Representation Learning and Skill Discovery with Empowerment

Anonymous authors
Paper under double-blind review

Keywords: Unsupervised Skill Discovery, Representation Learning, Empowerment

Summary

Representation learning and unsupervised skill discovery remain key challenges for training reinforcement learning agents. We show that the empowerment objective, which measures the maximum number of distinct skills an agent can execute from some representation, enables agents to simultaneously perform representation learning and unsupervised skill discovery. Our theoretical analysis shows that empowerment provides a principled objective for learning sufficient statistic representations of observations. In addition, we show that the empowerment objective, when combined with a new approach to mutual information maximization, enables agents to learn large skillsets.

Contribution(s)

1. We prove that for any encoder that maps observations to a learned representation, the average empowerment achieved by the encoder is upper bounded by the average empowerment achieved by an encoder that outputs sufficient statistics of observations.
   Context: Prior work has proven that the average empowerment produced by an observation encoder is upper bounded by the average empowerment conditioned on the state representation (Capdepuy, 2011). We prove this is a looser upper bound than our own. This bound is also not achievable in partially observable settings where agents are not able to learn mappings from observations to underlying states.
2. We introduce a new approach to maximizing the mutual information between skills and observations that uses bandit reinforcement learning.
   Context: None
3. We provide empirical evidence that the empowerment objective can be used to jointly learn (i) representations suitable for reinforcement learning and (ii) large sets of skills that can be executed from the learned representations.
   Context: None

Representation Learning and Skill Discovery with Empowerment

Anonymous authors
Paper under double-blind review

Abstract

Representation learning and unsupervised skill discovery remain key challenges for training reinforcement learning agents. We show that the empowerment objective, which measures the maximum number of distinct skills an agent can execute from some representation, enables agents to simultaneously perform representation learning and unsupervised skill discovery. We provide theoretical analysis that empowerment can help agents learn sufficient statistic representations of observations because the maximum number of distinct skills an agent can execute from a learned representation grows when that representation does not combine multiple observations associated with different sufficient statistics. To jointly learn representations and skills, we use a new approach to mutual information maximization that uses bandit reinforcement learning. Under this approach, the agent learns a bandit policy that maps the skill starting representation to a vector that contains the set of parameters that make up the skill-conditioned policy. The reward for a skill-conditioned policy action is the variational lower bound on mutual information conditioned on that policy, which measures the diversity of the skill-conditioned policy action. Empirically, we demonstrate that our approach can (i) learn significantly more skills than existing unsupervised skill discovery approaches and (ii) learn a representation suitable for downstream reinforcement learning applications.

1 Introduction

Representation learning and unsupervised skill discovery have shown to be helpful capabilities for reinforcement learning (RL) agents. Both capabilities can boost sample efficiency as compact representations can simplify the policy that an agent needs to learn (Laskin et al., 2020), and skills can assist with exploration (Nachum et al., 2019) and accelerate credit assignment (Levy et al., 2019). Despite the importance of both of these capabilities, prior work has mostly focused on only one of these two capabilities.

The purpose of this work is to demonstrate that a single objective, empowerment, enables agents to perform both representation learning and skill discovery simultaneously. The empowerment of a representation, which is the maximum mutual information between skills and observations conditioned on the representation under consideration, measures the maximum number of distinct policies or skills that can be executed from that representation over a certain time horizon. In the context of empowerment, a distinct skill is one that targets a set of one or more observations that is not targeted by other skills in the agent's skillset. For example, consider an agent that moves within a 2D room and observes its (x, y) position. The empowerment of the agent when it starts in the center of the room is the largest set of skills, in which each skill targets a unique precise region of the (x, y) space, assuming there is a small amount randomness in the transition function. Figure 1 (Left) (a) illustrates the trajectories that some of these skills could produce.

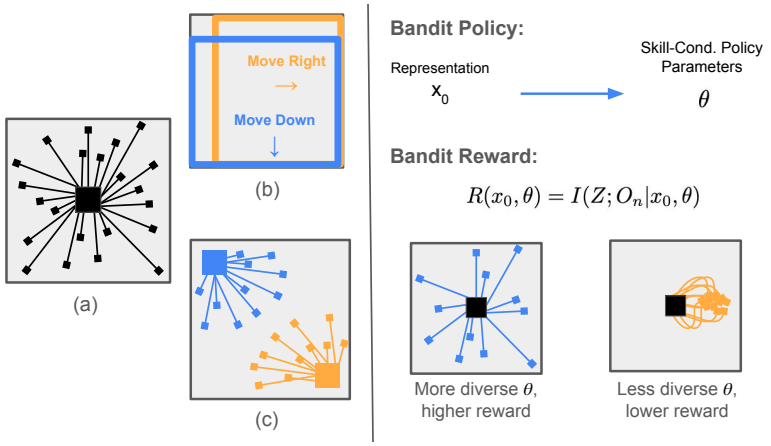

Figure 1: (Left) (a) Some of the distinct skills an agent can execute from the center of room that each target a unique (x, y) region. (b) Assuming all observations are encoded to the same representation, the orange box shows the (x, y) positions that could be targeted by a skill that moves the agent to the right, and the blue box shows the (x, y) positions that could be targeted by a skill that moves the agent down in the same scenario. Skills are highly stochastic (i.e., the boxes are large) because the starting observation can be anywhere in the room. (c) When different observations are not aliased, agents can learn different skillsets for different starting representations, such as when the agents learns skills to move up and to the left when it starts in the bottom right corner (orange square). (Right) Overview of the bandit RL approach to empowerment. The policy maps the skill starting representation to a vector, θ, containing the parameters of the skill-conditioned policy neural network. The reward for an skill-conditioned policy action θ is the mutual information of that policy, which measures the diversity of the policy.

We show that the empowerment objective enables agents to perform representation learning as it provides a principled way to learn sufficient statistic representations of observations, which are essential for downstream reinforcement learning (RL) tasks. The empowerment objective helps agents learn sufficient statistic representations because it discourages agents from encoding observations associated with different sufficient statistics to the same representation. This type of observation aliasing is discouraged because it reduces the number of distinct skills that can be executed from a learned representation. One reason is that unnecessary aliasing can make skills more stochastic, which in turn can result in redundant skills that target the same observations. Figure 1 (b) illustrates an extreme example of this in the 2D world in which the agent has learned to encode all observations into the same representation. In this scenario, skills that produce different actions (such as one skill that moves the agent right and another that moves the agent down) now target similar observations and become redundant. A second reason incorrectly mapping different observations to similar encodings reduces the number of distinct skills is that is forces similar skillsets to be applied to the aliased observations. For instance, in the 2D room, encoding observations where the agent starts in the top left of the room to the same representation as when the agent starts in the bottom right, forces the agent to execute similar skillsets in both situations, which can result in redundant skills where different skills cause the agent to target the same position on a wall. On the other hand, if these observations were mapped to different representations, the agent could learn larger skillsets tailored for the specific starting representation, such as moving down and to the right when the agent starts in the top left of the room.

We also show that the empowerment objective, when combined with a new approach to mutual information maximization, can be used to learn large skillsets. After the inconsistent performance of earlier methods that tried to discover skills using empowerment (Gregor et al., 2016; Eysenbach et al., 2019), recent work has moved away from maximizing a pure mutual information objective for learning skills, arguing that the empowerment objective is not sufficient and that additional bonus

terms need to be added (Laskin et al., 2022; Strouse et al., 2022; Zheng et al., 2025; Baumli et al., 2021; Kim et al., 2023) or that empowerment is fundamentally not capable of learning large skillsets in continuous settings (Park et al., 2022; 2024). We show that a key reason for the inconsistent performance of earlier empowerment methods was that they were not properly using reinforcement learning to optimize the mutual information objective. Earlier approaches treated the mutual information objective as a typical sequential decision making problem in RL in which the tuple of states and primitive actions is a sufficient statistic for reward. But in the mutual information objective in empowerment, this is not true as the reward depends on the state and the complete skill-conditioned policy rather than a single primitive action. Treating the tuple of states and actions as Markov with respect to reward when it is not results in nonstationary rewards which can make training unstable. In addition, we show that the reward function used by prior work is also flawed as it evaluates skill-conditioned policies using a loose lower bound on mutual information. Given that mutual information evaluates the diversity of a skill-conditioned policy, the loose lower bound means agents were often underestimating the diversity of skill-conditioned policies, which in turn made it difficult to learn a diverse skill-conditioned policy.

To overcome both of these issues, we introduce a bandit RL approach to maximizing the mutual information between skills and observations. The bandit policy the agent learns maps the skill starting representation to a vector of the neural network parameters that make up the skill-conditioned policy. The reward for proposing a particular skill-conditioned policy action from some starting representation is a variational lower bound on the mutual information between skills and observations *conditioned* on the proposed skill-conditioned policy, which provides a tighter bound on mutual information. Figure 1 (Right) provides an illustration of this bandit RL approach to mutual information maximization.

We evaluate whether our approach can jointly learn suitable representations for RL and skills in a variety of experiments. In the first set of experiments, we show that our approach can learn significantly larger skillsets than leading unsupervised skill discovery algorithms and several variants of our approach. In the second set of experiments, we demonstrate that the representations learned by our approach can serve as effective representations for downstream RL tasks.

2 Related Work

Related to our work are numerous other works in unsupervised skill discovery. Several of these works learn skills by maximizing the mutual information between skills and some function of observations (Mohamed & Rezende, 2015; Gregor et al., 2016; Eysenbach et al., 2019; Warde-Farley et al., 2019; Achiam et al., 2018; Hansen et al., 2020; Sharma et al., 2020; Zhang et al., 2021; Campos et al., 2020; Choi et al., 2021; Levy et al., 2023). Given the inconsistent performance of these methods, several other works emerged modifying the mutual information objective, typically adding particular bonus terms to the mutual information objective to help with exploration (Laskin et al., 2022; Zheng et al., 2025; Kim et al., 2023; Strouse et al., 2022; Baumli et al., 2021). Others works have claimed that empowerment is not capable of learning meaningful skillsets in continuous settings and instead argued that Lipschitz constraints (Park et al., 2022) or Wasserstein distances (Park et al., 2024) were superior objectives. Moreover, most prior work in unsupervised skill discovery does not focus on jointly learning representations to be used as inputs for skill-conditioned policies and downstream RL tasks. Prior work that has jointly learned representations and skills has used separate objectives for the two capabilities, and the representation learning objective involves image reconstruction which can be difficult settings with high-dimensional and noisy observations (Nair et al., 2018; Campos et al., 2020; Pong et al., 2020).

Also related to our approach are several works in representation learning. The most similar algorithms have been those that have used empowerment to learn representations (Klyubin et al., 2008; Capdepuy, 2011; Bharadhwaj et al., 2022). Our work builds on the work by Capdepuy (2011), which proved that the average maximum mutual information between primitive actions and observations conditioned on some learned representation is upper bounded by the average mutual information

conditioned on the state representation. But this is too loose of an upper bound in partially observable settings where agents cannot learn deterministic mappings from observations to states. We extend this result by proving a tighter and achievable bound that the average empowerment produced by an observation encoder is upper bounded by the average empowerment produced an encoder that outputs sufficient statistic representations. In addition, the mutual information term in our proof is between closed loop skills and observations, enabling our agents to simultaneously learn temporally extended actions and representations simultaneously. Other representation learning works similar to our own include methods that learn inverse dynamics models between observations and primitive actions (Lamb et al., 2023; Islam et al., 2022; Koul et al., 2024; Rudolph et al., 2024) as empowerment-based skill discovery algorithms learn similar models between observations and skills. A key difference from these works is again that our approach can also learn skills.

3 Background

3.1 Problem Setting

We assume that agents operate in partially observable settings with Markov observations. These environments are defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, p(s_0), p(s_{t+1}|s_t, a_t), p(o_t|s_t))$. $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$ represent the state, action, and observation spaces, respectively; $p(s_0)$ is the initial state distribution; $p(s_{t+1}|s_t, a_t)$ is the state transition dynamics; and $p(o_t|s_t)$ is the observation distribution. Note that states are not visible to the agent. In this setting, Markov observations means that given the current observation o_t and action a_t, the distribution over the next observation o_{t+1} is conditionally independent of the history of actions and observations $h_t = o_0, a_0, o_1, \ldots, a_{t-1}, o_t$: $p(o_{t+1}|h_t, o_t, a_t) = p(o_{t+1}|o_t, a_t)$. We also assume that all environments have one or more deterministic functions $f_x : \mathcal{O} \to \mathcal{X}$ that map an observation to a sufficient statistic $x \in \mathcal{X}$ of the observation with respect to the next observation. This means that for any $(o_t, x_t = f_x(o_t))$ tuple and any action a_t, the distribution over the next observation o_{t+1} given sufficient statistic x_t and a_t is conditionally independent of the observation o_t: $p(o_{t+1}|o_t, x_t, a_t) = p(o_{t+1}|x_t, a_t)$. Note that f_x is not provided to the agent.

3.2 Empowerment

We define the empowerment of an observation o_0 as the maximum mutual information between a policy random variable Π and a policy-terminating observation random variable O_n:

$$\mathcal{E}(o_0) = \max_{p(\pi|o_0)} I(\Pi; O_n|o_o). \tag{1}$$

Equation 1 means that empowerment measures the maximum number of distinct policies π that can be executed from observation o_0. Because it is unclear how to learn a distribution over policies $p(\pi|o_0)$, we will instead work with a lower bound of equation 1 that is common in prior work (see section A for proof of the lower bound):

$$\mathcal{E}(o_0) = \max_{\pi_z} I(Z; O_n|o_0, \pi_z) \tag{2}$$

In this definition, the empowerment of observation o_0 is the maximum mutual information between a skill random variable Z and skill-terminating observation random variable O_n conditioned on the skill-conditioned policy $\pi_z : \mathcal{O} \times \mathcal{Z} \to \mathcal{A}$, which is a mapping from observations and skills to actions. Note that the maximum, which is with respect to π_z, could also be with respect to the distribution over skills $p(z|o_0)$, but as in most prior work, we will assume this distribution is fixed. Specifically, we will assume skills are uniformly sampled from the range $[-1, 1]$ for each of the d dimensions of the skill space: $z \sim \mathcal{U}(-1, 1)^d$. Thus, line 2 defines empowerment as the largest number of distinct skills that can be executed from observation o_0 using some skill-conditioned policy π_z. The mutual information term in line 2 can be further defined

$$I(Z; O_n|o_0, \pi_z) = \mathbb{E}_{z \sim p(z), p(o_n|o_0, \pi_z, z)}[\log p(z|o_0, \pi_z, o_n) - \log p(z)] \tag{3}$$

Given that in continuous settings computing the mutual information $I(Z; O_n|o_0, \pi_z)$ is not tractable due to the posterior term $p(z|o_0, \pi_z, o_n)$, it is common to instead work with a variational lower bound of mutual information, $I^V(Z; O_n|o_0, \pi_z)$, in which the problematic posterior is replaced with a variational distribution $q_\psi(z|o_0, \pi_z, o_n)$ with trainable parameters ψ:

$$I^V(Z; O_n|o_0, \pi_z) = \mathbb{E}_{z\sim p(z), o_n\sim p(o_n|o_0,\pi_z,z)}[\log q_\psi(z|o_0, \pi_z, o_n) - \log p(z)]. \tag{4}$$

Note that for any skill-conditioned policy π_z, the gap between the true mutual information $I(Z; O_n|o_0, \pi_z)$ and the variational lower bound of mutual information $I^V(Z; O_n|o_0, \pi_z)$ is an average KL divergence between the true and variational posteriors: $I(Z; O_n|o_0, \pi_z) - I^V(Z; O_n|o_0, \pi_z) = \mathbb{E}_{o_n\sim p(o_n|o_0,\pi_z)}[D_{KL}(p(z|o_0, \pi_z, o_n)||q_\psi(z|o_0, \pi_z, o_n))]$ (Barber & Agakov, 2003; Poole et al., 2019). Thus, I^V can accurately measure the diversity of a skillset defined by the skill-conditioned policy π_z if the variational posterior is close to the true posterior.

3.3 Prior Approaches to Mutual Information Maximization

Earlier approaches to empowerment-based skill discovery sought to maximize the variational lower bound on mutual information $I^V(Z; O_n|o_0, \pi_z)$ using an approach that alternates between two steps (Gregor et al., 2016; Eysenbach et al., 2019; Hansen et al., 2020). In the first step of the update, the KL divergence between the posterior of the current skill-conditioned policy π_z^{Current}, $p(z|o_0, \pi_z^{\text{Current}}, o_n)$ and the variational posterior $q_\psi(z|o_0, o_n)$ is minimized. In the second step, I^V is optimized with respect to the skill-conditioned policy π_z using a typical skill-conditioned RL approach. For instance, in the first empowerment-based skill learning approach, VIC (Gregor et al., 2016), the reward function for training the skill-conditioned policy is 0 for the first $n-1$ actions and then the final step reward is the log variational posterior term: $R(o_n, z) = \log q_\psi(z|o_0, o_n)$.

There are two problems with this approach. The first issue is that in contrast to the typical skill-conditioned RL problem, the reward is not only a function of observations and skills (i.e., observations and skills are not jointly sufficient statistics with respect to reward). Instead, the reward is also a function of the full skill-conditioned policy π_z because the posterior term $q_\psi(z|o_0, o_n)$ depends on the true posterior $p(z|o_0, \pi_z, o_n)$, which depends on π_z. This is a problem because changes to the skill-conditioned policy for any skill z can cause changes in the reward, which can result in significant instability. The second problem is that the reward function is flawed because skill-conditioned policies π_z are evaluated with a variational mutual information I^V in which the posterior $q_\psi(z|o_0, o_n)$ is not trained to match the true posterior the skill-conditioned policy π_z under consideration. That is, for π_z that differ from the current skill-conditioned policy, there can be a gap between the true posterior of π_z, $p(z|o_0, \pi_z, o_n)$, and the variational posterior $q_\psi(z|o_0, o_n)$, which means I^V can be a loose lower bound on the true mutual information and that the agent may be underestimating the diversity of π_Z. This in turn can discourage the agent from changing its skill-conditioned policy even when those changes can produce a more diverse skill-conditioned policy.

4 Learning Sufficient Statistic Representations with Empowerment

In this section, we show that training an observation encoder to maximize the average empowerment of learned representations provides a principled way to learn sufficient statistic representations of observations. Sufficient statistics of observations are critical to using reinforcement learning in a learned representation space because they enable agents to replace potentially high-dimensional observations with more compact representations as policy inputs as discussed in section B.

4.1 Empowerment of a Learned Representation

Prior to providing our proof that empowerment can help observation encoders learn sufficient statistic representations, we first define the empowerment of a learned representation or context $c_0 \in \mathcal{C}$:

$$\mathcal{E}(c_0, f_c) = \max_{\pi_z} I(Z; O_n | c_0, f_c, \pi_z). \tag{5}$$

In line 5, $f_c : \mathcal{O} \to \mathcal{C}$ refers to the encoder that maps observations to the learned representation space, and $\pi_z : \mathcal{C} \times \mathcal{Z} \to \mathcal{A}$ is the skill-conditioned policy that maps contexts and skills to primitive actions. Note than this definition of empowerment also takes as input the observation encoder f_c because the skill-terminating observation o_n depends on actions that depend on f_c. The mutual information can be further defined:

$$I(Z; O_n | c_0, f_c, \pi_z) = \mathbb{E}_{z \sim p(z), o_n \sim p(o_n | c_0, f_c, \pi_z, z)}[\log p(z | c_0, f_c, \pi_z, o_n) - \log p(z)], \tag{6}$$

in which the channel distribution $p(o_n | c_0, f_c, \pi_z, z)$ is a marginal of the joint distribution $p(x_0, a_0, o_1, \ldots, x_{n-1}, c_{n-1}, a_{n-1}, o_n | c_0, f_c, \pi_z, z) = p(x_0|c_0, f_c)p(a_0|c_0, z)p(o_1|x_0, a_0) \ldots p(x_{n-1}|o_{n-1}, f_x)p(c_{n-1}|o_{n-1}, f_c)p(a_{n-1}|c_{n-1}, z)p(o_n|x_{n-1}, a_{n-1})$. Note that $p(x_0|c_0, f_c)$ represents the distribution of sufficient statistics x_0 that context c_0 is aliasing.

4.2 Theoretical Analysis

In this section, we provide our main theoretical result and then sketch out the proof. The full proof is provided in the appendix.

Theorem 1. *For any observation encoder f_c and encoder f_x that outputs sufficient statistics of observations, the average empowerment produced by the observation encoder f_c, $\mathbb{E}_{c_0 \sim p(c_0|f_c)}[\mathcal{E}(c_0, f_c)]$, is upper bounded by the average empowerment produced by the sufficient statistic encoder f_x, $\mathbb{E}_{x_0 \sim p(x_0|f_x)}[\mathcal{E}(x_0, f_x)]$.*

Proof Sketch. We first show that average maximum mutual information produced by the observation encoder is upper bounded by the average mutual information additionally conditioned on aliased sufficient statistics x_0: $I(Z; O_n | c_0, f_c, \pi_z^{*,c}) <= \mathbb{E}_{x_0 \sim p(x_0|c_0, f_c)}[I(Z; O_n | c_0, x_0, f_c, \pi_z^{c,*})]$, in which $x_0 \sim p(x_0|c_0, f_c)$ represents the aliased sufficient statistic and $\pi_z^{c,*}$ represents the mutual information maximizing policy for context c_0 and encoder f_c. This is an intuitive result because, as discussed in Figure 1 (b), skills can be less stochastic and redundant when executed from a known x_0 than from a c_0 aliasing multiple x_0. Next, because we need mutual information only in terms of sufficient statistics x_0 and the encoder f_x and not c_0 and f_c, we show that for any tuple of (c_0, x_0, z), there is a different skill-conditioned policy π_z' such that $I(Z; O_n | c_0, x_0, f_c, \pi_z^{c,*}) = I(Z; O_n | x_0, f_x, \pi_z')$. $\pi_z' : \mathcal{X} \times \mathcal{Z} \to \mathcal{A}$ now maps sufficient statistics and skills to actions. Finally, we can upper bound the resulting average of mutual information terms by replacing π_z' with the optimal skill-conditioned policy $\pi_z^{*,x}$ for sufficient statistic x_0 and encoder f_x. This last step is equivalent to discussion of Figure 1 (c), in which we noted that different representations can require different optimal skill-conditioned policies to maximize the number of distinct skills.

5 Maximizing Mutual Information with Bandit RL

This section discusses our bandit reinforcement learning approach to maximizing mutual information. We first set aside the representation learning component and show how we use bandits to maximize mutual information with respect to the skill-conditioned policy only. Then we explain how nearly the same bandit RL approach can be used to maximize mutual information with respect to both the skill-conditioned policy and observation encoder simultaneously.

To maximize mutual information with respect to the skill-conditioned policy we use a particular bandit reinforcement learning setup. The bandit policy $f_\lambda : \mathcal{O} \to \Theta_z$ maps the skill-starting representation, which for now is observations, to the neural network parameters θ_z (i.e., the weights and

biases) of the skill-conditioned policy MLP. The reward for proposing a skill-conditioned policy defined by θ_z in o_0 is the mutual information variational lower bound, $I^V(Z; O_n|o_0, \theta_z)$, in which the variational posterior $q_\psi(z|o_0, \theta_z, o_n)$ is conditioned on the proposed action θ_z and trained to match the true posterior $p(z|o_0, \theta_z, o_n)$. This bandit approach overcomes both problems from prior work: (i) o_0 and θ_z are together sufficient statistics with respect to reward I^V and (ii) the reward function I^V can represent a tight bound on mutual information.

The main challenge is how to implement this bandit RL approach in practice. A naive actor-critic approach, in which an actor represents the bandit policy f_λ and a critic $Q_\eta(o_0, \theta_z)$ maps o_0 and θ_z to an estimate of $I^V(Z; O_n|o_0, \theta_z)$, is not practical because this would require a variational posterior q_ψ and a critic Q_η that take as input the θ_z vector, which can be thousands of dimensions long. Instead, we will use a different actor-critic approach that "simulates" the gradient from the naive actor-critic method. The key insight is that in the naive approach, the gradient of the critic with respect to any parameter λ_j in the actor f_λ is $\frac{dQ}{\lambda_j} = \sum_{i=0}^{|\theta_z|-1} \frac{dQ}{d\theta_z^i} \frac{d\theta_z^i}{d\lambda_j}$, in which θ_z^i is the i-th entry of the θ_z vector. (Section D of the supplementary materials shows this for a 1-hidden layer critic.) This means that we can match the gradients from the naive approach if we can accurately estimate $\frac{dQ}{d\theta_z^i}$ (i.e., how mutual information changes from small changes to one parameter of θ_z assuming the other parameters are constant). We take this approach using a new actor-critic architecture in which we train parameter-specific critics $Q_{\eta^i}(o_0, \pi_z^i)$ to respectively approximate $I^V(Z; O_n|o_0, \theta_z^i)$ for $i = 0, \ldots, |\theta_z| - 1$, in which θ_z^i is a *scalar* representing the skill-conditioned policy in which all entries in θ_z take on their greedy values from the actor $f_\lambda(o_0)$ except the i-th parameter which takes on value θ_z^i. We then use the trained critics to update the actor f_λ so that it outputs more diverse skill-conditioned policies θ_z. Figure 4 provides a visual of the parameter-specific actor-critic architecture.

Algorithm 1 Actor-Critic Method for Maximizing $I(Z; O_n|o_0, \theta_z)$ w.r.t. θ_z

for all dimensions $i = 0, \ldots, |\theta_z| - 1$ in parallel **do**
  for M iterations **do**
    Update q_{ψ^i}: $\psi^i \leftarrow \psi^i - \alpha \nabla_{\psi_i}(D_{KL}(p(z|o_0, \theta_z^i, o_n)||q_{\psi^i}(z|o_0, \theta_z^i, o_n)))$ with noisy θ_z^i
  end for
  for M iterations **do**
    Update Q_{η^i}: $\eta_i \leftarrow \eta_i - \alpha \nabla_{\eta_i}((Q_{\eta^i}(o_0, \theta_z^i) - \text{Target})^2)$ with noisy θ_z^i,
    Target $= \mathbb{E}_{z \sim p(z), o_n \sim p(o_n|o_0, \theta_z^i, z)}[\log q_{\psi^i}(z|o_0, \theta_z^i, o_n) - \log p(z)]$
  end for
end for
Update f_λ: $\lambda \leftarrow \lambda + \alpha \nabla_\lambda(\sum_{i=0}^{|\theta_z|-1} Q_{\eta^i}(o_0, \theta_z^i = f_\lambda(o_0)[i]))$

Algorithm 1 provides the full algorithm for the actor-critic method for maximizing mutual information with respect to θ_z. The first step is to train in parallel and until convergence all the variational posteriors $q_{\eta^i}(z|o_0, \theta_z^i, o_n)$ to match the true posteriors $p(z|o_0, \theta_z^i, o_n)$ for noisy values of θ_z^i. The second step is to train all critics $Q_{\eta^i}(o_0, \theta_z^i)$ until convergence to approximate variational mutual information. The final step is to update the actor.

Next, we discuss how we can use nearly the same bandit RL approach to jointly maximize mutual information with respect to both the skill-conditioned policy θ_z and the observation encoder f_c. The average mutual information objective we are trying to maximize is

$$\max_{f_c, \theta_z} \mathbb{E}_{c_0 \sim p(c_0|f_c)}[I^V(Z; O_n|c_0, f_c, \theta_z)] \tag{7}$$

We maximize this mutual information by alternating between two actor-critic algorithms. In the first algorithm, we fix the observation encoder f_c and maximize the mutual information of a context c_0 with respect to θ_z. That is, we perform Algorithm 1 with c_0 replacing o_0. In the second actor-critic algorithm, we hold the skill-conditioned policies constant and train the observation encoder. In this second actor-critic, the actor f_μ maps a fixed vector v to a vector containing the parameters of the observation encoder neural network (also referred to as f_c). The parameter-specific critics $Q_{\kappa^i}(f_c^i)$

Algorithm 2 Actor-Critic Method for Maximizing $I(Z; O_n|c_0, f_c)$ w.r.t. f_c

for all dimensions $i = 0, \ldots, |f_c| - 1$ in parallel **do**
  for M iterations **do**
    Update q_{ω^i}: $\omega^i \leftarrow \omega^i - \alpha \nabla_{\psi_i}(D_{KL}(p(z|o_0, f_c^i, o_n)||q_{\omega_i}(z|o_0, f_c^i, o_n)))$ with noisy f_c^i
  end for
  for M iterations **do**
    Update Q_{κ_i}: $\kappa^i \leftarrow \kappa^i - \alpha \nabla_{\kappa_i}((Q_{\kappa^i}(f_c^i) - \text{Target})^2)$ with noisy f_c^i,
    Target $= \mathbb{E}_{c_0 \sim p(c_0|f_c^i), z \sim p(z), o_n \sim p(o_n|c_0, f_c^i, z)}[\log q_{\omega^i}(z|c_0, f_c^i, o_n) - \log p(z)]$
  end for
end for
Update f_μ: $\mu' \leftarrow \mu + \alpha \nabla_\mu (\sum_{i=0}^{|f_c|-1} Q_{\kappa_i}(f_c^i = f_\mu(v)[i]))$

approximates the average mutual information $\mathbb{E}_{c_0 \sim p(c_0|f_c^i)}[I^V(Z; O_n|c_0, f_c^i)]$ using the parameter-specific variational posteriors $q_{\omega_i}(z|c_0, f_c^i, o_n)$. Figure 5 provides a visual of the parameter-specific actor-critic architecture for training the observation encoder. Algorithm 2 provides the algorithm for training the observation encoder actor-critic.

Our approach currently has one main limitation, which is that it assumes the agent has learned a model of the transition dynamics, which can be challenging in noisy and high-dimensional settings. However, existing work (provided in the attached supplementary materials) has shown that mutual information can be optimized without needing to learn a (potentially high-dimensional) simulator of the environment. Instead, the mutual information between skills and observations can be maximized while only learning a transition model that predicts encodings of observations. We leave for future work to combine our approach with this model-based approach.

6 Experiments

Next, we discuss the experiments we implemented to evaluate our two main claims that (i) our bandit RL approach to mutual information maximization can learn larger skillsets than existing approaches to unsupervised skill discovery and (ii) our approach can learn sufficient statistic representations of observations suitable for downstream RL. We implemented separate sets of experiments to evaluate each claim. The first set of experiments are in reward free environments in which the agent is focused solely on unsupervised skill learning and, if applicable, representation learning. In this first set of experiments, we evaluate agents based on the average mutual information of their learned skillsets (i.e., how many unique skills are in their learned skillsets). The second set of experiments then implement downstream RL tasks using the learned representations and, if applicable, the skillsets learned during the first set of experiments. For the downstream RL tasks, we implement goal-conditioned RL (GCRL) tasks in which the agent is tasked with achieving a wide range of observations. Agents that learn sufficient statistic representations of observations during the pretraining phase should be able to learn effective policies mapping the learned representation and goals to actions.

6.1 Environments

For our experiments, we implemented several domains that vary along observation dimensionality and stochasticity but all have low-dimensional underlying state spaces. We focus on these simpler domains for two reasons. The first reason is that in simple domains it is easy to visualize whether the mutual information maximizing skill discovery algorithm is actually working and learning skills that target most of the reachable observations. Most existing skill discovery work does not evaluate in these settings and strictly applies their approaches to much larger domains like the Ant or Humanoid domains in MuJoCo (Todorov et al., 2012), in which it is difficult to visualize whether the agent is learning skills that target most combinations of torso and joint positions and velocities. Existing unsupervised skill discovery work also does not report the mutual information of their learned skillsets (Eysenbach et al., 2019; Zheng et al., 2025; Laskin et al., 2022) so it is unclear how well these algo-

rithms are working. The second reason we selected settings with lower dimensional underlying state spaces is to save on cost as our approach is compute intensive. Larger underlying state spaces can mean parameter vectors θ_z and f_c with more dimensions, which means more variational posteriors need to be trained in parallel.

We implemented the following six settings for the first set of experiments. The first setting was a simple two-dimensional square room with a two-dimensional observation space and a two-dimensional continuous action space. The second setting was a stochastic version of the first setting, in which two extra dimensions are added to the observation and these two dimensions are randomly sampled from the range $[-1, 1]$. The remaining four settings have high-dimensional observations that consist of 32x32 grayscale images (1,024 dimensions). The first of these settings is again a two-dimensional room in which the room is black and agent is white. The second high-dimensional settings is a stochastic version of the previous setting in which darker background pixels are random sampled from a range of black to gray colors. The third high-dimensional setting is a "plus" shaped intersection of a horizontal and vertical hallways. The final high-dimensional setting is a pushing task where the agent can move around an object if the object is within a certain distance. Figure 6 shows sample image observations from the high-dimensional settings. In all settings, the initial observation can be mostly anywhere in the environment. The number of primitive actions in each skill $n = 7$ for all tasks. Section G details the key hyperparameters for our approach in all settings. For the second set of experiments implementing the GCRL tasks, we used all the high-dimensional settings except for the push task.

6.2 Baselines

In the first set of experiments, we compare our full approach that jointly performs representation learning and skill discovery to six other existing algorithms, including three from prior work and three ablations of our approach. The three algorithms from prior work we compare to are the explicit version of Variational Intrinsic Control (VIC) (Gregor et al., 2016), Diversity Is All You Need (DIAYN) (Eysenbach et al., 2019), and Contrastive Successor Features (CSF) (Zheng et al., 2025). The main differences between these approaches and our approach is the learnable action space and how the posterior is trained. Instead of treating the skill-conditioned policy as the learnable action space as in our bandit RL approach, these treat the primitive action space as the trainable action space. In addition, instead of conditioning the posterior on the proposed skillset to achieving a tighter mutual information lower bound, these approach do not condition on the proposed skillset. VIC differs from DIAYN by using the skill-terminating observation in the mutual information term, while DIAYN samples observation from the entire skill trajectory. CSF differs from VIC and DIAYN by training the posterior using a contrastive lower bound on mutual information. In addition, CSF trains the skill-conditioned policy using a modified version of mutual information that subtracts an “anti-exploration" term. Note that CSF is a recent approach that reports state of the art results and is a mutual information-based version of METRA (Park et al., 2024), which is another recent leading approach. Moreover, the focus of these baselines is on unsupervised skill discovery and not on representation learning for downstream tasks, in contrast to our approach.

The three ablations of our approach that we compare against include (i) our approach without representation learning (i.e., the observation encoder is an identity function: $f_c(o_0) = o_0$), (ii) our approach but we do not condition the variational posterior on the skill-conditioned policy as in prior work, and (iii) our approach but we fix the observation encoder. (Note that we only implement (i) for the two low-dimensional observation settings as some representation learning is needed for the high-dimensional settings.) We compare to (i) because per Theorem 1, if our approach is working as expected the average empowerment of a learned representation should be close to the average empowerment of a sufficient statistic representation and in the low-dimensional settings the observation is a sufficient statistic. We compare to (ii) in order to evaluate the effect of training skill-conditioned policies using a loose lower bound on mutual information. The comparison to VIC also accomplishes this but (ii) does not have the non-stationary reward issue because the skill-conditioned policy is used as the trainable action space. We compare to (iii) to show the importance of training

the observation encoder with empowerment rather than simply using a randomly initialized function to encode observations.

In the second set of experiments, we implement four algorithms. One algorithm learns a goal-conditioned policy outputting primitive actions conditioned on a learned representation from the first phase of experiments. The second algorithm learns a goal-conditioned policy that outputs skills using the learned representation and skillsets learned during the first phase. The third algorithm trains a goal-conditioned policy outputting primitive actions using the representation from a fixed observation encoder. The fourth algorithm learns a goal-conditioned policy outputting primitive actions directly from pixels (i.e., does not use the observation encoder from the first phase).

6.3 Results

Table 1: Average ($\pm$std) variational mutual information of learned skillsets (nats)

Algorithm	**2D**	**Noisy 2D**	**Gray**	**Noisy Gray**	**Plus**	**Push**
Ours	8.0 ± 0.0	7.6 ± 0.1	5.7 ± 0.3	4.7 ± 0.3	4.5 ± 0.1	6.4 ± 0.4
VIC	4.1 ± 1.3	4.4 ± 1.0	0.3 ± 0.6	0.5 ± 0.5	0.5 ± 0.6	-0.1 ± 0.6
DIAYN	-0.4 ± 0.0	-0.4 ± 0.0	-0.4 ± 0.1	-0.4 ± 0.0	-0.3 ± 0.0	-0.7 ± 0.0
CSF	-0.4 ± 0.7	-0.6 ± 0.2	0.3 ± 0.9	-0.2 ± 0.4	-0.6 ± 0.3	0.1 ± 0.2
No Abs	7.7 ± 0.3	4.6 ± 0.8	N/A	N/A	N/A	N/A
Fixed Abs	7.5 ± 0.5	4.4 ± 0.7	2.4 ± 0.2	1.9 ± 0.2	2.4 ± 0.1	3.6 ± 0.4
Loose Bound	4.1 ± 0.8	3.6 ± 0.3	2.1 ± 0.7	2.1 ± 0.3	2.0 ± 0.3	2.8 ± 0.5

Table shows the variational mutual information results for all algorithms in all settings in the first set of experiments. Note that (i) the mutual information is shown in the logarithmic units of nats (e.g., in the 2D room domain, the agent learns 8.0 nats of skills or $\approx 2,980$ skills) and (ii) variational mutual information can be negative if it is a loose lower bound on mutual information. The results show strong across-the-board outperformance by our approach. Relative to the approaches that used loose lower bounds on mutual information to evaluate skill-conditioned policies π_z (i.e., VIC, DIAYN, CSF, and Loose Bound, which is the ablation that trains a variational posterior not conditioned on π_z), our approach learns far larger skillsets. For instance, the best performance of these approaches was by VIC and Loose Bound in the low-dimensional tasks where our approach still learned 3.9 more nats of skills (i.e., 49x more skills) and 3.2 more nats of skills (25x more skills) in the 2D and Noisy 2D domains, respectively. Relative to the ablation that uses a fixed observation encoder (i.e., Fixed Abs), our approach learned far larger skillsets except for the simplest low-dimensional setting where there was smaller outperformance, showing that training the observation encoder with empowerment performs better than using a randomly initialized function to encode observations. Interestingly, our approach also outperformed the ablation in the low-dimensional settings that simply used the low-dimensional observation as the policy input, which in theory should serve as an upper bound for our approach. We believe our approach performed better in practice because in domains such as the Noisy 2D room in which different observations can be close in the observation space but need to support different skill-conditioned policies, it is helpful to learn representations that separate these observations in order to output different θ_z. Further, Figures 7, 8, and 9 provide the learning curves for the first set of experiments, showing that our approach learns efficiently. For instance, in the low-dimensional tasks our approach can learn thousands of skills in around 1000 gradient steps to the two actors, while the image domains required around 3000 gradient steps for agents to reach their peak performance.

Qualitatively, the agents learn large distinct skillsets that target large portions of the reachable observation space. Figure 2 and section I provide various visuals showing the diverse skillsets that are learned.

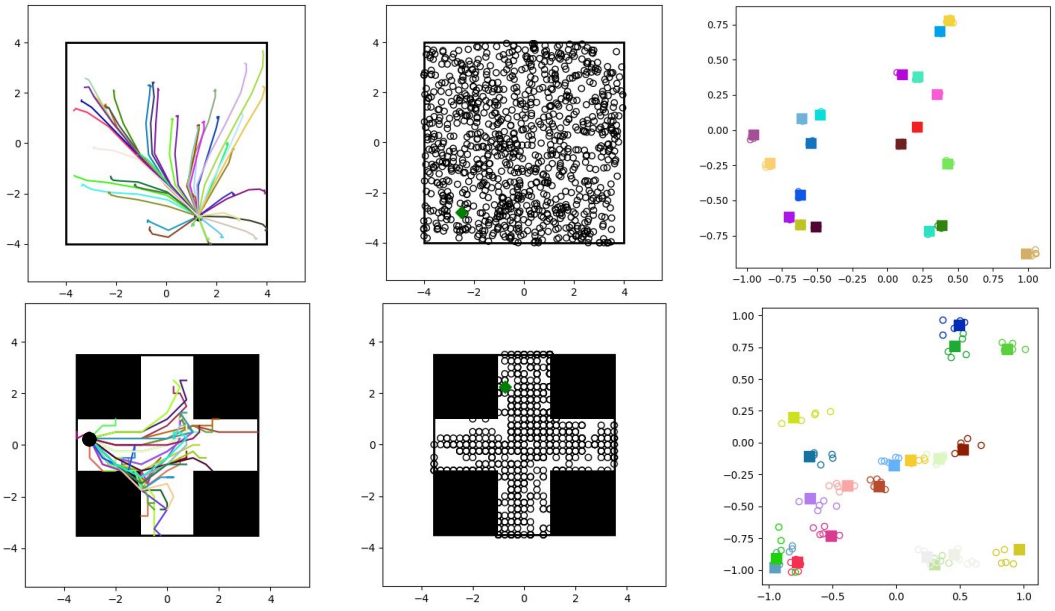

Figure 2: Some qualitative results from the 2D domain (top row) and Plus Intersection domain (bottom row). The left column shows the trajectories from a single starting observation produced by 45 randomly sampled skills. The center column shows the skill-terminating (x, y) positions from 1000 randomly sampled skills when starting at the green marker. The right column shows 20 randomly sampled skills (squares), and for each skill, 5 samples (circles) from the variational posterior $q_\psi(z|c_0, \pi_z, o_n)$. The large state space coverage and tight variational posterior around each skill shows the agents is learning large, diverse skillsets.

In addition to learning large skillsets, the second set of experiments provide evidence that our approach can learn sufficient statistics of observations as the theory suggests. Section J provides the learning curves for the second set of experiments, and section K provides visuals of the goal-conditioned trajectories. Per Figure 14, both algorithms that used the representations learned during the first phase of experiments were able to learn effective goal-conditioned policies as would be expected from an approach that learned a sufficient statistic representation. The hierarchical policy was able to learn with the best sample efficiency, consistent with previous hierarchical RL work (Levy et al., 2019; McClinton et al., 2021). In addition, we observed that the algorithm that used representations from a randomly initialized observation encoder failed at all tasks, providing additional evidence that empowerment is more effective at learning representations suitable for reinforcement learning than some randomly initialized function.

7 Conclusion

Representation learning and unsupervised skill discovery remain two important problems for reinforcement learning agents. Through theoretical analysis and experimentation, we show that the empowerment objective provides a potential solution for both problems. Future work should try to extend our results to partially observable settings with non-Markov observations and integrate model-based empowerment approaches so that a high-dimensional simulator of the environment is not needed.

Appendix

This section provides the proof for Theorem 1.

Proof.

$$\begin{aligned}
\mathbb{E}_{c_0 \sim p(c_0|f_c)}[\mathcal{E}(c_0, f_c)] &= \mathbb{E}_{c_0 \sim p(c_0|f_c)}[I(Z; O_n|c_0, f_c, \pi_z^{c,*})] && (8)\\
&\leq \mathbb{E}_{c_0 \sim p(c_0|f_c), x_0 \sim p(x_0|c_0, f_c)}[I(Z; O_n|c_0, x_0, f_c, \pi_z^{c,*})] && (9)\\
&\leq \mathbb{E}_{x_0 \sim p(x_0|f_x)}[I(Z; O_n|x_0, f_x, \pi_z^{x})] && (10)\\
&\leq \mathbb{E}_{x_0 \sim p(x_0|f_x)}[I(Z; O_n|x_0, f_x, \pi_z^{x,*})] && (11)\\
&= \mathbb{E}_{x_0 \sim p(x_0|f_x)}[\mathcal{E}(x_0, f_x)] && (12)
\end{aligned}$$

□

Line 8 inserts the definition of the empowerment of a context c_0 and observation encoder f_c. Line 9 uses the fact that mutual information is convex with respect to the channel distribution (Cover & Thomas, 2006). That is, if the channel distribution is a weighted average of other channels, then the mutual information of the mixed channel is upper bounded by the weighted average of the mutual information of the individual channels. In this case, the mixed channel is $p(o_n|c_0, f_c, \pi_z^{c,*}, z)$ and the individual channels are $p(o_n|c_0, x_0, f_c, \pi_z^{c,*}, z)$ (i.e., include the aliased sufficient statistic x_0) and are weighted by $p(x_0|c_0, f_c)$.

The purpose of line 10 is to replace each mutual information $I(Z; O_n|c_0, x_0, f_c, \pi_z^{c,*})$ with an equivalent mutual information term that removes c_0 from the conditioning variables and replaces f_c with the sufficient statistic observation encoder f_x. This is done by first swapping the skill-conditioned policy $\pi_z^{c,*}$ with a particular skill-conditioned policy π_z^x, which uses the same distribution over actions as $\pi_z^{c,*}$ when in representation x_t at time t while pursuing skill z. That is, $p(a_t|x_0, f_x, x_t, z) = p(a_t|c_0, x_0, f_c, x_t, z)$, in which $p(a_t|c_0, x_0, f_c, x_t, z)$ is the marginal of the joint distribution $p(c_t, a_t|c_0, x_0, f_c, x_t, z)$. (Note that as long as the distribution $p(a_t|c_0, x_0, f_c, x_t, z) = p(a_t|c_0, x_0, f_c, x_{t'}, z)$ for all a_t when $x_t = x_{t'}$ and $t' > t$, then π_z^x can remain a stationary policy that does not need to take an extra t as input. If this is not the case, π_z^x will also need to take the time step t as input to avoid the conflict of mapping the same policy to two different policy distributions). With π_z^x, we can show that for any (c_0, x_0, z), the original channel distribution $p(o_t|c_0, x_0, f_c, \pi_z^{c,*}, z)$ equals the channel distribution $p(o_t|x_0, f_x, \pi_z^x, z)$ for any step $t = 1, \ldots, n$. These marginal distributions are equal because the joint distributions are equal: $p(x_{t-1}, a_{t-1}, o_t, x_t|x_0, f_x, z) = p(x_{t-1}, a_{t-1}, o_t, x_t|c_0, x_0, f_c, z)$ for $t = 1, \ldots, n$. The joint distributions are true because (i) the marginals over the prior sufficient statistics $p(x_{t-1}|x_0, f_x, z) = p(x_{t-1}|c_0, x_0, f_c, z)$ for $t = 1, \ldots, n$, (ii) the policies are the same by definition: $p(a_{t-1}|x_0, f_x, z, x_{t-1}) = p(a_{t-1}|c_0, x_0, f_c, z, x_{t-1})$, and (iii) the distribution over the next observation and sufficient statistic $p(o_t, x_t|x_0, f_x, z, x_{t-1}, a_{t-1}) = p(o_{t+1}, x_{t+1}|c_0, x_0, f_c, z, x_{t-1}, a_{t-1})$ because these only depend on x_{t-1} and a_{t-1}. (i) is true because (a) it is true at $t = 0$ because $p(x_0|x_0, f_x) = p(x_0|c_0, x_0, f_c)$ as x_0 is a conditioning variable in both and (b) it is true for $t = 1, \ldots, n-1$ because the joint distributions $p(x_{t-1}, a_{t-1}, o_t, x_t|x_0, f_x, z) = p(x_{t-1}, a_{t-1}, o_t, x_t|c_0, x_0, f_c, z)$. The reason there is an inequality instead of an equality in line 10 is that if there are multiple $I(Z; O_n|c_0, x_0, f_z, \pi_z^{c,*})$ terms with different c_0 terms but the same x_0 (i.e., the same sufficient statistic x_0 is associated with different contexts c_0). In this case, if the mutual information is not equal for all terms, the largest $I(Z; O_n|x_0, f_x, \pi_z^x)$ can be used in place of the rest and the inequality in line 10 becomes a strictly less than.

In line 11, the skill-conditioned policy π_z^x is replaced with the mutual information maximizing policy $\pi_z^{x,*}$ for starting representation x_0 and encoder f_x. The inequality becomes a strictly less than if $\pi_z^{x,*}$ differs from π_z^x. The final line uses the definition of the empowerment of a context representation. Section C shows the same proof can be used to show the average empowerment of a sufficient statistic encoder is upper bounded by the average empowerment of states.

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.

David Barber and Felix Agakov. Information maximization in noisy channels : A variational approach. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/a6ea8471c120fe8cc35a2954c9b9c595-Paper.pdf.

Kate Baumli, David Warde-Farley, Steven Hansen, and Volodymyr Mnih. Relative variational intrinsic control. In *AAAI*, pp. 6732–6740. AAAI Press, 2021. ISBN 978-1-57735-866-4. URL http://dblp.uni-trier.de/db/conf/aaai/aaai2021.html#BaumliWHM21.

Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, and Sergey Levine. Information prioritization through empowerment in visual model-based RL. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=DfUjyyRW90.

Víctor Campos, Alex Trott, Caiming Xiong, Richard Socher, Xavier Giro-i Nieto, and Jordi Torres. Explore, discover and learn: unsupervised discovery of state-covering skills. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

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-conditioned reinforcement learning. 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. 1953–1963. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/choi21b.html.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954.

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.

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

Steven Hansen, Will Dabney, Andre Barreto, David Warde-Farley, Tom Van de Wiele, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJeAHkrYDS.

Riashat Islam, Manan Tomar, Alex Lamb, Hongyu Zang, Yonathan Efroni, Dipendra Misra, Aniket Rajiv Didolkar, Xin Li, Harm van Seijen, Remi Tachet des Combes, and John Langford. Agent-controller representations: Principled offline RL with rich exogenous information. In *3rd Offline RL Workshop: Offline RL as a "Launchpad"*, 2022. URL https://openreview.net/forum?id=0pFzg-8y-o.

Seongun Kim, Kyowoon Lee, and Jaesik Choi. Variational curriculum reinforcement learning for unsupervised discovery of skills. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16668–16695. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kim23n.html.

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.

Anurag Koul, Shivakanth Sujit, Shaoru Chen, Ben Evans, Lili Wu, Byron Xu, Rajan Chari, Riashat Islam, Raihan Seraj, Yonathan Efroni, Lekan Molu, Miro Dudik, John Langford, and Alex Lamb. Pclast: Discovering plannable continuous latent states, 2024. URL https://arxiv.org/abs/2311.03534.

Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Rajiv Didolkar, Dipendra Misra, Dylan J Foster, Lekan P Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=TNocbXm5MZ.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5639–5650. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/laskin20a.html.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 34478–34491. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/debf482a7dbdc401f9052dbe15702837-Paper-Conference.pdf.

Andrew Levy, Robert Platt, and Kate Saenko. Hierarchical reinforcement learning with hindsight. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryzECoAcY7.

Andrew Levy, Sreehari Rammohan, Alessandro Allievi, Scott Niekum, and George Konidaris. Hierarchical empowerment: Towards tractable empowerment-based skill learning, 2023. URL https://arxiv.org/abs/2307.02728.

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.

Shakir Mohamed and Danilo J. Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2125–2133, Cambridge, MA, USA, 2015. MIT Press.

Ofir Nachum, Haoran Tang, Xingyu Lu, Shixiang Gu, Honglak Lee, and Sergey Levine. Why does hierarchy (sometimes) work so well in reinforcement learning? *CoRR*, abs/1909.10618, 2019. URL http://arxiv.org/abs/1909.10618.

Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/7ec69dd44416c46745f6edd947b470cd-Paper.pdf.

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BGvt0ghNgA.

Seohong Park, Oleh Rybkin, and Sergey Levine. METRA: Scalable unsupervised RL with metric-aware abstraction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=c5pwL0Soay.

Vitchyr Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7783–7792. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/pong20a.html.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

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.

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.

DJ Strouse, Kate Baumli, David Warde-Farley, Volodymyr Mnih, and Steven Stenberg Hansen. Learning more skills through optimistic exploration. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cU8rknuhxc.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, pp. 5026–5033. IEEE, 2012. ISBN 978-1-4673-1737-5. URL http://dblp.uni-trier.de/db/conf/iros/iros2012.html#TodorovET12.

David Warde-Farley, Tom Van de Wiele, Tejas Kulkarni, Catalin Ionescu, Steven Hansen, and Volodymyr Mnih. Unsupervised control through non-parametric discriminative rewards. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1eVMnA9K7.

Jesse Zhang, Haonan Yu, and Wei Xu. Hierarchical reinforcement learning by discovering intrinsic options. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=r-gPPHEjpmw.

Chongyi Zheng, Jens Tuyls, Joanne Peng, and Benjamin Eysenbach. Can a MISL fly? analysis and ingredients for mutual information skill learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xoIeVdFO7U.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Proof of Empowerment Objective Lower Bound

Below is a proof that $I(Z; O_n|o_0, \pi_z) \leq I(\Pi; O_n|o_0)$.

In the below proof, let Z be the skill random variable with $z \sim p(z|o_0)$. π_z is a skill-conditioned policy in which $p(\pi_z|z) = p(\pi_z) = 1$.

$$I(\Pi; O_n|o_0) \geq I(Z, \pi_z; O_n|o_0) \tag{13}$$
$$= I(Z; O_n|o_0, \pi_z) \tag{14}$$

In line 13, the Data Processing Inequality (Cover & Thomas, 2006) is used because given o_0, $Z, \pi_z \rightarrow \Pi \rightarrow O_n$ form a Markov chain. This is true because the combination of a skill z and a skill-conditioned policy π produces some policy π that maps from observations to actions. Then, given o_0 and π, the distribution over the terminating observation o_n is conditionally independent of z and π_z as none of the intermediate states, actions, and observation depend on these quantities. In line 14, the skill-conditioned policy π_z has been moved to the list of conditioned variables given that $p(\pi_z) = 1$.

B Sufficient Statistic Representations and RL

Sufficient statistic representations of observations are critical to using reinforcement learning in a learned representation space because they enable agents to replace potentially high-dimensional observations as a policy input with more compact representations as discussed in section of the supplementary materials. This is because the distribution over future rewards given a sufficient statistic, action, encoding function, and policy is the same as the distribution over future rewards in which the original observation replaces the sufficient statistic (assuming rewards are functions of observations): $p(r_{t+1}, r_{t+1}, \ldots, r_{t+N}|x_t, a_t, f_x, \pi) = p(r_{t+1}, r_{t+1}, \ldots, r_{t+N}|x_t, o_t, a_t, f_x, \pi) = p(r_{t+1}, r_{t+1}, \ldots, r_{t+N}|o_t, a_t, f_x, \pi)$. This is because no future reward requires knowing o_t when sufficient statistic x_t is known. Equality in these distributions in turn means that the Q-values $Q(o_t, a_t) = Q(x_t, a_t)$ for all $(o_t, x_t = f_x(o_t), a_t)$ tuples are equal, which is why observations can be replaced by sufficient statistic representations.

C Proof that Empowerment of States Upper Bounds Empowerment of Sufficient Statistics

In this section, we prove that the average empowerment produced by a sufficient statistic encoder, $\mathbb{E}_{x_0 \sim p(x_0|f_x)}[\mathcal{E}(x_0, f_x)]$, is upper bounded by the average empowerment of state representations. This is the same proof as in 7 except the initial context variable c_0 is replaced with the initial sufficient statistic variable x_0 and the state representations s_t replace the sufficient statistic representations x_t. Note that this extends the prior work of (Capdepuy, 2011) which only considered the empowerment objective in which the mutual information was between open loop actions and observations.

Proof.

$$\begin{aligned}
\mathbb{E}_{x_0 \sim p(x_0|f_x)}[\mathcal{E}(x_0, f_x)] &= \mathbb{E}_{x_0 \sim p(x_0|f_c)}[I(Z; O_n|x_0, f_x, \pi_z^{x,*})] && (15)\\
&\leq \mathbb{E}_{x_0 \sim p(x_0|f_x), s_0 \sim p(s_0|x_0, f_x)}[I(Z; O_n|x_0, s_0, f_x, \pi_z^{x,*})] && (16)\\
&\leq \mathbb{E}_{s_0 \sim p(s_0)}[I(Z; O_n|s_0, \pi_z^{s})] && (17)\\
&\leq \mathbb{E}_{s_0 \sim p(s_0)}[I(Z; O_n|s_0, \pi_z^{s,*})] && (18)\\
&= \mathbb{E}_{s_0 \sim p(s_0)}[\mathcal{E}(s_0)] && (19)
\end{aligned}$$

□

Line 15 inserts the definition of the empowerment of a sufficient statistic x_0 and sufficient statistic encoder f_x. $\pi_z^{x,*}$ is the mutual information maximizing skill-conditioned policy. This proof will assume $\pi_z^{x,*}$ is a non-stationary policy that takes sufficient statistics, skills, and the step number (e.g., $0, 1, \ldots, n-1$) as input and outputs primitive actions.

Line 16 uses the fact that mutual information is convex with respect to the channel distribution (Cover & Thomas, 2006). That is, if the channel distribution is a weighted average of other channels, than the mutual information of the mixed channel is upper bounded by the weighted average of the mutual information of the individual channels. In this case, the mixed channel is $p(o_n|x_0, f_x, \pi_z^{x,*}, z)$ and the individual channels are $p(o_n|x_0, s_0, f_x, \pi_z^{x,*}, z)$ (i.e., include the state s_0) and are weighted by $p(s_0|x_0, f_x)$.

The purpose of line 17 is to replace each mutual information $I(Z; O_n|x_0, s_0, f_x, \pi_z^{x,*})$ with an equivalent mutual information term that removes x_0 and f_x from the conditioning variables. This is done by first swapping the skill-conditioned policy $\pi_z^{x,*}$ with a particular skill-conditioned policy π_z^s, which uses the same distribution over actions as $\pi_z^{x,*}$ when in state s_t at time t while pursuing skill z. That is, $p(a_t|s_0, s_t, t, z) = p(a_t|x_0, s_0, f_x, s_t, t, z)$, in which $p(a_t|x_0, s_0, f_x, x_t, t, z)$ is the marginal of the joint distribution $p(x_t, a_t|x_0, s_0, f_x, s_t, t, z)$. With π_z^s, we can show that for any (x_0, s_0, z), the original channel distribution $p(o_t|x_0, s_0, f_x, \pi_z^{x,*}, z)$ equals the channel distribution $p(o_t|s_0, \pi_z^s, z)$ for any step $t = 1, \ldots, n$. These marginal distributions are equal because the joint distributions are equal: $p(s_{t-1}, a_{t-1}, o_t, s_t|s_0, z) = p(s_{t-1}, a_{t-1}, o_t, s_t|x_0, s_0, f_x, z)$ for $t = 1, \ldots, n$. The joint distributions are true because (i) the marginals over the prior states $p(s_{t-1}|s_0, z) = p(s_{t-1}|x_0, s_0, f_x, z)$ for $t = 1, \ldots, n$, (ii) the policies are the same by definition: $p(a_{t-1}|s_0, s_{t-1}, z) = p(a_{t-1}|c_0, x_0, f_c, x_{t-1}, z)$, and (iii) the distribution over the next observation and state $p(o_t, s_t|s_0, s_{t-1}, a_{t-1}) = p(o_{t+1}, s_{t+1}|x_0, s_0, f_x, s_{t-1}, a_{t-1})$ because these only depend on s_{t-1} and a_{t-1}. (i) is true because (a) it is true at $t = 0$ $p(s_0|s_0) = p(s_0|x_0, s_0, f_c)$ as s_0 is a conditioning variable in both and (b) it is true for $t = 1, \ldots, n-1$ because the joint distributions $p(s_{t-1}, a_{t-1}, o_t, s_t|s_0, z) = p(s_{t-1}, a_{t-1}, o_t, s_t|x_0, s_0, f_x, z)$. The reason there is an inequality instead of an equality in line 17 is that if there are multiple $I(Z; O_n|x_0, s_0, f_x, \pi_z^{x,*})$ terms with different x_0 terms but the same s_0 (i.e., the same state s_0 is associated with different sufficient statistics x_0). In this case, if the mutual information is not equal for all terms, the largest $I(Z; O_n|s_0, \pi_z^x)$ can be used in place of the rest and the inequality in line 17 becomes a strictly less than.

In line 18, the skill-conditioned policy π_z^s is replaced with the mutual information maximizing policy $\pi_z^{s,*}$ for starting representation s_0. The inequality becomes a strictly less than if $\pi_z^{s,*}$ differs from π_z^s. The final line uses the definition of the empowerment of a state.

D Gradient of 1-Hidden Layer Critic w.r.t. Actor

In this section we derive the gradient of a 1-hidden layer MLP critic $Q_\eta(o_0, \theta_z = f_\lambda(o_0))$ with respect to some parameter λ_j in the bandit policy actor $f_\lambda(o_0)$. The critic will take the following form, which is visualized in Figure 3. The output $Q = a(\sum_{i=1}^{|h|} \mathbf{h} W_1)$, in which $a(\cdot)$ is a nonlinear function; $\mathbf{h}$ is the hidden layer vector with $|h|$ dimensions; and $\mathbf{h}W_1$ applies matrix multiplication between vector $\mathbf{h}$ and weight matrix W_1. Next, each entry $h_i \in \mathbf{h}$ is defined $h_i = a(\sum_{i=1}^{|\theta_z|} \theta_z W_0)$.

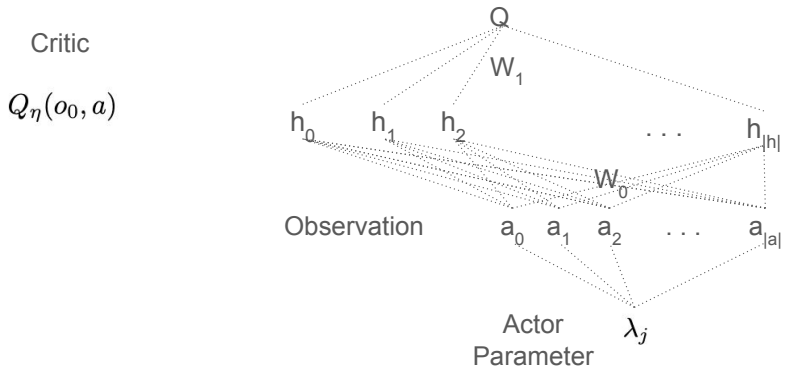

Figure 3: Figure visualizes the function form of a 1 hidden layer critic. We use this visual to show that the derivative of Q with respect to a parameter λ_j of the bandit policy actor depends on the derivatives of Q with respect to the individual entries in the skill-conditioned policy vector θ_z.

Note that in this definition the connection between the observation o_0 and h_i are ignored because o_0 has no dependence on the parameters of the bandit policy actor λ. Lastly, each entry $\theta_z^i \in \theta_z$ is defined $\theta_z^i = f(\lambda_j, o_0, \lambda_{/j})$. That is, each entry in θ_z is some function of the parameter λ_j under consideration, the initial observation o_0, and the other parameters (excluding λ_j) in λ.

With this functional form,

$$\begin{aligned}
\frac{dQ}{d\lambda_j} &= \frac{dQ}{d(\sum_{i=0}^{|h|-1} \mathbf{h}W_1)} \Big(\sum_{i=0}^{|h|-1} \frac{d(\sum_{i=0}^{|h|-1} \mathbf{h}W_1)}{dh_i} \frac{dh_i}{d(\sum_{k=0}^{|\theta_z|-1} \theta_z W_0)} \Big(\sum_{k=0}^{|\theta_z|-1} \frac{d(\sum_{k=0}^{|\theta_z|-1} \theta_z W_0)}{d\theta_z^k} \frac{d\theta_z^k}{d\lambda_j} \Big) \Big) \\
&= \sum_{k=0}^{|\theta_z|-1} \frac{d\theta_z^k}{d\lambda_j} \Big(\sum_{i=0}^{|h|} \frac{dQ}{d(\sum_{i=0}^{|h|-1} \mathbf{h}W_1)} \frac{d(\sum_{i=0}^{|h|-1} \mathbf{h}W_1)}{dh_i} \frac{dh_i}{d(\sum_{k=0}^{|\theta_z|-1} \theta_z W_0)} \frac{d(\sum_{k=0}^{|\theta_z|-1} \theta_z W_0)}{d\theta_z^k} \Big) \\
&= \sum_{k=0}^{|\theta_z|-1} \frac{dQ}{d\theta_z^k} \frac{d\theta_z^k}{d\lambda_j} \qquad (20)
\end{aligned}$$

Thus, the gradient of Q with respect to each parameter of the bandit policy actor depends on the gradients of Q with respect to each of the entries in θ_z (i.e., $\frac{dQ}{d\theta_z^k}$ for $k = 0, \ldots, |\theta_z| - 1$). Our approach uses this fact when simulating the gradient of this actor-critic using a new parameter-specific actor-critic architecture.

E Visualization of New Actor-Critic Architectures

Figure 1 visualizes how the parameter-specific critics attach to the bandit actor that outputs the parameters of the skill-conditioned policy.

Figure 5 visualizes how the parameter-specific critics attach to the bandit actor that outputs the parameters of the observation encoder.

F Environment Sample Observations

Figure 6 provides sample image observations from each of the high-dimensional tasks.

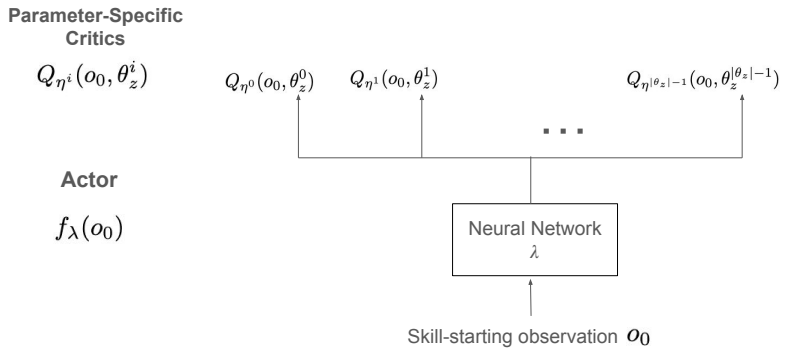

Figure 4: Visual of how the parameter-specific critics attach to the actor. In this case, the actor maps observations to the parameters of the skill-conditioned policy $\theta_z = [\theta_z^0, \theta_z^1, \ldots, \theta_z^{|\theta_z|-1}]$. For each dimension in θ_z, there is a critic $Q_{\eta^i}(o_0, \theta_z^i)$ that approximates the variational mutual information of executing the skill-conditioned policy θ_z^i from observation o_0. θ_z^i is a scalar representing the skill-conditioned policy, in which all parameters $j \neq i$ take on the greedy value from the actor (i.e., $f_\lambda(o_0)[j]$), while the i-th parameter takes on value θ_z^i.

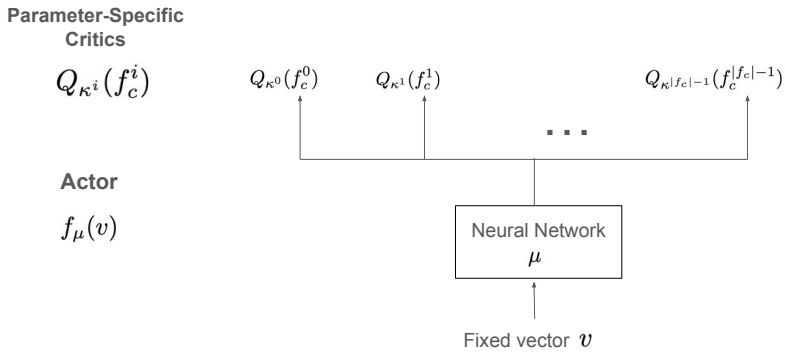

Figure 5: In this case, the actor maps a fixed vector v to the parameters of the observation encoder $f_c = [f_c^0, f_c^1, \ldots, f_c^{|f_c|-1}]$. For each dimension in f_c, there is a critic $Q_{\kappa^i}(f_c^i)$ that approximates the average variational mutual information $\mathbb{E}_{c_0 \sim p(c_0|f_c^i)[I^V(Z;O_n|c_0,f_c^i)]}$ of using the observation encoder f_c^i from context $c_0 \sim p(c_0|f_c^i)$.

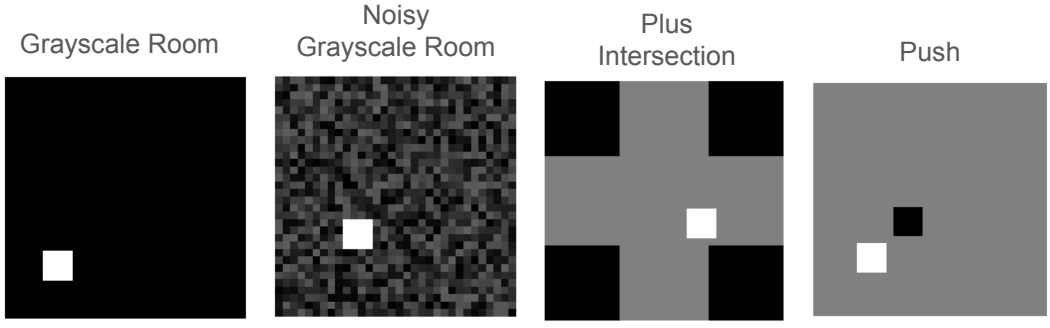

Figure 6: Sample image observations from each of the four high-dimensional settings.

Table 2: Environment-dependent Hyperparameters

Hyperparameter	2D	Noisy 2D	Gray	Noisy Gray	Plus	Pick-and-Place
Context Dim	5	5	5	5	7	7
Skill Dim	2	2	2	2	2	4
$\|\theta_z\|$	392	392	512	512	528	776
$\|f_c\|$	424	424	440	440	472	536

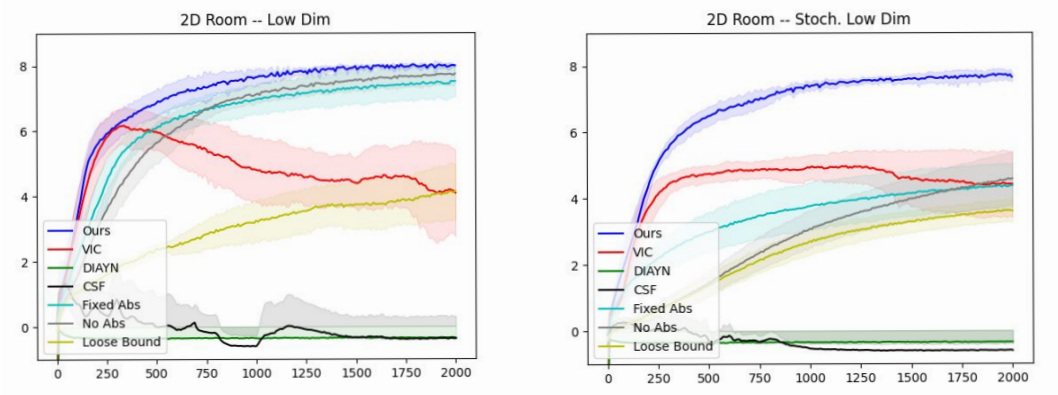

Figure 7: Learning curves for the low-dimensional tasks in the first set of experiments. The x-axis measures the number of updates to the skill-conditioned policy and observation encoder actors (i.e., the number of passes through Algorithms 1 and 2). The y-axis shows the average variational mutual information $I(Z; O_n|C_0)$.

G Hyperparameters

Figure 2 shows some of the notable domain-dependent hyperparameters including the dimension of the context space $\mathcal{C}$, dimension of the skill space $\mathcal{Z}$, the dimensionality of the skill-conditioned policy parameter vector $|\theta_z|$, and the dimensionality of the observation encode parameter vector $|f_c|$.

Other notable parameters that were used for all domains include: (i) $n = 7$, in which n the number of primitive actions contain in a skill, (ii) $M = 300$, in which M is the number of gradient updates to the variational posterior and then to the critic in Algorithms 1 and 2, (iii) learning rates of $1.5e^{-5}$ for the actors and $3e^{-4}$ for the critics and variational posteriors, and (iv) the skill-conditioned policy π_z was always implemented as a 2-hidden layer MLP with 16 neurons in each hidden layer.

H Learning Curves

Figures 7, 8, and 9 show the learning curves for the first set of experiments. The x-axis measures the number of updates to the skill-conditioned policy and observation encoder actors (i.e., the number of passes through Algorithms 1 and 2). The y-axis shows the average variational mutual information $I(Z; O_n|C_0)$.

I Additional Qualitative Results

Figures 10-13 provide qualitative results for the remaining domains. In each figure, the left image shows trajectories from 45 randomly sampled skills starting from a fixed observation. The center image shows skill-terminating (x, y) positions from 1000 randomly sampled skills when the agent

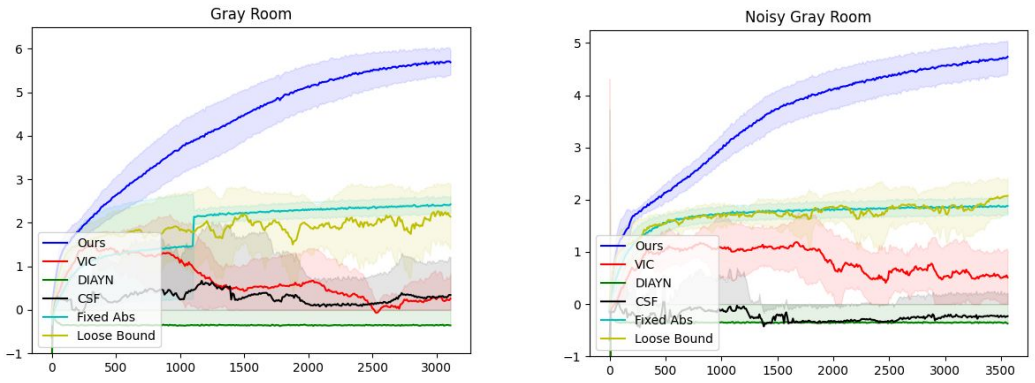

Figure 8: Learning curves for the regular and noisy grayscale rooms tasks in the first set of experiments. The x-axis measures the number of updates to the skill-conditioned policy and observation encoder actors (i.e., the number of passes through Algorithms 1 and 2). The y-axis shows the average variational mutual information $I(Z; O_n|C_0)$.

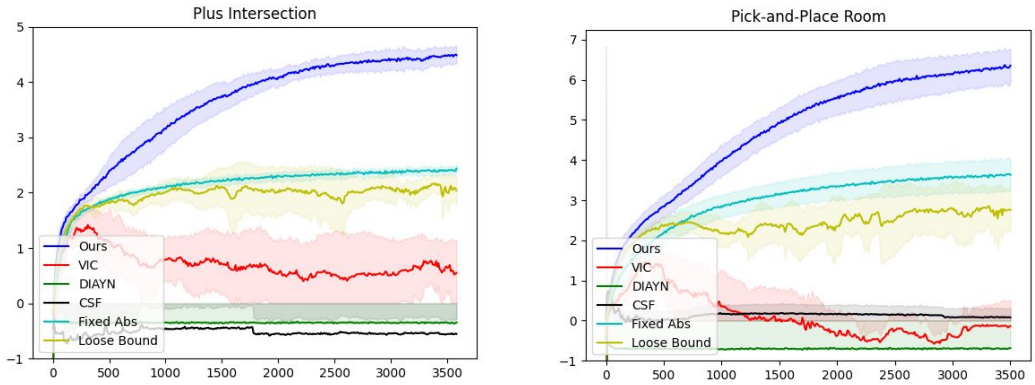

Figure 9: Learning curves for the plus intersection and push tasks in the first set of experiments. The x-axis measures the number of updates to the skill-conditioned policy and observation encoder actors (i.e., the number of passes through Algorithms 1 and 2). The y-axis shows the average variational mutual information $I(Z; O_n|C_0)$.

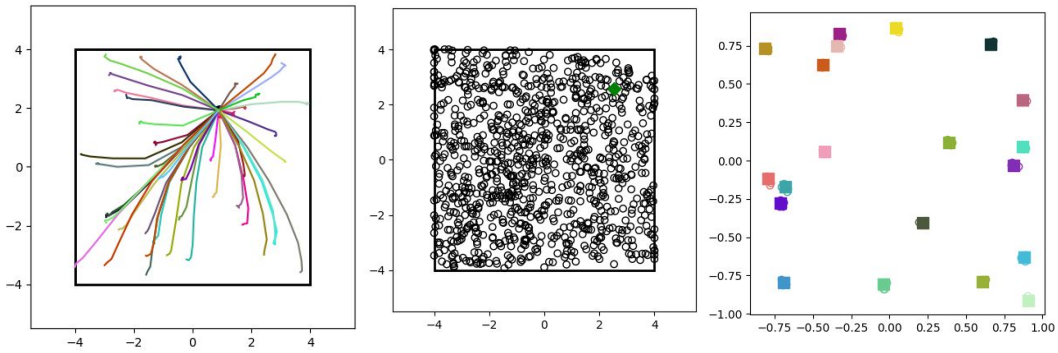

Figure 10: Qualitative results for the Noisy 2D room. Left image shows trajectories from 45 randomly sampled skills where the agent starts from the same observation. Center image shows skill-terminating (x, y) positions from 1000 randomly sampled skills when the agent starts at the green marker. Right image shows 20 skills (squares), and for each skills, 5 samples (circles) from the variational posterior $q_\psi(z|c_0, \pi_z, o_n)$.

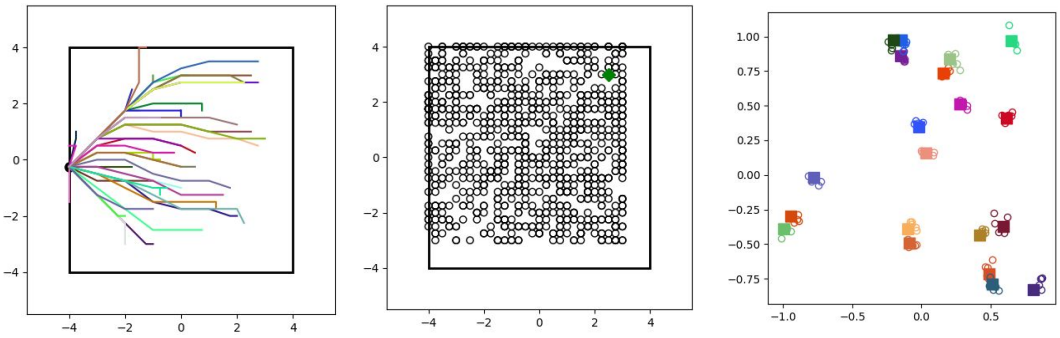

Figure 11: Qualitative results for the Noisy 2D room. Left image shows trajectories from 45 randomly sampled skills where the agent starts from the same observation. Center image shows skill-terminating (x, y) positions from 1000 randomly sampled skills when the agent starts at the green marker. Right image shows 20 skills (squares), and for each skills, 5 samples (circles) from the variational posterior $q_\psi(z|c_0, \pi_z, o_n)$.

starts at the green marker. The right image shows 20 skills (squares), and for each skills, 5 samples (circles) from the variational posterior $q_\psi(z|c_0, \pi_z, o_n)$

J Phase 2 Learning Curves

Figure 14 shows the phase 2 learning curves for the four algorithms in the three environments. The hierarchical policy should achieve lower cumulative reward as a result of the particular shortest path reward used (0 for goal achieved and -1 otherwise) and its temporally extended actions. The graphs also show that the hierarchical policy converges the fastest. The Fixed Abs algorithm in which the representation used was produced by a randomly initialized observation encoder failed at all tasks.

K Phase 2 Qualitative Results

Figures show the goal-conditioned trajectories in the Grayscale Room and Plus Intersection domains. Figure 15 shows the results for the algorithm learning a goal-conditioned policy outputting primitive actions that is conditioned on the learned representation space, while Figure 16 shows the

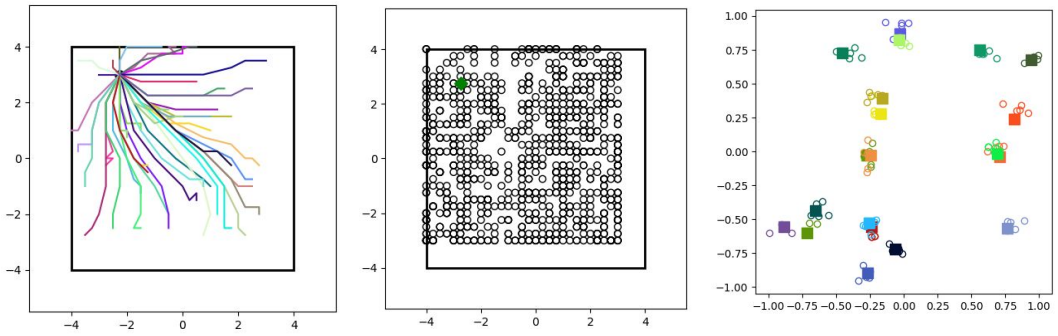

Figure 12: Qualitative results for the Noisy 2D room. Left image shows trajectories from 45 randomly sampled skills where the agent starts from the same observation. Center image shows skill-terminating (x, y) positions from 1000 randomly sampled skills when the agent starts at the green marker. Right image shows 20 skills (squares), and for each skills, 5 samples (circles) from the variational posterior $q_\psi(z|c_0, \pi_z, o_n)$.

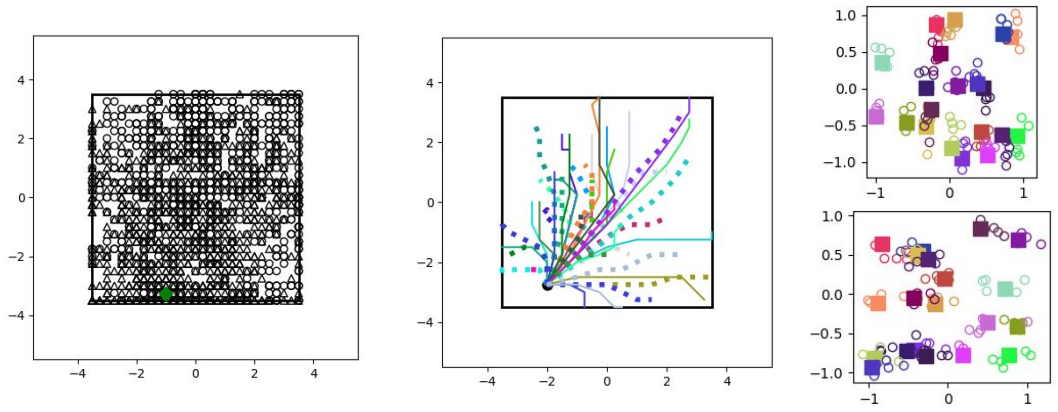

Figure 13: Qualitative results for the Noisy 2D room. Left image shows trajectories from 45 randomly sampled skills. Center image shows skill-terminating (x, y) positions from 1000 randomly sampled skills. Right image shows 20 skills (squares), and for each skills, 5 samples (circles) from the variational posterior $q_\psi(z|c_0, \pi_z, o_n)$

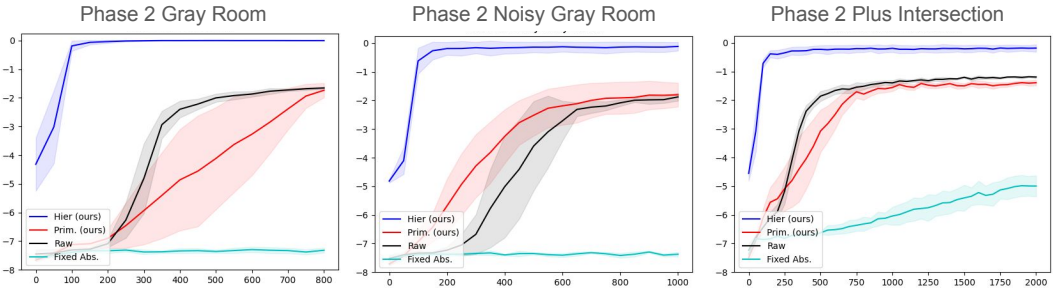

Figure 14: Learning curves for the phase 2 experiments. The x-axis shows the number of updates to the goal-conditioned policy and the y-axis shows the cumulative reward. The hierarchical policy should achieve lower cumulative reward as a result of the particular shortest path reward used and its temporally extended actions.

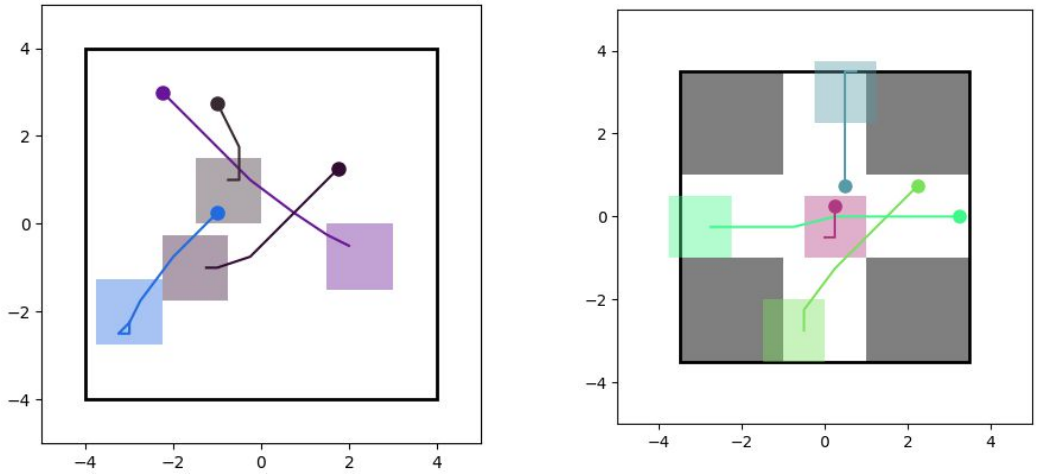

Figure 15: Phase 2 goal-conditioned trajectories for the grayscale room (Left) and Plus Intersection domains (Right) for the algorithm that learns a goal-conditioned policy outputting primitive actions and is conditioned on the learned representation space. Shaded regions are the episode goal and the line is the trajectory produced by the goal-conditioned policy.

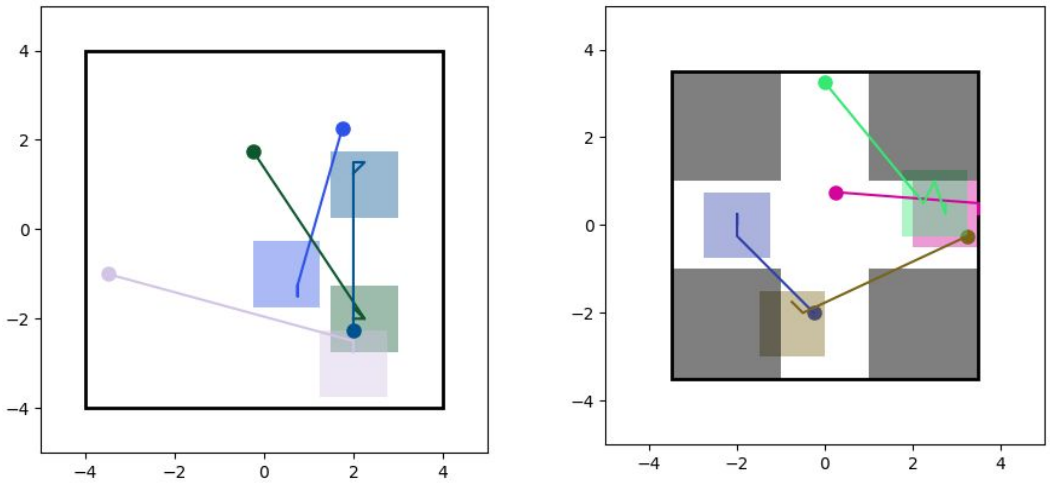

Figure 16: Phase 2 goal-conditioned trajectories for the grayscale room (Left) and Plus Intersection domains (Right) for the algorithm that learns a goal-conditioned policy outputting skills using the learned representation space and skills from pretraining. Shaded regions are the episode goal and the line is the trajectory produced by the goal-conditioned policy.

results for the hierarchical algorithm learning a goal-conditioned policy outputting skills using the learned representation space and skills from pretraining.