Learning Large Skillsets in Stochastic Settings with Empowerment

Andrew Levy
andrew_levy2@brown.edu
Brown University

Alessandro Allievi
alessandro.allievi@us.bosch.com
Robert Bosch LLC

George Konidaris
gdk@brown.edu
Brown University

Abstract

General purpose agents need to be able to execute diverse skillsets in stochastic settings. Given that the mutual information between skills and states measures the number of distinct skills in a skillset, a compelling objective for learning a large skillset is to find the skillset with the largest mutual information between skills and states. The problem is that the two main unsupervised approaches for optimizing this mutual information objective, Empowerment-based skill learning and Unsupervised Goal-Conditioned Reinforcement Learning, only optimize loose lower bounds on the mutual information, which can impede diverse skillset learning. We propose a new empowerment objective, *Skillset Empowerment*, that optimizes a tighter bound on the mutual information between skills and states. For any proposed skillset, the tighter bound on mutual information is formed by replacing the posterior distribution of the proposed skillset with a variational distribution trained to match the posterior of the proposed skillset. In addition, because our empowerment objective is not a reinforcement learning problem, we optimize the objective as a bandit problem in which actions are skillsets (e.g., the set of parameters that make up the skill-conditioned policy) and the reward for an action is our variational mutual information lower bound for the proposed skillset. We show empirically that our approach is able to learn large abstract skillsets in stochastic domains, including ones with high-dimensional observations, in contrast to existing approaches.

1 Introduction

General purpose agents that operate in the real world will need to be able to execute a large set of skills in highly stochastic settings. A futuristic household robot, for instance, will need to execute the vast number of skills involved in household chores like cooking and cleaning while the human members of the household may be moving as well as conversing with each other and the robot in seemingly random ways. Even the simple act of the robot moving its head to look in different directions will produce unpredictable outcomes as relevant objects to the robot may appear in unexpected places. An appealing approach for learning diverse skillsets, regardless of the level of randomness in the domain, is to find the skillset with the largest mutual information between skills and skill-terminating states because this mutual information measures the number of distinct skills in a skillset.

The problem is that the two most popular approaches for optimizing the mutual information between skills and states, Empowerment-based skill learning (Gregor et al., 2016; Eysenbach et al., 2018; Achiam et al., 2018; Choi et al., 2021; Sharma et al., 2019) and Unsupervised Goal-Conditioned Reinforcement Learning (GCRL) (Ecoffet et al., 2019; Mendonca et al., 2021; Nair et al., 2018; Pong et al., 2019; Campos et al., 2020; Pitis et al., 2020; Held et al., 2017; McClinton et al., 2021; Held et al., 2017; Kim et al., 2023), only maximize a loose lower bound of the mutual information between skills and states, which can make it challenging to learn a diverse skillset. In both existing empowerment and unsupervised GCRL approaches, this loose lower bound on mutual information

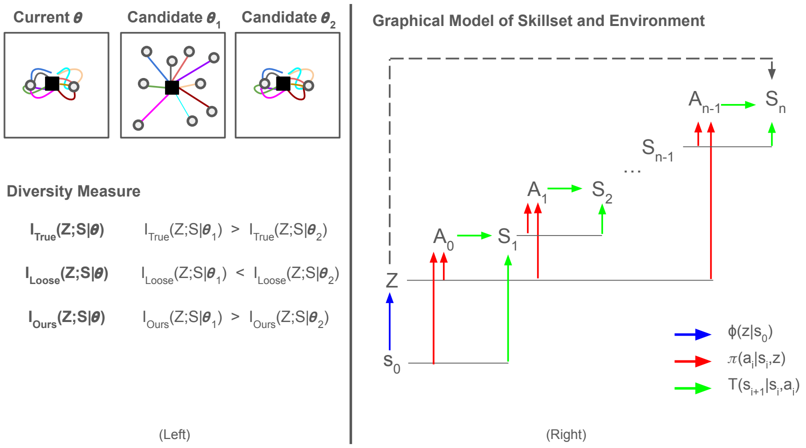

Current $\boldsymbol{\theta}$
Candidate $\boldsymbol{\theta}_1$
Candidate $\boldsymbol{\theta}_2$
Diversity Measure
$I_{True}(Z;S|\theta)$
$I_{True}(Z;S|\theta_1) > I_{True}(Z;S|\theta_2)$
$I_{Loose}(Z;S|\theta)$
$I_{Loose}(Z;S|\theta_1) < I_{Loose}(Z;S|\theta_2)$
$I_{Ours}(Z;S|\theta)$
$I_{Ours}(Z;S|\theta_1) > I_{Ours}(Z;S|\theta_2)$
(Left)
Graphical Model of Skillset and Environment
A_{n-1}
S_n
S_{n-1}
...
A_1
S_2
A_0
S_1
Z
s_0
$\phi(z|s_0)$
$\pi(a_i|s_i,z)$
$T(s_{i+1}|s_i,a_i)$
(Right)

Figure 1: (Left) Two candidate skillsets θ_1 and θ_2 are compared using different measures of diversity. θ_1 is the more diverse skillset as each of its nine skills target different states, whereas the skills in θ_2 target two different states (different skills are shown by different colored trajectories emanating from the black square agent). Thus, using the true mutual information, $I_{True}(Z;S|\theta_1) > I_{True}(Z;S|\theta_2)$. However, using the loose lower bound on mutual information employed by existing empowerment methods that penalizes differences from the current skillset, $I_{Loose}(Z;S|\theta_1) < I_{Loose}(Z;S|\theta_2)$ because θ_2 is more similar to the current skillset θ than θ_1. We introduce a tighter bound to the mutual information in which the more diverse skillset θ_1 would score higher than θ_2. (Right) Graphical model of a skillset and its effect on environment. Z, A_i, and S_i are random variables for skills, actions, and states, respectively. Solid arrows represent conditional probability distributions. A skillset is defined by the distribution over skills (blue arrow) and skill-conditioned policy distribution (red arrows). Green arrows represent the transition dynamics. Dotted arrow is the skill channel.

for some candidate skillset is formed by replacing the true posterior probability that computes the probability of a skill given the skill-terminating state and the candidate skillset with a potentially very different distribution. Existing empowerment approaches replace the true posterior of the candidate skillset with another distribution trained to match the true posterior of the *current* skillset, which may have significant differences with the candidate skillset. As Figure 1 (Left) illustrates, this can create a loose lower bound on mutual information for desirable diverse skillsets that differ significantly from the current skillset. Similarly, GCRL replaces the true posterior with a fixed posterior distribution that encourages the goal-conditioned policy to execute actions that achieve the assigned goal state (Choi et al., 2021). However, this implementation can produce a loose lower bound on mutual information in stochastic settings for diverse skillsets with abstract skills that target groupings of states.

Our main contribution, Skillset Empowerment, is an empowerment objective that optimizes a tighter lower bound on the mutual information between skills and states. In our variational lower bound, we replace the true posterior distribution within the mutual information term for a candidate skillset with a variational posterior that is trained to match the true posterior of the candidate skillset. The resulting empowerment objective is not a reinforcement learning problem because the reward depends on the full skill-conditioned policy. As a result, we optimize our objective as a bandit problem, in which the actions are candidate skillsets (e.g., the parameters of the skill-conditioned policy neural network) and the reward is our version of the skillset candidate's mutual information variational lower bound, which measures the diversity of the proposed skillset. We provide a particular architecture and optimization method that can efficiently optimize the objective despite the potentially large action space. Our experiments show that our approach can learn diverse abstract skillsets in stochastic settings, including ones with high-dimensional observations. To our knowledge, our approach is the first unsupervised skill learning method to successfully learn large skillsets in stochastic settings.

2 Background

2.1 The Skill Channel and Empowerment

We model an agent's skillset and its effect on the environment with the probabilistic graphical model shown in Figure 1 (Right). The random variables in the model include Z, which represents the selected skill; A_i, which is the primitive action executed by the agent at step i; and S_i, which represents the state at step i that is assumed to be visible to the agent. s_0 is the skill start state under consideration. The model's probability distributions that produce the random variables include the environment's Markovian transition dynamics, $T(s_{i+1}|s_i, a_i)$ that are independent of past states and actions, and the pair of distributions that define a skillset: $\phi(z|s_0)$ and $\pi(a_i|s_i, z)$. $\phi(z|s_0)$ represents the distribution over skills that an agent can execute in some skill start state s_0. $\pi(a_i|s_i, z)$ represents the agent's distribution over actions given some skill start state s_0 and skill z (i.e., the agent's skill-conditioned policy). The model assumes each skill consists of n primitive actions. Given a start state s_0 and specific skillset distributions $\phi(z|s_0)$ and $\pi(a_i|s_i, z)$, the joint distribution of the random variables $p(z, a_0, s_1, a_1, s_2, \ldots, a_{n-1}, s_n|s_0, \phi, \pi) = \phi(z|s_0)\pi(a_0|s_0, z)T(s_1|s_0, a_0)\pi(a_1|s_1, z)T(s_2|s_1, a_1)\ldots\pi(a_{n-1}|s_{n-1}, z)T(s_n|s_{n-1}, a_{n-1})$.

Given our goal of learning diverse skillsets that target many states, the relationship between the skill random variable Z and the skill-terminating state random variable S_n given some start state s_0 and skillset defined by the distribution over skills $\phi(z|s_0)$ and the skill-conditioned policy $\pi(a|s, z)$ is of particular interest. We refer to the relationship between skills and skill-terminating states as the skill channel. The skill channel is interesting because sampling the channel can indicate how diverse a skillset is. More diverse skillsets will be those in which different skills z produce different states s_n, while less diverse skillsets will be those in which different skills z produce the same distribution over s_n. Shannon (1948) showed that the diversity of a channel (i.e., the number of distinct inputs to a channel that can be decoded with arbitrarily low error at the output of the channel) can be measured using the mutual information of the channel. Thus, in the case of skillsets, the number of unique skills in a skillset defined by $(\phi(z|s_0), \pi(a|s, z))$ can be quantified using the mutual information between skills and states $I(Z; S_n|s_0, \phi, \pi)$, in which

$$I(Z; S_n|s_0, \phi, \pi) = H(Z|s_0, \phi, \pi) - H(Z|s_0, \phi, \pi) \tag{1}$$

$$= \mathbb{E}_{z\sim\phi(z|s_0), s_n\sim p(s_n|s_0,\pi,z)}[\log p(z|s_0, \phi, \pi, s_n) - \log \phi(z|s_0)]. \tag{2}$$

The goal of our work is to find the most diverse skillset that produces the largest mutual information of the skill channel. The maximum mutual information of a channel is known as the capacity of the channel. For channels that describe the relationship between action and sensor representations, which is the case for the skill channel where the skills are actions and the states are sensor representations, the capacity of the channel can also be referred to as Empowerment (Klyubin et al., 2005; Capdepuy, 2011; Salge et al., 2013a). Thus, the goal of our work is to find the empowerment of states s_0, which requires finding the skillset $(\phi(z|s_0), \pi(a|s, z))$ that produces the largest mutual information between skills and states:

$$\mathcal{E}(s_0) = \max_{\phi(z|s_0), \pi(a|s,z)} I(Z; S_n|s_0, \phi, \pi). \tag{3}$$

2.2 Existing Approaches

Next we discuss how both of the two dominant approaches to unsupervised skill learning, empowerment-based skill learning and unsupervised GCRL, optimize the mutual information of the skill channel and why these approaches are only optimizing loose lower bounds on the mutual information between skills and states.

A key problem with optimizing the mutual information of the skill channel for a candidate skillset (ϕ, π) is that it depends on the posterior distribution $p(z|s_0, \phi, \pi, s_n)$, which provides the probability of a skill z given the start state s_0, skillset (ϕ, π), and skill-terminating state s_n. As shown in line

4, the posterior is intractable to compute in large settings (e.g., domains with continuous state and action spaces) because it requires integrating over an infeasible number of trajectories.

$$p(z|s_0, \phi, \pi, s_n) = \frac{\int_{a_0, s_1, a_2, \ldots, s_{n-1}, a_{n-1}} p(z, a_0, s_1, \ldots, a_{n-1}, s_n|s_0, \phi, \pi)}{\int_{z, a_0, s_1, a_2, \ldots, s_{n-1}, a_{n-1}} p(z, a_0, s_1, \ldots, a_{n-1}, s_n|s_0, \phi, \pi)} \tag{4}$$

To overcome this problem, both existing empowerment and unsupervised GCRL methods replace the original posterior with a different variational distribution $q_\psi(z|s_0, s_n)$, similar to Mohamed & Rezende (2015) who first replaced the intractable posterior with a variational distribution when optimizing a different mutual information between open loop action sequences and terminating states. Replacing the true posterior $p(z|s_0, \phi, \pi, s_n)$ with the variational distribution $q_\psi(z|s_0, s_n)$ results in a lower bound on mutual information $I^V(Z; S_n|s_0, \phi, \pi)$ for the candidate skillset (ϕ, π). The gap between the actual mutual information $I(Z; S_n|s_0, \phi, \pi)$ and the variational lower bound on mutual information $I^V(Z; S_n|s_0, \phi, \pi)$ is an average of the KL divergences between the true posterior and the variational distribution (Barber & Agakov, 2003):

$$I(Z; S_n|s_0, \phi, \pi) - I^V(Z; S_n|s_0, \phi, \pi) = \mathbb{E}_{s_n \sim p(s_n|s_0, \phi, \pi)}[D_{KL}(p(z|s_0, \phi, \pi, s_n)||q_\psi(z|s_0, s_n))]. \tag{5}$$

Thus, replacing the true posterior with a similar variational distribution can produce a tight bound on mutual information, but replacing the true posterior with a markedly different one can produce a loose bound. Next, we discuss the variational posteriors used by existing empowerment approaches and unsupervised GCRL.

Existing Empowerment The typical empowerment-based skill-learning algorithm replaces the true posterior with a variational distribution trained to be similar to the posterior of the current *greedy* skillset $(\phi_{\text{greedy}}, \pi_{\text{greedy}})$ (Gregor et al., 2016; Eysenbach et al., 2018; Achiam et al., 2018; Lee et al., 2019; Choi et al., 2021; Strouse et al., 2021). Specifically, the parameters of the variational posterior ψ are trained to minimize the KL divergence between the posterior of the current skillset and the variational distribution:

$$\psi^* = \underset{\psi}{\text{argmin}}\, \mathbb{E}_{s_n \sim p(s_n|s_0, \phi_{\text{greedy}}, \pi_{\text{greedy}})}[D_{KL}(p(z|s_0, \phi_{\text{greedy}}, \pi_{\text{greedy}}, s_n)||q_\psi(z|s_0, s_n))]. \tag{6}$$

With this posterior, the empowerment objective becomes

$$\begin{aligned} \mathcal{E}^V(s_0) &= \max_{\phi(z|s_0), \pi(a|s,z)} I^V(Z; S_n|s_0, \phi, \pi), \\ I^V(Z; S_n|s_0, \phi, \pi) &= \mathbb{E}_{z \sim \phi(z|s_0), s_n \sim p(s_n|s_0, \pi, z)}[\log q_{\psi^*}(z|s_0, s_n) - \log \phi(z|s_0)]. \end{aligned} \tag{7}$$

Using this objective, candidate skillsets (ϕ, π) are evaluated using the variational lower bound on mutual information defined in line 7.

The problem with this implementation is that it will significantly penalize skillsets that differ from the current skillset, even when they are significantly more diverse than the current skillset. Examples similar to the one discussed in Figure 1 (Left) in which a candidate skillset can target more unique states than the current greedy skillset but because the skills z that achieve the terminating states s_n do not have high probability according to variational posterior $q_{\psi^*}(z|s_0, s_n)$ at the states s_n, the $\log q_{\psi^*}(z|s_0, s_n)$ can be very low, which can then result in a poor I^V for the candidate skillset (ϕ, π). The low I^V scores would then discourage the agent from selecting these skillsets. Instead, the skillsets that are favored are the ones similar to the greedy skillset because the (z, s_n) tuples generated by skillsets similar to the greedy skillset will be rewarded with higher $\log q_{\psi^*}(z|s_0, s_n)$ values. Several prior works have empirically demonstrated this result in which existing empowerment approaches tend to learn skillsets that do not change much from initialization (Campos et al., 2020; Park et al., 2022; 2023a;b; Strouse et al., 2021; Levy et al., 2023).

Unsupervised GCRL Unsupervised GCRL approaches replace the true posterior with a fixed variational posterior that encourages the skill-conditioned policy (or the goal-conditioned policy) to target the conditioned goal state. There are different options for the fixed variational posterior depending on the desired goal-conditioned reward. The variational posterior could take the form of a fixed standard deviation gaussian centered at the skill-terminating state s_n: $q(z|s_0, s_n) = \mathcal{N}(z; \mu = s_n, \sigma = \sigma_0)$ (Choi et al., 2021). This will encourage the learning of a goal-conditioned policy such that when given a goal state z, the goal-conditioned policy will target a skill-terminating state s_n close to z. Section D in the Appendix describes how the common goal threshold reward function in which the agent is rewarded for entering within a threshold of the goal can be implemented with a fixed posterior.

For the desirable diverse skillsets, the tightness of a variational lower bound on mutual information I^V that employs a fixed posterior can depend on the level of randomness in the domain. In deterministic settings, the variational lower bound $I^V(Z; S_n|s_0, \phi, \pi)$ for diverse skillsets can form a tight bound to the true mutual information $I(Z; S_n|s_0, \phi, \pi)$. Consider the I^V of an effective goal-conditioned policy π, in which I^V employs a tight fixed-variance gaussian variational posterior with a mean at the skill-terminating state s_n. Because the goal-conditioned policy is effective (i.e., for goal state z, $s_n \approx z$), the true posterior $p(z|s_0, \phi, \theta, s_n)$ will be similar to the fixed-variance gaussian variational posterior, and thus the gap between the true mutual information and I^V will be small. In deterministic settings, GCRL can thus be an effective way to learn diverse skillsets, which has been repeatedly demonstrated empirically (Andrychowicz et al., 2017; Mendonca et al., 2021; Nair et al., 2018; Pong et al., 2019; Campos et al., 2020; Pitis et al., 2020; Held et al., 2017; McClinton et al., 2021; Held et al., 2017; Kim et al., 2023).

However, in significantly stochastic settings, the lower bound can be quite loose for the most diverse skillsets. In significantly random domains, the most diverse skillsets may be ones that contain abstract skills that target distinct groupings of states. Abstract skillsets like these will have posteriors in which many skill-terminating states map to the same skill, which can be far different than the fixed posterior used in GCRL in which each terminating state is mapped to its own unique goal state. Thus, the GCRL lower bound on mutual information can be quite loose for these desirable abstract skillsets, which will discourage the agent from learning them. Instead, the GCRL objective may limit skillsets to the more deterministic parts of the environment because trying to achieve any goal state that cannot be reliably achieved may be heavily penalized.

3 Skillset Empowerment

To enable agents to learn diverse skillsets in stochastic domains, we introduce the empowerment objective, Skillset Empowerment. In this section, we first present the objective and explain how the objective maximizes a tighter bound on the mutual information of the skill channel. Then we provide a practical implementation of the objective.

3.1 Skillset Empowerment Objective

The Skillset Empowerment objective is defined as follows:

$$\begin{aligned}
&\mathcal{E}^{SE}(s_0) = \max_{\phi(z|s_0), \pi(a|s,z)} I^{SE}(Z; S_n|s_0, \phi, \pi), \\
&I^{SE}(Z; S_n|s_0, \phi, \pi) = \mathbb{E}_{z \sim \phi(z|s_0), s_n \sim p(s_n|s_0, \pi, z)}[\log q_{\psi^*}(z|s_0, \phi, \pi, s_n) - \log \phi(z|s_0)], \\
&\psi^* = \operatorname*{argmin}_{\psi} \mathbb{E}_{s_n \sim p(s_n|s_0, \phi, \pi)}[D_{KL}(p(z|s_0, \phi, \pi, s_n)||q_\psi(z|s_0, \phi, \pi, s_n))] \qquad (8)
\end{aligned}$$

Like existing approaches, Skillset Empowerment optimizes a variational lower bound on the mutual information of the skill channel in which the true posterior of a candidate skillset (ϕ, π) is replaced with a variational distribution. The key reason Skillset Empowerment optimizes a tighter bound on mutual information is the manner in which ψ^* is selected. For any candidate skillset (ϕ, π), Skillset Empowerment trains the variational posterior $q_\psi(z|s_0, \phi, \pi, s_n)$ to match the true posterior of the candidate skillset $p(z|s_0, \phi, \pi, s_n)$. Note that the variational posterior is now conditioned on the skillset distributions (ϕ, π). The next section will discuss how this is implemented in practice. This strategy for learning ψ^* results in a variational mutual information I^{SE} that is a tighter lower bound for any skillset (ϕ, π) because the KL divergence between the true and variational posterior cannot be larger than the KL divergence in existing approaches. This is true because in the selection of ψ^*, Skillset Empowerment can choose the ψ selected by existing empowerment approaches or a ψ that produced a fixed variational distribution like in unsupervised GCRL if that is the ψ that minimizes the KL divergence, but Skillset Empowerment is not limited to those options. The tighter mutual information lower bound for any skillset in turn means that, for Skillset Empowerment, the skillset with the largest mutual information will have a mutual information level at least as large as existing approaches. That is, Skillset Empowerment will learn skillsets that are as or more diverse than existing approaches.

3.2 Practical Implementation

Next, we provide a practical implementation of the objective that removes the max and min operators. We first discuss how we optimize the skillset distributions ϕ and π because that informs how we will train the variational parameters ψ. Before discussing how the skillset distributions are updated updated, note that the Skillset Empowerment objective is not a Reinforcement Learning problem because the final step reward would include the $\log q(z|s_0, \phi, \pi, s_n)$ term, which depends on the skill-conditioned policy π. Optimizing Skillset Empowerment with RL would be subject to potentially significant nonstationary rewards because each time the skill-conditioned policy is updated, the final step reward could change.

In order to optimize π and ϕ with deep learning, these skillset distributions need to be represented as vectors. In our implementation of Skillset Empowerment, we implement the skill-conditioned policy π as a vector of parameters that represents the weights and biases of a neural network f_π that forms the skill-conditioned policy. $f_\pi : \mathcal{S} \times \mathcal{Z} \rightarrow \mathcal{A}$ takes as input a state s and skill z outputs the mean of a gaussian skill-conditioned policy. That is, $\pi(a|s, z) = \mathcal{N}(a; \mu = f_\pi(s, z), \sigma = \sigma_0)$, in which σ_0 is a small fixed standard deviation. $\phi(z|s_0)$ represents the distribution over skills given some skill start state s_0. Because we implement the distribution over skills as a uniform distribution over a d-dimensional cube centered at the origin, we define ϕ as a scalar value representing the side length of that cube. For instance, in our tasks in which the skills are two-dimensional, skills are sampled from a square with side length ϕ centered at the origin. The probability density $\phi(z|s_0) = (1/\phi)^d$.

We optimize both ϕ and π using their own bandit problem. In the ϕ bandit problem, the bandit policy is $f_\eta : \mathcal{S} \rightarrow \phi$, which takes as input the skill start state s_0 and outputs ϕ (i.e., the size of the uniform skill space). In the π bandit problem, the bandit policy $f_\lambda : \mathcal{S} \times \phi \rightarrow \theta$ takes as input the skill start state s_0 and a ϕ value and outputs π, the vector of parameters that define the skill-conditioned policy. In the π bandit problem, the reward for an π action is the variational lower bound on mutual information $I^{SE}(Z; S_n|s_0, \phi, \pi)$ (i.e., the reward is how diverse the skillset (ϕ, π) is). Similarly, the reward in the ϕ bandit problem for a ϕ action is $I^{SE}(Z; S_n|s_0, \phi, \pi = f_\lambda(s_0, \phi))$, in which the skill-conditioned policy π is the greedy output from the π bandit policy. Both bandit policies are optimized using an actor-critic structure. That is, to determine the gradient for the bandit policy f_η, a critic $f_\gamma(s_0, \phi)$ is trained to approximate $I^{SE}(Z; S_n)(s_0, \phi, f_\lambda(s_0, \phi))$ for ϕ values around the current greedy output $f_\eta(s_0)$. Similarly, a critic $f_\alpha(s_0, \phi, \pi)$ is used to approximate I^{SE} for sets of parameters π that are near the current greedy vector $f_\lambda(s_0, \phi)$. Figure 5 in the Appendix provides a visual overview of the two bandit problems.

The current setup is not yet practical because trying to learn the critic $f_\alpha(s_0, \phi, \pi)$ for the π actor when the vector of parameters π may be thousands of dimensions is infeasible. f_α would need to be able to discern the difference in mutual information when small changes are made to numerous parameters in π. Instead, because the gradient with respect to the π bandit policy f_λ only needs to know how the I^{SE} responds to small changes in each of the individual parameters in π, we instead train $|\pi|$ critics, $f^1(s_0, \phi, \hat{\pi^1}), \ldots, f^{|\pi|}(s_0, \phi, \hat{\pi^{|\pi|}})$. Each of these critics only takes a scalar, $\hat{\pi}_i$, as input, in which $\hat{\pi}_i$ represents the i-th parameter of π (e.g., could be a noisy value of the current greedy i-th parameter). The remaining parameters of π are assumed to take on their greedy values. Thus, each critic $f^i(s_0, \phi, \hat{\pi^i})$ only needs to approximate how the mutual information changes from small changes to a single parameter of π. All of these critics are updated in parallel. In Figure 4 in the Appendix we show how each of these $|\pi|$ critics attach to the bandit policy f_λ that outputs π.

Prior to updating the critics for each parameter of the skill-conditioned policy π, the variational distribution parameters ψ need to be updated so that I^{SE} forms a tighter bound on the true mutual information. Because we need to estimate I^{SE} for small changes in each of the parameters of π, we train $|\pi|$ different variational parameters $\psi^1, \ldots, \psi^{|\pi|}$. Each set of variational parameters ψ^i is trained to minimize the KL divergence between $p(z|s_0, \phi, \hat{\pi^i}, s_n)$ and $q_{\psi^i}(z|s_0, \phi, \hat{\pi^i}, s_n)$

Algorithm 1 Skillset Empowerment

```
Initialize variational posterior parameters $\psi^1, \ldots, \psi^{|\pi|}$
Initialize $\pi$ critic parameters $\alpha^1, \ldots, \alpha^{|\pi|}$ and actor parameters $\lambda$
Initial $\phi$ critic parameters $\gamma$ and actor parameters $\eta$
repeat
    Update in parallel $\psi^i$ by minimizing $D_{KL}(p(z|s_0, \phi, \hat{\pi^i}, s_n)||q_{\psi^i}(z|s_0, \phi, \hat{\pi^i}, s_n))$ for noisy $(\phi, \hat{\pi^i})$
    Update in parallel $\alpha^i$ s.t. $f_{\alpha^i}(s_0, \phi, \hat{\pi^i}) \approx I^{SE}(Z; S_n|s_0, \phi, \hat{\pi^i})$ for noisy $(\phi, \hat{\pi^i})$
    Update $\pi$ actor with gradient $\nabla_\lambda f_\alpha(s_0, \phi, f_\lambda(s_0, \phi))$ for noisy $\phi$
    Update $\gamma$ s.t. $f_\gamma(s_0, \phi) \approx I^{SE}(Z; S_n|s_0, \phi, f_\lambda(s_0, \phi))$ for noisy $\phi$
    Update $\eta$ with gradient $\nabla_\eta f_\gamma(s_0, f_\eta(s_0))$
until convergence
```

Algorithm 1 provides the full algorithm for the practical implementation of Skillset Empowerment. In the first step, the $|\pi|$ sets of variational parameters $\psi^1, \ldots, \psi^{|\pi|}$ are updated in parallel so that $q_{\psi^i}(Z; S_n|s_0, \phi, \hat{\pi^i}, s_n)$ better approximates the true posterior $p(Z; S_n|s_0, \phi, \hat{\pi^i}, s_n)$, in which $\hat{\pi^i}$ refers to the greedy output of $f_\lambda(s_0, \phi)$, except for i-th dimension which can take on noisy values. In the next step, the $|\pi|$ critics $f_{\alpha^1}, \ldots, f_{\alpha^d}$ are updated so that $f_{\alpha^i}(s_0, \phi, \hat{\pi^i})$ better approximates $I^{SE}(Z; S_n|s_0, \phi, \hat{\pi^i})$. Next, the actor that outputs π, f_λ is updated to produce skillsets (ϕ, π) with more diversity. Last, the critic and actor responsible for learning ϕ are updated.

Section B of the Appendix discusses how the $|\pi|$ sets of variational posteriors and the $|\pi|$ skill-conditioned policy critics can be training efficiently in parallel with the help of multiple accelerators and the parallelization capabilities of modern deep learning frameworks (e.g., JAX). Section C discusses the limitation of our approach, which is that it requires of a simulator of the transition dynamics. The simulator can be either be provided or learned like in prior empowerment works (Jung et al., 2012; Karl et al., 2015; 2017).

4 Experiments

We implement several experiments in stochastic settings to evaluate whether Skillset Empowerment can learn larger skillsets than existing empowerment methods and unsupervised GCRL.

4.1 Environments

Given that existing unsupervised skill-learning approaches like unsupervised GCRL already excel at learning large skillsets in deterministic settings, our experiments focused on significantly stochastic domains where no particular state was achievable with high probability. A key constraint on the set of environments we could choose from was that the transition dynamics of the environment needed to be sampled in parallel (i.e., a simulator of the environment needed to be available) due to the large parallel computation requirements of the empowerment algorithms we tested. To our knowledge, there are no existing benchmarks that satisfy both our stochasticity and parallelization requirements so we implemented our own domains in JAX. We briefly describe the four environments we created next. Additional details are provided in section F of the Appendix. Visualizations of each environment when a random sequence of actions is applied is shown in Figure 2.

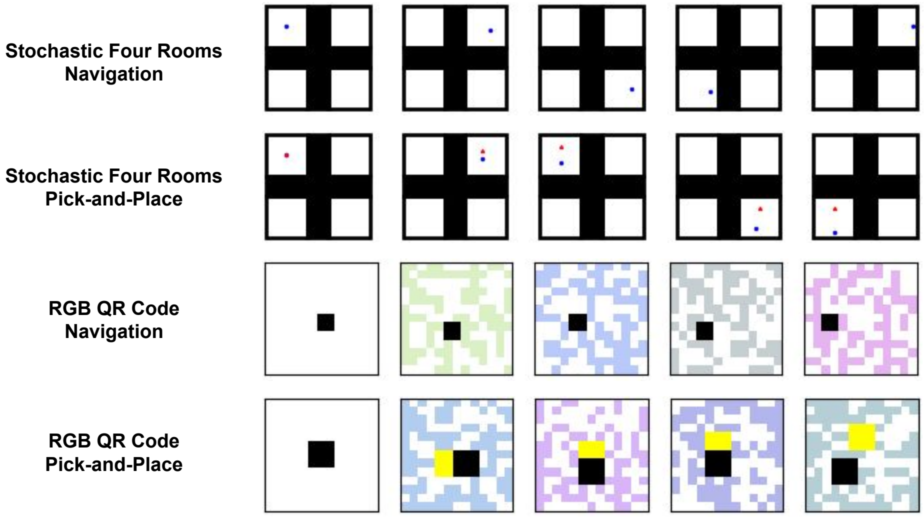

Stochastic Four Rooms Navigation
Stochastic Four Rooms Pick-and-Place
RGB QR Code Navigation
RGB QR Code Pick-and-Place

Figure 2: Sample state trajectories executed by a random policy in all four domains.

1. **Stochastic Four Rooms Navigation:** In this domain, a two-dimensional point agent navigates in an environment with four walled rooms by executing two-dimensional $(\Delta x, \Delta y) \in \mathcal{R}^2$ actions. The domain is highly stochastic because after each action is completed, the agent is moved to the same $(x \text{ offset}, y \text{ offset})$ location in a randomly sampled room. The abstract skills an agent should learn in this domains are skills that target $(x \text{ offset}, y \text{ offset})$ locations from the center of a room.

2. **Stochastic Four Room Pick-and-Place:** This is the same environment as the navigation task except there is now an object (the red triangle in the second row of Figure 2) that can be moved if the agent is within a certain distance of the object.

3. **RGB QR Code Navigation:** In this domain an agent learns to navigate amid a continually changing RGB-colored QR code background. Observations are 507-dimensional RGB images and are highly stochastic as the colored-QR code image fully changes after each action. Actions are discrete and consist of a horizontal movement (move east/west/stay) and a vertical movement (move north/south/stay).

4. **RGB QR Code Pick-and-Place:** This environment is the same as the navigation task except there is a now an object that can be moved if the object is in reach.

In all domains, there is a single skill start state and skills consist of 5 primitive actions.

4.2 Baselines

We compare our approach, Skillset Empowerment, to both a popular empowerment-based skill-learning method and unsupervised GCRL. For the prior empowerment-based method, we selected Variational Intrinsic Control (VIC) (Gregor et al., 2016). VIC, like the other popular approaches to optimizing empowerment including DIAYN (Eysenbach et al., 2018) and VALOR (Achiam et al., 2018), optimizes a loose lower bound on the mutual information of the skillset in which the variational posterior is trained to match the posterior of the current greedy skillset. For the unsupervised GCRL comparison, our focus was solely on whether goal-conditioned skills can be learned in stochastic domains and not on exploration which is the primary focus on recent unsupervised GCRL algorithms. Thus, we assist the unsupervised GCRL algorithm and provide it with the space of reachable states and thereby are just comparing our algorithm to supervised GCRL. We compare to a GCRL objective in which the variational posterior used is the tight diagonal gaussian centered

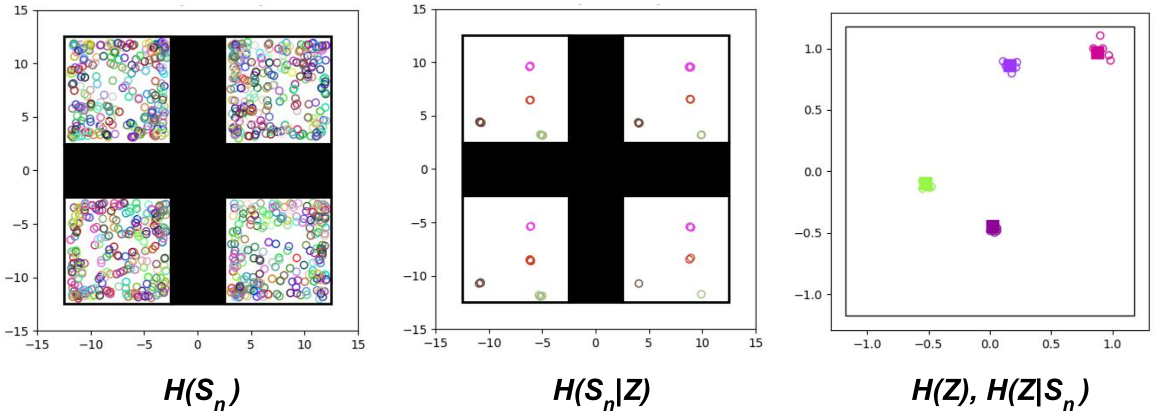


Figure 3: Entropy visualizations for the stochastic four rooms domain. Left image visualizes $H(S_n)$ by marking the skill-terminating state from 1000 skills randomly sampled. The center image visualizes $H(S_n|Z)$ by showing 12 samples of skill-terminating states from 4 specific skills randomly sampled. The right image visualizes (i) $H(Z)$ by showing the skill space (black rectangle) and (ii) $H(Z|S_n)$ by showing samples of the variational posterior (empty circles) for four different skills (filled squares)).

Table 1: Variational empowerment of learned skillsets for all baselines. The average Variational empowerment (measured in nats) over five random seeds and one standard deviation of error is shown.

Task	Ours	VIC	GCRL
Four Rooms Nav.	**5.1**± **0.3**	0.2± 0.4	0.3± 0.4
Four Room P.-and-P.	**8.7**± **0.3**	-0.1± 0.3	3.9± 0.6
RGB QR Nav.	**3.5**± **0.1**	-0.4± 0.0	-0.4± 0.3
RGB QR P.-and-P.	**6.0**± **0.2**	-0.6± 0.1	-2.6± 5.8

at the skill-terminating state discussed in section 2.2. For the higher dimensional QR code tasks, we implemented Reinforcement learning with Imagined Goals (RIG) (Nair et al., 2018). RIG performs GCRL in a latent space learned separately by a VAE. See section G for details on how the baselines were implemented.

4.3 Results

Our approach significantly outperforms both baselines in all tasks. Table 1, which notes the average variational empowerment (in nats) of the skillsets learned by each baseline, shows that our approach learns a much larger skillset than the baselines in all domains.

For additional evidence that our approach is able to effectively optimize the mutual information of the skill channel, we visualize the various entropies that appear in the symmetric definitions of empowerment—$H(S_n), H(S_n|Z), H(Z)$, and $H(Z|S_n)$—for the learned skillsets in all tasks. Figure 3 provides three images showing these visualizations for the stochastic four rooms navigation environment. The left image visualizes $H(S_n)$ by executing 1000 skills uniformly sampled from the learned skill space and marking the skill-terminating state with a colored circle. Per the image, $H(S_n)$ is large as the skill-terminating states produced by the sampled skills nearly uniformly cover the possible state space. The center image visualizes $H(S_n|Z)$ by focusing on four skills, uniformly sampled from the skill space, and for each skill sampling 12 skill-terminating states. Per the image, despite the room the room randomly changing at each transition, each skill targets a particular (x offset, y offset) position (i.e. $H(S_n|Z)$ is low), which is the correct abstract skill in this domain.

For example, the brown skill targets a specific position in the bottom left of each room, while the the pink skill targets a position towards the top right of each room. Thus, in this task, the agent is not only learning abstract skills but skills with the appropriate level of abstraction. The image on the right shows samples (empty circles) of the variational posterior distribution, $q_\psi(z|s_0, \phi, \pi, s_n)$ for four skills (filled squares) sampled from the learned skill space (inner black square). Per the image, $H(Z|S_n)$ is low (i.e., skills are targeting distinct groupings of states) because the samples from the variational posterior form narrow distributions around the sampled skill. If skills were targeting overlapping regions of states, the samples from the variational posterior would be more dispersed across the skill space.

The entropy visualizations for the remainder of the tasks are in section H of the Appendix. For instance, Figure 6 shows the same images for the stochastic four rooms pick-and-place task. The left image shows that $H(S_n)$ is large as the agent is able to execute skills that can achieve many of the possible (agent position, object position) tuples (object is shown by triangles). The center image, which visualizes $H(S_n|Z)$, shows that the agent is learning abstract skills that target (x offset, y offset) positions for both the agent and object. The image on the right visualizes $H(Z|S_n)$ for the now four-dimensional skill space and again the variational posterior forms narrow distributions around the sampled skill showing the agent is learning skills that target distinct groupings of states. Figures 7 and 8 in the Appendix provide the entropy visualization images for the RGB QR code navigation and pick-and-place tasks. Both of these figures show that despite significant stochasticity and high-dimensional observations, our approach learns diverse skillsets with the appropriate level of abstraction.

Moreover, the growth in empowerment from the navigation tasks to the pick-and-place tasks provides further evidence that our approach effectively optimizes empowerment. Moving from navigation tasks to pick-and-place tasks should boost empowerment by a significant factor as for every position the agent can reach, there may be a large number of object positions that can be achieved. The increase in empowerment by multiple nats confirms that empowerment grew by a large factor when an object was added to the environment. For instance, in the four rooms tasks, moving from the navigation task to the pick-and-place task increased the size of the skillset from 5.1 nats (164 skills) to 8.7 nats (6,000 skills)

On the other hand, the baselines were not able to learn large skillsets. For instance, in the stochastic four rooms task, the GCRL agent only learns skills to move to corners of the room as shown in Figure 9, which shows the skill-terminating states of 1000 random skills. More specifically, as shown in Figure 10, when given a goal state of some (x,y) position in one of the four rooms, the agent simply moves towards whichever room the goal is in regardless of where in the room the goal is. This behavior is likely taken to minimize the average distance to the goal as the goal-conditioned reward heavily penalizes the agent if it is far from the specific goal state. In the image-based QR code tasks, the VAE generally struggled to reconstruct the large variety of colored QR codes as shown in Figure 11, ultimately producing an image similar to a mean QR code. The overly abstract latent state space in turn made it challenging for the GCRL component to learn distinct skills. These results provide evidence that GCRL's loose lower bound on mutual information for diverse abstract skillsets GCRL does discourage agents from learning these desirable skillsets. On the other hand, because Skillset Empowerment can learn a tight bound on mutual information for diverse abstract skillsets, agents are encouraged to learn them. In addition, like GCRL, the performance of VIC agents also was poor. As with prior works, we observed stagnant skillsets.

5 Related Work

Developing a scalable way to optimize the mutual information of an action-perception channel has been a longstanding problem. Earlier empowerment work focused on the action-perception channel between open-loop action sequences and terminating-states. Klyubin et al. (2005; 2008) showed how the open-loop variant of empowerment can be computed in small domains with discrete state and action spaces using the Blahut-Arimoto algorithm (Blahut, 1972). Jung et al. (2012) scaled this method to continuous state space settings by integrating Monte Carlo approximation techniques. Mohamed & Rezende (2015) and Karl et al. (2017) further scaled the computation of open-loop empowerment to continuous action settings by optimizing a variational lower bound on mutual information in which a variational distribution replaces the true posterior. However, both of these approaches still optimize loose lower bounds on mutual information because they replace the true posterior for a candidate action sequence with a variational posterior trained to match the current action sequence. Gregor et al. (2016) extended empowerment to the skill channel, in which the mutual information between skills and skill-terminating states is optimized and the agent needs to learn both a distribution over skills and a skill-conditioned policy. Optimizing the mutual information of skill channel has two notable advantages over the open-loop action sequence channel. Because the distribution over the input variable is limited to simple distributions (e.g., uniform or gaussian distributions), when the open-loop channel is used the agent is limited in the variety of open-loop actions it can select. On the other hand, with the skill channel, even though the distribution of skills is limited (e.g., in our case, a uniform distribution in the shape of a d-dimensional cube), the marginal distribution over actions those skills can produce can be complex because of the skill-conditioned policy neural network that converts skills into actions. A second advantage of the skill channel is that it can learn closed loop policies.

Prior work has used the computation of empowerment for numerous downstream applications, including using empowerment as a state utility function (Klyubin et al., 2008; Salge et al., 2013b; Jung et al., 2012; Mohamed & Rezende, 2015; Karl et al., 2015; 2017), as an evolutionary signal to evolve sensors and actuators (Klyubin et al., 2005), as an objective for learning a state representation (Capdepuy, 2011; Bharadhwaj et al., 2022), as an intrinsic motivation reward (Oudeyer & Kaplan, 2007; Bharadhwaj et al., 2022), and as a way to measure human empowerment (Du et al., 2020; Myers et al., 2024). Empowerment-based skill learning methods that optimize the mutual information of the skill channel have used the learned distribution over skills $\phi(z|s_0)$ as temporally extended action spaces for downstream RL tasks (Eysenbach et al., 2018) or for learning hierarchical skills (Levy et al., 2023).

Also related to our work are the methods that learn abstract skills in settings with random environment distractors (Bharadhwaj et al., 2022; Fu et al., 2021; Ma et al., 2021; Zhang et al., 2020; Rudolph et al., 2024; Zou & Suzuki, 2024). However, these approaches only learn a single skill with the help of a reward function or require supervision in the form of a hand-crafted goal space. To our knowledge, our approach is the first unsupervised skill learning method to successfully learn large skillsets in stochastic settings.

6 Conclusion

Agents need to be able to execute large skillsets in settings with significant randomness. Skillset Empowerment takes a step in this direction by optimizing a tighter lower bound to the mutual information of the skill channel. Our experiments show that our approach is able to learn diverse abstract skillsets in domains with significant randomness.

References

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *CoRR*, abs/1807.10299, 2018. URL http://arxiv.org/abs/1807.10299.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *CoRR*, abs/1707.01495, 2017. URL http://arxiv.org/abs/1707.01495.

David Barber and Felix Agakov. The im algorithm: A variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pp. 201–208, Cambridge, MA, USA, 2003. MIT Press.

Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, and Sergey Levine. Information prioritization through empowerment in visual model-based rl, 2022. URL https://arxiv.org/abs/2204.08585.

R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. doi: 10.1109/TIT.1972.1054855.

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. *CoRR*, abs/2002.03647, 2020. URL https://arxiv.org/abs/2002.03647.

Philippe Capdepuy. *Informational principles of perception-action loops and collective behaviours*. PhD thesis, University of Hertfordshire, UK, 2011.

Jongwook Choi, Archit Sharma, Honglak Lee, Sergey Levine, and Shixiang Shane Gu. Variational empowerment as representation learning for goal-based reinforcement learning. *CoRR*, abs/2106.01404, 2021. URL https://arxiv.org/abs/2106.01404.

Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, and Anca Dragan. Ave: Assistance via empowerment. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4560–4571. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/30de9ece7cf3790c8c39ccff1a044209-Paper.pdf.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *CoRR*, abs/1901.10995, 2019. URL http://arxiv.org/abs/1901.10995.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *CoRR*, abs/1802.06070, 2018. URL http://arxiv.org/abs/1802.06070.

Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. Learning task informed abstractions. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3480–3491. PMLR, 18–24 Jul 2021. URL http://proceedings.mlr.press/v139/fu21b.html.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *CoRR*, abs/1611.07507, 2016. URL http://arxiv.org/abs/1611.07507.

David Held, Xinyang Geng, Carlos Florensa, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. *CoRR*, abs/1705.06366, 2017. URL http://arxiv.org/abs/1705.06366.

Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for continuous agent-environment systems. *CoRR*, abs/1201.6583, 2012. URL http://arxiv.org/abs/1201.6583.

Maximilian Karl, Justin Bayer, and Patrick van der Smagt. Efficient empowerment, 2015. URL https://arxiv.org/abs/1509.08455.

Maximilian Karl, Maximilian Soelch, Philip Becker-Ehmck, Djalel Benbouzid, Patrick van der Smagt, and Justin Bayer. Unsupervised real-time control through variational empowerment, 2017. URL https://arxiv.org/abs/1710.05101.

Seongun Kim, Kyowoon Lee, and Jaesik Choi. Variational curriculum reinforcement learning for unsupervised discovery of skills, 2023.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PLOS ONE*, 3(12):1–14, 12 2008. doi: 10.1371/journal.pone.0004018. URL https://doi.org/10.1371/journal.pone.0004018.

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, 2005. doi: 10.1109/CEC.2005.1554676.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric P. Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *CoRR*, abs/1906.05274, 2019. URL http://arxiv.org/abs/1906.05274.

Andrew Levy, Sreehari Rammohan, Alessandro Allievi, Scott Niekum, and George Konidaris. Hierarchical empowerment: Towards tractable empowerment-based skill learning, 2023.

Xiao Ma, SIWEI CHEN, David Hsu, and Wee Sun Lee. Contrastive variational reinforcement learning for complex observations. In Jens Kober, Fabio Ramos, and Claire Tomlin (eds.), *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 959–972. PMLR, 16–18 Nov 2021. URL https://proceedings.mlr.press/v155/ma21a.html.

Willie McClinton, Andrew Levy, and George Konidaris. HAC explore: Accelerating exploration with hierarchical reinforcement learning. *CoRR*, abs/2108.05872, 2021. URL https://arxiv.org/abs/2108.05872.

Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and achieving goals via world models. *CoRR*, abs/2110.09514, 2021. URL https://arxiv.org/abs/2110.09514.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning, 2015.

Vivek Myers, Evan Ellis, Benjamin Eysenbach, Sergey Levine, and Anca Dragan. Learning to assist humans without inferring rewards. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024. URL https://openreview.net/forum?id=pN8bDIqpBM.

Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *CoRR*, abs/1807.04742, 2018. URL http://arxiv.org/abs/1807.04742.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 02 2007. doi: 10.3389/neuro.12.006.2007.

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. *CoRR*, abs/2202.00914, 2022. URL https://arxiv.org/abs/2202.00914.

Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsupervised skill discovery, 2023a.

Seohong Park, Oleh Rybkin, and Sergey Levine. Metra: Scalable unsupervised rl with metric-aware abstraction, 2023b.

Silviu Pitis, Harris Chan, Stephen Zhao, Bradly C. Stadie, and Jimmy Ba. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. *CoRR*, abs/2007.02832, 2020. URL https://arxiv.org/abs/2007.02832.

Vitchyr H. Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning. *CoRR*, abs/1903.03698, 2019. URL http://arxiv.org/abs/1903.03698.

Max Rudolph, Caleb Chuck, Kevin Black, Misha Lvovsky, Scott Niekum, and Amy Zhang. Learning action-based representations using invariance, 2024. URL https://arxiv.org/abs/2403.16369.

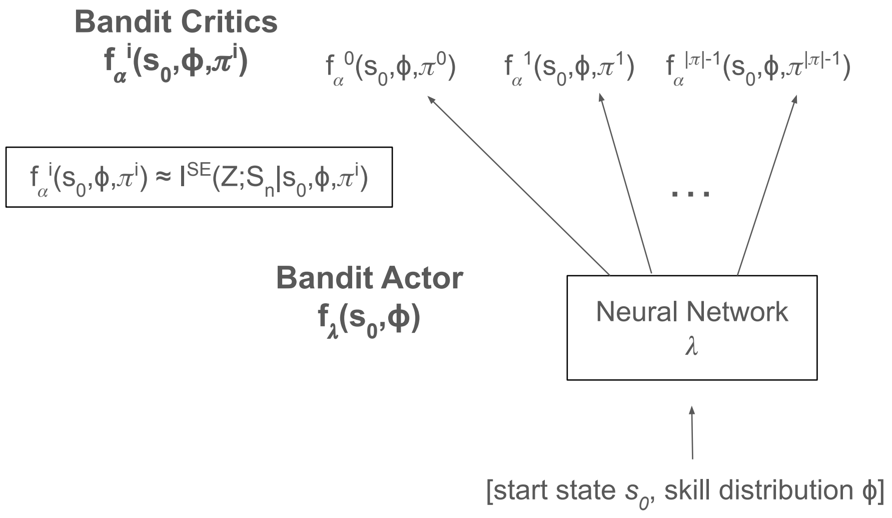

Bandit Critics
$f_{\alpha}^{i}(s_0,\phi,\pi^{i})$
$f_{\alpha}^{0}(s_0,\phi,\pi^{0})$
$f_{\alpha}^{1}(s_0,\phi,\pi^{1})$
$f_{\alpha}^{|\pi|-1}(s_0,\phi,\pi^{|\pi|-1})$
$f_{\alpha}^{i}(s_0,\phi,\pi^{i}) \approx I^{SE}(Z;S_n|s_0,\phi,\pi^{i})$
...
Bandit Actor
$f_{\lambda}(s_0,\phi)$
Neural Network
λ
[start state s_0, skill distribution ϕ]

Figure 4: Architecture for the actor-critic that trains the bandit policy f_λ that outputs the parameters of the skill-conditioned policy. The output of the actor has $|\theta|$ dimensions. For each of the dimensions, a critic $f_{\alpha^i}(s_0, \phi, \hat{\pi}_i)$ is trained to approximate the mutual information $I^{SE}(Z; S_n|s_0, \phi, \hat{\pi^i})$. During the actor update step, the actor uses the critic at each dimension to determine how to update each dimension of the skill-conditioned policy.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment - an introduction. *CoRR*, abs/1310.1863, 2013a. URL http://arxiv.org/abs/1310.1863.

CHRISTOPH Salge, CORNELIUS GLACKIN, and DANIEL POLANI. Approximation of empowerment in the continuous domain. *Advances in Complex Systems*, 16(02n03):1250079, 2013b. doi: 10.1142/S0219525912500798. URL https://doi.org/10.1142/S0219525912500798.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948. URL http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *CoRR*, abs/1907.01657, 2019. URL http://arxiv.org/abs/1907.01657.

DJ Strouse, Kate Baumli, David Warde-Farley, Vlad Mnih, and Steven Hansen. Learning more skills through optimistic exploration. *CoRR*, abs/2107.14226, 2021. URL https://arxiv.org/abs/2107.14226.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *CoRR*, abs/2006.10742, 2020. URL https://arxiv.org/abs/2006.10742.

Qiming Zou and Einoshin Suzuki. Compact goal representation learning via information bottleneck in goal-conditioned reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2024. doi: 10.1109/TNNLS.2023.3344880.

A Skill-Conditioned Policy π Actor-Critic Architecture

Figure 4 shows the actor-critic architecture for training the bandit policy that outputs π.

Table 2: Table shows various measures of the parallel computation demands for each environment.

Task	$\lvert\pi\rvert$	Update Time (s)	GPU Notes
Four Rooms Nav	386	37	1 A100 40GB or 2 H100 80GB SXM5
Four Rooms Pick	484	56	1 A100 40GB or 2 H100 80GB SXM5
RGB QR Nav	2528	10	8 A100 80GB SXM
RGB QR Pick	2528	10	8 A100 80GB SXM

B Parallelized Training

Although the number of skill-conditioned policy parameters may be large, the $|\pi|$ sets of variational posterior and π critic parameters can be trained efficiently using the parallelization capabilities of modern deep learning frameworks (e.g., JAX) and multiple accelerators. For instance, the update step for each set variational posterior parameters ψ^i requires (skill z, state s_n) tuples from skillsets $(\phi, \hat{\pi^i})$ in which small noise has been added to ϕ and the i-th parameter of π. Assuming access to multiple accelerators, this update step can occur in parallel for all $|\pi|$ sets of variational posterior parameters using the pmap and vmap functions in JAX. For instance, if the skill-conditioned policy π has 1000 dimensions and there are 4 GPUs available, each GPU can process the update to 250 sets of variational posterior parameters in parallel.

Table 2 provides some data on the parallel computation demands of our approach for each of our experiments. $|\pi|$ is the number of parameters in the skill-conditioned policy. Update Time reflects the time (in seconds) required to complete one whole update step (i.e., one iteration of the Repeat loop in Algorithm 1). Note that the update times shown for the four rooms tasks were when using a single A100 40GB device, while the update times for the RGB QR tasks reflect 8 A100 80GB SXM GPUs. When we used multiple GPUs, the update times were roughly 1/Num GPUs of the original time with a single GPU.

C Limitations

The major limitation of our approach is that it requires a simulator of the transition dynamics. In order to approximate each mutual information lower bound $I^{SE}(Z; S_n|s_0, \phi, \hat{\pi^i})$ for $i = 1, \dots, |\pi|$, we need to sample the (z, s_n) tuples that result from making small changes to the i-th parameter of π. Collecting these tuples in an online manner would require an intractable amount of interaction with the environment. Thus, we assume access to a model of the transition dynamics, which enables the large amount of (z, s_n) tuples to be sampled in parallel. A model of the simulator can also be learned as in prior empowerment methods (Jung et al., 2012; Karl et al., 2015; 2017).

D Goal Threshold Reward Function with a Fixed Posterior

For the typical goal threshold reward function in which the agent is only rewarded for entering within some threshold of the goal, a uniform distribution can be used in which the uniform distribution is centered at s_n and the half lengths of each dimension of the uniform distribution are given by the parameters $\sigma_1, \dots, \sigma_d$ for each of the d dimensions of the skill space (i.e., the goal state space): $q(z|s_0, s_n) = \mathcal{U}(z; \mu = s_n, \sigma = [\sigma_1, \dots, \sigma_d])$. If the goal state z is within the goal threshold of s_n, the $\log q(z|s_0, s_n)$ reward would be $1/\prod_{i=1}^{d} \sigma_i$. Otherwise, the reward could be set to 0.

E Overview of Bandit Problems

Figure summarizes the bandit problems for optimizing the skill-conditioned policy π and the distribution over skills ϕ.

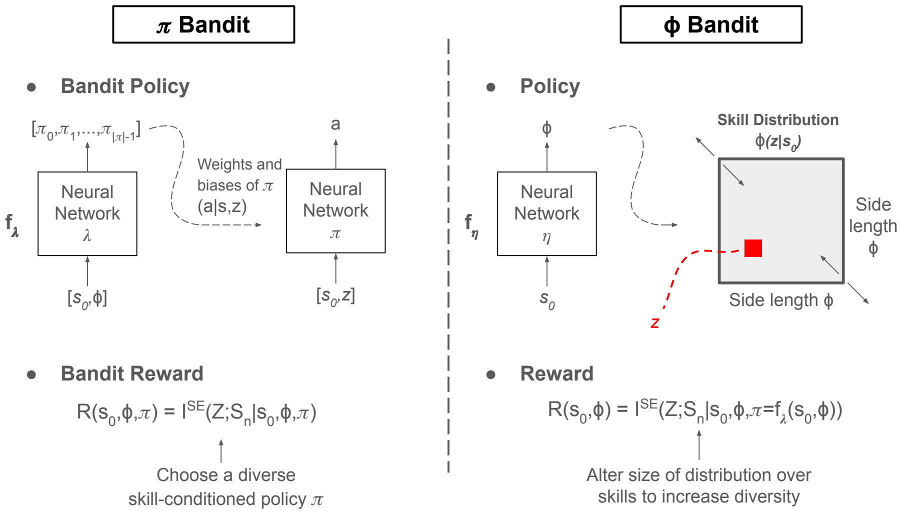

π Bandit
Bandit Policy
[π_0,π_1,...,π_|π|-1]
Weights and biases of π (a|s,z)
a
f_λ
Neural Network λ
Neural Network π
[s_0,ϕ]
[s_0,z]
Bandit Reward
R(s_0,ϕ,π) = I^SE(Z;S_n|s_0,ϕ,π)
Choose a diverse skill-conditioned policy π
ϕ Bandit
Policy
ϕ
f_η
Neural Network η
s_0
Skill Distribution ϕ(z|s_0)
Side length ϕ
Side length ϕ
z
Reward
R(s_0,ϕ) = I^SE(Z;S_n|s_0,ϕ,π=f_λ(s_0,ϕ))
Alter size of distribution over skills to increase diversity

Figure 5: Overview of the two bandit problems. In the π bandit problem, the task is to optimize the bandit policy $f_\lambda : \mathcal{S} \times \phi \rightarrow \phi$, which takes as input the skill start state s_0 and side length of the uniform distribution ϕ and output a vector π consisting of the parameters of the skill-conditioned policy neural network. That is, π is the vector of the weights and biases of the function f_π that determines of the mean of an action given a state and skill. The reward for a π action $R(s_0, \phi, \pi)$ is the mutual information lower bound $I^{SE}(Z; S_n|s_0, \phi, \pi)$ (i.e., the reward is how diverse the skillset (ϕ, π) is). In the ϕ bandit problem, the task is to learn the bandit policy $f_\eta : \mathcal{S} \rightarrow \phi$, which takes as input the skill start state s_0 and outputs ϕ, which is the side length of the d-dimensional cube uniform distribution centered at the origin. The reward for a particular ϕ action $R(s_0, \phi)$ is $I^{SE}(Z; S_n|s_0, \phi, \pi = f_\lambda(s_0, \phi))$. That is, the bandit policy is encouraged to output ϕ that results in diverse skillsets $(\phi, \pi = f_\lambda(s_0, \phi))$.

F Environment Details

We implemented the following environments.

1. **Stochastic Four Rooms Navigation:** This domain consists of a two-dimensional point agent in an environment with four walled rooms. The domain is highly stochastic because after each action is complete, the agent is placed in a room uniformly sampled. That is, the next location of the agent is the sum of (i) the agent's $(\Delta x, \Delta y)$ position relative to the center of its current room before the action, (ii) the next action, (iii) small gaussian noise, and (iv) the center of the new room. Thus, the "abstract" state in this domain is the $(\Delta x, \Delta y)$ position relative to the center of the current room. The observation space is two-dimensional and continuous consisting of the (x, y) position of the agent. The action space is also two-dimensional and continuous and consists of the the $(\Delta x, \Delta y)$ action. The action space for each dimension is the range $[-1, 1]$. The rooms are squares and have half lengths of size 5.

2. **Stochastic Four Room Pick-and-Place:** This is the same environment as the navigation task except there is now an object that can be moved. The observation space is four-dimensional consisting of the two-dimensional positions of the agent and object. The action space is also four-dimensional. The first two dimensions determine the change in the relative position of the agent. The final two dimensions determine the change in the relative position of the object. However, the final two dimensions of the action are only applied if both the x and y positions of the object are within 2.5 units of the agent. The agent starts in the same position as the object.

3. **RGB QR Code Navigation:** In this domain an agent learns to navigate amid a continually changing RGB-colored QR code background. The underlying dynamics in this environment are simple. The environment state, not visible to the agent, is the two-dimensional position of the agent. Actions are discrete and consist of a horizontal movement (i.e., move west, move east, or no change) and a vertical movement (i.e., move north, move south, or no change). In the underlying state space, the transition dynamics are deterministic. However, the observation space and dynamics are more complex. The observation space is high-dimensional consisting of $3x13x13$ RGB images (i.e., 507 total pixels), in which the dimensions represent (number of channels, image length, image width). The agent in every image is represented by a 3x2x2 black pixel. The domain is highly stochastic as the RGB-colored QR code background completely changes after each action.

4. **RGB QR Code Pick-and-Place:** This environment is the same as the navigation task except there is a now an object that can be moved. The object is always represented by a yellow square of pixels. In additional to the agent position changes, agent actions now also include a horizontal object position change (i.e., no change, move west, move east) and a vertical object position change (i.e., no change, move north, move south). The box can be moved if the both the underlying x and y positions of the box are within two units of the the agent. The agent starts in the same position as the object.

G Baseline Details

For the GCRL comparison, in the low-dimensional stochastic four rooms domains, we compared against the variant of GCRL that is a lower bound to Empowerment, in which the variational posterior consists of a tight diagonal gaussian centered at the skill-terminating state. The goal distribution is set to the distribution of all reachable state (e.g., all possible agent (x, y) positions in the stochastic four rooms navigation task). For the higher dimensional QR code tasks, we implemented Reinforcement learning with Imagined Goals (RIG) Nair et al. (2018). RIG is an unsupervised GCRL method that combines representation learning and GCRL. RIG uses a VAE to separately learn an encoder that maps state to distributions over skills and a decoder that maps latent states

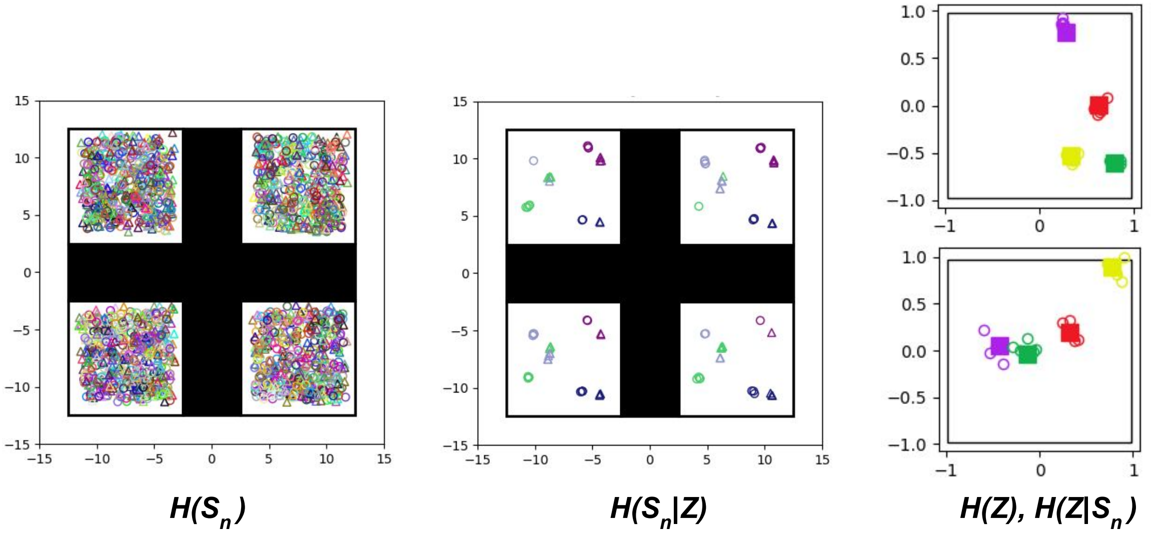


Figure 6: Images show the entropy visualizations for the stochastic four rooms pick-and-place domain. The left image shows the skill-terminating states s_n that result from 1000 skills uniformly sampled from the learned skill space. The near uniform coverage of the state space shows that $H(S_n)$ is large. The middle image focuses on four skills, uniformly sampled from the skill space, and for each skill shows 12 samples of skill-terminating states. Per the image, each skill targets an abstract state representing an offset from the center of a room for both the agent and object, showing that $H(S_n|Z)$ is low. The right image focuses on four skills and shows 5 samples from the variational posterior $q_\psi(z|s_0, l, \theta, s_n)$. Per the image, the samples form a narrow distribution around the executed skill, showing that $H(Z|S)$ is low.

to distributions over observations. RIG then performs GCRL in the learned embedding space (i.e, the agent learns skills that target specific latent states). Because the focus of this paper is not exploration, we make it easier on the representation learning component of RIG and provide it with a large dataset of reachable observations (e.g., images of the agent and object in a large variety of positions in the pick-and-place QR code environment.) The goal distribution for the GCRL phase is the prior distribution $p(z)$ from the VAE component.

H Visualizations

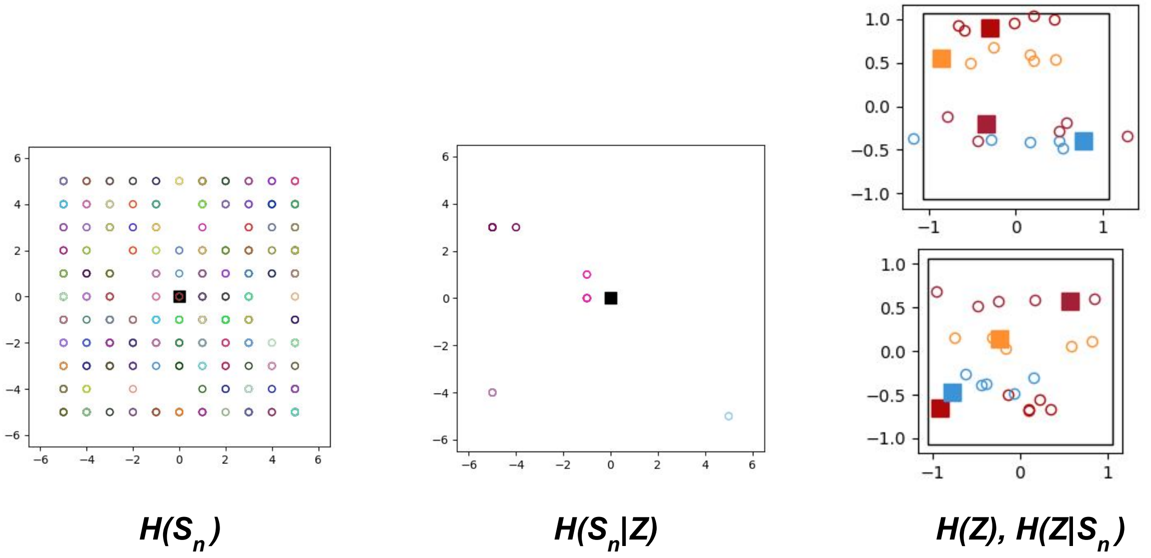
Figure 6 shows that our approach effectively optimizes empowerment in the stochastic four rooms pick-and-place domain by visualizing the entropies and conditional entropies of the learned skillset.

Figure 7 visualizes the entropy terms for the RGB QR code navigation task. Note that in this implementation, we used a four dimensional skill space which is twice as large as the two dimensional underlying state space. Per the image on the right, the algorithm corresponds by only using two dimensions of the latent state space, which is why you see the horizontal lines formed by the variational posterior.

Figure 8 visualizes the entropy terms for the RGB QR code pick-and-place task. Per the images the agent learns abstract skills to move itself and the object to specific regions of the underlying (x,y) space.

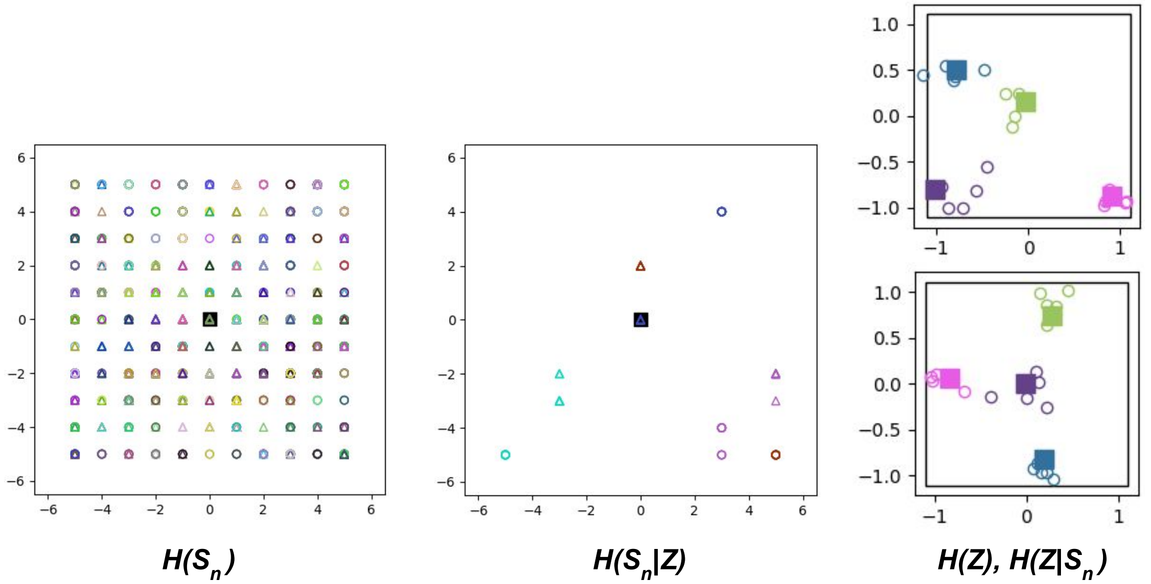
Figure 9 shows the state coverage of the GCRL agent in the stochastic four rooms task. Per the image, the agent only learns skills to target the corners of rooms.

Figure 10 shows the behavior of the skills learned by the GCRL agent. Each skill simply moves in the direction of the room of the goal state regardless of where in the room the goal is. For instance,



$H(S_n)$ $H(S_n|Z)$ $H(Z), H(Z|S_n)$

Figure 7: Entropy visualizations for the RGB QR code navigation task. Left image visualizes $H(S_n)$ by marking the skill-terminating states s_n produced by executing 1000 samples of skills from the learned skill space. Center image visualizes $H(S_n|Z)$ by executing four skills 12 times each and recording the skill-terminating states. Each skill targets an abstract (x,y) position. The right image shows samples from the variational posterior distribution. Note that in this case, the latent space is four dimensional even though the underlying state space is two dimensional. Because the agent does not need those extra dimensions, you see the horizontal lines in the variational posterior visualization.



$H(S_n)$ $H(S_n|Z)$ $H(Z), H(Z|S_n)$

Figure 8: Entropy visualizations for the RGB QR code pick-and-place tasks. Left image visualizes $H(S_n)$ by marking the skill-terminating states s_n produced by executing 1000 samples of skills from the learned skill space. Center image visualizes $H(S_n|Z)$ by executing four skills 12 times each and recording the skill-terminating states. Each skill targets an abstract (x,y) position for both the agent and object. The right image shows samples from the variational posterior distribution. Per the visuals, as expected, $H(S_n)$ is large while the conditional entropies $H(S_n|Z)$ and $H(Z|S_n)$ are small.

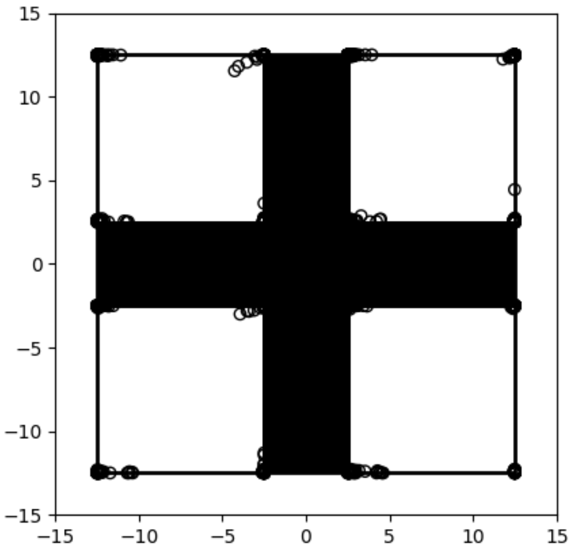


Figure 9: GCRL state coverage in stochastic four rooms domain.

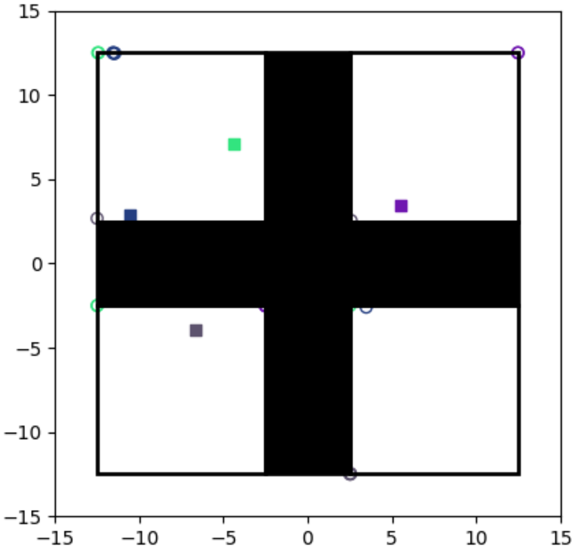


Figure 10: Image shows the skill-terminating states (empty circles) from four randomly selected goal states (filled squares). Each skill just moves in the direction of the room the skill is in.

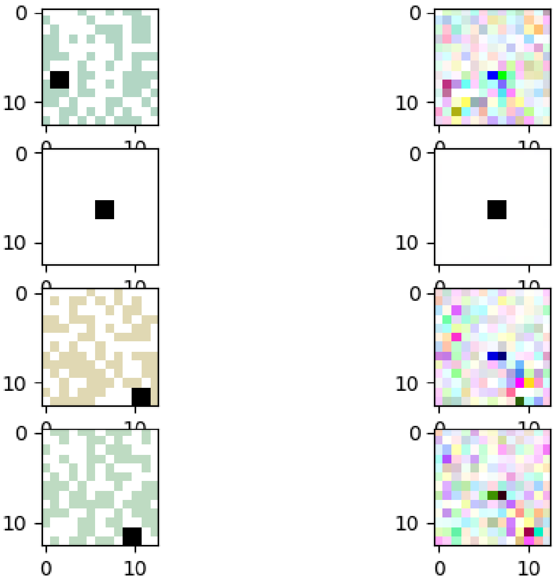


Figure 11: Image shows a sample of the VAE results in the RGB QR code navigation task. The left column shows sample images from the environment and the right column shows the results when those samples are encoded and then decoded. The VAE was able to decode the initial state of the environment, which is just a white background with the agent in the center, but struggled for other states.

given the purple goal (shown by purple square) in the lower left of the top right room, the agent just moves to the top right, which you can see by the purple circles in the top right room and bottom left room (hard to see). Similarly for the dark blue goal, the agent just moves to the top left, which you can see by the blue circles in the top left of the various rooms.

Figure 11 shows a sample of the VAE results in the RGB QR navigation domain. The VAE was able to decode the start state of a white background with the black agent in the center, but struggled with the other states. The area in which the agent was located would sometimes have slightly darker pixels but not enough to distinguish the state.