of pendulum dynamics for 3 time steps allows us to calculate s_{t+3dt} explicitly, which appears as:

$$\begin{aligned} \theta_{t+3dt} = {} & dt(\frac{gdt}{l}\sin(dt\dot{\theta}_t + \theta_t) + \frac{gdt}{l}\sin(\theta_t) + \dot{\theta}_t) \\ & + dt(\frac{gdt}{l}\sin(\theta_t) + 2\dot{\theta}_t) + \theta_t + 3dta_t + 3dta_{t+dt}, \end{aligned}$$

and,

$$\begin{aligned} \dot{\theta}_{t+3dt} = {} & \frac{gdt}{l}\sin(dt\dot{\theta}_t + \theta_t) + \frac{gdt}{l}\sin(\theta_t) + \dot{\theta}_t \\ & + \frac{gdt}{l}\sin(dt(\frac{gdt}{l}\sin(\theta_t) + 2\dot{\theta}_t) + \theta_t + 2dta_t) + dta_t + dta_{t+dt} + dta_{t+2dt}, \end{aligned}$$

where a_{t+dt} is the action at the next step. Our goal is to represent s_{t+3dt} by:

$$s_{t+3dt} = K + G(s_t) \cdot \begin{bmatrix} a_t \\ a_{t+dt} \\ a_{t+2dt} \end{bmatrix}, \tag{9}$$

which is achieved by linearisation θ_{t+3dt} and $\dot{\theta}_{t+3dt}$ around $a_t = 0$. A straightforward calculation shows that the matrix $G(s_t)$ is given by:

$$G(s_t) = \begin{bmatrix} 2dt & dt & 0 \\ \frac{dt^2 g}{l}\cos(dt(\frac{gdt}{l}\sin(\theta_t) + 2\dot{\theta}_t) + \theta_t) + 1 & 1 & 1 \end{bmatrix}. \tag{10}$$

This calculation can be continued beyond 3-steps, involving more complicated terms in the expansions. Given $G(s_t)$ in equation 10 one can represent s_{t+3dt} by equation 9, where the matrix K does not affect the mutual information Cover & Thomas (2012). Adding to equation 9 the normal noise, η, as suggested in Salge et al. (2013), we get the representation in equation 2. In contrast, we propose to learn $G(s)$ from samples via interaction with the environment for an arbitrary dynamics and an arbitrary number of steps (c.f. 3 steps in the above derivation). Using $G(s)$ from equation 10, we can calculate the analytical (ground truth empowerment landscape), which is shown at Figure 6

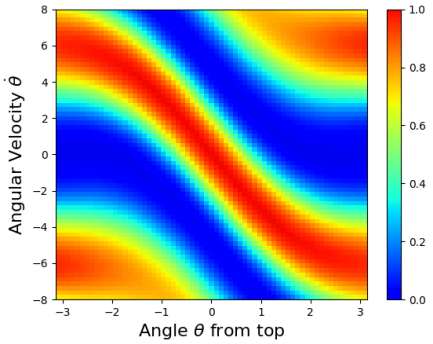


Figure 6: Analytical Empowerment landscape . The x-axis is θ (angle) and the y-axis is $\dot{\theta}$ (angular velocity). The black curve is the optimal trajectory for swinging up the pendulum, which starts from the bottom with $\theta = \pi\,\mathrm{rad}$ and $\dot{\theta} = 0\,\mathrm{rad}$.

B GAUSSIAN CHANNEL IN LATENT SPACE

To learn the empowerment function from pixels, we construct the Gaussian linear channel equation 2 in the latent space using an encoder. To verify the correctness of our latent dynamics model, we compare reconstruction of latent predictions with ground truth observations. In inverted pendulum, we use two consecutive images to capture the state information (angle and angular velocity). As shown in Figure 7, our model is capable of reconstructing future images from current images and action sequences.

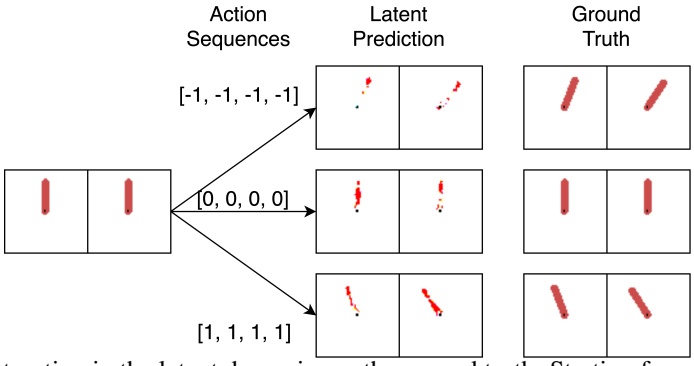


Figure 7: Reconstruction in the latent dynamics vs the ground truth. Starting from the same upright position, a different action sequence is taken in each row (action +1: counter-clockwise torque, 0: no torque, -1: clockwise torque). Reconstruction matches the actual observation in all cases.

C CART-POLE SWING-UP

The cart-pole environment is made up of a stiff pole with one end fixed to a pivot on a cart movable along one axis. The cart can be pushed left or right within a confined region. The state space of cart-pole has 4 components: cart position, cart velocity, pole angle, pole angular velocity. We expect

high empowerment when the pole is up right, because starting there, gravity is able to help the system reach a broader set of future states. To measure whether Algorithm 1 successfully brings the cart-pole to up-right position, we plot the average squared angle to the top against the number of training steps. We also include 2 different reward functions that directly aim this objective for comparison.

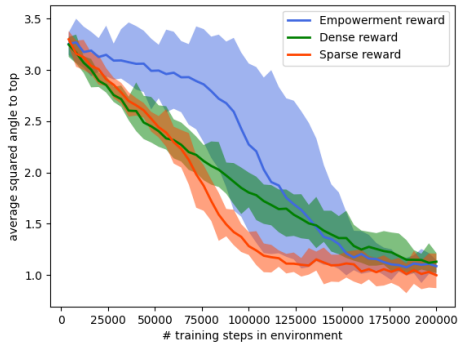


Figure 8: Cart-Pole - Comparison of unsupervised and supervised reward functions. The "dense reward" is -θ^2 and the “sparse reward” is $\mathbb{1}\{|\theta| < \frac{\pi}{10}\}$.

As shown in Figure 8, the empowerment reward, without the need of prior knowledge of the environment, is able to achieve similar final performance compared to direct reward signals. Moreover, it learns this objective at a speed on par with the dense reward.

D EMPOWERMENT ESTIMATION FOR SAFETY

A useful application of empowerment is its implication of safety for the artificial agent. A state is intrinsically safe for an agent when the agent has a high diversity of future states, achievable by its actions. This is because in such states, the agent can take effective actions to prevent undesirable futures. In this context, the higher its empowerment value, the safer the agent is. In this experiment, we first check that our calculation of empowerment matches the specific design of the environment. Additionally, we show that empowerment augmented reward function can affect the agent's preference between a shorter but more dangerous path and a longer but safer one.

Environment: a double tunnel environment implemented with to the OpenAI Gym API (Brockman et al., 2016). Agent (blue) is modeled as a ball of radius 1 inside a 20×20 box. The box is separated by a thick wall (gray) into top and bottom section. Two tunnels connect the top and bottom of the box. The tunnel in middle is narrower but closer to the goal compared to the one on the right.

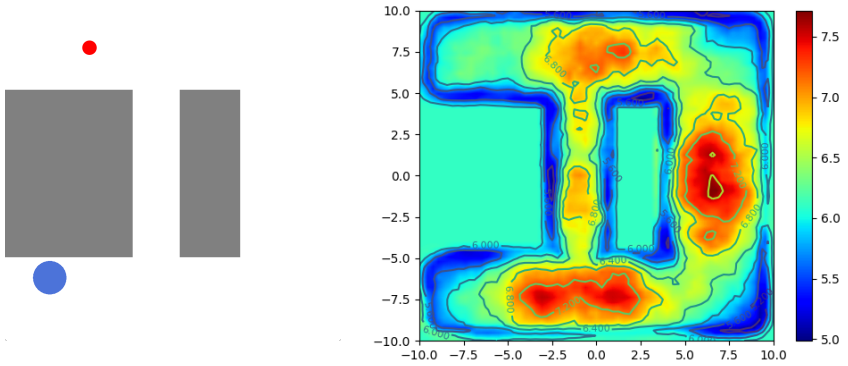


Figure 9: **Left sub-figure**: Double tunnel environment. The goal is marked in red. The agent in blue. **Right sub-figure**: Empowerment landscape for the tunnel environment. The values of empowerment reduce at the corner and inside the tunnel where the control of the agent is less effective compared to more open locations.

Control: In each time step, the agent can move at most 0.5 unit length in each of x and y direction. If an action takes the agent into the walls, it will be shifted back out to the nearest surface.

Reset criterion: each episode has a maximum length of 200 steps. The environment resets when time runs out or when the agent reaches the goal.

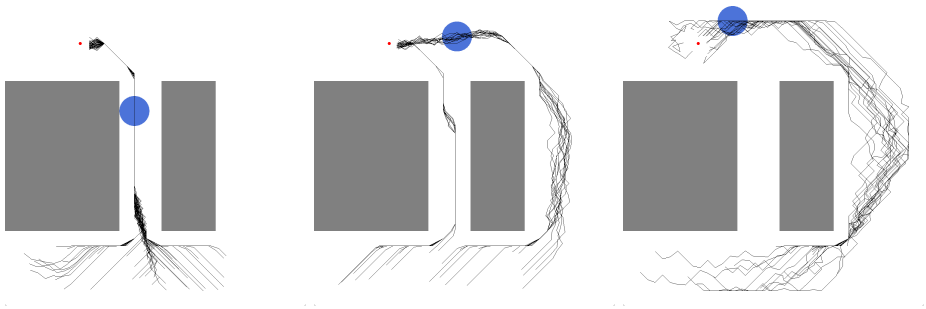
Since the tunnel in the middle is narrower, the agent is relatively less safe there. The effectiveness of the control of the agent is damped in 2 ways:

1. In a particular time step, it's more likely for the agent to bump into the walls. When this happens, the agent is unable to proceed as far as desired.
2. The tunnel is 10 units in length. When the agent is around the middle, it will still be inside the tunnel in the 5-step horizon. Thus, the agent has fewer possible future states.

We trained the agent with PPO algorithm (Schulman et al., 2017) from OpenAI baselines (Dhariwal et al., 2017) using an empowerment augmented reward function. After a parameter search, we used discount factor $\gamma = 0.95$ over a total of 10^6 steps. The reward function that we choose is:

$$R(s, a) = \mathbf{1}_{goal} + \beta \times Emp(s)$$

where β balances the relative weight between the goal conditioned reward and the intrinsic safety reward. With high β, we expect the agent to learn a more conservative policy.



(a) $\beta = 0$ (b) $\beta = 1/1600$ (c) $\beta = 1/800$

Figure 10: Trajectories of trained policy. As β increase, the agent develops a stronger preference over a safer route, sacrificing hitting time.

The results from the tunnel environment again support our proposed scheme. First, the empowerment landscape matches our design of the environment. Second, high quality empowerment reward successfully alters the behavior of the agent.

E Experiment Details

E.1 Dynamics-Aware Discovery of Skills (DADS)

For our experiments using DADS, we use policy and discriminator architecture of two hidden layers of size 256 with ReLU activations and a latent dimension of 2. Our discount factor is .99. As per suggestions from the original paper, we use 512 prior samples to approximate the intrinsic reward, run 32 discriminator updates and 128 policy updates per 1000 timesteps; learning rates of 3×10^{-4} and batch size of 256 are used throughout. We use a replay buffer to help stabilize training of size 20000, used only for policy updates.